Algoritmos y Tareas para la Mineria de Datos

Ferley Medina Rojas
Universidad Cooperativa de Colombia sede Neiva, Huila, Colombia,ferley.medina@campusucc.edu.co
Cristina Gómez Santamaría
Universidad Pontificia Bolivariana sede Medellín, Antioquia, Colombia cristina.gomez@upb.edu.co

ABSTRACT

Reviewing the current state of the data mining was found that this uses statistical algorithms, estimation or prediction of data and outputs in knowledge representation, which respond to a selection model to a data classification process to estimate new values for the selected attributes to model the dependencies according to the specific attributes, to displays the results of data mining and discovers new rules as a result of the extraction of data to achieve its applicability to certain objective.

Almost most of the fields of know and knowledge is present data mining, that is like it find in the telecommunications, multimedia, education, trade and the financial sector, government management and medicine, extracting data for use in research, trends, habits, behaviors, predictions and estimates.

That is why, it is essential that the selection of the use of different algorithms that meet condition of the problem of the which should to have a level of knowledge for make decision.

Keywords: Data mining, algorithms, applications

RESUMEN

Revisando el estado actual de la minería de datos se encontró que esta emplea algoritmos estadísticos, de estimación o predicción de datos y de salidas en la representación del conocimiento, los cuales responden a una selección del modelo, a un proceso de clasificación de datos, a estimar nuevos valores para los atributos seleccionados, para modelar las dependencias según los atributos específicos, visualizar los resultados obtenidos de la minería de datos y a descubrir nuevas reglas como producto de la extracción de los datos para lograr su aplicabilidad al objetivo determinado.
En casi la mayoría de los campos del saber y del conocimiento esta presente la minería de datos, es así como se encuentra en las telecomunicaciones, la multimedia, la educación, el comercio y el sector financiero, la gestión gubernamental y en la medicina, extrayendo datos para el uso de investigaciones, tendencias, hábitos, comportamientos, predicciones y estimativos.
Es por ello, que resulta fundamental la selección del uso de los diferentes algoritmos que satisfagan las condiciones del problema del cual se pretende tener un nivel de conocimiento para la toma de decisión.

Palabras claves: Minería de datos, algoritmos, aplicaciones

1. INTRODUCTION

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, redes bayesianas, de asociación, árboles de decisión, clustering, inferencia difusa, series

de tiempo y redes neuronales entre otros, que permiten conocer los antecedentes y realizar prospectivas para una toma de decisiones en la área involucrado. (Gorbea, 2013)

2. ALGORITMOS

Para la minería de datos antes de pensar en los algoritmos es necesaria la construcción de un modelo, que permita dar respuesta al tipo de problema, al entrenamiento de los datos y a los datos existentes. Que finalmente redundara en tener una información visible o escondida según los objetivos del procesamiento de los datos, proyecciones, estimaciones, predicciones o conocer las tendencias o preferencias de un conjunto de clientes, de un supermercado, entidad financiera. (Witten, Frank, & Hall, 2011)

2.1 CONSIDERACIONES A TENER EN CUENTA PARA LA SELECCIÓN DE LOS ALGORITMOS

La determinación de la selección del uso de algoritmo a ser aplicado en la minería de datos depende también de la etapa en la que se encuentre la extracción de los datos, por ello se deben involucrar las siguientes:

2.1.1 SELECCIONAR EL MODELO

La estructuración del problema teniendo en cuenta el tipo de datos y el objetivo que se quiere obtener, los datos se deben explorar, preparar (sacar los datos incompletos o llenar los faltantes) y dejar previsto una forma de entrenamiento de manera que se pueda validar, implementar y realizar los ajustes requeridos. (Riquelme, Ruiz, & Gilbert, 2006)

2.1.2 CLASIFICAR LOS DATOS

La clasificación de los datos se puede realizar mediante el uso de reglas de asociación o formulación de condiciones, con lo que se descubre las relaciones entre los atributos de un conjunto de datos que superan los umbrales determinados, está reglas se refieren al descubrimiento de relaciones de asociación y dependencias funcionales entre los diferentes atributos. A partir de la predefinición de las clases categóricas, tomado un dato y o ubicándolo dentro del rango determinado.

