Revista Ingeniería y Región
2014 (12): 31-40

Funcionalidades de la minería de datos

Features of data mining

Ferley Medina Rojas[1], y Cristina Gómez Santamaría[2]

Resumen

En este documento se ha revisado una metodología y los algoritmos utilizados para abordar un problema de predicción o clúster de datos de acuerdo a la información solicitada. La minería de datos emerge de las áreas de base de datos (data base), repositorio de datos (Data Warehouse) y de las grandes bases de datos (big Data), como un proceso de extracción de información fundamentado en la matemáticas y la estadística. Siendo necesario realizar la selección del modelo, la exploración de los datos, la clasificación de datos, la predicción de valores en función de los datos, el modelamiento de las dependencias para resolver el problema, el descubrimiento de nuevas reglas y visualizar los resultados, con lo que se realiza el análisis e interpretación de la información obtenida.
Es así, como algunas de las aplicaciones de la minería de datos son: en la educación, en la multimedia, en el comercio, en el sector financiero, en la medicina, en el sector agropecuario, en las ciencias sociales, en la gestión gubernamental, y en la tecnología. Para realizar el proceso de extracción de los datos solicitados de estas aplicaciones se requiere el uso de algunos algoritmos como los de regresión lineal, y logística, redes bayesianas, bayes naive, árboles y reglas de decisión, lógica e inferencia difusa y redes neuronales.

Palabras clave: Algoritmos; minería de datos; clasificación de la minería de datos

Abstract

This document has been revised methodology and algorithms used to address a problem of prediction or cluster data according to the information requested. Data mining areas emerge database (data base), data warehouse (Data Warehouse) and large databases (Big Data), as a process of information extraction based on the mathematics and statistics. Being necessary to perform model selection, data exploration, data classification, prediction of values based on the data, the modeling of dependencies to solve the problem, the discovery of new rules and visualize the results, with so the analysis and interpretation of the information obtained is obtained.

Some applications of data mining are: in education, in the media, in commerce, in the financial sector, in medicine, in agriculture, in social sciences, in public administration, and the technology. To made the extraction process request data, using some algorithm like, linear and logistic regression, Bayesian networks, naive Bayes, trees and decision rules, logic and neural networks and fuzzy inference is required.

Keywords: algorithms; mining data application; mining data classification

1 - Msc, Telemática. Estudiante Doctorado en Ingeniería. Universidad Cooperativa de Colombia sede Neiva, calle 11 No. 1g 31. ferley.medina@campusucc.edu.co
2 - PhD. Ingeniería. Área Telecomunicaciones. Universidad Pontificia Bolivariana sede Medellín, circular 1 No. 74 22, cristina.gomez@upb.edu.co

Recibido 29 de Julio de 2014 Aprobado 1 de Noviembre de 2014

Introducción

La minería de datos, como el proceso de extracción de la información dentro de los grandes volúmenes de datos (Big Data) (Londhe, *et al.*, 2013), de origen matemático, a través de algunos teoremas de eficacia de algoritmos y ecuaciones lineal, cuadratica entre otras, sumado, un tratamiento estadístico y un aprendizaje automático de los datos, permiten demostrar la utilidad de las cosas en las diferentes áreas del conocimiento y del saber, antes de la puesta de su funcionamiento.

Es así, como en la hora de tomar un problema se debe seleccionar el modelo, teniendo presente el objetivo a lograr, el tipo y la exploración de los datos; clasificación de los datos, de acuerdo al umbral establecido en el contexto de los datos y los algoritmos empleados (Rutkowski, *et al.*, 2014); predicción de valores en función de los datos, empleando funciones matemáticas o estadísticas para realizar extrapolación, interpolación, regresión, inducción, deducción, transducción en establecer los patrones de comportamiento; modelamiento de las dependencias para resolver el problema, basados en los datos de la predicción o de cluster de los valores en función de los datos se organiza el modelo que se ajuste al error establecido como permitido; descubrimiento de nuevas reglas, producto del entrenamiento, de las asociaciones y condiciones de los datos en el algoritmo seleccionado para la obtención de estas; y la visualización de los resultados como, un producto para la realización de los análisis e interpretación de la información obtenida (Mishra, *et al.*, 2012).

En este documento, se describen algunas de las aplicaciones que la minería de datos tiene como en la educación, para el agrupamiento de las páginas visitadas y las actividades que realizan los estudiantes en sus cursos (Peña, 2014); en la multimedia para realización de la recuperación de la información por contenidos similares; en el comercio y sector financiero, para la evaluación de los clientes; en la medicina, para la identificación de las relaciones en el suministro de fármacos; en el sector agropecuario, para el reconocimiento de patrones durante el desarrollo de las diferentes etapas de los cultivos; en las ciencias sociales y gestión gubernamental, para la evaluación de los resultados producto de aplicación de políticas sociales; y en la tecnología, como una forma de realizar la evaluación operativa o funcional de los nuevos avances.

La utilización de los algoritmos, es una de las formas de lograr los objetivos de la minería de datos, estos son empleados en la exploración (Zorrilla & Garzia, 2013), para realizar asociaciones y clustering; en la estadística, realizando funciones de regresión lineal y logística, redes bayesianas y bayes naive; en la clasificación de datos, formando árboles y reglas de decisión (Hajian & Domingo, 2013); en el aprendizaje automático, el uso de la lógica e inferencia difusa y las redes neuronales; y para finalizar, los usados en la serie temporal para el comportamiento de variables.

