UNIVERSIDAD AUTÓNOMA DEL ESTADO DE HIDALGO

INSTITUTO DE CIENCIAS BÁSICAS E INGENIERÍA

ÁREA ACADÉMICA DE COMPUTACIÓN Y ELECTRÓNICA

ESTRUCTURACIÓN GLOBAL DE POBLACIONES DISTRIBUIDAS CON TÉCNICAS DE RECONOCIMIENTO DE PATRONES

TESIS

QUE PARA OBTENER EL GRADO DE DOCTOR EN CIENCIAS COMPUTACIONALES

PRESENTA:
LAURO VARGAS RUÍZ

DIRECTORAS DE TESIS
DRA. ANILU FRANCO ARCEGA
DRA. MARÍA DE LOS ÁNGELES ALONSO LAVERNIA

PACHUCA DE SOTO, HGO. MAYO 2019.

UNIVERSIDAD AUTÓNOMA DEL ESTADO
Instituto de Ciencias Básicas e Ingeniería
School of Engineering and Basic Sciences
Área Académica de Computación y Electrónica
Department of Electronics and Computer Science

150 Años de VIDA INSTITUCIONAL 1869-2019

Mineral de la Reforma Hgo., a 04 de marzo del 2019

Número de control: ICBI-AACyE/375/2019
Asunto: Autorización de impresión de tesis.

M. EN C. JULIO CÉSAR LEINES MEDÉCIGO
DIRECTOR DE ADMINISTRACIÓN ESCOLAR DE LA UAEH

Por este conducto le comunico que el comité revisor asignado al C. Lauro Vargas Ruíz, alumno del Doctorado en Ciencias Computacionales, autoriza la impresión del proyecto de tesis titulado "Estructuración global de poblaciones distribuidas con técnicas de reconocimiento de patrones" en virtud de que se han efectuado las revisiones y correcciones pertinentes.

A continuación, se registran las firmas de conformidad de los integrantes del comité revisor.

PRESIDENTE	DR. OMAR LÓPEZ ORTEGA
SECRETARIO	DRA. KARINA ALEMAN AYALA
VOCAL	DRA. ANILU FRANCO ARCEGA
SUPLENTE	DRA. MARÍA DE LOS ÁNGELES ALONSO LAVERNIA

Sin otro particular reitero a Usted la seguridad de mi atenta consideración.

Atentamente
"Amor, Orden y Progreso"

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE HIDALGO

Dr. Omar López Ortega
Coordinador del Doctorado
en Ciencias Computacionales

Instituto de Ciencias Básicas
Área Académica de Computación

Dr. Hugo Romero Trejo
Jefe del Área Académica de
Computación y Electrónica
Vo. Bo.

HRT/APL

Ciudad del Conocimiento
Carretera Pachuca-Tulancingo km 4.5 Colonia
Carboneras, Mineral de la Reforma, Hidalgo,
México. C.P. 42184
Teléfono: +52 (771) 71 720 00 ext. 2250, 2251
Fax 2109
aacye_icbi@uaeh.edu.mx

A DIOS
Por la fuerza para continuar,
cuando mi cansancio me pedía postergar

A mi amada esposa, padres, hermanos y amigos

Les dedico este trabajo con todo mi reconocimiento a su ayuda, paciencia y amor,
durante todos estos años.

Agradecimientos

*A **DIOS***
Por la oportunidad de respirar el fresco amanecer y luego, seguir descubriendo.

Gracias **Papá DIOS**, por permitirme respirar y soñar, gracias a ti es que hoy uno de esos sueños se ha materializado. Gracias **Divina Madrecita María**, por acompañarme en cada momento de mi vida, ¡Siempre conmigo a pesar de todo!

Un doctorado lo realizan simultáneamente muchas personas junto al estudiante, de otra manera se antoja imposible concluirlo exitosamente. Mi agradecimiento sincero a quienes me han acompañado siempre; especialmente a mi bella esposa **Carmina** por su comprensión, paciencia y todo el amor con el que me anima a no desmayar. Por ser mi cómplice para todo, por ser mi más dura crítica y por mostrarme con el ejemplo cotidiano el valor de la lucha y el esfuerzo. ¡Te amo!

Gracias también a mis padres **Lauro** y **Úrsula** por creer siempre en mi; soy producto de su ejemplo permanente de trabajo, dedicación, total confianza y apoyo. A cada uno de mis **hermanos** les agradezco su respaldo permanente en tiempos difíciles, su confianza y compañia. Quizá no lo saben, pero han sido un reto permanente para dar lo mejor de mí.

Agradezco en gran medida a mis directoras de tesis, **Dra. Anilu** y **Dra. María de los Ángeles**, por la exigencia para dar más, por el escrutinio tan detallado a todo el trabajo, por guiar mi viaje para descubrir y aportar, por mantener siempre el rumbo de mi investigación y ayudarme a culminar exitosamente este gran reto.

Mi sincero agradecimiento a la Universidad Autónoma del Estado de Hidalgo, por ser mi barca al navegar en el mar de la ciencia, hasta llegar felizmente al puerto

del conocimiento. Seguiré mi viaje orgulloso de ser *Garza*.

Agradezco al Instituto Tecnológico Superior del Oriente del Estado de Hidalgo, por todas las facilidades otorgadas. Otras generaciones de ingenieros y maestros recibirán los beneficios de mi preparación. **Mtra. Sonia** gracias por la gestión administrativa que propició toda esta aventura y que se convirtió en la puerta abierta para incorporarme al maravilloso mundo de la investigación. **Ing. Román**, gracias por el apoyo desde tu Jefatura de División, no habría manera de alcanzar este logro sin ese hombro amigo.

Gracias también a mis amigos, que sin conocer -y sin importar- el tema de mi tesis doctoral, siempre me animaron y confiaron en que podría terminar este gran proyecto. Ahora ya podremos compartir nuevamente una tarde juntos.

¡Infinitas gracias a todos!

Resumen

El tratamiento automático de datos es un pilar fundamental en el éxito de toda organización, ya que el almacenamiento, recuperación y procesamiento de los mismos permite la generación de información y que luego, mediante un tipo especial de técnicas y procedimientos puede generar conocimiento útil.

El procesamiento de los datos puede verse obstaculizado por diversos factores, como pueden ser entre otros, el aumento masivo en la cantidad disponible de ellos o bien, las diferentes representaciones y clasificaciones que pueden existir para un mismo objeto de acuerdo a una gran diversidad de criterios.

Esas clasificaciones y descripciones distintas que pueden existir de un mismo objeto pueden ser válidas en un contexto en particular pero en otros no. Sin embargo, puede ser valioso el contar con una descripción general de los objetos así como, de alguna clasificación global los mismos, esto facilita el obtener un punto de vista más amplio de ellos.

En esta tesis, se presenta un método de estructuración de poblaciones distribuidas, el cual incluye diversas técnicas de Reconocimiento de Patrones para identificar las características que describirán a los objetos en estudio de una forma global incluyendo su clasificación correspondiente, lo que aportará una herramienta novedosa y útil para obtener conocimiento más general de dichos objetos.

Específicamente, las técnicas de Reconocimiento de Patrones consideradas en este nuevo método incluyen la Selección de Variables, la Clasificación Supervisada y el Agrupamiento. La primera de ellas ocupa el selector Filter univariado *Relief* y se utiliza para elegir las variables que describen adecuadamente a los objetos re-

presentados en las diferentes particiones que conforman una población distribuida en estudio, creando subconjuntos de éstas. En el caso de la Clasificación Supervisada, se incluye un algoritmo identificado como *Ibk*, que sirve para validar dichos subconjuntos. En una siguiente etapa, se presenta el desarrollo y aplicación de una estrategia para determinar la inclusión de las variables que conformarán la representación general, para que mediante un agrupador reconocido como *K-means* pueda ser generada la correspondiente clasificación de la representación global.

Se incluye una etapa de validación del método de estructuración de poblaciones distribuidas con diversos casos de estudio reales, para probar su eficacia.

Palabras clave: Método, Reconocimiento de Patrones, Selección de Variables, Clasificación Supervisada, Agrupamiento.

Abstract

The automatic processing of data is a fundamental pillar in the success of any organization, the storage, recovery and processing of them allows the generation of information and then, through a special type of processing techniques can generate useful knowledge.

The processing of data can be hampered by several factors, among others, the massive increase in the available amount of them or, the different representations and classifications that can exist for the same data according to a great diversity of criteria.

Those classifications and different descriptions that may exist of the same object may be valid in a particular context but not in others. However, it can be valuable to have a general description of the objects as well as some global classification, this makes it easier to have a broader view of them.

In this thesis, a method of structuring distributed populations is presented, which includes various techniques of Pattern Recognition to identify the characteristics that will describe the objects under study in a global way including their corresponding classification, which will provide a novel tool and useful to obtain more general knowledge of such objects.

Specifically, the techniques of Pattern Recognition considered in this new method include Feature Selection, Supervised Classification and Clustering. The first of them occupies the Univariate Filter selector *Relief* that is used to choose the variables that adequately describe the objects represented in the different partitions that make up a distributed population under study, creating subsets of these. In

the case of the Supervised Classification, an algorithm identified as *Ibk* is included, which serves to validate said subsets. In a next stage, a strategy is proposed and used to determine the inclusion of the variables that will make up the general representation, so that a clusterer recognized as *K-means* can generate the corresponding classification of the global representation.

It includes a stage of validation of the method of structuring distributed populations with several real case studies, to observe its effectiveness.

Keywords: Method, Pattern Recognition, Feature Selection, Supervised Classification, Clustering.

Índice

Índice de Figuras

Índice de Tablas

Capítulo 1

Introducción

No vayas por donde el camino te lleve. Ve en cambio por donde no hay camino y deja rastro.

Ralph Waldo Emerson

Las actividades productivas del ser humano generan una cantidad importante de datos, los cuales son almacenados con la finalidad de utilizarlos posteriormente para la mejora de procesos, servicios, comportamientos, etc. Para lograr ésto, se requiere su adecuado tratamiento preferentemente por medios electrónicos, por lo que es necesaria la estructuración de los mismos a fin de facilitar su representación, comprensión e interpretación, especialmente cuando se trata de grandes volúmenes de información.

Inicialmente, las representaciones de los datos incluían una sola estructura, pero con el paso del tiempo han surgido diferentes formas de organización, las que en combinación con el uso de redes de computadora han provocado que los esquemas deban adaptarse a nuevos entornos de trabajo, específicamente a aquellos donde los datos se generan, almacenan y procesan en diversos momentos y lugares.

Es precisamente la necesidad de procesar los datos que pudieran encontrarse en distintas ubicaciones, la que da origen a este trabajo de tesis. Este tipo de tratamiento no constituye una tarea trivial, debido en gran medida a la complejidad derivada de la alta dimensional y/o por las diferencias en las distintas representaciones en que pudieran encontrarse dichos datos.

Las representaciones de los datos se corresponden o asocian con conjuntos de

objetos, a los que típicamente se les denomina *Instancias* y se describen por diversas variables o características que detallan al objeto.

Las variables referidas a los objetos se pueden clasificar en *Descriptora* y *Objetivo*, estas últimas permiten establecer alguna clasificación para los objetos y también se reconocen como variable *Clase*. Por su parte, las descriptoras ayudan a comprender el comportamiento de cada instancia, de acuerdo a los valores que toma cada una de ellas.

El aumento en el número de instancias y/o de variables puede comprometer la manipulación de los datos, ya que las técnicas de procesamiento se vuelven muy costosas computacionalmente. Para afrontar el tratamiento de grandes volúmenes de datos es común recurrir a alguna técnica de particionamiento que permita dividir los conjuntos originales de datos en diversos subconjuntos. Básicamente, es posible crear fragmentos de un conjunto de datos separando subgrupos tanto de instancias como de variables, incluso de ambas formas simultáneamente. Sin embargo, todos estos subconjuntos conforman un solo conjunto general, al que se le reconoce como *Población*.

Las poblaciones entonces se pueden encontrar (*i*) centralizadas, cuando se utiliza una sola estructura para su representación completa, o bien (*ii*) distribuidas, cuando se utilizan múltiples fragmentos, a los que también se les identifica como *Particiones*.

Formalmente, una población distribuida PD está conformada por diversas particiones, de tal forma que $PD = \{P_1, P_2, ..., P_i\}$, o bien, $PD = P_1 \bigcup P_2 \bigcup ... \bigcup P_i$, dónde $i \geq 2$. Por otra parte, cada partición P_i está integrada por uno o más individuos, de modo que $P_i = \{I_1, I_2, ..., I_n\}$, $n \geq 1$ y cada individuo I_n está definido mediante un subconjunto de variables descriptoras $V = \{V_1, V_2, ..., V_k\}$, así como de una variable objetivo o clase C_x. Adicionalmente, una clase dada puede asociarse a más de un individuo.

De las poblaciones distribuidas, existen distintos tipos y su estructura se construye, principalmente, en función de la naturaleza de los datos que incluye; por tanto, pueden encontrarse ejemplos donde existen (*i*) múltiples particiones con distintos individuos representados en cada una de ellas, pero con las mismas variables descriptoras e igual clasificación, (*ii*) diferentes individuos en cada partición, con diferentes variables descriptoras y la misma clasificación, y (*iii*) un tipo especial que considera diversos subconjuntos con representaciones distintas de los mismos

individuos y con diferente clasificación en cada una. En todos los casos, conviene observar que los individuos podrían requerir una sola representación global, dado que independientemente del dominio o contexto del que se trate, el contar con un conocimiento más general de éstos brinda un mejor soporte a la toma de decisiones.

A partir de una población distribuida identificada por $PD = \{P_a, P_b, ..., P_i\}$, dónde $P_a = \{I_{a1}, I_{a2}, ..., I_{an}\}$, $P_b = \{I_{b1}, I_{b2}, ..., I_{bm}\}$ y en general, $P_i = \{I_{i1}, I_{i2}, ..., I_{ij}\}$, se tiene que $\mid P_a \mid = n$, $\mid P_b \mid = m$, ..., $\mid P_i \mid = j$. Adicionalmente, cada $I_{an}, I_{bm}, ..., I_{ij}$ es un individuo definido por un subconjunto de variables dado, así I_{an} se describe por $V_a = \{V_x, V_y, ..., V_w\}$, I_{bm} por $V_b = \{V_p, V_q, ..., V_s\}$, ..., I_{ij} por $V_i = \{V_e, V_f, ..., V_h\}$. Por otra parte, $\forall I_{an} \Leftrightarrow C_a$, $\forall I_{bm} \Leftrightarrow C_b$, ..., $\forall I_{ij} \Leftrightarrow C_i$.

En el primer tipo de población distribuida mencionado anteriormente, se cumple que $\mid P_a \mid \neq \mid P_b \mid \neq ... \neq \mid P_i \mid$, $I_{an} \neq I_{bm} \neq ... \neq I_{ij}$, $V_a = V_b = ... = V_i = \{V_x, V_y, ..., V_w\}$, además, $\forall I_{an}, I_{bm}, ..., I_{ij} \Leftrightarrow C_a$. Por otra parte, para el segundo tipo se debe satisfacer que $\mid P_a \mid \neq \mid P_b \mid \neq ... \neq \mid P_i \mid$, $I_{an} \neq I_{bm} \neq ... \neq I_{ij}$ pero $V_a \neq V_b \neq ... \neq V_i$, es decir, $V_a = \{V_x, V_y, ..., V_w\}$, $V_b = \{V_p, V_q, ..., V_s\}$, ..., $V_i = \{V_e, V_f, ..., V_h\}$, luego $\forall I_{an}, I_{bm}, ..., I_{ij} \Leftrightarrow C_a$. Posteriormente, para el tipo especial se observa que $\mid P_a \mid = \mid P_b \mid = ... = \mid P_i \mid$, por lo que $I_{an} = I_{bm} = ... = I_{ij}$ pero con la diferencia de que $V_a = \{V_x, V_y, ..., V_w\}, V_b = \{V_p, V_q, ..., V_s\}$, ..., $V_i = \{V_e, V_f, ..., V_h\}$, ésto es $V_a \neq V_b \neq$, ..., $\neq V_i$ y en relación a las clases, $\forall I_{an} \Leftrightarrow C_a$, $\forall I_{bm} \Leftrightarrow C_b$, ..., $\forall I_{ij} \Leftrightarrow C_i$, dónde $C_a \neq C_b \neq$, ..., $\neq C_i$.

A fin de comprender mejor la distribución de individuos en las particiones descritas, obsérvese que para el primer y segundo tipo también se cumple que $P_a \cap P_b \cap ... \cap P_i = \emptyset$. Por otra parte, en el tipo especial de población distribuida se verifica que $P_a \cap P_b \cap ... \cap P_i = PD$ además de que $\forall n, I_{an} = I_{bn} =, ... = I_{in}$, es decir, la *n-esima* representación en cada partición corresponde al mismo individuo.

El tipo especial de población distribuida ya mencionado es el que se estudia en esta tesis y puede ser utilizado en empresas, dependencias gubernamentales y organizaciones en general, quizá sin tener plena conciencia de lo que están tratando. Por ejemplo, un grupo de clientes bancarios en una ciudad pueden tener un comportamiento financiero dado, el que no necesariamente es el mismo en una tienda departamental; posteriormente, si algunas de esas personas pretendieran adquirir un nuevo auto, la agencia vendedora podría requerir un conocimiento de dichos individuos para fortalecer en gran medida su toma de decisiones. En este caso, la población distribuida incluye a los mismos individuos representados de dos maneras diferentes, por un lado, con la estructura que interesa a entidades bancarias y

por el otro, mediante el esquema de variables propias para clientes de un entorno comercial automotriz.

En el caso de este trabajo doctoral, con base en el Reconocimiento de Patrones (RP) se estudian a profundidad ocho métodos Filter de Selección de Variables (SV) para combinarlos con una técnica de Agrupamiento o Clustering y seis de Clasificación Supervisada (CS), a fin de crear un método capaz de procesar el tipo especial de poblaciones distribuidas y encontrar una representación general de un grupo de individuos que se encuentran originalmente incluidos en múltiples conjuntos de datos, descritos localmente de diferentes maneras.

Antecedentes

En su concepción original, las técnicas para SV, Agrupamiento y CS, se desarrollaron para el procesamiento de datos en particiones únicas, por lo que existe mucha literatura disponible sobre aportaciones teóricas con respecto a dichos temas. Adicionalmente, también se tiene disponible una cantidad importante de trabajos de RP que persiguen objetivos particulares relacionados a un ámbito en específico, como puede ser la identificación de patrones en pacientes con enfermedades concretas, como por ejemplo el cáncer (Escarcega et al., 2010) o la diabetes y sus diferentes tipos (Tomar y Agarwal, 2015) o bien para detectar comportamientos genéticos, sociales, económicos, físicos, incluso para resolver problemas de agricultura (Bolón-Canedo et al., 2015), (Tyagi y Mishra, 2013), (Li et al., 2017), (Zhao et al., 2018), (Li et al., 2016), (Mera et al., 2017).

Derivado de lo anterior, el desarrollo de técnicas para tratar el problema de procesar poblaciones distribuidas es un área cuya exploración aún no es tan profunda como el caso de problemas cuyos conjuntos de datos tienen representaciones únicas. En general, los trabajos encontrados persiguen objetivos relacionados con el procesamiento simultáneo de los datos y no con la búsqueda de una representación global de la población (Bolón-Canedo et al., 2013), (Morán-Fernández et al., 2015), (Morán-Fernández et al., 2017).

De manera concreta, se han identificado trabajos referidos al procesamiento simultáneo de particiones de datos como una estrategia para facilitar el tratamiento de conjuntos, originalmente muy grandes, dividiéndolos en fragmentos más pequeños. Sin embargo, las particiones obtenidas incluyen diferentes objetos en cada una pero compartiendo formas de representación semejantes, dónde la separación de dichos objetos se realiza mediante la clase de los mismos, por lo que los frag-

mentos tienen la misma estructura pero distinta clasificación (Bolón-Canedo et al., 2014). También existen otros trabajos que proponen estrategias para tratar particiones ya existentes y que están conformadas por grupos de objetos de la misma población original, por lo que comparten la misma clase y variables descriptoras, como consecuencia de la aplicación de una técnica de fragmentación horizontal (Zhao et al., 2013b), (Bolón-Canedo et al., 2105a).

En los casos anteriores, el particionamiento se realiza principalmente para aprovechar las características multitarea y multiprocesamiento de las computadoras actuales, mientras que en el caso de este trabajo doctoral, el objetivo es tratar diversos conjuntos que incluyen representaciones diferentes de los mismos objetos, para encontrar una estructura general de los mismos. Cabe mencionar que en todo momento, al considerar a todas las particiones presentes como subconjuntos, la aplicación de la operación intersección entre todas ellas siempre equivale al total de objetos que conforman la población distribuida. Esto garantiza que se procesan siempre los mismos individuos en cada fragmento.

Problemática

Los casos donde se requiere el tratamiento automático de datos estructurados en poblaciones distribuidas, abundan cada vez más, observándose hoy en día una clara tendencia de su aumento. Específicamente, en este tipo de población distribuida obtener conocimiento desde un punto de vista más amplio sobre los objetos en estudio puede no ser una tarea fácil. La idea de llevar a cabo este proceso responde a que en algunos casos es necesario poder reconocer un comportamiento general de los individuos que componen a las poblaciones en estudio, con el fin de usar esa información en la toma de decisiones.

Derivado de lo anterior, se requiere identificar las posibles correlaciones que pudieran existir entre las variables descriptoras presentes en todas las poblaciones con la idea de construir una sola representación de los individuos en estudio con una nueva clasificación global.

Haciendo una revisión de la literatura para encontrar métodos que generen esas representaciones globales, no se ha encontrado algún modelo que sirva para realizar este tipo de procesamiento. Mayoritariamente, el estado del arte presenta trabajos diseñados para aprovechar el procesamiento simultáneo de particiones de un conjunto original de datos (Ebrahimpour y Eftekhari, 2018). En los trabajos de algunos autores, se han observado esfuerzos por adaptar los métodos tradicionales

al contexto del tratamiento distribuido de datos. Sin embargo, aunque se observan algunas aportaciones de nuevas técnicas diseñadas en dicho entorno, aún no se ha identificado alguna que aborde al mismo tiempo, cada una de las siguientes condiciones:

- Los mismos objetos están representados en todas las poblaciones distribuidas, es decir, si un objeto $O_i \in P_1, \forall O_i \in P_2 \Rightarrow PD = O_i$.
- Las variables descriptoras utilizadas, son distintas y excluyentes en cada partición.
- Existencia de clases diferentes en cada población.
- Se busca una representación global a partir de todas las particiones, que produce mejor desempeño de clasificación que si se considera simplemente la unión directa de sus variables.

Solución propuesta

Este trabajo doctoral presenta un método automático para construir una estructura general mediante la integración de variables que proceden de distintas poblaciones distribuidas y proponer una nueva clasificación correspondiente. Para lograr lo anterior se requiere del uso de técnicas especializadas provenientes de la SV, el Agrupamiento y la CS.

El conocimiento previo sobre los objetos representados en las particiones locales está circunscrito a su ámbito particular, lo que limita su aplicación a dicho contexto. Sin embargo, de manera general, es decir, sin la pertenencia a algún dominio en específico, hace falta contar con algún método para generar una nueva representación global y su correspondiente clasificación, ésto es valioso especialmente por la diversidad de aplicaciones que se puede tener, debido a que se proporciona una comprensión más amplia de los individuos estudiados.

El objetivo primordial se centra en identificar las posibles correlaciones entre las características representativas que definen a los objetos incluidos en diversos conjuntos de datos locales para que posteriormente, se defina su inclusión en la representación global correspondiente.

De manera general, el método de estructuración de poblaciones distribuidas propuesto incluye un algoritmo Filter de SV aplicado en una etapa inicial y que

identifica a las variables que en una segunda fase serán evaluadas con ayuda de un clasificador supervisado, determinando así el subconjunto final de ellas y que conforman la representación global que se desea. Con lo anterior es posible procesar poblaciones donde esencialmente se encuentran incluidos a los mismos individuos y que cuentan con una clasificación local.

Hipótesis

El tratamiento y estudio de la correlación de variables provenientes de diversas poblaciones distribuidas con los mismos individuos permiten identificar a aquellas que sirven para construir una representación global de éstos, con el fin de emplear esa información en la toma de decisiones.

Objetivo general

Definir un método para organizar automáticamente la estructura de poblaciones de iguales objetos con diferentes definiciones mediante técnicas de Reconocimiento de Patrones, para generar representaciones de datos con una nueva clasificación global.

Objetivos específicos

1. Evaluar diferentes métodos Filter univariados de Selección de Variables a través de su aplicación en conjuntos de datos de diversos tipos, para determinar la factibilidad de uso en el método propuesto.

2. Desarrollar un criterio para la generación de subconjuntos de variables seleccionadas localmente, validados con diversos tipos de conjuntos de datos.

3. Proponer una estrategia para construir un algoritmo de integración de variables en el ámbito del tratamiento simultáneo de particiones, que se usará en el método de estructuración propuesto.

4. Validar el nuevo método de estructuración mediante casos de estudio con las características del tipo especial de poblaciones distribuidas, para la obtención de una representación global de los datos y su clasificación.

Justificación

Con el nuevo método de estructuración de poblaciones distribuidas, será posible construir de manera automática una representación general de toda la población, a partir de diversas particiones locales incluyendo una nueva clasificación global. Ésto permitirá proporcionar un conocimiento más completo de los objetos en estudio, con lo que se apoyará al mejoramiento de diversos procesos de toma de decisiones.

El uso de este método tiene lugar en sectores como el comercial, industrial, académico y otros, por ejemplo, el comportamiento de un cliente con una institución bancaria, no necesariamente es igual con otra, pero al obtener un conocimiento más amplio sobre dicho cliente a través de las variables relevantes para una nueva clasificación, entonces un nuevo banco puede mejorar en gran medida el proceso de otorgamiento de nuevos beneficios crediticios.

Otro contexto en el que se requiera contar con un conocimiento más general de diversos objetos, puede darse en la línea de producción, almacén y el departamento de ventas de una fábrica, donde un mismo producto puede ser descrito por distintas variables de acuerdo a cada área. Así, el producto podrá ser detallado en una partición dada, por un costo de producción, precio de venta o el porcentaje de descuento por volumen, mientras que en otra, se incluyen datos como la fecha de fin de producción, cantidad entregada al almacén, incluso su peso o color. Sin embargo, siempre se trata del mismo producto, en todo caso sería muy útil saber ¿Cuáles son las variables que representan de manera general a todos los productos? o bien ¿Cuál es la clasificación global de los mismos?

El principal beneficio del proceso de integración de variables mencionado es que permite obtener para una población distribuida, una organización general de los objetos en estudio cuya representación incluya una dimensionalidad menor que en la estructura original, pero sin perder la capacidad de descripción de dichos objetos.

Aportaciones

La principal contribución de este trabajo doctoral consiste en proveer un método para organizar automáticamente una representación global de poblaciones distribuidas, el cual considera las siguientes características:

- Procesamiento del tipo especial de poblaciones distribuidas que incluyen a los mismos objetos en cada una de las particiones incluidas, descritos con variables distintas y con diferentes clases.

- Identificación de una efectiva combinación de selector Filter univariado de SV - clasificador supervisado, con los que se obtienen altos desempeños de clasificación.

- Incorpora un nuevo criterio de evaluación de subconjuntos de variables de acuerdo a un ranking, para obtener subconjuntos de variables seleccionadas.

- Propone una nueva estrategia para la integración de variables provenientes de diversas particiones distribuidas para obtener de forma automática, una nueva representación general y su correspondiente clasificación global.

Estructura del documento

El presente documento incluye un total de siete capítulos para abordar el desarrollo del nuevo método de estructuración de poblaciones distribuidas mediante el uso de técnicas de RP, para lo que en el capítulo 2 se incluyen diversos conceptos que sustentan la base científica de los métodos utilizados.

En el Capítulo 3, se presenta una revisión del estado del arte referido al tratamiento simultáneo de conjuntos de datos, mientras que en el Capítulo 4, se presenta la descripción general del contexto en el que debe trabajar el método de estructuración de poblaciones distribuidas, donde se ha detallado el escenario que corresponde por una parte al entorno tradicional en el que operan los trabajos relacionados identificados hasta el momento y por otra, el que corresponde a las poblaciones distribuidas con individuos representados en todas las particiones pero, con distintas variables descriptoras y clases.

Con la finalidad de establecer el procedimiento para obtener subconjuntos de variables seleccionadas localmente, en el Capítulo 5 se presenta la experimentación que a manera de comparativa de comportamiento se realizó con ocho selectores univariados, a fin de justificar la elección del método de tipo Filter y del clasificador supervisado que se incluirá en el método propuesto. Los resultados obtenidos se contrastan con las conclusiones presentadas en los trabajos mencionados en el Capítulo 3

Posteriormente en el Capítulo 6, se propone un principio que debe satisfacerse para identificar el momento de detención en la evaluación de rankings y que de acuerdo a la aplicación de un método de búsqueda secuencial, permite seleccionar subconjuntos de variables en las particiones locales.

En el Capítulo 7, se presenta una estrategia para la integración de variables provenientes de diversas particiones locales, cuyo objetivo es construir la representación general de los datos con su correspondiente clasificación. Se incluye la incorporación de técnicas de Agrupamiento y de CS así como, una discusión sobre la representación general construida para que en el Capítulo 8, se aborde la validación del nuevo método completo, evidenciando la efectividad de la propuesta. Durante este último proceso, se incluyen diferentes conjuntos de datos reales extraídos tanto de una universidad que ha aportado datos sobre su proceso interno de evaluación docente, como de repositorios cuyos contenidos están relacionados con tareas de clasificación.

Adicionalmente, se han incluido otras secciones sustantivas, una de ellas está referida a las conclusiones generales en donde se comentan las áreas de oportunidad para continuar esta investigación como trabajo futuro, otra más es para las referencias utilizadas en todo el documento. Finalmente, se consideró una sección con los anexos que complementan la información proporcionada en los Capítulos 5 y 8.

Capítulo 2

Marco teórico

Nunca se hizo ningún material tan resistente
como el espíritu humano

Bernard Williams

Con la finalidad de contar con los elementos teóricos necesarios que desde distintos puntos de vista conforman la investigación, motivo de esta tesis, se abordan diferentes conceptos referidos a las ciencias computacionales y de manera más específica, aquellos relacionados con el tratamiento de datos que se utilizarán a lo largo de este trabajo doctoral.

El conjunto de elementos teóricos que inciden en el tema de esta tesis están relacionados con el RP, mismo que incluye tres principales tareas, una de ellas la Selección de Variables ya mencionada y que será abordada en la siguiente sección.

2.1. Reconocimiento de patrones

Formalmente, el Reconocimiento de Patrones se define como la disciplina científica que desarrolla y utiliza métodos para la descripción de objetos así como su clasificación y se conforma de tres principales tareas: Selección de Variables, Clasificación Supervisada y Agrupamiento, en algunos contextos el Clustering puede ser visto como Clasificación No Supervisada (Marques de Sa, 2001).

El problema fundamental para el cual, el RP se ha desarrollado, consiste en proveer herramientas para que a través del procesamiento de datos suministrados

sobre un tema en particular, sea posible identificar objetos mediante sus características y tomar acciones a partir de las categorías de dichos objetos (Duda et al., 2001).

Las características que representan a los objetos, comúnmente se utilizan para catalogar a dichos objetos en diferentes grupos excluyentes a los que en RP se les denomina *Clases*. Al tener establecidas las clases en un conjunto de datos dado, una primera actividad consiste en identificar la pertenencia de nuevos ejemplos, en alguna de las clases presentes, siendo ésta la acción fundamental de la CS.

Por otra parte, cuando los conjuntos de datos no contienen una clasificación previamente establecida, es posible sugerir la conformación de diversos grupos a partir de las características de los datos en estudio, ésta es la labor principal del Agrupamiento.

Para actividades tanto de CS como de Agrupamiento, es necesario contar con alguna herramienta para determinar si alguna instancia puede parecerse a otra o no. En RP, se utiliza una noción fundamental denominada *Semejanza*, a fin de establecer una medida útil para indicar qué tanto uno o más objetos son similares entre sí (Marques de Sa, 2001). Una manera de evaluar la semejanza, se enfoca en la similaridad o distancia (disimilaridad) relacionada a las características que describen a los objetos, conocidas éstas como *Variables*.

Formalmente, un objeto puede estar descrito por un número n de variables identificadas como x, de tal forma que O=$\{x_1,\ x_2,\ x_3,\ ...,\ x_n\}$, representa a un conjunto de n variables, dónde $n \geq 1$. Por otra parte, un conjunto de objetos puede conformarse por m instancias representadas por I=$\{i_1,\ i_2,\ i_3,\ ...,\ i_m\}$, dónde $m \geq 1$.

El estudio de las variables incluidas en la representación de objetos se aborda a través de las tres tareas del RP. De manera específica, en la SV se tiene como principal actividad identificar a aquellas que son relevantes en la definición de objetos. Además se busca que el subconjunto de variables seleccionadas pueda representar adecuadamente al conjunto de datos original en estudio.

Una de las razones fundamentales de utilizar SV es para abordar el problema de la reducción de la dimensionalidad (Liu y Motoda, 2008), dónde el objetivo es disminuir el tamaño de la representación de los objetos a tratar sin perder la capacidad de representación de los mismos, para reducir así los recursos y tiempo necesarios para su procesamiento. Normalmente, el uso de algoritmos de SV por

su naturaleza antecede a los algoritmos tanto de CS como de Agrupamiento (Tan et al., 2005).

La SV puede abordarse desde diferentes enfoques ya descritos en la literatura, éstos se reconocen como Filtro y Envoltorio (Filter y Wrapper respectivamente por su traducción en inglés), así como uno más derivado de éstos y reconocido como Híbrido, en el cual se aprovechan diversas características de los primeros dos.

Una subclasificación de los métodos de SV puede darse en función del número de variables que se consideran simultáneamente en la evaluación, catalogándose como *univariado* aquel que analiza de manera individual la capacidad informacional que tiene una variable en correlación con los valores de una clase para un conjunto específico de datos. Por su parte, los métodos *multivariados* consideran la relación existente entre subconjuntos de variables (Duda et al., 2001).

Los algoritmos Filter univariados muestran los resultados de su aplicación a través de una lista de variables ordenada descendentemente a la que se le denomina *Ranking*, dicho orden se establece de acuerdo a una medida de relevancia que cada variable tiene respecto a la clase, este coeficiente es obtenido en el proceso y suele llamársele *Mérito*. Una estrategia para calcular la relevancia consiste en el uso de estadísticas, como por ejemplo la varianza y/o la correlación entre variables descriptivas y objetivo.

El ranking derivado de un algoritmo Filter suele ordenarse de manera ascendente o bien descendente, en éste caso se muestra en primer lugar a la variable mejor evaluada, luego a la siguiente y así sucesivamente hasta listar a todas las variables.

Si se considera un conjunto original de datos con N variables, la idea principal consiste en elegir un subconjunto de k variables y dado que el orden del ranking es de acuerdo al mérito de cada una, las primeras serán también las k mejores y servirán para determinar el subconjunto más pequeño que represente a los objetos sin modificar su comportamiento. El valor de k es un entero positivo $\leq N$ y representa el número de variables con las que se obtiene el mejor desempeño de clasificación.

A partir de un ranking de variables, la búsqueda de subconjuntos de ellas se puede efectuar de manera iterativa, básicamente de dos formas (Kittler, 1978). La primera consiste en considerar inicialmente solo a la variable mejor rankeada para aplicarle un proceso de clasificación supervisada y obtener así, la precisión del cla-

sificador. El proceso se repite agregando la siguiente variable mejor calificada en el ranking y así sucesivamente, hasta completar todas las variables del conjunto original de datos. Esta forma de procesamiento se conoce como *Búsqueda Secuencial hacia Adelante* (SFS por sus siglas en inglés) y puede ser usada para elegir entonces a las k variables para representar al conjunto de datos completo. La segunda forma de búsqueda considera invertir el orden en el que las variables se pasan al clasificador comenzando las pruebas con todas las variables para posteriormente, ir eliminando una a una comenzando por la variable menos significativa de acuerdo al ranking de ellas y continuando de manera iterativa eliminando la siguiente menos significativa, hasta eliminarlas todas. Esta técnica es conocida como *Eliminación Secuencial hacía Atrás* (SBE por sus siglas en inglés).

Dado que la aplicación de algoritmos de SV busca la reducción de la dimensionalidad de un problema, en las tareas de clasificación tanto supervisada como no supervisada posteriores se obtienen entre otras, las siguientes ventajas (Chandrashekar y Sahin, 2014):

- Posible mejora del rendimiento y/o la eficacia de los métodos.
- Reducción del coste computacional para el tratamiento de datos.
- Facilita la comprensión del modelo final.

El desarrollo de métodos de SV, en general, está soportado por conceptos cuyo ámbito puede ser tanto estadístico como matemático. Entre los más utilizados para determinar la importancia de las variables descriptoras, a través de criterios para medir la relevancia o redundancia de éstas, están los derivados de la Teoría de la Información (Shannon, 1948).

A continuación, se revisan a detalle diversos conceptos teóricos que son incluidos en los métodos de SV implícitos en el presente trabajo.

2.2. Relevancia de variables

Una variable puede ser considerada como relevante si está correlacionada con el valor predictivo de una clase, en otro caso, se le considera irrelevante en cierta medida (Yu y Liu, 2004).

Dada la noción de relevancia de variables, adicionalmente, éstas pueden ser categorizadas como fuertemente relevantes, débilmente relevantes e irrelevantes con

respecto a su capacidad de incidir en la clase (John et al., 1994), (Zeng et al., 2015).

Una variable es fuertemente relevante cuando ésta no puede ser ignorada en el subconjunto de las que deben resultar seleccionadas para representar a los objetos en estudio, ya que afecta la distribución original de la clase; mientras que una variable débilmente relevante pudiera ignorarse, pero bajo condiciones específicas y en función del objetivo de clasificación deseado, podría ser necesaria. Por otra parte, se considera irrelevante a una variable cuando no aporta información para la definición de la clase y entonces puede ser removida del subconjunto seleccionado (Yu y Liu, 2004).

Basándose en la importancia de las variables, se han desarrollado una gran cantidad de métodos, que apoyados en diferentes herramientas conceptuales buscan encontrar el subconjunto óptimo de variables.

En la subsección siguiente, se abordan los algoritmos Filter univariados investigados a lo largo de esta tesis.

2.3. Algoritmos filter univariados de SV

Para la comprensión de los métodos de SV univariados utilizados en el presente trabajo y en particular, de los procesos de obtención de los rankings derivados de cada uno, se presenta la descripción teórica de los mismos.

Los algoritmos fueron elegidos debido a que su utilización es frecuentemente reportada por otros investigadores reconociendo su eficiencia, con la ventaja adicional de encontrarse disponibles en diversas plataformas de software para RP y/o para Minería de Datos.

2.3.1. Chi cuadrada (Chi Squared)

El método Chi Squared permite evaluar la dependencia de cada variable con respecto de la clase haciendo un análisis estadístico de ocurrencia de dos eventos dados. En tareas de SV, estos eventos se refieren a la frecuencia de aparición de valores distintos en cada variable y en la clase.

Su algoritmo se centra en calcular el estadístico Chi cuadrado para todas las variables descriptoras incluidas en el conjunto original de datos, ésto permite esta-

blecer su nivel individual de correlación con respecto a la clase (Dunham, 2003).

El proceso de cálculo comienza con el establecimiento de una hipótesis inicial denominada *Hipótesis nula* H_0 que deberá verificarse al final del mismo y de una contraparte conocida como *Hipótesis alternativa* H_1 que niega a H_0. Típicamente, la hipótesis nula sostiene la premisa de que la variable en análisis y la clase son independientes entre sí, mientras que la hipótesis alternativa indica que las variables estudiadas son dependientes; la independencia en este contexto implica que el conocimiento del valor de la variable predictora, no determina el valor que deberá tomar la variable clase (Jin et al., 2006).

El propósito es entonces, evaluar la dependencia o independencia de las variables en estudio mediante el contraste de las hipótesis H_0 y H_1, para lo que a continuación se presenta una secuencia de pasos que permiten determinar un ranking de variables a partir del estadístico Chi cuadrado.

H_0: Las variables X y Y son independientes.

H_1: $\neg\, H_0$

1. Establecer las hipótesis H_0 y H_1.
2. Crear una tabla que contiene el conteo de cada uno de los valores diferentes, observados para cada variable en el conjunto original de los datos.
3. Agregar una fila y una columna para registrar la sumatoria de cada frecuencia. A ésta se le denomina Tabla de Contingencia.
4. Construir una segunda tabla para registrar también, las frecuencias observadas -es decir, el conteo simple de ocurrencia de cada valor- así como sus frecuencias esperadas correspondientes, estas últimas mediante la Ecuación 2.1.
5. Calcular el estadístico Chi cuadrada mediante la expresión de la Ecuación 2.2.
6. Ordenar de manera descendente las variables en función del estadístico obtenido, ésta es la lista buscada.

$$E_{ij} = \frac{\sum_{j=1}^{c} X_i \sum_{i=1}^{r} X_j}{\sum_{i=1}^{r} \sum_{j=1}^{c} X_{ij}} \tag{2.1}$$

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(X_{ij}E_{ij})^2}{E_{ij}}, \tag{2.2}$$

dónde:
X_{ij} corresponde al valor *i,j* de la tabla de frecuencias observadas.
E_{ij} corresponde al valor *i,j* de la tabla de frecuencias esperadas.
r corresponde al número de filas de la tabla de contingencia.
c corresponde al número de columnas de la tabla de contingencia.

2.3.2. Ganancia de información (Info Gain)

El algoritmo Information Gain (Info Gain) evalúa las variables obteniendo la ganancia de información de cada una con respecto a la clase, para lo cual previamente discretiza las variables numéricas (Quinlan, 1986). Es común observar que esta medida determina la entropía en el conjunto original de datos, vista ésta como *"la medida del desorden de un sistema mediante la incertidumbre existente ante un conjunto de casos, del cual se espera uno solo"* (Ruiz-Sanchez, 2005).

La Ecuación 2.3 presenta el cálculo de la ganancia de información.

$$Ganancia\,(\varepsilon,\, X_i) = Ent\,(\varepsilon) - \sum_{v=1}^{|X_i|} \frac{|\varepsilon\,(x_{iv})|}{|\varepsilon|} \times Ent\,(\varepsilon\,(x_{iv})) \tag{2.3}$$

donde $|\varepsilon|$ es el número total de ejemplos y *Ent(ε)* es la entropía que se obtiene a través de la Ecuación 2.4; $|X|$ es el número de valores distintos de una variable X_i, $\varepsilon(x_{iv})$ es el subconjunto de ejemplos para el cual $X_i = x_{iv}$, donde $|\varepsilon(x_{iv})|$ es su cardinal.

$$Ent\,(\varepsilon) = -\sum_{l=1}^{k} \frac{frec\,(y_l,\, \varepsilon)}{|\varepsilon|} \times \log_2 \left(\frac{frec\,(y_l,\, \varepsilon)}{|\varepsilon|}\right) \tag{2.4}$$

En la Ecuación 2.4, $\frac{frec(y_l,\, \varepsilon)}{|\varepsilon|}$ es la probabilidad de que se dé un ejemplo con clase y_l y $\log_2 \left(\frac{frec(y_l,\, \varepsilon)}{|\varepsilon|}\right)$ es la información que transmite un ejemplo de clase y_l, la entropía es máxima cuando todas las clases presentan la misma proporción.

Obtener un ranking mediante el cálculo de la Ganancia de Información, implica:

1. Calcular la Ganancia de Información de cada variable con respecto de la clase.

2. Ordenar las variables de acuerdo al valor obtenido, este orden constituye el ranking derivado.

2.3.3. Proporción de ganancia de información (Gain Ratio)

La tarea central del algoritmo Gain Ratio (Quinlan, 1993) consiste en evaluar la importancia de cada variable midiéndola a través de un valor determinado por un cociente al que se le conoce como razón de beneficio con respecto a la clase.

Este algoritmo se diseñó debido a que el autor identificó un problema de comportamiento con el algoritmo Info Gain, el cual es inherente a la frecuencia de distribución de los datos en los conjuntos tratados, ya que si una variable contiene valores únicos entonces se generará un gran número de subconjuntos que incluyen una sola clase, lo que para identificar la importancia de las variables es inútil.

La idea es rectificar el valor de la ganancia de información mediante la normalización de dicha medida, lo que para una variable X dada, Gain Ratio se obtiene por la Ecuación 2.5.

$$Gain\ Ratio\ (X) = Gain\ (X)\ /\ SplitInfo\ (X) \tag{2.5}$$

Split Info (X) representa la información potencial que se obtiene para la variable X, al dividir un conjunto de datos S dado, en n subconjuntos, lo anterior mediante la Ecuación 2.6.

$$SplitInfo_X\,(S) = -\sum_{i=1}^{n} (\ |S_i|\ /\ |S|\ ) \log_2 (\ |S_i|\ /\ |S|\ ) \tag{2.6}$$

Por otra parte, la información que puede ser ganada sobre el conjunto de variables que conforman a las instancias en S, está denotada por la Ecuación 2.7.

$$Gain\,(X) = I\,(S) - E\,(X) \tag{2.7}$$

Con el cálculo de Gain Ratio mediante la Ecuación 2.5, los pasos a seguir para construir un ranking de las variables presentes en S pueden consistir en:

1. Calcular el valor Gain Ratio para todas las variables del conjunto de datos.

2. Ordenar descendentemente a las variables, en función del valor obtenido.

2.3.4. Puntuación Laplaciana (Laplacian Score)

En el caso del algoritmo Laplacian Score (LS), se asume que las instancias de la misma clase están cerca una de la otra y que la distribución de sus datos -entendida ésta como la estructura geométrica local- es crucial para la discriminación de variables. Así, la relevancia de una variable está dada por su capacidad de conservar su ubicación local (He et al., 2005).

Para comprender mejor la noción de relevancia de variables en el algoritmo LS, debe asumirse que "*una variable es consistente con la estructura de los datos, si ésta toma valores similares cuando los objetos están cercanos uno del otro y toma valores disimilares cuando los objetos están lejanos entre sí*" (Zhao y Liu, 2007b).

En el algoritmo LS, al obtener la cuantificación de la consistencia de una variable con relación a la estructura de los datos, es posible establecer un ranking de las variables donde la más significativa es la que tiene una mayor capacidad de conservar la ubicación local de los datos (Solorio-Fernandez et al., 2010b).

Fundamentalmente, este algoritmo se basa en los Eigenmapas Laplacianos (Belkin y Niyogi, 2001) y la proyección de la conservación de la estructura local, para lo que se obtiene un valor denominado como Lr para cada r-ésima variable. Un resumen de los pasos que deben seguirse para construir un ranking de variables, consiste en:

1. Construir un grafo de vecindad cercana con m nodos, donde el i-ésimo nodo se refiere a la instancia P_i; se considera una frontera entre los nodos i y j si P_i y P_j están cerca, entendiendo que P_i está entre los k vecinos más cercanos de P_j o bien, que P_j está entre los k vecinos más cercanos de P_i. Si el valor de la clase está disponible se puede establecer una frontera entre ambos nodos compartiendo la misma clase.

2. Si los nodos i y j están conectados, entonces puede obtenerse S denominada como la Matriz de Pesos del Grafo, misma que es utilizada para modelar la estructura local del espacio de datos. Para el cálculo de esta matriz puede utilizarse la Ecuación 2.8:

$$S_{ij} = e^{-\frac{\|P_i - P_j\|^2}{t}}, \tag{2.8}$$

   donde t es una constante elegible, es decir, puede establecerse su valor a discreción-; en otro caso S_{ij}= 0.

3. Para la *r*-ésima variable, se define la matriz denotada como grafo Laplaciano *L* mediante la Ecuación 2.9.

$$L = D - S, \tag{2.9}$$

dónde *D* es otra matriz denotada por la Ecuación 2.10

$$D=diag(S1), 1 = [1, ..., 1]^T \tag{2.10}$$

4. Finalmente, se puede calcular el Score Laplaciano de la *r*-ésima variable, mediante la Ecuación 2.11:

$$L_r = \frac{\tilde{f}_r^T \, L\widetilde{f}_r}{\tilde{f}_r^T \, D\widetilde{f}_r}, \tag{2.11}$$

dónde:
$\tilde{\mathbf{f}}_{\mathrm{r}} = \mathbf{f}_{\mathrm{r}} - \frac{\mathbf{f}_{\mathrm{r}}{}^{\mathrm{T}}\mathrm{D1}}{1^{\mathrm{T}}\mathrm{D1}}1$
$\mathrm{fr} = [\mathrm{fr}1, \mathrm{fr}2, \ldots, \mathrm{fr}m]^{\mathrm{T}}$

5. Los valores obtenidos por L_r se ordenan descendentemente conformando así el ranking de las variables.

2.3.5. Regla única (One R)

El algoritmo One R (Holte, 1993) es un método implementado mediante árboles de decisión, donde el conocimiento obtenido durante la fase de entrenamiento puede representarse a través de un árbol.

Es común observar a One R en tareas de SV y como método de clasificación supervisada. Para el caso de la SV, el algoritmo evalúa la calidad de cada variable mediante el clasificador One R y utiliza la variable de mínimo error cuadrático, para predecir el valor de la variable objetivo. En este método, se genera un árbol de decisión típicamente de un nivel simple, por lo que se tiene una sola regla de decisión, sin embargo es capaz de inferir reglas de clasificación partiendo de un conjunto de instancias; puede manejar valores faltantes y atributos numéricos mostrando su adaptabilidad a pesar de su simplicidad; es posible aplicar el algoritmo en conjuntos de datos tanto numéricos como nominales.

En tareas de Selección de Variables, los pasos a seguir para construir un ranking de variables, consisten en:

1. Para cada variable predictora en el subconjunto de datos de entrenamiento, debe crearse una regla mediante:

   - La construcción de una tabla de frecuencias; ésto se hace a través de un conteo simple del total de veces que aparece cada valor diferente de la variable en la distribución de la clase.
   - La elección de la clase más frecuente para cada valor de la variable.

2. Calcular el error mínimo de las reglas de cada variable predictora, ésto se realiza sumando las frecuencias mínimas de cada valor de la variable en la clase.

3. Seleccionar la regla con el error total más pequeño como la única regla, la variable predictora correspondiente es considerada la más significativa.

4. Ordenar las variables de forma descendente de acuerdo a su significancia, una vez finalizado el proceso, se habrá obtenido el ranking.

2.3.6. Relief

El algoritmo Relief (Kira y Rendell, 1992) consiste en asignar un peso de relevancia a cada variable evaluada, mismo que sirve para destacar su importancia con respecto a una clase. Se trata de un método que utiliza aleatoriamente instancias provenientes de un conjunto de entrenamiento y posteriormente, realiza el cálculo de los valores de relevancia correspondientes. Se identifican las diferencias entre una instancia cualquiera X y las dos más cercanas a ella, una dentro de su misma clase y otra de una clase diferente; a la más cercana en la misma clase se le denomina *Near hit* mientras que a la más cercana de otra clase, se le identifica como *Near miss*. En otras palabras, se estima la calidad de los atributos de acuerdo a qué tan bien distingue a una instancia de otras, i.e. qué tan cerca están entre sí las instancias en observación.

Considérese de un conjunto de datos G, un subconjunto de tamaño n y el establecimiento de un umbral de relevancia con valores entre 0 y 1 denominado τ y que se obtiene por medio de la Desigualdad de Chebychev (1867) referida a la probabilidad de que una variable aleatoria con varianza finita se ubique a cierta distancia de su esperanza matemática. Este algoritmo puede identificar aquellas variables que son estadísticamente relevantes con relación a una variable objetivo o clase; las diferencias en los valores de una variable i para un par de instancias P y Q se obtienen mediante las Ecuaciones 2.12 y 2.13

Para variables nominales, booleanas:

$$diff(Pi,\ Qi) = \begin{cases} 0\ si\ P_i = Q_i \\ 1\ si\ P_i \neq Q_i \end{cases} \tag{2.12}$$

Para variables numéricas:

$$diff(Pi,\ Qi) = \frac{(P_i - Q_i)}{nu_i}, \tag{2.13}$$

donde nu_i es un parámetro adicional utilizado para normalizar los valores resultantes en un intervalo de [0, 1].

Por otra parte, los pesos de relevancia de las variables se almacenan en un vector denominado *W*, donde para cada variable se calcula $Rel = .(\frac{1}{n})W$.

El algoritmo considera:

1. Seleccionar aleatoriamente una instancia.
2. Buscar los dos vecinos más cercanos.
   - Dentro de la misma clase (*Near hit*).
   - En otra clase (*Near miss*).
3. Actualizar el estimado de calidad o relevancia para todas las variables en función de los valores de cada uno con respecto de los respectivos *Near hit* y *Near miss*, ésto se obtiene mediante W_i - $(P_i - Near\ hit_i)^2$ + $(P_i - Near\ miss_i)^2$.
4. Al finalizar el proceso y dependiendo del ordenamiento de los valores almacenados en el vector de pesos, se construye el ranking de variables correspondiente.

La idea principal es que si una variable es relevante entonces P_i y $Near\ hit_i$ estarán menos separados en el espacio de vecindad de P. Por el contrario, una variable es menos relevante entre más separados se encuentran P_i de $Near\ miss_i$.

2.3.7. Máquinas de vectores de soporte (Support Vector Machine)

El algoritmo Support Vector Machine (SVM) está basado en la Teoría del Aprendizaje Estadístico y se utiliza para tareas de SV así como en CS. En este algoritmo se considera la inclusión de separadores lineales identificados como *Hiperplanos* en espacios de variables con dimensionalidad alta. Originalmente, el algoritmo se

diseñó para problemas de clasificación binaria (Vapnik y Lerner, 1963), dado que diversas representaciones de ejemplos denominados *Vectores* se ubican a un lado o al otro de un hiperplano, indicando con ésto la pertenencia a una clase dada o a la otra. Sin embargo, las contribuciones realizadas hoy en día permiten exitosamente su aplicación en SV, incluso en ambientes multiclase (Boser et al., 1992), (Cortes y Vapnik, 1995) y (Vapnik, 1998).

La idea comienza con la creación de un hiperplano en un espacio D-dimensional $\Re$D y que se define como $h(x)=\langle w,x\rangle + b$, donde $w \in \Re$D, el cual representa un vector ortogonal al hiperplano, $b \in \Re$ definiendo a la función signo como se observa en la Ecuación 2.14 (Hernandez Orallo et al., 2005).

$$signo\,(x) = \left\{ \begin{array}{ll} +1 & si\ x \geq 0 \\ -1 & si\ x < 0 \end{array} \right\} \tag{2.14}$$

En este ámbito de clasificación, las $x \in \Re$D son representaciones vectoriales de los ejemplos incluidos, mismos que incluyen un componente real por cada variable y donde el vector w se le denomina *Vector de Pesos*, dicho vector contiene un peso para cada variable representando su significatividad con respecto a qué tanto contribuye a la regla de clasificación definida. Esta importancia es la que se utiliza para construir el ranking de variables mediante SVM.

La construcción ideal del hiperplano consiste en situarlo en la posición más neutral posible en relación a las clases representadas en el espacio de características, con la idea principal de maximizar la distancia ortogonal existente entre la línea que lo define y hasta los vectores más cercanos de cada clase, estos últimos son los denominados *Vectores de Soporte*.

El algoritmo SVM, visto entonces como selector de variables, debe considerar lo siguiente:

1. Definir un espacio de características que represente a todas las variables en estudio.
2. Construir un hiperplano mediante una función de separación ortogonal a cada clase, que sea equidistante a los vectores de soporte.
3. Si los ejemplos no permiten construir un ejemplo linealmente separable se procede a construir nuevos espacios de características mediante transformaciones del espacio de variables originales, a través de funciones de núcleo o kernel, matemáticamente implica convertir funciones elipsoidales o curvas en lineales.

4. Encontrar a partir de las representaciones, el vector de pesos w.

5. En función de los valores de w, emitir el ranking de las variables.

De acuerdo a la literatura, este algoritmo ha demostrado altos niveles de eficiencia en conjuntos de datos de muy alta dimensionalidad, específicamente, debido a la facilidad de adaptarse a diferentes espacios de características, dependiendo básicamente de la construcción de las funciones de núcleo para establecer las fronteras de separación de los vectores.