2.1.3 ESTIMAR NUEVOS VALORES PARA LOS ATRIBUTOS SELECCIONADOS

El uso de la matemática y la estadística con las funciones de interpolación, extrapolación, regresión, y estimación crea la relación de variable dependiente e independiente, colocando a un atributo de entrada o un valor de predicción para obtener el valor estimado o de salida de acuerdo al error seleccionado en el modelo.

La inferencia con el uso de la inducción, que es el desconocimiento de la dependencia de las entrada y las salida o de la estructura del sistema, usando un número limitado de observaciones o de mediciones de las entrada y las salidas del sistema, para lo cual toda la información debe estar organizada y la muestra está definida por un par de entrada y salida. La deducción, como otra alternativa de la inferencia permite al modelo definido la aplicación de entradas para llegar a las salidas previstas de forma que se puedan confrontar las dependencias o las asociaciones de las variables. Para terminar el proceso de inferencia, la transducción, a partir del entrenamiento de los datos se puede encontrar relaciones o reglas que con llevan a las salidas previstas. (Kantardzic, 2011)

2.1.4 MODELAR DEPENDENCIAS SEGÚN LOS ATRIBUTOS ESPECIFICOS

Mediante el uso de la información disponible a través de los atributos y con el uso las funciones de inducción, estimación y predicción en las labores de exploración o clasificación de los datos se pueden determinar las dependencias para facilitar el entrenamiento del modelo o permitir un aprendizaje automático.

2.1.5 VISUALIZAR LOS RESULTADOS OBTENIDOS DE LA MINERÍA DE DATOS

Mostrar la información extracción de la minería de datos de una forma adecuada, permite su análisis e interpretación para medir el alcance del objetivo, la comprensibilidad de los patrones extraídos y la facilidad del modelo concebido.

2.1.6 DESCUBIR NUEVAS REGLAS COMO PRODUCTO DE LA EXTRACCIÓN DE LOS DATOS

La interpretación, el análisis y el entrenamiento de los datos, junto con las dependencias o las asociaciones de los atributos de entrada con los de salida y sus predicciones, acompañadas del porcentaje de error permitido se pueden obtener nuevas reglas entre los elementos de la muestra.

2.2 CLASIFICACION DE LOS ALGORITMOS

Los algoritmos se pueden clasificar en estadísticos, estimación o predicción de datos y de salidas en la representación del conocimiento.

2.2.1 ALGORITMOS ESTADISTICOS

Estos algoritmos se caracterizan por hacer uso de la estadística mediante sus funciones de estimación o proyección entre los cuales se encuentran:

Regresión lineal. A partir de la relación de una variable independiente y la otra dependiente forma una ecuación que representa a la serie de datos, ajustada con los ángulos de los coeficientes a y b.

Regresión logística. Es una variación del algoritmo de red neuronal, la curva de los datos se comprime mediante una transformación logística para minimizar el efecto de los valores extremos.

Redes bayesianas. 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 los nodos procesados son considerados en el ordenamiento fijo, este proceso no puede ser cíclico, por qué 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.

2.2.2 ALGORITMOS PARA LA ESTIMACIÓN O PREDICCIÓN DE DATOS

Para la minería de datos uno de los objetivos de la construcción del modelo, es hacer estimaciones o predicciones sobre una serie de acontecimientos reflejados en datos, que le permitan la evaluación de los comportamientos de unas variables a través de las tendencias de las preferencias de los clientes o de eventos en estudio. Los siguientes algoritmos contribuyen a este objetivo:

Redes neuronales. La arquitectura está definida por las capas de nivel de entrada, las neuronas de entrada definen todos los valores de los atributos de entrada y sus probabilidades.