1. Que hace la minería de datos

Emerger de las áreas de base de datos (data base), repositorio de datos (Data Warehouse), la estadística, el aprendizaje automático, la visualización de datos, la búsqueda y recuperación de la información y de la computación de alta ejecución, para elaborar procesos esenciales donde se aplican una serie de métodos inteligentes para poder extraer y descubrir patrones de los datos, mediante el uso de algoritmos como regresión lineal, regresión logística, de asociación, lógica difusa, árboles de decisión y redes neuronales entre otros, que permiten obtener un conocimiento histórico y prospectivo para una toma de decisiones en el área de estudio o de conocimiento (Gorbea, 2013).

En minería de datos, el proceso de abordar un problema se sugiere tener en cuenta la realización de las siguientes etapas:

1.1. Selección del modelo

Con la estructuración del problema basado en el tipo de datos y el objetivo que se quiere obtener, los datos se deben preparar y explorar de manera que se pueda llegar a la generación del modelo, con el cual se logra dar explicación al comportamiento de los datos, para posteriormente realizar su validación, implementación y actualización (Riquelme, *et al.*, 2006).

1.2. Clasificación de los datos

Agrupa las reglas de asociación, decisión, métodos, funciones, núcleos, vectores para descubrir las relaciones entre los atributos de un conjunto de datos de acuerdo a un porcentaje de error permitido (umbral), descubriendo relaciones de asociación, segmentación, dependencias funcionales y no funcionales entre los diferentes atributos. A partir de la predefinición o búsqueda de las clases categóricas o de reglas, toma un dato y lo ubica dentro del rango determinado, haciendo para ello el uso de algoritmos.

1.3. Predicción de valores en funcion de los datos

Basado en funciones matemáticas o estadísticas como:

La interpolación, la cual estima un valor dado dentro de los límites de una función, construida a partir de unos valores conocidos.

La extrapolación, la cual estima un valor dado por fuera de los límites de una función, construida a partir de unos valores conocidos.

Regresión, técnica estadística para crear relación de variable, dependiente e independiente, para que a partir de un atributo de entrada o un valor de predicción obtener el valor estimado o de salida de acuerdo a un valor de error permitido.

También, el uso de la inferencia a través de sus funciones de:

Inducción, es el desconocimiento de la dependencia de las entradas y las salidas o de la estructura de un sistema, usando un número limitado de observaciones o de mediciones de entradas y salidas del sistema, para lo cual toda la información está organizada y la muestra está definida por un par de entrada y salida.

Deducción, al modelo definido se le aplica, las entradas para llegar a las salidas previstas de forma que se pueda confrontar las dependencias o las asociaciones de las variables.

Transducción, a partir del entrenamiento de los datos se puede encontrar relaciones o reglas que con llevan a las salidas previstas (Kantardzic, 2011).

Permitiendo lograr el establecimiento de patrones de comportamientos, para el uso del seguimiento, cálculo, diseño, creación de diferentes actividades de los seres, en una variedad de las áreas del saber.

1.4. Modelamiento de las dependencias para resolver el problema

Usando los datos disponible en el proceso de aprendizaje inductivo, deductivo, transducción y estimación, se predicen las dependencias desconocidas, con que se crea el modelo para el entrenamiento de los datos que satisfaga el porcentaje de error permitido y que genere el nuevo modelo.

1.5. Descubrimiento de nuevas reglas

Con el entrenamiento de los datos, las dependencias o las asociaciones de los atributos de entrada con los de salida y sus predicciones, acompañadas del porcentaje de error permitido en las muestras de los datos, de acuerdo al algoritmo seleccionado se obtienen nuevas reglas y condiciones que pueden hacer variar las entradas o descubrir nuevas dependencias o asociaciones entre los datos de la muestra.

1.6. Visualización de los resultados

La información obtenida de la minería de datos se muestra para permitir su análisis e interpretación, medir el alcance del objetivo, la comprensibilidad de los patrones extraídos, y la facilidad del modelo. En la figura 1, se muestra al departamento del Huila con la división política de los 37 municipios, 37 corregimientos, las zonas agropecuarias actuales cultivadas en café y mojarra de acuerdo a las características de las variables de pH, humedad, oxígeno y las propiedades del suelo. Las cuales se pueden convertir en variables decisorias en la hora de establecer patrones.

2. Aplicaciones de la minería de datos

En la revisión de la literatura se ha encontrado algunos ejemplos de aplicaciones como son las siguientes:

En la educación, mediante el uso del algoritmo de clustering, para agrupar las páginas por su contenido, acceso, tipo, obteniendo un sistema de personalización, de recomendadores, de modificación, de irregularidades, regularidades, de los contenidos, descubrimiento de relaciones entre actividades, clasificaciones y diagnóstico incremental de estudiantes, con lo que la universidad tiene un perfil de ellos, para hacer una construcción adaptativa de los planes de enseñanza según las capacidades de los estudiantes, ofreciéndoles a estos diferentes materiales o ayudas didácticas, permitiéndole a la universidad planificar de acuerdo la necesidad de los estudiantes de tener a su disposición los medios educativos.