2.3.8. Incertidumbre simétrica (Symmetrical Uncertainty - SU)

Para describir este algoritmo, es importante observar que existen básicamente dos medidas para determinar la correlación entre un par de variables aleatorias, la correlación lineal y la no lineal. En el caso de la primera, una de las más conocidas es el Coeficiente de Correlación Lineal cuyo valor se encuentra entre -1 y 1 indicando el grado de independencia entre las variables observadas; mientras que en el caso de la segunda, las medidas están basadas en el concepto teórico de la Entropía, como una métrica de incertidumbre de una variable aleatoria definida por la ecuación 2.4. De lo anterior, este algoritmo utiliza un criterio también denominado *Incertidumbre Simétrica* (SU) (Press et al., 1988), definido por la Ecuación 2.15.

$$SU\,(X,\,Y) = 2\left[\frac{IG(X\,|Y)}{H\,(X)+H\,(Y)}\right] \tag{2.15}$$

Esta ecuación restringe sus valores a un rango de [0, 1], donde el 0 indica que las variables X y Y son independientes, mientras que el 1 significa que el valor de una variable puede predecir el valor de la otra. Considerando que una de las variables es la clase, es posible construir un ranking ordenando las variables en función de la independencia sugerida por la tendencia al valor 0 de $SU(X,Y)$.

Si se evaluara la SU entre las variables descriptoras, podrían establecerse criterios para identificar variables redundantes entre sí, con lo que sería factible considerar la eliminación de algunas de ellas, dado que las variables redundantes son prescindibles generalmente (Yu y Liu, 2004).

2.4. Resumen

La literatura correspondiente a los métodos Filter de SV es amplia, abordando aspectos importantes desde su concepción original hasta su aplicación en múltiples problemas y con diversos objetivos en particular. Los conceptos teóricos revisados,

permiten comprender los principios de funcionamiento de los algoritmos utilizados en esta tesis, soportados por elementos de la Teoría de la Información, matemáticos, estadísticos y/o heurísticos.

Para mejorar la comprensión de los selectores univariados descritos, se presenta un resumen de los aspectos relevantes de cada uno de ellos en la Tabla 2.1, se incluyen datos sobre el fundamento teórico y el autor del método.

Tabla 2.1: Descripción de los algoritmos Filter univariados revisados.

No	Filter	Fundamento	Autor	Medida / Criterio
1	Chi Squared	Estadístico	Dunham, 2003	Chi cuadrada (χ^2).
2	Info Gain	Teoría de la información	Quinlan, 1986	Entropía.
3	Gain Ratio	Teoría de la información	Quinlan, 1993	Normalización de la ganancia de información.
4	Laplacian Score	Matemático	He et al, 2005	Eigenmapas Laplacianos.
5	One R	Heurístico, estadístico	Holte, 1993	Mínimo error cuadrático.
6	Relief	Heurístico	Kira y Rendell, 1992	Umbral de relevancia mediante la desigualdad de *Chevychev*.
7	SVM	Matemático	Vapnik y Lerner, 1963	Función de decisión.
8	Symmetrical Uncertainty	Teoría de la información	Press et al, 1988	Normalización de la ganancia de información.

Posterior a su introducción, algunos de los algoritmos han sido objeto del estudio a profundidad de otros investigadores obteniéndose actualizaciones a fin de mejorar características determinantes de su funcionamiento.

Todos los selectores Filter univariados que han sido descritos se utilizan como soporte para la investigación que se desarrolla en esta tesis.

Capítulo 3

Estado del arte

La ciencia no sabe de países, porque el conocimiento le pertenece a la humanidad y es la antorcha que ilumina el mundo. La ciencia es el alma de la prosperidad de las naciones y la fuente de todo progreso

Louis Pasteur

A fin de dirigir correctamente los pasos seguidos en esta investigación, se ha realizado una búsqueda documental de diversos trabajos reportados, cuya naturaleza se centra en el tratamiento de datos de manera no centralizada o con una sola representación. Ésto con la finalidad de identificar un área de oportunidad, que permita realizar un aporte valioso al conocimiento y posibilite en un futuro, resolver problemas cuyo ámbito incluya entornos distribuidos.

Se presenta en la sección siguiente, la revisión de literatura cuya relación con el tema de esta tesis se refiere al ámbito del procesamiento simultáneo de conjuntos de datos.

3.1. Procesamiento simultáneo de conjuntos de datos

Los trabajos presentados a continuación evidencian el esfuerzo dirigido por otros autores en el contexto del tratamiento de diversos subconjuntos de datos, para lo que se incluyen diversos enfoques, técnicas y alcances.

3.1.1. Aprendizaje distribuido con reducción de datos

Este trabajo se ha desarrollado como un acercamiento al aprendizaje distribuido, mismo que incluye entre diversas tareas de selección de variables, una familia de algoritmos para reducción de instancias utilizando una medida de similaridad (Czarnowski, 2011).

El autor propone reducir el conjunto original de características considerándolo como un vector de atributos del conjunto de entrenamiento, procediendo a calcular el valor de su coeficiente de similaridad entre ellas y agrupando instancias dentro de clusters consistentes de vectores con valores idénticos a este coeficiente, en dichos clusters se selecciona una representación y el conjunto reducido de datos se ha formado. La selección en este caso cubre tópicos de reducción de instancias y de atributos.

En este trabajo, se conjuntan procedimientos de integración requeridos para la reducción de datos con un clasificador de aprendizaje, a los que se les llama agentes de población de datos. El documento considera los siguientes objetivos:

1. Proponer un conjunto de procedimientos para reducción de datos a través de instancias simultáneas y selección de atributos con la selección de prototipos, ejecutados por programas denominados *agentes*.

2. Diseñar e implementar un framework basado en agentes, para el aprendizaje de los clasificadores y que trabajen a partir de los datos distribuidos y reducidos.

A partir de una base de datos, mediante técnicas de fragmentación horizontal se construyen diversas particiones de la misma, se busca el balanceo de clases en los diferentes fragmentos, posteriormente, cada uno de ellos se ubica en computadoras independientes para aprovechar las bondades del procesamiento en paralelo, a nivel local cada fragmento es procesado con programas que bajo el esquema de agentes aplican diversos métodos tipo Filter para encontrar un subconjunto de variables susceptibles de pasar a una fase Wrapper. Para esta parte, los agentes mantienen comunicación entre sí y en conjunto encuentran el subconjunto óptimo, para validar los resultados, se somete el subconjunto de variables elegidas a procesos de clasificación supervisada y se compara la eficacia de los clasificadores con la clasificación de la partición original.

Un aspecto importante de este trabajo es que la fragmentación horizontal por su naturaleza produce fragmentos con las mismas variables, por lo tanto, no procesa

particiones de características diferentes entre sí.

3.1.2. Selección de variables masivamente en paralelo, un enfoque basado en la preservación de la varianza

El objetivo es presentar un nuevo algoritmo de selección de variables de gran escala, basado en el análisis de la varianza, el algoritmo trabaja mediante la evaluación de las capacidades para preservar la varianza de datos, se dirige al ámbito de la selección de variables supervisada aunque aborda también el caso de la selección de variables no supervisada (Zhao et al., 2013b).

El algoritmo propuesto está implementado como un procedimiento analítico de alto desempeño que puede leer datos de forma distribuida, así como ejecutar tareas de selección de variables en dos modos: Multiprocesamiento simétrico vía multihilo y Procesamiento masivamente en paralelo vía interfaz de paso de mensajes.

El trabajo realizado tiene la capacidad de proveer un enfoque unificado para selección de variables en ámbitos tanto supervisados como no supervisados, en el primer caso soporta tareas de regresión y clasificación. Maneja efectivamente características redundantes, determina cuántas características deben seleccionarse cuando la información objetivo está disponible para el modelo de selección, además, el algoritmo está optimizado y paralelizado basándose en el particionamiento de datos.

Las particiones que se generan son tanto horizontales como verticales, no es necesario el balanceo de clases, tampoco se requiere que los fragmentos contengan el mismo número de clases entre sí. También se aplican algoritmos tipo Filter en la fase de análisis de los fragmentos de manera local. De igual forma, se incluyen algoritmos basados en métodos tipo Wrapper para encontrar el subconjunto óptimo de características.

Si el contexto es en selección de variables no supervisada, lo que se calcula es la covarianza entre variables. Un particularidad de este trabajo es que un individuo no puede estar representado en más de una partición, de la misma manera los individuos que se encuentran en una partición pueden contener características que están en el conjunto original pero que pudieran no haberse considerado al crear los fragmentos, esto puede ser un riesgo de pérdida de variables representativas o informacionales.

3.1.3. Selección de variables distribuida utilizando particionamiento vertical para datos de alta dimensionalidad

Este trabajo tiene el objetivo de proponer una técnica para la aplicación del proceso de SV distribuida, de tal forma que se ejecute de forma paralela en diversos nodos utilizados para tratar fragmentos de un conjunto de datos que originalmente se considera de gran escala y alta dimensionalidad (Prasad et al., 2016).

Los autores proporcionan información teórica sobre una técnica distribuida basada en la aplicación de un selector de tipo Filter. Dicha estrategia distribuye verticalmente conjuntos de datos de alta dimensionalidad en varios nodos informáticos, lo anterior para identificar rankings de las diferentes variables enviadas a cada nodo. La idea principal consiste en realizar la distribución de dichos rankings de variables y procesarlos simultáneamente.

Durante todo el procesamiento realizado se utiliza el algoritmo Filter univariado de SV denominado *Info Gain* (Quinlan, 1986). Sin embargo, no se reporta el nombre del clasificador involucrado en la etapa de evaluación de los subconjuntos de variables seleccionadas.

En el estudio, se utilizaron cinco conjuntos de datos provenientes del repositorio UCI (Lichman, 2009), a partir de los cuales, la experimentación consiste en particionar verticalmente todos los conjuntos de datos en cinco fragmentos con el mismo número de variables. Posteriormente, se aplica el selector Info Gain a cada fragmento local para construir un ranking correspondiente a cada partición y así, obtener de cada una de ellas un subconjunto con las mejores k variables. A continuación, se procede a aplicar una estrategia de mezclado de variables, ésta comienza por ordenar descendentemente las variables de acuerdo a su mérito local para finalmente, elegir un subconjunto global con las primeras k variables globales.

De acuerdo a los resultados es posible observar que en todos los casos presentados, el procesar conjuntos de datos que han sido particionados previamente de manera vertical permite identificar un subconjunto de variables que presentan un mejor desempeño de clasificación con respecto a la representación original. Adicionalmente, el tiempo de procesamiento se reduce significativamente.

Una característica importante de este trabajo es que inicialmente se realiza un proceso de particionamiento vertical para realizar el procesamiento distribuido, esta fragmentación produce particiones que incluyen aproximadamente el mismo número de variables en cada una, lo que balancea el trabajo en cada nodo de proce-

samiento, incluso la clase original también está presente en cada fragmento por lo que no existe una clasificación local independiente o distinta una de otra, incluso la clase del subconjunto final es la misma.

En el trabajo no se busca encontrar una nueva clasificación sino encontrar un subconjunto de variables provenientes del conjunto original de datos que mejor represente a los objetos de acuerdo a su clase inicial, con la idea básica de reducir el tiempo de procesamiento necesario.

3.1.4. Un enfoque eficiente en tiempo para selección de variables distribuida particionando por variables

En este artículo, se propone un enfoque de tipo Filter distribuido para el procesamiento de datos particionados verticalmente, con la intención de reducir el tiempo necesario para realizar SV en conjuntos de datos de alta dimensionalidad (Morán-Fernández et al., 2015).

La idea principal es utilizar una estrategia de particionamiento vertical, es decir por variables, para después emplear repetidamente un algoritmo Filter de SV a todas las particiones generadas, este procedimiento se realiza para que en cada iteración se obtenga una votación para las variables seleccionadas; al terminar el ciclo SV se obtiene un subconjunto final de variables. Posteriormente, se propone una estrategia de fusión de variables basada en la complejidad teórica de los subconjuntos de variables en lugar de considerar solamente el error de clasificación.

En el trabajo se incluyeron cinco algoritmos para SV, éstos son: Correlation Based Feature Selection (CFS) (Hall, 1999), Consistency Based Filter (Cons) (Dash y Liu, 2003), INTERACT (Zhao y Liu, 2007a), Ganancia de Información (Info Gain) (Quinlan, 1986) y Relief-F (Kononenko, 1994). Por otra parte, se consideraron cuatro clasificadores identificados como: C4.5 (Quinlan, 1993), Naive Bayes (Titterington et al., 1981), KNN (Fix y Hodges, 1951) y SVM (Vapnik, 1995).

Los conjuntos de datos utilizados son cinco en total y provienen de tres repositorios distintos, tres ejemplos se encuentran en el repositorio UCI (Lichman, 2009), uno está en el repositorio bio−médico Kent Ridge de la Agencia para la Ciencia, Tecnología e Investigación (Ridge, 2016) y uno más pertenece al laboratorio de Sistemas de Descubrimiento de la Universidad Vanderbilt, USA (Statnikov et al., 2003). En todos los casos, se consideraron $2/3$ de las instancias como datos de entrenamiento y $1/3$ para validación. El software utilizado durante el estudio es la

plataforma Weka (Hall et al., 2009).

La experimentación inicia con la fragmentación vertical de un conjunto de datos, conformando así diversas particiones con un número equivalente de variables en cada una, éstas son ubicadas aleatoriamente. Posteriormente, se procede a aplicar los cinco selectores mencionados, dónde los primeros tres regresan un subconjunto de variables mientras que los algoritmos univariados Info Gain y Relief-F devuelven un ranking, los autores optaron por establecer un criterio de elección para las primeras p variables al usar estos dos métodos, dónde el valor de p es determinado por el número de variables seleccionadas por el método CFS.

El proceso de aplicación de todos los selectores se repite un total de cinco veces de manera iterativa. En cada iteración − denominada como round − las variables que han sido elegidas para ser removidas reciben un voto, construyendo así un vector de votos.

Al finalizar el procedimiento anterior, se realiza un conteo de votos, identificando a las variables que han recibido una votación por encima de un umbral definido previamente para que sean removidas; se observa que el máximo número de votos corresponde al total de rounds ejecutados, dicho umbral es definido considerando el error de clasificación y el porcentaje de variables retenidas.

Finalmente, la unión simple de las variables que resultan elegidas en cada nodo conforman el subconjunto global de variables seleccionadas; este último es evaluado con los cuatro clasificadores para observar su desempeño y determinar si hay o no mejoría con respecto al obtenido con el conjunto original de datos.

Los resultados muestran que el tiempo de ejecución disminuye considerablemente además de que con respecto a la precisión del clasificador, el enfoque propuesto iguala o mejora a los valores obtenidos al aplicarse a los conjuntos de datos originales.

Se observa que el tiempo requerido para el subproceso de establecimiento del umbral de selección de variables puede ser mayor al tiempo necesario para ejecutar la SV en el conjunto de datos original, es decir, sin particionar. Los autores proponen utilizar alguna medida de complejidad de datos para sustituir su estrategia original en el establecimiento de dicho umbral.

Dada la intención de mejorar el tiempo de procesamiento requerido así como la

precisión de clasificación, los resultados permiten observar que se alcanzó plenamente dicho objetivo. Sin embargo, aunque este trabajo propone un enfoque para el tratamiento de datos distribuido, no busca identificar una nueva clase para los objetos de acuerdo al subconjunto de variables que conforma la representación final de los mismos. Por otra parte el proceso de incorporación de variables de diversas particiones consiste en la unión simple de ellas, por lo que se ignora si alguna de ellas incide más que otra en la clase final o bien si entre ellas existe alguna relación.

3.1.5. Selección de variables distribuida, una aplicación para clasificación de microarreglos de datos

En el documento, se presenta un nuevo método para realizar el tratamiento de conjuntos de datos distribuidos, éste se centra en la aplicación de técnicas de SV antes y después de distribuir los datos, lo anterior en un contexto dirigido al ámbito genético (Bolón-Canedo et al., 2015).

El principal interés de los autores se centra en reducir el tiempo de procesamiento necesario para la SV en conjuntos de datos muy grandes, por lo que se utilizan dos estrategias de particionamiento vertical y un método de mezcla de variables para construir un subconjunto final de ellas.

Los algoritmos de SV que se consideraron son: Correlation Based Feature Selection (CFS) (Hall, 1999), Consistency Based Filter (Cons) (Dash y Liu, 2003), INTERACT (Zhao y Liu, 2007a), Ganancia de Información (Info Gain) (Quinlan, 1986) y Relief-F (Kononenko, 1994); mientras que los clasificadores utilizados se identifican como: C4.5 (Quinlan, 1993), Naive Bayes (Titterington et al., 1981), KNN (Fix y Hodges, 1951) y SVM (Vapnik, 1995).

Se incluyeron ocho conjuntos de datos, aunque en el documento no se menciona la fuente de éstos, si se indica que todos están referidos a un contexto genético de clasificación binaria. Por otra parte, se usó en todos los experimentos la plataforma Weka (Hall et al., 2009).

La experimentación inicia con el particionamiento vertical del conjunto original de datos buscando que los fragmentos contengan el mismo número de variables. Para realizar dicha separación se prueban dos estrategias de fragmentación, (*i*) una de forma aleatoria, en la que las variables pueden pertenecer a cualquier partición sin ningún criterio que indique lo contrario y (*ii*) otra que depende de un ordenamiento previo de las variables, mismo que puede ser ascendente o bien descendente

de acuerdo al mérito derivado de un algoritmo Filter univariado. Una vez elegida la estrategia de división se conforman las diversas particiones, incluyendo un número k de variables en cada una, este valor k se obtiene con la división del número total de instancias presentes en el conjunto original entre dos.

Con las particiones creadas se procede a la aplicación de los algoritmos de SV de forma local, para que posteriormente se efectúe una estrategia de mezcla de variables a fin de construir un subconjunto final de variables.

El proceso de mezcla de variables requiere establecer una partición base sobre la que habrán de agregarse más variables, lo cuál se realiza mediante la satisfacción de un criterio de aceptación, éste consiste en agregar a la estructura base una siguiente partición y evaluar el desempeño de clasificación del nuevo subconjunto con el total de las variables presentes hasta ese momento; si se supera el número de instancias correctamente clasificadas que se obtuvo con la partición base anterior entonces, la nueva partición se acepta, en caso contrario se rechaza.

La definición de la partición base depende de la manera en la que se construyeron los fragmentos, dicha construcción tiene dos posibilidades, en el primer caso (*i*) en el que las variables se hayan asignado a las particiones aleatoriamente, basta con identificar el subconjunto que presenta el mejor desempeño de clasificación para establecer la representación base, mientras que en un segundo caso (*ii*) dónde las variables fueron ubicadas mediante un ranking, entonces se toma el primer subconjunto de variables seleccionadas localmente como el que presenta el mejor desempeño de clasificación, dado que contiene a aquellas con el mayor mérito, ésta constituye la base. Posteriormente, se continúa el proceso de integración como se ha descrito.

El artículo ha propuesto un nuevo método para distribuir el proceso de SV aplicado a un ámbito en particular sobre datos genéticos, los resultados presentan iguales o mejores valores en el desempeño de clasificación con respecto al que se obtiene antes del proceso de la SV con los conjuntos de datos originales. Sin embargo, solo se ha probado con ejemplos referidos a microarreglos de datos genéticos que cuentan con un alto número de variables pero pocas instancias, así como dos clases.

3.1.6. Métodos de selección de variables centralizados vs distribuidos basados en medidas de complejidad de datos

El documento presenta una nueva metodología para realizar SV en conjuntos de datos de alta dimensionalidad mediante dos estrategias de particionamiento y comparar el desempeño de clasificación antes y después del procesamiento (Morán-Fernández et al., 2017).

A semejanza de los dos trabajos previos, en este estudio también se utilizan los selectores: Correlation Based Feature Selection (CFS) (Hall, 1999), Consistency Based Filter (Cons) (Dash y Liu, 2003), INTERACT (Zhao y Liu, 2007a), Ganancia de Información (Info Gain) (Quinlan, 1986) y Relief-F (Kononenko, 1994); así como los clasificadores identificados como: C4.5 (Quinlan, 1993), Naive Bayes (Titterington et al., 1981), KNN (Fix y Hodges, 1951) y SVM (Vapnik, 1995).

En este caso, se utilizaron 11 conjuntos de datos, de los cuales seis provienen del repositorio UCI (Lichman, 2009), uno del repositorio bio-médico Kent Ridge (Ridge, 2016), dos del repositorio de la Universidad de Arizona, USA (Arizona State University, 2016) y dos más del laboratorio de Sistemas de Descubrimiento de la Universidad Vanderbilt, USA (Statnikov et al., 2003).

El proceso de experimentación comienza con la fragmentación de los conjuntos de datos de dos maneras, una vertical y otra horizontal. En el primer caso, las particiones contienen el mismo número de instancias y de variables, las variables en cada fragmento son diferentes y se conserva la clase original; en el segundo caso, las particiones no contienen el mismo número de objetos, se separan en función de la clase y todas las particiones contienen el mismo número de variables, es decir las presentes en la representación inicial.

Posteriormente, se realizan cinco iteraciones, en cada una se aplican los algoritmos selectores a todas las particiones y se construye un vector de variables que almacena el número de votos que cada variable obtiene para ser eliminada del conjunto, es decir, reciben voto aquellas que no fueron seleccionadas por los métodos. Los métodos multivariados devuelven un subconjunto de variables seleccionadas, mientras que los univariados sólo regresan el ranking de ellas, el criterio para elegir las variables derivadas de estos últimos métodos se basa en determinar primero cuántas variables seleccionó CFS, este número de variables se usa para elegir las primeras k variables del ranking correspondiente.

En ambos tipos de fragmentación y a partir de las particiones, las variables

seleccionadas para recibir un voto para eliminarlas del subconjunto final deben cumplir con un criterio basado en la satisfacción de un umbral soportado por tres medidas de complejidad identificadas como (*i*) proporción discriminante de Fisher (F1), (*ii*) longitud de la región superpuesta (F2) y (*iii*) proporción del promedio intra e inter clase de la distancia de vecindad cercana (N2). Finalmente, las variables que conforman el subconjunto final son las que no recibieron voto de eliminación y las que no superaron el umbral en cada una de las particiones.

En general, el número promedio de variables seleccionadas utilizando particionamiento vertical fue menor que al usar separación horizontal, excepto al usar Info Gain y Relief-F ya que éstos seleccionaron menos variables con la distribución horizontal.

En términos de precisión de clasificación, los algoritmos Filter seleccionaron más variables en ambos tipos de fragmentación pero produjeron los mejores desempeños. Con respecto al tiempo de ejecución, éste fue variable según el selector Filter aplicado, en el caso de CONS y Relief-F fueron mayores cuando se utilizó la distribución vertical mientras que para CFS, INTERACT e Info Gain los tiempos fueron más rápidos.

Con respecto a las medidas de complejidad de datos utilizadas con las particiones distribuidas, F1 logró los mejores rendimientos de clasificación (más significativos en la distribución vertical), mientras que F2 seleccionó un número menor de variables.

En este estudio, se concluye que de manera general la precisión del clasificador mejora al usar particionamiento vertical en vez de horizontal.

La evaluación de la efectividad del método propuesto se basa únicamente en la comparación de la precisión del clasificador antes y después de fragmentar los conjuntos originales de datos y aunque existe cierta diversidad en el ámbito de los mismos, mayoritariamente se trata de ejemplos con dos o tres clases, solo hay dos casos cuyo número de clases es 11 y 26. Sería muy valioso saber cómo se comporta en general el método propuesto con conjuntos de datos que incluyen más clases.

3.1.7. Ampliando la selección de variables, un enfoque filtro distribuido

El objetivo de este trabajo es presentar una manera de utilización de métodos Filter de SV en conjuntos de datos distribuidos, lo anterior mediante la fragmentación horizontal de conjuntos de datos grandes, es decir, por instancias, para que en cada una de ellas se aplique un algoritmo Filter de SV y posteriormente, se utilice una estrategia de mezcla de particiones para combinar los resultados en un solo subconjunto de variables relevantes (Bolón-Canedo et al., 2013).

En este trabajo, también se incluyen cinco selectores, tres de ellos multivariados, a saber: Correlation Based Feature Selection (CFS) (Hall, 1999), Consistency Based Filter (Cons) (Dash y Liu, 2003), INTERACT (Zhao y Liu, 2007a), así como dos univariados, éstos son: Ganancia de Información (Info Gain) (Quinlan, 1986) y Relief-F (Kononenko, 1994); de igual manera se utilizan los clasificadores reconocidos como: C4.5 (Quinlan, 1993), Naive Bayes (Titterington et al., 1981), KNN (Fix y Hodges, 1951) y SVM (Vapnik, 1995).

Para este estudio se han utilizado seis conjuntos de datos extraídos del repositorio UCI (Lichman, 2009), dichos ejemplos no están referidos a un ámbito en particular, sin embargo, se observa que mayoritariamente incluyen dos clases aunque hay un ejemplo con tres y otro con 26 clases.

La experimentación se realiza mediante un proceso iterativo, en el que en cada iteración se hace una nueva fragmentación horizontal de los conjuntos de datos. Cada partición generada contiene aproximadamente el mismo número de instancias, las que son elegidas aleatoriamente en cada ocasión. Posteriormente, es aplicado un algoritmo Filter de SV para identificar un subconjunto de variables seleccionadas así como el complemento de ellas, es decir, un subconjunto de variables no seleccionadas, las que reciben un voto como variables rechazadas, mismo que se almacena en un vector de votos.

Finalmente, se realiza un subproceso de evaluación de las variables que han recibido votos para realizar una eliminación definitiva de aquellas que no cumplen con un umbral de rechazo, el cual es calculado mediante un promedio de votación derivado del vector de votos y su desviación estándar correspondiente.

Las variables que no fueron eliminadas conforman el subconjunto final que representa a los objetos en estudio y para validar su conformación se obtiene la precisión de un clasificador, la que es comparada con la que se obtiene utilizando

el conjunto original de datos.

De acuerdo a los resultados, los mejores valores de clasificación se obtienen con el enfoque distribuido a excepción del clasificador SVM. En particular, el selector Relief-F en combinación con el clasificador C4.5 es el que produce el subconjunto de variables con el que se alcanza la más alta precisión, superando un 4 % al mejor resultado del resto de clasificadores.

Con respecto del tiempo de procesamiento, los autores mencionan al selector Relief-F como el algoritmo que reporta la mejor reducción de tiempo requerido a lo largo de toda la experimentación.

En general, al utilizar la distribución de datos se demuestra que el tiempo de procesamiento se reduce significativamente y que la precisión de clasificación se mantiene o incluso mejora en comparación con la aplicación de los algoritmos de SV a los conjuntos de datos no particionados.

Los conjuntos de datos utilizados en la experimentación pertenecen a diferentes dominios. Sin embargo, los conjuntos de datos utilizados cuentan con pocas clases por lo que sería importante probar con ejemplos de mayor número de clases a fin de observar el comportamiento del método propuesto.

Finalmente, se aprecia que en todo momento se busca identificar las diferencias entre el tiempo de ejecución y la precisión del clasificador obtenida antes y después de distribuir el conjunto original de datos y no encontrar un subconjunto de variables que mejor represente a los objetos en estudio.

3.1.8. Selección de variables distribuida, un concepto dudoso de correlación difuso para conjuntos de datos de microarreglos de alta dimensionalidad

En este documento se propone un algoritmo completo para selección de variables distribuida, éste incluye el particionamiento vertical del conjunto de datos original para la aplicación de técnicas de SV en cada fragmento para un posterior proceso de integración que obtendrá un subconjunto global de ellas (Ebrahimpour y Eftekhari, 2018).

Para el desarrollo de la investigación se han incluido los algoritmos de SV identificados como: Correlation Based Feature Selection (CFS) (Hall, 1999), Consis-

tency Based Filter (Cons) (Dash y Liu, 2003), INTERACT (Zhao y Liu, 2007a), así como 2 univariados, estos son: Ganancia de Información (Info Gain) (Quinlan, 1986) y Relief-F (Kononenko, 1994), Maximum Relevancy Minimum Redundancy - Hesitant Fuzzy Sets (MRMR-HFS) (Ebrahimpour y Eftekhari, 2017), Occam´s Razor Feature Selection (OFS) (Ebrahimpour et al., 2017). Por otra parte, se utilizan los clasificadores reconocidos como: C4.5 (Quinlan, 1993), Naive Bayes (Titterington et al., 1981), KNN (Fix y Hodges, 1951) y SVM (Vapnik, 1995).

Los conjuntos de datos utilizados en este estudio son ocho y fueron tomados del repositorio bio-médico Kent Ridge (Ridge, 2016). En todos los casos se corresponden con el ámbito de datos genéticos.

La experimentación comienza con el particionamiento de los conjuntos iniciales de datos, lo que se realiza mediante la aplicación de algoritmos de SV para obtener un ranking de las variables de acuerdo al mérito de cada una, de acuerdo a dicho orden se realiza una fragmentación vertical construyendo subconjuntos de k variables seleccionadas localmente (SVSL). Posteriormente, se aplican todos los selectores a los fragmentos locales para que al finalizar dicho procedimiento se efectúe una etapa de integración de variables.

En la etapa de construcción de la representación final, se toma el subconjunto de variables obtenido de la primera partición −ya que ésta contiene a las variables más significativas− y se obtiene su desempeño de clasificación, éste valor se toma como el valor de clasificación de la precisión base y se utiliza para evaluar el rendimiento obtenido al unir dicha partición con el siguiente SVSL de mejor comportamiento, si se mejora el valor de clasificación entonces dicha partición es aceptada y se le considera la nueva partición base, en caso contrario se rechaza. El proceso continúa hasta evaluar todas las particiones, con lo que se obtiene el subconjunto de variables en su versión final.

Se observa que el algoritmo propuesto presenta mejores resultados al comparar el desempeño del conjunto original de datos con respecto al conjunto con las variables que resultaron de su proceso de integración, se reporta mejor comportamiento con el clasificador NB. Por otra parte, el ámbito de los datos utilizados es muy específico dado que está circunscrito al tema de microarreglos de datos genéticos. Finalmente, los ocho conjuntos de datos aunque se consideran de alta dimensionalidad por contar con un elevado número de variables, no están conformados por muchas instancias.

Se busca en todo momento mejorar el tiempo de procesamiento requerido para realizar la reducción de la dimensionalidad del conjunto original de datos. Por otra parte, los selectores univariados se han configurado para elegir siempre el 10 % de las variables presentes en las particiones, no hay un criterio que indique si este número es el que mejor representa a los datos en cada partición o incluso, en el conjunto original.

También se observa que la estrategia de integración de variables no considera a éstas una a una sino SVSL por SVSL, por lo que el criterio de aceptación de un siguiente SVSL en la partición base consiste en comparar su precisión de clasificación y la que se obtiene al unir estos SVSL y si ésta es mejor entonces se acepta todo el SVSL sino se rechaza, este proceso es iterativo hasta evaluar todos los SVSL. Esto es una diferencia con otros métodos de unión que evalúan variable por variable, con lo que las nuevas representaciones podrían incluir a menos de ellas.

3.2. Discusión del estado del arte

En la literatura se encuentran diversos trabajos que se relacionan con alguna parte de este trabajo doctoral, sin embargo no se ha encontrado ninguno que se desarrolle por completo en un contexto igual al planteado en esta tesis. Derivado de esta situación es que se ha realizado un análisis documental dirigido a aquellos cuyo procesamiento está distribuido y se aplica alguna técnica de SV y/o de clasificación:

La SV se desarrolló originalmente para el tratamiento de conjuntos de datos en particiones únicas, es decir que no se consideraban la fragmentación y/o distribución de los mismos, por lo que diversos investigadores han enfocado sus esfuerzos en el procesamiento de particiones para aprovechar bien sea la capacidad multitarea y/o multiprocesamiento de los equipos de cómputo actuales, así como la utilización de las redes de computadora. La tendencia natural es simplemente procesar partes de una misma base de datos en diferentes ubicaciones.

Se han reportado trabajos donde se aborda la separación de conjuntos de datos en fragmentos con tamaños semejantes, ésto para balancear la carga de trabajo de las computadoras encargadas del procesamiento; las características de estas particiones normalmente incluyen (*i*) diversos individuos descritos con las mismas variables y clasificación en todos los fragmentos, (*ii*) distintos individuos descritos con diferentes variables pero la misma clasificación, o bien (*iii*) los mismos

individuos descritos con diferentes variables en cada fragmento pero la misma clasificación de manera general. Sin embargo, no se ha encontrado algún trabajo que considere el escenario distribuido donde los mismos individuos se encuentran representados en todas las particiones, descritos a partir de diversas variables excluyentes así como con clasificaciones diferentes entre si.

Capítulo 4

Contexto general del método propuesto

Una vez que se ilumina la mente, no se puede oscurecer

Thomas Paine

Como se ha mencionado anteriormente, el constante aumento de los datos disponibles prácticamente sobre cualquier tema ha creado diversas maneras de almacenarlos, generándose así representaciones distribuidas o fragmentadas de los mismos. Sin embargo, una de las razones por las que se requiere su adecuado tratamiento, es para obtener utilidad de ellos, a pesar de que las tareas de recuperación, procesamiento e interpretación sean cada vez más complejas.

Para el procesamiento automático del tipo especial de poblaciones distribuidas motivo de esta tesis es necesario diseñar un método capaz de adecuarse a las características particulares de las mismas, que incluya por una parte, el análisis de las variables presentes en cada partición para determinar cuáles de ellas podrían representar adecuadamente a la población completa y por la otra, hace falta incluir alguna estrategia de integración de dichas variables, a fin de construir el subconjunto final de ellas, así como su correspondiente clasificación global.

En este capítulo se presenta el contexto del método de estructuración de poblaciones distribuidas y para facilitar su comprensión, se incluyen varias secciones para describir (*i*) el escenario correspondiente a su entorno de trabajo, (*ii*) la explicación de las etapas que lo conforman. Finalmente, (*iii*) una discusión del mismo.

4.1. Escenario de aplicación

Las características de las poblaciones distribuidas se ajustan a una diversidad de casos reales, en donde es útil contar con una descripción desde un punto de vista más amplio de los objetos incluidos, además de las ventajas inherentes a la reducción de la dimensionalidad de la representación original.

Generalmente, los ámbitos donde se encuentran las poblaciones mencionadas pertenecen a empresas y organizaciones que con el paso del tiempo han aumentado y diversificado la cantidad de sus datos disponibles, requiriéndose de distintas representaciones de los objetos a los que se refieren sus procesos internos.

El tipo especial de poblaciones distribuidas al que se dirige el método propuesto incluye particiones con diferentes representaciones de sus individuos y aunque pueden existir diversas maneras de abordar su tratamiento automático, éstas podrían variar en función del objetivo que se desee alcanzar. En el caso en particular del método propuesto en esta tesis, se pretende encontrar una representación general de dichos objetos y su correspondiente clase global.

El método de estructuración de poblaciones distribuidas debe procesar entonces particiones locales que incluyan el mismo número de objetos, dónde las variables además de ser diferentes entre sí tanto en significado como en cantidad, también deben ser excluyentes; es decir, aquella que pertenezca a una partición no podrá estar incluida en otra. Los tipos de valores que se soportan por el método pueden ser numéricos o categóricos. Para el caso de las clases, su concepto y cardinalidad tampoco son iguales en cada subconjunto de datos.

Específicamente, el contexto o entorno distribuido descrito se presenta en la Figura 4.1, dónde se observan las características de las poblaciones distribuidas en estudio, entre ellas se aprecian las diferentes variables y clases que la conforman. Además, estas particiones pudieran estar ubicadas en diversos sitios remotos.

Diversas técnicas de RP pueden ser utilizadas para obtener la representación general que se requiere. Sin embargo, no es trivial determinar qué método, herramienta o algoritmo debe emplearse. Los entornos distribuidos con este tipo de procesamiento es un área menos explorada en términos generales que el tratamiento de conjuntos de datos que cuentan con una sola partición.

Debido a lo anterior, el desarrollo de técnicas de RP se originó en un principio

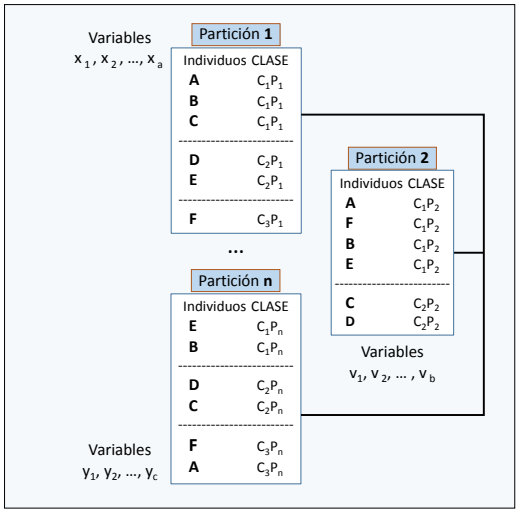

Figura 4.1: Características básicas de las poblaciones distribuidas.

para entornos donde la población de datos se encontraba representada en una sola partición, típicamente se incluían un número conocido de n variables descriptoras y una sola clase disponible.

Los métodos inicialmente desarrollados para RP han tenido que evolucionar, apoyándose en un principio con el uso de herramientas que en el ámbito del tratamiento de bases de datos ya existen, entre las que se encuentran algunas técnicas de fragmentación y posteriormente han llegado a la introducción de nuevos enfoques. Sin embargo, no hay alguno reportado en la literatura que persiga el fin específico de contar con una nueva representación general y una clasificación global

Las técnicas de fragmentación de conjuntos de datos pueden consistir en (*i*) crear subgrupos de variables (verticalmente), es decir, separando y agrupando columnas completas de la representación original, (*ii*) crear diversos subconjuntos de instancias con algún criterio en particular (horizontalmente), por ejemplo las que compartan valores semejantes en alguna(s) variable(s) y finalmente, (*iii*) de forma mixta (verticalmente - horizontalmente), agrupando fragmentos tanto de variables como de instancias, esta operación puede desarrollarse invirtiendo el orden, es decir, realizar primero la fragmentación horizontal y luego la vertical, el resultado final es el mismo, aunque el proceso no.

La fragmentación de conjuntos de datos si bien puede servir para descomponer

un conjunto original en diversas particiones, básicamente se utiliza para aprovechar la capacidad de multiprocesamiento de los equipos de cómputo actuales y acelerar el tratamiento de los datos. Sin embargo, posterior a la utilización de estas técnicas, básicamente la inercia no es proponer una representación global de los objetos sino específicamente enfocarse en la reunión de fragmentos que, dependiendo de cómo se construyeron éstos, con las acciones inversas se recuperaría la estructura original. No hay reducción de la dimensionalidad ni tampoco se sabe si las variables representan adecuadamente a los objetos en estudio, entre otras importantes desventajas.

Es por lo anterior que esta tesis se enfoca en la utilización de técnicas de RP para identificar la representación general de los objetos en observación. Por lo que, además de plantearse preguntas sobre qué técnicas tanto de SV como de Agrupamiento y CS deben utilizarse, también es importante determinar cómo deben ser las estrategias para: (*i*) conformar subconjuntos de variables seleccionadas localmente, *ii*) realizar la integración de variables derivadas de algoritmos de SV aplicados localmente así como, (*iii*) el proceso de validación de la representación global resultante

En complemento al uso de la fragmentación para acelerar el procesamiento de datos, en la literatura se han identificado las características representativas de diversos trabajos que han abordado el tratamiento de datos cuyos contextos incluyen distintas particiones. La Figura 4.2 muestra el escenario clásico encontrado en repetidas ocasiones, para este entorno de trabajo.

En este contexto clásico, algunos de los fragmentos conservan las mismas variables entre sí, mientras que en otras particiones esta situación no sucede. Lo anterior depende del tipo de fragmentación utilizada, ya que si se emplea el criterio de separación horizontal, entonces las particiones incluirán a las mismas variables descriptoras cada uno y si los objetos no están agrupados por clase entonces se tendrán instancias pertenecientes a todas ellas y distribuidas quizá de manera aleatoria. Otra particularidad es que si la fragmentación utilizada es vertical entonces se tendrán distintas variables en cada fragmento pero se conserva la misma clase en cada uno. Un tipo especial se presenta con el particionamiento mixto, en este caso los fragmentos resultantes pueden incluir iguales o diferentes variables y clases entre sí, volviendo complejo el procesamiento.

En resumen, algunas de las particularidades del escenario clásico son las siguientes:

1. Las variables descriptoras pueden cambiar o conservarse en cada partición

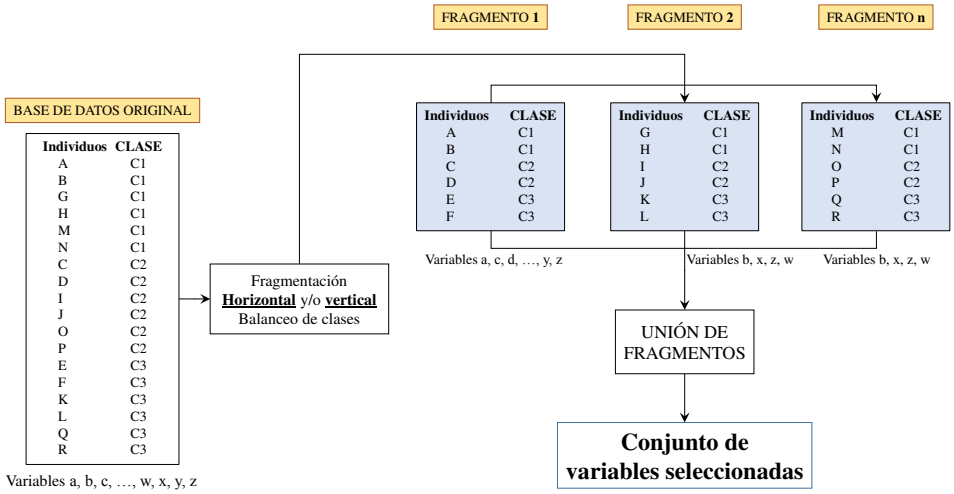

Figura 4.2: Escenario de trabajo clásico para el tratamiento de poblaciones grandes.

generada, esto depende del tipo de fragmentación utilizada.

2. Las clases en cada partición pueden ser las mismas.

3. Puede darse un balanceo en el número de instancias de cada partición, a fin de aprovechar las diferentes capacidades de procesamiento de las computadoras involucradas.

4. Típicamente, se realiza una estrategia de unión simple de variables provenientes de todos los fragmentos.

5. No se realiza la obtención de clases globales, dado que las variables resultantes se corresponden con la única clase conocida al inicio del proceso.

6. No se requiere ninguna validación adicional.

Al observar a mayor detalle, las características mencionadas en la lista anterior, se advierte una cierta tendencia a buscar el procesamiento simultáneo de las particiones involucradas, para después aplicar alguna estrategia de integración de las mismas.

El escenario clásico permite obtener algunas ventajas gracias al procesamiento simultáneo de particiones, especialmente cuando los fragmentos son del mismo tamaño. Sin embargo, esta manera de procesar conjuntos de datos grandes no siempre

se adapta plenamente a la forma en que los datos están distribuidos, pues como se ha mencionado previamente, las poblaciones pueden tener variaciones importantes.

Como se ha observado, existen diferencias entre el entorno distribuido que se aborda en esta tesis y el contexto clásico, mismas que se pretenden atender en la construcción del método de estructuración propuesto. En la sección 4.2 se presenta la descripción de éste.

4.2. Descripción general

A partir de las características del diagrama que se muestra en la Figura 4.1, en esta tesis se propone una alternativa para realizar el tratamiento de datos en poblaciones distribuidas.

Básicamente, la idea es obtener de cada representación inicial, un subgrupo de variables que describan adecuadamente a los individuos que se encuentran ahí representados, a fin de reducir la dimensionalidad del problema. Posteriormente, los subconjuntos de variables seleccionadas de manera local servirán como entrada para una etapa donde se evalúen mediante el cálculo de un factor de correlación, las posibles relaciones entre sus variables, completando con ésto un procedimiento de integración de ellas, con lo que se pretende obtener una representación general con una nueva clasificación global. La Figura 4.3 presenta el contexto completo en el que se desarrolla el método propuesto.

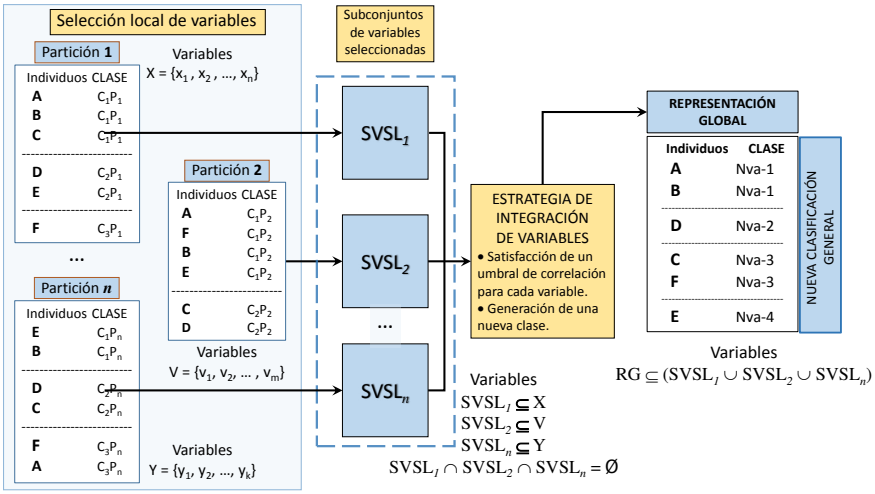

Figura 4.3: Contexto general para el tratamiento de poblaciones distribuidas.

Derivado de lo anterior, el método de estructuración de poblaciones distribuidas propuesto se compone de dos etapas fundamentales, la primera trata de la selección de variables realizada en las particiones de inicio, a esta actividad se le denomina *Selección local de variables*, mientras que la segunda incluye la construcción de la representación general, a la que se le reconoce como *Estrategia de integración de variables*, mediante la que también se obtiene la clasificación global correspondiente.

Ambas etapas requieren la utilización de las técnicas de RP que produzcan los mejores resultados, por lo que al finalizar se incluye un proceso de validación de los mismos.

El diagrama de la Figura 4.4 presenta el esquema general del método de estructuración de poblaciones distribuidas. Se muestran dos bloques principales, referidos a cada una de las etapas mencionadas.

Adicionalmente, al método le antecede una etapa opcional previa, no considerada como parte del mismo pero necesaria en caso de que los datos de entrada requieran alguna adecuación para su correcto tratamiento, a fin de garantizar que los mismos cumplen con la estructura necesaria y en caso de no ser así, serán preprocesados por esta fase de preparación.

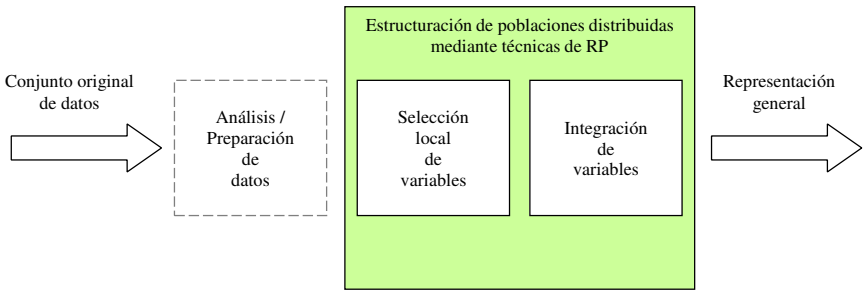

Figura 4.4: Esquema general del método de estructuración de poblaciones distribuidas.

La etapa de selección local de variables consiste en la aplicación de un algorit-

mo Filter univariado a cada una de las particiones presentes, cuyos resultados producirán subconjuntos de variables seleccionadas, determinando así cuáles y cuántas de ellas podrán pasar a la siguiente etapa.

Las variables presentes en los subconjuntos mencionados, seguirán también un orden importante, la primera variable de cada subconjunto será la que presentó mayor importancia de acuerdo a su clasificación local original y así sucesivamente hasta agotar a todas las que el algoritmo Filter seleccionó. La clase de cada subconjunto de variables seleccionadas es la misma que en la partición inicial correspondiente. Cabe mencionar que durante todo el proceso, en ningún momento se considera la eliminación de instancias para reducir la dimensionalidad, por lo que todos los objetos originales están representados.

Hasta este punto, se han creado las condiciones necesarias para evaluar la factibilidad de cada variable en ser incluida como parte de la representación general deseada. Sin embargo, aún hace falta establecer la manera en que la agregación de dichas variables podrá realizarse.

En consecuencia, la etapa de integración de variables comienza con establecer un punto de partida en la construcción de la representación general, y consiste en obtener el desempeño de cada subconjunto de variables seleccionadas localmente, con respecto de un clasificador supervisado. El resultado permitirá establecer cuál es el subconjunto con el mejor comportamiento de clasificación, cuál es el segundo mejor y así sucesivamente hasta contar con una lista ordenada de subconjuntos de acuerdo a este criterio.

El subconjunto de variables seleccionadas localmente con el mejor desempeño de clasificación es considerado como la partición de inicio, a partir de la que serán evaluadas todas las variables del resto de subconjuntos presentes. Cabe mencionar que la clase de la representación base será la que tiene el subconjunto elegido.

Para la evaluación de pertenencia de las nuevas variables en la representación general, se utilizará un segundo criterio, éste permitirá decidir si la variable se acepta o se rechaza, en caso de ser aceptada se debe obtener una clase nueva correspondiente a las variables presentes, para lo que deberá utilizarse un algoritmo de Agrupamiento. Cabe mencionar que si no se acepta la variable evaluada, se conserva la clase actual y se procede a seguir evaluando a las faltantes.

Al concluir el proceso anterior se contará con la nueva representación general

deseada y su respectiva clasificación, la que se corresponde con todas sus variables presentes. Este procedimiento de integración será descrito detalladamente en el Capítulo 7.

4.3. Discusión del contexto

La descripción del contexto en el que la aplicación del método de estructuración de poblaciones distribuidas debe trabajar muestra, por una parte, las diferencias que se encuentran con el procesamiento clásico de datos en entornos con particionamiento tradicional y por la otra, con el tratamiento de representaciones con particiones únicas.

Se observa que las técnicas tradicionales de RP deben adaptarse a los nuevos entornos distribuidos, ya que éstos son cada vez más comunes en una gran cantidad de organizaciones. De igual manera, contar con herramientas que permitan obtener representaciones que provean de un conocimiento más amplio de los objetos en estudio fortalece la mejor toma de decisiones para los usuarios finales de los datos.

El método propuesto aborda el tratamiento automático de poblaciones distribuidas para proporcionar precisamente ese conocimiento que desde un punto de vista más amplio permita describir y catalogar a los objetos en estudio.

En los Capítulos posteriores se presenta el desarrollo de las etapas del método de estructuración propuesto, así como la validación del mismo.

Capítulo 5

Selección local de variables

Las buenas preguntas superan a las respuestas fáciles.
Paul Samuelson

Identificar el subconjunto óptimo de variables que puedan representar adecuadamente a los objetos incluidos en un conjunto de datos, de acuerdo a alguna clasificación previamente conocida, ha sido la motivación de investigadores que han dedicado sus esfuerzos al estudio del comportamiento de éstas, sus características, a las relaciones entre sí mismas, así como a su capacidad informacional; con lo que se ha generado una gran variedad de métodos de SV. Sin embargo, es común observar que cada algoritmo se ha diseñado para alcanzar un fin en particular, por lo que no todos ellos trabajan de la misma forma o producen los mismos resultados.

Dado que los selectores Filter univariados se enfocan en generar un ranking de variables, se esperaría que la aplicación de diversos algoritmos en una misma partición dada, siempre produjera el mismo resultado; en otras palabras, que el ranking derivado de diversos selectores en una misma partición siempre fuera igual e independiente del método que se utilice; en la práctica esto no sucede, pues la construcción interna de cada método puede estar soportada por diferentes conceptos teóricos, lo que en un principio explica la existencia de las diferencias en los resultados. Adicionalmente, hay algoritmos que consideran algunos factores que otros no, por ejemplo la posible correlación individual entre una variable con su clase, la distribución estadística de los valores encontrados en una sola variable a lo largo de todas las instancias evaluadas y la utilización de algunas medidas de tendencia central como la media, mediana o la moda.

Después de obtener un ranking de variables hace falta determinar cuántas de éstas representan adecuadamente al conjunto de datos en estudio, para lo que es común utilizar un algoritmo de clasificación y una medida sobre su desempeño, por ejemplo, el número de instancias correctamente clasificadas reconocido como la *precisión del clasificador*; ésta medida puede servir para implementar un principio de discernimiento en la elección de las variables. La evaluación de rankings típicamente se realiza de acuerdo a algún método de búsqueda que proporciona un orden funcional a fin de que el clasificador proyecte un comportamiento.

Como se mostró en el Capítulo 2, existe actualmente una gran cantidad de algoritmos Filter univariados que han demostrado su eficacia en diversos contextos en particular, sin embargo, éstos generan diferentes rankings entre sí al aplicarlos sobre un mismo conjunto de datos dado. Adicionalmente, también están disponibles muchas técnicas que a partir de un ranking dado, pueden sugerir una regla para la elección de subconjuntos óptimos de variables. Un ejemplo de éstas consiste en establecer un mínimo de las k primeras variables mejor evaluadas.

Dada la amplia variedad de algoritmos de SV, de clasificación así como de métodos de búsqueda disponibles, en esta investigación se requiere determinar para un contexto general ¿Cuál es el método de SV que debe aplicarse en el tratamiento de poblaciones distribuidas? ¿Cómo determinar si algún selector es más adecuado que otro para este entorno?

Las interrogantes mencionadas no tienen una respuesta única dado que de inicio, no se encontraron trabajos que procesen el contexto distribuido tal como se ha planteado en esta tesis. Sin embargo, para determinar la técnica de SV más recomendable para incluirse en la construcción del algoritmo general que procesa poblaciones distribuidas se ha diseñado un proceso de revisión de diversas técnicas Filter univariadas, con el objetivo de realizar una comparación de su efectividad, evaluar su precisión y elegir la más adecuada.

Para soportar la elección del selector se incluye en principio (*i*) una revisión documental de trabajos que comparan diversos algoritmos de SV pertenecientes a diferentes enfoques, posteriormente, (*ii*) la descripción de los conjuntos de datos utilizados en la experimentación desarrollada, (*iii*) el diseño del experimento, (*iv*) el grupo de los algoritmos tanto de SV como de CS incluidos, (*v*) los resultados obtenidos y finalmente, (*vi*) una sección de discusión sobre el selector elegido.

5.1. Estado del arte

Se han revisado trabajos que abordan diferentes enfoques de SV para observar su efectividad con ejemplos de datos de diversa naturaleza y distintos métodos de búsqueda, o bien, se incluyen otros referidos a algún contexto en particular. Se presenta el objetivo de cada investigación, la descripción de sus algoritmos y de los conjuntos de datos utilizados, los resultados conseguidos y un análisis de éstos.

5.1.1. Una revisión de métodos de selección de variables

En este trabajo, se presenta una comparativa entre métodos de SV pertenecientes a distintos enfoques, con el objetivo de mostrar por una parte, una introducción genérica al campo de la reducción de variables demostrando que puede ser aplicado en una gran variedad de problemas de aprendizaje automático y por otra, la aplicabilidad de algunos métodos de SV en conjuntos de datos de diversos contextos incluyendo en particular uno diseñado por los autores (Chandrashekar y Sahin, 2014).

El estudio se centra en la evaluación de cuatro algoritmos de SV, dos Filter y dos Wrapper. Los primeros incluyen al Coeficiente de Correlación de Pearson (Guyon y Elisseeff, 2003) y Cálculo de la Información Mutua (Battiti, 1994), mientras que para los segundos se utilizó Sequential Floating Forward Selection (SFFS) del tipo Sequential Selection Algorithms (Pudil et al., 1994) y otro denominado CHC-GA como una variación de un algoritmo genético Heuristic Search Algorithms (Eshelman, 1991). Para la fase de evaluación del proceso de SV se utilizaron dos clasificadores, siendo éstos SVM (Boser et al., 1992) y Red de Función de Base Radial (Radial Basis Function Network - RBFN o RBF) (Broomhead y Lowe, 1988), (Haykin, 1998).

Para la fase de experimentación, los autores refieren una implementación hecha por ellos mismos de todos los métodos, donde las medidas de desempeño utilizadas fueron por una parte, la predicción de la precisión del clasificador y por la otra, el número de variables eliminadas; los algoritmos Wrapper incluyeron la incorporación del desempeño de los dos clasificadores mencionados como función objetivo.

Se reporta el uso de siete conjuntos de datos, de los cuales cinco provienen del repositorio UCI (Lichman, 2009) y dos más fueron generados por los autores; en estos últimos se incluye intencionalmente el registro de fallos en equipos de la marca MKS Instruments, datos que les eran esencialmente útiles en su ámbito de experimentación.