La capa de nivel oculta, las neuronas ocultas reciben de las neuronas de entrada las entradas, a las cuales se les asigna un peso de probabilidad dependiendo de su importancia, siendo negativo para que se desactive en lugar de activarse, para proporcionar a las neuronas de salida las salidas.

La capa de nivel de salida, representa los valores del atributo de predicción. Cuando en una salida se describe un estado ausente o existente es debido a que el atributo de predicción es binario. Para las neuronas de salidas de valor verdadero, falso y una de existe o de ausente, la predicción es un atributo de tipo discreto. Si en las neuronas de salida existen dos valores uno continuo y el otro existente o ausente es porque el atributo de la predicción es continuo. La función tangente hiperbólica y la sigmoidea, son utilizadas para la activación de las neuronas ocultas y las de salida.

Inferencia difusa. Trabaja con variables lingüísticas o con datos imprecisos, 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 se llaman consecuente o consecuencia. (Diaz & Morillas, 2004)

Serire temporal. Con la obtención de un conjunto de datos según las condiciones momemtaneas de un evento en estudio se analiza el comportamiento de los datos para preveer situaciones similares en función del tiempo, teniendo el algoritmo que ajustarse el mismo mismo, mediante técnica de interpolación para completar los datos faltantes.

2.2.3 ALGORITMOS DE SALIDA EN LA REPRESENTACION DEL CONOCIMIENTO

Los algoritmos de clustering, árboles de decisión y de asociación hacen parte de los usados para representar el conocimiento, producto de la extracción de los datos de la minería de datos. También se pueden utilizar en la etapa de clasificación debido al uso de condicionales con los cuales se pueden formas diferentes reglas.

Clustering. Basado en técnicas iterativas para agrupar los casos de un conjunto de datos dentro de un clúster que contienen características similares las cuales son útiles para la exploración de datos, la identificación de anomalías en los datos, la creación de predicciones y para representar el conocimiento.
Calculado el grado de perfección con que los clúster representan las agrupaciones para crear clúster que representa mejor los datos, se establece una iteración hasta que ya no sea posible mejorar los resultados de la redefinición de los clústeres. (M.Parimala, Lopez, & Senthilkumar, 2011)

Árboles de decisión. 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 usado. (Microsoft, 2012)

De asociación. Compuesto por 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 o para mostrar la representación de las salidas del conocimiento, basándose en la manifestación de un elemento específico identificado como importante. Con la creación de un conjunto de elementos a los cuales se les da una puntuación para representar el soporte y la confianza obteniendo una clasificación que deriva reglas relevante de los conjunto de datos.

3. CAMPOS DE ACCIÓN DE LA MINERA DE DATOS

El gran volumen de datos que genera las diferentes disciplinas del saber y del conocimiento ha originado la necesidad de crear técnicas, metodologías o herramientas que extraigan datos de esas grandes bodegas de almacenamiento de manera que produzcan conocimiento para ser aplicado en cada disciplina. Es así como se encuentra entre tantas las siguientes aplicaciones.

En las telecomunicaciones, es utilizado para mejorar la velocidad de entrenamiento de los datos que son obtenidos por redes móviles, mediante la técnica del clasificador integrado, el cual usa la técnica de muestreo, que incluye el muestreo aleatorio simple, muestreo estratificado y el muestreo poblacional. (JianPing, 2012)

En la educación, como sistemas de personalización, clasificaciones de estudiantes y de los contenidos, construcción adaptativa de planes de enseñanza, descubrimiento de relaciones entre actividades, diagnóstico incremental de los estudiantes, debido a sus capacidades para el descubrimiento de patrones de navegación regulares e irregulares. (Romero, Ventura, & Hervás, 2005)

En la multimedia, recupera la información 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. (Fernández, Veronica, Verónica, Nora, & Patricia, 2011)