Mediante el uso del algoritmo de reglas de asociación, se puede descubrir las relaciones entre los atributos de un conjunto de datos de acuerdo al umbral dado, determinando la asociación entre las distintas páginas Web visitadas. También aprovechando el registro de los log de las entradas a los servidores, se descubren patrones de navegación y se puede proponer uno, teniendo la relación entre las páginas visitadas y los documentos leídos.

El algoritmo de análisis de secuencia, determina las páginas visitadas durante una o varias sesiones de un

mismo usuario y su orden, encontrando patrones de navegación y de comportamiento que determinan caminos de aprendizajes (Romero, *et al.*, 2005).

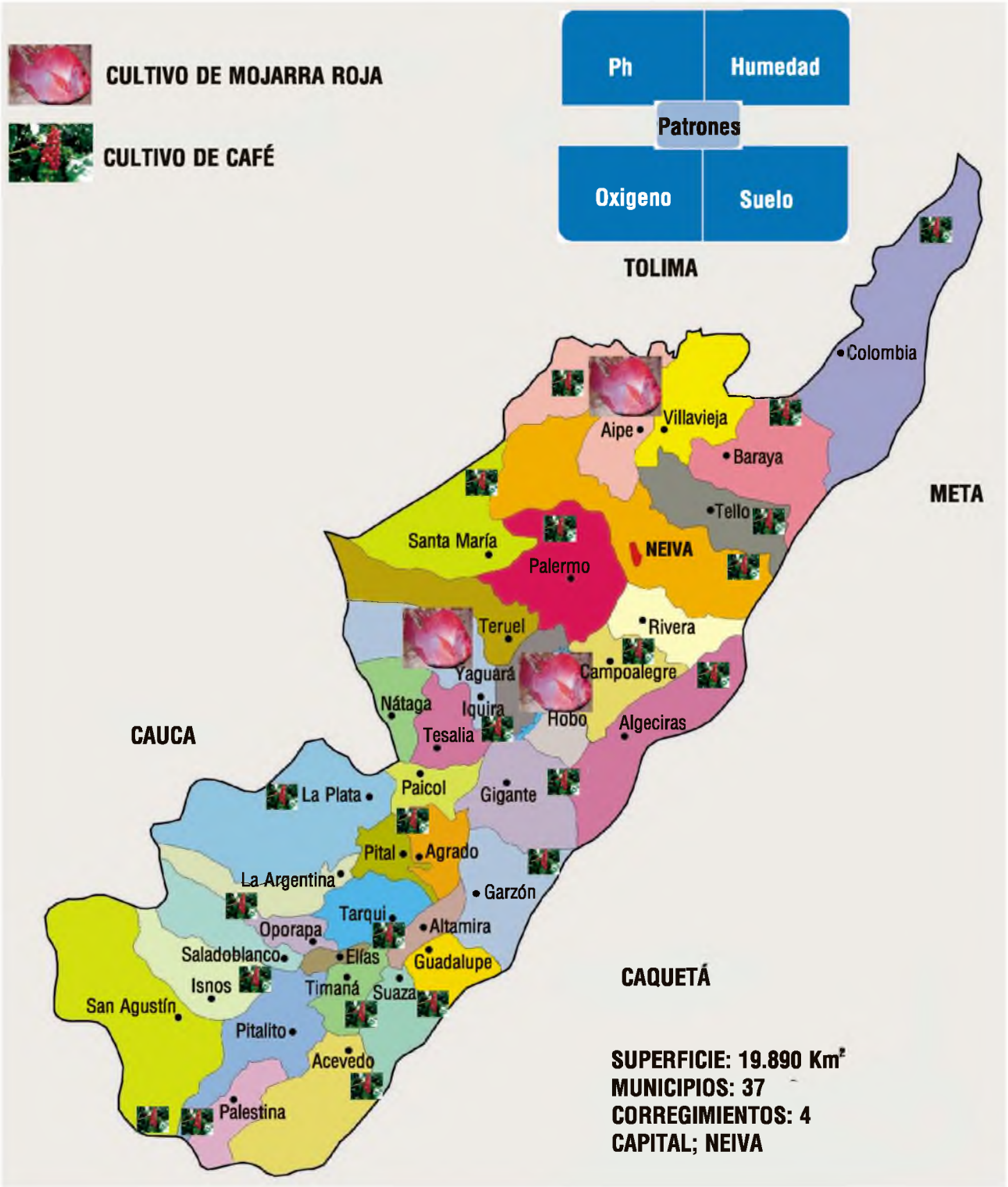

Figura 1. Zonificación de los cultivos de mojarra roja y café en el departamento del Huila

En la multimedia, la recuperación de la información se hace por contenidos, a través de la similitud de objetos, en una función distancia, desarrollada por la conformación de puntos en un espacio vectorial métrico lográndose la caracterización de interés en unos valores numéricos de los objetos. Esto funciona mediante soluciones algorítmicas que diseñan y construyen nuevas estructuras de datos que incluyen, la creación y recorrido de índices (Fernández, *et al.*, 2011).

