UNIVERSIDAD AUTÓNOMA DEL ESTADO DE HIDALGO

INSTITUTO DE CIENCIAS BÁSICAS E INGENIERÍA

CENTRO DE INVESTIGACIÓN EN TECNOLOGÍAS DE INFORMACIÓN Y SISTEMAS

GENERACIÓN DE REGLAS DE PRODUCCIÓN PONDERADAS

TESIS PRESENTADA EN OPCIÓN DEL TÍTULO DE MAESTRÍA EN CIENCIAS COMPUTACIONALES

AUTOR:

RICARDO SÁNCHEZ GUTIÉRREZ

DIRECTORA:

DRA. MARÍA DE LOS ÁNGELES ALONSO LAVERNIA

PACHUCA DE SOTO, HIDALGO, MÉXICO, JUNIO DEL 2011

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE HIDALGO
INSTITUTO DE CIENCIAS BÁSICAS E INGENIERÍA
ÁREA ACADÉMICA DE COMPUTACIÓN

MCC. 008/2011

Mineral de la Reforma, Hgo., a 24 de mayo de 2011

Ing. Ricardo Sánchez Gutiérrez
Presente

Por este conducto le comunico que el jurado asignado para la revisión de su trabajo de tesis titulado "**Generación de reglas de producción ponderadas**", que para obtener el grado de Maestro en Ciencias Computacionales fue presentado por usted ha tenido a bien, en reunión de sinodales autorizarlo para impresión.

A continuación se integran las firmas de conformidad de los integrantes del jurado.

PRESIDENTE:	DR. JOEL SUÁREZ CANSINO 25/05/2011
VOCAL:	DRA. MARÍA DE LOS ÁNGELES ALONSO LAVERNIA
SECRETARIO:	M. EN C. MARTHA IDALID RIVERA GONZÁLEZ
SUPLENTES:	M. EN C. MARIANO J. POZAS CÁRDENAS

TENTAMENTE
"AMOR, ORDEN Y PROGRESO"

Dr. Jair García Lamont
Coord. de la Maestría en Ciencias Computacionales

c.c.p Expediente/ apl

Centro de Investigación en Tecnologías de Información y Sistemas,
Carretera Pachuca - Tulancingo Km. 4.5, Ciudad Universitaria,
Colonia Carboneras, Mineral de la Reforma, Hidalgo, México, C.P. 42184
Tel. +52 771 7172000 exts. 6738, 6735, 6734 y 6731, Fax 6732
torres@uaeh.edu.mx

Agradecimientos

La elaboración de esta tesis y la conclusión del programa de estudios de la maestría ha requerido de un gran esfuerzo y mucha dedicación de mi parte, esto no hubiera sido posible sin el apoyo, ayuda y cooperación de todas y cada una de las personas que a continuación citaré.

En primer lugar, quiero agradecer a DIOS, por estar siempre a mi lado, por darme la capacidad de entendimiento e iluminarme en esos momentos difíciles, de fortalecer mi corazón y sobre todo por rodearme de todas y cada una de las personas que me han dado soporte y dejado algo importante en mi vida.

A mi familia, porque siempre han estado conmigo en las buenas y en las malas, en especial a mi padre Oscar Sánchez, quien me enseñó que lo mejor que uno puede hacer es aprender cosas nuevas, a mi madre Gloria Patricia Gutiérrez que es el motor de mi vida y un apoyo incondicional en toda mi vida y de la cual estoy orgulloso, a mis hermanos Oscar, Gustavo, Edgar, Cris e Itzel por enseñarme que el estudio es bueno, pero a veces, sólo a veces podemos dejarlo un momento y divertirse. En especial, quiero agradecerle a mi hermana Cristina que gracias a su empeño, dedicación y ganas de ser mejor persona, me obliga a ser mejor cada día, ya que me considera su ejemplo a seguir.

Igualmente, un sincero agradecimiento a mi asesora y directora de tesis la Dr. María de los Ángeles Alonso Lavernia, por el apoyo, paciencia, dedicación y amistad incondicional que me proporcionó a lo largo de todo el posgrado y sin su gran ayuda no hubiera podido llevar a cabo la realización de esta tesis. Igualmente quiero agradecer al Dr. Argelio V. De la Cruz por su apoyo desinteresado en estos últimos meses.

A todos los profesores del CITIS que contribuyeron a mi formación profesional y personal, especialmente, al Dr. Joel Suárez Cansino, al Mtro. Mariano Pozas y a la Mtra. Yira Muñoz.

A Cire, que en este momento es de lo mejor que tengo en mi vida.

A mis compañeros de maestría, por apoyarme, animarme, confiar en mí y brindarme su amistad, especialmente a Lizeth Trejo porque desde el propedéutico hemos forjado una gran amistad y conoce el tiempo y dedicación que le he puesto a esta tesis.

Un agradecimiento especial a Bernardo Perea Chamorro por su apoyo total y sus valiosos consejos en los últimos años. De igual manera, un agradecimiento al profesor Martin Colchado Díaz que es una ejemplo a seguir, sinónimo de superación y de inspiración.

A todas las personas que no nombre pero que han contribuido para que sea una mejor personas, especialmente, a mi amiga Patricia Carolina Rivera Herrada.

Dedicatoria

A mis Padres,

A mis hermanos, especialmente a Cris †

Resumen

Actualmente, la utilización de las Bases de Datos se ha convertido en una herramienta indispensable y fundamental en las organizaciones debido a que permite el almacenamiento de los datos de una manera segura y eficiente y representa una forma fácil de acceder a la información. A partir de los datos, se puede obtener un sinfín de información en forma de reportes, gráficas, listas, pero también, en forma de conocimiento gracias al auge que han tenido las técnicas de Base de Datos e Inteligencia Artificial en los últimos años. Estas áreas son utilizadas en diversos campos del conocimiento para facilitar y emular las actividades del hombre.

Sin embargo, a pesar de que el ser humano, de manera natural, se apoya de los datos y conocimiento simultáneamente para dar resultados y tomar decisiones en su vida cotidiana, en la mayoría de los problemas resueltos mediante técnicas computacionales, se utilizan las técnicas de Base de Datos y Sistemas Basados en Conocimiento de manera separada. Son escasos los problemas en los que se utilizan ambas fuentes, además, son pocas las herramientas computacionales que conjuntan estos dos tipos de información, esto es, datos y conocimiento. Menos aún, los sistemas que generan estructuras de representación de conocimiento a partir de datos, para posteriormente ser utilizadas como complemento al conocimiento especializado del experto en la solución de problemas.

La realización de esta tesis aporta un nuevo enfoque para la generación de conocimiento en forma de Reglas de Producción Ponderadas en el Sucedente a partir de los datos y desde el punto de vista tecnológico, una herramienta computacional donde se encuentre implementada esta nueva aproximación. Para lograr este propósito, se aplicaron técnicas de Minería de Datos como la relevancia de variables y la generación de asociaciones, para obtener el peso diferenciante y reglas de asociación de los datos, respectivamente. Posteriormente, las reglas de asociación obtenidas son traducidas a Reglas de Producción Ponderadas en el Sucedente, y respaldadas digitalmente para su uso futuro.

La importancia de esta investigación está dada en la posibilidad de poder emular el quehacer humano cuando resuelve sus problemas, utilizando el conocimiento especializado de éste y el conocimiento obtenido a partir de un conjunto de datos sobre el mismo tema.

ÍNDICE

Pág.

Índice de Figuras

Pág.

Índice de Tablas

Introducción

Las *Ciencias de la Computación* (CC) abarcan una amplia gama de áreas que se enfocan a los fundamentos teóricos de la Informática y la Computación, así como, al desarrollo de los sistemas computacionales, útiles para las organizaciones y centros académicos [Brookshear, 2005]. Dentro de las CC hay diferentes disciplinas que comprenden estudios tales como: los resultados mostrados por una computadora, el costo computacional que demandan los algoritmos computacionales, desarrollo de soluciones a problemas computacional específicos, entre muchos otros.

En las CC hay áreas que han tenido un gran crecimiento y aceptación en los últimos años entre investigadores, organizaciones y desarrolladores en general, dentro de las cuales se encuentran:

- La *Minería de Datos* (MD), proporciona algoritmos para la extracción de conocimiento o patrones de comportamiento de los datos, radicando su importancia en que los patrones ocultos en los datos muestran conocimiento que se desconoce y ayuda a las organizaciones e instituciones en la toma de decisión [Hernández, 2004]. Esta disciplina forma parte del *Descubrimiento de Conocimiento en Base de Datos* (DCBD), la cual aplica un procesamiento especializado a los datos para seleccionar, limpiar y transformar estos datos y así mejorar su calidad, dando también la posibilidad de evaluar y validar los modelos obtenidos en la etapa de MD para obtener patrones más confiables y legibles [Klösgen, 2002].
- La *Inteligencia Artificial* (IA), debido a que se enfoca al estudio del razonamiento y comportamiento humano y trata de emularlos para incorporar esta conducta inteligente a dispositivos electrónicos y sistemas de cómputo [Russell, 2003]. Algunas de las disciplinas que conforman la IA son: Aprendizaje Automático, Procesamiento del Lenguaje Natural, Visión Artificial, Vida Artificial y *Sistemas Expertos* (SE) o *Sistemas Basados en Conocimiento* (SBC). Esta última técnica, brinda la posibilidad de almacenar el conocimiento de un especialista en un repositorio llamado *Base de Conocimiento* (BC) e implementarlo en una computadora, la cual da la oportunidad de disponer del conocimiento, razonamiento

y diagnóstico de un especialista humano sin tener la necesidad de que el especialista esté presente, lo que puede ser de gran ayuda en poblaciones rurales lejanas, en espacios de alto riesgo, como hornos de fundición o simplemente, donde no se cuente con una persona con tales características [Leondes, 2000]. Los SBC son utilizados normalmente para:

1. Ayudar a la toma de decisiones, ya que a partir de una problemática determinada sugiere la solución que consideran más idónea a partir del conocimiento incluido en el sistema.
2. Ayudar a la configuración, ya que se encarga de la selección y planificación de los componentes que se necesitan en un proceso determinado.
3. Ayudar al diagnóstico porque a partir de unos síntomas, determinan las causas que lo producen. Son utilizados comúnmente para el diagnóstico de enfermedades.

Hoy en día es una práctica muy común, el automatizar la actividad humana en la computadora, para lo cual se utilizan, técnicas, modelos y metodologías para lograr este objetivo. Las áreas mencionadas han generado los medios para lograr implementaciones que apoyan al pronóstico, diagnóstico y, la toma de decisión en general.

Problemática científica y tecnológica

Frecuentemente, nos encontramos con problemas donde el ser humano se apoya de los datos y la experiencia para dar solución a los mismos, ejemplo de ello son: el diagnóstico médico y la solución de una estrategia para el proceso de enseñanza y aprendizaje. En la generalidad de las soluciones basadas en técnicas computacionales, estas alternativas se han utilizado de una manera eficiente, pero mayormente de forma separada, es decir, se pueden encontrar trabajos donde se resuelven problemas a partir de los datos y por otro lado, soluciones que implementan sistemas expertos a partir del conocimiento de un especialista. Sin embargo, son muy pocos los trabajos donde se combinan ambas fuentes, conocimiento y datos, y se utilizan comúnmente para la toma de decisiones [Roda, 2001], [Bao, 2002], [Wada, 2001].

Los trabajos que combinan ambas técnicas buscan regularmente obtener reglas de los datos y que estas reglas puedan ser utilizadas posteriormente, para algún análisis o toma de decisión. No obstante, en muy escasas situaciones se encuentran trabajos donde las reglas generadas son reglas de producción y más aún que éstas se encuentren ponderadas en el sucedente de las reglas como normalmente se utilizan en los SBC.

Esta problemática está dada, principalmente, porque son escasas las herramientas computacionales que logran trabajar conjuntamente con estas dos fuentes de información, *los datos y el conocimiento*, y además, porque en la gran mayoría de los casos, los especialistas en MD no lo son en el área de los SBC.

Propuesta de solución

Dada la problemática planteada, se propone un enfoque para la generación de conocimiento a partir de una base de datos en forma de *Reglas de Producción Ponderadas* (RPP) *en el Sucedente*.

Justificación

El trabajo que se presenta brinda, de alguna manera, contribuciones de tipo científico, tecnológico, social y económico lo cual se materializa de la siguiente manera:

Aportación científica. Se ha propuesto un nuevo enfoque para la generación de conocimiento en forma de reglas de producción ponderadas en el sucedente, con el cual permitirá resolver problemas, donde se tengan disponibles datos y conocimiento, como dos fuentes de información de entrada.

Aportación tecnológica. Para poner en práctica este nuevo enfoque ha sido necesario la implementación de un sistema informático orientado a la Web, lo cual permitirá generar conocimiento a partir de los datos para utilizarlo junto con el conocimiento experto.

Aportación social. Una vez validada su funcionalidad y pertinencia, podrá ser puesto a disposición de la comunidad científica y estudiantil para utilizarlo en la resolución de problemas y como apoyo al proceso de enseñanza y aprendizaje en las materias de Minería de Datos e Inteligencia Artificial.

Aportación económica. Dado que resulta ser un software de libre distribución, en las instituciones educativas que hagan uso de este tipo de herramientas, su adquisición no representará un problema económico.

Objetivos

Para llevar a cabo el desarrollo de esta tesis se han formulados algunos objetivos que se describen concisamente en esta sección.

Objetivo general

Desarrollar una herramienta computacional que pueda generar conocimiento a partir de una base de datos en forma de Reglas de Producción Ponderadas en el Sucedente para poder complementar el conocimiento de un experto con el conocimiento extraído de los datos, y así utilizarlas de manera conjunta en problemas donde se encuentren disponibles datos y conocimiento.

Objetivos específicos

1. Investigar y seleccionar un método de relevancia de las variables, con el objetivo de obtener el peso diferenciante de éstas.
2. Investigar y seleccionar un método para generar asociaciones que permita obtener *reglas de asociación* (RA).
3. Traducir las reglas de asociación obtenidas de los datos como reglas de producción ponderadas en el sucedente utilizando el peso diferenciante de las variables.
4. Respaldo de la información en archivos digitales del peso diferenciante de las variables, reglas de asociación, reglas de producción ponderadas en el sucedente, lista y parámetros de las proposiciones.
5. Analizar, diseñar y desarrollar la herramienta computacional que implementará los métodos de MD y el respaldo de la información obtenida por éstas.

Tareas desarrolladas

Las tareas que se tuvieron que llevar a cabo para poder realizar esta tesis, las cuales brindaron la posibilidad de desarrollar la herramienta computacional propuesta se enlistan a continuación:

- ***Posicionamiento y conocimiento sobre el contexto del problema***: Para el cumplimiento de esta tarea se cursaron y aprobaron materias como sistemas basados en conocimiento, base de datos y minería de datos.
- ***Análisis y estudio del estado del arte referente al contexto del problema****:* Una vez que se obtuvo un conocimiento general, se dispuso a obtener un conocimiento más específico sobre el contexto del problema y para esto, se realizó una revisión a los trabajos, artículos, herramientas, software y libros. Como resultado de esta revisión se obtuvo un amplio conocimiento del tema ya que se pudo profundizar en esté y se documento el estado del arte, que se presenta en ésta tesis.
- ***Análisis y asimilación de los algoritmos de minería de datos:*** Esta tarea dio la posibilidad de estudiar, analizar y comprender los algoritmos de MD que se implementaron posteriormente en la herramienta propuesta. Los algoritmos estudiados son:
    - De relevancia de variables, se analizó la teoría basada en Testores Típicos
    - Del análisis de asociaciones, se estudió el algoritmo llamado GUHA (General Unary Hypotheses Automaton)
- ***Estudio y comprensión de las herramientas de desarrollo web*****:** Al llevar a cabo esta tarea se adquirió el conocimiento relacionado con sintaxis, metodología de programación, tipos de datos, programación orientada a objetos, etc., para el manejo de herramientas como *xhtml*, *css*, *javascript* y *php*.
- ***Análisis y diseño del sistema:*** Se analizaron todos los requerimientos que necesitaría la herramienta para posteriormente elaborar su diseño conceptual y lógico.

- ***Programación del sistema:*** Una vez que se realizó la etapa del análisis y diseño, se llevó a cabo el desarrollo de la herramienta computacional con software enfocado al desarrollo Web.

Hipótesis

Con la ayuda de la herramienta computacional propuesta, el conocimiento que se encuentra inherente en los datos puede ser utilizado para complementar y ampliar el conocimiento de un especialista que se encuentra en una Base de Conocimiento.

Aportes teóricos o tecnológicos

Con esta tesis se propone dos aportaciones principales, las cuales son:

1) *Aporte teórico*: Un nuevo enfoque para generar reglas de producción ponderadas en el sucedente partiendo de los datos, utilizando técnicas de MD como son: la relevancia de variables y la generación de asociaciones.
2) *Aporte tecnológico:* Una herramienta computacional orientada a la web que implementa el enfoque propuesto.

Metodología

Se describe brevemente los elementos computacionales empleados para el desarrollo de esta tesis:

- Se hizo uso de dos algoritmos de MD para generar conocimiento y poder generar las reglas de producción ponderadas en el suceden
    - Uno de relevancia de variables llamado teoría de Test y Testores para obtener el peso diferenciante de las variables
    - Otro de Análisis de asociaciones nombrado GUHA, el cual permite generar reglas de asociación a partir de los datos
- Implemento una traducción de reglas de asociación a reglas de producción ponderadas en el sucedente y se utilizó teoría de incertidumbre para calcular el peso del sucedente de la reglas.

- Se diseñaron algunos formatos de archivos digitales y otros se tomaron del software para generar sistemas expertos HAries, con el objetivo de crear una base de conocimiento con dicho software.
- Para la implementación del enfoque propuesto se utilizo software orientado a la WEB como el lenguaje de programación PHP, JavaScript y HTML, con estas herramientas computacionales se desarrollo el sistema de generación de reglas de producción ponderadas en el antecedente.

Herramientas empleadas

En este apartado, se describen técnicas y herramientas computacionales que fueron utilizadas para llevar a cabo el desarrollo de esta tesis. Se emplearon tanto algoritmos de MD como software computacional para la elaboración de la herramienta propuesta.

Técnicas utilizadas

Para lograr los objetivos planteados y el desarrollo del sistema computacional, se aplicaron diferentes técnicas y otras pertenecientes a la MD como son *Relevancia de Variables* (RV) [Liu, 2008] y *Análisis de Asociaciones* (AA) [Zhang, 2002]. Y por último se ocupó la programación orientada a la Web para el desarrollo del software. Estas técnicas se comentan brevemente a continuación.

En cuanto a las técnicas de MD que se aplicaron fueron, uno para conocer la relevancia de variables y el otro para generar reglas de asociación. Dichas técnicas se comentan enseguida.

La RV forma parte de la MD y con la aplicación de esta técnica se pudo obtener el peso *diferenciante* (PD) de cada una de ellas. Este valor, permite conocer la importancia o relevancia que tienen las variables y está comprendido en un rango de 0 a 1 inclusive. Una vez obtenido este peso, se utilizó para poder asignar la ponderación de los objetivos en las reglas de producción.

Otra técnica que se utilizó y pertenece a la MD es el AA, la cual nos permitió generar RA y así, expresar combinaciones de valores de los atributos que suceden más frecuentemente en un conjunto de datos; dichas reglas constan de dos partes, principalmente, un antecedente y

un consecuente, donde el consecuente puede estar compuesto de más de un elemento. Una vez que se obtuvo este tipo de reglas se sometieron a una transformación o traducción para generar reglas de producción.

Por último, se utilizó la programación orientada a la Web para desarrollar la herramienta computacional, ya que permite crear páginas Webs dinámicas, esto gracias a que un servidor interpreta el código del lenguaje que está embebido en el HTML y se lo muestra al usuario a través del browser o navegador Web. Con la programación orientada a la Web se pueden implementar todo tipo de algoritmos que involucren cálculos matemáticos, manejo de estructuras de datos y manipulación de archivos, entre muchas otras posibilidades.

Herramientas utilizadas

La herramienta computacional fue desarrollada con el lenguaje de marcado de hipertexto HTML [Duckett, 2008] para la construcción de la página web, permitiendo la interacción entre la herramienta y el usuario, es decir, con HTML se realizó la interfaz del sistema apoyándose de otras herramientas como hojas de estilo para definir la presentación del documento y el lenguaje de JavaScript permite mejorar la interacción entre la interfaz y el usuario.

Por otra parte, para poder implementar los algoritmos de MD y el desarrollo de otros procesos, y que además, interactuará con HTML, se eligió el lenguaje de programación PHP [Suehring, 2009], el cual va embebido en el código HTML, para implementar el proceso de MD, manejo de estructuras de datos, manipulación de archivos, programación orientada a objetos, entre otras tareas.

Alcance

La herramienta desarrollada sustenta el aporte teórico propuesto en la presente tesis, con la cual se pueden aplicar varias tareas de MD para generar RPP.

Esta herramienta computacional se compone de cuatro módulos los cuales enlistan a continuación:

- *Relevancia de variables*.- En este módulo, se obtiene el peso diferenciante de las variables, el cual da la posibilidad de conocer la importancia que tiene cada una de

las variables y es utilizado para dar la ponderación a las reglas de producción, generadas posteriormente.

- *Generación de asociaciones*.- Aquí, se extraen de los datos reglas de asociaciones, las cuales representan conocimiento inherente de los datos, estas reglas serán utilizadas posteriormente.
- *Trasformación a reglas de producción ponderadas en el sucedente*.- Se generan de forma automática las reglas de producción ponderadas en el sucedente a partir de las reglas de asociación y la ponderación a cada reglas está dada por el peso diferenciante.
- *Respaldo de información*.- En este módulo, se guardan los resultados parciales y finales que genera el sistema como son peso diferenciante de las variables, reglas de asociación y reglas de producción ponderadas cada uno de estos resultados se guardan en archivos digitales independientes.

Limitaciones

Se enumeran las tareas o procesos que no forman parte de este trabajo.