En la experimentación reportada, se aprecian gráficas de desempeño del clasificador SVM con los selectores tipo Filter basados en el Coeficiente de Correlación de Pearson y el cálculo de la MI, sin embargo no se mencionan los resultados de estos algoritmos de SV con RBF por lo que no se puede apreciar si hay una mejora al usar un clasificador u otro; por otra parte se aprecia que en algunos conjuntos de datos el uso de la MI produce un máximo desempeño de clasificación utilizando menos variables que las existentes en el conjunto original, mientras que en otros ejemplos dicha situación se observa con el uso del Coeficiente de Correlación de Pearson. De acuerdo a esta experimentación, no hay información concluyente para determinar si un selector y/o clasificador es mejor que el otro.

Con respecto a los selectores Wrapper, se muestran los resultados de la aplicación en conjunto con los clasificadores SVM y RBF, sin embargo debido a la manera en que los autores implementaron el algoritmo genético, con éste no se obtuvo una gráfica de desempeño como con SFFS. Con CHCGA, sólo se obtiene el máximo desempeño de clasificación y el número de variables seleccionadas en ese momento. Se observa que el algoritmo SFFS con el clasificador RBF es capaz de obtener el mejor desempeño de clasificación con las primeras variables más significativas mientras que con SVM para alcanzar el mejor comportamiento se requiere un grupo mayor de variables aunque menor al total de ellas. También se aprecia que dos conjuntos de datos tienen un comportamiento lineal por lo que SFFS - SVM no encuentra un máximo desempeño antes de usar el 100 % de las variables.

Los autores muestran los resultados que obtuvieron en algunas de las pruebas realizadas pero no en todas, no se menciona la razón de dichas ausencias, debido a ésto no hay evidencia que permita indicar si algún selector y/o clasificador es mejor que otro, o bien si las técnicas empleadas se relacionan mejor con algún tipo de datos en particular, para conocer esto último se requeriría saber con más detalle la descripción de los conjuntos utilizados, especialmente los de propósito en específico.

5.1.2. Estudio comparativo de métodos de selección de variables utilizando Gain Ratio y CBFS

Se presenta en este estudio una comparativa entre dos métodos Filter, con el fin de mostrar la importancia y aplicación de algoritmos de SV previos a tareas de clasificación, para ello se utiliza un conjunto de datos de un contexto en particular (Gowda et al., 2010).

Los algoritmos de SV incluidos son Gain Ratio (Quinlan, 1993) y Correlation Based Feature Selection (CBFS) (Hall, 1999) así como dos clasificadores supervisados conocidos como Red Neuronal de Retro Propagación (Back Propagation Neural Network - BPNN o BPN) (Kelley, 1960) y RBFN (Broomhead y Lowe, 1988). El conjunto de datos utilizados pertenece a un contexto en particular denominado Pima Indians Diabetes Database (Lichman, 2009) constituido por ocho variables.

En el estudio, se observa que el algoritmo Gain Ratio se usa en un árbol de decisión C4.5 (Quinlan, 1993) como estructura para representar a las variables y determinar el punto de separación de las mismas así como para seleccionar las más importantes, mientras que CBFS es asociado a un algoritmo genético como método de búsqueda, a esta combinación se le denomina GA-CFS; una vez obtenidos los rankings de características se procede a utilizar los clasificadores mencionados.

En los resultados, se observa que Gain Ratio permite seleccionar cinco de las variables originales mientras que CBFS devuelve cuatro de ellas como relevantes, observando que la mejor precisión de los clasificadores se obtiene con la combinación de GA-CFS - BPNN, con un 88 % de instancias correctamente clasificadas. La principal desventaja de Gain Ratio respecto del método CBFS es que el primero no considera las interacciones entre las variables lo que si sucede en el segundo, dado que en este último se evalúa lo peor de un subconjunto de atributos mediante la consideración de la capacidad predictiva individual de cada característica a través del grado de redundancia entre ellas. Así, los resultados experimentales muestran que los subconjuntos de variables seleccionados por el algoritmo tipo Filter CBFS resultaron en el mejor desempeño de los clasificadores BPNN y RBFN para ese conjunto de datos, lo cual no asegura que este resultado sea el mismo al utilizar otros conjuntos de datos.

5.1.3. Un estudio comparativo de técnicas de selección de variables y aprendizaje automático para análisis emotivo

En este trabajo, se presenta una comparativa empírica entre cinco métodos de SV tipo Filter, a fin de explorar su aplicabilidad y determinar la sinergia existente con diferentes clasificadores basados en aprendizaje automático, todos ellos aplicados a un conjunto de datos de un ámbito en particular denominado Análisis de sentimientos o de opinión (Sharma y Dey, 2012).

En los algoritmos utilizados se han considerado los selectores: Info Gain (Quinlan, 1986), Gain Ratio (Quinlan, 1993), Chi Squared (Dunning, 1993), ReliefF

(Kononenko, 1994) y Document Frequency (Yang y Pedersen, 1997), así como los clasificadores: Naive Bayes (Titterington et al., 1981), Support Vector Machine (Boser et al., 1992), Maximum Entropy (Jaynes, 1957), Decission Tree (Quinlan, 1986), K-Nearest Neighbor (Fix y Hodges, 1951), Winnow (Littlestone, 1988) y Adaboost (Freund y Schapire, 1997).

El conjunto de datos utilizado es conocido como Cornell Movie Review Dataset, mismo que fue extraído del sitio de internet Movie Database (IMDb, 2012) y que no se encuentra estructurado con un número de variables de forma tradicional, sino que se diseñó para el tratamiento de texto, cada instancia está representada por una cadena y existen mil clasificadas como positivas y mil como negativas. Para realizar la comparativa fue necesario utilizar un modelo de vector de características, a fin de representar a las variables extraídas del conjunto de datos original, donde cada dimensión de los vectores se corresponden con una característica, obteniéndose diez mil de ellas, ésto permitió la aplicación de los algoritmos de SV.

Durante el proceso, se obtuvo una gráfica de desempeño para cada selector que incluye la precisión de los siete clasificadores versus el número de variables seleccionadas; a partir de estas gráficas, los autores muestran los valores de mejor desempeño por cada par selector-clasificador, encontrándose que Gain Ratio es el que se comporta mejor en todos los casos y que la combinación Gain Ratio-Naive Bayes es la que obtiene el valor más alto. También en los experimentos se puede observar que el clasificador basado en SVM tiene mejor comportamiento independientemente del selector.

Los autores alcanzaron exitosamente su objetivo al aplicar todos los selectores y clasificadores en su conjunto de datos y estudiar la integración de éstos, adicionalmente se observa que el mejor desempeño se obtuvo utilizando un algoritmo tipo Filter de SV derivado de la Teoría de la Información, en combinación con un clasificador basado en un enfoque probabilístico, lo que también demuestra que no necesariamente los algoritmos de SV y de clasificación deben estar construidos a partir de los mismos conceptos teóricos para producir los mejores desempeños. Con esta idea se pueden explorar a mayor detalle otros conjuntos de datos de diversos ámbitos, a fin de determinar si de manera general algún selector y/o clasificador se comporta bien o mejor que otros.

5.1.4. Análisis de técnicas de selección de variables para conjuntos de datos de tráfico de redes

En este trabajo, se realiza un estudio comparativo entre nueve métodos de SV tipo Filter para analizar el efecto que tiene la reducción de la dimensionalidad en relación al número de variables de un conjunto de datos de un contexto en particular, sobre la precisión obtenida por tres clasificadores supervisados a fin de determinar al selector que ofrece el mejor subconjunto de características relevantes Singh et al. (2013).

Sustentándose en la idea de que una sola técnica de SV no garantiza obtener el subconjunto de variables más adecuado, los autores incluyeron los algoritmos Chi Squared (Dunning, 1993), Correlation Feature Selection (Hall, 1999), Consistency Subset (Almuallim y Dietterich, 1991), Gain Ratio (Quinlan, 1993), Filtered (Hall et al., 2009), Filtered Subset, Info Gain (Quinlan, 1986), One R (Holte, 1993) y Symmetrical Uncertainty (Press et al., 1988); los clasificadores utilizados son Naïve Bayes (Titterington et al., 1981), J48 (C4.5) Quinlan (1993) y PART (variación de C4.5) (Quinlan, 1993); en todos los casos la implementación de los mismos se encuentra disponible en el software Weka (Hall et al., 2009). Además de la precisión del clasificador, en cada caso se reporta el número de variables seleccionadas en el punto de máximo desempeño, tiempo de procesamiento requerido y las tasas referidas a Falsos y Verdaderos Positivos (FP y TP por sus siglas en inglés respectivamente).

Para la fase de experimentación se utilizó un conjunto de datos referido a sistemas de detección de intrusiones en el tráfico de redes computacionales, extraído del repositorio UCI (Lichman, 2013), con un total de 41 variables.

Una vez aplicados los algoritmos de SV se utilizaron los tres clasificadores mencionados, encontrándose que el selector multivariado Filtered Subset resultó ser el que obtuvo el máximo desempeño equivalente al 97.552 % eligiendo el menor número de variables de entre todos los métodos, con un 17.07 % del total de ellas, ésto en combinación con el clasificador J48.

No obstante, aunque Filtered Subset elige el menor número de variables en su mejor desempeño, no es el selector que proporciona el máximo valor, de acuerdo a los resultados se observa que el algoritmo que produce el más alto porcentaje de instancias correctamente clasificadas es Consistency Subset con un total del 97.574 %, aunque requiere del 34.14 % de variables seleccionadas, ésto en combinación con el clasificador PART.

Con relación al objetivo que plantearon los autores, se observa que se logró identificar el mejor subconjunto de variables para obtener valores altos en la precisión de los clasificadores, sin embargo únicamente se utilizó un solo conjunto de datos referido a detección de intrusiones en el tráfico de redes, a fin de generalizar más los resultados se pueden incluir más ejemplos de este contexto y que se encuentran disponibles en internet como los proyectos WIDE y WAND entre otros, (Cho, 2016), (Wand, 2016).

5.1.5. Selección de variables basada en dispersión estructurada, un estudio comprensivo

En este trabajo, se aborda un grupo de algoritmos tanto tipo Filter, Wrapper como otros métodos denominados Structured Sparsity-inducing Feature Selection (SSFS), en español Selección de Variables para la Inducción de la Dispersión Estructurada, con el objetivo de proponer un estudio explicativo de las características de ellos, proponiendo en el proceso una taxonomía que explique la evolución de los mismos (Gui et al., 2016).

Los conjuntos de datos usados son 11 y provienen de diferentes fuentes. Uno de ellos se refiere a datos que se derivan de la navegación en páginas web en general y que se registran en una base de conocimiento, otro es sobre clasificación de imágenes, dos tratan sobre datos de rostro, otros dos más son sobre proteínas y vehículos respectivamente y los últimos cinco son de microarreglos de datos sobre diversos tipos de cáncer. A excepción de cuatro de los ejemplos referidos a cáncer que son tomados de lecturas reales en hospitales, el resto provienen de diferentes repositorios disponibles en internet.

Los autores agrupan diferentes métodos SSFS en dos categorías, la primera es la SV basada en vectores y la segunda es SV basada en matrices. Por otra parte, se combinó la SV con otros algoritmos de aprendizaje automático para aplicaciones en específico, tales como (*i*) aprendizaje multitarea, (*ii*) aprendizaje multietiqueta, (*iii*) aprendizaje multivista, (*iv*) clasificación y (*v*) agrupamiento. Se provee una explicación sobre los diversos usos en estas áreas de aplicación y las técnicas utilizadas en ellas.

El trabajo presenta dos grupos de experimentos, en el primero se incluyen nueve métodos de SV tradicionales, siendo éstos Chi Cuadrada (Dunning, 1993), Varianza, Score de Fisher (Duda et al., 2001), Coeficiente de Gini (Gini, 1912), Ganancia

de Información (Info Gain) (Quinlan, 1986), Mínima Redundancia Máxima Relevancia (mRMR) (Peng et al., 2005), Relief (Kira y Rendell, 1992), prueba T-Student (Davis, 1990) y prueba de Kruskal - Wallis (Hollander y Wolfe, 1973). En el segundo grupo se presentan seis métodos, siendo éstos L1 (Destrero et al., 2007), Discriminative Least Squares Regression for Feature Selection (DLSR-FS) (Xiang et al., 2012), Efficient and Robust Feature Selection (RFS) (Nie et al., 2010), Correntropy-Induced Robust Feature Selection (CRFS) (He et al., 2012), l2,0 -Norm Regularized/Constrained Feature Selection (FS20) (Cai et al., 2013) y Unsupervised Discriminative Feature Selection (UDFS) (Yang et al., 2011). En todos los casos, el desempeño de los selectores se evaluó con la precisión de un clasificador basado en SVM (Vapnik y Lerner, 1963).

En la experimentación reportada, se observan todas las curvas de desempeño mediante gráficas que incluyen la precisión del clasificador versus el número de variables seleccionadas, los autores presentan dichas gráficas con ocho números diferentes de variables seleccionadas. El valor correspondiente para la precisión de cada método, se calculó mediante el promedio de 20 lecturas aleatorias realizadas durante el proceso.

Finalmente, se obtienen las siguientes tres conclusiones: (*i*) El algoritmo mRMR es el que se comporta mejor que el resto de los métodos tradicionales, sin embargo no es así con los métodos SSFS, (*ii*) no hay un método simple que sea mejor que todos los involucrados, en orden de comportamiento general, el mejor es DLSR-FS, luego CRFS y en tercer lugar mRMR y (*iii*) los métodos SSFS se comportaron mejor que los métodos tradicionales aunque la diferencia no es significativa.

El documento aborda con un nivel importante de profundidad, la explicación de los métodos tanto tradicionales como los SSFS, esto proporciona una interesante referencia sobre diversos métodos de SV. Con relación al proceso de comparación de los métodos involucrados, se destaca que los conjuntos incluidos son de alta dimensionalidad lo que también permite evaluar el comportamiento de los selectores en este tipo de situaciones.

5.1.6. Una comparación de métodos de selección de variables multi etiqueta utilizando el paradigma de bosque aleatorio

Este trabajo tiene el objetivo de evaluar tres métodos tipo Wrapper para selección de variables, basados en el paradigma Random Forest y probados con siete bases de datos con diferentes dominios (Gharroudi et al., 2014).

Se evaluaron los métodos de selección de variables: BRRF Binary Relevance Random Forest, RFLP Random Forest Label Power-Set y RFPCT Random Forest of Predictive Clustering Trees, se demostró que el paradigma Random Forest es eficiente en procesos de selección de variables supervisados, no supervisados y semisupervisados, estos métodos son de tipo Wrapper.

Se logra determinar que el método BRRF es el más factible para selección de variables multiclase porque considera la relación entre clases.

La medida de importancia de las variables se basa en el decremento del desempeño predictivo y se calcula como el aumento relativo del error que se obtiene en dicha evaluación. Se evalúa el desempeño predictivo comparando el conjunto de variables seleccionadas contra la clasificación original.

Algunas limitantes importantes en este trabajo son que no se considera el procesamiento distribuido, no hay más de una partición que se evalúe simultáneamente, no incluye procesos de selección con métodos Filter lo que puede provocar procesos lentos y de alto costo computacional, además se observa una tendencia al utilizar particiones de datos con balanceo de clases.

5.1.7. Método de selección de variables híbrido para clasificación supervisada basado en el ranking del Score Laplaciano

El objetivo de este trabajo es proponer un método dirigido a Selección de Variables supervisada y basado en la combinación del rankeo de Score Laplaciano con una estrategia Wrapper (Solorio-Fernandez et al., 2010a).

El trabajo considera una sola partición de datos, a la que se aplica un método Filter de selección de variables denominado Score Laplaciano, mismo que obtiene un rankeo de las características y que posteriormente mediante un algoritmo warpper basado en Forward Selection con la variable mejor rankeada como punto de inicio, encuentra el subconjunto óptimo de ellas.

Se evalúa el desempeño predictivo comparando el conjunto de variables obtenidas contra la clasificación original, la precisión del modelo de clasificación se estima usando cross validation con 10 folds.

El autor realiza una comparativa de desempeño comparando este método contra los reconocidos algoritmos Info Gain, Relief, Correlation-based y Wrapper subset

evaluation. En todos los casos se reportan mejores tiempos al procesar diversas particiones de datos.

Con respecto a la efectividad, la comparativa se realizó utilizando los clasificadores C4.5 (decisión tree), Naive Bayes, KNN e Instance-based classifier, también se reportan resultados con alta efectividad.

En este trabajo no se procesan particiones que tengan mayoritariamente a los mismos individuos representados en ellas, no se observa procesamiento distribuido o en paralelo, tampoco se detalla en la fase Filter, el punto de corte a partir del cual se obtiene el mejor subconjunto de variables que posteriormente se someten al proceso Wrapper, sino que se evalúan todas bajo Forward Selection, lo que puede demandar mucho tiempo de procesamiento incluso en particiones de tamaño medio.

5.1.8. Análisis comparativo sobre la estabilidad de técnicas de SV utilizando tres frameworks sobre conjuntos de datos biológicos

Este trabajo, tiene como objetivo la realización de un estudio experimental sobre el desempeño de cinco selectores de variables de tipo Filter. Lo anterior, con el fin de observar la precisión de un clasificador supervisado para evaluar conjuntos de datos cuyo ámbito pertenece a la detección de fallas de software (Wald et al., 2013).

Los algoritmos de SV que se reportan en este estudio son los denominados Chi cuadrada (CS) (Dunning, 1993), Ganancia de Información (Info Gain) (Quinlan, 1986), Razón de ganancia (Gain Ratio - GR) (Quinlan, 1993), Incertidumbre simétrica (SU) (Press et al., 1988) así como Relief (Kira y Rendell, 1992), de este último se utilizaron dos versiones, una con el parámetro "peso de vecinos cercanos por distancia" activado y otra con éste desactivado. Finalmente, se incluyó un algoritmo adicional denominado Signal-To-Noise (S2N) (Chen y Wasikowski, 2008).

Se utilizaron cuatro conjuntos de datos de ámbito genético y con ejemplos de clasificación binaria, los autores no mencionan la fuente de los mismos.

La comparativa de los selectores mencionados comienza con la aplicación de los siete algoritmos a todos los conjuntos de datos, con lo que se obtiene un criterio útil para medir la estabilidad de los métodos. Los conjuntos de datos se sometieron a la inyección de ruido en la clase para volver a procesar la SV y comparar resulta-

dos.

Se utilizan tres frameworks para evaluar la robustez de los rakings de variables bajo diferentes circunstancias y para distintos tamaños de subconjuntos de variables (muestreos limpio vs limpio, ruidoso vs ruidoso y limpio vs ruidoso). Lo anterior, mediante la aplicación de tres técnicas comunes de muestreo (Random Under Sampling - RUS, Random Over Sampling - ROS y Synthetic minority oversampling technique - SMOTE), se evaluaron 24 combinaciones diferentes de niveles y distribución de ruido en todos los conjuntos de datos.

S2N supera al resto de algoritmos de SV en los tres frameworks y en todos los tamaños de subconjuntos, se demuestra que este selector es menos sensible a perturbación de datos sin importar el tamaño de muestreo e inyección de ruido. Sin embargo, Relief está en segundo lugar a lo largo de toda la experimentación, por lo que también es una alternativa de SV confiable en datos del ámbito mencionado. En contraparte, el algoritmo GR es el que presenta el peor resultado en términos de estabilidad y esto ocurre en todos los frameworks, lo que demuestra que es altamente sensible a ruido.

Aunque el estudio demuestra profundidad, las pruebas se dirigieron solamente a conjuntos de datos de un ámbito en particular y con sólo dos clases, sería conveniente probar con otro tipo de variables y mayor número de clases. Adicionalmente, es evidente que en el estudio se pretende obtener alguna conclusión respecto a la sensibilidad de los métodos evaluados con respecto al ruido, de ahí se desprende la elección de S2N en su experimentación. Sin embargo, es importante observar que en términos generales Relief se comporta bien en sus pruebas. Se presentan los resultados promediados de todos los frameworks, lo que valida que Relief también pudiera ser el método elegido para el fin deseado por los autores.

5.1.9. Estudio experimental sobre métodos de SV para detección de fallas de software

La principal intención de este trabajo es efectuar un estudio experimental sobre el desempeño de cinco algoritmos Filter de SV de tipo univariado, ésto sobre conjuntos de datos referidos al ámbito específico de detección de fallas en software (Gnana Singh et al., 2016).

Los algoritmos comparados entre sí son los conocidos: Ganancia de Información (Info Gain) (Quinlan, 1986), Razón de ganancia (Gain Ratio - GR) (Quinlan,

1993), Incertidumbre simétrica (SU) (Press et al., 1988) así como Relief (Kira y Rendell, 1992) y One R (Holte, 1993). Por otra parte, el clasificador incluido es Naive Bayes (NB) (Titterington et al., 1981).

Para la experimentación, se utilizaron 10 conjuntos de datos que pertenecen al mismo dominio, es decir, detección de fallas en software.

La comparativa comienza con la aplicación de los cinco selectores Filter univariados a todos los conjuntos de datos y a partir de los rankings obtenidos por cada uno se seleccionan cinco subconjuntos de variables, cada uno de éstos es evaluado con el clasificador NB mediante la comparación de la precisión obtenida.

El selector One R es el que obtiene en todos los casos los mejores resultados, debido a ésto, los autores lo refieren como el más adecuado para procesar datos referidos a detección de fallas de software.

En el estudio se han utilizado pocos selectores así como un solo clasificador, ésto no asegura que en términos generales One R funciona siempre mejor que los demás. Por otra parte, el número de variables seleccionadas no garantiza que se ha elegido el mejor subconjunto para alcanzar la máxima precisión del clasificador.

5.1.10. Otros estudios comparativos

En la literatura también se encuentran otros trabajos importantes cuyo propósito es estudiar diversos métodos tipo Filter y/o Wrapper, aunque con diferentes objetivos entre sí, uno de ellos por ejemplo, es el compararlos unos con otros sin importar que sean de un enfoque distinto, como se observa en (Bolón-Canedo et al., 2105b), (Singh-Rathore y Gupta, 2014), (Lazar et al., 2012), (Kumari-Bharti y Kumar-Singh, 2014) y en (Saeys et al., 2007).

En uno de los trabajos mencionados anteriormente, se evalúan siete algoritmos de SV tipo Filter denominados Chi Squared (Dunning, 1993), Info Gain (Quinlan, 1986), Gain Ratio (Quinlan, 1993), Relief (Kira y Rendell, 1992), SVM (Vapnik y Lerner, 1963), One R (Holte, 1993) y Principal Component Analysis (Harman, 1960), así como ocho tipo Wrapper conocidos como Classifier SubsetEval - Rank search (Dash et al., 2000), Classifier SubsetEval - Race search based (Dash et al., 2000), WrapperSubsetEval (Frank et al., 2011), FilterSubsetEval - Best search (Frank et al., 2011), FilterSubsetEval - Genetic search (Frank et al., 2011), CFSS subsetEval - Best search (Hall, 1999), CFSS subsetEval - Genetic search (Hall,

1999) y Logistic Regression Analysis (Cheng et al., 2006); mientras que los clasificadores incluidos son Naive Bayes (Titterington et al., 1981) y Random Forest (Breiman, 2001). Los autores incluyeron cuatro conjuntos de datos aunque con modificaciones en ellos se obtuvieron 14 versiones distintas (Singh-Rathore y Gupta, 2014).

En otro de los trabajos mencionados, no se evalúan métodos específicos sino el comportamiento de familias de métodos univariados, bivariados o multivariados, paramétricos o no paramétricos, rankeo de pares y aquellos con una función objetivo para espacio de búsqueda, por lo que no se utilizaron grupos de datos de algún dominio en específico, de igual manera tampoco se refiere alguna evaluación mediante clasificadores sino una medida independiente denominada Coeficiente de Correlación Grupal y el resultado se centra en establecer las técnicas más comunes de validación aplicables a todos los métodos de SV tipo Filter utilizados y en proveer una comparación conceptual entre funciones de score para los métodos (Lazar et al., 2012).

Adicionalmente se reportan otros trabajos donde no se evalúan métodos específicos. En uno de ellos los autores realizan su revisión en función de establecer qué algoritmos tipo Filter se pueden usar para hacer Minería de Textos, sugieren una clasificación que consiste en determinar aquellos que pertenecen al ámbito supervisado o al no supervisado así como si se produce un ranking o un espacio de búsqueda (Kumari-Bharti y Kumar-Singh, 2014).

En otro estudio, la intención es proveer información concreta sobre los enfoques Filter, Wrapper e híbridos además de discutir su uso y variedad en la aplicación de un contexto particular en el área de la Bioinformática (Saeys et al., 2007).

5.2. Comparativa para elegir un selector de variables

En esta sección se presenta el procedimiento utilizado para comparar el comportamiento de diferentes algoritmos de SV, evaluados por diversos clasificadores supervisados, a fin de establecer cuál de ellos logra elegir el mejor subconjunto de variables para obtener los máximos desempeños de clasificación.

5.2.1. Descripción de conjuntos de datos

Todos los ejemplos de datos utilizados, provienen del repositorio de aprendizaje automático UCI Lichman (2009), seleccionándose distintos tipos de ellos con

la intención de observar las diferencias en los resultados proporcionados por los algoritmos de SV y los clasificadores incluidos. En la Tabla 5.1 se presenta la descripción de los 34 conjuntos de datos elegidos para esta experimentación.

El lector encontrará que se incluyen casos con distintas cantidades de instancias, hay algunos que consideran alrededor de 100 instancias, así como otros que cuentan con mas de 30,000 objetos. Con respecto del total de variables, la base de datos más pequeña tiene 7, mientras que el caso mayor involucra tratar a 500 de ellas.

Por otra parte, las etiquetas de clase en cada conjunto de datos también es un aspecto a tomar en cuenta, y aunque en la mayoría de casos se trata de clasificación binaria, también se agregaron ejemplos con decenas de clases.

La razón para lo anterior, se debe a que se ha considerado valioso para este estudio contar con conjuntos de datos con un número representativo de variables y de instancias, para observar el comportamiento de los métodos al aplicarlos en ejemplos de alta, media y baja dimensionalidad.

Con respecto a los nombres de las variables y para estandarizar la fase de experimentación, estos fueron cambiados en todos los casos por nombres genéricos identificados por X = $\{x_1, x_2, x_3, x_4, ..., x_n, clase\}$ donde n corresponde al total de ellas en cada conjunto de datos.

Lo que se pretende en este estudio es identificar al algoritmo de SV que se incluirá en el método de estructuración de poblaciones distribuidas. Lo anterior para construir subconjuntos de variables seleccionadas localmente, es decir en cada una de las distintas particiones que conforman una población distribuida, a fin de identificar a las que mejor representarán a la población completa en una representación general.

En esta experimentación, solamente los ejemplos Iris y Madelon incluyen clases balanceadas, se eligieron a fin de determinar si esta característica sugiere algún comportamiento en específico para este tipo de base de datos. En el resto de casos, no se presenta balanceo.

Tabla 5.1: Descripción de los conjuntos de datos procesados

	Conjuntos de datos	Instancias	Variables	Clases
1	Abalone	4,177	8	28
2	Adults	48,842	14	2
3	Cylinder Bands	539	39	2
4	Bank	4,521	16	2
5	Bank Full	45,210	16	2
6	Breast Cancer	286	9	2
7	Car Evolution	1,728	6	4
8	Chess	3,196	36	2
9	Congressional Voting Records	435	16	2
10	Default Credit Card Clients	30,000	23	2
11	Dermatology	366	34	6
12	Ecoli	336	7	8
13	Egg Eye	14,980	14	2
14	Geographical Music Chromatic	1,059	116	33
15	Geographical Music Simple	1,059	68	33
16	German Credit	1,000	20	2
17	Glass	214	9	6
18	Hepatitis	155	19	2
19	Horse Colic	300	27	2
20	Iris	150	4	3
21	Letter Recognition	20,000	16	26
22	Lymphography	148	18	4
23	Madelon	2000	500	2
24	Messidor features	1151	19	2
25	MG Telescope	19,020	10	2
26	Mushroom	8,124	22	2
27	Nursery	12,960	8	5
28	Primary Tumor	339	17	21
29	Sensorless	58,509	48	11
30	Statlog - Australian Credit	690	15	2
31	Tic tac toe	958	9	2
32	Wisconsin Breast Cancer	699	10	2
33	Year Polish	43,405	64	2
34	Zoo	101	17	7

5.2.2. Algoritmos evaluados

Para la evaluación de métodos de SV, se incluyeron los ocho selectores conocidos como: Chi Squared, Gain Ratio, Info Gain, Laplacian Score, One R, Relief, SVM y Symmetrical Uncertainty. En todos los casos se trata de algoritmos de tipo Filter univariados y debido a que están diseñados con base en diferentes conceptos teóricos, el ranking que cada uno produce será diferente entre sí, lo que permitirá la comparación sin ventaja para alguno de ellos.

Los clasificadores utilizados para evaluar las listas de ranking son seis, se incluye entre ellos (*i*) una modificación del algoritmo SVM tradicional denominado Sequential Minimal Optimization (SMO) (Platt, 1998), (*ii*) un método de tipo probabilístico basado en el Teorema de Bayes (Naive Bayes - NB) (Titterington et al., 1981), (*iii*) dos métodos de tipo perezoso basados en el algoritmo del vecino más cercano (K Nearest Neighbor - KNN) (Fix y Hodges, 1951) identificados como KStar (Cleary y Trigg, 1995) e Ibk (Aha y Kibler, 1991), (*iv*) otro construido sobre C4.5 con una variante soportada por grafos, conocido como J48Graft (Quinlan, 1993) y (*v*) uno más diseñado sobre combinación de filtro de instancias - clasificador identificado como Filtered (Hall et al., 2009).

Los clasificadores mencionados, se eligieron por estar construidos con distintas bases teóricas y por ser reconocidos en la literatura correspondiente, como algoritmos con alta efectividad, de fácil implementación y de uso muy extendido.

Para todos los algoritmos de SV y CS, se utilizó la versión disponible en la plataforma Weka (Hall et al., 2009), a excepción del selector Laplacian Score, éste último requirió la implementación del algoritmo publicado por Solorio-Fernandez et al. (2010b), quién proporcionó el código fuente y colaboró en el proceso para obtener una versión ejecutable, compatible con la versión vigente de Java Virtual Machine (Oracle-Corporation, 2017).

5.2.3. Experimentación

Para implementar la fase de experimentación, se requiere una estrategia de búsqueda que permita la evaluación de subconjuntos de variables por todos los clasificadores, el método elegido es el denominado SBE mencionado en la sección 2.1, entre otras razones porque no se sabe a priori, con cuantas variables de cada conjunto se encontrará el mejor desempeño del clasificador, se presupone que en la generalidad, este valor máximo se identificará con la mayoría de ellas.

El método SBE, selecciona iterativamente subconjuntos de variables a partir de un ranking, el procedimiento comienza considerando al total de ellas para determinar el desempeño de algún clasificador, mediante la obtención del número de instancias correctamente clasificadas; en una siguiente iteración elimina a la primer variable menos significativa, lo que supone que no debería afectar significativamente la precisión del clasificador, el proceso continúa eliminando la siguiente variable en el ranking y así sucesivamente hasta haber eliminado a todas. En alguna de las iteraciones el desempeño de clasificación alcanza su valor máximo y posteriormente comienza a decrecer hasta llegar a un nivel mínimo.

Con la idea de aplicar los diferentes algoritmos de SV y de clasificación a todos los conjuntos de datos, se diseñó un procedimiento que permita para cada uno de ellos (*i*) utilizar los ocho selectores de variables, (*ii*) obtener los correspondientes ranking, (*iii*) determinar la precisión de los seis clasificadores para diferentes subconjuntos de variables mediante el método de búsqueda SBE, (*iv*) tabular y graficar los resultados para finalmente, (*v*) realizar el análisis, comparativa e interpretación de los resultados obtenidos. La Figura 5.1 muestra el diagrama de flujo correspondiente a este procedimiento.

Por otra parte, la medida de evaluación utilizada es la precisión del clasificador, la que se reconoce como el porcentaje de instancias correctamente clasificadas, con respecto del total de las existentes en cada conjunto original de datos.

Este algoritmo se ha diseñado para aplicarse con cualquier selector Filter y clasificador supervisado, no solamente con los algoritmos que se utilizarán en esta experimentación. Adicionalmente, el método de búsqueda SBE puede ser sustituido por SFS o algún otro, preferentemente de tipo secuencial.

La idea de aplicar en paralelo los seis clasificadores supervisados, permitirá tabular simultáneamente los resultados obtenidos en cada iteración. Sin embargo, no todos los clasificadores realizan su trabajo en el mismo tiempo de ejecución y considerando que el algoritmo se basa en un proceso de búsqueda exhaustiva, la duración de las pruebas requiere una cantidad considerable de tiempo. El objetivo de esta prueba no se centra en el rendimiento temporal sino en la efectividad obtenida, por lo que no se reportará la eficiencia sino la eficacia alcanzada en la experimentación.

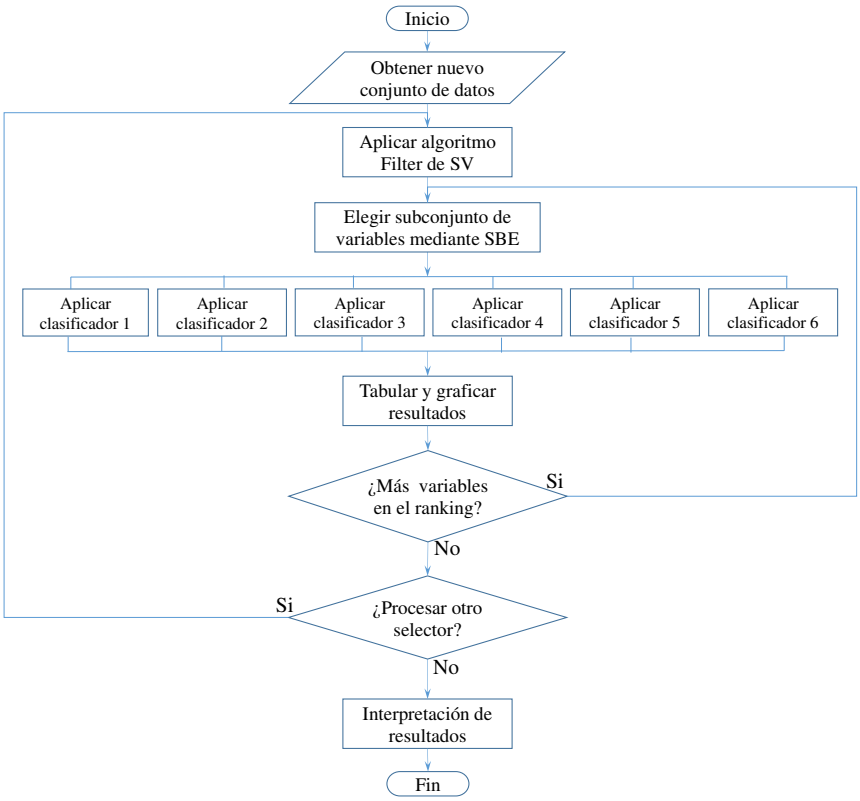

Figura 5.1: Procedimiento para aplicar algoritmos de SV y seis clasificadores.

La aplicación de este procedimiento se describe a detalle en la subsección siguiente, el proceso es semejante para cada uno de los grupos de datos involucrados. Se presentan en una primera instancia, los resultados obtenidos para el conjunto Messidor. Posteriormente se muestra la comparativa resumida de todos los métodos para el total de los conjuntos de datos incluidos.

5.2.4. Resultados

De acuerdo al procedimiento referido en la Figura 5.1, debe realizarse la aplicación de todos los selectores a cada conjunto de datos. Para mostrar el procedimiento completo de un conjunto de datos se eligió la base de datos Messidor, la cual trata de un conjunto con sólo 19 variables, las cuales pueden ser mostradas adecuadamente en las tablas y gráficas de resultados. La Tabla 5.2 muestra la lista

de los rankings obtenidos por los ocho métodos de SV mencionados. Cabe mencionar que cada ranking muestra sus resultados en orden descendente, de acuerdo a la importancia identificada para cada variable.

Tabla 5.2: Ranking de variables del conjunto de datos *Messidor* obtenidos de los algoritmos de SV

Chi Squared	Gain Ratio	Info Gain	Laplacian Score	OneR	Relief	SVM	Symmetrical
x_3	x_{15}	x_3	x_1	x_3	x_3	x_3	x_{15}
x_{15}	x_{16}	x_{16}	x_2	x_{15}	x_4	x_7	x_{16}
x_{16}	x_1	x_{15}	x_{19}	x_4	x_{19}	x_4	x_{14}
x_4	x_{14}	x_4	x_{14}	x_5	x_9	x_9	x_3
x_{14}	x_3	x_{14}	x_{15}	x_{16}	x_5	x_{15}	x_4
x_5	x_{13}	x_5	x_6	x_6	x_6	x_{10}	x_5
x_6	x_4	x_6	x_5	x_{14}	x_{10}	x_8	x_6
x_9	x_5	x_9	x_7	x_{13}	x_{15}	x_{16}	x_{13}
x_{13}	x_6	x_{13}	x_4	x_8	x_7	x_6	x_7
x_7	x_7	x_7	x_{13}	x_9	x_8	x_{13}	x_9
x_8	x_9	x_8	x_3	x_7	x_{11}	x_5	x_8
x_2	x_8	x_2	x_{16}	x_1	x_{12}	x_2	x_1
x_1	x_{12}	x_1	x_8	x_2	x_{16}	x_{14}	x_2
x_{19}	x_2	x_{19}	x_{12}	x_{12}	x_{14}	x_{11}	x_{19}
x_{12}	x_{19}	x_{12}	x_{11}	x_{17}	x_2	x_{18}	x_{12}
x_{18}	x_{18}	x_{18}	x_9	x_{19}	x_{13}	x_1	x_{18}
x_{17}	x_{17}	x_{17}	x_{10}	x_{11}	x_{18}	x_{12}	x_{17}
x_{11}	x_{11}	x_{11}	x_{17}	x_{10}	x_1	x_{17}	x_{11}
x_{10}	x_{10}	x_{10}	x_{18}	x_{18}	x_{17}	x_{19}	x_{10}

Cada método de SV, obtiene un ranking en particular para este conjunto de datos, aunque pudieran existir algunas coincidencias entre ellos no son necesariamente iguales. El hecho de que cada uno sea diferente sugiere la interrogante ¿Cuál método obtiene un mejor rankeo de variables? La respuesta a esta pregunta es la razón por la que se comparan entre sí diversos algoritmos selectores.

A continuación, se procede a evaluar los ocho rankings derivados de los algoritmos de SV, esto se realiza de acuerdo al método de búsqueda SBE, enviando los subconjuntos de prueba a los seis clasificadores. Para cada ranking el proceso ini-

cia utilizando a todas las variables, en una segunda iteración se elimina la variable menos significativa y ésto se repite hasta terminar de evaluar al total de variables. Al finalizar, se disponen todos los resultados en una tabla, para facilitar la observación del comportamiento de todos los clasificadores.

Como se puede observar, la Tabla 5.3 muestra la precisión obtenida por los seis clasificadores utilizando el ranking proporcionado por el algoritmo Chi Squared, con el conjunto de datos Messidor. En esta tabla, se resalta la precisión más alta obtenida por cada algoritmo clasificador. Sin embargo, no todos los clasificadores obtienen porcentajes de precisión similares. Algunos de ellos presentan mejores resultados que otros, ésta es una de las razones para comparar los resultados a través de diversos algoritmos de clasificación.

Tabla 5.3: Precisión de clasificadores mediante *SBE*, ranking derivado del selector *Chi Squared*, conjunto de datos *Messidor*

Iteración	Variables	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
1	Todas	**67.593**	56.820	61.251	64.639	61.338	63.249
2	Sin x_{10}	67.333	56.820	61.338	64.292	60.122	63.249
3	Sin x_{11}	66.725	56.646	62.467	64.813	59.340	63.249
4	Sin x_{17}	66.985	56.646	60.904	64.726	62.207	63.249
5	Sin x_{18}	67.159	56.646	63.162	64.987	67.333	63.249
6	Sin x_{12}	66.725	56.386	64.639	**65.161**	**67.593**	63.249
7	Sin x_{19}	67.246	56.386	64.466	64.205	65.508	63.249
8	Sin x_1	67.246	56.386	64.379	64.205	65.421	63.249
9	Sin x_2	65.074	56.299	63.944	63.076	66.030	63.249
10	Sin x_8	64.813	56.125	64.813	64.466	66.203	**63.423**
11	Sin x_7	64.031	56.386	63.510	64.118	65.421	63.336
12	Sin x_{13}	64.118	56.907	**65.421**	64.553	65.595	63.336
13	Sin x_9	63.510	56.820	64.987	63.336	66.811	62.815
14	Sin x_6	60.904	56.560	63.684	62.815	67.072	62.641
15	Sin x_5	60.122	56.212	62.467	60.990	64.205	62.815
16	Sin x_{14}	60.122	57.081	62.120	60.990	65.769	62.989
17	Sin x_4	60.295	56.386	60.904	61.685	58.645	62.815
18	Sin x_{16}	60.122	57.255	61.772	61.685	58.384	62.554
19	Sin x_{15}	59.253	**59.166**	56.212	60.990	59.687	59.687

Como se observa en la Tabla 5.3, el subconjunto de variables con el que se obtiene el mejor desempeño de clasificación, no es el mismo para cada clasificador. En particular, SMO indica que el conjunto completo de variables es el mejor subconjunto. Naive Bayes muestra que el mejor subconjunto se integra de una sola variable, en este caso se trata de x_3. KStar muestra que en su caso, el subconjunto óptimo incluye las ocho variables x_3, x_{15}, x_{16}, x_4, x_{14}, x_5, x_6 y x_9. Por otra parte, J48 e Ibk eligen un subconjunto con las 14 variables x_3, x_{15}, x_{16}, x_4, x_{14}, x_5, x_6, x_9, x_{13}, x_7, x_8, x_2, x_1 y x_{19}. Por último, Filtered sugiere que el mejor, es el subconjunto conformado por las 10 variables x_3, x_{15}, x_{16}, x_4, x_{14}, x_5, x_6, x_9, x_{13} y x_7.

De acuerdo a los resultados anteriores, es necesario establecer una estrategia que permita elegir la combinación selector - clasificador que mejor resultados proporcione, de acuerdo a los diferentes conjuntos de datos de experimentación. Por lo que de manera análoga, a partir de los rankings correspondientes al resto de los selectores aplicados a este mismo conjunto de datos, se han obteniendo las tablas que completan el proceso de evaluación.

Posterior a la construcción de las tablas de resultados, se han construido dos tipos de gráficas a fin de observar el comportamiento de los clasificadores a lo largo de toda la experimentación, definiéndose los siguientes criterios de observación:

1. Desempeño de clasificadores por el Ranking de cada algoritmo de SV, en donde se muestra el porcentaje de clasificación versus subconjuntos de variables creados de acuerdo al ranking, la gráfica incluye seis series de datos, cada una correspondiente a un clasificador.

2. Análisis de desempeño por clasificador, en donde se considera el porcentaje de clasificación versus porcentaje de variables seleccionadas, cada gráfica incluye ocho series de datos, cada serie corresponde a un ranking.

En el primer tipo de gráfica, puede observarse cuál de los clasificadores produce el valor de máximo desempeño para un ranking dado, así como cuántas y cuáles variables se requieren para alcanzarlo. Un ejemplo puede verse en la Figura 5.2, dónde se aprecia el comportamiento de los seis clasificadores de acuerdo al ranking proporcionado por el algoritmo Relief en el conjunto de datos Messidor.

En la Figura 5.2, el eje de ordenadas muestra las variables que están siendo evaluadas en cada iteración, así el primer punto corresponde a la evaluación con todas ellas, el segundo se obtiene sin considerar la variable x_{17} por ser la menos significativa, y así sucesivamente estas iteraciones continúan hasta eliminar a la variable más significativa (x_3). Se observa que para este ejemplo, el clasificador KStar que

al inicio de la evaluación comienza obteniendo una precisión de clasificación de 61.251 % de instancias correctamente clasificadas considerando a todas las variables, es el que produce el mejor desempeño, equivalente al 69.157 % cuando ha eliminado del conjunto original a trece de ellas y ha conservado a seis, (x_6, x_5, x_9, x_{19}, x_4 y x_3). Siendo esta última la más significativa.

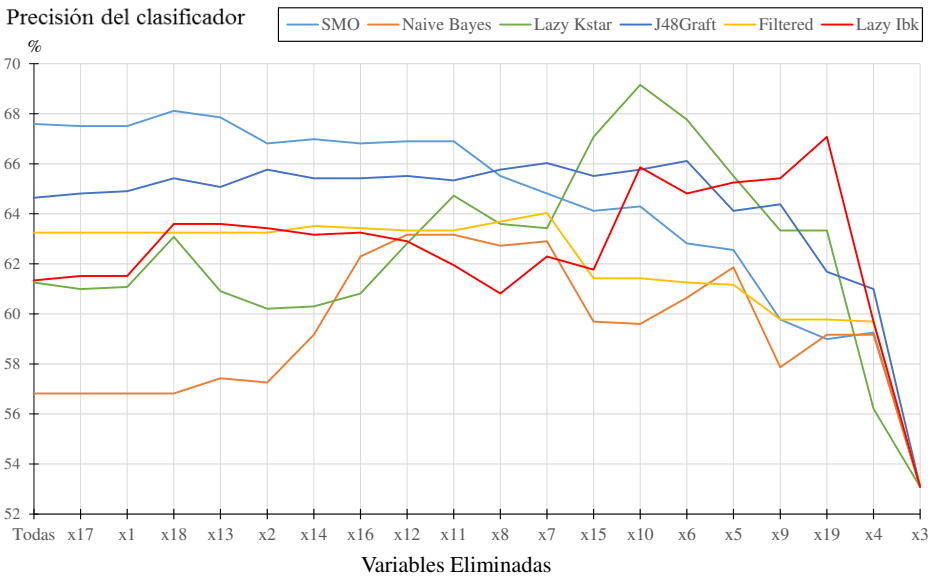

Figura 5.2: Curvas de desempeño de los seis clasificadores, conjunto de datos *Messidor*, ranking *Relief*.

En el segundo tipo de gráfica, se puede observar el desempeño de un mismo clasificador de acuerdo a cada uno de los rankings utilizados, un ejemplo se aprecia en la Figura 5.3, éste corresponde a la evaluación del clasificador KStar de acuerdo al ranking producido por los ocho métodos tipo Filter involucrados; en el eje de ordenadas se muestra el porcentaje de variables seleccionadas partiendo desde cero y hasta llegar al 100 % de ellas, el resultado confirma que el selector Relief produce el mejor desempeño de este clasificador 69.157 % utilizando un 31.6 % de variables seleccionadas.

El comportamiento ideal sugiere que la curva aumenta súbitamente al inicio de la gráfica y obtiene su valor máximo, comenzando posteriormente un descenso que se mantiene al llegar al total de variables seleccionadas, por tanto se busca que el

selector de variables a elegir como el mejor pueda producir ese desempeño.

De manera semejante, se han obtenido todas las gráficas correspondientes a la evaluación de los seis clasificadores para cada uno de los ocho algoritmos de SV estudiados, por lo que ahora es importante realizar la comparativa que permitirá saber que combinación de selector - clasificador obtiene la mejor clasificación en cada conjunto de datos observado.

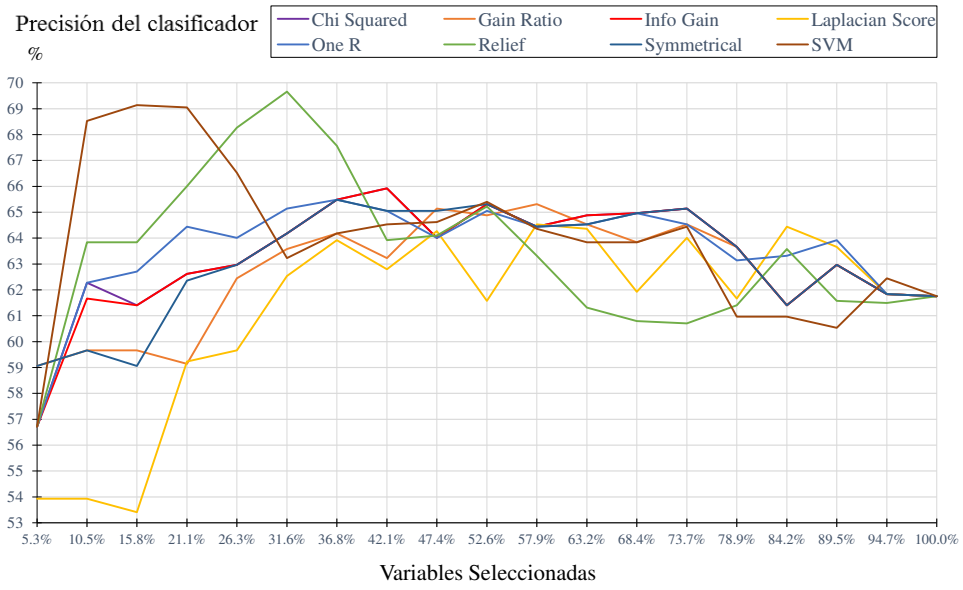

Figura 5.3: Comportamiento del clasificador *KStar* para los ocho rankings del conjunto de datos *Messidor*.

Posteriormente, para realizar la comparativa deseada, se han concentrado en una nueva tabla, los valores de máximo desempeño obtenidos por un mismo clasificador siguiendo el ranking producido por cada algoritmo de SV incluido, con lo que es posible observar de entre todos ellos, el valor más alto. Un ejemplo de esta comparación se muestra en la Tabla 5.4, éste corresponde al conjunto de datos Messidor, donde de manera concluyente se encontró que es el clasificador KStar es el que presenta el mejor desempeño en combinación con el selector Relief, con un total 69.157 % de instancias correctamente clasificadas.

Tabla 5.4: Resumen de máximo desempeño de todos los selectores - clasificadores, conjunto de datos *Messidor*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	67.593	59.166	65.421	65.161	67.593	63.423
Gain Ratio	67.593	56.820	64.813	64.987	67.333	63.423
Info Gain	67.593	59.166	65.421	65.161	67.593	63.423
Laplacian Score	67.941	57.689	64.031	66.725	66.203	63.249
One R	67.941	61.946	64.987	65.074	68.97	63.249
Relief	68.115	63.162	69.157	66.116	67.072	64.031
SVM	68.636	62.294	68.636	66.116	64.726	63.944
Symmetrical Uncertainty	67.593	56.820	64.987	65.161	67.593	63.423
Máximo por método	68.636	63.162	69.157	66.725	68.897	64.031
Máximo global	**69.157**					

Este mismo proceso se realizó para todos los conjuntos de datos (ver Apéndice A), por lo que se ha podido determinar en cada uno a la combinación selector - clasificador que produce los mejores desempeños, encontrándose que para algunos casos continúa siendo Relief - Kstar, sin embargo, hay casos donde Relief aunque sigue siendo el mejor selector obtiene mejor desempeño con el clasificador Ibk, incluso, se identificó un caso donde KStar se comporta mejor que el resto de clasificadores pero no con el ranking Relief sino con Info Gain o con One R.

El procedimiento realizado para establecer un criterio que permita seleccionar una combinación de selector - clasificador que represente la experimentación completa, consiste en construir una tabla formada por los valores promedio de los resultados parciales obtenidos en cada uno de los doce conjuntos de datos, a partir de estos valores es posible elegir el valor más alto y que corresponderá al máximo porcentaje de instancias correctamente clasificadas de acuerdo a un ranking aplicado al total de los conjuntos de datos.

Los valores finales se observan en la Tabla 5.5, éstos sustentan la elección del algoritmo Relief como el que produce el ranking que mejor desempeño obtiene con todos los clasificadores utilizados. Se observa que en cuatro de los seis clasificadores, se presenta el mejor comportamiento y específicamente con KStar, es donde se alcanza el máximo global.

Tabla 5.5: Promedio de clasificación en todos los conjuntos de datos evaluados

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	70.624	68.125	74.244	72.676	75.260	68.291
Gain Ratio	70.832	67.412	74.214	72.757	75.440	68.518
Info Gain	70.624	68.120	74.318	72.849	75.017	68.669
Laplacian Score	61.726	56.589	63.127	62.588	64.174	58.402
One R	70.443	69.095	74.559	72.781	75.402	68.516
Relief	70.733	68.817	76.021	73.237	75.654	68.725
SVM	62.251	61.366	66.208	65.585	64.865	63.296
Symmetrical Uncertainty	70.728	67.427	74.270	72.739	75.260	68.518
Máximo por método	70.832	69.095	76.021	73.237	75.654	68.725
Máximo global	**76.021**					

A partir de los resultados mostrados en la Tabla 5.5, se ha obtenido la gráfica de la Figura 5.4, que representa el comportamiento de todas las combinaciones selector - clasificador para el total de conjuntos de datos. La serie con mejor desempeño corresponde al selector Relief, encontrando su valor máximo con el clasificador KStar.

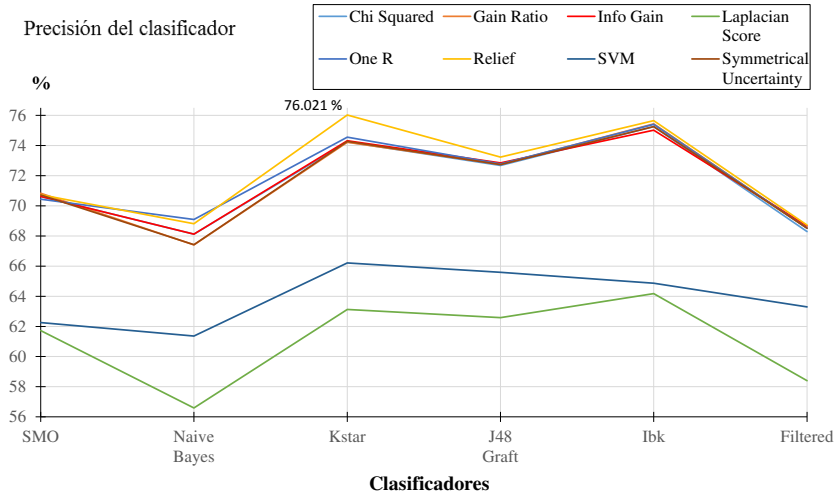

Figura 5.4: Resultados globales de comportamiento de selectores - clasificadores en todos los conjuntos de datos.

Es evidente bajo el sustento de los resultados obtenidos, que el clasificador KStar obtiene el valor máximo de clasificación general al combinarse con el selector Relief con un valor de 76.021 % de instancias correctamente clasificadas al considerar a todas las bases de datos. Por otra parte, al utilizar el mismo selector pero en combinación con el clasificador Ibk, se obtiene un 75.654 % de desempeño global, la diferencia es apenas del 0.367 %.

Ahora bien, los seis selectores Chi Squared, Gain Ratio, Info Gain, Laplacian Score, One R y Symmetrical Uncertainty, obtienen mejor desempeño con Ibk que con KStar y solamente con KStar y SVM es el caso contrario. Al calcular y promediar las diferencias en el desempeño de ambos clasificadores y restar del rendimiento de Kstar, el valor obtenido con Ibk, resulta un promedio de 0.514 %, indicando que en general, hay mayor precisión con Ibk que con KStar.

Dado que el objetivo de este trabajo, no se centra en evaluar el tiempo de ejecución de los algoritmos, no se han presentado datos temporales en el desempeño de los métodos, pero se ha observado durante toda la experimentación, que Ibk es sustancialmente más rápido que KStar y la eficacia aunque es menor, no tiene una diferencia significativa.

Derivado de lo anterior, la mejor combinación a lo largo de la experimentación con los 34 conjuntos de datos, se ha determinado que para continuar con esta investigación, se elige a la combinación selector - clasificador dada por Relief - Ibk.

5.3. Discusión del método elegido

En este trabajo, se ha desarrollado un estudio y comparativa entre ocho métodos de SV tipo Filter, considerando algunos de los más reportados en la literatura, para lo que se ha utilizado como medida de desempeño la precisión obtenida por seis clasificadores supervisados. Durante la revisión del estado del arte, se ha encontrado que existe cierta tendencia a evaluar diversos algoritmos para un conjunto de datos en un contexto en particular por lo que con la intención de que la validez de los resultados pueda ser considerada de forma más general, se han incluido 34 conjuntos de datos pertenecientes a diferentes contextos y se ha decidido que éstos provengan de un repositorio revisado y dirigido a tareas de clasificación supervisada.

El algoritmo Relief para SV ha mostrado ser el que produce un ranking donde la mayoría de clasificadores obtienen el mejor desempeño, produciendo el valor

más alto con los clasificadores KStar, J48, Ibk y Filtered. Particularmente, se ha elegido la combinación Relief - Ibk como la dupla con la que se obtiene el mejor comportamiento de toda la comparativa.