En el comercio y sector financiero, con el uso del algoritmo del método de análisis discriminante lineal se puede crear dos grupos de variables, una independiente y la otra dependiente, basado en que los datos son distribuidos de forma normal teniendo en cuenta para ello el coeficiente de correlación. También se puede utilizar el análisis de varianza para comparar los valores encontrados en las dos distribuciones y las series de tiempo para los rangos de fechas en las cuales se hace el estudio. Con lo anterior se evalúa el comportamiento de los clientes en una empresa o sector financiero, determinando según criterios, si son buenos o malos clientes y sus preferencias de compras, de acuerdo a los registros de información de ellos (Mylonakis, 2010).

En la medicina, para identificar las relaciones en el suministro de un fármaco sobre otro fármaco, basados en el enfoque de las propiedades de administración, metabolismo, reacción de un fármaco sobre el otro. Empleando el algoritmo de reglas lógicas, implementado en un árbol analítico con hojas etiquetadas por palabras enlazadas por un par de palabras dentro de una frase, almacenando las sentencias de información sintáctica y semántica en una base de datos, para posteriormente, mediante el uso de instrucciones de SQL, encontrar las relaciones de los fármacos (Tari, *et al.*, 2010).

En el sector agropecuario, en cultivos agrícolas (Díaz & Pérez, 2004) y pecuarios, haciendo reconocimiento de patrones producto de las tecnologías usadas como la red de sensores, los sistemas de posicionamiento geográficos, procesamiento de imágenes híper espectral que producen los registros de los datos de plagas, malezas, centros de mercadeo y de producción que ayuden a garantizar la sostenibilidad alimenticia del mundo, como también los ingresos económicos de sus agricultores. Para lo que es necesario garantizar el acceso a los datos desde una gran variedad de dispositivo móvil, para que su acceso tenga la aplicabilidad en el sitio de cultivo. El uso de algoritmos de inteligencia artificial, ha permitido el reconocimiento de patrones con los que se ha originado al máximo aprovechamiento de la tierra, el manejo sostenible del recurso del agua y la definición de políticas para la explotación de cultivos de madera (Bauckhage, 2013).

En las ciencias sociales y gestión gubernamental (Arabí, 2013), como un instrumento eficaz para soportar las mediciones de los patrones socio-económico que permite evaluar al mismo tiempo muchas preguntas, probar varias hipótesis o poder comparar diferentes puntos de estimación, de políticas gubernamentales, el comportamiento de los indicadores sociales alcanzados en las escuelas, estados o países. Dando un tratamiento adecuado a cada grupo de la población para lograr el desarrollo de la comunidad o de la sociedad. Utilizando el algoritmo de la regresión logística a cada nodo se le da la clase del nodo vecino con lo que se logra caracterizar al nodo, empleando etiquetas independientes y dependientes para determinar la probabilidad de cada clase (Grian & Tina, 2010).

En la tecnología, utilizado para mejorar la velocidad de entrenamiento de los datos que son obtenidos por los altos volúmenes de datos que arrojan las redes móviles, con el uso de algoritmos de la técnica del clasificador integrado, partiendo del concepto de la clasificación Bayesiana logra al mismo tiempo clasificar los datos y entrenarlos. Usando la técnica de muestreo, que incluye el muestreo aleatorio simple, (con la definición de una probabilidad fija sobre el conjunto de los datos se determina si el dato es procesado) el muestreo estratificado, (se divide la muestra en sub muestra de acuerdo al número de sub conjuntos encontrados) y el muestreo poblacional (similar al muestreo estratificado, pero, inicia a sub dividir según la secuencia de los datos) (JianPing, 2012).

La minera de datos ha sido aplicada en una variedad de ciencias, disciplinas y saberes como la educación, la multimedia, el comercio y sector financiero, en la medicina, en las actividades agropecuarias, agroforestales, en la ciencia social, en la tecnología y en la gestión gubernamental. En general su aplicación está marcada en todas las actividades que realiza y procesa el ser humano en donde se presente la generación de datos con los cuales se construye un modelo para proporcionar un conocimiento acerca de pronósticos, riesgos, probabilidades, recomendaciones, búsqueda de secuencias y agrupaciones (Microsoft, 2012).

3. Algoritmos en la minería de datos

Los algoritmos en la minería de datos, dependiendo del área a estudiar, del modelo, la complejidad de los datos, las reglas, agrupaciones, asociaciones, condiciones, correlaciones se seleccionan procediendo a su aplicación para encontrar la información que no era visible o que está incompleta (Witten, *et al.*, 2011).

3.1. Exploración

Se utiliza para encontrar las relaciones sistemáticas entre las variables cuando hay poco o ningún conocimiento de lo que el próximo resultado puede ser. La exploración sólo funciona como un primer paso para la predicción de un modelo.

3.1.1. De Asociación

Lo compone una serie de conjuntos de elementos y de reglas que describen como estos se agrupan dentro de los casos, siendo estas reglas usadas para predecir la presencia de un elemento en la base de datos, basándose en la manifestación de un elemento específico identificado como importante. A la creación de estos conjuntos, se les da una puntuación que representa el soporte y la confianza de su clasificación, de la cual se derivan las reglas relevantes de los conjunto de datos. Su mayor uso está en la predicción de la manifestación de los gustos de las compras de los clientes o casos similares (Nebot & Berlanga, 2012).