- Los datos de entrada se reciben en un archivo plano y deben estar categorizados por la no culminación de un sistema de adquisición y procesamiento que forma parte del proyecto.
- La calidad de los datos depende del usuario, ya que el sistema no aplica algoritmos que permiten seleccionar o limpiar los datos.
- No se aplica ningún algoritmo para reducir el número de variables, es decir, tratar con la dimensionalidad de los datos, aunque en un futuro fácilmente podría implementarse, ya que se tomaría como base la importancia de cada variable obtenida en este trabajo, para reducir la dimensionalidad del problema bajo estudio.

Estructura del documento

En este apartado se describe brevemente el contenido de cada unos de los capítulos que conforman la presente tesis. La tesis está compuesta de siete capítulos en el orden como se enlistan a continuación:

En el primer capítulo se puede encontrar un análisis relacionado al estado del arte y los trabajos, software y artículos similares al enfoque que se propone en esta tesis, el cual se

refiere a la generación de reglas de producción ponderadas en el sucedente a partir de un conjunto de datos sobre un problema en particular dado.

El segundo capítulo comprende información sobre los componentes que integran el enfoque propuesto, así como también de las características que deben cumplir los datos para poder ser sometidos a los diferentes módulos del sistema de cómputo. Los elementos que integran el enfoque propuesta son: un modulo para obtención del peso diferenciante de las variables, obtención de reglas de asociación, traducción a reglas de producción ponderadas en el sucedente y el respaldo de la información.

En el tercer capítulo aborda el cálculo del peso diferenciante de las variables, el cual permite identificar con un valor numérico la importancia que tienen las variables para el conjunto de datos, en este enfoque, se utiliza el peso diferénciate para obtener los pesos de los sucedentes de las reglas de producción generadas. También, se describe la teoría de testores y el algoritmo BT para el cálculo del peso diferenciante.

El cuarto capítulo trata sobre la generación de reglas de asociación y del método GUHA, el cual pertenece a la minería de datos. En el quinto capítulo habla de la teoría para la traducción a reglas de producción ponderadas en el antecedente, la cual para poderse llevar acabo necesita de los resultados del peso diferenciante y las reglas de asociación. En el penúltimo capítulo, que es el seis, se describen los archivos digitales que genera el enfoque y que pueden ser utilizados para diferentes objetivos, en particular, el enfoque muestra tres archivos que pueden ser importados al software HAries y así generar un sistemas experto, dichos archivos contienen la base de reglas, la base de proposiciones y los parámetros de las proposiciones.

Por último, en el capítulo siete se muestra la implementación del enfoque propuesto en forma de sistema de cómputo orientado a la web y desarrollado en el lenguaje de programación PHP.

Capítulo 1 Estado del Arte

1.1 Introducción

En este capítulo, se describen los antecedentes científicos que se han encontrado sobre el tema, es decir, los trabajos similares en forma de metodologías, software, aplicación, entre otras.

1.2 Trabajos similares

Los trabajos similares se dividieron en tres secciones, en la primera, están agrupados todos aquellos trabajos que descubren conocimiento a partir de una base de datos, en la segunda, se hace una reseña del software de propósito general que implementan algoritmos de MD y la tercera y última, comprende los trabajos que usan minería de datos y sistemas basados en conocimiento conjuntamente.

1.2.1 Descubrimiento de conocimiento a partir de base de datos

En esta sección, se presentan algunos trabajos que tienen características de generar conocimiento a partir de un conjunto de datos como son las bases de datos, los cuales utilizan técnicas de minería de datos, para la solución de problemas y la toma de decisiones.

1.2.1.1 Descubriendo conocimiento para el mejoramiento genético bovino usando técnicas de data mining

Este proyecto está enfocado a la ganadería, y utiliza técnicas de MD para determinar los criterios que establecen si un animal es un buen reproductor o no. Fue desarrollado en la Universidad Politécnica de Catalunya en el año del 2001 [Molina, 2001].

El trabajo trata de determinar los criterios genéticos tanto de los padres como del animal para poder alcanzar el mejor desarrollo escrotal a determinado número de días, esto, según el autor, hace que se tenga un posible gran reproductor. El software que se utiliza para llevar a cabo su investigación es libre o versión beta. Para determinar las variables más relevantes, y así reducir la dimensionalidad, se manejan tres algoritmos: Medida de Consistencia [Dash, 2000], Relieff [Robnik, 2003] y ES-RS-E [Zhang, 2004]. Los tres métodos generan un subconjunto de variables que representan las más relevantes, pero los

dos últimos calculan el peso diferenciante para cada una de las variables del subconjunto. Para encontrar los criterios genéticos más importantes que determinan cuando un animal es un buen reproductor, se utilizan CN2 [Clark, 1989] y C4.5 [Quinlan, 1993]. Estos algoritmos generan reglas de clasificación y son aplicados a datos supervisados[1].

El proyecto se aplica a un problema en particular en la ganadería y por lo cual es difícil generalizarlo para otro tipo de problemas ya que parte de una base de datos que contiene datos vacunos. Cada algoritmo utilizado en este trabajo es aplicado a los datos por separado, es decir, no forman parte de una herramienta computacional que las integre. En cuanto a la selección de variables, se utilizó estrictamente para obtener un subconjunto de variables que reducen la dimensionalidad, aunque se calculó el peso diferenciante de las variables, este solamente fue utilizado para identificar las variables menos relevantes y poderlas eliminar. Con respecto al tipo de reglas generadas, éstas fueron del tipo de clasificación de la forma IF-THEN, que no son reglas de producción ponderadas. Las reglas generadas no contribuyen al conocimiento de un SE ni se crea ninguna BC.

1.2.1.2 Obtención de patrones y reglas en el proceso académico de la Universidad de las Ciencias Informáticas utilizando técnicas de minería de datos

Este proyecto está enfocado a la educación y utiliza técnicas de MD para encontrar los criterios que permitan elevar la calidad de la educación de la Universidad de las Ciencias Informáticas (UCI). Fue desarrollado en el Instituto Superior Politécnico José Antonio Echeverría, Ciudad de la Habana, Cuba [González, 2007].

El objetivo de este trabajo es encontrar factores para mejorar el proceso de formación académica de un alumno de nivel licenciatura y elevar la calidad de la educación en la Universidad. Para encontrar dichos factores en forma de patrones, los autores se apoyaron de técnicas de MD como: clustering para clasificar a los estudiantes de acuerdo a su rendimiento académico, también utilizaron árboles de decisión y algoritmos de aprendizaje inductivo para encontrar patrones ocultos y reglas que los caracterizan. En este trabajo, se tratan de identificar a alumnos con riesgo de baja académica, alumnos con mayor potencial y, de esta forma, brindar una atención diferenciada a cada uno de los diversos tipos de estudiantes. Se utilizó software propietario para llevar a cabo el proceso de MD. Para la

[1] Es cuando los datos están agrupados en clases, y estas clases son representados por un atributo especial

relevancia de variables, se llevaron a cabo dos estudios, uno manual donde el especialista seleccionó las variables, que según su criterio eran más importantes y el otro automático, con el software Business Intelligence Development Studio. Con respecto al tipo de reglas generadas, se extrajeron reglas de clasificación para predecir los resultados académicos de los alumnos, para esto, se utilizó Microsoft SQL Server Analysis Service.

En esta investigación, no se desarrolló ningún software para llevar el proceso de MD, ya que éste fue realizado por software propietario de Microsoft, con la implicación de comprar licencias para el uso del software. Con la aplicación de los algoritmos de RV no se obtuvo el peso diferenciante, ya que se utilizó para reducir la dimensionalidad, también se utilizó la experiencia de un especialista para reducir la dimensionalidad de los datos y esto lo hace poco práctico para su generalización. En cuanto al tipo de reglas obtenidas de los datos, se generaron reglas clasificación, las cuales no están ponderadas, tampoco estas reglas contribuyen al conocimiento de un SBC, ni en esta investigación se creó un SBC.

1.2.1.3 Data Preparation Process for Construction Knowledge Generation through Knowledge Discovery in Databases

Esta investigación se enfoca a la industria de la construcción, ya que trata de predecir si una obra tendrá un retardo en su construcción o no, dependiendo de unas variables iníciales, los autores se apoyan de técnicas de MD y de DCBD para dar solución a este problema. Fue desarrollado por investigadores de la Universidad de Illinois, Urbana Illinois [Soilbelman, 2002].

El objetivo de este trabajo es identificar a tiempo aquellas situaciones que pueden originar posibles retardos en la industria de la construcción, y de esta forma, evitar posibles gastos financieros extras. Para poder llevar a cabo esto, se desarrolló un software que implementa el proceso de DCBD, desde el pre-procesamiento hasta la extracción de patrones. Dicho software contiene dos algoritmos de RV para seleccionar las variables más importantes. Los algoritmos se basan en el enfoque wrapper[2] y son C4.5 y Naive-bayes [Friedman, 1997]. También, se eliminaron de manera manual aquellas variables donde la mayoría de los datos no estaban almacenados, mal capturados o fuera del rango. Por otra parte, se utilizó el algoritmo de árbol de decisión C4.5 para crear reglas de clasificación, donde los

[2] El algoritmo de aprendizaje es envuelto dentro del proceso de selección

nodos finales del árbol definen si el conjunto de variables representa un retardo en la construcción o no. Además, se utiliza una red neuronal para predecir posibles retardos que pudieran causar las variables de entrada de la red neuronal artificial.

En esta investigación, la selección de variables se utilizó estrictamente para reducir la dimensionalidad de los datos y no se calculó el peso diferenciante de éstas, igualmente se eliminaron variables de forma manual, cosa que es impráctico a la hora de generalizar el enfoque y hacerlo automático. En cuanto al tipo de reglas, se obtuvieron reglas de clasificación, las cuales no son ponderadas, estas reglas fueron utilizadas estrictamente para predecir si la obra puede caer en un retardo o no. Con las reglas generadas, en este trabajo, no se construyó el conocimiento de un SBC ni se crea una BC.

1.2.1.4 Rule Generation for Protein Secondary Structure Prediction with Support Vector Machines and Decision Tree

Esta investigación se enfoca al área de Biología, se utilizan técnicas de MD para identificar las estructuras secundarias de proteínas. Este proyecto, fue realizado por profesores investigadores de la Universidad del Estado de Georgia, Atlanta [He, 2006].

La publicación se enfoca a la tarea de clasificación y predicción de estructuras secundarias de proteínas en el área de Biología. Estas estructuras secundarias permiten obtener información valiosa sobre la funcionalidad y propiedades estructurales de tales proteínas a partir de su secuencia de aminoácidos. En este trabajo, se aplica un algoritmo de clasificación llamado máquina de vectores soporte, para encontrar los tres tipos de estructuras secundarias existentes: hélice, lámina y espiral [Wang, 2005]. También, se aplica el algoritmo C4.5 para generar reglas de clasificación y así, predecir las estructuras secundarias a partir de nuevas secuencia de aminoácidos, encontradas con el algoritmo máquina de vectores soporte. Para la aplicación de los métodos mencionados, se utilizaron los software's SVM-Light y C4.5, respectivamente.

En este trabajo, no se aplicó ningún método automático o manual para la selección de variables, por lo tanto, no se conoció el peso de las variables que fueron utilizadas en el proceso de MD. En cuanto al tipo de reglas generadas, al aplicar un algoritmo de árbol de decisiones, fueron del tipo de reglas de clasificación, las cuales no se encontraban

ponderadas. Estas reglas no se utilizan para complementar la BC de un sistema experto ni tampoco se crea una BC.

1.2.1.5 Generating Weighted Fuzzy Production Rules Using Neural Networks

Este proyecto trata de resolver problemas de clasificación utilizando redes neuronales artificiales y lógica difusa. Se presenta un enfoque diferente para generar reglas de producción ponderadas difusas. La investigación fue desarrollada por investigadores de la Universidad de Hebei y del Instituto Hebei de Tecnología Industria, China [Fan, 2006].

El trabajo se enfoca a la tarea de clasificación y presenta un enfoque para generar reglas de producción ponderadas difusas. En primer lugar, a partir de un conjunto de datos se someten a un proceso de fusificación[3] a todas las variables. Después, se construye una red neuronal BackPropagation con tres capas, el número de nodos en la capa de entrada es igual número total términos lingüísticos, la capa oculta se compone de cuatro neuronas y la capa de salida está formada por el número de clases en el conjunto de datos. Las reglas de producción ponderadas difusas se obtienen del producto de los pesos en la capa de salida con los de la capa oculta. Los pesos de las RPP se encuentran en el antecedente de la regla [Wang, 2006].

En este trabajo no se realiza ningún proceso de selección de variables, así que los autores no conocen el peso de las variables y las utilizan con la misma importancia. Con respecto al tipo de reglas generadas, son reglas de producción ponderadas difusas, la ponderación de las reglas se encuentra en el lado del antecedente. Las reglas generadas no contribuyen ni forman parte del conocimiento de un SBC.

1.2.1.6 Comparación de trabajos encontrados

Se muestra una comparación en la Tabla 1.1 donde se consideran las técnicas, algoritmos y resultados alcanzados en los trabajos analizados anteriormente, los cuales están muy relacionados con el enfoque propuesto en este trabajo.

[3] Etiquetando el valor crisp de una variable de entrada con un término lingüístico y determinando el correspondiente grado de pertenencia.

Tabla 1.1. Comparación de trabajos que generan reglas a partir de un conjunto de datos.

TRABAJO	TÉCNICAS DE MD	ALGORITMOS	RELEVANCIA DE VARIABLE	TIPO DE REGLAS	REGLAS PONDERADAS	BASE DE CONOCI.	SW
[Molina, 2001]	Selección de variables, Árboles de decisión.	Consistencia, Relief, ES-RS-E, CN2 y C4.5.	Reducción de dimensionalidad	Clasificación	✗	✗	Utiliza software libre y beta, pero no se desarrolla ninguna herramienta computacional
[González, 2007]	Árboles de decisión y relevancia de variables	Suit de Microsoft	Manualmente	Clasificación	✗	✗	Software propietario, no se desarrolla ninguna aplicación
[Soilbelman, 2002]	Árboles de decisión, redes neuronales	Naivi-Bayes, C4-5 y Redes neuronales artificiales	Reducción de dimensionalidad	Clasificación	✗	✗	Se desarrolla una aplicación que involucra el proceso de DCBD
[He, 2006]	Árboles de decisión y Máquina de vectores soporte	MVS y C4.5	✗	Clasificación.	✗	✗	El software utilizado es de libre distribución.
[Wang, 2006].	Redes neuronales y lógica difusa	Backpropagation	✗	Clasificación	La ponderación se encuentra en el antecedente	✗	El software utilizado es de libre distribución

Nota.- el símbolo ✗, indica que no se realizó, generó o construyó el proceso correspondiente a la columna donde se encuentra.

1.2.2 Software de propósito general

Se describen algunas herramientas computacionales que aplican técnicas de minería de datos para la extracción de conocimiento y todo el proceso de DCBC partiendo de una base de datos. Entre los algoritmos que se pueden encontrar en estas herramientas están los de clasificación y predicción, análisis de clúster, análisis de asociaciones y relevancia de variables, haciendo énfasis en los algoritmos que generan reglas y la relevancia de variables. Las herramientas que se estudiaron en esta apartado son: TaryKDD, Orage Data Mining y Weka.

1.2.2.1 TariyKDD

TariyKDD es una herramienta computacional para descubrir conocimiento a partir de una DB, esta herramienta de propósito general fue desarrollada por la Universidad de Nariño, Colombia con software libre y distribuido bajo la licencia GNU. El software comprende módulos para conectar a la BD y archivos planos, pre-procesamiento de datos y por último,

un módulo que involucra el núcleo de MD [Timarán, 2009]. El núcleo de la MD en TariyKDD se compone de cincos algoritmos para la extracción de conocimiento en forma de reglas, los cuales son: *Apriori* [Agrawal, 1994], *FPGrowth* [Borgelt, 2005], y *EquipAsso* [Timarán, 2005] para análisis de asociaciones y *C4.5* y *Mate* para tareas de clasificación.

La selección de variables se realiza de forma manual y esto requiere de experiencia o conocimiento para eliminar las variables menos importantes. Los dos tipos de reglas que pueden generarse con esta herramienta son de clasificación y asociación, las cuales no se encuentran ponderadas.

Aunque se pueden guardar las reglas en archivos digitales, la herramienta no genera ni da la posibilidad de exportar el conocimiento extraído de los datos a una BC de un sistema experto.

1.2.2.2 Orange Data Mining Fruitful & Fun.

Es un framework de propósito general compuesto por técnicas de MD y aprendizaje automático enfocado a la extracción de conocimiento y que puede ser utilizado por diferentes usuarios. Este framework implementa tareas de MD y DCBD y es un software de libre distribución bajo licencia GNU, desarrollado por investigadores de la Universidad de Ljublana, Slovenia [Zupan, 2009]. Esta herramienta implementa, el método de relevancia de variables llamado ReliefF del cual se puede obtener un subconjunto de variables o el peso de cada una de ellas, pero trabaja con datos supervisados [Robnik, 2003]. De Orange, igualmente se pueden generar reglas de asociación, con el algoritmo Apriori [Agrawal, 1994] y reglas de clasificación con el algoritmo C4.5 y CN2 [Clark, 1989].

El algoritmo de selección de variables implementado trabaja con datos supervisados, es decir, requiere que se conozca la estructura de los datos, y esto determina su aplicación a datos donde previamente se identifique su distribución en grupos. Con respecto a las reglas generadas en Orange, éstas no se encuentran ponderadas. Los resultados generados por la herramienta se guardan en forma de reportes en archivos HTML y esto imposibilita su utilización para otro tipo de procesos. Esta herramienta tampoco da la posibilidad de crear un SBC con el conocimiento extraído de los datos.

1.2.2.3 Weka

Weka es una suit de librerías para la implementación y aplicación de algoritmos de MD y aprendizaje automático. Este software es de propósito general y fue desarrollado por investigadores de la Universidad de Waikato en Nueva Zelanda. Es de libre distribución ya que está bajo la licencia GNU [Mark, 2009]. Weka implementa varios métodos para generar reglas de clasificación como C4.5 y CN2, generar reglas de asociación como Relief y la relevancia de variables como algoritmos de filtro y wrapper.

Los algoritmos implementados para la relevancia de variable, se pueden utilizar para reducir la dimensionalidad de los datos o para conocer el peso diferenciante, pero la mayoría de estos métodos trabajan con datos supervisados. En cuanto a la generación de reglas, Weka implementa varios algoritmos para crear reglas de clasificación y de asociaciones.

Las reglas de asociación generadas no son del tipo de reglas de producción y no contienen un valor de certidumbre asociado a dichas reglas. Weka no genera ninguna base con el conocimiento extraído de los datos. El conocimiento generado se muestra en pantalla y da la posibilidad de guardar los resultados en una forma manual, esto es, copiando los resultados directamente de la pantalla, cosa que lo hace impráctico.

1.2.2.4 Comparación entre los software para la minería de datos

Se muestra en la Tabla 1.2 una comparación sobre las características más importantes a estudiar e estas herramientas desde el punto de vista del trabajo presentado.

Tabla 1.2. Comparación de software para minería de datos.

TITULO	TÉCNICAS	RELEVANCIA DE VARIABLES	TIPO DE REGLAS	BASE DE CONOCIMIENTO	REGLAS DE PRODUCCIÓN
[Timarán, 2009]	Análisis de asociaciones, Árboles de decisiones	Relevancia de variables manualmente por el usuario	Reglas de asociación y clasificación	✗	✗
[Zupan, 2009].	Análisis de asociaciones, Árboles de decisiones	Reducción de dimensionalidad	Reglas de asociación y clasificación	✗	✗
[Mark, 2009]	Análisis de asociaciones, Árboles de decisiones	Reducción de dimensionalidad y peso diferenciante	Reglas de asociación y clasificación	✗	✗

1.2.2 Trabajos que involucran minería de datos y sistemas basados en conocimiento

En esta sección, se presentan trabajos relacionados al descubrimiento de conocimiento a partir de base de datos que proponen un enfoque para crear una BC a partir del conocimiento extraído de los datos, el conocimiento generado es en forma de reglas.

1.2.3.1 Rule induction in constructing knowledge-based decision support

Esta investigación, se enfoca al desarrollo sustentable de países en vías de desarrollo, se utilizan técnicas de MD y SBC para crear un sistema experto con conocimiento extraído de los datos. Este proyecto, fue realizado en el Instituto de Ciencia y Tecnología Avanzada de Japón [Bao, 2002].

Este proyecto está dirigido al desarrollo sustentable y a la toma de decisiones, ya que a partir de una base de datos de un país desarrollado, de diferentes áreas, como pueden ser administración pública, desarrollo urbano y rural, cuidado de la salud o educación se genera un SE para su posterior implementación en países en vías de desarrollo. Los datos pueden ser supervisados[4] o no supervisados[5], y a partir de éstos, se generan reglas de inducción, las cuales son todas aquellas reglas que contienen el par condición - acción. Con las reglas encontradas se construyó la base de conocimiento de un SBC, el cual será utilizado por países en vías de desarrollo [Bao, 2002]. En este enfoque se utilizan dos algoritmos diferentes para generar las reglas dependiendo de los datos, si son supervisados se utiliza CABRO [Nguyen, 1999] y si no lo son se usa OSHAM [Bao, 1995]. En ambos casos, se obteniendo el mismo tipo de reglas. Una vez obtenidas las reglas, se construyen una BC con el software TESOR, que es un lenguaje para desarrollar sistemas expertos [Bao, 1996].

Los software utilizados por el autor no se encuentran implementados en una sola herramienta computacional. En cuanto a la selección de variables relevantes no se aplica ningún algoritmo de selección de variables, por tanto, no se calcula el peso de las variables. Las reglas generadas en este trabajo son reglas de clasificación, estas reglas no presentan un grado de certidumbre asociado a cada regla, como si las tienen las reglas de producción. Con las reglas generadas se crea un BC, pero está solamente contiene conocimiento

[4] Cuando los datos están agrupados en clases, y estas clases son representados por un atributo especial
[5] Cuando los datos no se presentan una agrupación

extraído de los datos y no contempla ni contribuye al conocimiento de un especialista como regularmente están constituidos los SBC.