Un resultado importante en general, es que seis de los ocho algoritmos de SV producen un desempeño en los clasificadores muy semejante, por encima de los selectores SVM y Laplacian Score, siendo este último con el que se obtuvo la menor capacidad de clasificación en general.

Con relación al comportamiento de los clasificadores, se observa que Ibk produce los más altos desempeños con seis de los ocho rankings evaluados, la excepción es con Relief y con SVM donde KStar es el mejor. Lo interesante de esta observación es que tanto KStar como Ibk son métodos basados en el algoritmo del vecino más cercano, con lo que se demuestra que de acuerdo a la literatura relacionada, se comprueba que estos clasificadores producen buenos resultados y que de alguna manera no se depende del ranking y/o el conjunto de datos utilizados. Por el contrario el clasificador Naive Bayes produce los mínimos porcentajes de clasificación en seis de los ocho rankings evaluados.

La estrategia utilizada y los resultados obtenidos, demuestran que Relief es un selector fuertemente recomendable en procesos de SV sin importar el contexto de los datos utilizados y que esta misma idea puede ser ampliada para comparar más selectores incluso considerar otros algoritmos Filter multivariados, así como, otros clasificadores. Además, es posible incluir una variedad mayor de conjuntos de datos e incorporar otras métricas diferentes al desempeño del clasificador. Esto constituye actividades futuras de investigación basadas en el presente trabajo.

En este capítulo, se ha determinado la combinación selector - clasificador que formará parte del método de estructuración de poblaciones distribuidas mediante técnicas de RP, sin embargo aún hace falta determinar a partir de un ranking, una estrategia para establecer el total de variables seleccionadas. Para abordar esta tarea, debe desarrollarse un criterio que permita detener el proceso de evaluación de rankings, buscando encontrar el máximo desempeño de algún clasificador, que se incluya para validar la pertinencia de las variables seleccionadas. En el capítulo 6 se aborda a mayor detalle, el desarrollo del criterio requerido.

Capítulo 6

Criterio para evaluación de rankings

La actividad más importante que un ser humano puede lograr es aprender para entender, porque entender es ser libre

Baruch Spinoza

Como resultado de su aplicación, los métodos Filter univariados de SV producen un ranking de las variables presentes en diversos conjuntos de datos. El orden obtenido depende en gran medida, del sustento científico con el que cada algoritmo está diseñado, por lo tanto, algunas técnicas podrían atribuirle más importancia a una variable que a otra.

Posterior a la aplicación de los algoritmos selectores, es frecuente que un experto en el contexto de los datos, decida cuáles de ellas son susceptibles de eliminarse; entre otros fines, para lograr la reducción de la dimensionalidad en la representación procesada. Sin embargo, elegir el subconjunto de variables óptimo a partir de un ranking y que represente adecuadamente a los objetos en observación, no siempre puede efectuarse con la intervención humana.

Dada la cantidad abrumadora de combinaciones posibles, que se pueden conformar con las variables de un conjunto de datos en específico, es justificable recurrir a la automatización de estrategias para el proceso de evaluación de distintos subconjuntos de éstas, a fin de encontrar el óptimo o más adecuado, de acuerdo a alguna clasificación o bien para alcanzar un objetivo en particular.

Derivado de lo anterior, pueden surgir diversas ideas para conformar el subconjunto de variables deseado. Una de ellas, indicaría que a partir de un ranking dado, es posible establecer un umbral, así aquellas variables que logren satisfacerlo, serán seleccionadas y el resto se rechazarán. En otra estrategia, se podría elegir una cantidad previamente establecida de ellas sin considerar más aspectos, como por ejemplo, la correlación entre sí mismas o incluso con la variable objetivo, entre otros.

En este capítulo, se incluye (*i*) la revisión de algunos trabajos en los que se proponen distintas maneras para elegir un subconjunto de variables a partir de un ranking, también, (*ii*) se describe la propuesta de un criterio para elegir un subconjunto de variables a partir de un ranking, buscando que dicha selección sea la que produce el máximo desempeño de un clasificador supervisado. Por otra parte y para analizar su desempeño, (*iii*) este criterio es sometido a un proceso de validación mediante la comparación con resultados obtenidos con otros algoritmos, éstos han sido reportados en la literatura correspondiente, como métodos de búsqueda efectivos. Finalmente, (*iv*) se realiza una discusión del criterio presentado.

Con relación a los trabajos con distintas propuestas en la elección de un subconjunto de variables, están por un lado (*a*) los métodos que requieren la intervención humana para que el investigador indique cuantas de ellas se desean y por otro, (*b*) los que a partir de la satisfacción de un criterio dado, catalogan cada variable como parte del subconjunto buscado, permitiendo automatizar el proceso. Sin embargo, no se ha encontrado algún aporte que se enfoque específicamente en el análisis de las curvas de desempeño de los clasificadores. Principalmente, los trabajos se concentran en satisfacer un umbral que es proporcionado por quienes conocen las naturaleza de los datos que se procesan, o en su defecto, dicho umbral puede ser determinado matemáticamente, lo que requiere cálculos complejos que dificultan el procesamiento, especialmente cuando la dimensionalidad del objeto de estudio es alta.

6.1. Estado del arte

Los siguientes trabajos incluyen diferentes enfoques de SV y distintas técnicas supervisadas de clasificación. Adicionalmente, se presentan algunas variantes en las métricas de distancia utilizadas así como la posible interacción entre variables. La descripción de los trabajos incluye a sus objetivos, algoritmos y conjuntos de datos empleados así como los resultados obtenidos y una reflexión de los mismos.

6.1.1. Un nuevo método supervisado de selección de variables para clasificación de patrones

Este trabajo se centra en la presentación de un nuevo algoritmo para la tarea de SV, éste ha sido desarrollado con una técnica soportada por criterios tomados de la Teoría de la Información (Shannon, 1948). El método presentado a diferencia de otros derivados del mismo enfoque, no se basa simplemente en la obtención de la MI o de la SU entre variables para determinar su relevancia o redundancia, sino que propone una función de evaluación que incluye dos criterios para medir la importancia de cada variable. En esta nueva propuesta, se asocia la relevancia con la distancia intercluster y la redundancia con la distancia intracluster, obtenidas mediante la MI y el *Coeficiente de Relevancia* (CR) respectivamente; este último, propuesto por los autores y que permite seleccionar a aquella variable que tiene la mayor relación incremental de información (Liu et al., 2014).

El algoritmo presentado, utiliza Sequential Forward Search (SFS) como método de búsqueda para evaluar las variables provistas por seis diferentes rankings, estos son Correlation Based Feature Selection (CFS) (Hall, 1999), Fast Correlation Based Feature Selection (FCBF) (Yu y Liu, 2004), modified Mutual Information Based Feature Selection (mMIFS-U) (Novovicova et al., 2007), Relief (Kira y Rendell, 1992), mRMR-U (Ponsa y Lopez, 2007) y el Feature Selection inFormation Criteria (FSFC) propuesto por los autores. En todos los casos se utilizaron tres clasificadores supervisados, el primero basado en el algoritmo clásico de Naive Bayes (Titterington et al., 1981), el segundo en 1-Nearest Neighbor (1NN) (Fix y Hodges, 1951) y uno más basado en C4.5 (Quinlan, 1993).

Se procesaron 12 diferentes conjuntos de datos, 10 de ellos fueron descargados del repositorio UCI Machine Learning (Lichman, 2009) con un promedio de 81 variables y 2,737 instancias, mientras que los restantes dos se extrajeron de ejemplos biomédicos, donde el primero tiene 2,000 variables y únicamente 62 instancias y el segundo consiste de 12,600 variables y 102 instancias.

La experimentación consistió en aplicar primero el método FCBF a todos los conjuntos de datos, debido a que por su naturaleza obtiene por sí mismo un número k de variables seleccionadas, luego, ese valor k es pasado como parámetro al resto de selectores utilizados, a fin de obtener subconjuntos de variables del mismo tamaño y con la intención de fomentar un posterior proceso imparcial para la evaluación de la precisión obtenida por los clasificadores.

En los resultados experimentales, se observa que el método FSFC de los autores

en combinación con el clasificador Naive Bayes, obtiene los más altos valores de precisión para 10 de los 12 conjuntos de datos utilizados; para el caso de la primera excepción donde FSFC no produce el máximo desempeño de clasificación, se trata de la base de datos con más instancias (8,124) y menos variables (22), donde el mayor número de instancias correctamente clasificadas se obtiene con el selector Relief, mientras que en el segundo caso, el conjunto de ejemplos tiene apenas 76 instancias y 44 variables y se presenta una situación de empate con tres selectores que coinciden en la obtención del mejor comportamiento de clasificación, éstos son: CFS, FCBF y mRMR-U.

Este trabajo, aunque presenta resultados que demuestran la efectividad de la propuesta, también sugiere que no hay relación entre el número de instancias y variables con el desempeño de selectoresclasificadores; Finalmente, es posible apreciar que en términos generales, el método propuesto produce mejor desempeño de clasificación no solo con Naive Bayes sino también con 1NN y con C4.5. Respecto del criterio de detención en el proceso de evaluación, se observa que se utiliza un algoritmo reportado previamente en la literatura (FCFB) para determinar el número de variables que serán seleccionadas, posteriormente se usa FSFC para obtener el ranking y finalmente, éste se evalúa completo mediante SFS con los clasificadores.

6.1.2. Un nuevo método de selección de variables que considera la interacción de variables

En este trabajo, se propone un algoritmo de SV orientado a determinar un factor de peso en la interacción entre características de una Base de Datos denominado IWFS, el cual está basado en el cálculo de métricas derivadas de la Teoría de la Información como la MI entre pares de ellas, considerando como variables interactivas a aquellas que se muestran irrelevantes o débilmente relevantes individualmente, pero cuando se combinan entre sí, pueden tener una alta correlación con respecto a la clase (Zeng et al., 2015).

Los autores comparan el método propuesto con los conocidos Correlation Based Feature Selection CFS (Hall, 1999), INTERACT (Zhao y Liu, 2007a), FCBF (Dash y Liu, 2003), mínima Redundancia Máxima Relevancia mRMR (Peng et al., 2005) y Relief-F (Kononenko, 1994). Durante el proceso, se consideró el uso de tres clasificadores que incluyen a los algoritmos C4.5 (Quinlan, 1993), IB1 (Fix y Hodges, 1951) y uno basado en reglas denominado PART (Frank y Witten, 1998).

Los algoritmos mencionados se utilizaron con dos tipos de conjuntos de datos, a saber, (i) con seis bases de datos artificiales, dónde tres de ellas se obtuvieron con

la herramienta de generación de datos RDG1 de Weka (Hall et al., 2009), tres más fueron descargadas del repositorio UCI (Lichman, 2009) y (ii) ocho casos referidos a ejemplos del mundo real, también tomados del repositorio UCI (Lichman, 2009). Para el primer tipo de datos, es decir, los artificiales, el ranking de variables considera como criterio de detención, el momento en que la MI de una variable i con respecto a la clase era igual a la MI de otra variable j, mientras que para los conjuntos de datos reales se utilizó un umbral determinado por las primeras k variables del ranking que producían la más alta precisión de un clasificador.

La diferencia en los criterios utilizados para la evaluación de variables tanto en la experimentación con datos sintéticos como en los reales, es que con los primeros, el objetivo principal era identificar características irrelevantes por lo que satisfacer la igualdad de MI respecto de la clase se consideró suficiente, mientras que con los segundos, se utilizó la precisión de los clasificadores mencionados.

De acuerdo a los resultados presentados por los autores, en la parte de la experimentación con datos sintéticos, su método IWFS remueve todas las variables irrelevantes para los seis conjuntos de datos evaluados. Encontraron también, que el algoritmo INTERACT falla con uno de los ejemplos generados y que CFS falla en todos los casos, mientras que el resto de selectores seleccionan variables correctamente en algunos casos pero fallan en otros, de manera específica mRMR y Relief-F eligen variables irrelevantes en sus resultados.

Para los experimentos con datos reales, los ocho conjuntos de datos fueron particionados en 70 % como datos de entrenamiento y 30 % como datos de prueba y la precisión del clasificador se determina con el promedio de 100 pruebas con diferentes inicializaciones aleatorias, donde IWFS obtiene el mejor desempeño comportándose aún mejor con el clasificador IB1.

Estos resultados favorecen el método propuesto por los autores, se observa que elimina variables redundantes y también identifica a aquellas que se pueden catalogar como interactivas. Sin embargo, es en el promedio de precisión de clasificación donde resulta más alto, lo que no garantiza que en todos los conjuntos de datos evaluados es el mejor, no hay algún comentario de los autores sobre los casos específicos donde otros métodos tienen mejor desempeño. Es factible considerar que, algunas características de dichos ejemplos como el tipo de variables incluidas o el número de elementos nulos, pueden influir en el comportamiento del método, ésto abre la posibilidad de profundizar en el estudio de esta propuesta con grupos de datos de naturalezas distintas.

6.1.3. Selección de variables basada en calidad de información

En este trabajo de investigación, se presenta un método de SV de tipo Filter denominado Quality Information Feature Selection (QIFS), centrado en la capacidad que tienen las variables de un objeto en estudio para describir a un objeto así como su pertenencia a una clase, lo cual está soportado por el cálculo del valor máximo de la vecindad más cercana entre objetos, mediante una fórmula construida por los autores, basada en conceptos derivados de la Teoría de la Información, misma que permite evaluar la calidad de las variables tanto de tipo categórico como numérico, así como, a conjuntos de datos con pocos atributos. El cálculo de la máxima vecindad de un ejemplo del conjunto de datos se obtiene a partir del establecimiento de la *Entropía* y la *Distancia* existente entre dicho ejemplo con respecto a otros; ésto determina la calidad de cada variable para identificar la pertenencia de diferentes ejemplos a una clase (Liu et al., 2017).

El nuevo método es comparado con cuatro algoritmos reportados previamente en la literatura, estos son los conocidos Neighborhood Rough Set (NRS) (Hu et al., 2008), Relief (Kira y Rendell, 1992), Spectral Feature Selection (SPEC) Zhao y Liu (2007b) y Similarity Preserving (SPSF-LAR) (Zhao et al., 2013a). En el proceso se incluyó el uso de los clasificadores CART (Breiman et al., 1984), Linear Support Vector Machine (LSVM) (Fan et al., 2008) y KNN (Fix y Hodges, 1951).

La experimentación se desarrolló desde dos perspectivas, dónde, (*i*) el primer tipo de pruebas se refiere a la utilización de seis conjuntos de datos de tamaño pequeño, extraídos del repositorio UCI (Lichman, 2009) con un promedio de 21 variables y 450 instancias y aplicándose los cuatro selectores mencionados, mientras que, (*ii*) en el segundo grupo de experimentos, se realizó un escalamiento con cuatro bases de datos de mayor tamaño, éste incluye a cuatro conjuntos de ejemplos con una media de 873 variables y 2,043 instancias, aplicándose solamente los selectores QIFS y NRS.

Para la fase de evaluación de los subconjuntos de variables, en el caso del clasificador CART se utilizó el *índice de Gini* (Gini, 1912) como criterio de parada, mientras que con LSVM se usó una función lineal y en el caso de KNN, se estableció k=5; en todos los casos se configuró una validación cruzada con un valor de 10 iteraciones.

Los resultados presentados evidencian una principal comparativa entre los métodos QIFS y NRS, dado que ambos proponen un subconjunto de variables de manera automática mientras que, el resto solamente producen un ranking sin se-

leccionar un número de ellas. Sin embargo, se analizan todos los rankings durante la fase evaluación de la precisión del clasificador.

En el caso de la experimentación realizada con bases de datos más pequeñas, los valores de precisión de los tres clasificadores muestran que globalmente el promedio de desempeño es mayor con QIFS; sin embargo también se aprecia de manera específica que con el clasificador CART solamente en cuatro de los seis conjuntos de datos QIFS es mejor, mientras que esto sucede en cinco de ellos con LSVM y con KNN.

Para las pruebas con conjuntos de datos más grandes, se destacan dos observaciones importantes, (*i*) la primera es que con bases de datos de pocas variables y miles de instancias NRS selecciona menos variables que QIFS pero para cuando existen miles de variables y pocas instancias sucede lo contrario, (*ii*) la segunda sugiere que el selector QIFS produce el mejor desempeño de clasificación para las cuatro bases de datos utilizando el clasificador KNN, mientras que con CART, sólo es mejor en tres de ellas y con LSVM solamente en dos, por lo que aunque QIFS en términos generales se comporte bien en conjuntos de datos grandes, no hay un criterio concluyente para indicar que es mejor en la generalidad.

Al interior del algoritmo de los autores que también selecciona un subconjunto de variables, el criterio consiste en satisfacer un umbral de calidad mediante la obtención de la entropía condicional de la máxima vecindad más cercana entre un subconjunto de variables dado el valor de una clase. Por lo que se observa que, en el trabajo no existe un método que durante el proceso de evaluación de los rankings derivados de los algoritmos de SV, determine cuantas variables habrán de seleccionarse, esta acción se realiza de manera previa al proceso de búsqueda secuencial hacia adelante utilizado, al llegar la experimentación a este punto, ya se sabe cuántas variables se evaluarán dado que se usó NRS para ello.

Como trabajo futuro, una línea de investigación puede darse en profundizar la evaluación de otros conjuntos de datos que incluyan un mayor número de atributos, para observar su comportamiento a fin de generalizar los resultados.

6.1.4. Selección de variables con distancia efectiva

El trabajo reporta la creación de tres nuevos algoritmos para SV de tipo Filter dirigidos a ámbitos no supervisados, utilizando el concepto de *Distancia efectiva* basado en representación escasa, el primero de ellos se denomina Effective

Distance-based Laplacian Score (EDLS), diseñado a partir del concepto de *Score Laplaciano* (He et al., 2005), mientras que, el segundo y tercero nombrados Effective Distance-based Sparsity Score 1 y 2 respectivamente (EDSS-1 y EDSS-2), creados sobre la *Ponderación de escasez* (Liu y Zhang, 2014), en todos los casos para medir la semejanza entre objetos incluidos en conjuntos de datos (Liu y Zhang, 2016).

La idea principal de la distancia efectiva, es que la estructura dinámica de los datos está dominada por un conjunto de rutas o caminos más probables entre ejemplos de un conjunto de datos y que se derivan de una matriz de conectividad unidireccional asimétrica P, donde la distancia de una instancia a hasta una instancia b no es igual de b hacía a.

El algoritmo EDLS propuesto modifica la manera tradicional de calcular el Score Laplaciano, considerando que la distancia efectiva equivale a la similaridad entre dos instancias dadas de un conjunto de datos y ésta debe ser minimizada, mientras que, con la misma intención EDSS-1 y EDSS-2 buscan reducir la distancia efectiva de una i-ésima variable en un conjunto de instancias dado, con la idea de que las variables que pueden respetar mejor la estructura dinámica de los datos, se consideran más discriminatorias.

Los métodos propuestos son comparados con otros cuatro algoritmos reportados en la literatura y diseñados a partir del concepto clásico de Distancia Euclidiana, conocidos como Ponderación de la Varianza (VS) (Webb, 2002), Laplacian Score (LS) (He et al., 2005), Sparsity Score 1 y 2 respectivamente (SS-1 y SS-2) (Liu y Zhang, 2014). En el uso de clasificadores, se reporta durante el proceso la inclusión de un algoritmo basado en K-medias, dónde, el desempeño se mide a través de la métrica *Score de Fisher* (F-Score).

Se utilizaron 12 conjuntos de datos, de los que 10 de ellos fueron descargados del repositorio UCI (Lichman, 2009) con un promedio de 31 variables y 313 instancias; mientras que los otros dos cuyo ámbito está referido a reconocimiento facial, uno se obtuvo de un repositorio de la Universidad de Yale (Yale, 1997) y otro fue tomado de un trabajo sobre parametrización de modelos estocásticos para identificación de rostro (Samaria y Harter, 1994) con una media para ambos de 1,024 variables y 282 instancias.

Aunque los tres algoritmos introducidos en este trabajo se diseñaron para realizar SV en tareas de clustering no supervisado, los autores dividieron su experi-

mentación en dos grupos: supervisado y no supervisado.

En el caso de la experimentación en el contexto no supervisado, se utilizaron los 10 conjuntos de datos del repositorio UCI y las pruebas se realizaron con el algoritmo basado en K-medias, midiendo su desempeño con F-Score y de manera específica, el criterio de detención en el proceso de evaluación consistió en seleccionar a las primeras n variables de los rankings derivados donde n es un valor controlado durante el proceso por los autores, lo que se repitió 10 veces mostrándose al final los resultados de clasificación más altos, así como el número de variables seleccionadas con los que se obtuvieron éstos, señalando que el total de clústeres concuerda con el verdadero número de clases de los conjuntos originales de datos en observación.

Los resultados muestran que el método EDSS-2 tiene mejor desempeño en siete conjuntos de datos, mientras que EDLS superó al resto con otras dos bases de datos y SS-2 solamente con una de ellas. Al final se incluye una fila de promedios por selector donde se observa que EDSS-2 tiene una clara ventaja en lo general.

Adicionalmente, se procesaron también los conjuntos de datos referidos a reconocimiento facial, encontrando que los tres algoritmos de los autores superan al resto de métodos y donde se aprecia que EDSS-2 es también el mejor.

En el contexto supervisado, se evaluaron todos los métodos en tareas de clasificación, tomando como punto de partida los resultados obtenidos en el ámbito no supervisado, aunque sólo se procedió a la graficación de la precisión de clasificación obtenida versus el número de variables seleccionadas para los tres algoritmos propuestos por los autores para dos de los conjuntos de datos; dichas gráficas incluyen una comparativa por tipo de método, asociando a LS con EDLS, SS-1 con EDSS-1 y SS-2 con EDSS-2, encontrando que en cada caso, los métodos basados en la distancia efectiva obtienen precisiones de clasificación más altas.

Cabe mencionar, que también se presentan gráficas comparativas entre los tres algoritmos nuevos para los dos conjuntos de datos seleccionados, observándose que EDSS-2 produce más altos desempeños de clasificación de EDLS y EDSS-1.

Este trabajo demuestra que el uso del concepto de Distancia Efectiva en sustitución de la Distancia Euclidiana al interior de diversos algoritmos existentes en la literatura, puede generar rankings de variables que producen mejores desempeños de clasificación. Sin embargo, aún se requiere la intervención del investigador para

interactuar con los métodos variando el número de variables seleccionadas, no se observa alguna nueva propuesta para realizar esta actividad de forma automática.

6.1.5. Un nuevo algoritmo bacteriano con control de aletoriedad para selección de variables

En este documento, se propone un algoritmo bacteriano perteneciente a un enfoque basado en poblaciones, adaptado a tareas de SV del tipo supervisado (BAFS), este método utiliza tres parámetros de control de aleatoriedad de la población a fin de seleccionar el subconjunto más pequeño de variables con dimensionalidad dinámica; estos parámetros son reconocidos como P_{te} que se utiliza para limitar la búsqueda excesiva del subconjunto óptimo, mientras que, P_{re} y P_{el} sirven para controlar la frecuencia de reproducción y las estrategias de eliminacióndispersión respectivamente (Wang y Niu, 2017).

Este nuevo método, busca reducir la complejidad computacional evitando la búsqueda redundante del subconjunto óptimo de variables, partiendo de la idea de que es inviable medir todos los posibles subconjuntos de ellas, especialmente con problemas de alta dimensionalidad. La población es representada mediante una matriz de vectores seleccionados aleatoriamente, cada uno de tamaño n, donde n es un número de variables elegidas. Un proceso de *Quimiotaxis y Caída* se usa para mejorar las capacidades de búsqueda de las bacterias, éstos son empleados alternadamente, también se incluye una estrategia de pesado aleatorio denominada *Ruleta de Ponderación*, que sirve para discriminar a las variables y eliminarlas del vector.

El proceso de reproducción permite reemplazar a las bacterias más pobres y mejorar la calidad de la población parcial en lugar de la mejor global, mientras que, la eliminacióndispersión sirve para enriquecer a la población y así fortalecer la capacidad de optimización global.

El algoritmo presentado, es comparado con otros cuatro métodos bacterianos reportados en la literatura, a saber son los conocidos como Bacterial Foraging Optimization (BFO) (Passino, 2002), BFO con paso lineal de quimiotaxis (BFOLDC) (Niu et al., 2011), BFO sin paso lineal de quimiotaxis (BFONDC) (Niu et al., 2011) y Bacterial Colony Optimization (BCO) (Wang et al., 2014). En todos los casos, se utilizó un clasificador basado en KNN con k=5, la métrica indicada como criterio de evaluación fue la tasa de error de clasificación.

Los selectores se aplicaron a 10 conjuntos de datos con un promedio de 8067

variables y 109 instancias, mismos que fueron tomados de dos repositorios especializados en dimensionalidad variable, que se encuentran en los sitios de la Universidad de Jinan, China (Jinan, 2002) y en el laboratorio de Sistemas de Descubrimiento de la Universidad Vanderbilt, USA (Statnikov et al., 2003).

Durante el proceso, se determinó que el tamaño de población es 50 y que el tiempo de iteración para BCO y BAFS fue de 200, para los otros tres algoritmos se establecieron los parámetros *Iteraciones de Quimiotaxis*=50, *Iteraciones de reproducción*=5 e *Iteraciones de eliminación*=2 y el máximo número de iteraciones para BFO, BFOLDC y BFONDC corresponde a 500, mientras que, los tres parámetros de control se inicializaron con P_{te}=40, P_{re}=25 y P_{el}=20, ésto para todos los conjuntos de datos. Los autores también controlaron el número de variables seleccionadas, estableciendo que no se podía exceder de 50 y que específicamente en uno de los conjuntos de datos no podría ser mayor a 10, lo anterior debido a la alta dimensionalidad de los ejemplos en estudio.

Los resultados presentados describen el comportamiento del clasificador KNN con cinco subconjuntos de variables seleccionadas, es decir, con 10, 20, 30, 40 y 50 de ellas a excepción del ejemplo en el que se seleccionaron 10, dónde, las mediciones se hicieron con 2, 4, 6, 8 y 10 atributos. En todas las bases de datos, se obtuvieron mejores desempeños de clasificación con su método BAFS, aunque en siete casos, BCO iguala el rendimiento pero con más variables.

Debido a lo anterior, los autores reportan que el método BAFS propuesto es mejor que el algoritmo BCO y que una característica relevante, es la mejora en el tiempo de procesamiento requerido, a diferencia del resto de algoritmos comparados.

Una desventaja encontrada a partir de la descripción del experimento, es que se observa que el algoritmo BAFS de los autores, requiere que se le indique el número de variables que deben seleccionarse, esta situación no es deseable en muchas aplicaciones, ya que se busca idealmente que los algoritmos realicen de forma automática dicha tarea.

6.2. Descripción del criterio para la evaluación de rankings

En esta sección, se propone una manera novedosa y sencilla de obtener un subconjunto de variables a partir de un ranking creado por un algoritmo Filter univariado de SV. La estrategia básica, consiste en analizar el comportamiento de diferentes subconjuntos de variables con un clasificador supervisado, para identificar su mejor desempeño

La idea para desarrollar el nuevo criterio, se obtiene a partir de la observación del comportamiento de seis algoritmos Filter univariados de SV y que fueron evaluados con seis clasificadores supervisados, aplicados a 34 conjuntos de datos con diversas características para validar su eficacia y precisión.

Con la utilización del criterio propuesto, se podrá establecer el tamaño de un subconjunto de variables seleccionadas y así prescindir del resto de ellas, buscando en todo momento obtener el mejor desempeño de clasificación.

El diseño se centra en determinar el momento en el que puede detenerse un proceso de evaluación de rankings, a través de algún clasificador supervisado y de acuerdo a algún método de búsqueda incorporado. El algoritmo está basado en la observación del comportamiento que presentan las curvas de desempeño de clasificación, obtenidas de los experimentos previos en donde se realizaron búsquedas exhaustivas.

En general, en las búsquedas exhaustivas es posible encontrar el máximo desempeño del clasificador, dado que se prueban todos los subconjuntos de variables que se pueden obtener de un ranking. Sin embargo, un problema serio, es el crecimiento del coste computacional a medida que la dimensionalidad aumenta, este incremento vuelve la búsqueda exhaustiva como algo improbable, bien sea por la demanda de tiempo o de los recursos de hardware requeridos.

En el caso del nuevo criterio, éste se ha construido con la idea de incorporar algún método de búsqueda secuencial, para aprovechar las bondades que de ellos se pueden aprovechar, específicamente, pueden ser los métodos de búsqueda SFS o SBE. En este punto de la investigación, se ha decidido explorar primero SFS como método de búsqueda.

Se tienen dos razones principales por las que se considera útil incorporar SFS

en el desarrollo del criterio. La primera es debido a que la evaluación comienza con una sola variable, la más significativa; esto permite que la primer prueba sea muy ágil y abre la posibilidad de obtener un buen desempeño, quizás el máximo que se busca o cercano a éste. La segunda razón es debido a que las primeras variables evaluadas son las más significativas y se asume que sean las que más inciden en el desempeño del clasificador. Si el máximo desempeño se logra identificar en las primeras iteraciones, entonces la experimentación completa habrá requerido menos recursos computacionales. Se espera que las últimas pruebas no tengan grandes variaciones en las medidas obtenidas, ya que se evalúan a las variables menos informacionales.

Si por el contrario, el proceso utilizara al método SBE, las primeras iteraciones incluirían a todas las variables, asumiendo que las pruebas correspondientes son más lentas. La primera variable que se elimina es la menos significativa, por lo que hay pocas posibilidades de incidir mucho en el comportamiento del clasificador y la prueba de clasificación sigue siendo lenta. Este comportamiento justifica el uso de SFS en vez de SBE. Adicionalmente, en ninguno de los dos casos hay algún indicio que sugiera en qué momento se encontrará el máximo global que se desea.

El nuevo criterio es obtenido entonces, a través de un método inductivo, derivado de la generalización del comportamiento observado en la realización de búsquedas exhaustivas a múltiples conjuntos de datos, para los que se ha seguido el ranking producido por la aplicación de diversos métodos Filter de SV.

En las búsquedas exhaustivas en general, se observan curvas de comportamiento que presentan crestas que representan un valor máximo en diferentes momentos de las pruebas, sin embargo algo que resalta y que da origen al diseño del criterio propuesto, es que en la generalidad, se encuentra solamente un valor máximo global, sin identificar un patrón específico relacionado al número de variables presentes en el conjunto de datos. Dicho de otra forma, no hay manera de saber a priori si el mejor desempeño se obtiene con pocas variables, muchas o todas.

Por otra parte, se ha observado que la disminución o aumento en el desempeño del clasificador no es espontáneo, siempre hay una tendencia negativa o positiva, identificándose cierta continuidad en los valores. Se advierte entonces, que durante un número de pruebas consecutivas puede determinarse si la curva va en aumento, o bien, en un descenso definitivo.

Lo anterior ha generado la idea de contabilizar, el número de veces en las que

se mantiene una tendencia negativa en el desempeño del clasificador, creando así la idea de un parámetro que servirá de factor decisivo para saber cuándo detener el proceso de evaluación de subconjuntos de variables. En el criterio propuesto, a este parámetro se le denomina como *ventana*.

El parámetro *ventana* se usará entonces, para detener el proceso de búsqueda exhaustivo al identificar una tendencia negativa en el desempeño del clasificador utilizado. Durante la experimentación se habrá de ajustar y definir el valor adecuado para este parámetro, buscando que éste funcione adecuadamente con diversos conjuntos de datos. No es posible establecer un valor fijo para *ventana*, dado que se pretende su aplicación a cualquier conjunto de datos y el total de variables incluidas cambia de un ejemplo a otro.

Se advierte que dicho valor debe ser calculado con relación al total de variables evaluadas, éste valor debe ser entero y en caso de obtenerse un valor decimal se deberá redondear al entero superior siguiente, a fin de tener pruebas completas.

En resumen, el valor del parámetro *ventana* sirve para indicar cuántas iteraciones consecutivas se realizan a partir de que se identifica un valor de máximo desempeño, si durante este total de pruebas no se mejora el desempeño, la evaluación termina; estableciendo el subconjunto óptimo de variables seleccionadas, para ese clasificador.

De manera inicial, debe establecerse un valor para el parámetro *ventana*, por lo que es posible que para un conjunto de datos cualquiera cuya dimensionalidad está definida por n variables, la *ventana* sea del 10 %, 20 %, 30 % ó 50 % de n.

El diagrama presentado en la Figura 6.1 muestra la propuesta del criterio de obtención del mejor subconjunto de variables de un conjunto de datos en particular, consiste de un proceso iterativo, donde en cada repetición se compara y actualiza el valor de máximo desempeño y se cuenta el total de veces en el que ésta medición no mejora, este número está definido por el parámetro *ventana*, si esto ocurre se habrá encontrado el momento de detener la prueba, en caso contrario, se restablece el conteo y el proceso continúa iterando hasta que no hay más variables que evaluar.

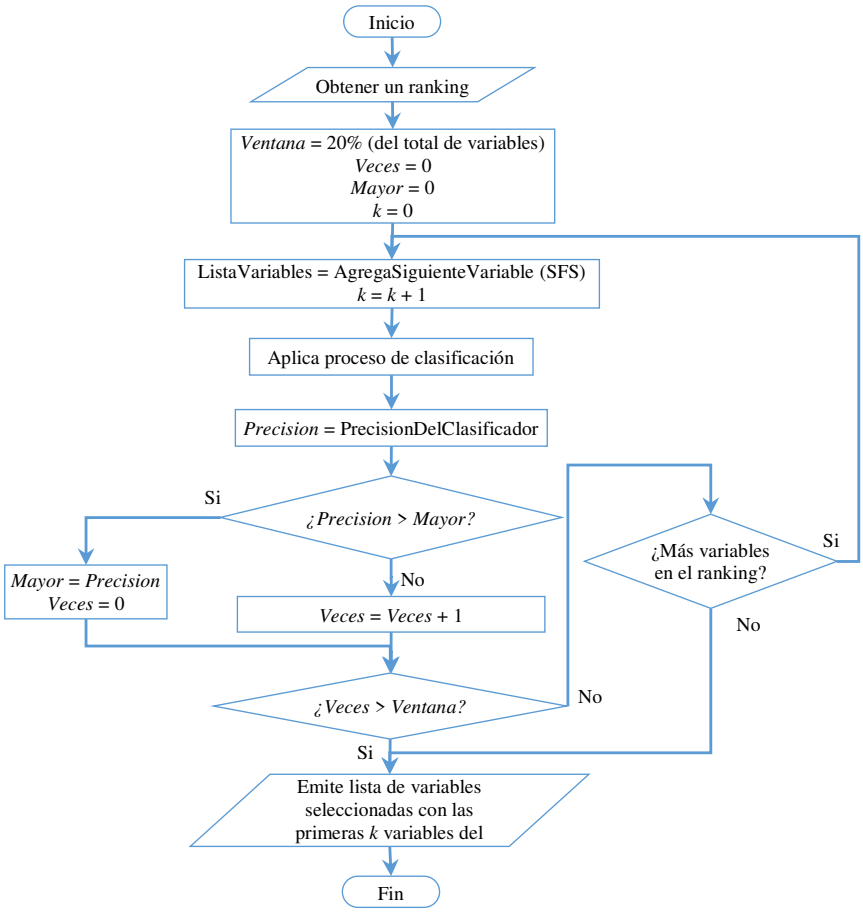

Figura 6.1: Criterio para la evaluación de rankings.

El procedimiento indica el establecimiento de la lista de variables que se envían al clasificador, para lo que en el caso de usar SFS debe agregarse la siguiente variable del ranking a evaluar. Esta etapa en particular, debe aplicarse en sentido contrario al utilizar SBE, refiriéndose a la eliminación de la siguiente variable en el ranking. Esto permite la aplicación del criterio con cualquiera de los métodos de búsqueda mencionados.

En resumen, hay dos maneras de detener la aplicación del nuevo criterio, la primera se da cuando en alguna iteración, se identifica que el total de veces consecutivas donde el desempeño de clasificación no aumenta, es decir, es igual al valor

del parámetro *ventana* y la segunda cuando nunca se cumplió la condición anterior, completándose así una búsqueda exhaustiva.

En la siguiente sección, se presentan los resultados de la aplicación del criterio propuesto, a los 34 conjuntos de datos utilizados en el Capítulo 5 y que se indican en la Tabla 5.1.

6.3. Validación del criterio propuesto

Esta nueva idea de seleccionar el subconjunto de variables a partir de un ranking dado, se somete a un proceso de validación, a fin de determinar su pertinencia al aplicarse a conjuntos de datos de diversas características.

Para abordar el proceso de validación del nuevo criterio, a continuación se presenta la descripción de (*i*) los conjuntos de datos utilizados, (*ii*) los selectores de variables y clasificadores incluidos, (*iii*) el proceso experimental con el criterio de evaluación de rankings, (*iv*) el establecimiento del parámetro *ventana*, (*v*) una comparativa con un método existente en la literatura, que propone la selección de subconjuntos de variables y finalmente, (*vi*) los resultados obtenidos.

6.3.1. Descripción de conjuntos de datos

Todos los conjuntos de datos utilizados para la validación del criterio provienen del Repositorio de Aprendizaje Automático UCI (Lichman, 2013) se han elegido 34 de ellos con diferentes características, a efecto de observar el comportamiento y la estabilidad del método propuesto, al procesar los rankings derivados de diversos selectores.

Los ejemplos incluidos, son de diferente dimensionalidad, hay algunos con clasificación binaria y otros con cardinalidad mayor, balanceados o no, con pocas o muchas instancias y con variables tanto numéricas como nominales. Esto sustentará la pertinencia del criterio propuesto en un amplio abanico de aplicaciones. La Tabla 5.1 del Capítulo 5 presentó la descripción de todos los conjuntos de datos utilizados.

6.3.2. Algoritmos utilizados

Durante la experimentación efectuada en la sección 5.2, se observó que algunos de los selectores no procesaban todos los conjuntos de datos incluidos, por lo

que se decidió que para esta serie de pruebas, se prescindiera de los métodos LS y SVM, conservando a los seis selectores: Chi Squared, Info Gain, Gain Ratio, One R, Relief y Symmetrical Uncertainty.

Los rankings derivados de los selectores mencionados, son evaluados mediante los seis clasificadores supervisados, reconocidos como: SMO, NB, KStar, Ibk, J48Graft y Filtered. Las implementaciones de todos estos algoritmos tanto de SV como de CS y que se utilizan en la experimentación del presente capítulo, se encuentran en la plataforma Weka (Hall et al., 2009).

A semejanza de la comparativa presentada en la sección 5.2, la medida de evaluación que en esta experimentación es utilizada, es la precisión del clasificador.

6.3.3. Experimentación

De acuerdo al procedimiento referido en la Figura 6.1, se han aplicado todos los selectores a cada conjunto de datos, obteniendo así los rankings correspondientes. Para mostrar el procedimiento completo con un ejemplo en particular y por razones de espacio, se presentan los resultados correspondientes al estudio de la base de datos Messidor Features. La Tabla 6.1 muestra los rankings obtenidos con los seis métodos de SV utilizados. En los rankings mostrados, se observa que todos los rankings son diferentes entre sí, por lo que no habrá coincidencias en las pruebas de clasificación que se desarrollarán con ellos, así el criterio de evaluación debería detenerse al identificar diferentes subconjuntos de variables como los óptimos para cada clasificador. A manera de ejemplo, obsérvese que la variable x_3 está indicada por cinco de los seis selectores como la más significativa, la excepción es el algoritmo Gain Ratio, que considera a x_{15} como la más importante en su ranking correspondiente.

Ahora bien, tomando un primer ranking como ejemplo, la Tabla 6.2 muestra los resultados de la búsqueda exhaustiva con los valores de la precisión obtenida por los seis clasificadores. Este caso corresponde al uso del ranking derivado del algoritmo Chi Squared, para el mismo conjunto de datos Messidor, se resalta la precisión más alta obtenida por cada clasificador.

Como puede comprobarse en la Tabla 6.2, no todos los clasificadores presentan porcentajes de precisión similares, tampoco se encuentra el mejor desempeño en la misma iteración, lo que implica que no se obtiene el valor máximo, con el mismo subconjunto de variables evaluado.

Tabla 6.1: Ranking de variables de cada algoritmo de SV aplicado al conjunto de datos *Messidor*

Chi Squared	Gain Ratio	Info Gain	OneR	Relief	Symmetrical Uncertainty
x_3	x_{15}	x_3	x_3	x_3	x_3
x_{15}	x_{16}	x_{16}	x_{15}	x_4	x_7
x_{16}	x_1	x_{15}	x_4	x_{19}	x_4
x_4	x_{14}	x_4	x_5	x_9	x_9
x_{14}	x_3	x_{14}	x_{16}	x_5	x_{15}
x_5	x_{13}	x_5	x_6	x_6	x_{10}
x_6	x_4	x_6	x_{14}	x_{10}	x_8
x_9	x_5	x_9	x_{13}	x_{15}	x_{16}
x_{13}	x_6	x_{13}	x_8	x_7	x_6
x_7	x_7	x_7	x_9	x_8	x_{13}
x_8	x_9	x_8	x_7	x_{11}	x_5
x_2	x_8	x_2	x_1	x_{12}	x_2
x_1	x_{12}	x_1	x_2	x_{16}	x_{14}
x_{19}	x_2	x_{19}	x_{12}	x_{14}	x_{11}
x_{12}	x_{19}	x_{12}	x_{17}	x_2	x_{18}
x_{18}	x_{18}	x_{18}	x_{19}	x_{13}	x_1
x_{17}	x_{17}	x_{17}	x_{11}	x_{18}	x_{12}
x_{11}	x_{11}	x_{11}	x_{10}	x_1	x_{17}
x_{10}	x_{10}	x_{10}	x_{18}	x_{17}	x_{19}

Tabla 6.2: Precisión de clasificadores mediante SFS, ranking derivado de Chi Squared, conjunto de datos Messidor

		Precisión del clasificador (%)					
Prueba	Variable agregada	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
1	x_3	59.253	59.166	56.212	60.990	59.687	59.687
2	x_{15}	60.122	57.255	61.772	61.685	58.384	62.554
3	x_{16}	60.295	56.386	60.904	61.685	58.645	62.815
4	x_4	60.122	57.081	62.120	60.990	65.769	62.989
5	x_{14}	60.122	56.212	62.467	60.990	64.205	62.815
6	x_5	60.904	56.560	63.684	62.815	67.072	62.641
7	x_6	63.510	56.820	64.987	63.336	66.811	62.815
8	x_9	64.118	56.907	65.421	64.553	65.595	63.336
9	x_{13}	64.031	56.386	63.510	64.118	65.421	63.336
10	x_7	64.813	56.125	64.813	64.466	66.203	63.423
11	x_8	65.074	56.299	63.944	63.076	66.030	63.249
12	x_2	67.246	56.386	64.379	64.205	65.421	63.249
13	x_1	67.246	56.386	64.466	64.205	65.508	63.249
14	x_{19}	66.725	56.386	64.639	65.161	67.593	63.249
15	x_{12}	67.159	56.646	63.162	64.987	67.333	63.249
16	x_{18}	66.985	56.646	60.904	64.726	62.207	63.249
17	x_{17}	66.725	56.646	62.467	64.813	59.340	63.249
18	x_{11}	67.333	56.820	61.338	64.292	60.122	63.249
19	x_{10}	67.593	56.820	61.251	64.639	61.338	63.249

A continuación lo que interesa no solo es determinar ¿Cuál es el clasificador con el que se produce mejor comportamiento? o bien ¿Con cuántas variables se logra?, sino verificar que el criterio propuesto, es capaz de determinar los máximos porcentajes de instancias correctamente clasificadas en cada caso. Para responder las preguntas planteadas es fundamental definir el valor del parámetro *ventana*, a continuación se presenta el establecimiento de su valor.

Establecimiento del parámetro *ventana*

Como se ha mencionado, el valor del parámetro *ventana* estará definido en función del total de variables que cualquier conjunto de datos en particular pudiera tener. Por conveniencia de espacio y continuidad de la experimentación desarrollada en este capítulo, se utilizará el ejemplo *Messidor*.

Para establecer el valor más adecuado de este parámetro se consideran varias posibilidades. La estrategia es utilizar valores del 10 %, 20 %, 30 % y 50 % de las variables presentes en el ejemplo de datos *Messidor*, para averiguar si se alcanza o no el máximo desempeño de clasificación, utilizando como base una búsqueda exhaustiva que permita observar el máximo global y comparar con los obtenidos con la *ventana*.

En la Tabla 6.2 se han mostrado los desempeños que corresponden a la búsqueda exhaustiva de dicha base de datos. Estableciendo el parámetro *ventana* = 10 %, el número de pruebas que deben satisfacerse en donde no hay mejoría en el desempeño del clasificador, corresponde a dos iteraciones. Con este valor se observa que el Clasificador SMO no identifica su máximo global, pues en la cuarta iteración ha identificado un máximo local = 60.295 y durante las siguientes dos iteraciones, este desempeño no mejora, deteniéndose el proceso con un desempeño muy por debajo del máximo global que para este caso es del 67.593. En el caso de Naive Bayes si se alcanza el máximo global por coincidir que en la segunda iteración está ubicado el 59.166, pero esto no es una regla. Para Kstar, también se alcanza aunque está ubicado en la novena iteración. En el caso de J48 Graft, Ibk y Filtered tampoco se alcanza aunque los máximos globales están en diferentes iteraciones. En resumen, solo en dos de los seis clasificadores se identificó su máximo global.

Utilizando ahora una *ventana* del 20 %, en dos de los clasificadores no se encuentra el máximo global, pero los máximos locales identificados están por debajo del global por menos del 0.5 %. Este es un dato a tomar en cuenta para decidir el mejor valor de la *ventana*.

Al establecer el parámetro *ventana* con el 30 % se aumenta la identificación de los máximos globales de todos los clasificadores pero el número de iteraciones ya es cercana a la exhaustividad. Sólo con Naive Bayes se evitaron suficientes pruebas, mientras que con KStar, Filtered solo se dejaron de hacer 5 y 3 iteraciones para completar la evaluación exhaustiva.

Con la utilización del 50 %, solo Naive Bayes tiene reducción en el número de iteraciones, en el resto de clasificadores se completa la exhaustividad. Con base en estos resultados se establece el parámetro *ventana* en el 20 %, por mostrar un comportamiento balanceado entre precisión y número de iteraciones.

Aplicación del criterio *ventana*

De acuerdo al algoritmo del nuevo criterio, se requiere determinar el valor que corresponde al parámetro *ventana* para el conjunto de datos Messidor, que contiene 19 variables, dónde el 20 % de ellas equivalente a cuatro por redondeo del valor exacto de 3.8 al entero siguiente. Adicionalmente, se utiliza una variable denominada *mayor*, inicializándose con el valor cero y que como se describió anteriormente, servirá para establecer el máximo desempeño encontrado iteración por iteración.

Observando los resultados presentados en la Tabla 6.2, para el caso del clasificador Naive Bayes, en la primer iteración la precisión muestra un 53.084 % de instancias correctamente clasificadas, que en ese momento corresponde al mejor desempeño, actualizando a la variable *mayor* con ese valor; para la segunda prueba, se obtiene un 59.166 % y como este número es mayor que la variable *mayor*, entonces ésta se actualiza con este último número. En la tercer prueba, sucede que el desempeño obtenido es menor que el valor almacenado como *mayor*, por lo que no se actualiza, a partir de este momento se comienza a contar el número de veces consecutivas donde no se supera el rendimiento de la iteración anterior. Posteriormente, durante la cuarta, quinta y sexta iteración, no se mejora el rendimiento de clasificación, alcanzando con esto el valor de la *ventana* y deteniendo el proceso.

En un proceso exhaustivo, se observa que para el mismo clasificador Naive Bayes, el mejor desempeño obtenido se obtuvo en la segunda iteración por lo que en las 18 pruebas restantes, nunca se supera dicho valor. En lugar de terminar la prueba completa, el criterio propone detener la evaluación del ranking en la sexta iteración, ahorrándose 14 iteraciones. Adicionalmente, por utilizar el método de búsqueda SFS, cada nueva iteración agrega una nueva variable volviendo el procedimiento cada vez más lento, lo que implica que para este caso que se describe, el nuevo criterio ha trabajado con el menor número de variables, requiriendo muy

poco tiempo de procesamiento y menos recursos computacionales.

De acuerdo al criterio, en la Tabla 6.2 se observa que para Naive Bayes se requieren mucho menos iteraciones para identificar el subconjunto de variables con las que se obtiene el mejor desempeño; ésto mismo sucede con el resto de clasificadores, aunque con más pruebas. Sin embargo, hay una excepción con el clasificador SMO, ya que el nuevo criterio requiere completar la búsqueda exhaustiva para encontrar el valor máximo.

Para fortalecer la comprensión del comportamiento del nuevo criterio; la Tabla 6.3 presenta los valores obtenidos para la variable *mayor*, ésta almacena en todo momento el mejor desempeño de clasificación por cada iteración. Una vez que el valor máximo es identificado, este valor ya no se actualiza. Adicionalmente, se han enmarcado las iteraciones que describen el tamaño de la *ventana*, con ésto se observa el momento de detención en la evaluación del ranking, encontrando el mejor desempeño global por cada clasificador.

Tabla 6.3: Máxima precisión de clasificadores por iteración, método de búsqueda SFS, ranking Chi Squared, conjunto de datos Messidor

		Precisión del clasificador (%)					
Prueba	Variable agregada	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
1	x_3	59.253	59.166	56.212	60.990	59.687	59.687
2	x_{15}	60.122	59.166	61.772	61.685	58.384	62.554
3	x_{16}	60.295	59.166	60.904	61.685	58.645	62.815
4	x_4	60.122	59.166	62.120	60.990	65.769	62.989
5	x_{14}	60.122	59.166	62.467	60.990	64.205	62.815
6	x_5	60.904	59.166	63.684	62.815	67.072	62.641
7	x_6	63.510	59.166	64.987	63.336	66.811	62.815
8	x_9	64.118	59.166	65.421	64.553	65.595	63.336
9	x_{13}	64.031	59.166	65.421	64.118	65.421	63.336
10	x_7	64.813	59.166	65.421	64.466	66.203	63.423
11	x_8	65.074	59.166	65.421	63.076	66.030	63.249
12	x_2	67.246	59.166	65.421	64.205	65.421	63.249
13	x_1	67.246	59.166	65.421	64.205	65.508	63.249
14	x_{19}	66.725	59.166	65.421	65.161	67.593	63.249
15	x_{12}	67.159	59.166	65.421	65.161	67.593	63.249
16	x_{18}	66.985	59.166	65.421	65.161	67.593	63.249
17	x_{17}	66.725	59.166	65.421	65.161	67.593	63.249
18	x_{11}	67.333	59.166	65.421	65.161	67.593	63.249
19	x_{10}	67.593	59.166	65.421	65.161	67.593	63.249

De manera análoga, para este conjunto de datos se han obteniendo las tablas que completan el proceso de evaluación de los seis rankings mencionados. De acuerdo a los resultados, el nuevo criterio evita las búsquedas exhaustivas, detiene el proceso de evaluación del ranking e identifica el mejor desempeño de los clasificadores.

La experimentación con un solo conjunto de datos no es suficiente para generalizar la estabilidad del método, por lo que se procedió a aplicar el algoritmo propuesto con el resto de ejemplos. La Tabla 6.4 presenta los rankings derivados de los seis selectores utilizados con el conjunto de datos Dermatology, éste cuenta con 34 variables y seis clases a diferencia de Messidor que incluye solo 19 variables y dos clases.

Tabla 6.4: Ranking de variables de los seis algoritmos de SV, conjunto de datos *Dermatology*

Chi Squared	Gain Ratio	Info Gain	OneR	Relief	Symmetrical Uncertainty
x_{34}	x_{31}	x_{21}	x_{21}	x_{21}	x_{21}
x_{33}	x_{27}	x_{20}	x_{29}	x_{22}	x_{22}
x_{27}	x_{33}	x_{22}	x_{33}	x_{20}	x_{20}
x_{29}	x_{6}	x_{34}	x_{25}	x_{33}	x_{33}
x_{12}	x_{29}	x_{33}	x_{12}	x_{27}	x_{27}
x_{31}	x_{12}	x_{29}	x_{20}	x_{16}	x_{29}
x_{15}	x_{15}	x_{27}	x_{22}	x_{28}	x_{12}
x_{25}	x_{25}	x_{12}	x_{6}	x_{12}	x_{25}
x_{6}	x_{8}	x_{25}	x_{27}	x_{29}	x_{6}
x_{22}	x_{30}	x_{6}	x_{8}	x_{25}	x_{8}
x_{20}	x_{22}	x_{16}	x_{16}	x_{6}	x_{15}
x_{8}	x_{20}	x_{8}	x_{28}	x_{8}	x_{9}
x_{21}	x_{7}	x_{28}	x_{15}	x_{14}	x_{28}
x_{30}	x_{21}	x_{9}	x_{9}	x_{9}	x_{16}
x_{7}	x_{9}	x_{15}	x_{10}	x_{15}	x_{10}
x_{16}	x_{24}	x_{10}	x_{14}	x_{10}	x_{24}
x_{9}	x_{10}	x_{24}	x_{24}	x_{5}	x_{14}
x_{28}	x_{28}	x_{14}	x_{19}	x_{4}	x_{31}
x_{10}	x_{14}	x_{5}	x_{2}	x_{3}	x_{26}
x_{24}	x_{16}	x_{26}	x_{4}	x_{19}	x_{5}
x_{26}	x_{26}	x_{3}	x_{3}	x_{24}	x_{7}
x_{14}	x_{23}	x_{19}	x_{7}	x_{26}	x_{30}
x_{5}	x_{11}	x_{31}	x_{23}	x_{2}	x_{3}
x_{3}	x_{5}	x_{7}	x_{31}	x_{7}	x_{3}
x_{19}	x_{3}	x_{4}	x_{13}	x_{31}	x_{23}
x_{4}	x_{2}	x_{23}	x_{26}	x_{30}	x_{19}
x_{2}	x_{19}	x_{2}	x_{30}	x_{23}	x_{2}
x_{23}	x_{13}	x_{30}	x_{1}	x_{18}	x_{4}
x_{11}	x_{4}	x_{11}	x_{5}	x_{11}	x_{11}
x_{1}	x_{34}	x_{1}	x_{17}	x_{17}	x_{1}
x_{18}	x_{1}	x_{18}	x_{32}	x_{1}	x_{18}
x_{17}	x_{18}	x_{17}	x_{11}	x_{13}	x_{13}
x_{13}	x_{17}	x_{32}	x_{34}	x_{32}	x_{17}
x_{32}	x_{32}	x_{13}	x_{18}	x_{34}	x_{32}

Para este ejemplo, la Tabla 6.5 muestra los resultados de la precisión obtenida con los seis clasificadores durante búsquedas exhaustivas con el ranking derivado del algoritmo Symmetrical Uncertainty en el conjunto de datos Dermatology. A diferencia de Chi Squared que está basado en una apreciación estadística, SU se apoya en conceptos de la Teoría de la Información. Lo anterior permitirá observar el comportamiento del criterio con diferentes algoritmos de SV.