3.1.2. Clustering

Basado en técnicas iterativas, para agrupar los casos de un conjunto de datos dentro de clúster que contienen características similares las cuales son útiles para la exploración de datos, la identificación de anomalías, en los datos y la creación de predicciones, para futuros comportamientos. El entrenamiento del modelo se hace a partir de las relaciones existentes entre los datos y la de los clústeres que identifica el algoritmo. El cálculo del grado de perfección con que los clúster representan las agrupaciones para crear el clúster que mejor representa los datos, se realiza mediante el uso de iteraciones, hasta cuando ya no sea posible mejorar los resultados de la redefinición de los clústeres (Parimala M., *et al.*, 2011).

3.2. Estadísticos

Estos algoritmos, funcionan sobre datos a los cuales se les trata de encontrar una función matemática como, la lineal, cuadrática, exponencial y logarítmica que mejor los correlacione, siendo necesario que su coeficiente de correlación estén los más cercano posible al menos uno (-1) o al uno positivo (+1), teniendo su relevancia en la minería de datos para corregir los datos faltantes en una muestra, clasificarla o realizar procesos de predicción.

3.2.1. Regresión Lineal

Utilizado para la clasificación o análisis de la asociación, la predicción a partir de la relación de una variable independiente y otra dependiente, formando una ecuación lineal que mejor representa la serie de datos, con los coeficientes a y b que ajustan el ángulo y la ubicación de la recta de regresión hasta que la suma de los valores asociados a todos alcance un valor mínimo y su coeficiente de correlación adecuado (Romero & Ventura, 2010).

3.2.2. Regresión Logística

Su nombre está dado por el hecho que la curva de los datos se comprime mediante una transformación logística (convierte a la función exponencial en lineal) para minimizar el efecto de los valores extremos, es una variación del algoritmo de red neuronal, basado en la estadística que se emplea para determinar la contribución de varios factores a un par de resultados. Es la probabilidad de que algunos eventos ocurran como una función lineal de un conjunto de variables predictoras (Westreich, *et al.*, 2010).

3.2.3. Redes Bayesianas

Llamadas redes de creencias o probabilísticas. Permite invertir las dependencias de los atributos; con la distribución previa, proporciona la forma de incorporar información externa al proceso de análisis de los datos, dando una distribución posterior, la cual vuelve hacer actualizada con la distribución previa. La forma de calcular la probabilidad posterior, la probabilidad de la condición por la probabilidad a priori, todo este conjunto divido por la condición marginal.

Todos los atributos son nominales y todos los valores son faltantes. Los atributos ocultos que tienen valores que no se pueden observar, son generados por algoritmos avanzados de aprendizaje que los agrega a los datos actuales caracterizando mejor al modelo para llegar al fondo del problema.

El algoritmo inicia dando un orden determinado a los atributos, los cuales se denominan nodos. El procesamiento de cada nodo se hace a la vez que va sumando los bordes de los nodos previamente procesados hasta llegar al nodo actual. En cada paso se adiciona el borde que maximiza el puntaje de la red. Cuando no hay un cambio en el puntaje de la red se inicia con el siguiente nodo, como un mecanismo para evitar el sobre ajuste. El número de padres de cada

nodo se puede restringir con un valor predefinido, debido a que solo los borde de nodos procesados son considerados en el ordenamiento fijo, este proceso no puede ser cíclico, porque los resultados dependen de cómo se haya establecido el ordenamiento inicial, para lo cual el algoritmo se puede ejecutar varias veces con diferentes formas ordenamientos al azar. Utilizado para procesos de clasificación y predicción de datos (Xiong, *et al.*, 2013).

3.2.4. Bayes Naive

Método heurístico basado en los teoremas de Bayes, sin las dependencias que puedan existir, en donde se emplea la teoría de la probabilidad, para encontrar lo más probable (Ranchi & Jaunpur, 2013). Utilizado para la clasificación de texto, descubriendo relaciones entre las columnas de entrada y las de predicción a partir del hecho que las columnas son independientes. También como clasificador para descartar redundante e irrelevante predictores (Vidaurre, *et al.*, 20013).

3.3. Clasificación de Datos

La clasificación es un proceso de aprendizaje de una función que contiene un dato dentro de algunas de las clases predefinidas. Cada clasificación está basada sobre el algoritmo de aprendizaje inductivo, el cual da como entrada un conjunto de muestras, que consiste en un vector de valores de atributos o vectores de características, y su clase correspondiente. El objetivo del aprendizaje es, crear un modelo de clasificación, conocer el clasificador que se está prediciendo, los valores disponibles de los atributos de entrada, y las clases para alguna entrada (según la muestra).

En otras palabras clasificación, es el proceso de asignación de etiquetas de valores discretos (clases) para un registro sin marcar, y un clasificador, es el modelo (resultado de la clasificación) que predice una clase de atributo de una muestra cuando los otros atributos son dados. Al hacerlo, las muestras son divididas dentro de un grupo predefinido. En un modelo de clasificación, se tiene en cuenta la fuente de los datos, el modelo, la información que se pretende descubrir o hallar, la conexión entre las clases y otras propiedades de las muestras, que pueden ser definidas por un diagrama de flujo o como tan complejas sin estructura como un proceso manual.

Las metodologías de minería de datos restringen la discusión para formalizar modelos de clasificación que sean ejecutables. Un modelo puede ser obtenido por entrevista a expertos relevantes o expertos. También de forma inductiva por la generalización de ejemplos específicos. Diferentes metodologías de clasificación son aplicadas hoy en día en casi todas las disciplinas donde está la tarea de clasificación, porque la gran cantidad de datos requieren automatización de los procesos.