1.2.3.2 Automatic knowledge acquisition from complex processes for the development of knowledge-based systems

Este proyecto está dirigido al tratamiento de aguas residuales de una planta, a la cual, se le realizan procesos Químicos. Se aplican técnicas de MD y SBC para crear un sistema experto para controlar dicho proceso y fue realizado por investigadores de la Universidad de Girona, Girona, España [Roda, 2001].

En esta investigación se llevan a cabo procesos químicos que son aplicados a aguas negras en tiempo real. En una primera fase se crea la BC con conocimiento adquirido de las entrevistas que se le realizaron al especialista de la planta y la revisión de la literatura disponible sobre el tema, dicha información fue estructurada en un árbol de decisiones, todo esto con el software G2 [Hangos, 2002], la forma de representación del conocimiento, que se emplea, es en reglas de clasificación. En una segunda etapa, se extrae conocimiento en forma de reglas de clasificación a partir de los datos, que contienen las características que presentó el agua diariamente, antes de obtener estas reglas, se utilizó el software Linneo de aprendizaje automático para clasificar la información en grupos, basado en el algoritmo k-means, y donde el administrador de la planta etiquetó dichos grupos [Sánchez, 1997]. Con el software GAR se unieron las reglas de clasificación de los datos con el conocimiento de la BC y también, se le dio mantenimiento a la base de conocimiento [Riaño, 1997]. Las reglas de clasificación si complementan y modifican el conocimiento que ya se encuentra incorporado en la BC. Las herramientas utilizadas en este trabajo son de libre distribución, cada uno para una tarea específica, G2 para crear el sistema experto, Lineo para encontrar grupos de los datos y GAR para incorporar las reglas de clasificación a la BC.

Las herramientas computacionales, utilizadas, no conforman un entorno común de aplicación. Con respecto a la selección de variables, el administrador de la planta seleccionó las variables, que según su criterio, eran las más relevantes, es decir, no se llevó a cabo ningún proceso automatizado para la relevancia de variables y por tanto no se obtuvo el peso diferenciante que determina, con un valor, tan importante es una variable en

un conjunto de datos específicos. Las reglas obtenidas son del tipo de reglas de clasificación las cuales no se encuentran ponderadas.

1.2.3.3 Knowledge acquisition from both human expert and data

Esta investigación presenta un enfoque para crear un SE con conocimiento extraído de los datos. Para ello, el autor utiliza técnicas de MD y SBC. Este proyecto, fue realizado por investigadores de la Universidad de Osaka, Japón [Wada, 2001].

En este trabajo se presenta un enfoque para integrar un algoritmo de aprendizaje inductivo[6] para extraer conocimiento de los datos, como lo son los árboles de decisiones, con la adquisición de conocimiento utilizando el método Ripple Down Rules, el cual construye un SBC con el experiencia de una persona para clasificar el nuevo conocimiento que se integre al sistema experto, proveniente de los datos [Richards, 2002]. Con la ayuda de un especialista se crea la BC y éste se encarga de incorporar dicho conocimiento. En cambio, si el conocimiento viene de los datos, esto se re realiza por medio del algoritmo árbol de decisiones C4.5, entonces se utiliza el principio de la longitud de descripción mínima para seleccionar la mejor hipótesis que describen los datos de acuerdo a la complejidad y así, incorporarlo a la BC del sistema experto [Grünwald, 2005].

Los programas utilizados en esta investigación son de libre distribución, pero estos no se encuentran implementados en una sola herramienta informática. Con respecto a la aplicación de algoritmos de relevancia de variables, en este proyecto no se aplica ninguno, todas las variables son seleccionadas con la misma importancia y por tal razón, no se calcula el peso de las variables. Las reglas generadas que se utilizaron para crear el sistema experto son del tipo de clasificación, las cuales no muestran un peso asociado que representa el grado de certidumbre que tiene la regla. Estas reglas de clasificación, complementan y contribuyen la BC pero el conocimiento con el que ya cuenta la BC, también proviene de los datos.

[6] Es la capacidad de obtener nuevos conceptos, más generales, a partir de ejemplos. Este tipo de aprendizaje conlleva un proceso de generalización/especialización sobre el conjunto de ejemplos de entrada.

1.2.3.4 Pronóstico para la inyección de tenso-activos en pozos de petróleo a partir de una metodología que integra técnicas de Inteligencia Artificial y Minería de Datos

Este proyecto presenta una metodología que combina técnicas de inteligencia artificial y MD para solucionar un problema referente a la inyección de tenso-activos a pozos petroleros, creando un sistema basado en conocimiento para pronosticar si se inyectan o no tenso-activos a dichos pozos. [Alonso, 2009].

La metodología presentada en este trabajo consiste en pronosticar bajo qué premisas o variables se debe inyectar-tenso activos a los pozos de petróleo.

Las variables se refieren a la composición geológica y las características del tenso-activo. Primeramente, se aplican diversos algoritmos de clasificación para obtener la estructura que guardan los datos, posteriormente, se realiza un estudio de relevancia de variables con dos objetivos principales: i) Reducir la dimensionalidad de los datos y ii) Obtener el peso diferenciante de las variables, para esto, utilizan la teoría de Test y Testores [Carrasco, 2004a], [Carrasco, 2004b] y [Cumplido. 2006]. También se utiliza el algoritmo GUHA [Hájek, 2004] de generación de hipótesis para generar reglas de asociación de los datos y descubrir las relaciones existentes entre las distintas variables y el objetivo que es el efecto resultante de la inyección de los pozos con tenso-activos. Posteriormente, la reglas obtenidas son convertidas en reglas de producción ponderadas, la ponderación de cada regla está dada por los pesos de las variables, con el fin incluirse en un sistema experto utilizando el Lenguaje de Representación del Conocimiento HAries [De la Cruz, 2002].

El trabajo presentado por Alonso es de propósito específico, relacionado con pozos de petróleo, el cual utiliza varios algoritmos de minería de datos que se encuentran implementados en diferentes herramientas computacionales. El Sistema Integral de Reconocimiento de Patrones (SIRP), fue utilizado para la clasificación de los pozos y la relevancia de variables, el algoritmo GUHA para la generación de reglas de asociación, la conversión a reglas de producción ponderadas se realiza de una forma manual a partir de las reglas de asociación y con el peso diferenciante de las variables se calcula la ponderación a estas reglas. Posteriormente, se incorporaran a la BC a través del sistema de adquisición del lenguaje.

1.2.3.5 Comparación entre los trabajos encontrados

La Tabla 1.3 muestra una comparativa de los trabajos analizados, considerando los mismos aspectos que en las tablas anteriores.

Tabla 1.3. Comparación de trabajos que generan reglas con técnicas de MD y SBC.

TRABAJO	TÉCNICAS DE MD	RV	TIPO DE REGLAS	REGLAS PONDERADAS	BC	SW
[Bao, 2002]	Árboles de decisión		Clasificación y jerarquía de conceptos			Utiliza software libre
[Roda, 2001]	Árboles de decisión y clasificación	Manualmente por el especialista	Clasificación			Utiliza software libre
[Wada, 2001]	Arboles de decisión, selección de variables		Clasificación			Utiliza software libre
[Alonso, 2009]	Análisis de asociaciones y relevancia de variables	Reduccion de dimensionalidad y peso diferenciante	Asociación y de producción			Utiliza software libre

Capítulo 2. Generación de Reglas de Producción Ponderadas en el Sucedente

2.1 Introducción

En este capítulo, se describen los pasos necesarios para generar reglas de producción ponderadas en el sucedente a través de la metodología que comprende el enfoque propuesto.

El enfoque está compuesto, principalmente, de tres procesamientos. El primero, permite calcular el peso diferenciante de las variables, lo cual representa el poder discriminante que poseen éstas para establecer la diferencia entre una clase y otra de la población bajo estudio. El segundo, permite generar reglas de asociación a partir de los datos, que no son más que patrones de comportamiento descubiertos en tales datos a los cuales también se les denomina *Hipótesis*. Por último, el tercer procesamiento es el encargado de trasformar las reglas de asociación a reglas de producción ponderadas en el sucedente utilizando el peso diferenciante de las variables para ponderar las reglas de producción.

La aplicación de cada paso de la metodología generará un conjunto de resultados, los cuales serán respaldados para su uso posterior, como Base de Conocimiento o de manera individual para un procesamiento específico. Estos respaldos se realizarán en formato texto para facilitar su reutilización o modificación en un futuro.

A continuación, se presentan los conceptos elementales y las bases teóricas para la realización del trabajo, así como la metodología propuesta para la generación de conocimiento.

2.2 Metodología para la generación de Reglas de Producción Ponderadas en el Sucedente

La metodología que se presenta, permite brindar una estrategia estructurada y secuencial para generar reglas de producción ponderadas en el sucedente a partir de un conjunto de datos. Su concepción parte del objetivo de que éstas puedan ser implementadas, posteriormente, en un sistema experto. Todo ello, con el propósito de brindar una forma de generar conocimiento a partir de los datos para ser utilizado conjuntamente con el de un especialista humano en la toma de decisión automatizada.

La metodología está integrada por cuatro módulos, tres de los cuales están dedicados a los procesamientos mencionados al inicio de este capítulo y uno para las operaciones de respaldo. Cada módulo ejecutará una tarea específica dentro del sistema y dependerá directamente de los otros, ya que los resultados de un módulo serán los datos de entrada de otro. Los módulos de la metodología se muestran en la Figura 2.1.

Figura 2.1. Módulos de la metodología para la generación de conocimiento.

Es necesario mencionar que los procesos de adquisición de los datos y de pre-procesamiento no forman parte de este trabajo, por lo que se presupone, para el desarrollo del mismo, que la matriz de datos fue adquirida y pre-procesada antes de comenzar a aplicar los procesamientos propios del enfoque que se presenta.

2.2.1 Característica de los datos en el enfoque propuesto

Los algoritmos que fueron incluidos en la metodología, trabajan sobre:

- Las *variables categóricas* las cuales están constituidas por un conjunto de características que representan categorías. Una propiedad de las categorías es que se diferencian unas de otras y son mutuamente excluyentes. Otra propiedad que pueden presentar, es que son exhaustivas, esto quiere de decir, que las categorías

para una variable debe incluir todas las posibles alternativas de variación en la variable.

- Las *variables binarias* son un tipo particular de variables categóricas. Estas variables sólo pueden tomar dos valores y generalmente, son consideradas como la presencia o ausencia de la característica, donde la negación de uno es la presencia del otro.

Las Tablas 2.1 y 2.2 muestran algunos ejemplos de estos tipos de variables.

Tabla 2.1 Ejemplos de variables categóricas.

Variable	Categoría 1	Categoría 2	Categoría 3	Categoría 4
Profesión	abogado	médico	plomero	…
Complexión	delgada	normal	sobrepeso	obeso
Estado civil	soltero	casado	viudo	divorciado

Tabla 2.2 Ejemplos de variables binarias.

Variable	Categorías
Fuma	Si/No
Tiene Hijos	Si/No
Alcohólico	Si/No

En la Tabla 2.1 se muestran tres variables categóricas, la variable Complexión presenta la propiedad mutuamente excluyente, debido a que una persona no puede presentar una complexión delgada y normal al mismo tiempo. Por otro lado, la variable Estado civil presenta la propiedad de exhaustividad, debido a que contempla todas las características del estado civil que una persona puede presentar.

Por su parte, la Tabla 2.2 muestra algunos ejemplos de variables binarias, donde los posibles valores son Si o No.

Si el problema a estudiar presenta variables numéricas, éstas deberán ser sometidas a un proceso de discretización (categorización o dicotomización) con el objetivo de establecer las categorías de todos los valores de dicha variable, esta operación pertenece a la fase de preprocesamiento de los datos en el área de *Descubrimiento de Conocimiento en Base de Datos* (DCBD) [Klösgen, 2002]. Una forma sencilla de llevar a cabo la discretizacion es definiendo rangos en los valores de la variable numérica y etiquetando cada rango con una categoría, que se identifica con un número. La Tabla 2.3 muestra un ejemplo de cómo se discretizan las variables de una matriz observacional[7] referente a estudios médicos aplicados al personal del Servicio Geológico Mexicano.

Tabla 2.3. Matriz observacional que comprende estudios médicos aplicados al personal del SGM.

SEXO	EDAD	COLESTEROL	GLUCOSA	PRESIÓN	PERIMETRO	IMC
M	*47*	*197*	*91*	*120*	*99*	*31*
M	*43*	*192*	*100*	*120*	*98*	*31*
M	*46*	*182*	*99*	*100*	*108*	*33*
M	*43*	*190*	*102*	*100*	*112*	*40*
M	*34*	*205*	*93*	*100*	*75,5*	*24*
M	*38*	*197*	*97*	*90*	*86*	*26*
M	*28*	*153*	*95*	*110*	*77*	*25*
M	*26*	*158*	*97*	*100*	*83*	*23*
M	*36*	*163*	*86*	*110*	*92*	*27*
M	*35*	*169*	*93*	*110*	*84*	*25*
M	*44*	*167*	*97*	*120*	*84*	*27*
M	*45*	*169*	*112*	*110*	*85*	*26*
M	*34*	*155*	*87*	*90*	*85*	*29*
M	*24*	*156*	*84*	*110*	*96*	*31*
M	*48*	*292*	*96*	*120*	*78*	*23*
H	*46*	*220*	*99*	*140*	*128*	*40*
H	*46*	*167*	*111*	*140*	*125*	*42*
H	*48*	*191*	*117*	*130*	*111*	*34*
H	*39*	*161*	*92*	*150*	*102*	*33*
H	*60*	*207*	*151*	*130*	*92*	*24*
H	*59*	*154*	*158*	*120*	*96*	*26*
H	*56*	*155*	*93*	*140*	*94*	*26*
H	*27*	*170*	*94*	*110*	*102*	*27*
H	*26*	*176*	*98*	*110*	*86*	*25*
H	*36*	*160*	*90*	*110*	*93*	*26*
H	*50*	*207*	*123*	*110*	*95*	*21*
H	*53*	*228*	*109*	*110*	*95*	*27*
H	*45*	*201*	*98*	*120*	*83*	*24*
H	*43*	*186*	*107*	*115*	*93*	*28*
H	*46*	*206*	*63*	*130*	*105*	*30*

El ejemplo contiene una variable binaria y seis variables numéricas, las cuales se sometieron al proceso de discretizacion. La Tabla 2.4 muestra los resultados de haber aplicado el proceso de discretización de las siete variables originales.

[7] Contiene datos recopilados sobre un estudio realizado a población que será sometido a un tipo de análisis.

Tabla 2.4. Matriz observacional con variables discretizadas.

GÉNERO	EDAD	COLESTEROL	GLUCOSA	PRESIÓN	PERÍMETRO	IMC
0 M	**1** de 24-31	**1** de 153-181	**1** de 63-82	**1** de 90-102	**1** de 75-85	**1** de 21-25
1 H	**2** de 32-38	**2** de 182-210	**2** de 83-102	**2** de 103-115	**2** de 86-96	**2** de 26-30
	3 de 39-45	**3** de 211-239	**3** de 103-122	**3** de 116-128	**3** de 97-107	**3** de 31-35
	4 de 46-52	**4** de 240-268	**4** de 123-142	**4** de 129-141	**4** de 108-118	**4** de 36-40
	5 de 53-60	**5** de 269-297	**5** de 143-162	**5** de 142-154	**5** de 119-129	**5** de 41-45

2.2.2 Módulos de la Metodología

La metodología propuesta para la generación de conocimiento en forma de reglas RPP comienza con un módulo que se encarga del cálculo del PD de las variables de un conjunto de datos, para esto se ejecuta un algoritmo de RV basado en la teoría de Testores Típicos [Lazo-Cortes, 2001].

El segundo módulo extrae, de los datos bajo estudio, RA o hipótesis, las cuales son proposiciones probabilísticas sobre la ocurrencia de ciertos variables en un conjunto de datos. El algoritmo GUHA se utilizó para la generación de hipótesis, las cuales son reglas conjuntivas y la variable de la derecha es la variable objetivo del problema a resolver y por esta razón, las reglas presentan una sola variable en el lado derecho de la regla [Hájek, 2004].

La traducción a reglas de producción ponderadas en el sucedente, es el proceso encargado de transformar las RA a RPP, de lo cual se ocupa el tercer módulo. La ponderación de cada regla será calculada con los PD de las variables que aparecen en el antecedente de cada regla de asociación, para esto, se apoya de la teoría de manejo de incertidumbre en sistemas expertos [Stefik, 1995].

El cuarto y último módulo se encarga de hacer respaldos de los resultados obtenidos con la aplicación de los procesos anteriores de la metodología. Se resguardan en archivos de texto para facilitar su reutilización posteriormente. Estos resultados son: el peso diferenciante de las variables, las reglas de asociación y la base de conocimiento conformada por las reglas de producción ponderadas en el sucedente y los nombres de las categorías de las variables.

Capítulo 3. Módulo de obtención del peso diferenciante de las variables

La Selección de Variables forma parte del DCBD y es una pieza fundamental en cualquier estudio sobre un conjunto de datos porque permite conocer la relevancia que tienen las variables o atributos, y así, poder saber qué variables son más importantes en el estudio de una colección de datos [Hernández, 2004].

La selección de variables no sólo se enfoca a la relevancia de atributos sino que también se pueden resolver problemas como reducir el tamaño de los datos, esto es, eliminar variables o atributos que puedan ser irrelevantes o redundantes. Una buena selección de variables puede mejorar la calidad del modelo, al permitir que el método de minería de datos se centre en las variables más importantes pudiendo expresar el modelo resultante en función de menos variables.

La RV es una técnica de MD que permite cuantificar la importancia que tienen las variables o atributos, en dicho conjunto de datos, la cuantificación representa la importancia que tiene está con respecto a los datos.

En este trabajo, se considera sólo la relevancia de variables, ya que para este enfoque es suficiente con cuantificar la importancia que tienen las variables en los datos a estudiar. Aunque en un futuro, puede incluirse la reducción de dimensionalidad.

3.1 Relevancia de variables

La relevancia de las variables es un valor numérico al cual se le denomina *Peso de la Variable*. Este valor está comprendido en el rango entre 0 y 1, inclusive, lo cual significa que si el peso es igual a 1, la variable es muy importante y por lo tanto, es imprescindible. En cambio, si el peso es 0, indica que la variable no tiene importancia y por lo tanto, es prescindible. Luego, valores entre 0 y 1, representan gradaciones de la importancia de las variables, en el mismo sentido de los valores extremos.

El estudio de RV se puede realizar desde diferentes puntos de vista, uno es la búsqueda de variables más representativas y la otra es la búsqueda de variables más diferenciantes ente clases. Para la primera búsqueda, se calcula el *Peso Informacional* y para la segunda, se

calcula el *Peso Diferenciante* [Santos, 2004]. Dependiendo de lo que se quiera obtener y de cómo están estructurados los datos, se utiliza una u otra aproximación

El peso informacional es la capacidad que tiene una variable para establecer diferencias entre individuos de una misma clase, mientras que el peso diferenciante es el poder discriminante que tiene una variable a la hora de establecer si un individuo pertenece a una clase u otra, para este último caso, se asume que los individuos analizados pertenecen a clases diferentes.

La razón por la que en esta tesis se calcula el PD de las variables está dada porque el algoritmo busca aquellas variables de mayor poder diferenciante a la hora de establecer si un individuo pertenece a una clase u otra, encontrando las relaciones que tienen las otras variables y la variable que clasifica a los datos, la cual es el objetivo del problema que se pretende resolver.

3.2 Datos de entrada al algoritmo

El algoritmo de relevancia de variables utilizado para este enfoque es el llamado *Testores Típicos* [Lazo-Cortes, 2001], a través del cual se puede calcular el peso diferenciante de las variables. Esté método es un algoritmo clásico de Reconocimiento de Patrones enfocado a la lógica combinatoria [Carrasco, 2002], [Somuano, 2004].

El algoritmo utiliza para su procesamiento una matriz polivalente X o conjunto de datos de m filas (individuos o instancias) denotado por I y n columnas (variables o atributos) denotado por V, la cual debe estar clasificada en l números de clases o grupos. El ejemplo de la matriz observacional presentada en la Tabla 2.4 se muestra en toda su extensión en la Tabla 3.1. Es de notar que las variables de esta matriz cumplen con el requisito impuesto por el algoritmo del cálculo de la relevancia, el cual se refiere a que las mismas tienen que ser de tipo cualitativas o binarias, esto es, la matriz debe ser polivalente.

La matriz polivalente de la Tabla 3.1 tiene siete variables, las cuales fueron discretizadas previamente. La variable GÉNERO presenta dos categorías con los valores 0 y 1, indicando 0 para la mujer y 1 para el hombre. Las variables restantes se discretizarón en cinco categorías, enumeradas a partir del uno hasta el cinco, que son los valores que aparecen en dichas columnas de la Tabla 3.1.

Tabla 3.1. Matriz polivalente del estudio médico.

GÉNERO	EDAD	COLESTEROL	GLUCOSA	PRESIÓN	PERÍMETRO	IMC
0	4	2	2	3	3	3
0	3	2	2	3	3	3
0	4	2	2	1	4	3
0	3	2	2	1	4	4
0	2	2	2	1	1	1
0	2	2	2	1	2	2
0	1	1	2	2	1	1
0	1	1	2	1	1	1
0	2	1	2	2	2	2
0	2	1	2	2	1	1
0	3	1	2	3	1	2
0	3	1	3	2	1	2
0	2	1	2	1	1	2
0	1	1	2	2	2	3
0	4	5	2	3	1	1
1	4	3	2	4	5	4
1	4	1	3	4	5	5
1	4	2	3	4	4	3
1	3	1	2	5	3	3
1	5	2	5	4	2	1
1	5	1	5	3	2	2
1	5	1	2	4	2	2
1	1	1	2	2	3	2
1	1	1	2	2	2	1
1	2	1	2	2	2	2
1	4	2	4	2	2	1
1	5	3	3	2	2	2
1	3	2	2	3	1	1
1	3	2	3	2	2	2
1	4	2	1	4	3	2

La matriz polivalente X, Tabla 3.1, tiene un conjunto de variables que en esta teoría se representa como un vector de las variables en la forma $\vec{V} = \{v_1, v_2, \dots, v_n\}$. Esta notación será utilizada a continuación para presentar los conceptos de *Testores* y *Testores Típicos*.