Tabla 6.5: Precisión de clasificadores mediante SFS, ranking Symmetrical Uncertainty, conjunto de datos *Dermatology*

		Precisión del clasificador (%)					
Prueba	Variable agregada	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
1	x_{21}	50.273	50.273	50.273	50.273	50.273	50.273
2	x_{22}	61.749	61.749	61.202	61.202	61.202	61.202
3	x_{20}	62.568	62.568	62.295	62.022	62.295	61.749
4	x_{33}	76.776	77.049	76.776	75.410	76.776	75.410
5	x_{27}	76.503	76.503	76.776	75.137	76.776	75.137
6	x_{29}	76.230	77.049	77.049	75.410	77.049	75.137
7	x_{12}	76.503	76.503	76.776	75.410	76.776	75.137
8	x_{25}	76.503	76.776	76.776	75.410	76.776	75.137
9	x_{6}	76.776	76.776	76.776	75.410	76.776	75.137
10	x_{8}	76.776	76.776	77.049	75.410	76.776	74.863
11	x_{15}	80.601	80.601	79.508	79.508	79.508	77.596
12	x_{9}	84.426	83.880	81.421	81.967	81.421	79.781
13	x_{28}	83.607	86.066	84.426	79.508	82.787	77.869
14	x_{16}	81.421	86.339	84.153	81.694	83.060	80.328
15	x_{10}	84.973	86.885	84.426	82.787	82.240	81.421
16	x_{24}	85.246	86.885	84.426	82.787	82.514	81.148
17	x_{14}	92.896	91.803	90.710	89.617	89.891	87.978
18	x_{31}	93.989	93.169	93.443	92.077	92.077	90.164
19	x_{26}	94.536	93.989	93.716	91.803	91.803	89.891
20	x_{5}	98.907	98.361	98.087	95.902	96.995	94.536
21	x_{7}	98.907	98.634	98.087	95.628	97.814	93.989
22	x_{30}	98.907	98.634	98.087	95.628	97.541	93.989
23	x_{34}	98.087	98.361	98.361	95.355	97.814	93.716
24	x_{3}	98.087	98.361	97.268	95.628	96.995	93.989
25	x_{23}	97.814	98.634	96.721	95.628	96.721	93.989
26	x_{19}	97.541	98.087	96.175	95.628	95.628	93.989
27	x_{2}	98.087	97.814	95.902	95.628	96.448	93.989
28	x_{4}	96.175	97.814	96.175	95.628	96.175	93.989
29	x_{11}	96.175	97.541	95.902	95.628	96.175	93.989
30	x_{1}	95.902	97.268	95.355	95.355	96.175	93.716
31	x_{18}	96.175	97.541	96.721	95.355	95.628	93.716
32	x_{13}	95.902	97.541	96.995	95.355	96.448	93.989
33	x_{17}	95.902	96.721	96.448	95.355	95.902	94.262
34	x_{32}	96.175	97.268	95.902	95.628	96.175	94.536

En la Tabla 6.5, se resalta la precisión más alta obtenida por cada clasificador. Se elige este ejemplo para apreciar un comportamiento que aunque presenta varias crestas durante la experimentación, el criterio es capaz de identificar con éxito el

momento en que la evaluación del ranking debe detenerse, identificando al subconjunto de variables que produce el máximo desempeño. En este ejemplo se utiliza una *ventana* igual a siete.

Para comprender la aplicación del criterio de detención al evaluar los seis rankings en estudio, la Tabla 6.6 muestra los valores de máximo desempeño de todos los clasificadores por cada iteración, se señala la *ventana* y el momento de parada.

Tabla 6.6: Máxima precisión de los clasificadores por iteración, ranking derivado de Symmetrical Uncertainty, conjunto de datos *Dermatology*

		Precisión del clasificador (%)					
Prueba	Variable agregada	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
1	x_{21}	50.273	50.273	50.273	50.273	50.273	50.273
2	x_{22}	61.749	61.749	61.202	61.202	61.202	61.202
3	x_{20}	62.568	62.568	62.295	61.202	62.295	61.749
4	x_{33}	76.776	77.049	76.776	75.410	76.776	75.410
5	x_{27}	76.776	77.049	76.776	75.410	76.776	75.410
6	x_{29}	76.776	77.049	77.049	75.410	77.049	75.410
7	x_{12}	76.776	77.049	77.049	75.410	77.049	75.410
8	x_{25}	76.776	77.049	77.049	75.410	77.049	75.410
9	x_{6}	76.776	77.049	77.049	75.410	77.049	75.410
10	x_{8}	76.776	77.049	77.049	75.410	77.049	75.410
11	x_{15}	80.601	80.601	79.508	79.508	79.508	77.596
12	x_{9}	84.426	83.880	81.421	81.967	81.421	79.781
13	x_{28}	84.426	86.066	84.426	81.967	82.787	79.781
14	x_{16}	84.426	86.339	84.426	81.967	83.060	80.328
15	x_{10}	84.973	86.885	84.426	82.787	83.060	81.421
16	x_{24}	85.246	86.885	84.426	82.787	83.060	81.421
17	x_{14}	92.896	91.803	90.710	89.617	89.891	87.978
18	x_{31}	93.989	93.169	93.443	92.077	92.077	90.164
19	x_{26}	94.536	93.989	93.716	92.077	92.077	90.164
20	x_{5}	98.907	93.989	98.087	95.902	96.995	94.536
21	x_{7}	98.907	98.634	98.087	95.902	97.814	94.536
22	x_{30}	98.907	98.634	98.087	95.902	97.814	94.536
23	x_{34}	98.907	98.634	98.361	95.902	97.814	94.536
24	x_{3}	98.907	98.634	98.361	95.902	97.814	94.536
25	x_{23}	98.907	98.634	98.361	95.902	97.814	94.536
26	x_{19}	98.907	98.634	98.361	95.902	97.814	94.536
27	x_{2}	98.907	98.634	98.361	95.902	97.814	94.536
28	x_{4}	98.907	98.634	98.361	95.902	97.814	94.536
29	x_{11}	98.907	98.634	98.361	95.902	97.814	94.536
30	x_{1}	98.907	98.634	98.361	95.902	97.814	94.536
31	x_{18}	98.907	98.634	98.361	95.902	97.814	94.536
32	x_{13}	98.907	98.634	98.361	95.902	97.814	94.536
33	x_{17}	98.907	98.634	98.361	95.902	97.814	94.536
34	x_{32}	98.907	98.634	98.361	95.902	97.814	94.536

Puede observarse en la Tabla 6.6, que aunque el conjunto de datos Dermatology que es considerablemente más grande que Messidor (casi el doble de variables y el triple tanto de instancias como de clases), el método propuesto evita alrededor del 17 % de pruebas experimentales y además éstas son las de mayor complejidad computacional por incluir cada vez más variables, lo que se traduce en ahorro de tiempo y recursos de hardware para encontrar el mejor subconjunto de variables.

Con el fin de fortalecer la comprensión del procedimiento anterior, la Figura 6.2 presenta la visualización gráfica del comportamiento de los seis clasificadores, cabe mencionar que en este ejemplo aunque existen varias crestas el método logra determinar el máximo desempeño cuando se presenta primero y evitando la culminación de la búsqueda exhaustiva en todos los casos.

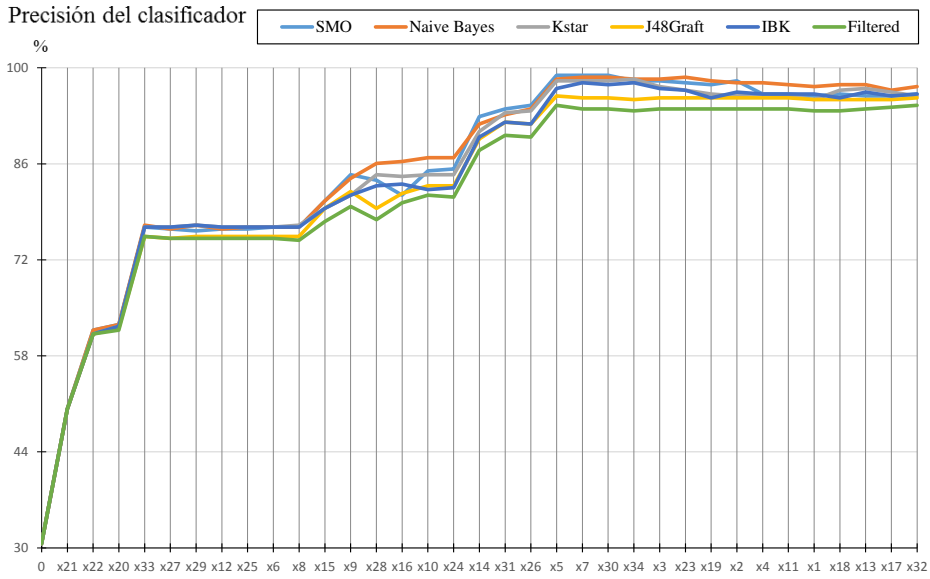

Figura 6.2: Desempeño de todos los clasificadores mediante SFS, ranking Symmetrical Uncertainty, conjunto de datos *Dermatology*.

Debido al número de iteraciones realizadas, la gráfica en la Figura 6.2 es muy densa, por lo que se ha acotado una sección de la misma, para apreciar de mejor manera el comportamiento de los clasificadores incluidos. En la gráfica de la Figura 6.3 se presenta con más claridad, el momento en el que el criterio propuesto identifica el máximo desempeño de clasificación.

Se observa que, para los clasificadores y específicamente en el caso de Ibk, una vez identificado el mejor desempeño, se aprecian otras crestas que indican máximos locales, aunque éstas son posteriores, el criterio no presenta problema en identificar el máximo global.

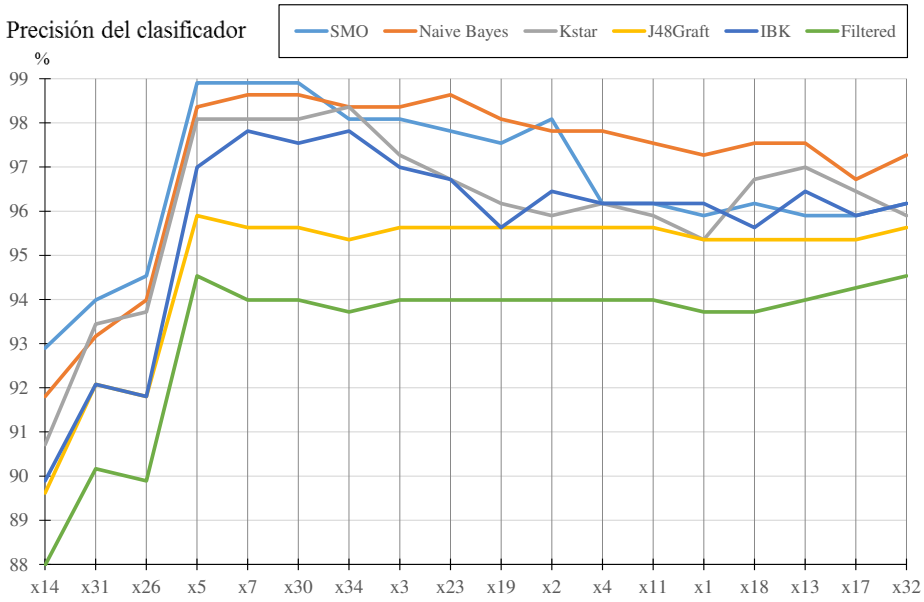

Figura 6.3: Detalle, comportamiento de clasificadores mediante SFS, ranking Symmetrical Uncertainty, conjunto de datos *Dermatology*.

Hasta este momento, los conjuntos de datos reportados incluyen únicamente variables de tipo numérico y el primero de ellos (Messidor) se refiere a un caso de clase binaria y el otro es multiclase (Dermatology). A continuación se presentan resultados para observar la estabilidad del nuevo criterio propuesto con otras variantes, la primera contiene simultáneamente variables tanto numéricas como categóricas y la segunda incluye únicamente variables categóricas; en ambos casos se incluyen datos de clases binarias y multiclase.

Para el estudio de ejemplos con variables tanto numéricas como categóricas, se incluyen las bases de datos Adults y Lymphography, la primera está conformada por 48,842 instancias, seis variables numéricas, ocho categóricas y dos valores de clase mientras que la segunda contiene 148 instancias, tres variables numéri-

cas, 15 categóricas y cuatro clases. La razón de mostrar estos resultados es porque se encontró que algunos conjuntos de datos con ambos tipos de variables, pueden producir curvas de desempeño con crestas demasiado largas que pueden alterar el criterio de detención en el proceso de evaluación del ranking, sin embargo en otros casos el algoritmo funciona y muestra una estabilidad aceptable.

La Tabla 6.7 muestra, los rankings de los seis selectores aplicados al conjunto de datos Adults. Se observa al igual que en ejemplos anteriores, que éstos son diferentes entre sí, aunque existe cierta semejanza entre Chi Squared e Info Gain, lo que no sugiere alguna relación entre estos algoritmos. Cabe mencionar que en este ejemplo, todos los métodos proponen a la misma variable x_3 como la menos significativa, por lo que se asume que tiene menor capacidad informacional que el resto de ellas y la coincidencia en los rankings lo confirma. Esta situación es inherente a los datos, por lo que no necesariamente debe suceder igual siempre.

Tabla 6.7: Ranking de variables de los seis algoritmos de SV, conjunto de datos *Adults*

Chi Squared	Gain Ratio	Info Gain	OneR	Relief	Symmetrical Uncertainty
x_8	x_{11}	x_8	x_{11}	x_7	x_{11}
x_6	x_{12}	x_6	x_{12}	x_8	x_6
x_{11}	x_6	x_{11}	x_4	x_4	x_8
x_4	x_8	x_1	x_5	x_6	x_{12}
x_5	x_{10}	x_4	x_2	x_1	x_5
x_7	x_5	x_5	x_6	x_9	x_4
x_1	x_4	x_7	x_1	x_2	x_1
x_{13}	x_1	x_{13}	x_7	x_5	x_{10}
x_{12}	x_{13}	x_{12}	x_8	x_{13}	x_7
x_{10}	x_7	x_{10}	x_{10}	x_{10}	x_{13}
x_2	x_2	x_2	x_9	x_{11}	x_2
x_9	x_9	x_9	x_{14}	x_{14}	x_9
x_{14}	x_{14}	x_{14}	x_{13}	x_{12}	x_{14}
x_3	x_3	x_3	x_3	x_3	x_3

La Tabla 6.8 incluye los valores de precisión de los clasificadores mediante prueba exhaustiva, utilizando el ranking proporcionado por el selector Info Gain, se resalta el máximo desempeño en cada clasificador. Aplicando el nuevo método propuesto.

Tabla 6.8: Precisión de clasificadores mediante SFS, ranking derivado de Info Gain, conjunto de datos *Adults*

		Precisión del clasificador (%)					
Prueba	Variable agregada	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
1	x_8	76.07	76.07	76.07	76.07	76.07	76.07
2	x_6	76.07	71.63	76.07	76.07	76.07	76.07
3	x_{11}	79.97	79.32	80.91	81.05	81.05	81.04
4	x_1	79.97	79.38	81.14	81.24	80.63	80.95
5	x_4	84.05	80.33	84.83	84.91	82.96	85.11
6	x_5	84.05	82.60	84.98	85.19	82.96	85.20
7	x_7	84.00	83.64	85.36	85.11	81.15	85.61
8	x_{13}	84.01	83.85	85.42	85.40	80.18	85.69
9	x_{12}	84.56	83.02	85.82	86.08	80.64	86.45
10	x_{10}	84.61	83.14	85.77	86.02	80.64	86.44
11	x_2	84.87	83.19	85.82	86.19	80.42	86.62
12	x_9	84.88	83.22	85.69	86.13	80.26	86.77
13	x_{14}	84.92	83.26	85.71	86.14	80.15	86.68
14	x_3	84.96	83.25	79.19	86.11	79.52	86.68

Aunque no es el propósito de este trabajo evaluar el tiempo de procesamiento requerido por cada algoritmo de clasificación, debe comentarse que SMO y KStar se muestran más lentos que el resto de clasificadores.

La Tabla 6.9 muestra los valores de máxima clasificación por iteración, en este caso una vez satisfecho el criterio de detención se ha parado el proceso, comparándose con los resultados de la Tabla 6.8, se observa que el método ha logrado en todos los casos coincidir con el máximo desempeño de clasificación mediante búsqueda exhaustiva, demostrando su validez con un ejemplo de datos con tipos numéricos y categóricos y de ámbito de clases binarias.

Tabla 6.9: Máxima precisión de los clasificadores por iteración, ranking derivado de Info Gain, conjunto de datos *Adults*

		Precisión del clasificador (%)					
Prueba	Variable agregada	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
1	x_8	76.07	76.07	76.07	76.07	76.07	76.07
2	x_6	76.07	76.07	76.07	76.07	76.07	76.07
3	x_{11}	79.97	79.32	80.91	81.05	81.05	81.04
4	x_1	79.97	79.38	81.14	81.24	81.05	81.04
5	x_4	84.05	80.33	84.83	84.91	82.96	85.11
6	x_5	84.05	82.60	84.98	85.19	82.96	85.20
7	x_7	84.05	83.64	85.36	85.19	82.96	85.61
8	x_{13}	84.05	83.85	85.42	85.40	82.96	85.69
9	x_{12}	84.56	83.85	85.82	86.08		86.45
10	x_{10}	84.61	83.85	85.82	86.08		86.45
11	x_2	84.87	83.85	85.82	86.19		86.62
12	x_9	84.88		85.82	86.19		86.77
13	x_{14}	84.92			86.19		86.77
14	x_3	84.96			86.19		86.77

De acuerdo al número de variables que es de 14, el valor redondeado del 20 % para definir el tamaño de la *ventana* es de tres, el número de pruebas que se evitaron fue de 11, éstas de haberse realizado incluirían más variables que las que si se realizaron, este número implica un porcentaje del 14.67 % de pruebas evitadas aunque el impacto en tiempo de procesamiento es mayor, dado que dado que los últimos tests son los más tardados, dado que incluyen más variables.

Cabe mencionar que en la Tabla 6.9, se observa que la aplicación del criterio en la evaluación del ranking Info Gain con el clasificador SMO, no logra determinar el máximo global con el valor de la *ventana* del 20 %, ya que en la iteración 6 se obtiene un valor máximo local equivalente al 84.05 % de precisión, mismo que se mantiene durante las siguientes tres iteraciones, deteniendo el proceso en este punto y fallando por 0.91 % con respecto del máximo general de este clasificador. Sin embargo, aunque es posible ajustar el tamaño de la variable *ventana* y establecerla en 30 %, con lo que se obtendría la mayor precisión; debe evaluarse si el aumento de dicho parámetro se justifica, ya que esto no evita ninguna iteración y se completa la evaluación exhaustiva. Por el contrario aceptar la disminución de la precisión con la diferencia mencionada, disminuye en un 40 % el tiempo de procesamiento

total para este caso en particular.

Los resultados de la aplicación de los métodos de SV al conjunto de datos Lymphography no son tan estables y precisos como con Adult, sin embargo se encuentra una estabilidad aceptable, los rankings para los seis selectores se presentan en la Tabla 6.10, en este caso no hay coincidencia de todos los algoritmos al identificar a la variable menos significativa, aunque para algunos sí; también se advierte una mayor semejanza entre Info Gain y Symmetrical.

Tabla 6.10: Ranking de variables de los seis algoritmos de SV, conjunto de datos *Lymphography*

Chi Squared	Gain Ratio	Info Gain	OneR	Relief	Symmetrical Uncertainty
x_1	x_9	x_{13}	x_{13}	x_{13}	x_{13}
x_{13}	x_7	x_{18}	x_2	x_2	x_{18}
x_{12}	x_{13}	x_{14}	x_{15}	x_{15}	x_9
x_9	x_{18}	x_{15}	x_{18}	x_8	x_7
x_{11}	x_2	x_2	x_{10}	x_{10}	x_2
x_{14}	x_8	x_1	x_8	x_{18}	x_{15}
x_7	x_4	x_{11}	x_{14}	x_7	x_8
x_{18}	x_{10}	x_{12}	x_9	x_5	x_{10}
x_{15}	x_{11}	x_{10}	x_7	x_{11}	x_{11}
x_2	x_{15}	x_9	x_{17}	x_{17}	x_1
x_{10}	x_1	x_8	x_5	x_9	x_{12}
x_8	x_{17}	x_7	x_3	x_{16}	x_{14}
x_4	x_{12}	x_5	x_6	x_{12}	x_5
x_5	x_5	x_{17}	x_1	x_{14}	x_{17}
x_{17}	x_{14}	x_{16}	x_{16}	x_6	x_{16}
x_{16}	x_{16}	x_4	x_4	x_3	x_4
x_3	x_3	x_3	x_{12}	x_1	x_3
x_6	x_6	x_6	x_{11}	x_4	x_6

Para este caso, por razones de espacio y por ser análogo al procedimiento mostrado anteriormente, en la Tabla 6.11 sólo se presentan los resultados de los valores de máximo desempeño en cada iteración del nuevo método propuesto resaltando el máximo valor de la precisión del clasificador, se encontró que no produce resultados satisfactorios para ningún clasificador con el criterio del 20 % para el tamaño de la *ventana*; si este valor se aumenta al 30 % el método es capaz de coincidir en cinco de los seis clasificadores como se aprecia en la Tabla 6.12.

Tabla 6.11: Máxima precisión de los clasificadores por iteración, ranking derivado de One R, conjunto de datos *Lymphography*, *ventana* = 20 %

		Precisión del clasificador (%)					
Prueba	Variable agregada	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
1	x_1	54.73	54.73	54.73	54.73	54.73	54.73
2	x_{13}	77.03	76.35	72.97	75.00	72.97	75.00
3	x_{12}	77.03	76.35	74.32	77.03	73.65	77.03
4	x_9	79.05	77.70	75.68	77.03	76.35	77.03
5	x_{11}	79.73	78.38	75.68	77.03	76.35	77.03
6	x_{14}	79.73	78.38	75.68	77.03	76.35	77.03
7	x_7	79.73	78.38	75.68	77.03	76.35	77.03
8	x_{18}	79.73	79.73	75.68		76.35	
9	x_{15}	83.11	79.73				
10	x_2	83.78	79.73				
11	x_{10}	85.81	79.73				
12	x_8	85.81	79.73				
13	x_4	85.81					
14	x_5	85.81					
15	x_{17}	85.81					
16	x_{16}						
17	x_3						
18	x_6						

Tabla 6.12: Máxima precisión de los clasificadores por iteración, ranking derivado de One R, conjunto de datos *Lymphography*, *ventana* = 30 %

		Precisión del clasificador (%)					
Prueba	Variable agregada	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
1	x_1	54.73	54.73	54.73	54.73	54.73	54.73
2	x_{13}	77.03	76.35	72.97	75.00	72.97	75.00
3	x_{12}	77.03	76.35	74.32	77.03	73.65	77.03
4	x_9	79.05	77.70	75.68	77.03	76.35	77.03
5	x_{11}	79.73	78.38	75.68	77.03	76.35	77.03
6	x_{14}	79.73	78.38	75.68	77.03	76.35	77.03
7	x_7	79.73	78.38	75.68	77.03	76.35	77.03
8	x_{18}	79.73	79.73	75.68	77.03	76.35	77.03
9	x_{15}	83.11	79.73	77.70		76.35	
10	x_2	83.78	79.73	83.11			
11	x_{10}	85.81	79.73	85.14			
12	x_8	85.81	79.73	85.14			
13	x_4	85.81	79.73	85.14			
14	x_5	85.81		85.14			
15	x_{17}	85.81		85.14			
16	x_{16}	87.84		87.84			
17	x_3	87.84		87.84			
18	x_6	87.84		87.84			

Aumentar el tamaño de la *ventana* mejora la efectividad del criterio, pero puede afectar significativamente su desempeño, ya que implica realizar más iteraciones con un mayor número de variables cada vez y mayor costo computacional.

A continuación, se presentan los ejemplos de variables categóricas Chess y Nursery. El primero incluye 3,196 instancias, 36 variables y dos clases y el segundo consta de 12,960 instancias, ocho variables y cinco clases. Las Tablas 6.13 y 6.14 muestran los rankings correspondientes.

Tabla 6.13: Ranking derivado de seis algoritmos de SV, conjunto de datos *Chess*

Chi Squared	Gain Ratio	Info Gain	OneR	Relief	Symmetrical Uncertainty
x_{21}	x_{21}	x_{21}	x_{33}	x_{33}	x_{21}
x_{10}	x_{29}	x_{10}	x_{10}	x_{21}	x_{10}
x_{33}	x_{10}	x_{33}	x_{21}	x_{10}	x_{33}
x_{8}	x_{14}	x_{8}	x_{8}	x_{15}	x_{32}
x_{15}	x_{33}	x_{15}	x_{7}	x_{35}	x_{8}
x_{32}	x_{32}	x_{32}	x_{35}	x_{1}	x_{15}
x_{7}	x_{16}	x_{16}	x_{18}	x_{6}	x_{16}
x_{18}	x_{8}	x_{18}	x_{6}	x_{34}	x_{29}
x_{16}	x_{28}	x_{7}	x_{32}	x_{32}	x_{18}
x_{35}	x_{27}	x_{29}	x_{15}	x_{2}	x_{7}
x_{6}	x_{15}	x_{35}	x_{13}	x_{7}	x_{27}
x_{27}	x_{25}	x_{6}	x_{16}	x_{11}	x_{35}
x_{29}	x_{3}	x_{27}	x_{27}	x_{16}	x_{31}
x_{31}	x_{18}	x_{31}	x_{9}	x_{27}	x_{6}
x_{22}	x_{7}	x_{22}	x_{23}	x_{18}	x_{14}
x_{13}	x_{35}	x_{13}	x_{3}	x_{31}	x_{3}
x_{3}	x_{23}	x_{3}	x_{29}	x_{23}	x_{22}
x_{23}	x_{31}	x_{23}	x_{31}	x_{22}	x_{23}
x_{9}	x_{6}	x_{9}	x_{14}	x_{9}	x_{13}
x_{14}	x_{22}	x_{14}	x_{11}	x_{29}	x_{9}
x_{11}	x_{19}	x_{11}	x_{25}	x_{8}	x_{11}
x_{24}	x_{13}	x_{24}	x_{22}	x_{3}	x_{25}
x_{30}	x_{9}	x_{34}	x_{2}	x_{30}	x_{30}
x_{34}	x_{30}	x_{30}	x_{12}	x_{14}	x_{24}
x_{25}	x_{11}	x_{25}	x_{36}	x_{4}	x_{19}
x_{5}	x_{24}	x_{5}	x_{17}	x_{17}	x_{34}
x_{26}	x_{34}	x_{26}	x_{30}	x_{20}	x_{5}
x_{19}	x_{5}	x_{19}	x_{34}	x_{25}	x_{26}
x_{28}	x_{26}	x_{28}	x_{19}	x_{5}	x_{28}
x_{12}	x_{17}	x_{12}	x_{28}	x_{28}	x_{12}
x_{17}	x_{12}	x_{17}	x_{26}	x_{19}	x_{17}
x_{4}	x_{4}	x_{4}	x_{5}	x_{36}	x_{4}
x_{20}	x_{20}	x_{20}	x_{20}	x_{12}	x_{20}
x_{1}	x_{1}	x_{1}	x_{1}	x_{26}	x_{1}
x_{36}	x_{2}	x_{36}	x_{24}	x_{13}	x_{36}
x_{2}	x_{36}	x_{2}	x_{4}	x_{24}	x_{2}

Tabla 6.14: Ranking derivado de seis algoritmos de SV, conjunto de datos *Nursery*

Chi Squared	Gain Ratio	Info Gain	OneR	Relief	Symmetrical Uncertainty
x_8	x_8	x_8	x_8	x_8	x_8
x_2	x_2	x_2	x_2	x_2	x_2
x_1	x_1	x_1	x_1	x_1	x_1
x_5	x_7	x_7	x_7	x_5	x_7
x_7	x_5	x_5	x_5	x_7	x_5
x_4	x_4	x_4	x_4	x_6	x_4
x_3	x_6	x_3	x_6	x_4	x_6
x_6	x_3	x_6	x_3	x_3	x_3

En el ejemplo Chess, con *ventana* igual a 20 %, se encontró rápidamente el punto de detención para los clasificadores KStar e Ibk y se observó eficacia para SMO, Naive Bayes, J48Graft y Filtered; aumentar la *ventana* al 30 % identifica el mejor desempeño en todos los casos. La Tabla 6.15 presenta la máxima precisión de clasificación durante el proceso y el momento de detención de la evaluación.

Como se ha mostrado, el nuevo criterio propuesto tiene una alta eficacia para evitar la realización de búsquedas exhaustivas, ésto aplica también con conjuntos de datos que incluyen variables de diversos tipos. Se ha encontrado, que en general para aquellos casos donde se consideran sólo variables numéricas, el método es estable y encuentra con rapidez el máximo desempeño; determinando así, cuantas variables conforman el subconjunto seleccionado.

Para ejemplos donde predominan las variables categóricas, se sugiere cambiar SFS por SBE como método de búsqueda, ya que se observa una tendencia a encontrar los mejores comportamientos de clasificación, eliminando a las variables menos significativas en lugar de considerar primero a las más informacionales.

6.3.4. Comparativa con otra estrategia existente

Con la intención de fortalecer el proceso de validación del nuevo criterio, se ha diseñado una comparativa referida a la efectividad de los resultados que éste produce así como la de otros métodos ya reportados en la literatura.

El proceso de equiparación, se ha diseñado para observar las diferencias, ventajas y desventajas de la utilización del nuevo criterio, con un conjunto de datos en particular y utilizar el subconjunto de variables seleccionadas, para contrastarlo

con el que otros métodos ya validados propongan, con el mismo ejemplo de datos.

Tabla 6.15: Máxima precisión de los clasificadores por iteración, ranking derivado de Relief, conjunto de datos *Chess*.

		Precisión del clasificador (%)					
Prueba	Variable agregada	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
1	x_{33}	68.24	68.24	68.24	68.24	68.24	68.24
2	x_{21}	77.75	77.75	77.75	77.75	77.75	77.75
3	x_{10}	90.43	90.43	90.43	90.43	90.43	90.43
4	x_{15}	90.43	90.43	90.43	90.43	90.43	90.43
5	x_{35}	93.02	90.64	93.15	93.15	93.15	93.15
6	x_{1}	93.02	90.64	93.34	93.74	93.74	93.74
7	x_{6}	94.81	94.21	94.74	95.09	95.24	95.09
8	x_{34}	94.81	94.21	95.87	95.71	95.87	95.71
9	x_{32}	94.81	94.43	97.15	97.40	97.31	97.40
10	x_{2}	94.81	94.43	97.15	97.40	97.31	97.40
11	x_{7}	94.81	94.43	97.15	97.40	97.31	97.40
12	x_{11}	94.81	94.43	97.15	97.40	97.31	97.40
13	x_{16}	94.96	94.43	97.15	97.40	97.31	97.40
14	x_{27}	94.96	94.43	97.59	97.87	97.59	97.93
15	x_{18}	94.96	94.43	97.68	97.87	97.59	97.93
16	x_{31}	95.31	94.43	97.72	97.87	97.59	97.93
17	x_{23}	95.31	94.43	97.72	97.93	97.59	98.00
18	x_{22}	95.31	94.43	97.72	97.93	97.59	98.00
19	x_{9}	95.31	94.43	97.72	97.93	97.59	98.00
20	x_{29}	95.46	94.43	97.72	97.93	97.59	98.00
21	x_{8}	95.49		97.72	97.93	97.59	98.00
22	x_{3}	95.78		97.72	97.93	97.59	98.00
23	x_{30}	95.87		97.72	97.97	97.59	98.03
24	x_{14}	95.87		97.72	98.25	97.62	98.37
25	x_{4}	95.87		97.81	98.78	97.78	98.84
26	x_{17}	95.87		98.34	99.16	98.09	99.22
27	x_{20}	95.87		98.56	99.37	98.12	99.44
28	x_{25}	95.87		98.65	99.37	98.25	99.44
29	x_{5}	95.87		98.65	99.37	98.25	99.44
30	x_{28}	95.87		98.65	99.37	98.25	99.44
31	x_{19}	95.87		98.65	99.37	98.25	99.44
32	x_{36}	95.87		98.65	99.37	98.25	99.44
33	x_{12}	95.87		98.65	99.37	98.25	99.44
34	x_{26}	95.87		98.65	99.37	98.25	99.44
35	x_{13}			98.65	99.37	98.25	99.44
36	x_{24}			98.65	99.37	98.25	99.44

Conjunto de datos utilizado

El conjunto de datos que se utilizará en esta comparativa, podría ser seleccionado aleatoriamente. Sin embargo, de acuerdo al comportamiento de los rankings derivados de los selectores utilizados en las secciones previas, se ha decidido implementar esta nueva comparativa con el conjunto de datos *Chess*, ya descrito.

La razón para utilizar el ejemplo *Chess*, se debe a que previamente se completaron los procesos de búsquedas exhaustivas para todos los rankings, incluidos en la experimentación de la subsección 6.3.3, por lo que se conoce en todos los casos, el valor correspondiente al máximo desempeño global así como el subconjunto de variables con las que se obtiene éste.

Adicionalmente, se observó que la curva de desempeño en la evaluación de algunos rankings con el clasificador Ibk, presenta en general dos crestas, lo que vuelve este caso interesante, para averiguar el comportamiento y resultados, tanto del criterio propuesto como de otros métodos que se incluyan.

Algoritmos utilizados

El diseño de esta comparativa incluye tres métodos que producen un subconjunto de variables seleccionadas, se trata de (*i*) el nuevo criterio propuesto, (*ii*) el método reconocido como *Greedy Stepwise* (Caruana y Freitag, 1994) y (*iii*) un algoritmo genético denominado *Genetic search* (Witten y Frank, 2005).

Para el caso del nuevo criterio, se ha decidido utilizar los rankings correspondientes al uso de seis selectores Filter, estos son: Chi Squared, Gain Ratio, Info Gain, One R, Relief, y Symmetrical Uncertainty.

Con respecto a los métodos Greedy stepwise y Genetic search, éstos se han utilizado en combinación con cinco selectores multivariados, identificados como: Correlation Based Feature Selection, Classifier Subset, Consistency, Filtered y WrapperSubsetEval.

Para todos los algoritmos de SV y métodos de búsqueda mencionados, se ha utilizado la implementación disponible en la plataforma weka (Hall et al., 2009).

Con el fin de facilitar la presentación de resultados, la Tabla 6.16 muestra una lista de abreviaturas, éstas serán utilizadas posteriormente con los nombres de los selectores univariados y multivariados incluidos.

En resumen, los algoritmos univariados de SV, generarán diversos rankings de las variables del conjunto *Chess* para que el nuevo criterio propuesto culmine el proceso de selección de variables en cada uno. En el caso de los métodos multivariados, al utilizarse en combinación con los métodos de búsqueda, Greedy stepwise

y Genetic search, producirán otras propuestas de subconjuntos de variables.

Tabla 6.16: Lista de abreviaturas de todos los selectores utilizados.

	Num	Método de SV	Abreviatura
Univariados	1	Chi Squared	ChS
	2	Gain Ratio	GR
	3	Info Gain	IG
	4	One R	OR
	5	Relief	R
	6	Symmetrical Uncertainty	SU
Multivariados	7	Correlation Based Feature Selection	CFS
	8	Classifier Subset Eval	ClasfS
	9	Consistency Subset Eval	ConS
	10	Filtered Subset Eval	FiltS
	11	Wrapper Subset Eval	Wrap

A partir de los resultados mencionados anteriormente, es posible realizar la comparativa de todos ellos. La Tabla 6.17 presenta la distribución de los algoritmos de SV y métodos de búsqueda descrita.

Tabla 6.17: Combinación de selectores y métodos de búsqueda utilizados.

		Método de búsqueda		
		Filter	Greedy Stepwise	Genetic search
Selectores	1	ChS	CFS	CFS
	2	GR	ClasfS	ClasfS
	3	IG	ConS	ConS
	4	OR	FiltS	FiltS
	5	R	Wrap	Wrap
	6	SU		

En la subsección siguiente, se describe el procedimiento para la realización de la comparativa, así como una discusión de los resultados obtenidos.

Comparativa

Para comenzar con esta comparativa, se procede a aplicar los seis selectores univariados mencionados, al conjunto de datos *Chess*. A semejanza de los capítulos previos, en esta experimentación tampoco se reportará el tiempo requerido por cada algoritmo, el análisis estará centrado en la observación del comportamiento y

eficacia de los métodos utilizados.

Los resultados de la aplicación de los selectores univariados, se han organizado para presentarse en la Tabla 6.18 y como se espera, se observa que todos los rankings que se generaron, son diferentes entre sí. Ahora interesa determinar el subconjunto de variables seleccionadas por cada uno, así como sus correspondientes valores de desempeño y porcentaje de variables eliminadas.

Tabla 6.18: Concentrado de Rankings, conjunto de datos *Chess*.

Variable	ChS	GR	IG	OR	R	SU
1	x_{21}	x_{21}	x_{21}	x_{33}	x_{33}	x_{21}
2	x_{10}	x_{29}	x_{10}	x_{10}	x_{21}	x_{10}
3	x_{33}	x_{10}	x_{33}	x_{21}	x_{10}	x_{33}
4	x_{8}	x_{14}	x_{8}	x_{8}	x_{15}	x_{32}
5	x_{15}	x_{33}	x_{15}	x_{7}	x_{35}	x_{8}
6	x_{32}	x_{32}	x_{32}	x_{35}	x_{1}	x_{15}
7	x_{7}	x_{16}	x_{16}	x_{18}	x_{6}	x_{16}
8	x_{18}	x_{8}	x_{18}	x_{6}	x_{34}	x_{29}
9	x_{16}	x_{28}	x_{7}	x_{32}	x_{32}	x_{18}
10	x_{35}	x_{27}	x_{29}	x_{15}	x_{2}	x_{7}
11	x_{6}	x_{15}	x_{35}	x_{13}	x_{7}	x_{27}
12	x_{27}	x_{25}	x_{6}	x_{16}	x_{11}	x_{35}
13	x_{29}	x_{3}	x_{27}	x_{27}	x_{16}	x_{31}
14	x_{31}	x_{18}	x_{31}	x_{9}	x_{27}	x_{6}
15	x_{22}	x_{7}	x_{22}	x_{23}	x_{18}	x_{14}
16	x_{13}	x_{35}	x_{13}	x_{3}	x_{31}	x_{3}
17	x_{3}	x_{23}	x_{3}	x_{29}	x_{23}	x_{22}
18	x_{23}	x_{31}	x_{23}	x_{31}	x_{22}	x_{23}
19	x_{9}	x_{6}	x_{9}	x_{14}	x_{9}	x_{13}
20	x_{14}	x_{22}	x_{14}	x_{11}	x_{29}	x_{9}
21	x_{11}	x_{19}	x_{11}	x_{25}	x_{8}	x_{11}
22	x_{24}	x_{13}	x_{24}	x_{22}	x_{3}	x_{25}
23	x_{30}	x_{9}	x_{34}	x_{2}	x_{30}	x_{30}
24	x_{34}	x_{30}	x_{30}	x_{12}	x_{14}	x_{24}
25	x_{25}	x_{11}	x_{25}	x_{36}	x_{4}	x_{19}
26	x_{5}	x_{24}	x_{5}	x_{17}	x_{17}	x_{34}
27	x_{26}	x_{34}	x_{26}	x_{30}	x_{20}	x_{5}
28	x_{19}	x_{5}	x_{19}	x_{34}	x_{25}	x_{26}
29	x_{28}	x_{26}	x_{28}	x_{19}	x_{5}	x_{28}
30	x_{12}	x_{17}	x_{12}	x_{28}	x_{28}	x_{12}
31	x_{17}	x_{12}	x_{17}	x_{26}	x_{19}	x_{17}
32	x_{4}	x_{4}	x_{4}	x_{5}	x_{36}	x_{4}
33	x_{20}	x_{20}	x_{20}	x_{20}	x_{12}	x_{20}
34	x_{1}	x_{1}	x_{1}	x_{1}	x_{26}	x_{1}
35	x_{36}	x_{2}	x_{36}	x_{24}	x_{13}	x_{36}
36	x_{2}	x_{36}	x_{2}	x_{4}	x_{24}	x_{2}

Con los rankings obtenidos, ahora se realiza la aplicación del nuevo criterio, cabe mencionar que el algoritmo implícito incluye a SFS, como método de búsqueda interno. Al finalizar dicha aplicación, se ha logrado determinar los 11 subconjuntos de variables seleccionadas, que se presentan en la Tabla 6.19.

Tabla 6.19: Subconjuntos de variables seleccionadas por ranking univariado.

Variable	ChS	GR	IG	OR	R	SU
1	x_{21}	x_{21}	x_{21}	x_{33}	x_{33}	x_{21}
2	x_{10}	x_{29}	x_{10}	x_{10}	x_{21}	x_{10}
3	x_{33}	x_{10}	x_{33}	x_{21}	x_{10}	x_{33}
4	x_{8}	x_{14}	x_{8}	x_{8}	x_{15}	x_{32}
5	x_{15}	x_{33}	x_{15}	x_{7}	x_{35}	x_{8}
6	x_{32}	x_{32}	x_{32}	x_{35}	x_{1}	x_{15}
7	x_{7}		x_{16}	x_{18}	x_{6}	x_{16}
8	x_{18}		x_{18}	x_{6}	x_{34}	x_{29}
9	x_{16}		x_{7}	x_{32}	x_{32}	x_{18}
10	x_{35}		x_{29}	x_{15}	x_{2}	x_{7}
11	x_{6}		x_{35}		x_{7}	x_{27}
12	x_{27}		x_{6}		x_{11}	x_{35}
13			x_{27}		x_{16}	x_{31}
14					x_{27}	x_{6}

Cada uno de los subconjuntos de variables seleccionadas, derivados de los selectores univariados utilizados, han sido evaluados con el clasificador supervisado Ibk. La Tabla 6.20, muestra el desempeño de clasificación obtenido en cada caso, así como el porcentaje de variables eliminadas.

Tabla 6.20: Eficiencia obtenida mediante selectores univariados.

%	ChS	GR	IG	OR	R	SU
Desempeño	96.62	94.34	96.62	96.37	97.59	96.56
Variables Eliminadas	66.67	83.33	63.89	72.22	61.11	61.11

Para el caso de los subconjuntos de variables sugeridos por los algoritmos mutlivariados en combinación con los métodos de búsqueda *Greedy stepwise* y *Genetic search*, estos se presentan en la Tabla 6.21.

Tabla 6.21: Subconjuntos de variables seleccionadas por los algoritmos multivariados.

Variable	*Greedy stepwise*					*Genetic search*				
	CFS	ClasfS	ConS	Filts	Wrap	CFS	ClasfS	ConS	Filts	Wrap
1	x_3		x_{10}	x_{10}		x_3	x_{16}	x_1	x_8	x_{16}
2	x_{10}		x_{14}	x_{21}		x_8		x_2	x_{10}	
3	x_{15}		x_{21}	x_{33}		x_{10}		x_3	x_{11}	
4	x_{16}		x_{28}			x_{11}		x_4	x_{15}	
5	x_{21}		x_{32}			x_{15}		x_5	x_{16}	
6	x_{28}		x_{33}			x_{16}		x_6	x_{21}	
7	x_{32}					x_{21}		x_7	x_{24}	
8	x_{33}					x_{24}		x_8	x_{32}	
9						x_{28}		x_9	x_{33}	
10						x_{32}		x_{10}		
11						x_{33}		x_{11}		
12								x_{12}		
13								x_{13}		
14								x_{14}		
15								x_{15}		
16								x_{16}		
17								x_{17}		
18								x_{18}		
19								x_{19}		
20								x_{20}		
21								x_{21}		
22								x_{22}		
23								x_{23}		
24								x_{24}		
25								x_{25}		
26								x_{26}		
27								x_{27}		
28								x_{28}		
29								x_{29}		
30								x_{30}		
31								x_{31}		
32								x_{32}		
33								x_{33}		
34								x_{34}		
35								x_{35}		
36								x_{36}		

Puede observarse, que en el caso del método de búsqueda *Greedy stepwise*, con dos de los selectores se produce un conjunto vacío, mientras que con *Genetic search* y para los mismos algoritmos de SV, el subconjunto derivado solo integra una sola variable.

A partir de los subconjuntos de variables seleccionadas que han sido generados por los métodos multivariados, los resultados de clasificación obtenidos con el mismo clasificador Ibk, se presentan en la Tabla 6.22.

Tabla 6.22: Eficiencia obtenida mediante selectores multivariados.

Variable	Greedy stepwise					Genetic search				
	CFS	ClasfS	ConS	Filts	Wrap	CFS	ClasfS	ConS	Filts	Wrap
Desempeño	94.18	0.00	94.34	90.43	0.00	94.06	55.41	96.28	94.09	55.41
Variables Eliminadas	77.78	0.00	83.33	91.67	0.00	69.44	97.22	0.00	75.00	97.22

Para mejorar la observación de los resultados generados por el criterio propuesto en contraste con los obtenidos por los métodos de búsqueda *Greedy stepwise* y *Genetic search*, la Tabla 6.23 muestra un resumen de ellos, dónde se aprecia

Tabla 6.23: Resumen de resultados de desempeño y eliminación de variables.

	Num	Método de SV	Porcentaje de Desempeño	Porcentaje de Eliminación
Criterio	1	ChS	96.62	66.67
	2	GR	94.34	83.33
	3	IG	96.62	63.89
	4	OR	96.37	72.22
	5	R	97.59	61.11
	6	SU	96.56	61.11
Greedy Step.	7	CFS	94.18	77.78
	8	ClasfS	0.00	0.00
	9	ConS	94.34	83.33
	10	FiltS	90.43	91.67
	11	Wrap	0.00	0.00
Genetic s.	12	CFS	94.06	69.44
	13	ClasfS	55.41	97.22
	14	ConS	96.28	0.00
	15	FiltS	94.09	75.00
	16	Wrap	55.41	97.22

Al analizar los resultados de la Tabla 6.23, se puede observar que de manera general el máximo desempeño de clasificación se ha obtenido mediante la utilización del criterio propuesto, ésto en combinación con el selector univariado Relief y con un total de 61.11 % de variables eliminadas. En el caso de los métodos de búsqueda *Greedy stepwise* y *Genetic search*, el mejor comportamiento del clasificador corresponde al 94.34 % y 96.28 % respectivamente y aunque con *Greedy stepwise* se mejora la eliminación de variables, es preferible mantener un desempeño alto a eliminar muchas variables. Por otra parte, cualquiera de los resultados obtenidos por el criterio es produce mejor desempeño de clasificación que el más alto obtenido con *Greedy stepwise*.

Comparando el criterio con *Genetic search*, aunque la diferencia en la precisión del clasificador no es tan grande, pues corresponde a un 96.28 % de instancias correctamente clasificadas, para lograr este valor debe utilizar el total de variables, es decir, no hace eliminación de variables. Lo que impacta en la eficiencia del método de búsqueda. En complemento, aunque con *Genetic search* se obtienen también altos porcentajes de eliminación de variables en combinación con dos selectores, el desempeño es muy pobre alcanzando un total de 55.41 % en ambos casos.

Se ha hecho un ejercicio de comparación global de todos los selectores utilizados por cada método de búsqueda, obteniéndose que con el criterio en combinación con los algoritmos Filter univariados, se alcanza un promedio general del 94.49 % de desempeño general de clasificación, eliminando para ello un 63.38 % de variables. Con respecto del método de búsqueda *Greedy stepwise*, el promedio de desempeño llega al 55.79 % con una eliminación del 50.56 % de variables. Para el caso de *Genetic search*, se obtiene un 79.05 % general de instancias correctamente clasificadas y un porcentaje de eliminación de variables correspondiente al 67.78 %.

En resumen, se obtuvieron mejores resultados al emplear el nuevo criterio propuesto, con respecto de *Greedy stepwise* y *Genetic search*. Una línea futura para continuar la investigación en este ámbito, podría incluir más selectores multivariados y otros métodos de búsqueda, así como considerar particiones de datos de muy alta dimensionalidad.

6.3.5. Resultados obtenidos en la validación del criterio

Los resultados han evidenciado por una parte, la efectividad del nuevo criterio como método de análisis del comportamiento de los rankings y por otra, aportar una solución de fácil implementación computacional, para identificar un subcon-

junto de variables que produzca altos valores de clasificación. Adicionalmente, se realizaron algunas pruebas experimentales sustituyendo SFS por SBE, dado que en algunos casos el máximo desempeño se identificaba al final de la gráfica. El criterio trabaja de manera adecuada y aceptable con ambos métodos de búsqueda.

Se encontró también que, aunque para un mismo conjunto de datos, no con todos los clasificadores se producen curvas de desempeño cuya gráfica semeja a una campana de Gauss, en la mayoría de los casos si se presenta esta situación, por lo que una línea futura de investigación consiste en determinar bajo qué condiciones y características de las variables se observa dicho comportamiento, a fin de asegurar que el método propuesto garantiza sus resultados.

Con el fin de identificar el valor del parámetro *ventana* adecuado para el funcionamiento del criterio, se realizaron diversos experimentos utilizando cuatro alternativas: 10 %, 20 %, 30 % y 50 %.

En la Tabla 6.24 se presentan los resultados obtenidos durante toda la experimentación, se observa el porcentaje de eficacia obtenido en todas las pruebas. Cada fila indica el valor del parámetro dónde el criterio elige el subconjunto de variables que alcanzan el mejor desempeño de clasificación; como se observa, una ventana equivalente al 20 % produce el mejor resultado, de ahí que este valor será el recomendado para aplicar el criterio propuesto para la evaluación de rankings.

De manera general, el criterio de detención propuesto ha demostrado una efectividad adecuada para todos los conjuntos de datos en estudio. Sin embargo, se presentaron algunas pruebas donde al utilizar el tamaño de *ventana* del 20 %, no se logró identificar el máximo desempeño global de algún clasificador, produciendo una detención temprana. En esas pruebas se aumentó el valor del parámetro al 30 % mejorando considerablemente la eficacia.

Al aumentar el porcentaje de variables que debe considerar el valor del parámetro *ventana* se aumenta el número de casos donde el criterio identifica el mejor subconjunto de variables para los clasificadores, encontrándose que al utilizar un valor del 30 % se produce una efectividad acumulada del 80.56 %. Sin embargo, se recomienda no configurar *ventana* $> 20\%$, ya que la mejora es menor al 6 % en el número de casos exitosos y el incremento puede disminuir significativamente el rendimiento del criterio, aunque siempre será mejor que el realizar búsquedas exhaustivas.

Tabla 6.24: Eficacia del criterio propuesto de acuerdo al tamaño del parámetro *ventana*.

Ventana	Porcentaje de casos de éxito	
	Individual	Acumulado
10 %	8.01 %	8.01 %
20 %	66.88 %	74.92 %
30 %	6.35 %	80.56 %
50 %	12.56 %	93.12 %
100 %	6.88 %	100 %

De acuerdo a los resultados obtenidos en los 34 conjuntos de datos, el criterio muestra un buen comportamiento, reconociendo 917 de 1,224 experimentos realizados, esto equivale a la identificación del máximo desempeño de clasificación en un 74.92 % de los casos evaluados, para lograr lo anterior se utilizó una *ventana* de tamaño igual al 20 %. La información sobre los casos en donde el criterio, al utilizar el parámetro *ventana* = 20 % logró identificar el mejor desempeño para cada combinación selector - clasificador se presenta en la Tabla 6.25.

Adicionalmente, cabe mencionar que los resultados correspondientes a la base de datos *Madelon* (identificada con el número 23 en la Tabla 6.25) destacan por sobre el resto de conjuntos de datos, debido a que a pesar de su tamaño, el criterio logra identificar el mejor subconjunto de variables en 35 de las 36 pruebas realizadas para todas las combinaciones selector - clasificador. Lo anterior permite demostrar la aplicabilidad del criterio presentado, al utilizarlo con particiones de datos que incluyen cientos de variables y miles de instancias.

Tabla 6.25: Resumen de experimentos donde se identificó máximo desempeño en la combinación Selector-Clasificador.

BD	SMO	NB	KStar	J48 Graft	Ibk	Filtered	Total
1	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	36
2	gr,ig,o, sy	cs,gr,ig, o,sy	cs,gr,ig, o,sy	cs,ig,o, r,sy	cs,ig,o, r,sy	gr,o,r, sy	28
3	gr	cs,ig	o	gr,o,r,sy	cs,ig,r,sy	gr,o,r,sy	16
4	cs,gr,ig, o,sy	cs,gr,ig, o,r,sy	cs,gr,ig, r,sy	cs,gr,ig, r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, r	31
5	cs,gr,ig, o,sy	cs,gr,ig, r,sy	cs,gr,ig, r,sy	gr,o,r, sy	cs,gr,ig, o,r,sy	cs,ig,r	28
6	gr,ig,o	cs,gr,ig, sy	cs,gr,ig, o,sy	cs,gr,o	cs,gr,ig, r,sy	gr,o	22
7	cs,gr,ig, r,sy	cs,gr,ig, r,sy	cs,gr,ig, r,sy	cs,gr,ig, r,sy	cs,gr,ig, r,sy	cs,gr,ig, r,sy	30
8	r	cs,gr,ig, o,r,sy	r	cs,gr,ig, r,sy	cs,ig,sy	cs,ig,r, sy	20
9		cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy		cs,gr,ig, o,r,sy		18
10	r	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	31
11	cs,gr, o,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,s	cs,gr, o,r,sy	cs,gr,ig, o,r,sy	o,sy	29
12	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	36
13	gr,ig, o,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	r	cs,gr,ig, o,r,sy	29
14	o,r,sy	cs,gr,ig, o,r,sy	o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	o,r,sy	27
15	cs,gr,ig, o,r,sy	cs,gr,ig, sy	o,r	ig,o, r,sy	cs,gr,ig, o,r,sy	o,r	24
16	cs,gr,ig, r,sy	o,r	cs,gr,ig, r,sy	cs,gr,ig, r,sy	cs,ig,r, sy	cs,ig,r, sy	25
17	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	36
18		cs,ig,r	cs,ig,r, sy	cs,gr,ig, o,sy	cs,gr,ig, o,r	cs,ig,o, r	21
19		cs,gr,ig, o,r,sy	cs,gr,ig, r,sy	cs,gr,ig, sy	cs,gr,ig, o,r,sy	cs,gr,ig, sy	25
20	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	36
21	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	36
22	gr	gr,ig	o,r	r	gr,sy	ig,o,r	11
23	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, r,sy	35
24	o,r	cs,ig,o, r,sy	cs,gr,ig, o,r,sy	gr,o,r, sy	o,r	cs,gr,ig, r,sy	24
25	gr,sy	cs,gr,ig, o,r,sy	gr,r,sy	gr,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	26
26	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	36
27	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	36
28	cs,gr,ig, r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, r,sy	cs,gr,ig, o,sy	cs,gr,ig, r,sy	32
29	cs,gr,ig, o,r,sy	cs,ig,o, r,sy	sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy	30
30		cs,gr,ig, o,sy	cs,gr,ig, o,r,sy		cs,ig,sy		14
31	cs,gr,ig, o,r,sy		o,r	r	r	cs,gr,ig, o,r,sy	16
32	cs,gr,ig, o,r	cs,gr,ig, o,r	cs,gr,ig, o,r,sy	cs,gr,ig,o, r,sy	gr,o,r	cs,gr,ig, o,r,sy	31
33	cs,gr,ig, o,r,sy	cs,gr,ig, o,r,sy			cs,gr,ig, o,r,sy		18
34	o,r	cs,ig,o, sy	cs,gr,ig, o,sy	cs,gr,o, r	cs,gr,o	cs,gr,ig, o,r,sy	24
							917
Aplicaciones exitosas del criterio, con ventana = 20 %							74,92 %

6.4. Discusión del criterio propuesto

Este nuevo criterio propuesto, ha mostrado tener eficacia con conjuntos de datos de distintas características, se ha probado con bases de datos de diversos tamaños, se han utilizado datos de tipo numérico y categórico, se han incluido ejemplos de clasificación binaria y de orden mayor, se han evaluado particiones con muchas y pocas instancias. En términos generales, los resultados han permitido encontrar en la mayoría de los casos, el mejor subconjunto de variables y en aquellos casos donde no ha sido así, la precisión no está muy lejana del máximo global.

Por otra parte, el número de variables eliminadas se ha observado del orden del 60 %, lo que además de reducir en una cantidad significativa la dimensionalidad de los conjuntos de datos, permite su posterior tratamiento con menos recursos computacionales.

En general, el criterio propuesto utiliza una *ventana* igual al 20 % de las variables presentes en la descripción original de los objetos en estudio, permitiendo identificar el subconjunto que obtiene el mejor desempeño de clasificación, sin recurrir a la utilización de búsquedas exhaustivas. Si el valor de ventana aumenta, se podrían reconocer más casos de éxito. Sin embargo, esto implicaría realizar más iteraciones del criterio y la diferencia no es tan significativa.

Las características de los conjuntos de datos utilizados y los resultados obtenidos, permiten concluir que no hay incidencia entre el tipo de datos incluido y la aplicación del criterio propuesto. En todo caso, mientras se disponga de un ranking derivado de un selector Filter univariado, el criterio funcionará adecuadamente.

Se sugiere preferir el uso de SFS antes que SBE por la diferencia en el número de variables al inicio de las pruebas, que con SFS son menores, Por otra parte, no se ha explorado, bajo que condiciones pudiera valer la pena explorar simultáneamente ambas estrategias de búsqueda, con la finalidad de establecer si en términos generales, con alguno de estos métodos de búsqueda se identifica el subconjunto óptimo de variables con mayor prontitud. Ésta puede ser una línea de investigación futura y verificar si aún es posible realizar menos pruebas que al emplear búsquedas exhaustivas.

Adicionalmente, se propone comparar el nuevo criterio con otras estrategias de búsqueda y utilizar conjuntos de datos aún más grandes, para evaluar su funcionamiento con ejemplos catalogados como de muy alta dimensionalidad.