Los árboles de decisiones y las reglas de decisión son metodologías aplicadas en muchas aplicaciones del mundo, ejemplo las tendencias en el mercado financiero y la identificación de objetos en grandes base de datos de imágenes. Un caso particular puede ser un grupo de clientes ordenados por aquellos quienes pagan sus facturas dentro los treinta días y quienes toman más de treinta días para el pago.

3.3.1. Árboles de Decisión

Un método particularmente eficiente de producción de clasificadores desde datos es generar un árbol de decisión. El árbol de decisión es la representación del método lógico. Un árbol de decisión puede ser empleado en un modelo jerárquico de aprendizaje supervisado donde la región local es identificada en una secuencia de divisiones recursivas a través de nodos de decisión con función de prueba. El método de aprendizaje supervisado que construye árboles de decisión, lo hace desde un conjunto de entradas-salidas de las muestras.

El árbol de decisión en el algoritmo de inducción se describe principalmente en el aprendizaje automático, aplicando la literatura de la estadística. Esto es un eficiente método no paramétrico usado para el tratamiento de datos continuos y discretos, con distintos métodos para crear el árbol, admitiendo varias tareas de análisis, la regresión, la clasificación y la asociación.

El árbol se crea con la determinación de las correlaciones entre una entrada y el resultado deseado, terminado de hacer todas las correlaciones, se usa la ecuación que calcula la obtención de la información, identificando un atributo único de mayor puntuación (entropía de Shannon, la red bayesiana con prioridad K2 y la red bayesiana con una distribución Dirichlet uniforme de prioridades) que separa los resultados, de casos en subconjuntos que son analizados de forma recursiva hasta que no se pueda dividir más el árbol. Cada caso tiene una única red bayesiana anterior y una única medida de confianza para dicha red.

Un modelo puede contener varios árboles para distintos atributos de predicción; un árbol varias bifurcaciones, su profundidad y forma está dado por el método de puntuación y del resto de parámetros usados (Barros, *et al.*, 2012).

3.3.2. Reglas de Decision

Se construyen de acuerdo a la descripción mínima de principio de longitud mediante algoritmos ávidos utilizando el enfoque de programación dinámica. Una técnica es utilizar una regla para cada clase. Cada regla es un conjunto de pruebas, una para cada atributo. Para los atributos numéricos la prueba comprueba si los valores se encuentran dentro de un intervalo dado; para cada uno de los nominales ésta comprueba si está en un cierto subconjunto de esos valores de los atributos.

Estos dos tipos de pruebas, la de intervalos y la de subconjunto, aprenden desde el entrenamiento de los datos correspondiente a cada una de las clases. Para un atributo numérico, los puntos del fin del intervalo son valores mínimo y máximo que ocurren en el entrenamiento de los datos para esa clase. Para un nominal, el subconjunto contiene esos valores que ocurre para ese atributo en el entrenamiento del dato para la clase individual. Las reglas, están representando diferentes clases usualmente sobrepuesta, y el tiempo de predicción es predecido con la prueba más parecida.

Usado en la representación del conocimiento de grandes volúmenes de datos estadísticos o de experimentos para construir clasificadores que puedan predecir características de nuevos objetos basados en la información existente para lo cual puede utilizar tablas de decisiones (Chikalov, *et al.*, 2011).

3.4. Aprendizaje Automático

Esto es posible para relacionar el problema de aprendizaje desde las muestras de datos para una noción general de la inferencia en la filosofía clásica. Cada proceso predictivo de aprendizaje consiste en dos fases principales: aprendizaje o estimación de las dependencias desconocidas desde un conjunto de muestras dadas, usando dependencias estimadas para predecir nuevas salidas para futuras entradas de valores del sistema.

El aprendizaje automático, es una combinación de inteligencia artificial y estadística, siendo un área muy útil en la investigación, provee un número de diferentes problemas y algoritmos para su solución. Este algoritmo varía en sus preguntas, en el conjunto de datos disponibles para el entrenamiento, y en las estrategias de aprendizaje y representación de los datos. Todos estos algoritmos, sin embargo aprenden por búsqueda a través de un espacio n dimensional de un conjunto de datos dados para encontrar una generalización aceptable. Una de las tareas más fundamentales en el aprendizaje automático es el aprendizaje automático inductivo donde una generalización es obtenida desde un conjunto de muestras y esta es formalizada usando diferentes técnicas y modelo.

El aprendizaje inductivo se define como el proceso de estimación de una dependencia desconocida de entrada- salida o la estructura de un sistema, usando limitado número de observaciones o mediciones de entradas y salidas del sistema. En la teoría del aprendizaje inductivo, todo dato en un proceso de aprendizaje está organizado y cada par de instancia de entrada-salida usa un simple término conocido como una muestra. En general el escenario de aprendizaje involucra, un generador de entradas aleatorias para el vector X, un sistema que retorna unas salidas Y para dar entrada al vector X y un aprendizaje automático que estima despliegues desconocidos (entrada X, salida Y) desde un sistema observado de muestras (entrada X, salida Y) (Padhy, *et al.*, 2012).