Si se denota con la letra τ al subconjunto de las variables de la matriz X, esto es, $\tau \subseteq \vec{V}$ y se representa como $\tau = \{v_1, \dots, v_s\}$, donde la cantidad de variables es menor que n.

Entonces, se puede decir que τ es Testor para la matriz polivalente X si al eliminar todas las columnas que no están en τ, la submatriz generada por τ, no tiene individuos iguales entre clases diferentes, o lo que es lo mismo, se mantiene la diferencia entre individuos de clases distintas.

Algunos Testores posibles para la matriz polivalente de la Tabla 3.1 son mostrados a continuación.

τ_1 = {GÉNERO COLESTEROL GLUCOSA PRESIÓN PERÍMETRO IMC}
τ_2 = {GÉNERO EDAD GLUCOSA PRESIÓN PERÍMETRO IMC}
τ_3 = {GÉNERO EDAD COLESTEROL PRESIÓN PERÍMETRO IMC}
τ_4 = {GÉNERO EDAD COLESTEROL GLUCOSA PRESIÓN IMC}
τ_5 = {GÉNERO EDAD COLESTEROL GLUCOSA PERÍMETRO IMC}
τ_6 = {GÉNERO EDAD COLESTEROL GLUCOSA PRESIÓN PERÍMETRO}
τ_7 = {GÉNERO GLUCOSA PRESIÓN PERÍMETRO IMC}
τ_8 = {GÉNERO COLESTEROL PRESIÓN PERÍMETRO IMC}
τ_{36} = {GÉNERO EDAD COLESTEROL GLUCOSA}
τ_{37} = {GÉNERO PERÍMETRO IMC}
τ_{38} = {GÉNERO PRESIÓN IMC}
τ_{39} = {GÉNERO PRESIÓN PERÍMETRO}
τ_{40} = {GÉNERO COLESTEROL IMC}
τ_{41} = {GÉNERO EDAD PERÍMETRO}
τ_{42} = {GÉNERO EDAD IMC}

Los Testores encontrados pueden contener, a su vez, otros Testores, como es el caso de τ_4 que contiene a τ_{36} y el Testor τ_1 contiene a τ_{37}.

Por su parte, un Testor Típico es un Testor para la matriz polivalente X, si ningún subconjunto propio de τ es un Testor, es decir, que τ no contenga a otros Testores.

De los Testores presentados anteriormente, los siguientes son Testores Típicos.

τ_{36} = {GÉNERO EDAD COLESTEROL GLUCOSA}
τ_{37} = {GÉNERO PERÍMETRO IMC}
τ_{38} = {GÉNERO PRESIÓN IMC}
τ_{39} = {GÉNERO PRESIÓN PERÍMETRO}

τ_{40} = {GÉNERO COLESTEROL IMC}

τ_{41} = {GÉNERO EDAD PERÍMETRO}

τ_{42} = {GÉNERO EDAD IMC}

3.3.1 Cálculo del Peso Diferenciante

El peso diferenciante se define como la relación que existe entre el número de apariciones de la variable en los Testores Típicos y el número total de Testores Típicos encontrados en la matriz polivalente, esto es, la ocurrencia de la variable en el total de los Testores Típicos. Esta relación se expresa como sigue:

$$P_D(v_j) = TO_j/TO$$

Donde:

$P_D(v_j)$ es el peso diferenciante de la variable v_j,

TO_j es el número de Testores Típicos en los que aparece la variable j y

TO es el número de Testores Típicos obtenidos de la matriz polivalente X.

De acuerdo a la expresión, el peso diferenciante de cada variable de la matriz polivalente de la Tabla 3.1, se muestra en la Tabla 3.2.

Tabla 3.2 Peso diferenciante de cada variable.

Variable	Peso Diferenciante
GÉNERO	1
IMC	0.571
EDAD	0.428
PERÍMETRO	0.428
PRESIÓN	0.285
COLESTEROL	0.285
GLUCOSA	0.143

La variable que presentó el PD más alto fue la variable GÉNERO, esto quiere decir que es más importante para el conjunto de datos y que divide a la población en sus dos categorías

que son en mujer y hombre. Otras variables, que presentan un alto poder discriminante entre clases y que en menor medida dividen a la población, son las variables IMC, EDAD y PERÍMETRO.

3.4 Implementación del Algoritmo

La teoría de Testores Típicos es muy útil para hacer selección de variables a partir de los datos, incluso permite datos incompletos. Sin embargo, la complejidad del cálculo de los Testores Típicos en una matriz polivalente es tal que tiene un crecimiento exponencial en correspondencia con el número de variables que presenta la matriz. Todo ello se debe al número de subconjuntos y las comparaciones que se tiene que hacer el algoritmo para descubrir los Testores Típicos existentes.

Por esta razón, se han desarrollado diferentes enfoques para calcular los Testores Típicos en una matriz de datos, algunos de ellos son: algoritmos heurísticos, basados en procesamiento paralelo y distribuido [Cumplido. 2006], [Carrasco, 2003], [Carrasco, 2004a] y [Vega-Alvarado, 2001].

Un algoritmo frecuentemente usado para encontrar Testores Típicos es el algoritmo BT, el cual no trabaja directamente con la matriz polivalente. En su lugar, trabaja con una matriz resultante de un procesamiento aplicado a la matriz polivalente llamada *Matriz Básica* (MB) [Sánchez, 2002] y [Torres, 2006].

Los pasos que se deben seguir para generar la MB a partir de la matriz polivalente, se describen a continuación:

Paso 1.- Se crea una *Matriz de Diferencias* (MDI) como resultado de la comparación de un individuo de la clase bajo análisis con los individuos del resto de las clases, el criterio de comparación utilizado está dado por:

$$C_k\big(v_k(I_i),\quad v_k(I_j)\big) = \begin{cases} 0 & si\ v_k(I_i) = v_k(I_j) \\ 1 & si\ v_k(I_i)\ !\ = v_k(I_j) \end{cases}$$

Paso 2.- Se crea la MB, a partir de la MDI, con base a los siguientes criterios:

Criterio 1.- Sea un individuo â $= (a_i \cdots \ a_n)$ y ĉ $= (c_i \cdots \ c_n)$, dos filas de la MDI. Se dice que â es *subfila* de ĉ o que ĉ es *superfila* de â, si se cumple que $\hat{a}_i \leq c_i, para \ \forall \ i = 1, \ldots, m.$

Criterio 2.- La fila â se denomina *Básica* si no existe fila alguna que sea subfila de â, es decir, si no es superfila de otra.

Entonces, se establece que una MB está formada por todas las filas básicas de la MDI.

A continuación, se presenta un ejemplo del cálculo de la MB, a partir de la matriz polivalente Z de la Tabla 3.3. La matriz polivalente se encuentra clasificada en dos grupos, los tres primeros individuos pertenecen a la clase C_0 y los individuos restantes a la segunda clase denotada por C_1.

Tabla 3.3. Matriz polivalente Z.

Z	V_1	V_2	V_3	V_4	V_5	
I_1	1	1	0	1	1	C_0
I_2	0	1	0	1	1	
I_3	1	0	1	0	1	
I_4	1	1	1	1	1	C_1
I_5	0	0	0	1	1	

Al realizar las comparaciones entre individuos de diferentes clases, se genera la matriz de diferencias que se muestra en la Tabla 3.4, aplicando el paso 1, mencionado anteriormente.

Tabla 3.4. Matriz de diferencias de Z.

MDI	V_1	V_2	V_3	V_4	V_5
I_{14}	0	0	1	0	0
I_{15}	1	1	0	0	0
I_{24}	1	0	1	0	0
I_{25}	0	1	0	0	0
I_{34}	0	1	0	1	0
I_{35}	1	0	1	1	0

La Tabla 3.5 presenta la MB, que se generó utilizando la MDI y aplicando el paso 2. La MB será utilizada por el algoritmo BT para obtener los Testores Típicos.

Tabla 3.5. Matriz Básica de la matriz polivalente Z.

MB	V_1	V_2	V_3	V_4	V_5
a_1	0	0	1	0	0
a_2	0	1	0	0	0

A continuación, se muestra el algoritmo BT que se encarga de encontrar los Testores Típicos a partir de una MB.

INICIO Algoritmo_BT(MB)

GENERAR la primera lista â no nula de longitud m, (0, 0 …, 1)

MIENTRAS â no sea una lista unitaria (1,1 …,1) REALIZAR

SI â es una lista Testor en MB según propiedad_1 ENTONCES

Aplicar propiedad_3

SI a es una lista Testor Típico ENTONCES

IMPRIMIR a

Aplicar la propiedad_2

DE LO CONTRARIO

Aplicar propiedad_4

GENERAR la lista siguiente a la obtenida

FIN MIENTRAS

FIN Algoritmo_BT

A continuación se describen la función GENERAR y las propiedades que aparecen en el algoritmo BT.

Función GENERAR

Primero se crea un vector auxiliar llamado lista Testor y se representa como $\hat{\alpha} = (\alpha_i \dots \ \alpha_m)$ y los valores que puede tener esta lista está dada por $\hat{\alpha} \in \{0, \ 1\}$, donde un valor igual a 0, es decir $\hat{\alpha} = 0$, indica que la variable no está presente y un valor igual a 1, es decir, $\hat{\alpha} = 1$, quiere decir que está presente.

Si â $= (a_i \ldots \ a_m)$ es una fila básica, se tiene:

$$a_j \wedge \alpha j = \begin{cases} 1 & Si\ a_i = 1\ y\ \alpha_j = 1 \\ 0 & Si\ a_i = 0\ y\ \alpha_j = 0 \end{cases}$$

Las propiedades que utiliza el algoritmo BT, se enlistan a continuación:

1. Si se encuentra una fila tal que $a_j \wedge \alpha j = 0$ para todo j, entonces, la lista $\hat{\alpha}$ no es un Testor.
2. Si $\hat{\alpha} = (\alpha_i \ldots \ \alpha_m)$es una lista Testor Típico y k es el subíndice del último, las siguientes $2^{m-k} - 1$ listas no son Testor Típicos (en el orden natural).
3. Si $\hat{\alpha} = (\alpha_i \ldots \ \alpha_m)$es una lista Testor y k es el subíndice del último 1, entonces los siguientes $2^{,m-k} - 1$ son listas Testor pero no Testores típicos.
4. Si â $= (a_i \ldots \ a_m)$ es una fila básica y k el subíndice del último 1 en â, supongamos que $\hat{\alpha} = (\alpha_i \ldots \ \alpha_m)$no es lista Testor y es la primera de las listas no nulas que cumple con el orden natural la condición $a_j \wedge \alpha j = 0$ para todo $j = 1, \ldots, m$. Entonces, las siguientes $2^{,m-k} - 1$ listas tampoco son listas Testor.

De la MB de la Tabla 3.5 se obtiene el Testor Típico conformado por las variables $V_2\ y\ V_3$, el cual se muestra en la Tabla 3.6, con los datos de la matriz polivalente de la Tabla 3.3. Se puede observar que no existen individuos entre las clases C_0 y la clase C_1.

Tabla 3.6. Testor Típico para la matriz polivalente Z.

Z	V_2	V_3	
I_1	1	0	C_0
I_2	1	0	
I_3	0	1	
I_4	1	1	C_1
I_5	0	0	

Capítulo 4. Obtención de Reglas de Asociación

En la segunda fase de la metodología propuesta, se encuentra la generación de reglas de asociación. Este proceso da la posibilidad de generar hipótesis a partir de un conjunto de datos, proporcionado por el usuario.

El análisis de asociaciones es una técnica importante dentro de la MD, gracias a que permite generar reglas de asociación, las cuales expresan patrones de comportamiento entre los datos en función de la aparición conjunta de valores de dos o más variables, es decir, expresan las combinaciones de valores de las variables que suceden más frecuentemente [Hernández, 2004].

Dicho de otra forma, una regla de asociación es una proposición probabilística sobre la ocurrencia de ciertas relaciones en una base de datos. Generalmente, estas reglas tienen asociadas ciertas medidas que garantizan la calidad de dichas reglas, lo cual se valida a partir de un umbral o valor límite establecido para dichas medidas. El valor del umbral puede ser proporcionado por el usuario o puede estar implícito en la implementación del algoritmo.

4.1 Método GUHA

El algoritmo utilizado para la generación de asociaciones es el método GUHA, el cual es un método de MD que genera automáticamente hipótesis a partir de los datos. El algoritmo GUHA busca las dependencias, asociaciones, o relaciones entre las variables estudiadas y las presenta en forma de reglas [Hájek, 2004]. Este algoritmo puede procesar conjuntos de datos con cientos o miles de individuos y decenas de variables. Las hipótesis creadas son interesantes con respecto al punto de vista de los datos y del estudio del problema.

El método GUHA trabaja con un conjunto de datos que tiene m número de individuos y con n número de variables y ésta es representada en forma de matriz denotada por:

$$\{X_{ij}\} \quad \begin{matrix} j = 1, \dots, n \\ i = 1, \dots, m \end{matrix}$$

donde $x_{ij} = f_j(i)$, es el valor de la $j - esima$ variable para el $i - esimo$ individuo. Cuando las variables son binarias éstas sólo presentan dos valores 0 y 1 y cuando son categóricas, presentan un valor entero que representa la categoría.

El algoritmo implementado trabaja con variables binarias y variables categóricas. También, trabaja con datos que tengan una estructura de grupo o datos supervisados, es decir, se selecciona una variable objetivo para su estudio, la cual clasifica los datos según las categorías que ella presenta.

GUHA genera varios tipos de reglas, una de ellas es de la forma $A \sim B$ y se describe como: A está asociado con B. Otro tipo de regla que genera este método es $A \rightarrow B$ que representa la relación A es una causa de B (casi todos los objetos que satisfacen A también satisfacen B).

Cabe señalar que las reglas que se utilizan en esta aproximación y las cuales serán generadas a partir de la aplicación del método GUHA, son del tipo de reglas de asociación porque nos permite identificar las ocurrencias entre variable asociadas a un variable objetivo.

4.1.1 Notación del algoritmo

La notación que utiliza el método GUHA es:

- **Predicado**.- Es identificado por el nombre de la variable o por cada una de las categorías que tiene la variable
- **Literal**.- También llamada *Fórmula Atómica* y representa un predicado o un predicado negativo.
- **Fórmula Abierta**.- Es construida a partir de predicados usando los conectivos lógicos: $\&, v, \longrightarrow, -$ (conjunción, disyunción, implicación y negación). El cumplimiento de una fórmula se representa con la letra griega Fi y está denotada como $\varphi(\chi)$ para $u \in X$, donde u es un individuo de la matriz X y se define por el cumplimiento de las literales usando la tabla de verdad de la lógica clásica.
- **Cuantificadores**.- Están representados por los símbolos $\forall$, $\forall_p$, $\exists, \sim y \rightarrow$ llamados *Cuantificador universal*, *casi universal*, *existencial*, *asociación* e *implicación*.

Los cuantificadores se dividen en dos tipos:

- Tipo 1.- El cuantificador universal, casi universal y existencial, y se aplican a una sola fórmula abierta.
- Tipo 2.- Este tipo de cuantificadores son aplicados a pares de fórmulas abiertas, utilizando la implicación y la asociación.

4.1.2 Tabla de Contingencia

Cada par de fórmulas abiertas $\varphi(\chi)$ y $\psi(\chi)$ determina cuatro números importantes llamadas frecuencias:

- a es el número de individuos satisfechos por la operación $\varphi(\chi)\ \&\ \psi(\chi)$ en X
- b es el número de individuos satisfechos por la operación $\varphi(\chi)\ \&\ -\psi(\chi)$ en X
- c es el número de individuos satisfechos por la operación $-\varphi(\chi)\ \&\ \psi(\chi)$ en X
- d es el número de individuos satisfechos por la operación $-\varphi(\chi)\ \&\ -\psi(\chi)$ en X

A estas cuatro frecuencias, usualmente, se les llama *Tabla de Contingencia*, la cual se representan como se muestra en la Tabla 4.1.

Tabla 4.1. Tabla de Contingencia.

a	b	R
c	d	S
k	l	M

Donde r, s, k, l y m son condiciones relacionadas a las fórmulas abiertas $\varphi(\chi)$ y $\psi(\chi)$ y se deben de cumplir para decir que una expresión $\varphi(\chi)\ \&\ \psi(\chi)$ es verdadera. A continuación se presenta la forma de obtener estos valores.

$$r = a + b$$

$$s = c + d$$

$$k = a + c$$

$$l = b + d$$

$$m = a + b + c + d$$

4.2 Reglas de asociación generadas por GUHA

Recurriendo al ejemplo presentado en el capítulo tres referente a estudios médicos aplicados al personal del Servicio Geológico Mexicano. Se ha utilizado la matriz polivalente de este estudio presentado en la Tabla 3.1 para generar varias reglas de asociación que sirvan como ejemplo en esta sesión.

Para este caso, se seleccionó como variable objetivo la variable ESTADO DE SALUD la cual presenta cuatro categorías que son mostradas en la Tabla 4.2.

Tabla 4.2. Categorías de la variable objetivo ESTADO DE SALUD.

Categoría	Etiqueta	Descripción
1	Mujeres No saludables	Individuos que presentan valores altos en EDAD, IMC, PERÍMETRO ABDOMINAL, PRESIÓN y son mujeres
2	Mujeres saludables	Individuos que presentan valores aceptables en EDAD, IMC, PERÍMETRO ABDOMINAL, PRESIÓN y son mujeres
3	Hombres No saludables	Individuos que presentan valores altos en EDAD, IMC, PERÍMETRO ABDOMINAL, PRESIÓN y son hombres
4	Hombres Saludable	Individuos que presentan valores aceptables en EDAD, IMC, PERÍMETRO ABDOMINAL, PRESIÓN y son hombres

Los datos son clasificados, dependiendo de los valores de las variables GÉNERO, EDAD, IMC, PERÍMETRO ABDOMINAL, PRESIÓN, estas variables, obtuvieron valores altos en el estudio de los pesos diferenciantes.

A continuación, en la Tabla 4.3 se presenta la matriz polivalente de la Tabla 3.1, pero agregando la variable objetivo ESTADO DE SALUD, la cual clasifica los datos, en los valores presentados en la Tabla 4.2.

Tabla 4.3. Matriz polivalente del estudio médico, agrupada según la variable ESTADO DE SALUD.

GÉNERO	EDAD	COLESTEROL	GLUCOSA	PRESIÓN	PERÍMETRO	IMC	ESTADO DE SALUD
0	4	2	2	3	3	3	1
0	3	2	2	3	3	3	1
0	4	2	2	1	4	3	1
0	3	2	2	1	4	4	1
0	2	2	2	1	1	1	2
0	2	2	2	1	2	2	2
0	1	1	2	2	1	1	2
0	1	1	2	1	1	1	2
0	2	1	2	2	2	2	2
0	2	1	2	2	1	1	2
0	3	1	2	3	1	2	2
0	3	1	3	2	1	2	2
0	2	1	2	1	1	2	2
0	1	1	2	2	2	3	2
0	4	5	2	3	1	1	2
1	4	3	2	4	5	4	3
1	4	1	3	4	5	5	3
1	4	2	3	4	4	3	3
1	3	1	2	5	3	3	3
1	5	2	5	4	2	1	4
1	5	1	5	3	2	2	4
1	5	1	2	4	2	2	4
1	1	1	2	2	3	2	4
1	1	1	2	2	2	1	4
1	2	1	2	2	2	2	4
1	4	2	4	2	2	1	4
1	5	3	3	2	2	2	4
1	3	2	2	3	1	1	4
1	3	2	3	2	2	2	4
1	4	2	1	4	3	2	4

Algunas de las reglas de asociación generadas por el método, se muestran a continuación:

RA 1: GÉNERO (0) & COLESTEROL (1) ~ ESTADO DE SALUD (2)

RA 2: GÉNERO (0) & PERÍMETRO (1) ~ ESTADO DE SALUD (2)

RA3: GÉNERO (1) & PERÍMETRO (2) ~ ESTADO DE SALUD (4)

RA 4: GÉNERO (1) & IMC (2) ~ ESTADO DE SALUD (4)

RA5: GÉNERO (1) & PRESIÓN (2) ~ ESTADO DE SALUD (4)

RA 6: GÉNERO (0) & IMC (2) ~ ESTADO DE SALUD (2)

RA 7: EDAD (5) & GÉNERO (1) ~ ESTADO DE SALUD (4)

RA 8: GÉNERO (1) & IMC (1) ~ ESTADO DE SALUD (4)

RA 9: EDAD (2) & PRESIÓN (1) ~ ESTADO DE SALUD (2)

RA 10: GÉNERO (1) & PRESIÓN (3) ~ ESTADO DE SALUD (4)

RA 11: GÉNERO (1) & IMC (3) ~ ESTADO DE SALUD (3)

RA 12: EDAD (4) & PERÍMETRO (5) ~ ESTADO DE SALUD (3)

RA 13: GÉNERO (0) & EDAD (3) ~ ESTADO DE SALUD (1)

RA 14: PERÍMETRO (1) & GÉNERO (0) ~ ESTADO DE SALUD (2)

Analizando la regla número 1, indica que si la categoría de la variable GÉNERO es 0 (mujer), y el nivel de COLESTEROL tiene categoría 1 (Bajo, esto es, entre 151 y 181), entonces estos valores se asocian a la categoría 2 (Mujer saludable) de la variable ESTADO DE SALUD.

Analizando la regla número 5, indica que si la categoría de la variable GÉNERO es 1 (hombre), y el nivel de PRESIÓN tiene categoría 2 (Bajo, esto es, entre 103 y 115), entonces estos valores se asocian a la categoría 4 (Hombre saludable) de la variable ESTADO DE SALUD.

En la lista de RA presentadas, se puede observar que la variable que aparece más frecuentemente en el lado derecho de las reglas es la variable GÉNERO y en menor medida, las variables IMC, EDAD, PERÍMETRO y las variables que casi no aparecen son COLESTEROL y GLUCOSA. La aparición de cada una de las variables tiene una correspondencia con los PD a cada variable presentado en la Tabla 2.3. Esto se debe a que los datos son los que indican su comportamiento y las relaciones que tienen las variables, además, son los datos que aparecen más frecuentemente en el conjunto de datos.