Capítulo 7

Integración de variables

No hay sustituto del trabajo duro

Thomas Alva Edison

Hasta ahora, se han abordado las etapas del método de estructuración de poblaciones distribuidas, referidas a la selección de las variables en las particiones locales. Este capítulo se enfoca en la manera de integrar las variables seleccionadas localmente, para obtener la nueva representación general de la población distribuida en estudio y proporcionar una clasificación global adecuada.

En la búsqueda de una nueva representación general a partir de variables seleccionadas localmente, en el contexto de las poblaciones distribuidas descritas en este documento, probablemente se puedan desarrollar múltiples formas de conformar la partición global deseada. Sin embargo, no es tarea fácil definir ¿cómo integrar de manera adecuada los resultados de las particiones locales en un entorno global?.

Una primer estrategia pudiera considerar, simplemente unir a todas las variables seleccionadas de cada partición local y conformar con ellas un nuevo conjunto de datos. Aunque, al hacer esto se corre el riesgo de incluir variables que no aporten información en la definición de objetos a nivel global, o bien que pudieran resultar redundantes con las de otros subconjuntos, lo que podría llevar a incluir a aquellas que en sus particiones locales son relevantes, pero en un ámbito más amplio pudieran no serlo. Por otra parte, el considerar a todas las variables locales no permite alcanzar uno de los objetivos fundamentales de la SV, que consiste en reducir la dimensionalidad del problema original.

Derivado de lo anterior, se tiene la necesidad de evaluar la factibilidad y pertinencia de cada variable considerada, pues a priori no se sabe si existe entre ellas alguna propiedad o característica que las haga prescindibles o imprescindibles.

Este capítulo, propone para la última etapa del nuevo método de estructuración de poblaciones distribuidas, una estrategia para la integración de variables seleccionadas localmente, la cual está basada en dos ideas principales, por una parte se consideran primero a las variables más significativas en sus ámbitos locales y por la otra, la aceptación de cada nueva variable, debe satisfacer algún criterio de correlación entre sí misma y las variables que pudieran existir previamente, en la representación general.

En las siguientes secciones, se presenta (*i*) la estructura del algoritmo diseñado para la integración de las variables provenientes de diversas particiones locales, con el fin de construir una representación general incluyendo su correspondiente nueva clasificación y (*ii*) una discusión de la aplicabilidad de dicho algoritmo.

7.1. Algoritmo de integración de variables

En esta sección, se describe la estructura general del algoritmo para la integración de variables a partir de diversas particiones locales. El subconjunto final de variables, definirá a la nueva representación general de la población distribuida en estudio y su correspondiente clasificación global.

Un actividad relevante en la construcción de la estrategia mencionada y que sirve como medio de validación de la misma, es la medición del desempeño de clasificación tanto de los Subconjuntos de Variables Seleccionadas Localmente (SVSL) como de la Representación General (RG) que se obtiene al final del proceso.

A partir de las poblaciones distribuidas descritas en esta tesis, -las que incluyen como una de sus particularidades, que las clases locales en sus distintas particiones son diferentes entre sí-, una característica que también se considera fundamental, es que cada una de las variables descriptoras locales, guardan un cierto grado de correlación con su clase.

Por lo anterior, la estrategia propuesta confiere especial importancia a dos aspectos fundamentales, por una parte (*i*) el cálculo de un Factor de Correlación (FC) de las variables descriptoras con respecto de una clase, para decidir si se acepta o no a cada variable en estudio en la RG y por la otra (*ii*) la generación de una nueva

clasificación que se corresponda con el conjunto de variables en un momento en particular.

La idea principal consiste de un proceso iterativo que evalúa a las variables de cada SVSL presente en la población distribuida, para decidir en función de los aspectos mencionados, su aceptación o rechazo en la nueva RG. De manera gráfica, el algoritmo general de integración de variables se presenta en la Figura 7.1.

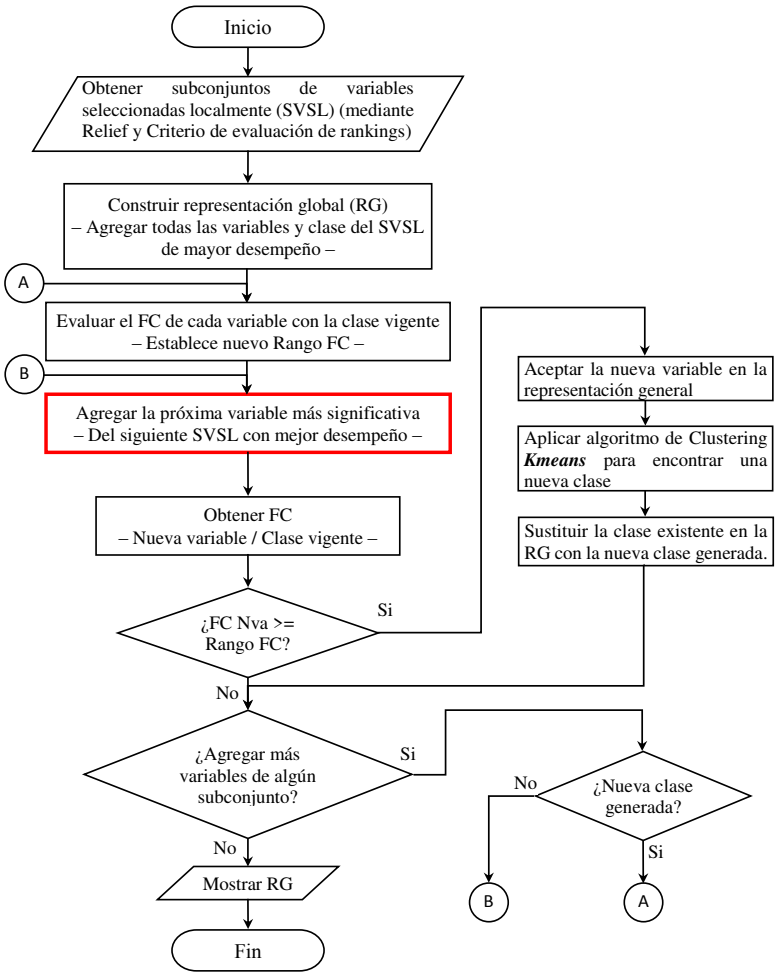

Figura 7.1: Estrategia para integración de variables.

Descripción del algoritmo de integración de variables

Los SVSL utilizados deberán ser obtenidos a partir de las particiones locales de datos que conforman a la población distribuida en estudio. Para lo que se requiere aplicar en éstas, al selector univariado Relief. Con el ranking derivado en cada partición se procede a aplicar el criterio de evaluación de rankings referido en el Capítulo 6. Como resultado de lo anterior se han generado los SVSL que se toman como entrada de la estrategia de integración de variables.

El análisis del comportamiento de clasificación de todos los SVSL, es el que permite identificar al SVSL que establecerá la primer estructura para construir la RG que se desea. Por lo que la estrategia utilizará al SVSL con mejor desempeño como la base de la RG incluyendo en este punto inicial, su clase local como la clase global.

La inclusión posterior de variables de otros SVSL, consiste en evaluar la satisfacción de un criterio basado en el cálculo del FC de cada nueva variable, con respecto de la clase incluida en la RG. Identificando un rango de FCs donde el valor mínimo es identificado como FC_{min} mientras que el más alto se reconoce como FC_{max}. FC_{min} es el valor que debe ser igualado o superado por cada nueva variable evaluada.

Si el FC obtenido con la nueva variable cumple con el criterio de superar el FC_{min} establecido, entonces es aceptada en la nueva RG y se procede a generar una nueva clasificación general mediante alguna técnica de Agrupamiento. La nueva clase generada ahora es considerada como la clase actual de la RG sustituyendo a la anterior.

El procedimiento es iterativo hasta agotar la evaluación de todas las variables presentes en los distintos SVSL. La RG obtenida en la última iteración conforma la nueva partición general con su correspondiente clasificación global.

A continuación, se presenta la descripción detallada de todos los pasos que conforman la estrategia propuesta.

1. Obtener subconjuntos de variables seleccionadas localmente (SVSL) y establecer un ranking de ellas en cada uno, Éstos deberán haber sido procesados mediante la aplicación del selector univariado Relief, así como del criterio de evaluación de rankings referido en el Capítulo 6.

2. Evaluar la precisión del clasificador Ibk para cada SVSL, a fin de establecer un orden descendente de todos ellos, de acuerdo a su desempeño individual.

3. Utilizar todas las variables y la clase original del SVSL que obtuvo el máximo desempeño de clasificación en el punto anterior, para generar una nueva partición donde se construirá una representación general (RG) de los datos así como una nueva clasificación global.

4. Determinar el FC de cada variable presente en la RG actual con respecto de su clase y determinar el valor de FC_{min}.

5. A partir del siguiente subconjunto con mejor desempeño (del punto 2), agregar a la RG, a aquella variable mejor posicionada en su ranking correspondiente.

6. Obtener el FC de la nueva variable (FC_{nva}) con respecto a la clase actual de la RG.

7. Si FC_{nva} es mayor o igual a FC_{min} continuar con el punto 8, en otro caso continuar con el punto 11.

8. Aceptar la nueva variable como parte de la RG.

9. Aplicar a la RG, un algoritmo de agrupamiento para obtener una nueva clasificación, este método deberá ser *Kmeans* (Lloyd, 1982) configurando su valor de k igual al número de las clases presentes en la RG actual.

10. Sustituir en la RG la clase anterior con la nueva clase generada en el punto 9.

11. Si aún hay más variables que agregar, considerar dos posibilidades, (*i*) si se generó una nueva clase (variable aceptada), regresar al punto 4, (*ii*) si no se generó una nueva clase (variable rechazada), continuar en el punto 5. Sin embargo, si ya no hay mas variables que agregar, continuar con el punto 12.

12. Mostrar RG, este conjunto contiene el subconjunto de variables que representan adecuadamente a los objetos en estudio y contiene la nueva clasificación global.

La utilización del algoritmo de integración de variables con diversos casos reales, se aborda en la sección 8.2. Sin embargo, para mejorar la comprensión del mismo, a continuación se explica a detalle su aplicación, con un ejemplo extraído del

conjunto de datos denominado *Wisconsin breast cancer*, proveniente del repositorio UCI (Lichman, 2009). La descripción del conjunto original mencionado, puede observarse en la Tabla 7.1.

Tabla 7.1: Descripción del conjuntos de datos Wisconsin breast cancer

Conjunto de datos	Instancias	Variables	Clases	Tipo
Wisconsin Breast Cancer	699	10	2	Num

Por razones de espacio y simplicidad en la siguiente ejemplificación, se utilizan nueve variables identificadas como x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9 y x_{10}, así como las primeras 40 instancias del conjunto original *Wisconsin breast cancer*. A esta partición se le denomina en adelante como *Wisconsin_R*, misma que se presenta en la Tabla 7.2.

Tabla 7.2: Conjunto de datos *Wisconsin_R*

Inst	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	Class
1	5	1	1	1	2	1	3	1	1	2
2	5	4	4	5	7	10	3	2	1	2
3	3	1	1	1	2	2	3	1	1	2
4	6	8	8	1	3	4	3	7	1	2
5	4	1	1	3	2	1	3	1	1	2
6	8	10	10	8	7	10	9	7	1	4
7	1	1	1	1	2	10	3	1	1	2
8	2	1	2	1	2	1	3	1	1	2
9	2	1	1	1	2	1	1	1	5	2
10	4	2	1	1	2	1	2	1	1	2
...					...					...
40	2	5	3	3	6	7	7	5	1	4

Posteriormente y para conformar un caso mínimo de población distribuida, este fragmento se ha dividido en dos. La primer partición incluye a las primeras cuatro variables y la segunda a las siguientes cinco. La clasificación original se ha sustituido por una nueva clase obtenida por el agrupador *Kmeans* configurado con k=2. Lo anterior debido a que se requiere que las clases en ambos subconjuntos de variables seleccionadas sean diferentes entre sí. La conformación final de ambos SVSL, se observa en las Tablas 7.3 y 7.4.

Tabla 7.3: SVSL0 de *Wisconsin_R*

Inst	x_2	x_3	x_4	x_5	Class
1	5	1	1	1	C1
2	5	4	4	5	C0
3	3	1	1	1	C1
4	6	8	8	1	C0
5	4	1	1	3	C1
6	8	10	10	8	C0
7	1	1	1	1	C1
8	2	1	2	1	C1
9	2	1	1	1	C1
10	4	2	1	1	C1
...		...			...
40	2	5	3	3	C1

Tabla 7.4: SVSL1 de *Wisconsin_R*

Inst	x_6	x_7	x_8	x_9	x_{10}	Class
1	2	1	3	1	1	C2
2	7	10	3	2	1	C3
3	2	2	3	1	1	C2
4	3	4	3	7	1	C3
5	2	1	3	1	1	C2
6	7	10	9	7	1	C3
7	2	10	3	1	1	C3
8	2	1	3	1	1	C2
9	2	1	1	1	5	C2
10	2	1	2	1	1	C2
...			...			...
40	6	7	7	5	1	C3

Cabe mencionar, que dado que el conjunto Wisconsin_R es un fragmento de una partición más grande con dos clases identificadas y que se usarán la mayoría de sus variables, cualquier evaluación que se realice sobre ellas mismas, provocará que probablemente el algoritmo de integración, acepte a todas. Sin embargo, se usa estas particiones como ejemplo demostrativo.

Para comenzar el proceso, se asume que las variables que conforman los SVSL 0 y 1 respectivamente, ya se encuentran ordenadas previamente, ésto de acuerdo a algún mérito o relevancia con respecto a su clase. Ahora, de acuerdo a las conclusiones obtenidas en el Capítulo 5, se procede a obtener la precisión del clasificador Ibk. En ambos casos, los resultados se presentan en la tabla 7.5.

Tabla 7.5: Precisión del clasificador en el ejemplo *Wisconsin_R*.

Partición	Variables	Clases	Precisión
SVSL0	4	2	97.5 %
SVSL1	5	2	100.0 %

A partir de los resultados de precisión mostrados y continuando con el algoritmo de integración de variables, el SVSL1 es el subconjunto elegido para conformar la base para construir la RG, dado que es el que obtuvo la precisión de clasificación más alta. Por lo tanto, en este momento la nueva RG está conformada por cinco variables descriptoras y la clase.

A continuación, se inicia el proceso iterativo para evaluar la agregación o rechazo de las variables provenientes del SVSL0, para lo que se requiere obtener el FC de cada variable presente en la nueva RG con respecto a su clase.

Los valores correspondientes a los FC de todas las variables incluidas en RG inicial, se presentan en la Tabla 7.6.

Tabla 7.6: FC de variables, RG inicial, ejemplo Wisconsin_R

Variables en RG inicial				
x_6	x_7	x_8	x_9	x_{10}
0.714	0.878	0.580	0.742	**0.273**
	Max			Min

A partir del FC_{min} identificado, en la primer iteración se debe igualar o superar un nuevo FC equivalente a 0.273 para aceptar una nueva variable, en este caso, se trata de x_2 del SVSL0, que de acuerdo a sus valores, obtiene un FC=0.566, por lo que ésta es aceptada en la nueva RG, creándose así una nueva estructura y clase. La Tabla 7.7 muestra la versión correspondiente a la RG en la primera iteración, con

la aceptación de la variable x_2 así como la nueva clase generada por el agrupador *Kmeans*.

Tabla 7.7: RG para la iteración 1, en el ejemplo *Wisconsin_R*

Inst	x_6	x_7	x_8	x_9	x_{10}	x_2	Class
1	2	1	3	1	1	5	C2
2	7	10	3	2	1	5	C3
3	2	2	3	1	1	3	C2
4	3	4	3	7	1	6	C3
5	2	1	3	1	1	4	C2
6	7	10	9	7	1	8	C3
7	2	10	3	1	1	1	C3
8	2	1	3	1	1	2	C2
9	2	1	1	1	5	2	C2
10	2	1	2	1	1	4	C2
...				...			...
40	6	7	7	5	1	2	C3

Para culminar esta primer iteración, se realiza la actualización de los FC correspondientes a todas las variables presentes en la RG actual, éstos se muestran en la Tabla 7.8.

Tabla 7.8: FC de variables, RG primera iteración, ejemplo Wisconsin_R

Variables en RG, primera iteración					
x_6	x_7	x_8	x_9	x_{10}	x_2
0.675	0.878	0.659	0.738	**0.250**	0.624
	Max			Min	

En el rango obtenido, se observa que el FC_{min} equivale a 0.250, por lo que en la segunda iteración, debe superarse dicho valor al evaluar a la variable x_3. Una vez realizado el cálculo, se obtiene para la nueva variable, un FC=0.765 superando el FC_{min} requerido, con lo que también es aceptada la variable en observación, requiriendo también la aplicación del agrupador *Kmeans* para realizar la actualización de la clase para la RG. Cabe mencionar que su FC es cercano al máximo de 0.878, siendo el segundo más alto en el rango, lo que evidencia que existe una alta correlación entre esta variable y la clase actual.

En esta segunda iteración, la RG está conformada por las siete variables descriptoras más su clase correspondiente y que se presentan en la tabla 7.9

Tabla 7.9: RG para la segunda iteración, en el ejemplo *Wisconsin_R*

Inst	x_6	x_7	x_8	x_9	x_{10}	x_2	x_3	Class
1	2	1	3	1	1	5	1	C0
2	7	10	3	2	1	5	4	C1
3	2	2	3	1	1	3	1	C0
4	3	4	3	7	1	6	8	C1
5	2	1	3	1	1	4	1	C0
6	7	10	9	7	1	8	10	C1
7	2	10	3	1	1	1	1	C0
8	2	1	3	1	1	2	1	C0
9	2	1	1	1	5	2	1	C0
10	2	1	2	1	1	4	2	C0
...				...				...
40	6	7	7	5	1	2	5	C1

Para culminar esta segunda iteración, se procede a actualizar la lista de los FC de las variables presentes en la RG actual con la clase vigente. La Tabla 7.10 muestra los valores correspondientes, dónde se observa que la variable evaluada, es la que mayor correlación guarda con la clase actual.

Tabla 7.10: FC de variables, RG segunda iteración, ejemplo Wisconsin_R

Variables en RG, segunda iteración						
x_6	x_7	x_8	x_9	x_{10}	x_2	x_3
0.782	0.738	0.685	0.781	**0.25**	0.738	0.826
				Min		Max

Ahora, corresponde evaluar a la variable x_4 del SVSL0, por lo que se procede a obtener su correspondiente FC con respecto de la clase actual de la RG, dicho valor equivale a 0.815, que es mayor que FC_{min} y de hecho es muy semejante a FC_{max}, por lo que también x_4 es aceptada. Cabe mencionar que hasta ahora, se ha presentado una tendencia a obtener un FC alto por cada nueva variable, lo que sugiere que la variación en las clases generadas en cada nueva iteración, es mínima.

Aceptar a la nueva variable x_4, da origen a una tercera iteración. La Tabla 7.11 presenta la estructura de la RG utilizada en esta evaluación incluyéndose la actualización de la clase correspondiente. Mientras que la actualización del FC de las variables en la nueva RG con respecto de su clase, se muestran en la Tabla 7.12.

Tabla 7.11: RG para la tercera iteración, en el ejemplo *Wisconsin_R*

Inst	x_6	x_7	x_8	x_9	x_{10}	x_2	x_3	x_4	Class
1	2	1	3	1	1	5	1	1	C0
2	7	10	3	2	1	5	4	4	C1
3	2	2	3	1	1	3	1	1	C0
4	3	4	3	7	1	6	8	8	C1
5	2	1	3	1	1	4	1	1	C0
6	7	10	9	7	1	8	10	10	C1
7	2	10	3	1	1	1	1	1	C0
8	2	1	3	1	1	2	1	2	C0
9	2	1	1	1	5	2	1	1	C0
10	2	1	2	1	1	4	2	1	C0
...				...					...
40	6	7	7	5	1	2	5	3	C1

Tabla 7.12: FC de variables, RG tercera iteración, ejemplo Wisconsin_R

Variables en RG, segunda iteración							
x_6	x_7	x_8	x_9	x_{10}	x_2	x_3	x_4
0.782	0.738	0.685	0.781	**0.25**	0.738	0.826	0.815
				Min		Max	

El siguiente paso es evaluar a la última variable del SVSL0, se trata de x_5, la que con respecto de la clase vigente en la RG actual, produce un FC igual a 0.731 que supera al FC_{min} de 0.250, provocando que x_5 sea aceptada en la RG final. La Tabla 7.13 presenta la RG en su última versión, incluyendo la clase global para este ejemplo.

Tabla 7.13: RG final, en el ejemplo *Wisconsin_R*

Inst	x_6	x_7	x_8	x_9	x_{10}	x_2	x_3	x_4	x_5	Class
1	2	1	3	1	1	5	1	1	1	C0
2	7	10	3	2	1	5	4	4	5	C1
3	2	2	3	1	1	3	1	1	1	C0
4	3	4	3	7	1	6	8	8	1	C1
5	2	1	3	1	1	4	1	1	3	C0
6	7	10	9	7	1	8	10	10	8	C1
7	2	10	3	1	1	1	1	1	1	C0
8	2	1	3	1	1	2	1	2	1	C0
9	2	1	1	1	5	2	1	1	1	C0
10	2	1	2	1	1	4	2	1	1	C0
...					...					...
40	6	7	7	5	1	2	5	3	3	C1

Para finalizar el proceso y a manera de comprobación, se han sometido al clasificador Ibk, los dos conjuntos de datos *Wisconsin_R* y RG final, encontrándose los resultados mostrados en la Tabla 7.14.

Tabla 7.14: Precisión del clasificador Ibk, con *Wisconsin_R* y RG final

Partición	Variables	Clases	Precisión
Wisconsin_R	9	2	90.0%
RG final	9	2	92.5%

De acuerdo a estos últimos resultados, el algoritmo de integración de variables ha identificado el subconjunto de éstas, que representa adecuadamente a la población distribuida Wisconsin_R, mejorando la precisión de clasificación al establecer el orden en el que las variables deben ser consideradas.

7.2. Discusión

En la sección anterior, se ha presentado una estrategia para integrar variables provenientes de las particiones locales que representan a una población distribuida. Los resultados presentados en la Tabla 7.14 han mostrado su efectividad, al establecer una partición base de la que se puede iniciar un proceso de evaluación sobre la pertinencia de cada nueva variable, llegando a una partición propuesta con

una representación general de todos los objetos en estudio incluyendo una nueva clasificación global.

Para el caso presentado, los desempeños de clasificación incluidos en la Tabla 7.14 muestran que para el caso del clasificador Ibk se mejora el resultado, al clasificar instancias con la RG respecto de la partición original utilizada. Esto puede ser extrapolado con otros conjuntos de datos pertenecientes a distintos ámbitos. Sin embargo, para respaldar esta aseveración, hace falta un proceso de validación más robusto que el aquí presentado, éste se aborda en el Capítulo 8.

Capítulo 8

Validación

Es más sabio averiguar que suponer

Mark Twain

En los capítulos previos se ha propuesto un método para crear la estructura global que represente adecuadamente a una población distribuida; ésto mediante la utilización de técnicas de RP.

El método de estructuración propuesto se enfoca entonces en el tratamiento de poblaciones distribuidas donde los objetos se representan simultáneamente en diversas particiones locales, pero con distintas variables descriptoras y clasificaciones en cada una.

En este capítulo, se aborda la fase de validación general, para lo cual el lector encontrará (*i*) una descripción de los conjuntos de datos utilizados así como la estrategia utilizada para generar particiones distribuidas cuando originalmente los datos se concentran en un solo conjunto, (*ii*) el proceso de validación del método de estructuración global en diversos casos, comenzando con uno referido al proceso de evaluación docente en una universidad, donde los datos por su naturaleza ya se encuentran distribuidos de origen, se incluyen también 10 casos de ejemplos sintéticos y que cuentan originalmente con una sola representación. Finalmente, (*iii*) una sección de discusión de resultados.

8.1. Conjuntos de datos para validación

Para la validación del método propuesto y a fin de demostrar su eficacia, se han elegido diversos conjuntos de datos reales con distintas características. Por una parte, se ha identificado un caso que se ajusta de origen, a las condiciones de particiones distribuidas descritas para este trabajo doctoral y por la otra, se han utilizado 10 ejemplos, extraídos de los repositorios de aprendizaje automático UCI (Lichman, 2009) y de la Universidad del Estado de Arizona (Arizona State University, 2016), desarrollado para un artículo sobre SV (Li et al., 2016).

Las subsecciones siguientes describen los conjuntos de datos utilizados y para el caso de los casos sintéticos, se incluye una estrategia de particionamiento creada para preparar y transformar los conjuntos de datos representados por una partición, en poblaciones distribuidas.

8.1.1. Ejemplo ITESA

El primer ejemplo está referido a un proceso de evaluación docente, desarrollado en el Instituto Tecnológico Superior del Oriente del Estado de Hidalgo (ITESA). En esta base de datos, se cuenta de inicio con una población distribuida, cuyas características incluyen la representación de los mismos objetos en todas sus particiones. En cada una de éstas, se utilizan variables descriptoras y clases locales distintas.

La evaluación docente mencionada se realiza regularmente cada ciclo escolar, cuya periodicidad es semestral. Las tablas de datos seleccionadas corresponden al semestre Enero - Junio 2014, debido al interés de dicha institución por conocer el comportamiento de sus docentes en ese periodo. Sin embargo, la misma experimentación puede replicarse para otros ciclos.

Las variables consideradas en las particiones de esta base de datos representan a las diversas preguntas sobre el desempeño de los docentes evaluados, las cuales están incluidas en dos cuestionarios individuales, distintos entre sí; el primero de ellos está dirigido a los estudiantes, mientras que el segundo es para las autoridades académicas. Cada instrumento busca evaluar aspectos diferentes de la práctica académica y administrativa de los profesores (Ver apéndice B).

En cada partición, los valores originales tanto de las variables descriptoras como de la clase se expresan a través de una escala diseñada para medir rasgos actitu-

dinales (Likert, 1932). Sin embargo, los docentes que pertenecen a una clase de acuerdo a la evaluación realizada por alumnos, no necesariamente están clasificados de la misma forma según el criterio de las autoridades académicas.

La opinión de los estudiantes sobre el desempeño en el aula de 94 docentes, se concentra en una partición denominada *Alumnos*, ésta incluye en su estructura 48 preguntas y se puede elegir entre las opciones A, B, C, D y E, correspondiendo A al valor más alto o de mejor desempeño y E al menor. Cabe mencionar que en esta partición, únicamente se han encontrado los valores A y B para la clase, por lo que es una partición local binaria (2 clases).

Otra partición reconocida como *Autoridades*, reúne el punto de vista que diversas autoridades académicas tienen sobre el desempeño de los mismos 94 profesores, el cuestionario correspondiente contiene 18 preguntas, diferentes en su totalidad a las de la opinión de los estudiantes. En este caso, los docentes se clasifican con valores desde A hasta E, lo que se refiere a una partición local multiclase.

Cada uno de los 94 docentes tiene entonces dos evaluaciones, con la posibilidad de que sus resultados sean semejantes o diferentes entre sí; dado que los estudiantes pueden considerar a un profesor como bueno, pero para las autoridades académicas pudiera no ser bueno o regular, sino malo. La Tabla 8.1 presenta la descripción estructural de las particiones en esta base de datos.

Tabla 8.1: Descripción de particiones en la base de datos ITESA

	Partición	Instancias	Variables	Tipo	Clases
1	Alumnos	94	48	Nominal	2
2	Autoridades	94	18	Nominal	5

Como se ha mencionado, los datos recopilados originalmente son de tipo nominal, con valores desde A hasta E, por lo que ha sido necesario numerizar. En la sustitución de valores se cambia la letra A por un número 1, B por 2 y sucesivamente hasta E por 5. En adelante, las versiones numerizadas de las tablas originales, son las que se utilizarán como *Alumnos* y *Autoridades* respectivamente.

En la tabla *Alumnos*, por cada instancia que representa la evaluación de un docente, el valor para su clase se ha obtenido mediante el promedio de la opinión de todos los estudiantes encuestados para ese profesor, mientras que en el caso de *Autoridades*, sólo se considera la valuación emitida por el jefe directo del profesor,

dado que en algunos casos un mismo maestro participa en dos o más áreas académicas, generando más evaluaciones.

En ningún caso, la distribución de clases está balanceada y aunque por tratarse de los mismos docentes y porque en ambos casos se refiere a la evaluación de su actuación profesional, se podría esperar que aquellos que tienen una etiqueta de clase en una tabla, la conserven en la otra, sin embargo, ésto no es una regla.

En este ejemplo, se pretende obtener una sola representación de los profesores evaluados, ésta debe incluir las preguntas adecuadas para describirlos desde un punto de vista más amplio y de acuerdo a una nueva clasificación global, a fin de proporcionar un panorama general en el conocimiento del perfil de los docentes.

8.1.2. Ejemplos provenientes de repositorios de datos

Como en el caso ITESA, es necesario identificar ejemplos otros ejemplos con dominio es cuya distribución y características de sus objetos sea equivalente. Sin embargo, en esta tesis no se presenta algún otro conjunto de datos del tipo mencionado, entre otras cosas, debido a que normalmente esos casos se refieren al ámbito empresarial o gubernamental en general, dónde comúnmente se manejen aspectos de confidencialidad y seguridad que pudieran limitar el acceso a los datos. Esta es la razón principal de incluir el procesamiento de datos sintéticos.

Por lo expuesto anteriormente, los ejemplos sintéticos que se describen en esta subsección incluyen en su estructura una partición única, éstos serán sometidos a un proceso de particionamiento inicial, mismo que los dispondrá en particiones distribuidas.

Se han elegido conjuntos de datos con diferentes características entre sí, a fin de validar el método de estructuración con ejemplos pertenecientes a distintos ámbitos, los que además están constituidos por representaciones de diversas variables y dimensionalidad.

Los conjuntos de datos utilizadas se presentan en la Tabla 8.2, dónde se han marcado con un asterisco aquellos que provienen del repositorio UCI (Lichman, 2009), el resto de ejemplos se obtuvieron del sitio web para SV de la Universidad del Estado de Arizona (Arizona State University, 2016).

Tabla 8.2: Descripción de conjuntos de datos a particionar.

	Ejemplo	Instancias	Variables	Tipo	Clases
*1	Chess	3,196	36	Nominal	2
2	Isolet	1,560	617	Numérico	26
3	Lung discrete	73	325	Numérico	7
*4	Madelon	2,000	500	Numérico	2
*5	Molecular	3,190	60	Nominal	3
*6	Mushroom	8,124	22	Nominal	2
*7	Spambase	4,601	57	Numérico	2
8	Warp	130	2,400	Numérico	10
9	Wine Quality W	4,898	11	Nominal	7
10	Yale	165	1,024	Numérico	15

En el ejemplo *Chess* hay 36 variables que describen un tablero de ajedrez e incluye 3,196 instancias, cada una representa una posición en dicho tablero. La clasificación es binaria y se refiere al estado *ganador* o *no ganador* de las diferentes posiciones, con el 52 % y 48 % de instancias, respectivamente.

El caso del conjunto *Isolet* está referido al dominio de reconocimiento del habla, por lo que se incluyen 26 clases y cada una está referida a la identificación de una letra del alfabeto, las que se describen por 617 variables, en total se han tratado 1,560 ejemplos.

Para *Lung discrete*, el dominio está catalogado en el área biológica concretamente como un caso de micro arreglo de datos genéticos referido al tema de padecimientos de pulmón, los datos están definidos por 325 variables y 73 ejemplos incluidos en siete clases.

El ejemplo *Madelon* está referido a la predicción de cáncer a partir de datos de espectrometría de masas. Fue diseñado para evaluar algoritmos de SV, por lo que no se provee información de los objetos que describen sus 500 variables, con la intención de no influenciar los resultados que se pudieran obtener. Los 2,000 objetos que contiene se agrupan en dos clases posibles, cada uno representa un paciente *sano* o *enfermo*, la distribución de instancias es balanceada.

La base de datos *Molecular* trata de secuencias de unión de empalmes genéticos en primates, con 60 variables que describen 3,190 observaciones de comportamientos celulares y conformadas en tres grupos posibles, consideradas como *donante*, *aceptador* o *ninguno*, con una distribución de clases que se corresponde con 25 %, 25 % y 50 % de manera respectiva.

Mushroom es un conjunto de datos en el que se describen 22 características físicas de 8,124 hongos clasificados en dos categorías, con una distribución de 48.2 % de ejemplos *venenosos* y 51.8 % *comestibles*. Una característica especial en este caso específico, es que se tiene identificada una variable con 2,480 datos faltantes, lo que permitirá observar su influencia en el comportamiento del método de estructuración presentado en esta tesis.

El ejemplo denominado *Spambase* incluye la descripción mediante 57 variables para 4,601 correos electrónicos, éstos se clasifican en *spam* o *no spam* con un 39.4 % y 60.6 % de forma correspondiente.

Con respecto al ejemplo *Warp*, el dominio corresponde a reconocimiento facial incluyendo 10 clases, para lo que se utilizan 2,400 variables y se ejemplifican 130 casos.

Wine Quality W es un ejemplo que trata sobre 11 variables que describen a las características del vino blanco, mismas que sirven para identificar siete clases de éste, se incluyen 4,898 instancias.

El conjunto de datos *Yale* es otro ejemplo circunscrito al reconocimiento facial, en este caso se tienen 15 clases descritas mediante 1,024 variables, para lo que se consideran 165 instancias.

Por otra parte, para estandarizar el proceso de validación del método de estructuración global de poblaciones distribuidas, y a semejanza del tratamiento de los conjuntos de datos utilizados en los capítulos previos, en todas las particiones incluidas en esta experimentación se ha modificado la nomenclatura de las variables desde sus representaciones originales, cambiándose por nombres genéricos identificados por $X = \{x_1, x_2, x_3, x_4, ..., x_n, class\}$ donde n corresponde al total de variables en cada base de datos.

Dado que se han considerado conjuntos de datos que requieren particionamiento previo para su tratamiento, en la siguiente sección, se presenta la estrategia utilizada para la construcción de particiones de datos.

8.1.3. Estrategia de particionamiento

En la literatura se han reportado diversas maneras de particionar un conjunto de datos, en algunos casos se considera la importancia de las variables implícitas y en

otros no. Sin embargo, esta tarea no es trivial dado que a menos que se conozca el contexto del objeto de estudio, se requiere la utilización de técnicas de RP para determinar la trascendencia de las variables y decidir cuáles de ellas deben incluirse en una misma partición.

Con el objetivo de fragmentar conjuntos de datos con una sola representación se requiere de una estrategia para conformar diversas particiones a fin de que éstas puedan tratarse con el método para procesamiento de poblaciones distribuidas propuesto, se ha utilizado una estrategia de particionamiento, la que es mostrada en su concepción general mediante el algoritmo presentado en la Figura 8.1.

Esta estrategia toma como punto de partida el conjunto original de datos y para decidir cuántos grupos de variables deben construirse, así como la distribución de las mismas, se comienza por aplicar una técnica de pivotaje a los datos (transposición), considerando las columnas como filas y viceversa, se ignora para este proceso la variable referida a la clase de los objetos incluidos.

	Etapa	Descripción
1	Validación del Conjunto Original de Datos (CO)	Revisión de la estructura del conjunto original de datos CO, integrado por una sola tabla.
2	Trasponer CO para generar CT	Se transpone CO, las filas se vuelven columnas y viceversa. Este nuevo conjunto se reconoce como CT.
3	Agrupamiento en CT	Se aplica el agrupador *K-means* al conjunto CT, dónde k = número de clases en CO.
4	Creacion de particiones	Cada cluster identificado en CT es separado, formando así una nueva partición local, ésta se transpone para recuperar la estructura original.
5	Agrupamiento en particiones locales	Se aplica *K-means* a cada partición local para crear las clases correspondientes.

Figura 8.1: Descripción de la estrategia de particionamiento utilizada.

El algoritmo de agrupamiento utilizado en adelante es el conocido *K-means* (Lloyd, 1957), (Lloyd, 1982). Éste se elige por sus características de efectividad, velocidad y amplia utilización reportadas en la literatura (MacQueen, 1967), (Lopez-Escobar, 2007), (Hernández-Valadez, 2006).

Una vez que se cuenta con el conjunto transpuesto de los datos, se aplica *K-means*, asignando a su parámetro interno k un valor igual al número de clases presentes en el conjunto original de datos, lo cual permitirá obtener k grupos que posteriormente se separarán.

Los grupos ya separados deben volverse a transponer para recuperar su estructura original. En este punto, se han generado k particiones del conjunto inicial, sin contar con clasificación propia en cada uno; solo se conoce la clase original que representa a los objetos con el total de variables. Por lo anterior, hace falta determinar una nueva clase local en cada partición.

Ahora se aplica el mismo agrupador *K-means* a las particiones locales, se configura k igual al número original de clases. Como resultado se obtiene una nueva clasificación local para cada partición, éstas son independientes entre sí.

En este punto de la estrategia, se ha conseguido conformar el escenario de población distribuida que se tratará a partir de ahora. Las representaciones que se han generado, incluyen a los mismos objetos representados en todas ellas con variables excluyentes y sus correspondientes clases locales.

La estrategia de particionamiento presentada permite fragmentar conjuntos de datos que originalmente incluyen una sola representación para que éstos se apeguen a las características de poblaciones distribuidas mencionadas en esta tesis.

A continuación, se presenta la aplicación de la estrategia de particionamiento con los ejemplos de datos que provienen de representaciones únicas, a fin de conformar poblaciones distribuidas de éstas.

8.1.4. Preparación de poblaciones distribuidas

Para contar con ejemplos de datos que se puedan utilizar en el proceso de validación del método de estructuración propuesto en esta tesis, debe aplicarse la estrategia de particionamiento descrita anteriormente en conjuntos de datos cuya representación original conste de una partición.

Para abordar la construcción de las representaciones distribuidas para los conjuntos de datos descritos en la Tabla 8.2, en la subsección siguiente se presenta el proceso completo para la obtención de las particiones derivadas del ejemplo *Chess*.

Para el resto de casos y por razones de espacio, se presenta un resumen en el que se detalla la estructura de los fragmentos que habrán de tratarse.

Conjunto de datos *Chess*

Este conjunto de datos se ajusta a una clasificación binaria, por lo que para este caso en particular, se establece para el agrupador *K-means*, el parámetro k con un valor igual a 2.

En el proceso de pivotaje del conjunto original *Chess* −que cuenta con 36 variables y 3,196 instancias− se genera una nueva representación transpuesta, ésta partición resultante ahora está conformada por 36 filas y 3,196 columnas, la que puede observarse en la Figura 8.2. En el proceso se ignora a la variable correspondiente a la clase.

var / inst	x_1	x_2	x_3	...	x_{36}
i_1	f	f	f	…	n
i_2	f	f	f	…	n
i_3	f	f	f	…	n
...	…	…	…	…	…
i_{3196}	t	f	t	…	n

→

inst / var	i_1	i_2	i_3	...	i_{3196}
x_1	f	f	f	…	t
x_2	f	f	f	…	f
x_3	f	f	f	…	t
...	…	…	…	…	…
x_{36}	n	n	n	…	n

Figura 8.2: Esquema de transposición del conjunto de datos *Chess*.

A partir de la versión transpuesta de la partición original *Chess* se aplica el agrupador *K-means* obteniéndose dos clústers. A continuación, las filas que conforman cada clúster deberán separarse del resto para crear nuevos subconjuntos, creando así dos fragmentos de datos.

Posteriormente, en cada nuevo fragmento corresponde recuperar la estructura original de los datos, por lo que debe volverse a transponer cada uno de ellos. Con este procedimiento se obtienen dos particiones locales del conjunto original, éstas contienen sus correspondientes variables excluyentes y entre ambas suman el total de las 36 originales; con respecto de las instancias, en cada partición están las 3,196. Hasta este punto, las particiones todavía no poseen clasificación.

De acuerdo a la estrategia de particionamiento, ahora se debe aplicar nuevamente el agrupador *K-means* a cada partición, generándose así una clasificación local. La Tabla 8.3 muestra el total de variables y clases obtenidas en las particiones 0 y 1 respectivamente.

Tabla 8.3: Variables y clases en las particiones 0 y 1 del ejemplo *Chess*

Partición	Variables	Clases
0	7	2
1	29	2

Las listas de variables que conforman las particiones 0 y 1 respectivamente, se presentan en la Tabla 8.4

Tabla 8.4: Variables incluidas en las particiones creadas en el ejemplo *Chess*

Partición	Variables										
0	7	x_{13}	x_{15}	x_{18}	x_{26}	x_{34}	x_{35}	x_{36}			
1	29	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
		x_{11}	x_{12}	x_{14}	x_{16}	x_{17}	x_{19}	x_{20}	x_{21}	x_{22}	x_{23}
		x_{24}	x_{25}	x_{27}	x_{28}	x_{29}	x_{30}	x_{31}	x_{32}	x_{33}	
Total	36										

Estas particiones 0 y 1 están listas para aplicarles el procedimiento de integración de variables, mismo que forma parte del método de estructuración de poblaciones distribuidas. Este proceso se abordará más adelante.

Distribución de particiones en ejemplos sintéticos

De manera análoga al ejemplo de datos *Chess*, se ha aplicado el proceso de fragmentación al resto de los conjuntos de datos sintéticos descritos en la Tabla 8.2, conformándose en cada caso versiones distribuidas de éstas. La Tabla 8.5 muestra el detalle del número de particiones y variables de todas las representaciones de objetos involucradas.

Tabla 8.5: Variables por cada partición de las poblaciones sintéticas utilizadas

Partición	Conjunto de datos									
	Chess	Isolet	Lung discrete	Madelon	Molecular	Mushroom	Spambase	Warp	Wine Quality W	Yale
0	7	56	39	140	11	9	3	225	1	49
1	29	10	47	360	25	13	54	217	2	87
2		30	32		24			244	1	145
3		10	27					200	1	45
4		19	42					233	1	152
5		31	69					170	1	43
6		34	69					215	4	110
7		39						424		36
8		37						138		29
9		22						334		51
10		34								73
11		25								81
12		29								31
13		12								51
14		31								41
15		8								
16		10								
17		20								
18		24								
19		24								
20		29								
21		25								
22		29								
23		1								
24		18								
25		10								
Instancias	3196	1560	73	2000	3190	8124	4601	130	4898	165
Variables	36	617	325	500	60	22	57	2400	11	1024
Clases	2	26	7	2	3	2	2	10	7	15

Las listas completas de variables que conforman a cada una de las nuevas particiones no se presentan por razones de espacio y por ser un proceso que se ha mostrado previamente en el caso *Chess*, por lo que sólo se muestran algunas de ellas. La Tabla 8.6 permite observar el caso de la distribución para *Lung discrete*, así como la Tabla 8.7 incluye la conformación de variables por partición en el conjunto de datos *Molecular* y la Tabla 8.8 hace lo propio con *Spambase*.

Tabla 8.6: Variables incluidas en las particiones creadas en el ejemplo *Lung discrete*

Partición	Variables															
0	x_{10}	x_{12}	x_{66}	x_{70}	x_{97}	x_{134}	x_{161}	x_{163}	x_{179}	x_{183}	x_{198}	x_{211}	x_{233}	x_{238}	x_{239}
	x_{242}	x_{247}	x_{251}	x_{256}	x_{270}	x_{285}	x_{286}	x_{288}	x_{289}	x_{292}	x_{294}	x_{295}	x_{297}	x_{299}	x_{300}
	x_{301}	x_{303}	x_{304}	x_{305}	x_{308}	x_{310}	x_{315}	x_{323}	x_{325}						
1	x_{19}	x_{29}	x_{40}	x_{45}	x_{47}	x_{55}	x_{56}	x_{57}	x_{58}	x_{59}	x_{60}	x_{61}	x_{62}	x_{63}	x_{64}
	x_{67}	x_{68}	x_{71}	x_{74}	x_{79}	x_{80}	x_{81}	x_{82}	x_{83}	x_{84}	x_{91}	x_{92}	x_{95}	x_{98}	x_{100}
	x_{101}	x_{103}	x_{104}	x_{105}	x_{115}	x_{118}	x_{124}	x_{125}	x_{126}	x_{127}	x_{132}	x_{133}	x_{136}	x_{137}	x_{140}
	x_{141}	x_{143}													
2	x_{72}	x_{75}	x_{77}	x_{78}	x_{85}	x_{86}	x_{87}	x_{88}	x_{89}	x_{90}	x_{93}	x_{96}	x_{102}	x_{106}	x_{107}
	x_{109}	x_{110}	x_{111}	x_{113}	x_{114}	x_{117}	x_{119}	x_{120}	x_{121}	x_{122}	x_{123}	x_{128}	x_{129}	x_{130}	x_{135}
	x_{139}	x_{142}													
3	x_1	x_3	x_4	x_5	x_6	x_7	x_{13}	x_{14}	x_{15}	x_{16}	x_{24}	x_{26}	x_{28}	x_{31}	x_{32}
	x_{34}	x_{35}	x_{37}	x_{38}	x_{166}	x_{200}	x_{202}	x_{206}	x_{253}	x_{273}	x_{284}	x_{311}			
4	x_2	x_{18}	x_{22}	x_{39}	x_{65}	x_{69}	x_{73}	x_{146}	x_{155}	x_{158}	x_{164}	x_{188}	x_{191}	x_{195}	x_{197}
	x_{199}	x_{201}	x_{207}	x_{214}	x_{217}	x_{220}	x_{221}	x_{224}	x_{232}	x_{243}	x_{246}	x_{248}	x_{249}	x_{252}	x_{255}
	x_{258}	x_{262}	x_{272}	x_{274}	x_{278}	x_{282}	x_{296}	x_{307}	x_{314}	x_{316}	x_{319}	x_{321}			
5	x_8	x_9	x_{11}	x_{17}	x_{20}	x_{23}	x_{25}	x_{27}	x_{30}	x_{36}	x_{46}	x_{147}	x_{148}	x_{149}	x_{151}
	x_{152}	x_{153}	x_{157}	x_{159}	x_{160}	x_{162}	x_{169}	x_{172}	x_{175}	x_{178}	x_{180}	x_{187}	x_{190}	x_{192}	x_{194}
	x_{196}	x_{204}	x_{208}	x_{209}	x_{215}	x_{218}	x_{225}	x_{226}	x_{230}	x_{231}	x_{236}	x_{237}	x_{240}	x_{241}	x_{245}
	x_{254}	x_{257}	x_{260}	x_{261}	x_{263}	x_{265}	x_{266}	x_{267}	x_{268}	x_{269}	x_{271}	x_{276}	x_{280}	x_{287}	x_{290}
	x_{291}	x_{293}	x_{298}	x_{309}	x_{312}	x_{313}	x_{317}	x_{318}	x_{324}						
6	x_{21}	x_{33}	x_{41}	x_{42}	x_{43}	x_{44}	x_{48}	x_{49}	x_{50}	x_{51}	x_{52}	x_{53}	x_{54}	x_{76}	x_{94}
	x_{99}	x_{108}	x_{112}	x_{116}	x_{131}	x_{138}	x_{144}	x_{145}	x_{150}	x_{154}	x_{156}	x_{165}	x_{167}	x_{168}	x_{170}
	x_{171}	x_{173}	x_{174}	x_{176}	x_{177}	x_{181}	x_{182}	x_{184}	x_{185}	x_{186}	x_{189}	x_{193}	x_{203}	x_{205}	x_{210}
	x_{212}	x_{213}	x_{216}	x_{219}	x_{222}	x_{223}	x_{227}	x_{228}	x_{229}	x_{234}	x_{235}	x_{244}	x_{250}	x_{259}	x_{264}
	x_{275}	x_{277}	x_{279}	x_{281}	x_{283}	x_{302}	x_{306}	x_{320}	x_{322}						
Total	325														

Tabla 8.7: Variables incluidas en las particiones creadas con el ejemplo *Molecular*

Partición	Variables											
0	x_{14}	x_{27}	x_{28}	x_{29}	x_{33}	x_{34}	x_{37}	x_{40}	x_{46}	x_{52}	x_{55}	
1	x_2	x_6	x_{17}	x_{20}	x_{30}	x_{31}	x_{32}	x_{35}	x_{38}	x_{39}	x_{41}	x_{42}
	x_{43}	x_{44}	x_{45}	x_{47}	x_{48}	x_{49}	x_{50}	x_{51}	x_{53}	x_{54}	x_{56}	x_{57}
	x_{59}											
2	x_1	x_3	x_4	x_5	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{15}
	x_{16}	x_{18}	x_{19}	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{36}	x_{58}	x_{60}
Total	60											

Tabla 8.8: Variables incluidas en las particiones creadas con el ejemplo *Spambase*

Partición	Variables											
0	x_{19}	x_{21}	x_{57}									
1	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{20}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}
	x_{27}	x_{28}	x_{29}	x_{30}	x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	x_{36}	x_{37}	x_{38}
	x_{39}	x_{40}	x_{41}	x_{42}	x_{43}	x_{44}	x_{45}	x_{46}	x_{47}	x_{48}	x_{49}	x_{50}
	x_{51}	x_{52}	x_{53}	x_{54}	x_{55}	x_{56}						
Total	57											

Con el proceso de particonamiento en los conjuntos de datos mencionados se crearon con éxito versiones distribuidas de todos ellos, cuyas clasificaciones fueron definidas por las variables descriptoras incluidas en cada partición local, por lo que no necesariamente son iguales de una partición a otra. A partir de este punto, todas las bases de datos cumplen con las características requeridas para la aplicación del método de estructuración de poblaciones distribuidas.

En la sección siguiente se presenta la aplicación del método propuesto en todos los conjuntos de datos descritos.

8.2. Aplicación del método propuesto

Una vez que se cuenta con las condiciones requeridas para todos los ejemplos referidos en la sección 8.1, se presenta a continuación la aplicación del método de estructuración de poblaciones distribuidas. En primer instancia, se aborda el caso de evaluación docente universitaria y posteriormente, el caso de los ejemplos sintéticos.

Al finalizar la aplicación del método para cada uno de los conjuntos de datos, se presenta una tabla que a manera de comparativa muestra la diferencia entre el desempeño de clasificación antes y después del proceso.

Para el caso ITESA, dado que su estructura inicial ya está distribuida, es necesario construir un conjunto de datos que incluye a todas las variables originales en una misma representación para comparar su desempeño con el que se obtiene en la RG correspondiente. Para el resto de ejemplos, se toman las versiones antes del particionamiento y las que se obtienen al final del proceso para realizar la evaluación respectiva.

8.2.1. Caso *ITESA*

Con el ejemplo *ITESA*, el proceso de integración de variables provenientes de los subconjuntos de variables seleccionadas de las particiones *Alumnos* y *Autoridades*, inicia mediante la identificación de la partición que conformará la estructura base sobre la que se construirá la nueva RG.

Para establecer la estructura base mencionada, se requiere la aplicación del selector Relief en ambas particiones locales. En la Tabla 8.9, se presenta el ranking obtenido en cada caso.

Tabla 8.9: Ranking obtenido por Relief en las particiones *Alumnos* y *Autoridades*

Posición	Alumnos	Autoridades
1	x_{22}	x_8
2	x_{14}	x_9
3	x_3	x_{10}
4	x_{46}	x_{12}
5	x_{18}	x_6
6	x_{23}	x_3
7	x_{12}	x_{14}
8	x_{19}	x_7
9	x_1	x_2
10	x_{47}	x_{16}
11	x_7	x_{13}
12	x_{48}	x_{17}
13	x_{21}	x_{11}
14	x_5	x_{18}
15	x_{37}	x_5
16	x_{36}	x_4
17	x_4	x_{15}
18	x_{15}	x_1
19	x_{20}	
20	x_{31}	
21	x_{29}	
22	x_{24}	
23	x_{11}	
24	x_{17}	
25	x_{25}	
26	x_{38}	
27	x_{13}	
28	x_6	
29	x_{30}	
30	x_{16}	
31	x_{40}	
32	x_{34}	
33	x_{45}	
34	x_{42}	
35	x_{28}	
36	x_9	
37	x_8	
38	x_{39}	
39	x_{10}	
40	x_{44}	
41	x_2	
42	x_{27}	
43	x_{43}	
44	x_{41}	
45	x_{32}	
46	x_{33}	
47	x_{35}	
48	x_{26}	

A partir de estos rankings, se utiliza el método de búsqueda SFS en combinación con el clasificador supervisado Ibk, para establecer en cada partición un subconjunto de variables con las que se obtiene el mayor número de instancias correctamente clasificadas. La Tabla 8.10 muestra los valores obtenidos de forma exhaustiva, observándose un comportamiento ascendente al inicio de la lista, posteriormente se identifica un punto máximo y finalmente, se aprecia un descenso que continúa hasta considerar todas las variables.

De acuerdo a la estrategia presentada en la Figura 7.1 y para concluir con la primer fase de SV, deben construirse subconjuntos de variables seleccionadas localmente, por lo que ahora corresponde establecer el valor del parámetro *ventana* para ambas particiones.

En la Tabla 8.10 puede observarse un recuadro con un número de elementos que equivale al 20 % de las variables presentes en cada partición, conformando así, el valor del parametro *ventana* para cada una de ellas. Se han incluido todas las pruebas de manera exhaustiva para mostrar que la aplicación del criterio propuesto para la selección de subconjuntos de variables funciona adecuadamente, encontrando el mejor subconjunto en ambas particiones locales.

Tabla 8.10: Desempeño del clasificador IBK mediante SFS y el ranking Relief

Iteración	Alumnos		Autoridades	
	Variable agregada	Desempeño	Variable agregada	Desempeño
1	x_{22}	98.936	x_8	67.021
2	x_{14}	98.936	x_9	69.149
3	x_3	96.809	x_{10}	64.894
4	x_{46}	96.809	x_{12}	68.085
5	x_{18}	97.872	x_6	76.596
6	x_{23}	96.809	x_3	78.723
7	x_{12}	97.872	x_{14}	76.596
8	x_{19}	97.872	x_7	78.723
9	x_1	98.936	x_2	78.723
10	x_{47}	97.872	x_{16}	**91.489**
11	x_7	**100.000**	x_{13}	87.234
12	x_{48}	100.000	x_{17}	89.362
13	x_{21}	100.000	x_{11}	89.362
14	x_5	100.000	x_{18}	90.426
15	x_{37}	100.000	x_5	93.617
16	x_{36}	98.936	x_4	91.489
17	x_4	98.936	x_{15}	90.426
18	x_{15}	98.936	x_1	91.489
19	x_{20}	98.936		
20	x_{31}	98.936		
21	x_{29}	98.936		
22	x_{24}	97.872		
23	x_{11}	98.936		
24	x_{17}	97.872		
25	x_{25}	97.872		
26	x_{38}	97.872		
27	x_{13}	97.872		
28	x_6	97.872		
29	x_{30}	97.872		
30	x_{16}	97.872		
31	x_{40}	97.872		
32	x_{34}	97.872		
33	x_{45}	97.872		
34	x_{42}	97.872		
35	x_{28}	97.872		
36	x_9	97.872		
37	x_8	97.872		
38	x_{39}	97.872		
39	x_{10}	97.872		
40	x_{44}	97.872		
41	x_2	97.872		
42	x_{27}	98.936		
43	x_{43}	98.936		
44	x_{41}	98.936		
45	x_{32}	98.936		
46	x_{33}	98.936		
47	x_{35}	97.872		
48	x_{26}	98.936		

Una vez obtenidos los dos subconjuntos de variables seleccionadas localmente, reconocidos como SVSL0 y SVSL1 respectivamente. La asignación de la etiqueta SVSL0 será para el subconjunto con el mejor desempeño de clasificación, que como puede observarse en la Tabla 8.10 será para el que se conformará con las variables seleccionadas de la partición *Alumnos*, dado que su mejor desempeño es del 100 % mientras que en el caso de *Autoridades* es del 91.489 %.

Con respecto de la eficiencia del criterio de evaluación de rankings, en el caso de la partición *Alumnos*, se realizaron 22 de 49 iteraciones o pruebas, lo que equivale a no realizar las últimas 27 de ellas, esto corresponde a un ahorro de procesamiento del 55.10 % que se traduce tanto en tiempo como en recursos computacionales. Para el caso de la partición *Autoridades* el porcentaje de pruebas no realizadas es del 21.05 % y aunque es menor que para *Alumnos*, también es significativo.

Los resultados obtenidos en la primer etapa de la estrategia de la Figura 7.1 referidos a la selección de variables en las particiones locales para conformar los SVSL0 y SVSL1 se presentan en la Tabla 8.11. Se han incluido los subíndices *a* y *b* a las variables de ambos subconjuntos, para distinguirlas unas de otras.