3.4.1. Lógica e Inferencia Difusa

Se puede utilizar en escenarios en los que no existe un modelo de solución simple o matemático preciso. También en el uso del conocimiento de un experto en un tema determinado cuando este es ambiguo o impreciso. Además, en algunas partes desconocidas o en una variable que produce desajuste de otras, de un sistema que se pretende controlar y no se pueden medir de forma fiable. Trabaja con variables lingüísticas o con datos imprecisos, con reglas de tipo IF-THEN, definidas a partir de la opinión de expertos o de un sistema de aprendizaje (red neuronal). El conjunto de las entradas se hacen mediante antecedentes o premisas y de las salidas llamadas consecuente o consecuencia (Diaz & Morillas, 2004) La lógica difusa es el puente entre la alta precisión y la alta complejidad de la difusividad, siendo empleada para el control automático de: vehículos, tráfico, compuertas, máquinas lavadoras y ascensores. Debido a la no linealidad de la lógica booleana en el diseño de sistemas automáticos, se utiliza los sistemas difusos (Hullermeier, 2011).

3.4.2. Redes Neuronales

Su utilidad es brindar análisis a datos de entradas complejos o relaciones complejas entre muchas entradas y relativamente pocas salidas. La red está restringida al mismo atributo de entrada en todos los nodos de cada capa oculta, contiene varias redes dependiendo del número de columnas de entrada y de los estados que contiene cada una de ellas, sus conexiones nodo-objetivo representa reglas de

asociación entre entrada y objetivo. Se utiliza en la clasificación y reconocimiento de patrones de voz, imágenes, fraudes económicos, en la predicción del tiempo atmosférico y del mercado financiero (Swiderski, *et al.*, 2012).

3.5. Serie Temporal

Utilizado para predecir tendencias basadas únicamente en el conjunto de datos original empleado para crear el modelo, en predicción cruzada para crear un modelo general que se pueda aplicar a múltiples series, en la previsión del tiempo de valores continuos.

En aplicaciones médicas, para predecir el tiempo de supervivencia que tiene un paciente que padece de una enfermedad determinada de acuerdo a la sintomatología presentada. En la industria, para establecer la predicción del fallo del componente de una máquina especificada. En general en donde está implicado la supervivencia o la probabilidad de fracaso de un evento (Salvo, *et al.*, 2013).

4. Conclusiones

La aplicación de la minería de datos está marcada en las actividades o disciplinas que realiza y procesa el ser humano de las cuales se generan datos como en el caso de la educación, la multimedia, el comercio, sector financiero, la medicina, la ciencia social, la tecnología, la gestión gubernamental, el sector agropecuario y agroforestal, entre otras; con los cuales se construye un modelo para proporcionar un conocimiento y entrenamiento acerca de pronósticos, riesgos, probabilidades, recomendaciones, búsqueda de secuencias y agrupaciones, similitud de objetos, comportamiento de clientes, la interacción, medición y relación de variables, atributos, patrones y actividades, que contribuyen a facilitar al ser humano la toma de decisiones, la evaluaciones, así también como las investigaciones.

La aplicación de los algoritmos de la minería de datos está definida según el problema que se desee resolver, los tipos, flujos y tamaños de datos, el área de exploración o de investigación, encontrándose que en un modelo se puede aplicar más de un algoritmo o la combinación entre ellos de acuerdo a la etapa en la que se encuentre, o a los resultados esperados (patrones, estimaciones, clasificaciones de usuarios).

5. Referencias Bibliográficas

1. Arabí, U., 2013. Ethical data mining and social science data exploration and descriptcion: scope and limitations in social science research. In: Ethical data mining applications for socio-economic development. United States of america: Idea Group Inc (IGI), pp. 22-39.
2. Barros, R. C., Basgalupp, M. P., Carvalho, A. C. P. L. F. d. & Freitas, A. A., 2012. A Survey of Evolutionary Algorithms for Decision-Tree Induction. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on V42 Issue 3 , pp. 291-312.
3. Bauckhage, C. a. K. K., 2013. Data Mining and Pattern Recognition in Agriculture V 27 No. 4. KI - Kunstliche Intelligenz, pp. 313-324.
4. Chikalov, I., Moshkov, M. & Zielosko, B., 2011. Online Learning Algorithm for Ensemble of Decision Rules. Rough Sets, Fuzzy Sets, Data Mining and Granular Computing V 6743, pp. 310-313.
5. Diaz, D. B. & Morillas, R. A., 2004. MIneria de datos y lógica difusa. Una aplicación al estudio de la rentabilidad económica de las empresas agroalimentarias en Andalucía. Estadística Española Vol 46, Núm. 157, pp. 409-430.
6. Díaz, J. L. A. & Pérez, G. R., 2004. Estado del arte en la utilizacón de técnicas avanzadas para la busqueda de información no trivial a partir de datos en los sistemas de abastecimientos de agua potable. Universidad Politécnica de Valencia. Departamento de ingeniria hidráulica y medio abte. [Online] Available at: http://www.lenhs.ct.ufpb.br/html/downloads/serea/trabalhos/A15_15.pdf
7. Fernández, J. *et al.*, 2011. Indexación y recuperación de información multimedia. La plata Buenos aires, Universidad Michoacana de San Nicolás de Hidalgo, pp. 324-328.
8. Gorbea, P. S., 2013. Tendencias transdisciplinarias en los estudios métricos de la información y su relación con la gestión de la información y del conocimiento. Perspectivas em Gestão & Conhecimento V. 3, N 1, pp. 13-27.
9. Grian, g. & Tina, E. r., 2010. Leveraging label independent Features for classification in sparsely labeled networks: an emprical study. In: Advances in social network mining and analysis. Pennsylvania: Springer Verlag Heidelberg, pp. 1-19.
10. Hajian, S. & Domingo, F. J., 2013. A Methodology for Direct and Indirect Discrimination Prevention in Data Mining. IEEE Transactions on Knowledge and Data Enginneering Vol 25 No. 7, pp. 1445-1459.
11. Hullermeier, E., 2011. Fuzzy sets in machine learning and data mining. Applied Soft Computing V 11 No. 2, pp. 1493-1505.
12. JianPing, G., 2012. The research on model of group behavior based on mobile network mining and high-speed data streams. Emerging computacion and information technologies for education V 146, pp. 473 - 480.
13. Kantardzic, M., 2011. Data mining conceptos, models methods, and algorithms secend edition. New Jersey: John Wiley & Sons, Inc..
14. Londhe, S. R., Mahajan, R. A. & Bhoyar, B. j., 2013. Survey on Mining High Utility Itemset Transactional