Capítulo 5. Traducción a Reglas de Producción Ponderadas en el Sucedente

Los Sistemas Basados en Conocimiento o también nombrados Sistemas Expertos son una técnica de gran importancia en el área de IA, debido a que tienen la capacidad de dar solución a problemas donde se cuente con suficiente conocimiento especializado sobre el tema [Russell, 2003]. Los SBC son sistemas informáticos que usan el conocimiento del experto y los procedimientos de inferencia para resolver problemas con un grado de dificultad mayor a los sistemas convencionales.

Este tipo de sistemas contienen conocimiento, de un especialista humano, almacenado en un repositorio que será utilizado posteriormente por el mismo sistema, además, puede brindar conclusiones, tomar decisiones y dar una explicación sobre los resultados alcanzados. Todas estas posibilidades son agrupadas en diferentes componentes, que son: la Base de Conocimiento, el Motor de Inferencia y el Sistema Explicatorio.

Una BC es un repositorio de conocimiento proveniente de la experiencia de un especialista humano referente a un problema particular y que será utilizado por el SBC. Este repositorio se compone de hechos y relaciones entre éstos [Stefik, 1995]. Los hechos son sucesos que describen el problema y al igual que las relaciones son representadas en la computadora en alguna forma o *Estructura de Representación del Conocimiento* (ERC). Una ERC es una manera de representar la realidad, en forma de símbolos para ser tratado computacionalmente y someterla a algún proceso de cómputo [Pajares, 2006].

El motor de inferencia, es la parte que razona su manera para la solución de problemas haciendo búsquedas a través el contenido de la BC. También, un motor de inferencia tiene la capacidad de representar y guardar resultados intermedios, administrar la memoria y los recursos computacionales [Stefik, 1995].

El sistema explicatorio brinda una explicación del cómo y por qué se ha llegado a una determinada conclusión, diciendo qué reglas se han usado, en qué orden, etc. [Pajares, 2006].

Para representar el conocimiento se han utilizado muchas ERP, con el objetivo de codificarlo manipularlo en una computadora. Algunas de las ERP que se han diseñado para formalizar el conocimiento son [Pajares, 2006]:

- Lógica proposicional: Forma clásica y la más básica de representar el conocimiento, cada proposición o hecho es representado con símbolos y relacionados con conectivos lógicos como: conjunción, disyunción. implicación o negación, y utilizando reglas de inferencia como el Modus Ponens.
- Lógica de predicado: Otra forma clásica de representar el conocimiento donde se introducen cuantificadores, que permiten hacer predicados sobre las proposiciones. Los cuantificadores utilizados son:
    - Cuantificador existencial ∃
    - Cuantificador universal ∀
- Reglas: Son la forma más común de representar el conocimiento en una computadora. El antecedente está integrado por uno o más hechos, este último caso, unidos por un operado de conjunción y el consecuente se compone de un solo hecho. Estas reglas establecen una relación entre el antecedente y el consecuente para generar nueva información o probar la autenticidad de una sentencia.
- Marcos: Es una colección de atributos que define el estado de un objeto y su relación con otros marcos. Los marcos son organizados jerárquicamente permitiendo establecer un mecanismo de herencia y es lo que constituye su sistema de inferencia.
- Guiones: Representan una secuencia de acciones unidas entre sí por una relación de causalidad.
- Redes semánticas: Son utilizadas para definir el significado de un concepto mediante su relación con otros conceptos, se representan en forma de grafos donde los conceptos son los nodos y los enlaces definen las relaciones entre ellos.
- Lógica de incertidumbre: Trata con la información cuando presenta incertidumbre, vaguedad o está incompleta. Algunos campos que han tratado la incertidumbre son:

- Teoría estadística
- Lógica difusa

5.1 Estructura de Representación a utilizar

La estructura de representación de conocimiento que se utiliza en este enfoque es del tipo de reglas de producción, ya que es una de las ERC más utilizadas en el área de los SBC y representan lo más cercano a la forma en que razonan los seres humanos. Además, casi todos los sistemas actuales, incluyendo los híbridos, utilizan reglas de producción.

Las reglas de producción son representadas, de manera general, en la forma:

Si antecedente **Entonces** consecuente

Donde:

Antecedente o las premisas son el conjunto de condiciones que se deben cumplir para evaluar la regla.

Consecuente es una acción que se derivan como respuesta al cumplimiento del antecedente.

El antecedente son los hechos que se encuentran en la BC y puede estar constituido por un hecho o más relacionados por medio de los operadores lógicos conjunción (∧) ó disyunción (∨). De igual forma, el consecuente es representado por un solo hecho, Un ejemplo de esto:

Si $C_{1v1} \wedge C_{1v2}$ **Entonces** C_{1vo}

Donde:

C_{1v1} y C_{1v2}, son una de tantas categorías que pertenecen a sus respectivas variables. Por ejemplo, la variable GÉNERO tiene dos categorías, las cuales son: mujer y hombre. Las categorías que aparecen en el antecedente nunca aparecerán en el consecuente.

C_{1vo} es una de todas las categorías que pertenecen a la variable objetivo. Por ejemplo, la variable ESTADO DE SALUD y sus categorías son: mujer no saludable, mujer saludable, hombre no saludable y hombre saludable.

5.2 Conceptos previos

Las reglas de producción ponderadas en el sucedente, tienen la siguiente sintaxis:

$$A \Rightarrow S(P_C, P_I)$$

Donde:

A es el Antecedente de la regla y puede ser proposición simple o compuesta, esta última representada como Conjunción Elemental.

S es el Sucedente de la regla, el cual consta de una proposición simple que no aparece en el antecedente de ninguna regla.

P_C *y* P_I son los pesos para el cumplimiento y el incumplimiento del antecedente, respectivamente y representan el grado de certidumbre que cuantifica la relación de producción establecida por el símbolo ⇒ entre el antecedente y el sucedente de la regla.

Estas reglas siguen el formato de la Reglas de Producción Generalizadas (RPG) del Lenguaje de Representación del Conocimiento HAries [De la Cruz, 1996], cuyas proposiciones pueden aparecer tanto en el antecedente como en el sucedente. De acuerdo a su aparición en estos dos lugares de la regla, las proposiciones toman como atributo natural una de las siguientes categorías: *Pregunta*, *Intermedio* u *Objetivo*. Donde:

- Una proposición es pregunta si ésta se encuentra en el antecedente de una regla pero no aparece en el sucedente de ninguna.
- Una proposición es intermedio si forma parte tanto del antecedente como el sucedente de alguna regla, pero no de la misma.
- Una proposición es objetivo si aparece en el sucedente de al menos una regla pero nunca aparece en el antecedente de ninguna regla.

Cabe aclarar que las reglas que se generan con el enfoque propuesto tienen la característica particular de que no cuentan con proposiciones intermedio. Esto debido, a que la categoría de la variable objetivo, solamente aparece en el lado derecho de la regla de asociación. Por este motivo, nunca se encontrará una variable objetivo en ambas partes de las reglas de producción ponderadas en el sucedente.

5.3 Traducción de las reglas de asociación

Las RA generadas, partiendo de los datos, son sometidas a un proceso de traducción para generar reglas de producción ponderadas en el sucedente. En esta sección se describe este proceso.

Las reglas de producción se generan a partir de las RA obtenidas por la aplicación del método GUHA, como se explicó en el capítulo 4. A continuación, se describen los pasos para convertir RA a reglas de producción.

1. Las categorías de las variables que aparecen en las RA, tanto en el antecedente como en el sucedente, se mantienen igual.
2. El operador AND de la RA denotado por el símbolo & se mantiene igual para las reglas de producción.
3. El operador de asociación representado por $\sim$ en la RA, es remplazado por el símbolo $\Rightarrow$ para las reglas de producción.

Los pasos anteriores se pueden ver reflejados en la siguiente línea.

$$A \,\&\, B \sim S \;\; \boldsymbol{transformada\ a} \;\; A \,\&\, B \Rightarrow S$$

Llevado a cabo el proceso de conversión de las reglas de asociación a reglas de producción, es necesario ponderarlas de acuerdo al PD de cada variable, que representa la contribución de cada antecedente al sucedente de la regla.

5.4 Cálculo del peso del sucedente de las reglas de producción ponderadas en el sucedente

El peso que se asigna al sucedente es la contribución del cumplimiento del antecedente, el cual será obtenido del PD de las variables que intervienen en el antecedente.

El peso de las reglas de producción ponderadas tiene un valor entre -1 y 1, inclusive. Donde un valor de: 1 indica verdad absoluta, -1 falsedad absoluta, 0 desconocimiento total y el resto de los valores están asociados a gradaciones de la veracidad, correspondientes a la afirmación o negación, respectivamente.

En este enfoque, únicamente se calcula el peso de cumplimiento P_C, el cual está asociado al cumplimiento del antecedente, esto se debe a que, las RA describen cómo se asocia el cumplimiento de una condición a la presencia de las categorías en el antecedente y no al incumplimiento de dicha condición. Por lo tanto, el peso de incumplimiento P_I, no se calculará en esta aproximación, pero se le asignará un valor de 0 indicando con esto que se desconoce el valor del P_I para el incumplimiento de la regla.

El tratamiento de la incertidumbre en los SE, se han desarrollado muchos otros métodos, algunos métodos están basados en probabilidades, redes bayesianas, otros en factores de certeza y más recientemente en la teoría de la lógica difusa [Klir, 1995], [Pajares, 2006]. La teoría que sustenta la estrategia de contribución utilizada en este enfoque, se basa en la lógica de la incertidumbre, y más precisamente, en la lógica difusa [Stefik, 1995].

Aunque se describen todos los operadores, en esta aproximación solamente hay antecedentes unidos con el operador lógico de conjunción. A continuación, se muestra la forma en que se lleva a cabo el cálculo del peso del sucedente, en las reglas de producción ponderadas en el sucedente. El valor del peso depende directamente del conectivo lógico con el que estén relacionadas las proposiciones que intervienen en el antecedente. Luego, si las proposiciones están unidas por el operador de conjunción, el cálculo de este peso es diferente a si están unidas por un operador de disyunción.

5.4.1 Conjunción

El cálculo de P_c, para la conjunción se efectúa escogiendo el peso diferenciante mínimo de las variables que intervienen en el antecedente de cada regla [Stefik, 1995]. El cálculo del peso para la conjunción se representa con la siguiente expresión:

$$Conjuncion(C_{1v1}, \dots C_{nvn}) = min(P(C_{1v1}), \dots, P(C_{nvn}))$$

La conjunción es un operador lógico que es verdadero si todos los elementos de la expresión son verdaderos. La Figura 5.1 muestra que para la conjunción de A y B denotado por $A \wedge B$ es verdadera siempre que el valor de A sea verdadero. Por esta razón, se elije el valor mínimo, que es el que hace que se cumpla la conjunción.

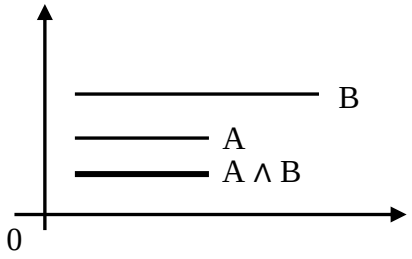

Figura 5.1. Operación conjunción.

Un ejemplo sería:

$$\mathrm{A}_{(0.37)}\ \&\ \mathrm{B}_{(0.70)} \sim \mathrm{S}\ \boldsymbol{transformada\ a}\ \mathrm{A}\ \&\ B \Longrightarrow S(0.37, 0)$$

Entonces, para obtener el peso de cumplimiento P_C del sucedente S, se obtiene el valor mínimo del antecedente. En este caso, A tiene en un peso de 0.37 y de B tiene un peso de 0.70, el mínimo de estos dos valores es 0.37 y por consiguiente, el peso de cumplimiento P_C asociado a S es igual 0.37, y el peso de incumplimiento P_I es cero.

5.4.2 Disyunción

Para la disyunción, se escoge el valor de peso diferenciante máximo de las categorías que intervienen en el antecedente de la regla [Stefik, 1995]. El cálculo del peso para la disyunción se representa con la siguiente expresión:

$$Disyuncion(v_1, \dots v_n) = max(P(v_1), \dots, P(v_n))$$

La disyunción es un operador lógico que es verdadero si uno o ambos operadores son verdaderos. La Figura 5.2 muestra que para la disyunción entre A y B denotado por $A \vee B$ es verdadera siempre que uno de los dos operadores A o B son verdaderos, de ahí que la expresión $A \vee B$ se mantenga verdadera mientras que B sea verdadero. Por esta razón, se elije el valor máximo, que es el que mantiene verdadera la expresión la disyunción.

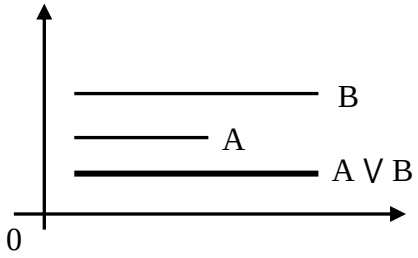

Figura 5.2. Operación disyunción.

Un ejemplo sería:

$$\mathrm{A}_{(0.37)} \,||\, \mathrm{B}_{(0.70)} \sim \mathrm{S}\ \boldsymbol{transformada\ a}\ \mathrm{A} \,||\, \mathrm{B} \Longrightarrow S(0.70, 0)$$

Entonces, para obtener el peso de cumplimiento P_C del sucedente S, se obtiene el valor máximo del antecedente. En este caso, de *A* que tiene un peso de 0.37 y de *B*, que tiene un peso de 0.70, el máximo de estos dos valores es 0.70 y por consiguiente, el peso de cumplimiento P_C asociado a S es igual 0.70, y el peso de incumplimiento P_I es cero.

5.4.2 Negación

Para la negación, sólo se cambia el signo del valor de la categoría que está asociada al operador negación –.

$$Negacion(C_{nvn}) = -(P(C_{nvn}))$$

Cuando se niega una categoría con un valor de peso definido, el grado de creencia que se tiene sobre la categoría es el mismo pero con signo opuesto, lo cual queda representado en la Figura 5.3. Así, cuando se cumple un hecho con absoluta seguridad (1) y se niega, se puede decir que no se cumple dicho un hecho con absoluta seguridad (-1).

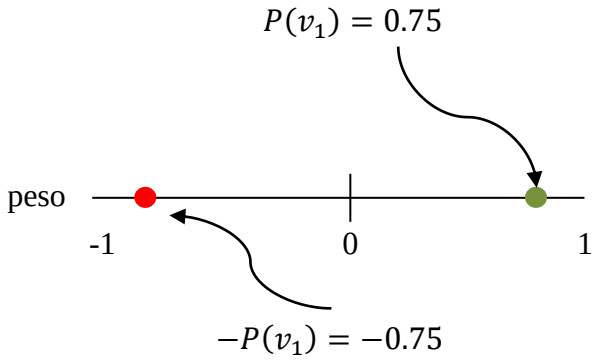

Figura 5.3. Operación negación.

La siguiente expresión incluye la negación de uno de sus operadores (A).

$$-A_{(0.37)}\ \&\ B_{(0.70)} \sim S\ \boldsymbol{transformada\ a}\ A\ \&\ B \Longrightarrow S(-0.37, 0)$$

El valor de A está negado con el símbolo –, esto quiere decir, que se debe de cambiar el signo del peso de A y después realizar la operación de conjunción que es el mínimo entre los dos valores de las variables. Entonces, el valor de A quedo como -0.37 y el de B es

igual a 0.70, por tanto, el mínimo entre estos dos valores es -0.37 y este es el peso de certidumbre P_C de S y el peso de incertidumbre P_I es igual a cero.

5.5 Reglas temporales

Las reglas temporales o RPP temporales, son un proceso intermedio e interno que nos permitirá crear, al final, reglas de producción ponderadas en el sucedente. Estas reglas se pueden ver como un hibrido de reglas de asociación y reglas de producción ponderadas en el sucedente, que para este caso sólo nos servirán para transformar las RA a RPP.

A continuación, se ejemplifica la traducción a reglas de producción ponderadas temporales, partiendo de las reglas obtenidas en el capítulo 4 y de la forma de calcular la ponderación a la regla, explicado en la sesión anterior.

La siguiente regla es del tipo de reglas de asociación que se someterá al proceso de traducción a RPP temporales.

RA 1: GÉNERO (0) & COLESTEROL(1) ~ ESTADO DE SALUD(2),

En el antecedente intervienen dos categorías, GÉNERO(0) con un PD de 1 y COLESTEROL(1) con PD de 0.285, relacionadas por el operador lógico AND. Como ya se vio, el operador & se mantiene para las reglas de producción temporales y el símbolo de asociación ~ es cambiado por el de relación de producción $\Longrightarrow$. El peso de cumplimiento P_C para la categoría ESTADO DE SALUD(2) se obtiene del valor mínimo entre los valores de las categorías GÉNERO(0) y COLESTEROL(1), el cual es 0.285, es decir el $P_C = 0.285$ y el peso de incumplimiento P_I para la categoría ESTADO DE SALUD(2) es de cero, es decir $P_I = 0$. Todo lo anterior, se puede representar en una RPPT como sigue:

RT1: GÉNERO (0) & COLESTEROL (1) ⇒ ESTADO DE SALUD (2) (0.285, 0)

Haciendo lo mismo para las demás reglas de asociación generadas anteriormente y tomando los PD de las variables de la Tabla 2.4, se obtuvieron las siguientes reglas de producción ponderadas temporales:

RT 1: GÉNERO (0) & COLESTEROL (1) ⇒ ESTADO DE SALUD (2) (0.285, 0)

RT 2: GÉNERO (0) & PERÍMETRO (1) ⇒ ESTADO DE SALUD (2) (0.428, 0)

RT 3: GÉNERO (1) & PERÍMETRO (2) ⇒ ESTADO DE SALUD (4) (0.428, 0)

RT 4: GÉNERO (1) & IMC (2) ⇒ ESTADO DE SALUD (4) (0.571, 0)

RT 5: GÉNERO (1) & PRESIÓN (2) ⇒ ESTADO DE SALUD (4) (0.285, 0)

RT 6: GÉNERO (0) & IMC (2) ⇒ ESTADO DE SALUD (2) (0.571, 0)

RT 7: EDAD (5) & GÉNERO (2) ⇒ ESTADO DE SALUD (4) (0.428, 0)

RT 8: GÉNERO (1) & IMC (1) ⇒ ESTADO DE SALUD (4) (0.571, 0)

RT 9: EDAD (2) & PRESÍON (1) ⇒ ESTADO DE SALUD (2) (0.285, 0)

RT 10: GÉNERO (1) & PRESIÓN (3) ⇒ ESTADO DE SALUD (4) (0.285, 0)

RT 11: GÉNERO (1) & IMC (2) ⇒ ESTADO DE SALUD (3) (0.571, 0)

RT 12: EDAD (4) & PERÍMETRO (5) ⇒ ESTADO DE SALUD (3) (0.428, 0)

RT 13: GÉNERO (0) & EDAD (3) ⇒ ESTADO DE SALUD (1) (0.428, 0)

RT 14: PERÍMETRO (1) & GÉNERO (0) ⇒ ESTADO DE SALUD (2) (0.428, 0)

5.6 Creación de la Base de Conocimiento

Retomando la definición de Base de Conocimiento que es donde se almacena la información necesaria del SE sobre un problema en alguna ERP. La BC es interpretada como la unión de dos componentes fundamentales:

- *Base de hechos* que es conjunto formado por hechos o conceptos provenientes de un dominio especifico del conocimiento
- *Base de reglas*, que es un conjunto de relaciones entre los elementos de la base de hechos.

En este enfoque, se genera la Base de Proposiciones y la Base de Reglas para su incorporación en la base de conocimiento de un SE.

5.6.1. Base de Proposiciones

Todas las categorías de las variables que aparecen en las reglas de producción temporales tanto en el antecedente como en el sucedente, forman la *Base de Proposiciones*.

El proceso para identificar las proposiciones en las RPP temporales, es como sigue:

Paso 1: De todo el conjunto de reglas, identificar todas las categorías que aparecen en la regla, comenzando por las categorías que se presentan en el sucedente.

Paso 2: Enumerar las categorías según aparición, sin que se repitan, empezando por el número 1.

Paso 3: Identificar las categorías de que se presentan en al antecedente.

Paso 4: Enumerar las categorías según aparición, sin que se repitan, continuando con la numeración del paso 2.

Tabla 5.1 Lista de proposiciones encontradas en las RPP temporales

Proposición	Descripción
P1	Mujer saludable
P2	Hombre saludable
P3	Hombre no saludable
P4	Mujer no saludable
P5	Es una mujer
P6	Presenta colesterol entre 153 y181
P7	Tiene un perímetro abdominal ente 75 y 85
P8	Es un hombre
P9	Tienen un perímetro abdominal ente 86 y 96
P10	Su IMC está entre 26-30
P11	Presenta presión arterial entre 103 y 115
P12	Su edad está entre 53 y 60 años
P13	Su IMC está entre 21 y 25
P14	Su edad está entre 32 y 38 años
P15	Presenta presión arterial entre 90 y 102
P16	Presenta presión arterial entre 116 y 128
P17	Su edad está entre 46 y 52 años
P18	Tiene un perímetro abdominal ente 119 y 129
P19	Su edad está entre 39 y 45 años

La lista de proposiciones para las reglas presentadas en la sección, se puede ver en la Tabla 5.1.

5.6.1. Base de Reglas

En esta sección, se obtienen las reglas de producción ponderas en el sucedente tal y como serán implementadas en un sistema experto y que formarán parte de la base de reglas de la BC.

Se debe partir de las RPP temporales y de la base de proposiciones, para poder generar la base de reglas que contendrá las reglas de producción ponderadas en el sucedente. El siguiente algoritmo muestra el proceso para traducir a RPP.