Tabla 8.11: Subconjuntos de variables seleccionadas localmente, caso *ITESA*

Variable	SVSL0 Alumnos	SVSL1 Autoridades
1	x_{22a}	x_{8b}
2	x_{14a}	x_{9b}
3	x_{3a}	x_{10b}
4	x_{46a}	x_{12b}
5	x_{18a}	x_{6b}
6	x_{23a}	x_{3b}
7	x_{12a}	x_{14b}
8	x_{19a}	x_{7b}
9	x_{1a}	x_{2b}
10	x_{47a}	x_{16b}
11	x_{7a}	
Mejor desempeño	100 %	91.489 %

De acuerdo a estos resultados, entonces el SVSL0 se considera como la partición base, sobre la que se debe construir la representación global al incluir todas sus variables descriptoras. En la Tabla 8.12 se muestra la estructura de la RG inicial.

Tabla 8.12: Estructura de la RG inicial

Variables										
x_{22a}	x_{14a}	x_{3a}	x_{46a}	x_{18a}	x_{23a}	x_{12a}	x_{19a}	x_{1a}	x_{47a}	x_{7a}	Clase

Continuando con el proceso, el cálculo del FC de las variables incluidas en la Tabla 8.12 con respecto de su clase, se presentan en la Tabla 8.13, dónde se observa que el FC de la variable x_{12a} corresponde al valor más alto de esta medida (FC_{max} = -0.670) mientras que el más bajo es el de la variable x_{1a} (FC_{min} = -0.715).

Tabla 8.13: FC de las variables de la RG inicial con respecto de su clase

Variables										
x_{22a}	x_{14a}	x_{3a}	x_{46a}	x_{18a}	x_{23a}	x_{12a}	x_{19a}	x_{1a}	x_{47a}	x_{7a}
-0.701	-0.712	-0.698	-0.707	-0.696	-0.692	-0.670	-0.705	-0.715	-0.697	-0.688
						Max		Min		

Ahora, para aplicar el proceso de integración de variables, se procede a evaluar a las variables incluidas en el SVSL1, comenzando con la más significativa, a fin de decidir su aceptación en la RG; en este caso se trata de la variable x_{8b}, que produce un FC = -0.087 y dado que es mayor a FC_{min} = -0.715, entonces es aceptada y se genera una nueva clasificación, conformando así la siguiente RG. La estructura correspondiente a esta primer iteración, se presenta en la Tabla 8.14.

Tabla 8.14: Estructura de la RG en la iteración 01, caso *ITESA*

Partición	Clases	Variables									
RG	5	12	x_{22a}	x_{14a}	x_{3a}	x_{46a}	x_{18a}	x_{23a}	x_{12a}	x_{19a}	x_{1a}
			x_{47a}	x_{7a}	x_{8b}						

Este proceso iterativo continúa, aceptando a las variables x_{9b}, x_{10b}, x_{12b}, x_{6b}, x_{3b} y x_{14b}, hasta que en la iteración 07 la variable x_{7b} no es aceptada, debido a que no supera el FC con respecto de la última clase obtenida, a saber, para x_{7b} se obtiene un FC = -0.201 y debe superar un FC_{min} = -0.104. Se procede ahora a evaluar a la siguiente variable x_{2b}, calculándose un FC = 0.143 que permite aceptarla. Finalmente, al evaluar a la variable restante x_{16b} también resulta admitida.

Al finalizar el proceso de integración, se ha acumulado un total de 20 variables, quedando la estructura final como se puede observar en la Tabla 8.15.

Tabla 8.15: Estructura de la RG en la iteración final, caso *ITESA*

Partición	Clases	Variables									
RG	5	20	x_{22a}	x_{14a}	x_{3a}	x_{46a}	x_{18a}	x_{23a}	x_{12a}	x_{19a}	x_{1a}
			x_{47a}	x_{7a}	x_{8b}	x_{9b}	x_{10b}	x_{12b}	x_{6b}	x_{3b}	x_{14b}
			x_{2b}	x_{16b}							

Un resumen del desempeño y descripción de los subconjuntos de datos incluidos en esta experimentación, se presentan en la Tabla 8.16. Cabe mencionar que para fines comparativos, se generó una versión de esta base de datos a través de la unión de las dos particiones iniciales *Alumnos* y *Autoridades*, en una sola representación que incluye a todas sus variables originales y que suman 66 en total. La clase incluida en esta representación se obtuvo utilizando el agrupador *K-means*, estableciendo k=5 por ser el número de clases que originalmente se tienen etiquetadas en ambas particiones.

Tabla 8.16: Desempeño del clasificador Ibk, caso *ITESA*

Conjunto de datos	Clasificación	Clases	Desempeño	Variables
Original	**Generada**	5	**84.04**%	**66**
Partición 0 (Alumnos)	Local	2	98.94%	48
Partición 1 (Autoridades)	Local	5	91.49%	18
SVSL 0 (Alumnos)	Local	2	100.00%	11
SVSL 1 (Autoridades)	Local	5	90.43%	10
Final	**Nueva**	5	**87.23**%	**20**

La experimentación realizada con el conjunto de datos *ITESA* ha permitido encontrar una representación global de los docentes evaluados, integrándose un subconjunto de variables provenientes de las dos particiones iniciales, con las que se obtiene una nueva clasificación global.

Con respecto de la validación del proceso, se encontró que la nueva RG presenta un desempeño de clasificación mayor que en el estado inicial, con un valor final de 87.23%. Adicionalmente, aunque el objetivo de este trabajo doctoral no

es elilminar variables, cabe mencionar que el desempeño mencionado se alcanzó mediante una eliminación del 69.70 % de variables. Esto significa una reducción importante de la dimensionalidad del problema original, pues se ha prescindido de 46 variables, sin perdida de la capacidad de descripción de los 94 profesores a los que se hace referencia además de incluir una nueva clasificación global para todos ellos.

8.2.2. Caso *Chess*

El conjunto de datos *Chess*, se describe completo en esta subsección. Sin embargo, dado que para los casos *Madelon*, *Molecular*, *Mushroom* y *Spambase*, el proceso es el mismo, sólo se presentarán las tablas de resultados finales.

A partir de la versión distribuida de la base de datos *Chess*, es posible aplicar la estrategia de construcción de la representación global para poblaciones distribuidas, misma que comienza con el análisis de las particiones locales, dado que internamente no se sabe que variables son las de mayor mérito de acuerdo a su clase local, por lo que el selector *Relief* es aplicado, obteniendo así un ranking de variables en cada partición. La Tabla 8.17 muestra las variables de cada partición de acuerdo a su importancia.

Tabla 8.17: Ranking de variables en las particiones creadas en el ejemplo *Chess*

Partición	Variables										
0	7	x_{34}	x_{18}	x_{15}	x_{13}	x_{36}	x_{35}	x_{26}			
1	29	x_{36}	x_6	x_{11}	x_1	x_{32}	x_{31}	x_{27}	x_{10}	x_{29}	x_4
		x_{30}	x_7	x_3	x_{20}	x_{16}	x_8	x_{23}	x_{21}	x_{12}	x_{19}
		x_{14}	x_{28}	x_{25}	x_{17}	x_9	x_2	x_{22}	x_5	x_{24}	

36 Variables

Con los rankings obtenidos para cada partición, se realiza la selección de las variables que conforman los Subconjuntos de Variables Seleccionadas Localmente, SVSL0 y SVSL1 respectivamente. Lo anterior de acuerdo al criterio de detención en evaluación de rankings presentado en la Figura 6.1; mismo que produce los dos SVSL presentados en la Tabla 8.18.

Tabla 8.18: Subconjuntos de variables seleccionadas localmente, ejemplo *Chess*

SVSL	Variables				Precisión
0	2	x_{34}	x_{18}		100,00 %
1	3	x_{36}	x_6	x_{11}	100,00 %
	5	Variables			

En ambos SVSL, la precisión del clasificador Ibk se mantiene en el 100.00 %, el proceso continúa con la construcción de la RG, para lo que debe elegirse el SVSL con mejor desempeño y dado que en ambos casos se tiene el mismo valor, se decide iniciar con la alternativa de considerar a SVSL0 como la partición base y de acuerdo al algoritmo presentado en la Figura 7.1, se evaluarán las variables de SVSL1 para decidir si se agregan o no a la RG. El establecimiento de SVSL1 como partición base da lugar a otro comportamiento, éste se abordará más adelante en esta misma sección,

A continuación, debe obtenerse el FC de cada variable presente en el SVSL0 dado que la RG se genera a partir de dicho subconjunto, la Tabla 8.19 muestra el correspondiente valor de los factores de correlación de cada variable descriptora con respecto de la clase inicial.

Tabla 8.19: Factores de correlación, variables descriptoras - clase de RG inicial, ejemplo *Chess*

FC en RG inicial		
x_{34}	x_{18}	Clase
0.875	0.767	1.000

De los valores obtenidos en el cálculo del FC en la RG inicial, se observa que el menor corresponde a 0.767, lo que establece que al evaluar la primer variable del SVSL1, debe obtenerse un FC mayor o igual para que la variable sea aceptada.

Al comparar el valor de FC de la nueva variable y el FC mínimo de la RG inicial, se agrega o no dicha variable a la RG. En el caso de aceptarse, también se genera una nueva clase mediante el agrupador *K-means* y se vuelven a calcular los nuevos FC correspondientes; mientras que en el caso de que sea rechazada, se conserva la estructura actual de la RG inicial y se procede a calcular el FC de la siguiente variable en SVSL1, repitiendo este procedimiento hasta completar todas las posibles nuevas variables.

La Tabla 8.20 presenta los valores de FC de las variables de SVSL1 con respecto de la clase de la RG inicial, Se observa que para cada una, fueron menores a 0.767, por lo que en ningún caso se agregaron variables a la RG, por lo que se conserva la estructura del SVSL0 como RG final.

Tabla 8.20: Factores de correlación de los SVSL, ejemplo *Chess*

SVSL0			SVSL1		
x_{34}	x_{18}	Clase	x_{36}	x_6	x_{11}
0.875	0.767	1.000	-0.054	-0.042	0.018

Los resultados mostrados hasta ahora se corresponden con la alternativa de haber conformado la partición base con el SVSL0. Sin embargo, el SVSL1 presenta el mismo desempeño de clasificación aunque con más variables, por lo que a continuación se muestra el proceso de integración, considerando el SVSL1 como la RG inicial.

La Tabla 8.21 muestra los valores correspondientes de FC entre las variables del SVSL1 y su respectiva clase local. Se puede observar que el FC mínimo es un valor negativo de -0.519 para la variable x_{36} mientras que el resto son positivos; este valor es el que deben superar o igualar los FC de las variables de SVSL0 al ser evaluadas.

Tabla 8.21: Factores de correlación entre las variables descriptoras y la clase en SVSL1 como RG inicial, ejemplo *Chess*

RG inicial			
x_{36}	x_6	x_{11}	Clase
-0.519	0.451	0.553	1.000

Al evaluar la primer variable del SVSL0 (x_{34}), se obtiene un FC de -0.014, este valor es mayor que el FC minimo de la RG inicial de -0.519 por lo que la nueva variable es aceptada, creándose una nueva clase mediante el agrupador *Kmeans* y la RG ahora es conformada por la estructura representada en la Tabla 8.22, dónde también se observan los FC actualizados.

Tabla 8.22: Factores de correlación RG iteracion 1, ejemplo *Chess*

RG Iteración 1				
x_{36}	x_6	x_{11}	x_{34}	Clase
-0.435	0.377	0.478	0.375	1.000

Para el resto de iteraciones, se encontró que en todos los casos las nuevas variables fueron aceptadas, por lo que la RG final tiene la estructura mostrada en la Tabla 8.23.

Tabla 8.23: Estructura de la RG final, ejemplo *Chess*

RG final					
x_{36}	x_6	x_{11}	x_{34}	x_{18}	Clase global

La RG final también es sometida a un proceso de clasificación para determinar su precisión, a fin de compararse con el desempeño que se obtiene utilizando la representación original del conjunto de datos *Chess*, encontrándose que hay un mejoramiento global y una disminución sustancial de variables. Debe observarse que además se cuenta ahora con una clasificación global de todos los objetos en estudio. Un resumen del comportamiento de todos los conjuntos de datos involucrados se muestra en la Tabla 8.24.

Tabla 8.24: Desempeño del clasificador, caso *Chess*

Conjunto de datos		Clasificación	Clases	Desempeño	Variables
Original		**Inicial**	2	**96.28**%	**36**
Partición 0		Local	2	100.00%	7
Partición 1		Local	2	100.00%	29
SVSL 0		Local	2	100.00%	2
SVSL 1		Local	2	100.00%	3
Final	**1**	**General**	**2**	**100.00**%	**2**
	2	**General**	**2**	**100.00**%	**5**

Las dos versiones finales obtenidas de RG, permiten alcanzar un desempeño de clasificación del 100.00%, aunque hay diferencia en el número de variables que las conforman. Sin embargo, con la idea de integrar variables de dos o más parti-

ciones, se establece la segunda alternativa como la que corresponde a una nueva representación del ejemplo de datos *Chess*.

8.2.3. Caso *Isolet*

Los resultados alcanzados en la experimentación con el ejemplo *Isolet* incluyen las evaluaciones en el desempeño de las particiones locales, de los SVSL derivados de estos y de la RG obtenida.

Debido al número de variables y por razones de espacio, los resultados por partición, SVSL y RG se presentan por separado. En el caso de las particiones locales, éstos se muestran en la Tabla 8.25.

Tabla 8.25: Desempeño del clasificador en particiones, caso *Isolet*

Conjunto de datos	Clasificación	Clases	Desempeño	Variables
Original	**Inicial**	26	**83.56**%	**617**
Partición 0	Local	26	86.15%	56
Partición 1	Local	26	76.47%	10
Partición 2	Local	26	76.73%	30
Partición 3	Local	26	94.81%	10
Partición 4	Local	26	81.47%	19
Partición 5	Local	26	73.65%	31
Partición 6	Local	26	76.54%	34
Partición 7	Local	26	72.82%	39
Partición 8	Local	26	68.72%	37
Partición 9	Local	26	75.51%	22
Partición 10	Local	26	67.82%	34
Partición 11	Local	26	65.96%	25
Partición 12	Local	26	78.85%	29
Partición 13	Local	26	74.17%	12
Partición 14	Local	26	78.97%	31
Partición 15	Local	26	80.51%	8
Partición 16	Local	26	90.38%	10
Partición 17	Local	26	80.32%	20
Partición 18	Local	26	77.18%	24
Partición 19	Local	26	74.87%	24
Partición 20	Local	26	76.86%	29
Partición 21	Local	26	78.59%	25
Partición 22	Local	26	75.26%	29
Partición 23	Local	26	100%	1
Partición 24	Local	26	77.18%	18
Partición 25	Local	26	95.83%	10

Con respecto a los resultados de desempeño y variables incluidas en cada SVSL, éstos se pueden observar en la Tabla 8.26. En la parte final se incluyen los valores correspondientes al desempeño y variables que conforman la RG respectiva.

Tabla 8.26: Desempeño del clasificador en SVSL, caso *Isolet*

Conjunto de datos	Clasificación	Clases	Desempeño	Variables
SVSL0	Local	26	86.41 %	52
SVSL1	Local	26	76.47 %	10
SVSL2	Local	26	76.73 %	30
SVSL3	Local	26	94.81 %	10
SVSL4	Local	26	81.47 %	19
SVSL5	Local	26	73.65 %	31
SVSL6	Local	26	77.12 %	32
SVSL7	Local	26	72.82 %	39
SVSL8	Local	26	68.72 %	36
SVSL9	Local	26	75.51 %	22
SVSL10	Local	26	67.95 %	32
SVSL11	Local	26	65.96 %	25
SVSL12	Local	26	78.85 %	29
SVSL13	Local	26	74.17 %	12
SVSL14	Local	26	78.97 %	31
SVSL15	Local	26	80.51 %	8
SVSL16	Local	26	90.38 %	10
SVSL17	Local	26	80.32 %	20
SVSL18	Local	26	77.18 %	24
SVSL19	Local	26	75.77 %	23
SVSL20	Local	26	76.92 %	28
SVSL21	Local	26	78.59 %	25
SVSL22	Local	26	75.26 %	29
SVSL23	Local	26	100 %	1
SVSL24	Local	26	77.24 %	17
SVSL25	Local	26	95.83 %	10
RG	**General**	26	**85.32 %**	**327**

Como se puede observar, se ha alcanzado una reducción importante en el número de variables ya que se conservan 327 de las 617 variables originales, esto equivale a prescindir de un 47.00 % de ellas. Sin embargo, el desempeño de clasificación aumenta un 1.76 % pasando del 83.56 % al 85.32 %.

8.2.4. Caso *Lung discrete*

Para la experimentación con la base de datos *Lung discrete* se han obtenido las evaluaciones del desempeño del clasificador Ibk, tanto para las particiones locales, como para los SVSL que se obtienen de éstas, así como de la RG resultante.

En este caso, también se presentan por separado los resultados por partición, para los SVSL y para la RG . En el caso de las particiones locales, éstos se muestran en la Tabla 8.27.

Tabla 8.27: Desempeño del clasificador en particiones, caso *Lung discrete*

Conjunto de datos	Clasificación	Clases	Desempeño	Variables
Original	**Inicial**	7	**83.56**%	**325**
Partición 0	Local		71.23%	39
Partición 1	Local		84.93%	47
Partición 2	Local		76.71%	32
Partición 3	Local		75.34%	27
Partición 4	Local		72.6%	42
Partición 5	Local		80.82%	69
Partición 6	Local		56.36%	69

Con respecto a los resultados de desempeño y variables incluidas en cada SVSL, éstos se pueden observar en la Tabla 8.26. En la parte final se incluyen los valores correspondientes al desempeño y variables que conforman la RG respectiva.

Tabla 8.28: Desempeño del clasificador en SVSL, caso *Lung discrete*

Conjunto de datos	Clasificación	Clases	Desempeño	Variables
SVSL0	Local		78.08%	21
SVSL1	Local		89.04%	31
SVSL2	Local		75.34%	2
SVSL3	Local		73.97%	5
SVSL4	Local		80.82%	28
SVSL5	Local		76.71%	12
SVSL6	Local		90.41%	43
RG	**General**	7	**93.15**%	**132**

En las tablas anteriores se puede apreciar que también se se logra una eliminación considerable del número de variables en la RG con respecto de la representación original, dado que se utilizan únicamente 132 de 325 de ellas, lo que equivale a un 59.38%. Por parte de la precisión de clasificación, ésta aumenta un 9.59%, pasando del 83.56% al 93.15%.

8.2.5. Caso *Madelon*

Para el ejemplo *Madelon*, el resultado final se presenta en la Tabla 8.29. Se aprecia una eliminación de variables bastante amplia, conservando solamente 12 de las 500 variables de inicio, esto equivale a prescindir de un 97.60 % de ellas sin perder precisión en el clasificador, por el contrario el desempeño aumenta un 43.0 % pasando del 54.15 % al 97.15 %.

El ranking derivado del selector *Relief*, generó una curva de comportamiento con su máximo global al utilizar las variables con mayor mérito. Este comportamiento es el que sustenta la eliminación tan significativa de variables.

Tabla 8.29: Desempeño del clasificador, caso *Madelon*

Conjunto de datos	Clasificación	Clases	Desempeño	Variables
Original	**Inicial**	2	**54.15** %	**500**
Partición 0	Local	2	74.95 %	140
Partición 1	Local	2	73.60 %	360
SVSL 0	Local	2	94.75 %	6
SVSL 1	Local	2	97.10 %	6
Final	**General**	2	**97.15** %	**12**

8.2.6. Caso *Molecular*

Con respecto al desempeño de clasificación, aunque la diferencia no es tan significativa pasando del 74.67 % al 75.36 %, está mejoría se alcanza con una reducción del 35 % de variables presentes en la RG, ya que ésta incluye a 39 de las 60 presentes en la representación original.

Tabla 8.30: Desempeño del clasificador, caso *Molecular*

Conjunto de datos	Clasificación	Clases	Desempeño	Variables
Original	**Inicial**	3	**74.67** %	**60**
Partición 0	Local	3	73.79 %	11
Partición 1	Local	3	67.52 %	25
Partición 2	Local	3	67.08 %	24
SVSL 0	Local	3	78.34 %	8
SVSL 1	Local	3	69.28 %	8
SVSL 2	Local	3	67.08 %	23
Final	**General**	3	**75.36** %	**39**

8.2.7. Caso *Mushroom*

Para la base de datos *Mushroom*, la RG obtenida considera la eliminación de 15 de las 22 variables presentes en la representación original, lo que equivale a una reducción del 68.18 % de ellas.

Este caso en específico, maneja una precisión de clasificación del 100 % en la representación original, misma que se conserva en la RG final.

Tabla 8.31: Desempeño del clasificador, caso *Mushroom*

Conjunto de datos	Clasificación	Clases	Desempeño	Variables
Original	**Inicial**	2	**100.00** %	**22**
Partición 0	Local	2	100.00 %	9
Partición 1	Local	2	100.00 %	13
SVSL 0	Local	2	100.00 %	2
SVSL 1	Local	2	100.00 %	5
Final	**General**	2	**100.00** %	**7**

8.2.8. Caso *Spambase*

En este ejemplo se ha obtenido una reducción final del 73.68 % de variables, dado que originalmente se tienen 57 y se incluyen al final 15 de ellas en la RG. lo que permite alcanzar un aumenta del 90.78 % al 99.63 % en la precisión del clasificador para la RG. Se observa que con la aplicación del método propuesto.

Tabla 8.32: Desempeño del clasificador, caso *Spambase*

Conjunto de datos	Clasificación	Clases	Desempeño	Variables
Original	**Inicial**	2	**90.78** %	**57**
Partición 0	Local	2	99.52 %	3
Partición 1	Local	2	99.33 %	54
SVSL 0	Local	2	99.76 %	2
SVSL 1	Local	2	99.54 %	13
Final	**General**	2	**99.63** %	**15**

8.2.9. Caso *Warp*

Para el caso de los resultados obtenidos en la experimentación con la base de datos *Warp*, también se obtienen los datos de la precisión del clasificador Ibk, tanto

para las particiones locales, como para los SVSL que se derivan de éstas, así como de la RG resultante.

Este ejemplo es el que cuenta de origen con el mayor número de variables, por lo que también se presentan por separado los resultados, éstos se han agrupado para las particiones, los SVSL y finalmente la RG. La Tabla 8.33 muestra la precisión del clasificador y el número de variables incluidas en todas las particiones locales.

Tabla 8.33: Desempeño del clasificador por partición, caso *Warp*

Conjunto de datos	Clasificación	Clases	Desempeño	Variables
Original	**Inicial**	10	**50.00**%	**2,400**
Partición 0	Local	10	84.62%	225
Partición 1	Local	10	87.69%	217
Partición 2	Local	10	85.38%	244
Partición 3	Local	10	86.15%	200
Partición 4	Local	10	89.23%	170
Partición 5	Local	10	84.62%	233
Partición 6	Local	10	87.69%	215
Partición 7	Local	10	90.00%	424
Partición 8	Local	10	93.08%	138
Partición 9	Local	10	85.38%	334

Con respecto a los resultados de desempeño y variables incluidas en cada SVSL, éstos se pueden observar en la Tabla 8.34. En la parte final se incluyen los valores correspondientes al desempeño y variables que conforman la RG respectiva.

Tabla 8.34: Desempeño del clasificador en SVSL, caso *Warp*

Conjunto de datos	Clasificación	Clases	Desempeño	Variables
SVSL0	Local	10	88.46%	106
SVSL1	Local	10	89.23%	161
SVSL2	Local	10	86.92%	186
SVSL3	Local	10	77.69%	15
SVSL4	Local	10	77.69%	18
SVSL5	Local	10	84.62%	43
SVSL6	Local	10	86.15%	62
SVSL7	Local	10	92.31%	373
SVSL8	Local	10	92.31%	79
SVSL9	Local	10	88.46%	202
RG	**General**	10	**90.00**%	**1,075**

Como se observa en las Tablas 8.33 y 8.34, la reducción de la dimensionalidad de la RG con respecto de la representación original es considerable, dado que de las 2,400 variables iniciales solamente se utilizan 1,075 en el conjunto de datos final, lo que equivale a un 55.21 % de eliminación.

Por parte de la precisión del clasificador incluido, el desempeño mejora en un 40.00 % ya que con la configuración de inicio se tiene un 50.00 % y con la RG se obtiene un 90.00 %.

8.2.10. Caso *Wine Quality W*

De manera análoga al resto de conjuntos de datos, con el ejemplo *Wine Quality W* también se ha obtenido la precisión del clasificador Ibk, tanto para las particiones locales, como para los SVSL que se derivan de éstas, así como de la RG resultante.

Este ejemplo incluye tan solo 11 variables y un número importante de instancias, la intención de procesar este ejemplo es mostrar la funcionalidad del método propuesto con conjuntos de datos de poca dimensionalidad. En este caso es posible mostrar en una misma tabla los resultados referidos al desempeño de clasificación tanto de las particiones, como de los SVSL y la RG. La Tabla 8.35 muestra dichos valores.

Tabla 8.35: Desempeño del clasificador por partición, caso *Wine Quality W*

Conjunto de datos	Clasificación	Clases	Desempeño	Variables
Original	**Inicial**	**7**	**65.39 %**	**11**
Partición 0	Local	7	99.94 %	1
Partición 1	Local	7	99.59 %	2
Partición 2	Local	7	100.00 %	1
Partición 3	Local	7	99.98 %	1
Partición 4	Local	7	100.00 %	1
Partición 5	Local	7	99.96 %	1
Partición 6	Local	7	94.51 %	4
SVSL 0	Local	7	99.94 %	1
SVSL 1	Local	7	99.59 %	2
SVSL 2	Local	7	100.00 %	1
SVSL 3	Local	7	99.98 %	1
SVSL 4	Local	7	100.00 %	1
SVSL 5	Local	7	99.96 %	1
SVSL 6	Local	7	94.51 %	4
RG	**General**	**7**	**93.96 %**	**5**

Como se puede observar en la Tabla 8.35, la eliminación de variables disminuye de 11 en el conjunto original de datos a 5 de ellas en la RG correspondiente, produciendo una disminución del 54.55 %.

Con respecto a la variación en la precisión del clasificador incluido, el desempeño mejora notablemente, ya que se observa un aumento del 28.56 % dado que antes del proceso de integración de variables, se tenía un 65.39 % de instancias correctamente clasificadas y con la RG se obtiene un 93.96 %.

8.2.11. Caso *Yale*

De manera análoga al resto de conjuntos de datos, con el ejemplo *Yale* también se ha obtenido la precisión del clasificador Ibk, tanto para las particiones locales, como para los SVSL que se derivan de éstas, así como de la RG resultante.

Este ejemplo incluye 1,024 variables, por ser una cantidad considerable también se presentan por separado los resultados de clasificación, éstos se han agrupado para las particiones, los SVSL y finalmente la RG. La Tabla 8.36 muestra la precisión del clasificador y el número de variables incluidas en todas las particiones locales.

Tabla 8.36: Desempeño del clasificador por partición, caso *Yale*

Conjunto de datos	Clasificación	Clases	Desempeño	Variables
Original	**Inicial**	**15**	**66.06 %**	**1,024**
Partición 0	Local	15	84.85 %	49
Partición 1	Local	15	75.76 %	87
Partición 2	Local	15	77.58 %	145
Partición 3	Local	15	83.03 %	45
Partición 4	Local	15	66.06 %	152
Partición 5	Local	15	82.42 %	43
Partición 6	Local	15	79.39 %	110
Partición 7	Local	15	83.64 %	36
Partición 8	Local	15	85.45 %	29
Partición 9	Local	15	82.42 %	51
Partición 10	Local	15	79.39 %	73
Partición 11	Local	15	78.18 %	81
Partición 12	Local	15	82.42 %	31
Partición 13	Local	15	70.91 %	51
Partición 14	Local	15	70.91 %	41

Por otra parte, los resultados de desempeño y variables incluidas en cada SVSL se muestran en la Tabla 8.37. En la parte final se incluyen los valores correspondientes al desempeño y variables que conforman la RG respectiva.

Tabla 8.37: Desempeño del clasificador en SVSL, caso *Yale*

Conjunto de datos	Clasificación	Clases	Desempeño	Variables
SVSL 0	Local	15	85.45 %	45
SVSL 1	Local	15	80 %	60
SVSL 2	Local	15	79.39 %	89
SVSL 3	Local	15	83.64 %	36
SVSL 4	Local	15	72.12 %	98
SVSL 5	Local	15	83.03 %	37
SVSL 6	Local	15	83.64 %	66
SVSL 7	Local	15	84.85 %	34
SVSL 8	Local	15	83.64 %	15
SVSL 9	Local	15	82.42 %	51
SVSL 10	Local	15	81.21 %	65
SVSL 11	Local	15	77.58 %	43
SVSL 12	Local	15	83.64 %	28
SVSL 13	Local	15	72.73 %	48
SVSL 14	Local	15	69.09 %	26
RG	**General**	**15**	**78.79 %**	**736**

Como se observa en las Tablas 8.36 y 8.37, también se presenta una eliminación importante en el número de variables presentes en la RG con respecto de la representación original, ya que de las 1,024 existentes en el conjunto inicial de datos solamente se conservan 736 de ellas, produciendo una disminución del 28.13 %.

Con respecto a la variación en la precisión del clasificador incluido, el desempeño mejora en un 12.73 % ya que antes de aplicar procesar la integración de variables, se tenía un 66.06 % de desempeño y con la RG se obtiene un 78.79 %.

8.3. Discusión de resultados

El método de estructuración global de poblaciones distribuidas, ha sido aplicado a diversos conjuntos de datos, presentando un promedio de 63.60 % de eliminación de variables y como se puede observar en la Tabla 8.38, en todos los casos se encuentra una mejora o conservación en el desempeño de clasificación de la RG final con respecto a los conjuntos de datos originales.

También se observa que de manera general, se ha obtenido un 13.83 % de mejora global en el desempeño de clasificación, pasando de 77.14 % en los conjuntos originales a 90.96 % en las RG obtenidas.

En el conjunto de datos *ITESA*, la eliminación de variables es significativa y el desempeño de clasificación mejora un 3.19 %. En este ejemplo, de las 48 variables de la tabla *Alumnos*, solo 11 se incluyeron en la RG; mientras que de *Autoridades* se consideraron 10 en el proceso de integración y únicamente 9 se agregaron a la RG final. La nueva clasificación contiene profesores en las cinco categorías posibles.

En el caso específico de la base de datos *Mushroom*, desde el inicio se obtiene una precisión de clasificación del 100.00 %, ésta se conserva hasta el final del proceso pero se elimina un total de 68.18 % de variables.

Tabla 8.38: Resultados finales de experimentación

Conjunto de datos	Desempeño del clasificador (Ibk)						Variables eliminadas
	Conjunto original			Nueva RG			
	Desempeño	Clases	Vars	Desempeño	Clases	Vars	
1.- Chess	96.28 %	2	36	100.00 %	2	5	86.11 %
2.- Isolet	83.56 %	26	617	85.32 %	26	327	47.00 %
3.- Itesa	84.04 %	5	66	87.23 %	5	20	69.70 %
4.- Lung discrete	83.56 %	7	325	93.15 %	7	132	59.38 %
5.- Madelon	54.15 %	2	500	97.15 %	2	12	97.60 %
6.- Molecular	74.67 %	3	60	75.36 %	3	39	35.00 %
7.- Mushroom	100.00 %	2	22	100.00 %	2	7	68.18 %
8.- Spambase	90.78 %	2	57	99.63 %	2	15	73.68 %
9.- Warp	50.00 %	10	2,400	90.00 %	10	1,075	55.21 %
10.- Wine Quality W	65.39 %	7	11	93.96 %	7	5	54.55 %
11.- Yale	66.06 %	15	1,024	78.79 %	15	736	28.13 %
Promedio	77.14 %			90.96 %			63.60 %

Los resultados alcanzados, permiten observar que el método de estructuración presentado, en todos los casos permite la integración de variables provenientes de distintas particiones, conformando una representación general de los objetos en estudio, con una nueva clasificación global correspondiente. Además de mejorar o conservar el desempeño del clasificador incluido.

Conclusiones y trabajo futuro

Es más sabio averiguar que suponer

Mark Twain

Conclusiones

En esta tesis se ha propuesto un nuevo método para la estructuración global de poblaciones distribuidas con técnicas de RP, demostrando su efectividad al reducir de manera importante el número de variables necesarias para obtener una adecuada representación global del problema original, ésta incluye una nueva clasificación general de todos los objetos en estudio.

El método propuesto resuelve un problema en el tratamiento de datos particionados, que como se abordó en el Capítulo 3, no se ha identificado algún trabajo, cuyo objetivo en particular pretenda construir una representación general de la población distribuida.

Un aporte específico en el ámbito de la SV consiste en la identificación de la combinación más adecuada de uno de los algoritmos selectores Filter univariados que fueron estudiados y un clasificador supervisado. Esta combinación, en términos generales provee altos desempeños de clasificación en conjuntos de datos pertenecientes a diversos contextos.

en el Capitulo 6 se propone un Criterio utilizado durante el proceso de evaluación de rankings, el cual logra identificar el momento en el que se ha obtenido el máximo desempeño de clasificación, deteniendo el testeo de más subconjuntos de variables, identificando así el óptimo para dicho selector-clasificador.

El criterio mencionado es capaz de modificar su configuración a través de un parámetro denominado *ventana*, lo que permite aumentar o disminuir la exactitud de la identificación de subconjuntos de variables, ésto en balance con el número de pruebas realizadas, evitando así la realización de búsquedas exhaustivas.

Por otra parte, se ha implementado una estrategia de integración de variables provenientes de diversas particiones locales, ésta evalúa la factibilidad de incluir nuevas variables en la representación general que se busca e incluye la generación de la nueva clasificación global correspondiente.

Adicionalmente, el método propuesto incluye una estrategia para preparar los conjuntos de datos si éstos no cumplen con la organización de población distribuida que se requiere. Dicha estrategia es capás de construir las particiones necesarias a fin de garantizar el correcto procesamiento del conjunto de datos original.

Por otra parte, en el Capítulo 8 se ha demostrado la efectividad del método propuesto, se han incluido diversos casos reales a fin de observar su comportamiento utilizando conjuntos de datos pertenecientes a diferentes contextos, en todos los casos se obtuvo una representación general con una disminución de variables y se ha mejorado la capacidad de clasificación.

El uso de diversas técnicas tanto de SV, como de CS y Agrupamiento, así como la noción de correlación de variables han permitido la construcción y validación del modelo presentado con éxito.

De manera específica, uno de los casos de estudio utilizados incluye a una institución como usuario final del modelo, para quién se ha encontrado una representación general que incluye puntos de vista de dos actores sobre un mismo grupo de individuos. Lo anterior ha permitido contar con un panorama más amplio de los objetos en estudio y obtener conocimiento más general de los mismos.

Trabajo futuro

Durante el desarrollo de la presente investigación se han identificado diversas áreas de oportunidad para profundizar en algunos aspectos del método desarrollado, a fin de explorar otras aplicaciones del mismo o bien mejorar su construcción interna.

Una de las áreas de interés se centra en considerar al nuevo método como un selector de variables, dado que ha mostrado efectividad al obtener un subconjunto de variables final que representa adecuadamente a una población. Lo anterior permitirá evaluar y contrastar su funcionamiento con relación a otros selectores, incluso se pueden considerar técnicas Wrapper que también sugieren subconjuntos de variables seleccionadas.

En el ámbito de la SV, en la literatura se reportan buenos resultados al implementar diversos métodos de búsqueda para aplicarse en rankings de variables. Lo anterior conforma otra área de oportunidad, dado que el método propuesto incluye una estrategia diseñada a partir de una búsqueda secuencial, ésta puede modificarse explorando alternativas aleatorias o de otro tipo, lo cuál pudiera reducir el tiempo de procesamiento general.

Por otra parte, al incluir el agrupador *K-means* aparte de su probada efectividad se busca también la eficiencia del método. sin embargo, es posible explorar con agrupadores que se reporten como más ágiles o bien con mayor eficacia. La experimentación en esta área puede redundar en el enriquecimiento del método al mejorar la clasificación de nuevas instancias.

Con respecto a la validación del nuevo método se puede explorar la utilización de otros conceptos teóricos enfocados a la conformación de clusters desde el punto de vista tanto interno como externo, lo que podría sugerir la utilización de diversos agrupadores en lugar incluir únicamente a *K-means*.

En la parte de la aplicación del método, es posible buscar aplicaciones específicas en otros ámbitos, como puede ser la búsqueda de una clasificación general de aquellos clientes de interés para instituciones bancarias pero que se encuentran representados en múltiples repositorios. De igual manera el nuevo método tiene cabida en entornos empresariales tanto administrativos como operativos, en líneas de producción, en contextos gubernamentales, etc. En todos los casos es valioso contar con reportes concretos de los resultados que se pudieran obtener.

Referencias

Y así, del mucho leer y del poco dormir, se le secó el celebro de manera que vino a perder el juicio.

Miguel de Cervantes Saavedra

AHA, D. y KIBLER, D. Instance-based learning algorithms. *Machine Learning*, vol. 6, páginas 37–66, 1991.

ALMUALLIM, H. y DIETTERICH, T. G. Learning with many irrelevant features. En *Proceedings of the ninth national conference on artificial intelligence (AAAI-91)*, páginas 547 – 552. AAAI Press, 1991.

ARIZONA STATE UNIVERSITY, A. Feature selection. `http://featureselection.asu.edu/datasets.php`, 2016.

BATTITI, R. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, vol. 5(4), páginas 537–550, 1994.

BELKIN, M. y NIYOGI, P. Laplacian eigenmaps and spectral techniques for embedding and clustering. En *Advances in Neural Information Processing Systems* (editado por T. G. Dietterich, S. Becker y Z. Ghahramani). Chicago, University of, 2001.

BOLÓN-CANEDO, V., SÁNCHEZ-MAROÑO, N. y ALONSO-BETANZOS, A. Distributed feature selection: An application to microarray data classification. *Applied Soft Computing*, vol. 30, páginas 136 – 150, 2015. ISSN 1568-4946.

BOLÓN-CANEDO, V., SÁNCHEZ-MAROÑO, N. y ALONSO-BETANZOS, A. *Feature selection for high-dimensional data*. Springer, 2105a.

BOLÓN-CANEDO, V., SÁNCHEZ-MAROÑO, N. y ALONSO-BETANZOS, A. *Feature selection for high-dimensional data*, capítulo A Critical Review of Feature Selection Methods, páginas 29–60. Springer, 2105b.

BOLÓN-CANEDO, V., SÁNCHEZ-MAROÑO, N. y CERVIÑO-RABUÑAL, J. Scaling up feature selection: A distributed filter approach. En *15th Conference of the Spanish Association for Artificial Intelligence, CAEPIA, Lecture Notes in Computer Science* (editado por B. et al.), páginas 121 – 130. Springer-Verlag Berlin Heidelberg, 2013.

BOLÓN-CANEDO, V., SÁNCHEZ-MAROÑO, N. y CERVIÑO-RABUÑAL, J. Toward parallel feature selection from vertically partitioned data. En *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. ESANN, 2014.

BOSER, B. E., GUYON, I. M. y VAPNIK, V. N. A training algorithm for optimal margin classifiers. En *Proceedings of the Fifth Annual Workshop on Computational Learning Theory COLT 92* (editado por D. Haussler). ACM Digital Library, 1992.

BREIMAN, L. Random forests. *Machine Learning*, vol. 45(1), páginas 5–32, 2001.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. y STONE, C. J. *Classification and regression trees*. Wadsworth, 1984.

BROOMHEAD, D. S. y LOWE, D. Radial basis functions, multi-variable functional interpolation and adaptative networks. Informe técnico, Royal signals and Radar Establishment, 1988.

CAI, X., NIE, F. y HUANG, H. Exact top-k feature selection via l2, 0 -norm constraint. En *International Conference on Artificial Intelligence*, páginas 1240 – 1246. Springer, 2013.

CARUANA, R. y FREITAG, D. Greedy attribute selection. En *Machine Learning Proceedings 1994* (editado por W. W. Cohen y H. Hirsh), páginas 28 – 36. Morgan Kaufmann, 1994. ISBN 978-1-55860-335-6.

CHANDRASHEKAR, G. y SAHIN, F. A survey on feature selection methods. *Computers and Electrical Engineering*, páginas 16–28, 2014.

CHEBYCHEV, P. Des valeurs moyennes. *Journal de mathematiques pures et appliquees*, vol. 2(12), páginas 177 – 184, 1867.

CHEN, X.-W. y WASIKOWSKI, M. Fast: A roc-based feature selection metric for small samples and imbalanced data classification problems. En *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, páginas 124–132. ACM, New York, NY, USA, 2008. ISBN 978-1-60558-193-4.

CHENG, Q., VARSHNEY, P. K. y ARORA, M. K. Logistic regression for feature selection and soft classification of remote sensing data. *Letters on Geoscience and Remote Sensing*, vol. 3(4), páginas 491–494, 2006.

CHO, K. Mawi working group traffic archive. `http://mawi.wide.ad.jp/mawi/`, 2016.

CLEARY, J. G. y TRIGG, L. E. K*: An instance-based learner using an entropic distance measure. En *12th International Conference on Machine Learning*, páginas 108 – 114. Waikato, University of, 1995.

CORTES, C. y VAPNIK, V. Support vector networks. *Machine Learning*, vol. 20(3), páginas 273–297, 1995.

CZARNOWSKI, I. Distributed learning with data reduction. En *Transactions on CCI IV*, páginas 3 – 121. Springer-Verlag Berlin Heidelberg, 2011.

DASH, M. y LIU, H. Consistency-based search in feature selection. *Artificial Intelligence*, vol. 151(1), páginas 155–176, 2003.

DASH, M., LIU, H. y MOTODA, H. Consistency based feature selection. En *Knowledge Discovery and Data Mining, Current issues and New Applications* (editado por T. Terano, H. Liu y A. L. P. Chen), páginas 98–109. Springer, 2000.

DAVIS, J. C. *Statistics and Data Analysis in Geology*. John Wiley & Sons, Inc, 2nd edición, 1990.

DESTRERO, A., DE MOL, C., ODONE, F. y VERRI, A. A regularized approach to feature selection for face detection. En *8th Asian Conference on Computer Vision*, páginas 881 – 890. None, 2007.

DUDA, R. O., HART, P. E. y STORK, D. G. *Pattern Classification*. John Wiley & Sons, 2001.

DUNHAM, M. H. *Data Mining, Introductory and advanced topics*. Prentice Hall, 2003.

DUNNING, T. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, vol. 19(1), páginas 61–74, 1993.

EBRAHIMPOUR, M. K. y EFTEKHARI, M. Ensemble of feature selection methods: A hesitant fuzzy sets approach. *Applied Soft Computing*, vol. 50, páginas 300 – 312, 2017. ISSN 1568-4946.

EBRAHIMPOUR, M. K. y EFTEKHARI, M. Distributed feature selection: A hesitant fuzzy correlation concept for microarray high-dimensional datasets. *Chemometrics and Intelligent Laboratory Systems*, vol. 173, páginas 51 – 64, 2018. ISSN 0169-7439.

EBRAHIMPOUR, M. K., ZARE, M., EFTEKHARI, M. y AGHAMOLAEI, G. Occam's razor in dimension reduction: Using reduced row echelon form for finding linear independent features in high dimensional microarray datasets. *Engineering Applications of Artificial Intelligence*, vol. 62, páginas 214 – 221, 2017. ISSN 0952-1976.

ESCARCEGA, D., RAMOS, F., ESPINOSA, A. y BERUMEN, J. A hybrid methodolgy for pattern recognition in signaling cervical cancer pathways. En *Advances in Pattern Recognition, MCPR 2010, Lecture Notes in Computer Science*, páginas 301 – 310. Springer-Verlag Berlin Heidelberg, 2010.

ESHELMAN, L. J. The chc adaptive search algorithm; how to have safe search when engaging in nontraditional genetic recombination. En *Foundations of Genetic Algorithms.* (editado por R. GJE), páginas 265–283. Morgan Kauffman, 1st edición, 1991.

FAN, R.-E., CHANG, K.-W., HSIEH, C.-J., WANG, X.-R. y LIN, C.-J. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, vol. 9, páginas 1871–1874, 2008.

FIX, E. y HODGES, J. L. J. Discriminatory analysis nonparametric discrimination consistency properties. Informe técnico, Project number 21-49-004, 1951.

FRANK, E., WITTEN, I. y HALL, M. *Data Mining, Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, 3rd edición, 2011.

FRANK, E. y WITTEN, I. H. Generating accurate rule sets without global optimization. En *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, páginas 144–151. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998. ISBN 1-55860-556-8.

FREUND, Y. y SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, vol. 55((SS971504)), páginas 119–139, 1997.

GHARROUDI, O., ELGHAZEL, H. y AUSSEM, A. A comparison of multi-label feature selection methods using the random forest paradigm. En *Canadian Conference on Artificial Intelligence, LNAI* (editado por P. Sokolova, M; Van Beek), páginas 95 – 106. Springer International Publishing Switzerland, 2014.

GINI, C. *Variabilita u Mutabilita, Contributo allo Studio delle Distribuzioni e delle Relaziono Statistiche*. Tipografia di Paolo Cuppin, 1912.

GNANA SINGH, D. A. A., FERNANDO, A. E. y LEAVLINE, E. J. Experimental study on feature selection methods for software fault detection. En *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, páginas 1–6. 2016.

GOWDA, A., MANJUNATH, A. S. y JAYARAM, M. A. Comparative study of attribute selection using gain ratio and correlated based feature selection. *International journal of information technology and knowledge management*, vol. 2(2), páginas 271–277, 2010.

GUI, J., SUN, Z., JI, S., TAO, D. y TAN, T. Feature selection based on structured sparsity: A comprehensive study. *IEEE Transactions on Neural Networks and Learning Systems*, vol. pp(99), páginas 1 – 18, 2016.

GUYON, I. y ELISSEEFF, A. An introduction to variable and feature selection. *Journal of Machine Learning Research*, vol. 3, páginas 1157–1182, 2003.

HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTMANN, P. y WITTEN, I. H. The weka data mining software: An update. *SIGKDD Explorations*, vol. 11(1), 2009.

HALL, M. A. *Correlation-based Feature Selection for Machine Learning*. Waikato, University of, 1999.

HARMAN, H. H. *Modern Factor Analysis*. University of Chicago, 1st edición, 1960.

HAYKIN, S. *Neural networks: a comprehensive foundation*. Prentice Hall, 2nd edición, 1998.

HE, R., TAN, T., WANG, L. y ZHENG, W. S. l2, 1 regularized correntropy for robust feature selection. En *Conference on Computer Vision and Pattern Recognition 2012*, páginas 2504 – 2511. Institute of Electrical and Electronics Engineers, 2012.

HE, X., CAI, D. y NIYOGI, P. Laplacian score for feature selection. En *Proceedings of the 18th International Conference on Neural Information Processing Systems*, NIPS'05, páginas 507–514. MIT Press, Cambridge, MA, USA, 2005.

HERNANDEZ ORALLO, J., RAMÍREZ QUINTANA, M. J. y FERRI RAMÍREZ, C. *Introduccion a la Mineria de Datos*. Madrid: Pearson Educacion S. A, 2005.

HERNÁNDEZ-VALADEZ, E. *Algoritmo de clustering basado en entropía para descubrir grupos en atributos de tipo mixto*. Proyecto Fin de Carrera, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, México, DF, 2006. Tesis de Maestría en Ciencias en la Especialidad Ingeniería Eléctrica, opción Computación, desarrollada en el Departamento de Ingeniería Eléctríca sección Computación.

HOLLANDER, M. y WOLFE, D. A. *Nonparametric statistical methods*. John Wiley & Sons, Inc, 1973.

HOLTE, R. C. Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, vol. 11, páginas 63–91, 1993.

HU, Q., YU, D., LIU, J. y WU, C. Neighborhood rough set based feature subset selection. *Information Sciences*, vol. 178(18), páginas 3577–3594, 2008.

IMDB. Internet movie database. `http://www.imdb.com`, 2012.

JAYNES, E. T. Information theory and statistical mechanics. *The physical review*, vol. 106(4), páginas 620–630, 1957.

JIN, X., XU, A., BIE, R. y GUO, P. Machine learning techniques and chi-squared feature selection for cancer classification using sage gene expression profiles. *Lecture Notes in Bioinformatics in BioDM*, vol. 3916, páginas 106–115, 2006.

JINAN, U. Computational intelligence group at jinan university. `http://cilab.ujn.edu.cn/datasets.htm`, 2002.

JOHN, G. H., KOHAVI, R. y PFLEGER, K. Irrelevant features and the subset selection problem. En *Machine Learning: Proceedings of the Eleventh International Conference*, páginas 121 – 129. Morgan Kaufmann, 1994.

KELLEY, H. J. Gradient theory of optimal flight paths. *Aerospace Research Central Journal*, vol. 30(10), páginas 947–954, 1960.

KIRA, K. y RENDELL, L. A. The feature selection problem: Traditional methods and a new algorithm. En *Tenth national conference on artificial intelligence*, páginas 129–134. MIT Press, 1992.

KITTLER, J. Feature set search algorithms, in:. En *Pattern Recognition and Signal Processing* (editado por C. Chen), páginas 41–60. Sijtho and Noordho, 1978.

KONONENKO, I. Estimating attributes: Analysis and extensions of relief. En *European Conference on Machine Learning* (editado por F. Bergadano y L. d. Raedt), páginas 171 – 182. Springer, 1994.

KUMARI-BHARTI, K. y KUMAR-SINGH, P. A survey on filter techniques for feature selection in text mining. *Advances in Intelligent Systems and Computing*, páginas 1545–1559, 2014.

LAZAR, C., TAMINAU, J., MEGANCK, S., STEENHOFF, D., COLETTA, A., MOLTER, C., DE SCHAETZEN, V., DUQUE, R., BERSINI, H. y NOWÉ, A. A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Transactions on Computational Biology and Informatics*, vol. 9(4), páginas 1106–1119, 2012.

LI, J., CHENG, K., WANG, S., MORSTATTER, F., TREVINO, R. P., TANG, J. y LIU, H. Feature selection: A data perspective. *CoRR*, vol. abs/1601.07996, 2016.

LI, Z., LU, W., SUN, Z. y XING, W. A parallel feature selection method study for text classification. *Neural Computing and Applications*, vol. 28(suppl 1), páginas 513–524, 2017.

LICHMAN, M. Uci machine learning repository. `https://archive.ics.uci.edu/ml/index.html`, 2009.

LIKERT, R. A technique for the measurement of attitudes. *Archives of psychology*, vol. 22(140), páginas 1–55, 1932.

LITTLESTONE, N. Learning quickly when irrelevant attributes abound, a new linearthreshold algorithm. *Machine Learning*, vol. 2, páginas 285–318, 1988.

LIU, H. y MOTODA, H. *Computational methods of feature selection*. Chapman & Hall/CRC, Taylor & Francis Group, 2008.

LIU, H., WU, X. y ZHANG, S. A new supervised feature selection method for pattern classification. *Computational Intelligence*, páginas 342–361, 2014.

LIU, J., LIN, M., WU, S. y ZHANG, J. Feature selection based on quality of information. *Neurocomputing*, páginas 11–22, 2017.

LIU, M. y ZHANG, D. Sparsity score: A novel graph-preserving feature selection method. *International Journal of Pattern Recognition and Artificial Intelligence*, páginas 1–29, 2014.

LIU, M. y ZHANG, D. Feature selection with effective distance. *Neurocomputing*, páginas 100–109, 2016.

LLOYD, S. P. Least squares quantization in pcm. Informe técnico, Mathematical Research Department at Bell Laboratories, 1957.

LLOYD, S. P. Least squares quantization in pcm. *IEEE Transactions on information theory*, vol. IT-28(2), páginas 129–137, 1982.

LOPEZ-ESCOBAR, S. *Algoritmos de Agrupamiento Global para Datos Mezclados*. Proyecto Fin de Carrera, Instituto Nacional de Astrofísica, Óptica y Electronica, Santa María Tonantzintla, Puebla, 2007. Tesis de Maestría en Ciencias con Especialidad en Ciencias Computacionales.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. En *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1:Statistics, páginas 281–297. University of California Press, Berkeley, Ca, 1967.

MERA, D., BOLON-CANEDO, V., COTOS, J. y ALONSO-BETANZOS, A. On the use of feature selection to improve the detection of sea oil spills in sar images. *Computers & Geosciences*, vol. 100, páginas 166 – 178, 2017. ISSN 0098-3004.

MORÁN-FERNÁNDEZ, L., BOLON-CANEDO, V. y ALONSO-BETANZOS, A. A time efficient approach for distributed feature selection partitioning by features. En *Advances in Artificial Intelligence - 16th Conference of the Spanish Association for Artificial Intelligence, CAEPIA 2015, Proceedings* (editado por J. M. Puerta, J. A. Gámez, B. Dorronsoro, E. Barrenechea, A. Troncoso, B. Baruque y M. Galar), vol. 9422 de *Lecture Notes in Computer Science*, páginas 245 – 254. Springer, 2015. ISBN 978-3-319-24597-3.

MORÁN-FERNÁNDEZ, L., BOLON-CANEDO, V. y ALONSO-BETANZOS, A. Centralized vs. distributed feature selection methods based on data complexity measures. *Knowledge-Based Systems*, páginas 27–45, 2017.

NIE, F., HUANG, H., CAI, X. y DING, C. Efficient and robust feature selection via join l2, 2 -norms minimization. En *Neural Information Processing Systems* (editado por J. D. Lafferty, C. K. I. Williams, J. Shawe Taylor, R. S. Zemel y A. Culotta), páginas 1813 – 1821. None, 2010.

NIU, B., FAN, Y., WANG, H., LI, L. y WANG, X. Novel bacterial foraging optimization with time-varying chemotaxis step. *International Journal of Artificial Intelligence*, páginas 257–273, 2011.

NOVOVICOVA, J., SOMOL, P., HAINDL, M. y PUDIL, P. Conditional mutual information based feature selection for classification task. En *12th Iberoamerican Conference on Pattern Recognition, Image Analysis and Applications*, páginas 417 – 426. Springer-Verlag, 2007.

ORACLE-CORPORATION. Java virtual machine. `https://www.java.com/es/`, 2017.

PASSINO, K. M. Biomimicry of bacterial foraging for distributed optimization and control. *IEEE Control Systems Magazine*, páginas 52–67, 2002.

PENG, H., LONG, F. y DING, C. Feature selection based on mutual information, criteria of max-dependency, max-relevance and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27(8), páginas 1226–1238, 2005.

PLATT, J. C. Sequential minimal optimization: A fast algorithm for training support vector machines. Tech report, Microsoft, 1998.

PONSA, D. y LOPEZ, M. Feature selection based on a new formulation of the minimal redundancy maximal relevance criterion. En *Proceedings for the Third Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 2007.

PRASAD, B. R., BENDALE, U. K. y AGARWAL, S. Distributed feature selection using vertical partitioning for high dimensional data. En *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, páginas 807–813. 2016.

PRESS, W. H., TEUKOLSKY, S. A., VETTERLING, W. T. y FLANNERY, B. P. *Numerical Recipes in C*. Cambridge: Press Syndicate of the University of Cambridge, 1st edición, 1988.

PUDIL, P., NOVOVICOVA, J. y KITTLER, J. Floating search methods in feature selection. *Pattern Recognition Letters*, vol. 15(11), páginas 119–1125, 1994.

QUINLAN, J. R. Induction of decision trees. *Machine Learning*, vol. 1(1), páginas 81–106, 1986.

QUINLAN, J. R. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc, 1993.

RIDGE, K. Bio-medical dataset repository. http://datam.i2r.a-star.edu.sg/datasets/krbd/, 2016.

RUIZ-SANCHEZ, R. selecion de atributos mediante proyecciones. tesis. Informe técnico, Sevilla, University of, 2005.

MARQUES DE SA, J. P. *Pattern Recognition, Concepts, Methods and Applications*. Springer, Oporto, Portugal, 2001.

SAEYS, Y., INZA, I. y LARRANAGA, P. A review of feature selection techniques in bioinformatics. *Bioinformatics*, vol. 23(19), páginas 2507–2517, 2007.

SAMARIA, F. S. y HARTER, A. C. Parameterisation of a stochastic model for human face identification. En *Proceedings of 1994 IEEE Workshop on Applications of Computer Vision*, páginas 138 – 142. IEEE Conference Publications, 1994.

SHANNON, C. E. A mathematical theory of communication. *The bell system technical journal*, páginas 379–423, 623–656, 1948.

SHARMA, A. y DEY, S. A comparative study of feature selection and machine learning techniques for sentiment analysis. En *Research in Applied. Computation Symposium RACS'12*. ACM, 2012.

SINGH, R., KUMAR, H. y SINGLA, R. K. Analysis of feature selection techniques for network traffic dataset. En *International Conference on Machine Intelligence and Research Advancement* (editado por I. C. Society), páginas 42 – 46. IEEE, 2013.