Database. International Journal of Innovative Research & Development Vol 2 Issue 13, pp. 43-47.
15. M.Parimala, Lopez, D. & Senthilkumar, N., 2011. A Survey on Density Based Clustering Algorithms for Mining Large. International Journal of Advanced Science and Technology V 31, pp. 59-66.
16. Microsoft, 2012. msdn.microsoft.com. [Online] Available at: msdn.microsoft.com/es-es/library/ms174949.aspx [Accessed 01 10 2013].
17. Mishra, P., Pandhy, N. & Panigrahi, R., 2012. The Survey of Data Mining Applications and Feature Scope. Asian Journal of Computer Science and Information Technology 2:4, pp. 68-77.
18. Mylonakis, J., 2010. Evaluating the likelihood of using linear discriminant analysis as a commercial bank card owners credit scoring model. International Business Research, V 3, No. 2 April 2010, pp. 9-20.
19. Nebot, V. & Berlanga, R., 2012. Finding association rules in semantic web data. Knowledge-Based Systems V 25, pp. 51-62.
20. Padhy, N., Mishra, P. & Panigrah, R., 2012. The Survey of Data Mining Applications. International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012, pp. 43-58.
21. Peña, A. A., 2014. Educational Data Minining: A Survey and a Data Mining-Based Analysis of recent Works. Expert Systems with Applications, pp. 1432-1462.
22. Ranchi, J. & Jaunpur, U. P., 2013. Data Mining Approach to Detect Heart. International Journal of Advanced Computer Science and information Techonology V 2 No. 4, pp. 56-66.
23. Riquelme, J. C., Ruiz, R. & Gilbert, K., 2006. Minería de datos: Conceptos y Tendencias. Revista Iberoamericana de Inteligencia Vol. 10 No 29, pp. 11-18.
24. Romero, C. & Ventura, S., 2010. Educational Data Mining: A Review of the. Transactions on Systems, Man, and Cybernetics--Part C: Applications and Reviews V 40 No. 6, pp. 601-618.
25. Romero, M. C., Ventura, S. S. & Hervás, M. C., 2005. Estado actual de la aplicación de la mineria de datos a los sistemas de enseñanza basada en WEB. Madrid, Universidad de Cordoba, pp. 49-56.
26. Rutkowski, L., Jaworski, M., Pietruczuk, L. & Duda, P., 2014. Decision Trees for Mining Data Streams Based on the Gaussian Approximation. Knowledge and Data Engineering, IEEE Transactions on , vol.26, no.1, pp. 108-119.
27. Salvo, R. D. et al., 2013. Multivariate time series clustering on geophysical data recorded at Mt. Etna from 1996 to 2003. Journal of Volcanology and Geothermal Research , Volume 251, pp. 65-74.
28. Swiderski, B., Kurek, J. & Osowski, S., 2012. Multistage classification by using logistic regression and neural networks for assessment of financial condition of company. Decision Support Systems V 52, pp. 538-547.
29. Tari, L. et al., 2010. Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism. Bioinformatics, Vol 26 , pp. 547 - 553.
30. Vidaurre, D., Bielza, C. & Larranaga, P., 20013. AN L1-REGULARIZED NA¨IVE BAYES-INSPIRED. International Journal on Arti?cial Intelligence Tools V 22 No 4, pp. 1350019-1 1350019-18.
31. Westreich, D., Lessler, J. & Funk, M. J., 2010. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression ". Journal of Clinical Epidemiology V 63 No. 8, pp. 826-833.
32. Witten, L. H., Frank, E. & Hall, M. A., 2011. Data mining practical machine learning tools and techniques, Third edition. Estados Unidos: Morgan Kaufmann Publications.
33. Xiong, W., Cao, Y. & Liu, H., 2013. Study of Bayesian Network Structure Learning. Applied Mathematics & Information Sciences V 7 No. 1L, pp. 49-54.
34. Zorrilla, M. & Garzia, S. D., 2013. A Service Oriented Architecture to Provide Data Mininig Services for Non-expert Data Miner. Decision Support Systems Vol. 55, issue 1, pp. 399-411.