PARA todas las reglas RTj HACER
    PARA todas las proposiciones en la regla RT_j HACER
        Buscar el número que le corresponde en la base de proposiciones
        Remplazar la proposición por su número correspondiente
        Convertir la ponderación de la reglas a una escala de 0 a 100
        Crear Rpp_j
    FIN PARA
FIN PARA

Todas las reglas temporales mostradas en la sesión anterior, se someten al algoritmo descrito anteriormente para obtener RPP, mostradas a continuación:

Rpp1) 5 & 6 ⇒ 1(28, 0)
Rpp2) 5 & 7 ⇒ 1(42, 0)
Rpp3) 8 & 9 ⇒ 2(42, 0)
Rpp4) 8 & 10 ⇒2(57, 0)
Rpp5) 8 & 11 ⇒2(28, 0)
Rpp6) 5 & 10 ⇒1(57, 0)
Rpp7) 12 & 8 ⇒2(42, 0)
Rpp8) 8 & 13 ⇒2(57, 0)
Rpp9) 14 & 15⇒1(28, 0)
Rpp10) 8 & 16 ⇒ 2(25, 0)
Rpp11) 8 & 10 ⇒ 3(57,0)
Rpp12) 17 & 18 ⇒3(42, 0)
Rpp13) 5 & 19 ⇒ 4(42, 0)
Rpp14) 7 & 5 ⇒ 1(42, 0)

Capítulo 6. Respaldo de la Información

El respaldo de la información es la última etapa de la metodología presentada. Este módulo proporciona la posibilidad de respaldar los resultados generados por cada uno de las otras etapas de la metodología. Estos archivos entonces, almacenan el PD, las reglas de asociación, la base de reglas y la base de proposiciones.

El respaldo de la información es un proceso importante en cualquier sistema informático ya que éste permite generar archivos digitales que almacenan información relevante que posteriormente puede ser utilización para su revisión, manipulación y uso en procesos manuales o automáticos futuros.

El tipo de archivos digitales que genera el sistema son archivos texto, los cuales pueden ser manipulados con casi cualquier editor de texto. Algunas de las características que presentan estos archivos son: la portabilidad y fácil manipulación, entre otras.

En siguientes secciones se describen el contenido y la estructura de cada uno de los archivos digitales generados por esta metodología.

6.1 Respaldo del peso diferenciante

El archivo generado, almacena el PD de cada una de las variables, de un conjunto de datos dado, calculado por el algoritmo de relevancia de variables Test Típicos.

El archivo se estructura como sigue:

Paso 1: En la primera línea, poner el número total de variables

Paso 2: Colocar en una nueva línea, el nombre de la variable seguido de su PD calculado entre corchetes

Paso 3: Realizar el paso 2 para todas las variables que se sometieron al proceso de relevancia de variables

La estructura del archivo del peso diferenciante, descrito anteriormente, se presenta en la Figura 6.1.

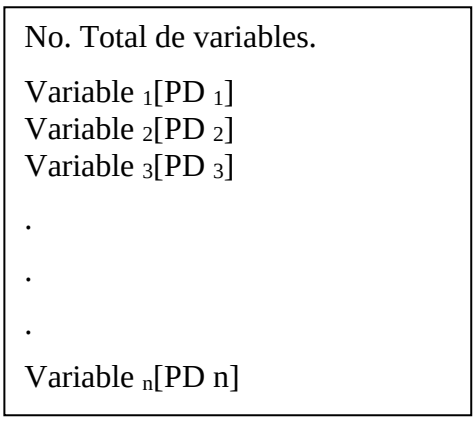

Figura 6.1. Estructura del archivo para el peso diferenciante de las variables.

En la Figura 6.2 se muestra el archivo digital para el ejemplo que se ha presentado y retomando los pesos diferenciantes de las variables calculados en el capítulo 3 que se mostraron en la Tabla 3.2.

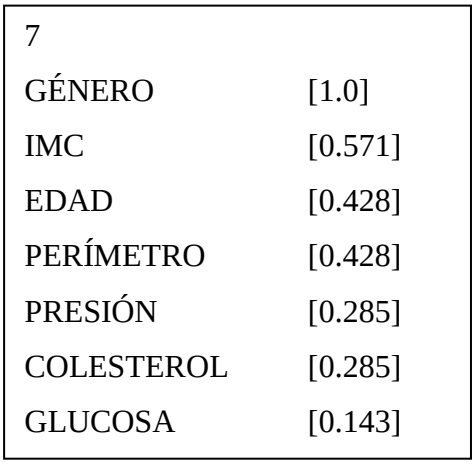

Figura 6.2. Archivo con el peso diferenciante de las variables.

6.2 Respaldo de reglas de asociación

El archivo presentado en esta sección, almacena todas las RA generadas a partir del conjunto de datos proporcionado por el usuario y las cuales fueron calculadas por el método GUHA de generación de asociaciones.

El archivo digital presenta la siguiente estructura.

Paso 1: En la primera línea, poner el número total de RA

Paso 2: Colocar en una nueva línea, una regla de asociación con el prefijo Ra y un número consecutivo que identifica a cada regla

Paso 3: Realizar el paso 2 para todas las RA que se obtuvieron con el método GUHA

La estructura del archivo de reglas de asociación se presenta en la Figura 6.3.

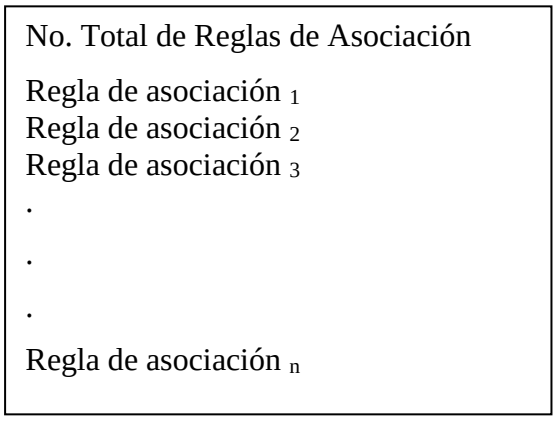

Figura 6.3. Estructura del archivo para las reglas de asociación.

En la Figura 6.4 se muestra el archivo digital con RA, para el ejemplo que se ha presentado y retomando las reglas de asociación presentadas en el capítulo 4.

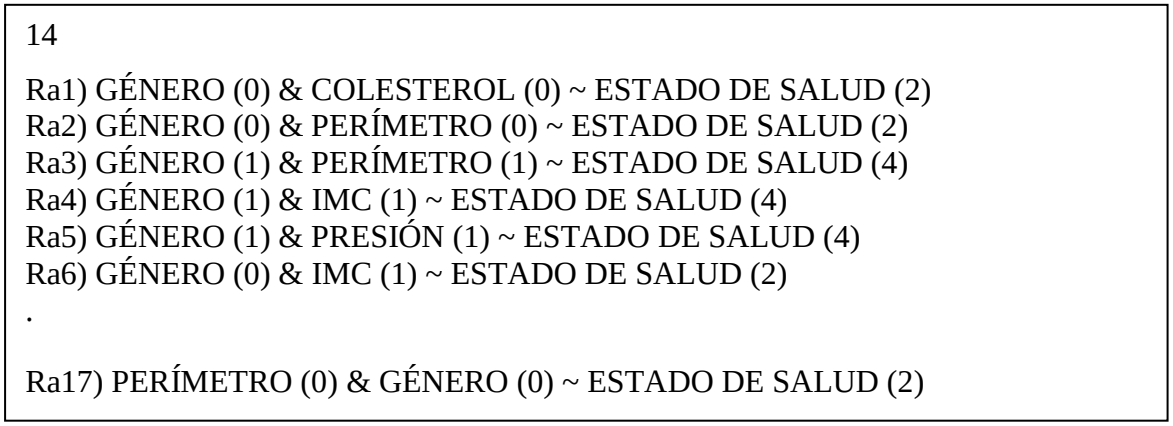

Figura 6.4. Archivo con las reglas de asociación.

6.3 Respaldo de reglas de producción ponderadas en el sucedente

El archivo que se genera en este módulo, contiene las reglas de producción ponderadas en el sucedente generadas por el proceso de traducción a reglas de producción.

El archivo creado, presenta la siguiente estructura.

Paso 1: En la primera línea, se aparece el número total de RPP

Paso 2: Colocar en una nueva línea, una regla de producción ponderada anteponiendo el prefijo Rpg#), donde el símbolo # es un número consecutivo que identifica a cada regla

Paso 3: Realizar el paso 2 para todas las RPP que se obtuvieron con el proceso de traducción

La estructura del archivo se presenta en la Figura 6.5, el cual sigue la estructura que deben tener los archivos que utiliza el lenguaje HAries para este propósito, y la extensión que debe tener tal archivo es rpg.

```
No. Total de Reglas de Producción Ponderadas

Rpg1) 5 & 6 ⇒ 1(28, 0)
Rpg2) 5 & 6 ⇒ 2(42, 0)
Rpg3) 8 & 9 ⇒ 1(57, 0)
Rpg2) 8 & 11 ⇒ 3(28, 0)
Rpg3) 5 & 6 ⇒ 1(28, 0)
.
Rpgn) 7 & 5 ⇒ 4(42, 0)
```

Figura 6.5. Estructura del archivo para las reglas de producción ponderadas en el sucedente.

Las RPP son almacenadas en un archivo plano con la extensión rpg, con el objetivo de ser exportadas en el lenguaje HAries. En la Figura 6.6 se muestra un ejemplo de este tipo de archivo.

```
14
Rpg1) 5 & 6 ⇒ 1(28, 0)
Rpg2) 5 & 7 ⇒ 1(42, 0)
Rpg3) 8 & 9 ⇒ 2(42, 0)
Rpg4) 8 & 10 ⇒2(57, 0)
Rpg5) 8 & 11 ⇒2(28, 0)
Rpg6) 5 & 10 ⇒1(57, 0)
Rpg7) 12 & 8 ⇒2(42, 0)
.
.
.
Rpg14) 7 & 5 ⇒ 1(42, 0)
```

Figura 6.6. Archivo con las reglas de producción ponderadas en el sucedente.

6.4 Otros archivos generados por el sistema

En esta sección, se describen las estructuras y contenidos de otros archivos que están comprendidos en el enfoque y que junto con el de reglas de producción ponderadas en el sucedente, son los archivos de entrada para el SE, con el objetivo de ampliar y complementar su BC. Estos archivos son: los textos de proposiciones y sus parámetros.

El lenguaje hibrido de representación del conocimiento HAries que permite por medio de un ambiente la construcción de aplicaciones basadas en conocimiento. Además, HAries se compone de un esquema de representación de conocimiento, varias máquinas de inferencia para obtener el valor de las proposiciones, un sistema de adquisición para introducir y modificar la BC y un sistema explicatorio [De la Cruz, 1996].

6.4.1. Archivo de texto de proposiciones

En este archivo, se almacena la descripción de todas las proposiciones que han aparecido en las reglas de producción ponderadas en el sucedente, tanto en el antecedente como en el consecuente. El archivo generado tiene la extensión TxP con el propósito de que pueda ser importada por el leguaje de representación del conocimiento HAries. Su estructura se muestra en la Figura 6.7.

```
No. Total de Proposiciones

1
Descripción positivo de la proposición 1 ¦
Descripción negativa de la proposición 1¦

2
Descripción positivo de la proposición 2¦
Descripción negativa de la proposición 2¦
.
.
.
n
Descripción positivo de la proposición n¦
Descripción negativa de la proposición n¦
```

Figura 6.7. Estructura del archivo de textos de proposiciones.

Para generar este archivo se siguen los siguientes pasos:

Paso 1: En la primera línea, se pone el número total de proposiciones

Paso 2: Colocar en una nueva línea, el número que identifica a la proposición correspondiente

Paso 3: Colocar en una nueva línea, una descripción positiva referente a la proposición con un separador de línea representado con el símbolo “ ¦ ”

Paso 4: Poner en una nueva línea, una descripción negativa referente a la proposición con un separador de línea “ ¦ ”

Paso 5: Realizar el paso 2, 3 y 4 para todas las proposiciones que aparecen en las RPP.

En ejemplo de un archivo de proposiciones con extensión TxP es el que se presenta en la Figura 6.8, el cual contiene la descripción de las proposiciones que aparecieron en las reglas de producción ponderadas en el sucedente.

```
19
1
Mujer saludable¦
Negación Prop. 1¦
2
Hombre saludable¦
Negación Prop. 2¦
3
Hombre no saludable¦
Negación Prop. 3¦
4
Mujer no saludable¦
Negación Prop. 4¦
.
.
.
19
Su edad está entre 39 y 45 años¦
Negación Prop. 19¦
```

Figura 6.8. Archivo de texto de proposiciones para el ejemplo utilizado.

6.4.2. Archivo de parámetros de las proposiciones

El archivo de los parámetros de las proposiciones contiene información específica sobre las propiedades que presentan las proposiciones con respecto a HAries. El archivo generado tiene la extensión PRO.

Este archivo de parámetros sirve para indicar al lenguaje HAries qué propiedades tiene cada elemento de la base de proposiciones y es sumamente importante pues define el comportamiento de cada una de las proposiciones definidas.

El archivo tiene la siguiente estructura:

Paso 1: En la primera línea aparece la cantidad total de proposiciones.

Paso 2: Para i=1 hasta el número total de proposiciones:

Las siguientes líneas están compuestas por una serie de columnas, cada una de las cuales representa una propiedad asociada a la proposición y cada línea corresponde a una proposición diferente. Las propiedades que pueden presentar las proposiciones son:

1. Atributo natural: propiedad que toma la proposición con respecto al lugar donde aparece en las reglas de producción ponderadas en el sucedente, estas pueden ser pregunta (P) y Objetivo(O).
2. Se indica como fue dada la propiedad anterior, por las reglas de forma natural (N), o por el usuario.
3. Tipo de valores: propiedad que indica el tipo de valor que representa la proposición y esta puede ser:

I → Implícito	N → Numérico
G → Gráfico	B → Numérico y gráfico
D → Discreta	E → Extremal
F → Fija	A →Asignada

- Tipo de dinamismo: E si es estática y D si es dinámica
- Primera regla de producción donde aparece como sucedente, en el caso de que sea objetivo, se pone 0 si no aparece en ninguna.
- Variable asociada: número de variable asociada a la proposición, 0 si no tiene ninguna
- Proposición asociada: se coloca la cantidad de proposiciones
- Acciones antes de evaluar la proposición: se coloca la cantidad de acciones
- Relación proposición-acción: I para Independiente y D para Dependiente
- Acciones después de evaluar la proposición: se coloca la lista de acciones
- Relación contextual
- Relación de evaluación alternativa.

Acabe aclarar que, debido a la naturaleza del enfoque propuesto, se desconocen ciertas propiedades o parámetros asociados a las proposiciones. Para indicar el desconocimiento de esos parámetros se indica implícitamente con un valor de 0.

En la Figura 6.9 se presenta un ejemplo de este archivo, con las proposiciones de la Figura 6.8.

```
19

O N I E 1 0 0 0 0  0 0 0 0
O N I E 3 0 0 0 0  0 0 0 0
O N I E 11 0 0 0 0  0 0 0 0
O N I E 13 0 0 0 0  0 0 0 0
P N I E 0 0 0 0 0  0 0 0 0
P N I E 0 0 0 0 0  0 0 0 0
P N I E 0 0 0 0 0  0 0 0 0
P N I E 0 0 0 0 0  0 0 0 0
P N I E 0 0 0 0 0  0 0 0 0
P N I E 0 0 0 0 0  0 0 0 0
P N I E 0 0 0 0 0  0 0 0 0
P N I E 0 0 0 0 0  0 0 0 0
P N I E 0 0 0 0 0  0 0 0 0
P N I E 0 0 0 0 0  0 0 0 0
P N I E 0 0 0 0 0  0 0 0 0
P N I E 0 0 0 0 0  0 0 0 0
P N I E 0 0 0 0 0  0 0 0 0
P N I E 0 0 0 0 0  0 0 0 0
P N I E 0 0 0 0 0  0 0 0 0
```

Figura 6.9. Archivo de parámetros de las proposiciones.

Capítulo 7. Sistema de generación de conocimiento

En el presente capítulo, se muestra la implementación del enfoque propuesto para la generación de reglas de producción ponderadas en el sucedente, a través de una herramienta computacional orientada a la Web, la cual ha sido desarrollada con los lenguajes de programación PHP, JavaScript y HTML.

La aplicación computacional está estructurada en varios módulos los cuales permiten llevar a cabo el proceso de extracción de conocimiento, empezando por la asignación de categorías para su procesamiento, luego el cálculo de la relevancia de variables seleccionadas, la generación de reglas de asociación y finalmente, obteniendo reglas de producción ponderadas en el sucedente y además, con la posibilidad de respaldar los resultados de cada módulo. Todas estas operaciones están representadas en opciones en el menú principal de la aplicación.

La información que contiene los datos de inicio, que aquí se utilizan, está relacionada con el estudio médico realizado al personal del Servicio Geológico Mexicano, que se explicó en el capítulo 2 y los datos observacionales se muestran la Tabla 2.3.

El menú principal está compuesto de seis opciones, colocadas horizontalmente, las cuales sirven para acceder y llevar a cabo todo el proceso de generación de reglas de producción ponderadas en el sucedente, tal y como se describió anteriormente en el capítulo 2. Las opciones que comprenden al menú principal son:

- **Principal**.- Muestra información general del sistema, así como, la información de los algoritmos implementados para generar este tipo de reglas.
- **Datos iniciales**.- Permite subir al sistema los datos generales del problema a resolver como son: una descripción general de los datos, el nombre de cada una de las variables, el nombre de todas las categorías de las variables, especificando cuál es la variable objetivo y, por supuesto, los datos del problema en forma de matriz polivalente.

- **Peso diferenciante**.- En esta opción, se calcula el peso diferenciante de las variables de la matriz polivalente subida anteriormente al sistema. El procedimiento para calcular se describió en el capítulo 3.
- **Reglas de asociación**.- Genera todas las reglas de asociación posibles de la matriz polivalente. El algoritmo de generación de asociaciones se presentó en el capítulo 4.
- **Reglas de producción**.- Hace una conversión de las reglas de asociación a reglas de producción ponderadas en el sucedente, generando paralelamente una lista de proposiciones, que no son más que todas las categorías que aparecen tanto en el sucedente como en el antecedente de todas las reglas extraídas de los datos.
- **Respaldo de información**.- La aplicación pone a disposición cinco archivos digitales para que el usuario pueda descargar y reutilizar dicha información. Los archivos que se crean contienen: el peso diferenciante de las variables, las reglas de asociación, las reglas de producción ponderadas en el sucedente, la lista de proposiciones y los parámetros que describen a las proposiciones.

7.1 Página Principal

La pantalla principal del sistema está dividida en tres secciones bien definidas que se pueden ver, gráficamente, en la Figura 7.1.

- Sección 1: Se encuentra en la parte superior de la pantalla y con ésta se pueden acceder a los algoritmos implementados en el sistema por medio de las opciones del menú.
- Sección 2: Esta sección se encuentra en la parte central de la pantalla, en la cual está disponible el contenido informativo del sistema y es donde aparecerá el contenido de cada una de las opciones del submenú.
- Sección 3: Está ubicada en la parte inferior de la pantalla y contiene información sobre los responsables del enfoque y la herramienta computacional.

La página principal del sistema contiene la información general sobre el enfoque propuesto para generar reglas de producción ponderadas en el sucedente. En esta página, se muestran unas imágenes que representan los cuatro módulos que involucran el proceso para la

generación de conocimiento, los cuales implementan métodos y algoritmos de Minería de Datos.

Figura 7.1. Pantalla principal del sistema.

En la parte superior de la página principal se encuentra el menú principal, que contiene las opciones necesarias para la implementación de dicho enfoque; en la parte central están cuatro imágenes que representan los métodos de minería de datos, los cuales calculan o generan: el peso diferenciante de las variables, reglas de asociación, reglas de producción ponderadas en el sucedente y el respaldo de la información en archivos digitales. Al dar un clic sobre cada una de las imágenes, se desplegará más información referente a la opción seleccionada en forma de ventana como se puede observar en la Figura 7.2.

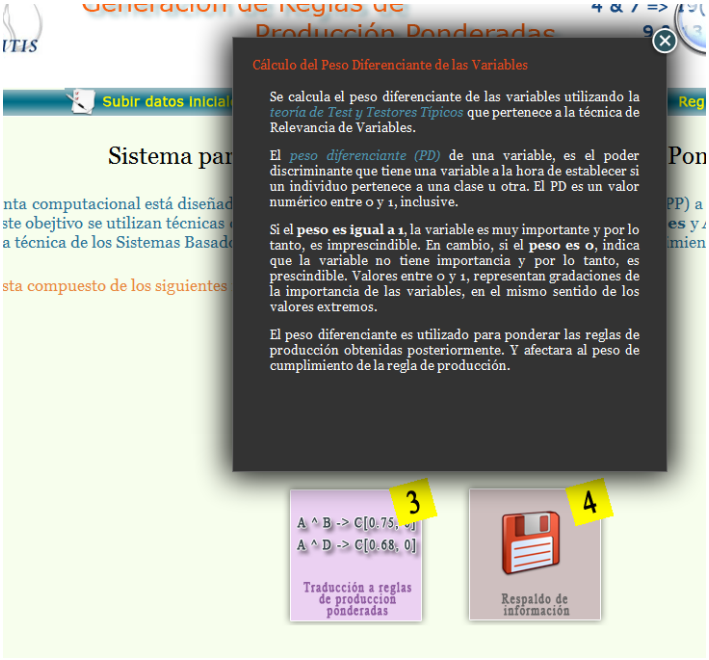

Figura 7.2 Información sobre el contenido del sistema.

7.2 Datos iniciales

Para poder ejecutar los módulos relacionados con el peso diferenciante, las reglas de asociación, las reglas de producción ponderadas en el sucedente y la obtención del respaldo de la información, primeramente, se debe cargar los datos alusivos al problema a procesar.

Para realizar la tarea de incorporar los datos al sistema, se debe ir a la opción *Datos iniciales* en el menú principal. Esta opción permite subir un archivo digital que contiene información general del problema a resolver y la matriz polivalente, como se presenta en la Figura 7.3. Una vez cargados los datos en el ambiente, se debe proporcionar cierta información referente a la matriz polivalente y para esto, se despliega un formulario en la parte derecha de la ventana como se presenta en la Figura 7.4. Si se desea, la información del formulario también puede ser suministrada a través de un archivo digital, el cual debe contener el nombre de cada una de las categorías que componen las variables y el nombre de la variable objetivo del problema.