SINGH-RATHORE, S. y GUPTA, A. A comparative study of feature-ranking and feature-subset selection techniques for improved fault prediction. En *ISEC '14*. ACM, 2014.

SOLORIO-FERNANDEZ, S., CARRASCO-OCHOA, J. A. y MARTÍNEZ TRINIDAD, J. F. Hybrid feature selection method for supervised classification based on laplacian score ranking. En *Mexican Conference on Pattern Recognition*, páginas 260 – 269. Springer-Verlag Berlin Heidelberg, 2010a.

SOLORIO-FERNANDEZ, S., CARRASCO-OCHOA, J. A. y MARTÍNEZ TRINIDAD, J. F. Seleccion de variables para clasificacion no supervisada utilizando un enfoque hibrido filter-wrapper, tesis. Informe técnico, Instituto Nacional de Astrofisica, Optica y Electronica, 2010b.

STATNIKOV, A., ALIFERIS, C. F. y TSAMARDINOS, I. Gene expression model selector gems. http://www.gems-systems.org, 2003.

TAN, P.-N., STEINBACH, M. y KUMAR, V. *Introduction to Data Mining*. Addison Wesley, 2005.

TITTERINGTON, D. M., MURRAY, G. D., MURRAY, L. S., SPIEGELHALTER, D. J., SKENE, A. M., HABBEMA, J. D. F. y GELPKE, G. J. Comparison of discrimination techniques applied to a complex dataset of head injured. *Journal of the Royal Statistical Society. Series A*, vol. 144(2), páginas 145–175, 1981.

TOMAR, D. y AGARWAL, S. Hybrid feature selection based weighted least squares twin support vector machina approach for diagnosing breast cancer, hepatitis and diabetes. *Advances in Artificial Neural Systems*, vol. 2015, páginas 1–10, 2015.

TYAGI, V. y MISHRA, A. A survey on different feature selection methods for microarray data analysis. *International Journal of Computer Applications*, vol. 67(16), páginas 36 – 40, 2013.

VAPNIK, V. *Statistical learning theory*. Wiley, 1998.

VAPNIK, V. y LERNER, A. Pattern recognition using generalized portrait method. *Automation and remote control*, vol. 24(6), páginas 774 – 780, 1963.

VAPNIK, V. N. *The Nature of Statistical Learning Theory*. Springer-Verlag, Berlin, Heidelberg, 1995. ISBN 0-387-94559-8.

WALD, R., KHOSHGOFTAAR, T., ABU SHANAB, A. y NAPOLITANO, A. Comparative analysis on the stability of feature selection techniques using three frameworks on biological datasets. En *2013 12th International Conference on Machine Learning and Applications*, vol. 1, páginas 418–423. 2013.

WAND, N. R. G. Wits: Waikato internet traffic storage. http://wand.net.nz/wits/, 2016.

WANG, H., JING, X. y NIU, B. A weighted bacterial colony optimization for feature selection. En *Intelligent Computing in Bioinformatics ICIC 2014, Lecture Notes in Computer Science*, páginas 379 – 389. Springer International Publishing Switzerland, 2014.

WANG, H. y NIU, B. A novel bacterial algorithm with randomness control for feature selection in classification. *Neurocomputing*, páginas 176 – 186, 2017.

WEBB, A. R. *Statistical Pattern Recognition*. John Wiley & Sons Inc, 2nd edición, 2002.

WITTEN, I. y FRANK, E. *Data Mining, Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, 2nd edición, 2005.

XIANG, S., NIE, F., MENG, G., PAN, C. y ZHANG, C. Discriminative least squares regression for multiclass classificaction and feature selection. *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23(11), páginas 1738–1754, 2012.

YALE, U. Yalefaces. `http://cvc.cs.yale.edu/cvc/projects/yalefaces/yalefaces.html`, 1997.

YANG, Y. y PEDERSEN, J. O. A comparative study on feature selection in text categorization. En *Proceedings of the International Conference on Machine Learning* (editado por D. Fisher), páginas 412 – 420. Morgan Kaufmann Publishers Inc., 1997.

YANG, Y., SHEN, H. T., MA, Z., HUANG, Z. y ZHOU, X. l2,1 -norm regularized discriminative feature selection for unsupervised learning. En *International Conference on Artificial Intelligence*, páginas 1589 – 1594. Springer, 2011.

YU, L. y LIU, H. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, páginas 1205–1224, 2004.

ZENG, Z., ZHANG, H., ZHANG, R. y YIN, C. A novel feature selection method considering feature interaction. *Pattern Recognition*, páginas 2656–2666, 2015.

ZHAO, L., CHEN, Z., HU, Y., MIN, G. y JIANG, Z. Distributed feature selection for efficient economic big data analysis. *IEEE Transactions on Big Data*, vol. 4(2), páginas 164–176, 2018. ISSN 2332-7790.

ZHAO, Z. y LIU, H. Searching for interacting features in subset selection. En *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence*, páginas 1156–1161. ", 2007a.

ZHAO, Z. y LIU, H. Spectral feature selection for supevised and unsupervised learning. En *International Conference on Machine Learning (ICML '07)* (editado por G. Zoubin), páginas 1151–1157. ACM, 2007b.

ZHAO, Z., WANG, L., LIU, H. y YE, J. On similarity preserving feature selection. *IEEE Transactions on Knowledge and Data Engineering*, páginas 619–632, 2013a.

ZHAO, Z., ZHANG, R., COX, J., DULING, D. y SARLE, W. Massively parallel feature selection: an approach based on variance preservation. *Mach learning*, páginas 195–220, 2013b.

Apéndices

Apéndice A

Máximo desempeño de todos los selectores-clasificadores

Aprender sin reflexionar es malgastar la energía.

Confucio

En este apéndice se presentan diversas tablas cuyo contenido corresponde al resumen de los máximos desempeños identificados para cada combinación selector - clasificador, en cada conjunto de datos incluido en la Tabla 5.1. Este desempeño se refiere al total de instancias presentes en sus correspondientes representaciones originales y que fueron correctamente clasificadas por los métodos supervisados incluidos.

En cada caso se ha resaltado en un tono gris el mejor desempeño de cada clasificador, facilitando la identificación del selector con el que se obtuvo dicho valor. Adicionalmente, se destaca en color amarillo la mejor puntuación obtenida en cada conjunto de datos.

Los selectores se presentan ordenados alfabéticamente, mientras que los clasificadores no siguen este orden. Para aquellos casos en los que algún selector y/o clasificador no pudieron procesar los datos, se muestran valores cero en las filas o columnas correspondientes.

Tabla A.1: Máximo desempeño selector - clasificador, conjunto de datos *Abalone*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	26.119	26.335	27.316	24.707	23.127	26.885
Gain Ratio	26.071	26.335	26.550	24.707	23.127	26.000
Info Gain	26.071	26.335	26.550	24.707	23.127	26.957
Laplacian Score	25.257	25.257	26.957	21.858	21.068	26.526
One R	26.071	26.335	26.550	25.210	24.874	26.790
Relief	25.784	26.335	26.957	25.210	24.874	26.670
SVM	0.000	0.000	0.000	0.000	0.000	0.000
Symmetrical Uncertainty	26.071	26.335	26.263	24.707	23.127	26.622
Máximo por método	26.119	26.335	27.316	25.210	24.874	26.957
Máximo global	**27.316**					

Tabla A.2: Máximo desempeño selector - clasificador, conjunto de datos *Adults*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	84.962	83.852	85.820	86.194	84.780	86.772
Gain Ratio	84.962	83.262	85.818	86.194	85.517	86.772
Info Gain	84.962	83.852	85.820	86.194	82.959	86.772
Laplacian Score	0.000	0.000	0.000	0.000	0.000	0.000
One R	84.962	83.262	86.080	86.140	85.494	86.778
Relief	84.962	83.999	85.709	86.143	82.568	86.677
SVM	0.000	0.000	0.000	0.000	0.000	0.000
Symmetrical Uncertainty	84.962	83.262	85.818	86.194	85.699	86.772
Máximo por método	84.962	83.999	86.080	86.194	85.699	86.778
Máximo global	**86.778**					

Tabla A.3: Máximo desempeño selector - clasificador, conjunto de datos *Cylinder Bands*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	76.438	72.727	80.148	64.750	77.922	64.935
Gain Ratio	76.438	72.171	80.148	64.750	76.252	65.306
Info Gain	76.438	72.356	80.148	64.750	77.922	64.935
Laplacian Score	0.000	0.000	0.000	0.000	0.000	0.000
One R	75.696	72.171	81.076	77.365	76.067	70.130
Relief	76.623	72.171	80.891	71.429	78.479	71.985
SVM	0.000	0.000	0.000	0.000	0.000	0.000
Symmetrical Uncertainty	76.994	72.171	80.148	64.750	77.922	64.750
Máximo por método	76.994	72.727	81.076	77.365	78.479	71.985
Máximo global	**81.076**					

Tabla A.4: Máximo desempeño selector - clasificador, conjunto de datos *Bank*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	89.294	89.538	90.002	89.759	88.476	89.958
Gain Ratio	89.294	89.294	90.002	89.759	88.476	89.914
Info Gain	89.294	89.538	90.046	89.626	88.476	90.091
Laplacian Score	89.294	88.476	88.963	89.272	88.476	89.317
One R	89.294	89.294	89.051	89.383	89.294	89.737
Relief	89.294	90.091	89.317	89.604	88.476	89.803
SVM	0.000	0.000	0.000	0.000	0.000	0.000
Symmetrical Uncertainty	89.294	89.162	90.002	89.759	88.476	89.914
Máximo por método	89.294	90.091	90.046	89.759	89.294	90.091
Máximo global	**90.091**					

Tabla A.5: Máximo desempeño selector - clasificador, conjunto de datos *Bank Full*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	89.286	89.279	90.033	90.551	88.974	90.409
Gain Ratio	89.286	89.286	90.095	90.387	89.286	90.390
Info Gain	89.286	89.270	90.135	90.423	88.974	90.401
Laplacian Score	0.000	0.000	0.000	0.000	0.000	0.000
One R	89.286	89.562	90.042	90.451	89.286	90.476
Relief	89.286	89.576	89.792	90.421	88.302	90.356
SVM	0.000	0.000	0.000	0.000	0.000	0.000
Symmetrical Uncertainty	89.286	89.195	90.095	90.387	88.974	90.401
Máximo por método	89.286	89.576	90.135	90.551	89.286	90.476
Máximo global	**90.551**					

Tabla A.6: Máximo desempeño selector - clasificador, conjunto de datos *Breast Cancer*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	70.629	75.175	75.175	75.525	75.175	75.525
Gain Ratio	72.378	73.776	75.175	75.525	75.175	75.525
Info Gain	70.280	75.175	74.825	75.525	73.077	75.525
Laplacian Score	70.280	73.430	73.430	75.520	72.380	75.520
One R	72.378	74.476	75.175	75.874	74.126	75.874
Relief	71.329	73.077	73.427	75.525	74.476	75.525
SVM	0.000	0.000	0.000	0.000	0.000	0.000
Symmetrical Uncertainty	70.629	75.175	75.175	75.525	75.175	75.525
Máximo por método	72.378	75.175	75.175	75.874	75.175	75.874
Máximo global	**75.874**					

Tabla A.7: Máximo desempeño selector - clasificador, conjunto de datos *Car Evolution*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	93.750	85.532	91.898	93.229	94.213	93.229
Gain Ratio	93.750	85.532	91.898	93.229	94.213	93.229
Info Gain	93.750	85.532	91.898	93.229	94.213	93.229
Laplacian Score	93.750	85.530	87.560	92.360	93.520	92.360
One R	93.750	85.532	87.558	92.361	93.519	92.361
Relief	93.750	85.532	91.898	93.229	94.213	93.229
SVM	0.000	0.000	0.000	0.000	0.000	0.000
Symmetrical Uncertainty	93.750	85.532	91.898	93.229	94.213	93.229
Máximo por método	93.75	85.532	91.898	93.229	94.213	93.229
Máximo global	**94.213**					

Tabla A.8: Máximo desempeño selector - clasificador, conjunto de datos *Chess*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	95.995	90.426	97.028	99.374	96.621	99.437
Gain Ratio	95.932	94.337	97.059	99.374	96.371	99.437
Info Gain	95.995	90.426	97.028	99.374	96.621	99.437
Laplacian Score	95.960	87.890	97.060	99.370	96.530	99.440
One R	95.651	90.426	97.622	99.374	97.215	99.437
Relief	95.87	94.431	98.655	99.374	98.248	99.437
SVM	0.000	0.000	0.000	0.000	0.000	0.000
Symmetrical Uncertainty	95.995	94.086	97.028	99.374	96.558	99.437
Máximo por método	95.995	94.431	98.655	99.374	98.248	99.440
Máximo global	**99.440**					

Tabla A.9: Máximo desempeño selector - clasificador, conjunto de datos *Congressional Voting Records*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	96.092	95.632	95.632	96.322	95.632	96.322
Gain Ratio	96.092	95.632	95.632	96.322	95.862	96.322
Info Gain	96.092	95.632	95.632	96.322	95.862	96.322
Laplacian Score	0.000	0.000	0.000	0.000	0.000	0.000
One R	96.092	95.632	95.632	96.322	95.632	96.322
Relief	96.092	95.632	95.862	96.322	96.552	96.322
SVM	96.322	96.322	96.322	96.322	96.322	96.322
Symmetrical Uncertainty	96.092	95.632	95.632	96.322	95.862	96.322
Máximo por método	96.322	96.322	96.322	96.322	96.552	96.322
Máximo global	**96.552**					

Tabla A.10: Máximo desempeño selector - clasificador, conjunto de datos *Default Credit Card Clients*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	80.937	81.960	80.993	82.223	81.983	82.123
Gain Ratio	80.937	80.200	80.993	82.223	81.66	82.123
Info Gain	80.937	81.960	80.993	82.223	81.983	82.123
Laplacian Score	0.000	0.000	0.000	0.000	0.000	0.000
One R	80.933	81.960	81.290	82.223	81.983	82.123
Relief	80.933	81.810	80.250	82.087	77.880	82.133
SVM	80.927	81.960	80.923	81.973	81.983	82.013
Symmetrical Uncertainty	80.937	81.960	80.993	82.223	81.973	82.123
Máximo por método	80.937	81.960	81.290	82.223	81.983	82.133
Máximo global	**82.223**					

Tabla A.11: Máximo desempeño selector - clasificador, conjunto de datos *Dermatology*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	98.087	98.361	98.361	95.628	97.814	94.536
Gain Ratio	98.634	98.634	97.268	95.628	97.541	94.536
Info Gain	97.814	98.087	97.268	95.902	96.721	94.536
Laplacian Score	97.541	98.361	98.087	95.628	97.541	94.536
One R	96.175	97.814	96.721	95.628	96.448	94.536
Relief	96.995	97.814	97.541	95.628	96.995	94.536
SVM	98.087	98.907	97.814	95.628	97.541	94.536
Symmetrical Uncertainty	98.907	98.634	98.361	95.902	97.814	94.536
Máximo por método	98.907	98.907	98.361	95.902	97.814	94.536
Máximo global	**98.907**					

Tabla A.12: Máximo desempeño selector - clasificador, conjunto de datos *Ecoli*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	83.631	85.417	81.845	84.226	80.357	81.548
Gain Ratio	83.631	85.417	80.952	84.226	80.357	81.548
Info Gain	83.631	85.417	80.952	84.226	80.357	81.548
Laplacian Score	83.631	85.417	80.952	84.226	80.357	81.548
One R	83.631	85.417	80.952	84.226	80.357	81.548
Relief	83.631	85.417	80.952	84.226	80.357	81.548
SVM	83.631	85.417	80.952	84.226	80.357	81.548
Symmetrical Uncertainty	83.631	85.417	80.952	84.226	80.357	81.548
Máximo por método	83.631	85.417	81.845	84.226	80.357	81.548
Máximo global	**85.417**					

Tabla A.13: Máximo desempeño selector - clasificador, conjunto de datos *EggEye*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	55.127	55.120	96.776	84.753	83.652	78.852
Gain Ratio	55.127	55.120	96.776	84.753	83.652	78.852
Info Gain	55.127	55.120	96.776	84.753	83.652	78.852
Laplacian Score	55.127	55.120	96.776	84.753	83.652	78.852
One R	55.127	55.120	96.776	84.753	83.652	78.852
Relief	55.127	55.120	96.776	84.753	83.652	78.852
SVM	55.127	55.120	96.776	85.134	83.658	78.852
Symmetrical Uncertainty	55.127	55.120	96.776	84.753	83.652	78.852
Máximo por método	55.127	55.120	96.776	85.134	83.658	78.852
Máximo global	**96.776**					

Tabla A.14: Máximo desempeño selector - clasificador, conjunto de datos *Geographical Music Chromatic*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	41.832	30.406	34.844	30.878	41.643	22.852
Gain Ratio	41.832	30.406	34.844	30.878	41.643	22.852
Info Gain	41.832	30.406	34.844	31.067	41.643	22.852
Laplacian Score	42.021	24.268	30.878	30.973	38.999	22.852
One R	41.832	32.767	36.166	31.822	41.643	23.041
Relief	41.832	36.449	40.415	32.956	42.115	23.607
SVM	0.000	0.000	0.000	0.000	0.000	0.000
Symmetrical Uncertainty	41.832	30.406	34.844	30.878	41.643	22.852
Máximo por método	42.021	36.449	40.415	32.956	42.115	23.607
Máximo global	**42.115**					

Tabla A.15: Máximo desempeño selector - clasificador, conjunto de datos *Geographical Music Simple*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	40.51	35.694	38.81	32.011	44.004	22.852
Gain Ratio	40.51	35.694	38.81	32.011	44.004	22.852
Info Gain	40.51	35.694	38.81	32.106	44.004	22.852
Laplacian Score	40.51	33.994	35.977	31.161	40.699	23.607
One R	40.604	34.939	39.754	33.522	40.982	22.852
Relief	40.699	34.655	40.699	32.956	41.832	23.607
SVM	0.000	0.000	0.000	0.000	0.000	0.000
Symmetrical Uncertainty	40.51	35.694	38.81	32.106	44.004	22.852
Máximo por método	40.699	35.694	40.699	33.522	44.004	23.607
Máximo global	**44.004**					

Tabla A.16: Máximo desempeño selector - clasificador, conjunto de datos *German Credit*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	76.700	75.800	74.800	74.600	72.300	74.300
Gain Ratio	76.700	75.700	72.700	73.200	72.000	72.600
Info Gain	76.600	75.900	74.800	74.600	72.300	74.300
Laplacian Score	75.800	75.400	73.900	73.400	72.000	73.500
One R	75.900	75.700	71.200	72.400	72.000	73.000
Relief	76.1	76.200	74.800	72.900	73.700	74.400
SVM	0.000	0.000	0.000	0.000	0.000	0.000
Symmetrical Uncertainty	76.400	75.900	74.800	74.600	72.100	73.600
Máximo por método	76.700	76.200	74.800	74.600	73.700	74.400
Máximo global	**76.700**					

Tabla A.17: Máximo desempeño selector - clasificador, conjunto de datos *Glass*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	58.879	55.607	80.374	69.626	78.037	71.028
Gain Ratio	60.748	51.869	80.374	70.093	78.037	71.028
Info Gain	58.879	55.607	80.374	70.561	75.701	72.43
Laplacian Score	61.215	51.402	79.439	69.159	71.495	71.028
One R	57.477	60.280	80.374	70.561	79.439	72.43
Relief	59.813	49.533	78.505	71.963	76.636	71.495
SVM	64.486	55.607	79.439	69.626	71.495	71.028
Symmetrical Uncertainty	59.813	52.336	80.374	70.093	78.037	71.028
Máximo por método	64.486	60.280	80.374	71.963	79.439	72.43
Máximo global	**80.374**					

Tabla A.18: Máximo desempeño selector - clasificador, conjunto de datos *Hepatitis*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	85.807	86.452	87.097	84.516	83.226	84.516
Gain Ratio	85.807	87.742	87.742	83.871	82.581	85.807
Info Gain	85.807	87.742	87.742	83.871	82.581	85.807
Laplacian Score	0.000	0.000	0.000	0.000	0.000	0.000
One R	85.161	86.452	87.097	85.161	83.226	85.161
Relief	85.807	85.161	86.452	83.871	85.161	82.581
SVM	88.390	86.450	87.740	84.520	83.230	85.160
Symmetrical Uncertainty	85.807	87.742	87.742	83.871	82.581	85.807
Máximo por método	88.390	87.742	87.742	85.161	85.161	85.807
Máximo global	**88.390**					

Tabla A.19: Máximo desempeño selector - clasificador, conjunto de datos *Horse Colic*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	82.667	82.333	87.000	88.667	85.667	88.000
Gain Ratio	82.667	82.667	86.667	88.667	85.667	88.000
Info Gain	82.667	82.667	87.000	89.000	85.667	88.667
Laplacian Score	0.000	0.000	0.000	0.000	0.000	0.000
One R	82.000	82.667	87.333	88.333	85.667	88.000
Relief	82.000	85.000	86.000	89.000	86.000	88.000
SVM	83.000	84.333	85.000	89.333	85.667	88.667
Symmetrical Uncertainty	82.667	82.333	87.000	88.667	85.667	88.000
Máximo por método	83.000	85.000	87.333	89.333	86.000	88.667
Máximo global	**89.333**					

Tabla A.20: Máximo desempeño selector - clasificador, conjunto de datos *Iris*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	96.000	96.667	95.333	95.333	96.000	94.000
Gain Ratio	96.000	96.000	96.000	95.333	96.000	95.333
Info Gain	96.000	96.000	96.000	95.333	96.000	95.333
Laplacian Score	96.000	96.667	95.333	95.333	96.000	94.000
One R	96.000	96.000	96.000	95.333	96.000	95.333
Relief	96.000	96.000	96.000	95.333	96.000	95.333
SVM	96.000	96.667	94.667	94.667	95.333	94.000
Symmetrical Uncertainty	96.000	96.000	96.000	95.333	96.000	95.333
Máximo por método	96.000	96.667	96.000	95.333	96.000	95.333
Máximo global	**96.667**					

Tabla A.21: Máximo desempeño selector - clasificador, conjunto de datos *Letter Recognition*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	82.435	66.06	96.415	88.505	96.95	79.260
Gain Ratio	82.435	65.895	96.205	88.470	96.305	79.260
Info Gain	82.435	66.06	96.415	88.505	96.95	79.260
Laplacian Score	0.000	0.000	0.000	0.000	0.000	0.000
One R	82.435	66.06	96.415	88.505	96.95	79.260
Relief	82.435	66.06	96.495	88.600	96.95	79.830
SVM	82.435	64.010	95.995	88.250	96.204	78.760
Symmetrical Uncertainty	82.435	66.060	96.415	88.505	96.950	79.260
Máximo por método	82.435	66.060	96.495	88.600	96.950	79.830
Máximo global	**96.950**					

Tabla A.22: Máximo desempeño selector - clasificador, conjunto de datos *Lymphography*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	87.838	83.784	87.838	79.73	87.162	80.405
Gain Ratio	87.838	83.784	87.838	80.405	87.162	79.054
Info Gain	87.838	83.784	87.838	79.054	87.162	80.405
Laplacian Score	86.490	80.410	82.430	80.410	79.730	79.730
One R	87.838	83.108	87.162	79.73	86.487	79.054
Relief	83.784	82.432	87.838	81.081	87.162	81.081
SVM	0.000	0.000	0.000	0.000	0.000	0.000
Symmetrical Uncertainty	87.838	83.784	87.838	79.054	87.162	78.378
Máximo por método	87.838	83.784	87.838	81.081	87.162	81.081
Máximo global	**87.838**					

Tabla A.23: Máximo desempeño selector - clasificador, conjunto de datos *Madelon*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	61.900	61.550	87.500	78.2	86.800	67.400
Gain Ratio	61.900	60.950	87.500	78.200	86.800	67.400
Info Gain	61.900	61.550	87.500	78.200	86.800	67.400
Laplacian Score	0.000	0.000	0.000	0.000	0.000	0.000
One R	61.800	61.450	71.200	74.500	67.250	65.300
Relief	62.550	61.450	88.250	82.450	88.850	67.200
SVM	0.000	0.000	0.000	0.000	0.000	0.000
Symmetrical Uncertainty	61.900	61.550	87.500	78.200	86.800	67.400
Máximo por método	62.550	61.550	88.250	82.450	88.850	67.400
Máximo global	**88.850**					

Tabla A.24: Máximo desempeño selector - clasificador, conjunto de datos *Messidor*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	67.593	59.166	65.421	65.161	67.593	63.423
Gain Ratio	67.593	56.820	64.813	64.987	67.333	63.423
Info Gain	67.593	59.166	65.421	65.161	67.593	63.423
Laplacian Score	67.941	57.689	64.031	66.725	66.203	63.249
One R	67.941	61.946	64.987	65.074	68.97	63.249
Relief	68.115	63.162	69.157	66.116	67.072	64.031
SVM	68.636	62.294	68.636	66.116	64.726	63.944
Symmetrical Uncertainty	67.593	56.820	64.987	65.161	67.593	63.423
Máximo por método	68.636	63.162	69.157	66.725	68.897	64.031
Máximo global	**69.157**					

Tabla A.25: Máximo desempeño selector - clasificador, conjunto de datos *MG Telescope*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	79.154	75.258	83.465	85.505	81.425	83.854
Gain Ratio	79.154	76.614	83.554	85.505	81.425	83.854
Info Gain	79.154	75.258	83.465	85.505	81.425	83.854
Laplacian Score	0.000	0.000	0.000	0.000	0.000	0.000
One R	79.154	75.258	82.886	85.379	81.194	83.854
Relief	79.154	76.614	85.195	85.478	81.425	84.090
SVM	79.154	76.614	84.974	85.636	81.493	83.996
Symmetrical Uncertainty	79.154	76.614	83.554	85.505	81.425	83.854
Máximo por método	79.154	76.614	85.195	85.636	81.493	84.09
Máximo global	**85.636**					

Tabla A.26: Máximo desempeño selector - clasificador, conjunto de datos *Mushroom*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	100.000	98.855	100.000	100.000	100.000	100.000
Gain Ratio	100.000	98.941	100.000	100.000	100.000	100.000
Info Gain	100.000	98.855	100.000	100.000	100.000	100.000
Laplacian Score	0.000	0.000	0.000	0.000	0.000	0.000
One R	100.000	98.855	100.000	100.000	100.000	100.000
Relief	100.000	98.855	100.000	100.000	100.000	100.000
SVM	0.000	0.000	0.000	0.000	0.000	0.000
Symmetrical Uncertainty	100.000	98.941	100.000	100.000	100.000	100.000
Máximo por método	100.000	98.941	100.000	100.000	100.000	100.000
Máximo global	**100.000**					

Tabla A.27: Máximo desempeño selector - clasificador, conjunto de datos *Nursery*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	93.079	90.324	96.767	97.076	98.380	97.053
Gain Ratio	93.079	90.324	96.767	97.076	98.380	97.053
Info Gain	93.079	90.324	96.767	97.076	98.380	97.053
Laplacian Score	0.000	0.000	0.000	0.000	0.000	0.000
One R	93.079	90.324	96.767	97.076	98.380	97.053
Relief	93.079	90.324	96.767	97.076	98.380	97.053
SVM	0.000	0.000	0.000	0.000	0.000	0.000
Symmetrical Uncertainty	93.079	90.324	96.767	97.076	98.380	97.053
Máximo por método	93.079	90.324	96.767	97.076	98.380	97.053
Máximo global	**98.380**					

Tabla A.28: Máximo desempeño selector - clasificador, conjunto de datos *Primary Tumor*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	47.198	50.443	43.068	41.298	43.068	41.298
Gain Ratio	47.788	50.148	43.068	41.298	43.068	41.298
Info Gain	47.198	50.443	43.068	41.298	43.068	41.298
Laplacian Score	0.000	0.000	0.000	0.000	0.000	0.000
One R	46.903	50.148	41.003	39.823	41.593	39.823
Relief	47.198	50.443	43.363	41.593	41.003	41.593
SVM	0.000	0.000	0.000	0.000	0.000	0.000
Symmetrical Uncertainty	47.198	50.443	43.068	41.298	43.068	41.298
Máximo por método	47.788	50.443	43.363	41.593	43.068	41.593
Máximo global	**50.443**					

Tabla A.29: Máximo desempeño selector - clasificador, conjunto de datos *Sensorless*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	88.711	81.952	0.000	99.469	99.945	97.89
Gain Ratio	88.725	81.529	0.000	99.445	99.944	97.908
Info Gain	88.711	81.952	0.000	99.469	99.945	97.894
Laplacian Score	0.000	0.000	0.000	0.000	0.000	0.000
One R	88.728	81.952	0.000	99.445	99.944	97.903
Relief	88.696	82.389	98.115	99.282	99.937	97.905
SVM	0.000	0.000	0.000	0.000	0.000	0.000
Symmetrical Uncertainty	88.725	81.529	99.884	99.469	99.945	97.896
Máximo por método	88.728	82.389	99.884	99.469	99.945	97.908
Máximo global	**99.945**					

Tabla A.30: Máximo desempeño selector - clasificador, conjunto de datos *Statlog Australian Credit Approval*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	85.797	85.507	85.507	88.261	85.507	87.101
Gain Ratio	85.797	85.507	85.507	88.261	87.246	87.101
Info Gain	85.797	85.507	85.507	88.261	85.507	87.101
Laplacian Score	85.797	77.681	81.014	86.087	85.217	86.377
One R	85.797	85.507	85.507	86.232	87.101	86.812
Relief	85.797	86.232	86.812	86.522	87.101	86.377
SVM	85.797	85.507	87.101	86.087	86.957	86.522
Symmetrical Uncertainty	85.797	85.507	85.507	88.261	85.507	87.101
Máximo por método	85.797	86.232	87.101	88.261	87.246	87.101
Máximo global	**88.261**					

Tabla A.31: Máximo desempeño selector - clasificador, conjunto de datos *Tic Tac Toe*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	65.344	71.608	84.342	88.727	83.612	69.937
Gain Ratio	65.344	71.608	84.342	88.727	83.612	69.937
Info Gain	65.344	71.608	84.342	88.727	83.612	69.937
Laplacian Score	65.344	71.608	85.908	89.144	83.82	69.937
One R	65.344	71.608	84.342	88.727	83.612	69.937
Relief	65.344	71.608	90.605	88.727	87.996	69.937
SVM	65.344	71.608	85.177	88.727	83.612	69.937
Symmetrical Uncertainty	65.344	71.608	84.342	88.727	83.612	69.937
Máximo por método	65.344	71.608	90.605	89.144	87.996	69.937
Máximo global	**90.605**					

Tabla A.32: Máximo desempeño selector - clasificador, conjunto de datos *Wisconsin*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	96.996	96.137	96.567	95.422	95.279	95.422
Gain Ratio	96.996	96.137	96.567	95.851	95.422	96.137
Info Gain	96.996	96.423	96.567	95.422	95.422	95.422
Laplacian Score	96.710	95.994	95.565	94.707	95.136	94.564
One R	96.996	96.137	96.567	95.422	95.279	94.993
Relief	96.996	96.710	95.994	95.565	96.137	96.137
SVM	96.996	96.28	95.851	95.708	95.994	95.565
Symmetrical Uncertainty	96.996	96.137	96.567	95.422	95.279	96.137
Máximo por método	96.996	96.71	96.567	95.851	96.137	96.137
Máximo global	**96.996**					

Tabla A.33: Máximo desempeño selector - clasificador, conjunto de datos *Year Polish*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	95.183	95.183	0.000	95.874	94.316	95.625
Gain Ratio	95.183	95.183	0.000	95.802	95.183	95.625
Info Gain	95.183	95.183	0.000	95.874	95.183	95.625
Laplacian Score	0.000	0.000	0.000	0.000	0.000	0.000
One R	95.183	95.183	0.000	96.054	95.183	95.625
Relief	95.183	95.183	0.000	95.851	95.183	95.706
SVM	95.183	95.183	0.000	95.858	95.183	95.63
Symmetrical Uncertainty	95.183	95.183	0.000	95.86	95.183	95.625
Máximo por método	95.183	95.183	0.000	96.054	95.183	95.706
Máximo global	**96.054**					

Tabla A.34: Máximo desempeño selector - clasificador, conjunto de datos *Zoo*.

Método Filter	Instancias correctamente clasificadas (%)					
	SMO	Naive Bayes	Lazy KStar	J48 Graft	Lazy Ibk	Meta Filtered
Chi Squared	96.040	96.040	98.0198	95.050	98.0198	95.050
Gain Ratio	96.040	96.040	98.0198	95.050	98.0198	95.050
Info Gain	96.040	96.040	98.0198	95.050	98.0198	95.050
Laplacian Score	96.040	96.040	97.0300	96.040	98.0200	96.040
One R	96.040	96.040	97.030	96.040	98.0198	96.040
Relief	96.040	96.040	96.0396	95.050	97.0297	95.050
SVM	0.000	0.000	0.000	0.000	0.000	0.000
Symmetrical Uncertainty	96.040	96.040	98.0198	95.050	98.0198	95.050
Máximo por método	96.040	96.040	98.020	96.040	98.0200	96.040
Máximo global	**98.0200**					

Apéndice B

Cuestionarios de evaluación docente

"Si buscas resultados distintos, no hagas siempre lo mismo."

Albert Einstein

La evaluación docente efectuada en el Instituto Tecnológico Superior del Oriente del Estado de Hidalgo, consiste fundamentalmente de dos cuestionarios. El primero, está dirigido a los estudiantes a fin de recopilar su punto de vista sobre la actuación de los profesores respecto de su actividad en el aula. El segundo se ha diseñado para que las autoridades académicas puedan expresar su criterio al evaluar las actividades académico-administrativas que cada maestro realiza.

A continuación se detalla cada uno de los instrumentos presentados a los diferentes actores:

Cuestionario 1 - **ALUMNOS**

Para responder a las siguientes 48 preguntas, considere la siguiente escala:
A) Excelente
B) Muy bueno
C) Bueno
D) Regular
E) Malo

Dominio de la disciplina
1.- ¿Explica de manera clara los contenidos de la asignatura?
2.- ¿Relaciona los contenidos de la asignatura con los contenidos de otras?
3.- ¿Resuelve las dudas relacionadas con los contenidos de la asignatura?
4.- ¿Propone ejemplos o ejercicios que vinculan la asignatura con la práctica profesional?
5.- ¿Explica la utilidad de los contenidos teóricos y prácticos para la actividad profesional?

Planificación del curso
6.- ¿Cumple con los acuerdos establecidos al inicio de la asignatura?
7.- ¿Durante el curso establece las estrategias adecuadas necesarias para lograr el aprendizaje deseado?
8.- ¿El programa presentado al principio de la asignatura se cubre totalmente?
Ambientes de aprendizaje 9.- ¿Incluye experiencias de aprendizaje en lugares diferentes al aula (talleres, laboratorios, empresa, comunidad, etc.)?
10.- ¿Utiliza para el aprendizaje las herramientas de interacción de las tecnologías actuales de la información (correo electrónico, chats, plataformas, etc.)?
11.- ¿Organiza actividades que me permiten ejercitar mi expresión oral y escrita?
12.- ¿Relaciona los contenidos de la asignatura con la industria y la sociedad a nivel local, regional, nacional e internacional?
13.- ¿Usa ejemplos y casos relacionados con la vida real?

Estrategias, métodos y técnicas
14.- ¿Adapta las actividades para atender los diferentes estilos de aprendizaje de los estudiantes?
15.- ¿Promueve el autodidactismo y la investigación?
16.- ¿Promueve actividades participativas que me permiten colaborar con mis compañeros con una actitud positiva?
17.- ¿Estimula la reflexión sobre la manera en que aprendes?
18.- ¿Se involucra en las actividades propuestas al grupo?
19.- ¿Presenta y expone las clases de manera organizada y estructurada?
20.- ¿Utiliza diversas estrategias, métodos, medios y materiales?

Motivación
21.- ¿Muestra compromiso y entusiasmo en sus actividades docentes?
22.- ¿Toma en cuenta las necesidades, intereses y expectativas del grupo?
23.- ¿Propicia el desarrollo de un ambiente de respeto y confianza?
24.- ¿Propicia la curiosidad y el deseo de aprender?

25.- ¿Reconoce los éxitos y logros en las actividades de aprendizaje?
26.- ¿Existe la impresión de que se toman represalias con algunos estudiantes?
27.- ¿Hace interesante la asignatura?

Evaluación
28.- ¿Identifica los conocimientos y habilidades de los estudiantes al inicio de la asignatura o de cada unidad?
29.- ¿Proporciona información para realizar adecuadamente las actividades de evaluación?
30.- ¿Toma en cuenta las actividades realizadas y los productos como evidencias para la calificación y acreditación de la asignatura?
31.- ¿Considera los resultados de la evaluación (asesorías, trabajos complementarios, búsqueda de información, etc.) para realizar mejoras en el aprendizaje?
32.- ¿Da a conocer las calificaciones en el plazo establecido?
33.- ¿Da oportunidad de mejorar los resultados de la evaluación del aprendizaje?
34.- ¿Muestra apertura para la corrección de errores de apreciación y evaluación?
35.- ¿Otorga calificaciones imparciales?

Comunicación
36.- ¿Desarrolla la clase en un clima de apertura y entendimiento?
37.- ¿Escucha y toma en cuenta las opiniones de los estudiantes?
38.- ¿Muestra congruencia entre lo que dice y lo que hace?

Gestión del curso
39.- ¿Asiste a clases regular y puntualmente?
40.- ¿Fomenta la importancia de contribuir a la conservación del medio ambiente?
41.- ¿Promueve mantener limpias y ordenadas las instalaciones?
42.- ¿Es accesible y está dispuesto a brindarte ayuda académica?
Tecnología de la información y de la comunicación 43.- ¿Emplea las tecnologías de la información y de la comunicación como un medio que facilite el aprendizaje de los estudiantes?
44.- ¿Promueve el uso de diversas herramientas, particularmente las digitales, para gestionar (recabar, procesar, evaluar y usar) información?
45.- ¿Promueve el uso seguro, legal y ético de la información digital?

Satisfacción general
46.- ¿En general, pienso que es un buen docente?
47.- ¿Estoy satisfecha o satisfecho por mi nivel de desempeño y aprendizaje logrado gracias a la labor del docente?

48.- ¿Yo recomendaría a este docente?

Cuestionario 2 - **AUTORIDADES**

Área: Formación

1.- ¿Cuenta con el título y la cédula de licenciatura y posgrado.?
A) Cuenta con el título y la cédula de nivel doctorado
B) Cuenta con dos títulos y cédulas de nivel maestría?
C) Cuenta con el título y la cédula de nivel maestría?
D) Cuenta con el título y la cédula de nivel licenciatura y una especialización?
E) Cuenta con el título y la cédula de nivel licenciatura?
2.- ¿Acredita un diplomado en formación docente o profesional , o realiza una estadía técnica en su área de formación, y asiste a un evento (congreso, simposio, convención, entre otros), y acredita cursos de formación docente y/o profesional y muestra evidencias de aplicación de mejora en su práctica educativa como resultado de la formación adquirida?
A) Acredita un diplomado en formación docente o profesional , o realiza una estadía técnica en su área de formación, y asiste a un evento (congreso, simposio, convención, entre otros), y muestra evidencias de aplicación de mejora en su práctica educativa como resultado de la formación adquirida?
B) Acredita por lo menos dos cursos de actualización docente o profesional, o uno docente y uno profesional, y asiste a un evento (congreso, simposio, convención, entre otros), y muestra evidencias de aplicación de mejora en su práctica educativa como resultado de la formación adquirida?
C) Acredita por lo menos dos cursos de actualización docente o profesional, o uno docente y uno profesional, y asiste a un evento (congreso, simposio, convención, entre otros)?
D) Acredita por lo menos dos cursos de actualización docente o profesional, o uno docente y uno profesional?
E) Acredita por lo menos un curso de actualización docente o profesional?
3.- ¿Diseña un curso de actualización docente o profesional que considere: la guía de sesión, cronograma y los materiales didácticos, y participa en la impartición de cursos entregando el reporte final?
A) Diseña un curso de actualización docente o profesional que considere: la guía de sesión, cronograma y los materiales didácticos, y participa en la impartición del mismo entregando el reporte final
B) Diseña al menos un curso de actualización docente o profesional que considere: la guía de sesión, cronograma y los materiales didácticos
C) Imparte un curso de actualización docente y/o profesional, elabora la guía de

sesión, el cronograma y los materiales didácticos
D) Imparte al menos un curso de actualización docente o profesional en el cual se le proporciona el material de apoyo
E) No imparte ni diseña cursos de actualización docente o profesional

Área: Gestión
4.- ¿Participa en la academia y grupos de trabajo o comités de evaluación y pláticas de inducción o en exámenes de nuevo ingreso o promoción de las carreras del instituto y pertenece a asociaciones académicas o profesionales nacionales o internacionales?
A) Participa en la academia y grupos de trabajo o comités de evaluación y pláticas de inducción o en exámenes de nuevo ingreso o promoción de las carreras del instituto y pertenece a asociaciones académicas o profesionales internacionales
B) Participa en la academia y grupos de trabajo o comités de evaluación y pláticas de inducción o en exámenes de nuevo ingreso o promoción de las carreras del instituto y pertenece a asociaciones académicas o profesionales nacionales
C) Participa en la academia y grupos de trabajo o comités de evaluación y pláticas de inducción o en exámenes de nuevo ingreso o promoción de las carreras del instituto
D) Participa en la academia y grupos de trabajo o comités de evaluación
E) Participa en la academia
5.- ¿Realiza el avance programático de curso y/o diseña la instrumentación didáctica en el seno de las academias por áreas del conocimiento (con docentes que dan la misma asignatura), y los presenta en tiempo y forma, y da seguimiento a lo programado, y realiza las acciones necesarias para propiciar el logro de las competencias profesionales de los estudiantes, y participa en el diseño o seguimiento curricular de los planes de estudio o diseño de especialidades?
A) Realiza el avance programático de curso y/o diseña la instrumentación didáctica en el seno de las academias por áreas del conocimiento (con docentes que dan la misma asignatura), y los presenta en tiempo y forma, y da seguimiento a lo programado realizando las acciones necesarias para propiciar el logro de las competencias profesionales de los estudiantes, y participa en el diseño o seguimiento curricular de los planes de estudio o diseño de especialidades.
B) Realiza el avance programático de curso y/o diseña la instrumentación didáctica en el seno de las academias por áreas del conocimiento (con docentes que dan la misma asignatura), y los presenta en tiempo y forma, y da seguimiento a lo programado, y realiza las acciones necesarias para propiciar el logro de las competencias profesionales de los estudiantes.
C) Realiza el avance programático de curso y/o diseña la instrumentación didáctica

en el seno de las academias por áreas del conocimiento (con docentes que dan la misma asignatura), y los presenta en tiempo y forma, y da seguimiento a lo programado.
D) Realiza el avance programático de curso y/o diseña la instrumentación didáctica en el seno de las academias por áreas del conocimiento (con docentes que dan la misma asignatura), y las entrega en tiempo y forma.
E) Realiza el avance programático de curso y/o diseña la instrumentación didáctica de manera individual, y las entrega en tiempo y forma
6.- ¿Entrega al Departamento Académico una relación de bibliografía básica y de consulta sugerida para mantener actualizado el acervo bibliográfico de su asignatura, y prepara presentaciones electrónicas o antologías para sus clases, y propone el uso de tecnologías de la información y las implementa en su instrumentación didáctica, y utiliza plataformas complementarias para fortalecer el aprendizaje a través de publicaciones, carga de archivos, foros de discusión, etc., y prepara apuntes o guías de estudio o manual de prácticas o prototipos didácticos para sus asignaturas?
A) Entrega al Departamento Académico una relación de bibliografía básica y de consulta sugerida para mantener actualizado el acervo bibliográfico de su asignatura, y prepara presentaciones electrónicas o antologías para sus clases, y propone el uso de tecnologías de la información y las implementa en su instrumentación didáctica, y utiliza plataformas complementarias para fortalecer el aprendizaje a través de publicaciones, carga de archivos, foros de discusión, etc., y prepara apuntes o guías de estudio o manual de prácticas o prototipos didácticos para sus asignaturas.
B) Entrega al Departamento Académico una relación de bibliografía básica y de consulta sugerida para mantener actualizado el acervo bibliográfico de su asignatura, y prepara presentaciones electrónicas o antologías para sus clases, y propone el uso de tecnologías de la información, y las implementa en su instrumentación didáctica, y utiliza plataformas complementarias para fortalecer el aprendizaje a través de publicaciones, carga de archivos, foros de discusión, etc.
C) Entrega al Departamento Académico una relación de bibliografía básica y de consulta sugerida para mantener actualizado el acervo bibliográfico de su asignatura, y prepara presentaciones electrónicas o antologías para sus clases, y propone el uso de tecnologías de la información, y las implementa en su instrumentación didáctica.
D) Entrega al Departamento Académico una relación de bibliografía básica y de consulta sugerida para mantener actualizado el acervo bibliográfico de su asignatura, y prepara presentaciones electrónicas o antologías para sus clases
E) Entrega al Departamento Académico una relación de bibliografía básica y de consulta y/o software necesarios para mantener actualizado el acervo bibliográfico

de su asignatura
7.- ¿Asiste/participa en: ?
- Eventos institucionales (honores a la bandera, ceremonia de inicio de cursos, graduaciones).
- Pláticas o reuniones informativas convocadas por grupos de trabajo con el fin de mejorar la vida institucional.
- Con su grupo a pláticas de formación integral como por ejemplo - Cursos sobre normas de certificación o cursos de seguridad e higiene - La difusión de eventos culturales o que promuevan los valores universales.
o realiza las siguientes actividades:
I. Es miembro de una comisión de seguridad e higiene o participa como expositor en pláticas o reuniones informativas por grupos de trabajo con el fin de mejorar la vida institucional.
II. Organiza pláticas o reuniones informativas de por grupos de trabajo con el fin de mejorar la vida institucional y Participa en la organización de eventos culturales o que promuevan los valores universales.
A) Asiste a / Participa en: - Eventos institucionales (honores a la bandera, ceremonia de inicio de cursos, graduaciones).
- Pláticas o reuniones informativas convocadas por grupos de trabajo con el fin de mejorar la vida institucional.
- Asiste con su grupo a Pláticas de formación integral como por ejemplo cursos sobre normas de certificación o cursos de seguridad e higiene.
- Asiste a la difusión de Eventos culturales o que promuevan los valores universales.
O realiza las siguientes actividades:
I. Es miembro de una comisión de seguridad e higiene o participa como expositor en pláticas o reuniones informativas por grupos de trabajo con el fin de mejorar la vida institucional.
II. Organiza pláticas o reuniones informativas de por grupos de trabajo con el fin de mejorar la vida institucional y Participa en la organización de eventos culturales o que promuevan los valores universales.
O colabora en la parte operativa de eventos culturales o que promuevan los valores universales.

Área: Vinculación
Aplica para docentes con nombramiento de 20 horas en adelante, sin cargo administrativo, o aquellos que hayan realizado vinculación adicional a su carga de trabajo.
8.- ¿Desarrolla, administra, Concerta y asesora proyecto de residencia profesional,

de servicio social, de investigación, innovación o incubación de empresas que den respuesta a las necesidades planteadas por los diferentes sectores de la sociedad?
A) Desarrolla, administra y asesora al menos un proyecto de residencia profesional o de servicio social y al menos un proyecto de investigación, innovación o incubación de empresas que den respuesta a las necesidades planteadas por los diferentes sectores de la sociedad.
B) Concerta y asesora al menos un proyecto de residencia profesional o de servicio social y al menos un proyecto de investigación, innovación o incubación de empresas que den respuesta a las necesidades planteadas por los diferentes sectores de la sociedad.
C) Concerta y asesora al menos un proyecto de residencia profesional o de servicio social o al menos un proyecto de investigación, innovación o incubación de empresas que den respuesta a las necesidades planteadas por los diferentes sectores de la sociedad.
D) Asesora al menos un proyecto de residencia profesional o de servicio social y al menos un proyecto de investigación, innovación o incubación de empresas que den respuesta a las necesidades planteadas por los diferentes sectores de la sociedad.
E) Asesora al menos un proyecto de residencia profesional, de servicio social o de investigación, innovación o incubación de empresas que den respuesta a las necesidades planteadas por los diferentes sectores de la sociedad.
9.- ¿Desarrolla y realiza un curso de capacitación, participa en el diseño de un diplomado, realiza un proyecto, da asesoría técnica, o realiza actividades, para los sectores empresarial, gubernamental y social?
A) Desarrolla y realiza un curso de capacitación o participa en el diseño de un diplomado o realiza un proyecto, para los sectores empresarial, gubernamental y social.
B) Realiza una o más propuestas de cursos de capacitación, diplomados, asesoría técnica o desarrollo de proyecto, etc., para los sectores empresarial, gubernamental y social.
C) Da asesoría técnica a los sectores empresarial, gubernamental y social a través de residencias profesionales.
D) Participa en actividades que generen la vinculación con los sectores gubernamental, empresarial y social (concursos, jurado, asistencia a eventos, etc.).
E) Identifica organismos de los sectores empresariales, gubernamentales y sociales relacionados con su materia en conjunto con la academia a la cual pertenece.
10.- ¿Planea y ejecuta actividades para el fortalecimiento del aprendizaje como reforzamiento al total de sus asignaturas con representantes de los sectores empresarial, gubernamental y social (conferencias, visitas industriales, talleres, foros, seminarios, paneles, entre otras), articulándolo en conjunto con otros docentes y estudiantes de la carrera?

A) Planea y ejecuta actividades para el fortalecimiento del aprendizaje como reforzamiento al total de sus asignaturas con representantes de los sectores empresarial, gubernamental y social (conferencias, visitas industriales, talleres, foros, seminarios, paneles, entre otras), articulándolo en conjunto con otros docentes y estudiantes de la carrera.
B) Planea y ejecuta actividades para el fortalecimiento del aprendizaje como reforzamiento a tres asignaturas que imparte (conferencias, visitas industriales, talleres, foros, seminarios y paneles, entre otras).
C) Planea y ejecuta actividades para el fortalecimiento del aprendizaje como reforzamiento a dos de las asignaturas que imparte (conferencias, visitas industriales, talleres, foros, seminarios y paneles, entre otras).
D) Planea y ejecuta una actividad para el fortalecimiento del aprendizaje como reforzamiento a una de las asignaturas que imparte (conferencias o visitas industriales o talleres o foros o seminarios o paneles, entre otras).
E) Planea actividades para el fortalecimiento del aprendizaje como reforzamiento a sus asignaturas (conferencias o visitas industriales o talleres o foros o seminarios, o paneles entre otras).

Área: Investigación

Aplica a docentes que tengan asignadas horas de descarga o nombramiento para realizar investigación por parte de la institución, o aquellos que hayan realizado investigación disciplinar o educativa adicional a su carga de trabajo.
11.- ¿Presenta informe final o parcial del proyecto de investigación registrado ante la DGEST?
A) Presenta informe técnico final de proyecto de investigación registrado ante la DGEST.
B) Presenta informe técnico de avance del proyecto de investigación registrado ante la DGEST.
C) Tiene un proyecto de investigación en desarrollo registrado ante la DGEST.
D) Presenta un protocolo de proyecto de investigación como responsable técnico.
E) Presenta un protocolo de proyecto de investigación como colaborador.
12.- ¿Involucra y asesora al o los estudiante (s) en el proyecto de investigación, avala bitácora de actividades realizadas y el estudiante genera producto(s) como resultado de las actividades realizadas en la investigación?
A) Asesora al o los estudiante (s) en el proyecto de investigación, avala bitácora de actividades realizadas y el estudiante genera producto(s) como resultado de las actividades realizadas en la investigación.
B) Asesora al o los estudiante (s) en el proyecto de investigación, avala bitácora de actividades realizadas y presenta evidencias de las aportaciones del o los estudiante

(s).
C) Asesora al o los estudiante (s) en el proyecto de investigación y avala bitácora de actividades realizadas.
D) Establece el plan de trabajo de las actividades a realizar por el o los estudiante(s).
E) Involucra de manera oficial al o los estudiante(s) en proyecto(s) de investigación.
13.- ¿Participa en grupos o redes de investigación (área de conocimiento de su competencia o educativos)?
A) Pertenece a una red de investigación.
B) Colabora con cuerpos académicos institucional e interinstitucionales.
C) Colabora en un cuerpo académico institucional.
D) Pertenece a un grupo de investigación y genera un producto (paneles, foros, ponencias, conferencias, artículos, etc.).
E) Pertenece a un grupo de investigación institucional.
14.- ¿Difunde y divulga los resultados de sus proyectos de investigación disciplinar o educativa?
A) Publica resultados de proyectos de investigación (artículos en revistas científicas indexada, capítulo de libro o libros, patentes, etc.).
B) Presenta avances o resultados de su investigación en evento internacional o artículo de revista científica de divulgación internacional.
C) Presenta avances o resultados de su investigación en evento nacionales o artículo de revista científica de divulgación nacional.
D) Presenta avances o resultados de su investigación en evento local, regional o estatal, o artículo de revista científica de divulgación estatal.
E) Presenta avances o resultados de su investigación en evento institucional.

Área: Tutoría

Aplica a docentes que tengan asignados alumnos dentro del programa institucional de tutorías.
15.- ¿Cuenta con un diplomado referente a la tutoría y cursos que apoyan la actividad tutoral?
A) Cuenta con un diplomado referente a la tutoría y cursos que apoyan la actividad tutoral.
B) Cuenta con un diplomado referente a la tutoría.
C) Acredita tres cursos que apoyan la actividad tutoral.
D) Acredita dos cursos que apoyan la actividad tutoral.
E) Acredita un curso que apoya la actividad tutoral.
16.- ¿Elabora el Programa de Acción Tutoral considerando: el diagnóstico, la com-

petencia general y específica, los contenidos, el cronograma de actividades, los recursos necesarios, las estrategias, y la evaluación, y lo presenta en tiempo y forma?
A) Elabora el Programa de Acción Tutoral considerando: el diagnóstico, la competencia general y específica, los contenidos, el cronograma de actividades, los recursos necesarios, las estrategias, y la evaluación, y lo presenta en tiempo y forma.
B) Elabora el Programa de Acción Tutoral considerando: el diagnóstico, la competencia general y específica, los contenidos, el cronograma de actividades, los recursos necesarios, y las estrategias, y lo presenta en tiempo y forma.
C) Elabora el Programa de Acción Tutoral considerando: el diagnóstico, la competencia general y específica, los contenidos, el cronograma de actividades y los recursos necesarios, y lo presenta en tiempo y forma.
D) Elabora el Programa de Acción Tutoral considerando el diagnóstico, la competencia general y específica, los contenidos y el cronograma de actividades, y lo presenta en tiempo y forma.
E) Elabora el Programa de Acción Tutoral considerando: el diagnóstico, la competencia general y específica, los contenidos, y lo presenta en tiempo y forma.
17.- ¿Realiza sesiones planeadas, realiza diagnóstico y canaliza estudiantes en riesgo, y promueve programas de apoyo para la formación integral del tutorado y aplica estrategias?
A) Realiza sesiones planeadas, realiza diagnóstico y canaliza estudiantes en riesgo, y promueve programas de apoyo para la formación integral del tutorado y aplica estrategias.
B) Realiza sesiones planeadas, realiza diagnóstico y canaliza estudiantes en riesgo, y promueve programas de apoyo para la formación integral del tutorado.
C) Realiza sesiones planeadas, realiza diagnóstico y canaliza estudiantes en riesgo.
D) Realiza sesiones planeadas y realiza diagnóstico.
E) Realiza sesiones planeadas.
18.- ¿Entrega un reporte con las actividades desarrolladas en el Programa de Acción Tutoral, presenta evidencias del seguimiento, hace contrastación entre lo planeado y lo realizado, identifica sus áreas de oportunidad y elabora una propuesta de realimentación del Programa?
A) Entrega un reporte con las actividades desarrolladas en el Programa de Acción Tutoral, presenta evidencias del seguimiento, hace contrastación entre lo planeado y lo realizado, identifica sus áreas de oportunidad y elabora una propuesta de realimentación del Programa.
B) Entrega un reporte con las actividades desarrolladas en el Programa de Acción Tutoral, presenta evidencias del seguimiento, realimenta el programa e identifica sus áreas de oportunidad.

C) Entrega un reporte con las actividades desarrolladas en el Programa de Acción Tutoral, presenta evidencias del seguimiento, y realimenta el programa.
D) Entrega un reporte con las actividades desarrolladas en el Programa de Acción Tutoral y presenta evidencias del seguimiento.
E) Entrega un reporte con las actividades desarrolladas en el Programa de Acción Tutoral.