En el comercio y sector financiero, es usada para evaluar el comportamiento de los clientes en una empresa o sector financiero, determinando según unos criterios, si son buenos o malos clientes y sus preferencias de compras. (Mylonakis, 2010)

En la medicina, identifica las interacciones de fármaco a fármaco en procesos críticos de administración de fármaco y el desarrollo de fármacos con la implementación de reglas lógicas (Tari, Saadat, Liang, Cai, & Baral, 2010)

En la gestión gubernamental, es 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. (Arabí, 2013)

4. CONCLUSIONES

Cuando se ha logrado la consecución de la estructuración del problema objeto de estudio se debe seleccionar un algoritmo acorde para realizar la clasificación de los datos, que defina el modelamiento del problema, mostrando sus resultados o salidas, para darle un tratamiento analítico, interpretativo, objetivo y comprensivo, que genere un modelo fácil para el descubrimiento de nuevas reglas o siendo acertado en el tratamiento de los datos en estudio.

La aplicación de la minera de datos está marcada en 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 riesgos, pronósticos, estimaciones, probabilidades, tendencias, recomendaciones, búsqueda de secuencias y agrupaciones de clientes o de eventos. Algunos ejemlplos, la educación, la multimedia, el comercio, el sector financiero, la medicina, las telecomunicaciones, la gestión gubernamental.

5. REFERENCIAS

Arabí, U. (2013). Ethical data mining and social science data exploration and descriptcion: scope and limitations in social science research. En H. Rahman, & I. Ramos, Ethical data mining applications for socio-economic development (págs. 22-39). United States of america: Idea Group Inc (IGI).

Diaz, D. B., & Morillas, R. A. (2004). MIneria de datos y lógica difusa. Una aplicación al estudio de la rentabilidad económica d elas empresas agroalimentarias en Andalucía. Estadística Española Vol 46, Núm. 157, 409-430.

Díaz, J. L., & Pérez, G. R. (01 de 10 de 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. Obtenido de http://www.lenhs.ct.ufpb.br/html/downloads/serea/trabalhos/A15_15.pdf

Fernández, J., Veronica, G.-c., Verónica, L., Nora, R., & Patricia, R. (01 de 05 de 2011). Indexación y recuperación de información multimedia. XIII workshop de investigadores en ciencias de la computacion

(págs. 324-328). La plata Buenos aires: Universidad Michoacana de San Nicolás de Hidalgo. Obtenido de http://sedici.unlp.edu.ar/handle/10915/20049

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, 13-27.

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, 473 - 480.

Kantardzic, M. (2011). Data mining conceptos, models methods, and algorithms secend edition. New Jersey: John Wiley & Sons, Inc.

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, 59-66.

Microsoft. (01 de 02 de 2012). msdn.microsoft.com. Recuperado el 01 de 10 de 2013, de msdn.microsoft.com/es-es/library/ms174949.aspx

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, 9-20.

Riquelme, J. C., Ruiz, R., & Gilbert, K. (2006). Minería de datos: Conceptos y Tendencias. Revista Iberoamericana de Inteligencia No 29, 11-18.

Romero, M. C., Ventura, S. S., & Hervás, M. C. (2005). Actas del III Taller Nacional de Mineria de Datos y Aprendizaje, TAMIDA 2005. Estado actual de la aplicación d ela mineria de datos a los sistemas de enseñanza basada en WEB (págs. 49-56). Universidad de Cordob.

Tari, L., Saadat, A., Liang, S., Cai, J., & Baral, C. (2010). Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism. Bioinformatics, Vol 26 , 547 - 553.

Witten, L. H., Frank, E., & Hall, M. A. (2011). Data mining practical machine learning tools and techniques, Third edition. Estados Unidos: Morgan Kaufmann Publications.

AUTORIZACIÓN Y RENUNCIA

El autor autoriza LACCEI para publicar el documento en la conferencia. Ni LACCEI o los editores son responsables ni por contenido o por las implicaciones que son expresadas en este documento.