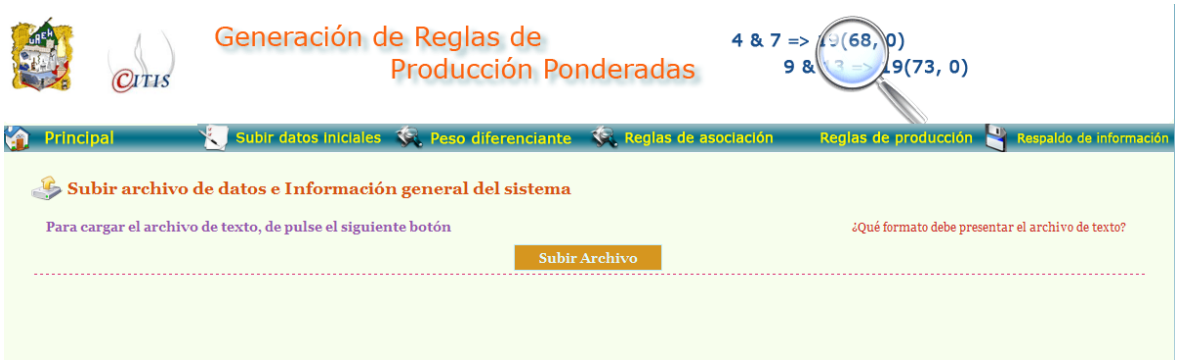

Figura 7.3. Pantalla para subir datos iniciales.

Para cargar el archivo digital que contiene la matriz polivalente, se debe dar un clic en el botón "Datos Iniciales" y el sistema mostrará una ventana de diálogo la cual permite al usuario buscar el archivo que contiene los datos del problema e incorporarlo al sistema. Una vez seleccionado el archivo, se despliega en la parte inferior de la ventana información referente al archivo de datos. En la parte izquierda**,** se muestra la descripción general del problema, las variables que integran el conjunto de datos y la matriz polivalente. En la parte derecha, se despliega un formulario donde se le pide al usuario que proporcione el nombre de cada una de las categorías de las variables y seleccione la variable objetivo del problema, esto es, la variable que clasifica a los datos, tal como se muestra en la Figura 7.4.

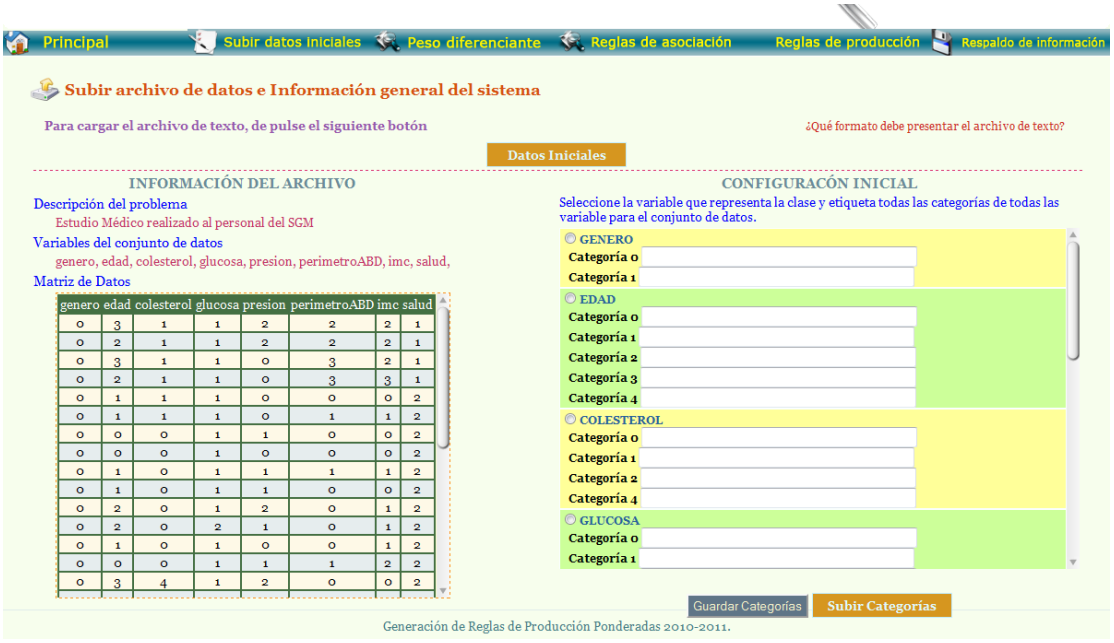

Figura 7.4. Pantalla que muestra información sobre el problema a resolver.

Otra forma de suministrar al sistema las descripciones de cada categoría de las variables e indicar la variable objetivo del problema, es dando un clic en el botón “Subir Categorías” e inmediatamente se presentará una ventana de diálogo para poder cargar el archivo que contiene información sobre las categorías y la variable objetivo, dicha información se solicita a través del botón que aparece en la parte derecha de la ventana que se muestra en la Figura 7.5.

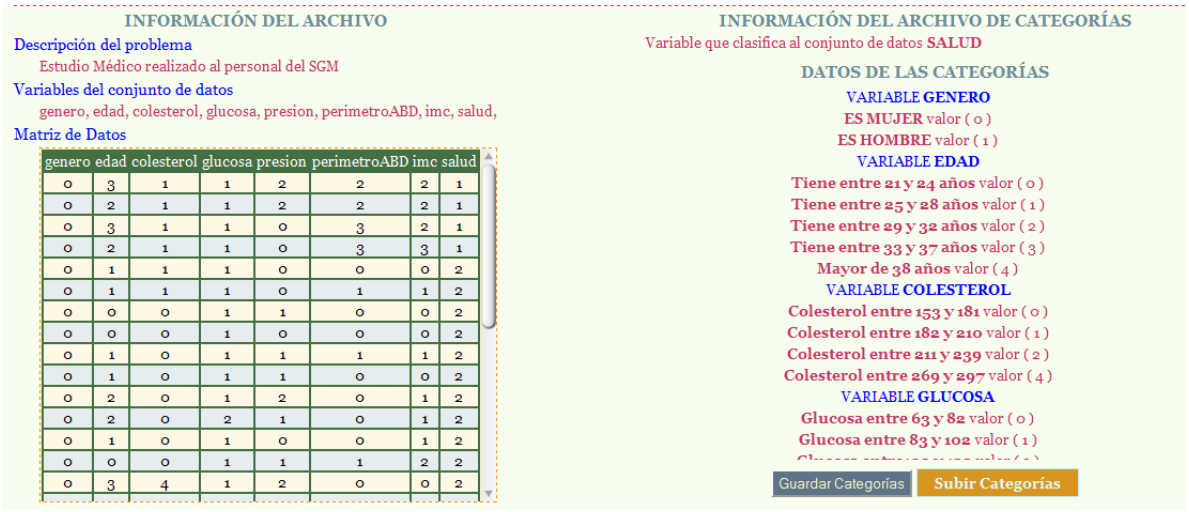

Figura 7.5. Información de las categorías cargas por un archivo.

En la opción Datos Iniciales, se le pide al usuario dos archivos digitales, uno para cargar la matriz polivalente y el otro para cargar información de las categorías de las variables y definir la variable objetivo. Los formatos que deben tener estos dos archivos y sus tipos de extensión se indican a continuación:

- *Archivo de la matriz polivalente:* este archivo contiene la estructura siguiente:

  → Línea 1.- Una descripción del problema

  → Línea 2.- El nombre de las variables o columnas de la matriz polivalente separadas por comas

  → Línea 3.- Ésta línea va en blanco

  → Línea 4.- A partir de esta línea, se encuentran los datos de la matriz polivalente, las columnas están separadas por comas y cada fila de la matriz ocupa una nueva línea en el archivo de texto. Este archivo puede tener las siguientes

extensiones txt, dat, inf o tmp y un ejemplo de éste se puede observar en la Figura 7.6.

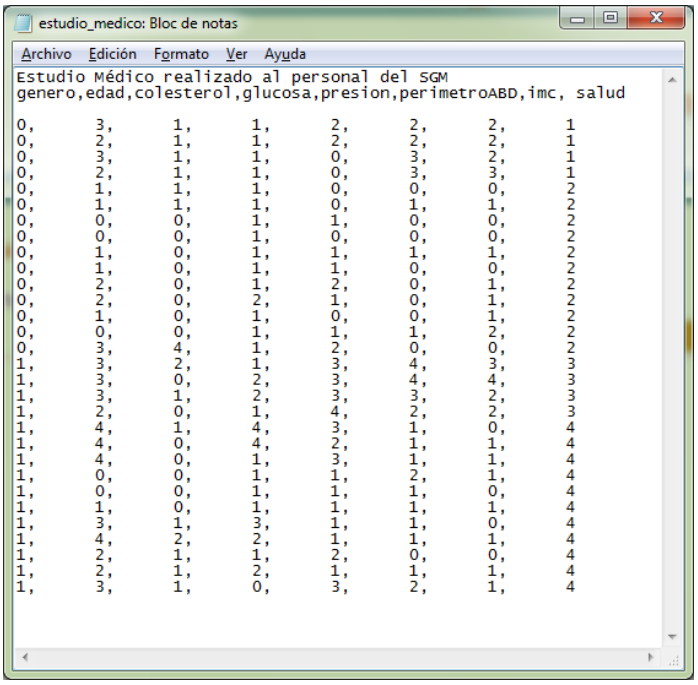

Figura 7.6. Formato del archivo con la matriz polivalente.

- *Archivo con la descripción de las categorías*: el contenido de este archivo está distribuido como sigue:

  → Línea 1.- Aparece el nombre de la variable objetivo encerrada entre los caracteres "@" y "|" y termina con el número de columna que le corresponde en la matriz polivalente

  → En las demás líneas y para cada una de las variables, se tiene:

  - Entre el carácter "@" y el carácter "," va el nombre de la variable y después la cantidad de categorías que la componen
  - Para cada categoría de la variable, el nombre de la categoría seguido del carácter "|" y el valor que le corresponde en la matriz polivalente. Un ejemplo que muestra como quedaría este archivo, se observa en la Figura 7.7 y la extensión del mismo es CAT.

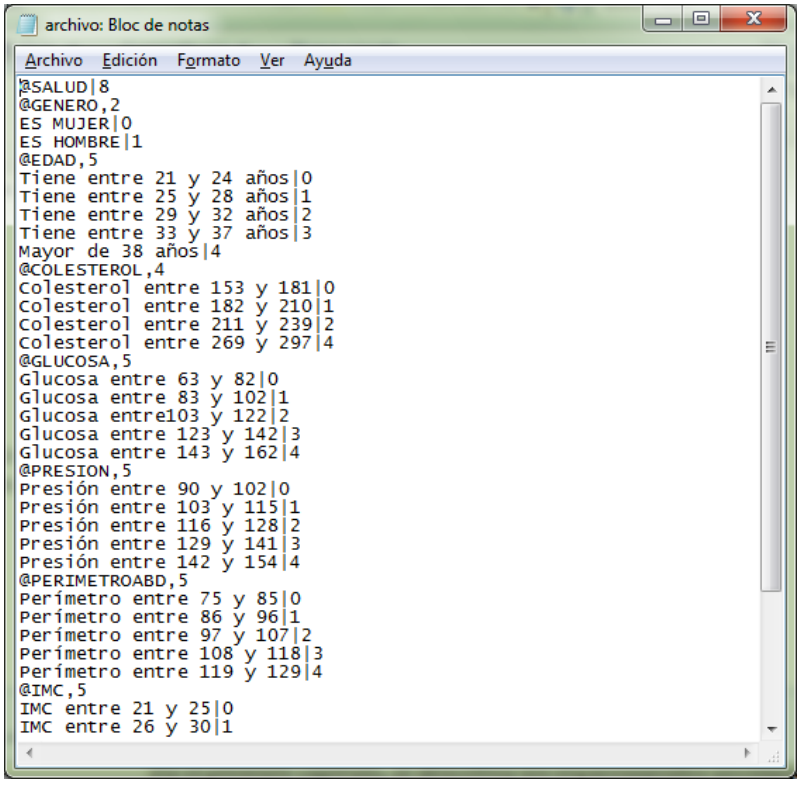

Figura 7.7. Archivo de categorías.

Una vez suministrada toda la información en esta opción, se está listo para aplicar los diferentes algoritmos a los datos suministrados por el usuario. El contenido del archivo de categorías es resultado del proceso de categorización representado en la Tabla 2.4.

7.3 Peso diferenciante

Para acceder a esta opción se debe de dar clic al submenú “Peso Diferenciante”, Si aún no se ha cargado la matriz polivalente, se despliega un cuadro de diálogo, como el de la Figura 7.8, en este caso, debería dirigirse primero a la opción del menú “Datos iniciales”.

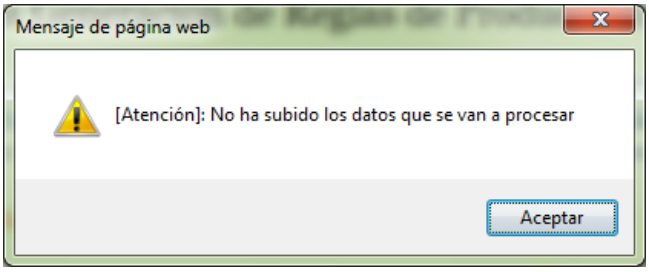

Figura 7.8. Cuadro de diálogo, mensaje de alerta.

Para calcular el peso diferenciante de las variables, sólo es necesario dar un clic en el botón naranja “Calcular peso” y debajo de éste se mostrará una tabla con dos columnas, la

primera, que contiene cada una de las variables y en la segunda, aparecerán los pesos asociados a dichas variables, lo cual se muestra en la Figura 7.9. Recordando que el peso diferenciante es el poder discriminante que tienen las variables para definir si un individuo pertenece a una clase u otra y que un valor cercano a 1 indica que su poder es más discriminante, no así, con un peso cercano a 0, que indica que su poder es muy poco discriminante.

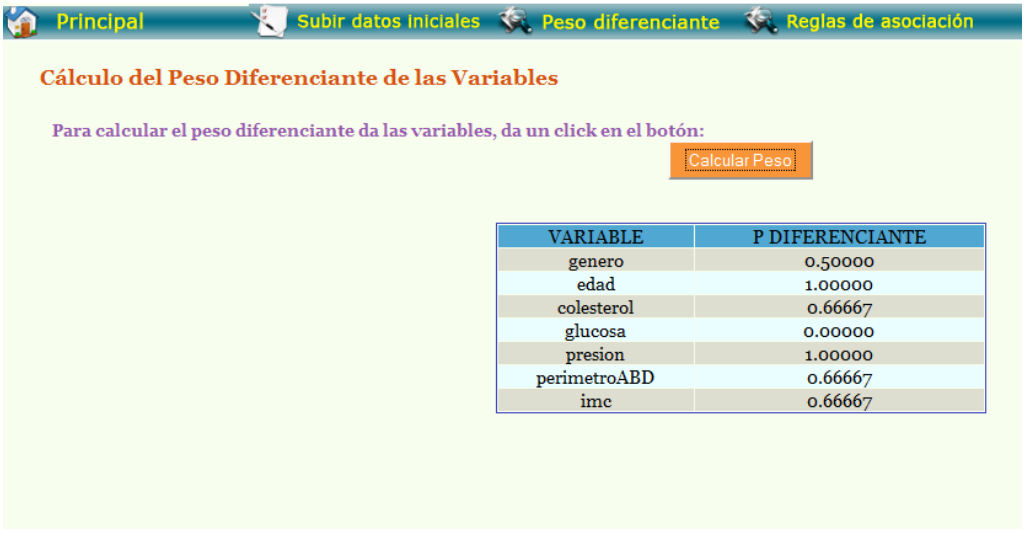

Figura 7.9. Pantalla para el cálculo del peso diferenciante.

7.4 Reglas de asociación

Esta opción permite generar las reglas de asociación tomando como datos de entrada a la matriz polivalente cargada en la opción “Datos iniciales”. Si aún no se ha cargado y se intenta entrar a esta pantalla, aparecerá un mensaje de diálogo, como el de la Figura 7.8, indicando que se debe de cargar primero los datos del problema.

Antes de poder generar las reglas de asociación, se debe configurar algunos parámetros propios del método GUHA. Los parámetros a configurar son;

→ número de hipótesis máximas a generar

→ número total de filas, que como mínimo, debe cumplir cada hipótesis generada, éste valor es llamado *base*

→ el nivel de significancia, es una probabilidad que sirve para aceptar o rechazar una hipótesis

→ el número de categorías que aparecerán en el antecedente, que deben ser dos como máximo.

Una vez que se han proporcionado información sobre los parámetros, se puede dar un clic en el botón “Generar reglas”, lo que hará que en la parte inferior se muestren las reglas encontradas por el método, desplegando primero las que contengan un antecedente y después, las de dos antecedentes, si es que se especificó en los parámetros. Las reglas generadas son mostradas en la Figura 7.10.

Figura 7.10. Pantalla para la generación de reglas de asociación

7.5 Reglas de producción ponderadas en el sucedente

Para acceder a esta opción, se debe elegir “Reglas de producción” en el menú principal. Para generar las reglas se debió de haber aplicado anteriormente al conjunto de datos, los algoritmos de peso diferenciante y reglas de asociación, si no es así, se desplegará el mensaje de advertencia que se muestra en la Figura 7.11.

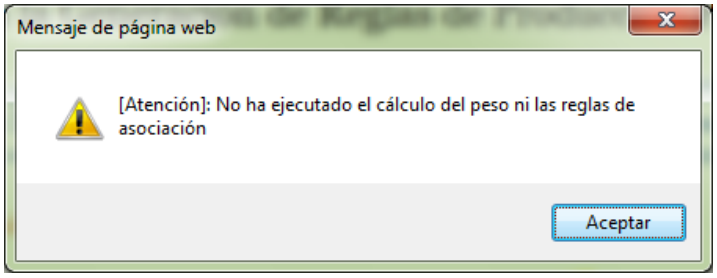

Figura 7.11. Mensaje de advertencia para la generación de reglas.

Con el simple hecho de acceder a la pantalla, se mostrarán dos tablas en la ventana, la tabla de la izquierda muestra todas las reglas de producción ponderadas en el sucedente y la tabla de la derecha contiene todas las proposiciones que aparecen en dichas reglas, para estas últimas, se mostrarán las proposiciones objetivo y después, todas las demás que aparecen en el antecedente. Todo lo anterior, se muestra en la Figura 7.12.

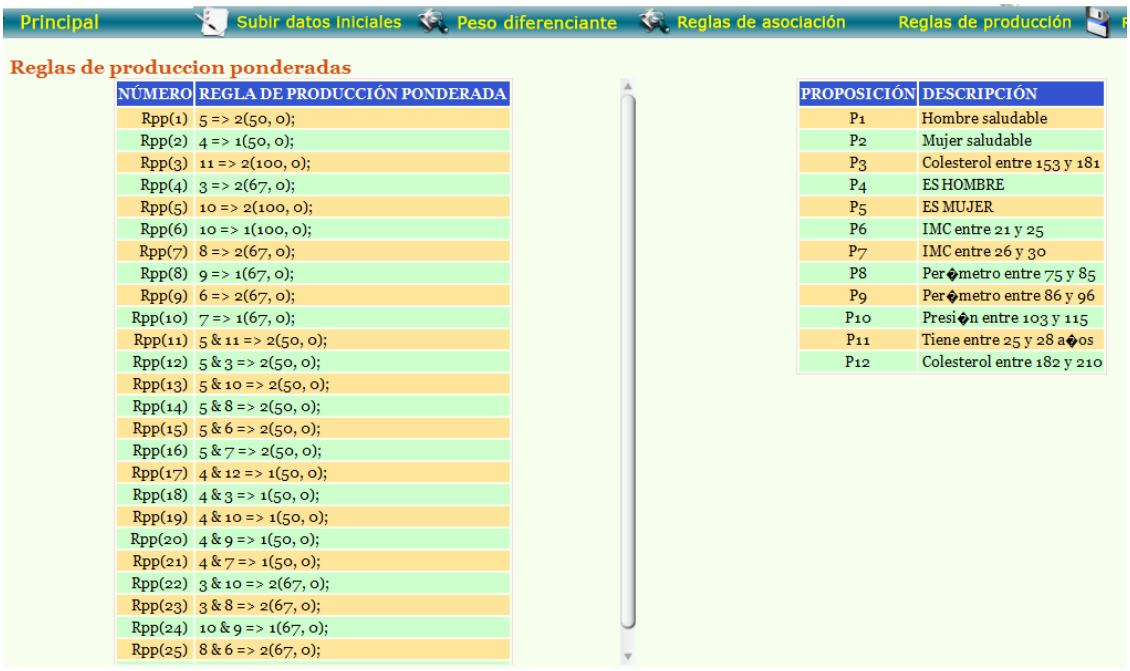

Figura 7.12. Pantalla para la generación a reglas de producción ponderadas en el sucedente.

Las reglas de producción ponderadas en el secedente junto con la lista de proposiciones conforman la base de conocimiento de un sistema experto para el Lenguaje de Representación del Conocimiento HAries [Alonso, 2004].

7.6 Respaldo de la información

En esta opción, que se puede acceder desde el submenú “Respaldo de información”, está disponible una vez que se haya aplicado cualquier algoritmo implementado en el sistema. Aquí, se pueden descargar todos los archivos digitales como se muestra en la Figura 7.13.

Figura 7.13. Pantalla para el respaldo de la información.

Son cinco los archivos digitales que se pueden ser descargados de la aplicación y ser utilizados en otros sistemas o procesos, éstos archivos son: el peso diferenciante, las reglas de asociación, la base de reglas de producción ponderadas en el sucedente, la base de proposiciones y los parámetros de las proposiciones. Una descripción de estos archivos se presenta a continuación:

- *Peso diferenciante:* Guarda el peso diferenciante de las variables, en este archivo aparece el número total de variables, las variables y sus correspondientes pesos, en el orden mencionado y como se muestra en la Figura 7.14.

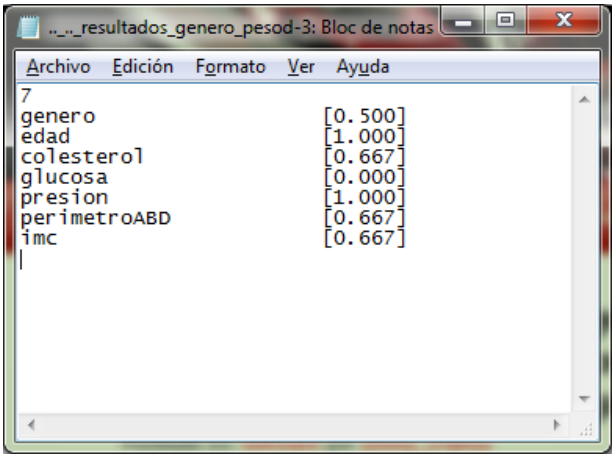

Figura 7.14. Archivo digital del peso diferenciante.

- *Reglas de asociación*: Contiene todas las reglas de asociación generadas por el sistema, en la primera línea aparece el número total de reglas y a partir de la tercera línea, se enlistan todas las reglas de asociación obtenidas por el método. Un ejemplo de este archivo se puede observar en la Figura 7.15.

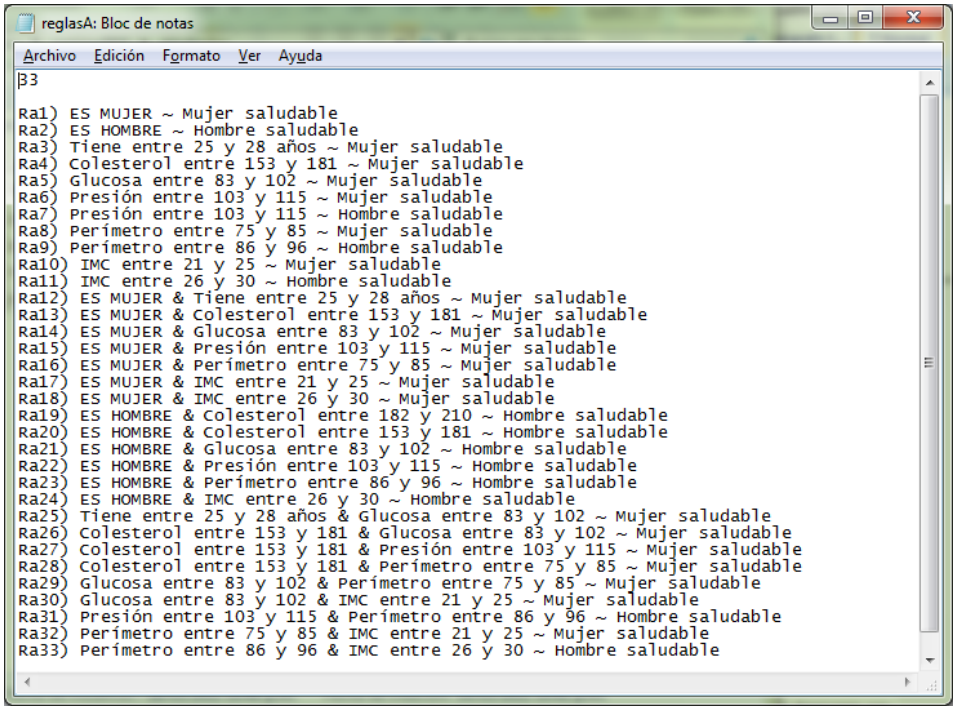

Figura 7.15. Archivo digital de las reglas de asociación.

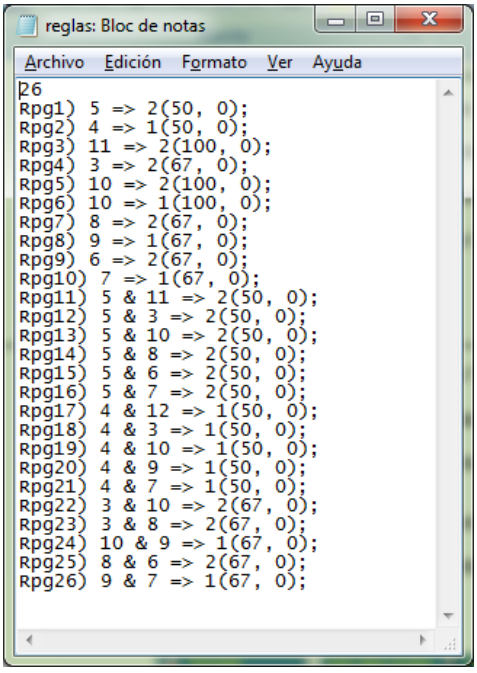

Figura 7.16. Archivo digital de la base de reglas.

- *Base de reglas*: Este archivo contiene las reglas de producción ponderadas en el sucedente con extensión RPG, con el objetivo de poder ser importadas por el lenguaje HAries para complementar el conocimiento con el del experto humano, este archivo se muestra en la Figura 7.16.

- *Base de proposiciones*: Este archivo guarda todas las proposiciones que aparecen en las reglas de producción ponderadas en el sucedente, en formato establecido por el programa HAries y con extensión TXP, un ejemplo de este archivo se muestra en la Figura 7.17.

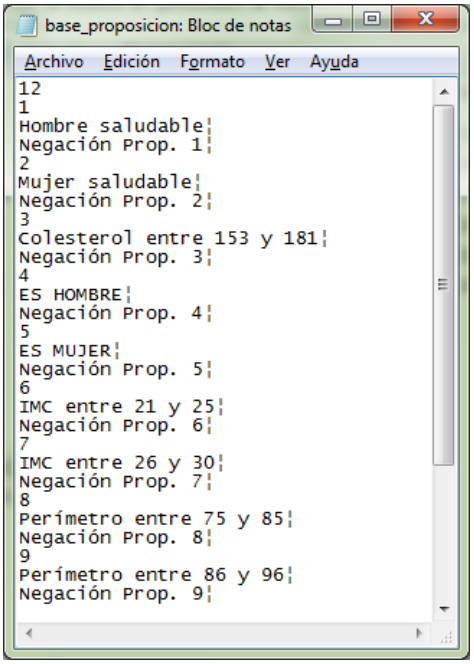

Figura 7.17. Archivo digital de base de proposiciones.

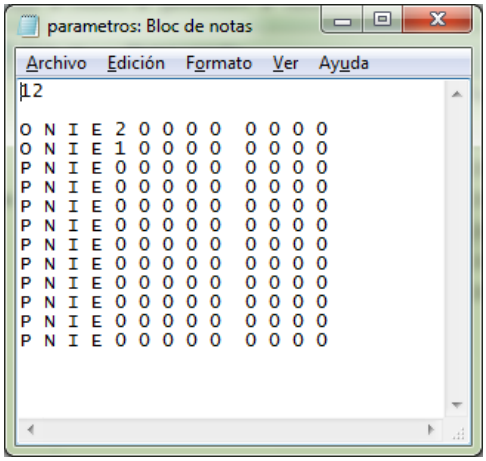

Figura 7.18. Archivo digital de los parámetros de las proposiciones.

- *Parámetros de las proposiciones*: El archivo de configuración indica los atributos o parámetros que están asociadas a cada una de las proposiciones que conforman la base de hechos, este archivo es necesario para ejecutar el sistema experto pues en él se define cómo se evaluarán las proposiciones y relaciones adicionales que se pueden establecer entre ellas. Un ejemplo de este archivo se muestra en la Figura 7.8.

Conclusiones

Los métodos de Minería de Datos aplicados permitieron extraer conocimiento de los datos y representarlos en un patrón de Reglas de Producción Ponderadas en el Sucedente.

Durante el desarrollo del trabajo, se pudo corroborar que las variables más importantes en la población estudiada, se manifestaron como tal en los dos métodos aplicados: la relevancia de variables y la generación de asociaciones. En el primero, a través de un valor numérico y en el segundo, por su frecuencia en la aparición en las reglas.

Finalmente, se evidenció que la computadora, equipada con las herramientas, técnicas y metodologías adecuadas, logra resolver los problemas apoyándose de ambas fuentes de información, *datos* y *conocimiento*, al igual que lo hace el ser humano.

Trabajos Futuros

Los trabajos futuros son tareas que pueden ser desarrolladas a corto y mediano plazo y que pueden complementar al trabajo presentado en esta tesis para mejorarla y enriquecerla y así, robustecer y aumentar la calidad del proyecto.

Los trabajos futuros que se han planteado, se enlistan a continuación:

- Generación de diversos formatos de Reglas de Producción Ponderadas en el Sucedente, para poder ser implementadas en otro lenguaje de representación de conocimiento.
- Creación de reglas de producción ponderadas en el sucedente con múltiples de ellos, tomando en cuenta un peso para cada uno de los sucedentes.
- Numeración personalizada de las reglas y proposiciones generadas.
- Selección de las reglas y las proposiciones, para de esta forma, dar la posibilidad al especialista de que haga su propia elección de reglas y proposiciones.
- Análisis de la calidad de las reglas de producción ponderadas en el secedente generadas.

Referencias

[Agrawal, 1994] Agrawal Rakesh y Srikant Ramakrishnan. "Fast Algorithms for Mining Assocition Rules" *Morgan Kaufmann Publishers Inc, Proceeding of the 20th International Conference on Very Large Data Bases*, (September 1994): Págs. 487-499.

[Alonso, 2004] Alonso M. A, De la Cruz A. V y Gutiérrez G. "Knowledge Representation Language: HAries". *Memorias de la 8th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2004).* Orlando, Florida. USA. Julio 2004. Págs. 358-361.

[Alonso, 2009] Alonso María A, De la Cruz Argelio V y Barcelo Grettel. "Pronóstico para la inyección de tenso-activos en pozos de petróleo a partir de una metodología que integra técnicas de inteligencia artificial y minería de datos". *INCI, Vol.3, No.10*. Oct 2009. Págs. 703-709.

[Bao, 1995] Bao Tu, "An Approach to Concept Formation Based on Formal Concept Analysis". *IEICE Trans. Information and Systems E78-D(5).* 1995. Págs. 553-559.

[Bao, 1996] Bao Tu. "Integrating Inductive Learning and Knowledge Acquisition in the Expert System Generator TESOR" *Cognizant Communication Corporation, 3rd World Congress on Expert Systems*, (Soul 2006): Págs. 925-932.

[Bao, 2002] Bao Tu. "Rule Induction in Constructing Knowledge-Based Decision Support." *Springer US, Decision Support System for Sustainable Development*, (2002): Págs. 263-276.

[Brookshear, 2005] Brookshear Glenn. Computer Science An Overview. Boston: Pearson International Edition: 9th Edition. 2005. Pág. 615.

[Carrasco, 2002] Carrasco Jesús. "Sensitivity Analysis in Logical-Combinatorial Pattern Recognition" *Computación y Sistemas, Vol 6, No 001*. 2002. Págs. 62-66.

[Carrasco, 2003] Carrasco Jesús y Martínez J-F. "Editing and training for ALVOT, an evolutionary approach". *Intelligent Data Engineering and Automated Learning, Volume 2690*. Springer-Verlag Berlin . 2003. Págs. 452-456.

[Carrasco, 2004a] Carrasco Jesús y Shulcloper JR. “Sensitivity analysis of fuzzy Goldman typical testors”. *Fuzzy Sets and Systems, Volume 141, Issue 2*. Elserver Science. 2004. Págs. 241-257.

[Carrasco, 2004b] Carrasco Jesús y Martínez J-F. “Feature selection for natural disaster texts classification using testers”. *Intelligente Data Engineering and Automated Learning Ideal 2004, Book Series: Lecture Notes in Computer Science. Vol. 3177*. 2004. Págs. 429-429.

[Clark, 1989] Clark Peter y Niblet Tim**.** “The CN2 Induction Algorithm”. *Machine Learning Volume 3, Inssue 4*. Kluwer Academic Publishers. (1989): Págs. 261-283.

[Cumplido. 2006] Cumplido René, Carrasco Jesús y Feregrino Claudia. “On the design and implementation of a high performance configurable architecture for testor identification”. *Progress in Pattern Recognition, Image Analysis and Aplications, Volume 4225*, Springer-Verlag Berlin. 2006. Págs. 665-673.

[Dash, 2000] Dash Manoranjan, Liu Huan y Motoda Hiroshu. “Consistency Based Feature Selection”. *Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, (200): Págs. 98-109.

[De la Cruz, 1996] De la Cruz, Argelio. V. “Representaciones del Conocimiento para la Construcción de Sistemas Expertos con Inteligencia Artificial”. *Tesis para optar por el grado de Doctor en Ciencias Técnicas*, *Instituto Politécnico Superior José Antonio Echeverría*. Ciudad de la Habana, Cuba. 1996.

[De la Cruz, 2002] De la Cruz Argelio V y Alonso Maria A. “The HAries environment (v6.00) for the development of intelligence Systems”. *Hífen 26*. 2002. Págs. 184-186.

[Duckett, 2008] Jon Duckett. Beginning Web Programming with HTML, XHTML, and CSS. Wrox. 2008. Pág. 768.

[Friedman, 1997] Friedman Nir, Geiger Dan y Goldszmidt Moises, "Bayesian Network Classifiers". *Springer Netherlands, Machine Learning. Volume 29, issue 2*. Pág. 131-163.

[González, 2007**]** González Ernesto, Espinosa Ivet y Pérez Zady. “Obtención de patrones y reglas en el proceso académico de la Universidad de las Ciencias Informáticas utilizando

técnicas de minería de datos.” *III Conferencia Científica Universidad de las Ciencias Informáticas,* (Octubre 2007).

[Grünwald, 2005] Grünwald Peter. Introducing the Minimum Description Length Principle en Grünwald Peter, Myung In Jae y Mark A**.** Pitt. *Advances in Minimum Description Length: Theory and Applicactions,* Cambridge: The MIT Press. 2005. Págs 1 - 21.

[Hájek, 2004] Hájek Petr, Holeña Martin y Rauch Jan. The GUHA Method and Foundations of (Relational) Data Mining en De Swart Harrie, Orlowska Ewa, Schmidt Gunther y Roubens Marc. Theory and Applications of Relational Structures as Knowledge Instruments COST Action 274, TARSKI, Heidelberg Springer Berlin. (Jan 2004).

[Hangos, 2002] Hangos Katalin, Lakner Rozália y Gerzson Miklós. “G2: An Example of a Real-Time Expert System”, *Springer US, Intelligent Control System, Applied Optimization,* (2002). Págs. 227-250.

[He, 2006] He Jieyue, Hu Hac-Jin, Harrison Robert, Tai Phang y Pan Yi. “Rule Generation for Protein Secondary Structure Prediction With Support Vector Machines and Decision Tree”. *IEEE Transactions on Nanobioscience, Vol. 5, No. 1,* (March 2006). Págs. 46-53.

[Hernández, 2004] Hernández José, Ramírez M. José y Ferri Cesar. Introducción a la Minería de Datos. Madrid: Pearson Educación. 2004. Pág. 680.

[Klir, 1995] Klir George y Yuan Bo. Fuzzy Sets and Fuzzy Logic Theory and Applications. New Jersey. Prentice Hall. 1995. Pág. 547.

[Klösgen, 2002] Klösgen Willi y Żytkow Jan. Handbook of Data Mining and Knowledge Discovery. New York: Oxford University Press. 2002. Pág. 1026.

[Lazo-Cortes, 2001] Lazo-Cortes Manuel, Ruiz-Shulcloper José y Alba-Cabrera Eduardo. “A overview of the evolution of the concept of testor”. *Pattern Recognition, Volume 34, Issue 4,* April 2001. Págs. 753-762.

[Leondes, 2000] Leondes Cornelius. Knowledge Based System Techniques and Applications. San Diego: ACADEMIC PRESS. 2000. Pág. 1449.

[Liu, 2008] Liu Huan y Motoda Hiroshi. Computational Methods of Feature Selection. Florida: Chapman & Hall/CRC. 2008. Pág. 411.

[Mark, 2009] Hall Mark, Frank Eibe, Holmes Geoffrey, Pfahringer Bernhard, Reutemann Peter y Witten Ian. "The WEKA Data Mining Software: An Update." *SIGKDD Explorations*. Volume 11, Issue 1. (2009).

[Molina, 2001] Molina Carlos y Ribeiro Sabine. "Descubriendo Conocimiento para el Mejoramiento Genético Bovino usando Técnicas de Data Mining." *In Proc. of the 4th. Congreso Catalán de Inteligencia Artificial (CCIA'2001).* Barcelona, España. (October 2001): Págs. 123-130.

[Quinlan, 1993] Quinlan J-Ross. C4.5 Programs for Machine Learning. San Francisco CA. USA. Morgan Kaufman Publishers (1993). Pág. 302.

[Pajares, 2006] Pajarez Gonzalo y Santos Matilde. Inteligencia Artificial e Ingeniería del Conocimiento. Madrid, España. Alfaomega Grupo. 2006. Pág. 364.

[Riaño, 1997] Riaño David y Cortés Ulises. "Rule Generation and Compactation in the WWTP" *Computacion y Sistemas. Vol. 1, No. 2,* (1997). Págs. 77 – 89.

[Richards, 2002] Debbie Richards. "Ripple down rules: a technique for acquiring knowledge" en Mora Manuel, Forgionne Guissepi y Gupta Jatinder. *Decision making support systems: achievements, trends and challenges for the New Decade*. Hershey, PA: IGI Publishing. 2002. Págs. 207-226.

[Robnik, 2003] Robnik Marko y Kononenko Igor. "Theoretical and empirical analysis of Relief and RReliefF" *Springer Netherlands Machine Learning*, Volume 53, Number 1-2 (octuber 2003), Págs. 23-69.

[Roda, 2001] Roda. I, Comas. J and Poch. M, Automatic Knowledge Acquisition from Complex Processes for the Development of Knowledge-Based System, Ind. Eng. Chem. Res. 2001. Págs. 3353-3360.

[Russell, 2003] Russell Stuart y Norvig Peter. Artificial Intelligence A Modern Approach. New Jersey: Pearson Education, Inc: Second edition. 2003. Pág. 1081.

[Sánchez, 1997] Sánchez Miquel, Cortés Ulises, Béjar Javier, De Grácia Joan, Lafuente Javier y Poch Manel. *"*Concept Formation in WWTP by Means of Classification Techniques: A Compared Study". *Applied Intelligence, volume 7, number 2*. Págs. 147-165.

[Sánchez, 2002] Sánchez Guillermo y Lazo Manuel. "Modificaciones al Algoritmo BT para Mejorar sus Tiempos de Ejecución". *Revista Ciencias Matemáticas, Vol. 20, No. 2*. Cuba. 2002. Págs. 129-136.

[Santos, 2004] Santos José, Carrasco Ariel y Martinez José "Feature Selection using Typical Testors applied to Estimation of Stellar Parameters". *Computación y Sistemas, Vol. 8, No. 1*. Págs. 15-23.

[Silberschatz, 2002] Silberschatz Avi, Korth Hank y Sudarshan S. Fundamentos de Bases de Datos. Madrid: Mc Graw Hill, 4ta edición. 2002. Pág. 797.

[Soibelman, 2002] Soibelman Lucio y Kim Hyunjoo. "Data Preparation Process for Construction Knowledge Generation through Knowledge Discovery in Databases". *Journal of Computing in Civil Engineering Vol. 16, No. 1*, ASCM (2002). Págs. 39-48.

[Somuano, 2004] Somuano Ochoa J, Reyes Salgado. G, "Reducción de Dimensiones usando el Enfoque Lógico Combinatorio". *3er. Congreso Internacional sobre Innovación y Desarrollo Tecnológico, CIINDET 2004. IEEE Sección Morelos*. Noviembre 2004. Págs. 15-19.

[Stefik, 1995] Stefik Mark, Reasoning about Uncertainty and Vagueness en Stefik Mark. *Introduction of Knowledge System*. San Francisco California: Morgan Kaufmann Publisher, Inc. 1995. Págs. 460-539.

[Suehring, 2009] Suehring Steve, Converse Tim y Park Joyce**.** PHP6 and MySQL Bible. Indianapolis: Wiley Publishing, Inc. 2009. Pág. 873.

[Torres, 2006] Torres María, Torres Aurora y Ponce de León Eunice. "Algoritmos Genéticos y Testores Típicos en el Problema de Selección de Subconjuntos de Características". *Revista Iberoamericana de Sistemas, Cibernética e Informática, Vol. 3, No. 2*. 2006. Págs. 1-5.

[Vega-Alvarado, 2001] Vega-Alvarado L y Ortiz-Posadas M. “Análisis de una Muestra de Pacientes con Lacio Paladar Hendido usando un Algoritmo de Tipicidad y Contraste. *Memoria II Congreso Latinoamericano de Ingeniera Biomédica*, Habana 2001.

[Wada, 2001] Wada Takuya, Motoda Hiroshi y Washio Takashi. “Knowledge Acquisition from Both Human Expert and Data.” *Advances in Knowledge Discovery and Data Mining*, Volume 2035/2001, (Enero 2001). Berlin. Págs. 550-561.

[Wang, 2005] Wang Lipo. Support Vector Machine: Theory and Applications. Poland: Springer. 2005. Pág. 431.

[Wang, 2006] Wang Shu-Tian y Chen Jun-Min. “Weighted Fuzzy Production Rules Using Neural Networks”. *Machine Learning and Cybernetics, 2006 International Conference on.* (August, 2006). Dalian, China. Págs. 3059 – 3062.

[Zhang, 2002] Zhang Chengqui y Zhang Shichao. Association Rule Mining Models and Algorithms. Pittsburgh: Springer. 2002. Pág. 236.

[Zhang, 2004] Zhang M y Yao J.T. “A Rough sets Based Approach to feature Selection”. *Fuzzy Information, 2004. Processing NAFIPS '04. IEEE Annual Meeting of the* , vol.1. (Septiembre 2004). Págs. 434-439.

Referencias electrónicas

[Timarán, 2009] Timarán Ricardo. "TariyKDD. " http://tariykdd.berlios.de/. University of Nariño. (Consultado el 11 de septiembre 2009).

[Zupan, 2009] Zupan. B, G. Leban Demsar y Janez Tomaz. ”Orange Data Mining Fruitful & Ful.” http://www.ailab.si/orange/. (Consultado el 14 de Septiembre 2009).