IMPLEMENTACIÓN DE MINERÍA DE DATOS BASADA EN REDES BAYESIANAS PARA LA TOMA DE DECISIONES EN LOS REGISTROS ACADÉMICOS

Alex Fernando Bonilla Gordillo[1], Miguel Angel Ojeda Schuldt[2], Fabricio Echeverría Briones[3]

[1] Ingeniero en Computación en Sistemas Tecnológicos 2006.
[2] Ingeniero en Computación en Sistemas de Información 2006.
[3] Director de Tópico, Ingeniero en Computación, Escuela Superior Politécnica del Litoral, 1998, Profesor de ESPOL desde el 2000, pechever@espol.edu.ec

RESUMEN

La minería de datos es una actividad de extracción cuyo objetivo es el descubrir hechos contenidos en las bases de datos. En la mayoría de los casos se refiere a un trabajo automatizado. Si hay alguna intervención humana a lo largo del proceso, este no es considerado como minería de datos.

Se persigue desarrollar un modelo o conjunto de reglas que permita estimar y proyectar la apertura de paralelos para las distintas materias, este trabajo tiene como objetivos:

- Determinar un método para encontrar las relaciones existentes en un conjunto de hechos.
- Determinar un método para resolver las proyecciones probabilísticas “que sí” de un almacén de datos.
- Sustentar el uso de la minería de datos para la toma de decisiones.

SUMMARY

Datamining is an extraction activity, its main goal is to discover facts contained in different databases. In most cases we assume it as automatic processes. If there is any human intervention during the whole process then it’s not considered as Datamining.

This project wants to develop a model or set of rules which allow to estimate and planify how many courses from the different subjects should be openned, this project should cover the following goals:

- Determine a method to find relations inside a set of facts.
- Determine a method to solve the statement “what if?” from a Datawarehouse.
- Support the Datamining use for decission taking.

INTRODUCCION

A lo largo de varios años, se desarrollaron un gran número de métodos de análisis de datos basados en la estadística. Sin embargo, en la medida en que se incrementaba la cantidad de información almacenada en las bases de datos, estos métodos empezaron a enfrentar problemas de eficiencia y escalabilidad y es aquí donde aparece el concepto de minería de datos.

Gran parte de esta información es histórica, es decir, representa transacciones o situaciones que se han producido. Además, ya que los datos pueden proceder de fuentes diversas y pertenecer a diferentes dominios, parece clara la inminente necesidad de analizar los mismos para la obtención de información útil para la organización.

En muchas situaciones, el método tradicional de convertir los datos en conocimiento consiste en un análisis e interpretación realizada de forma manual. El especialista analiza los datos y elabora un informe o hipótesis que refleja las tendencias o pautas de los mismos. Este conocimiento, validado convenientemente, puede ser usado por los superiores para tomar decisiones importantes y significativas para la organización.

Esta forma de actuar es lenta, cara y altamente subjetiva. De hecho, el análisis manual es impracticable en dominios donde el volumen de los datos crece exponencialmente: la enorme abundancia de datos desborda la capacidad humana de comprenderlos sin la ayuda de herramientas potentes. Consecuentemente, muchas decisiones importantes se realizan, no sobre la gran cantidad de datos disponibles, sino siguiendo la propia intuición del usuario al no disponer de las herramientas necesarias. Este es el principal cometido de la minería de datos: resolver problemas analizando los datos presentes en las bases de datos.

Hasta no hace mucho, el análisis de los datos de una base de datos se realizaba mediante consultas efectuadas con lenguajes generalistas de consulta, como el SQL, y se producía sobre la base de datos operacional, es decir, junto al procesamiento transaccional en línea (OLTP) de las aplicaciones de gestión. No obstante, esta manera de actuar sólo permitía generar información resumida de una manera previamente establecida, poco flexible y, sobre todo, poco escalable a grandes volúmenes de datos.

La tecnología de bases de datos ha respondido a este reto con una nueva arquitectura surgida recientemente: el almacén de datos (DataWarehouse). Se trata de un repositorio de fuentes heterogéneas de datos, integrados y organizados bajo un esquema unificado para facilitar su análisis y dar soporte a la toma de decisiones. Esta tecnología incluye operaciones de procesamiento analítico en línea (OLAP), es decir, técnicas de análisis como pueden ser el resumen, la consolidación o la agregación, así como la posibilidad de ver la información desde distintas perspectivas.

Sin embargo, a pesar de que las herramientas OLAP soportan cierto análisis descriptivo y de sumarización que permiten transformar los datos en otros datos agregados o cruzados de manera sofisticada, no generan reglas, patrones, pautas, es decir, conocimiento que pueda ser aplicado a otros datos. No obstante, en muchos contextos, como los negocios, la medicina o la ciencia, los datos por sí solos tienen un valor relativo. Lo que de verdad es interesante es el conocimiento que puede inferirse a partir de los datos y, más aún, la capacidad de poder usar este conocimiento.

Todos estos problemas y limitaciones de las aproximaciones clásicas han hecho surgir la necesidad de una nueva generación de herramientas y técnicas para soportar la extracción de conocimiento útil desde la información disponible, y que se engloban bajo la denominación de minería de datos. El resultado de la minería de datos son conjuntos de reglas, ecuaciones, árboles de decisión, redes neuronales, redes bayesianas.

Durante mucho tiempo, la Facultad ha hecho grandes esfuerzos para hacer coincidir la apertura de los paralelos versus la demanda real, en ciertos semestres existían paralelos donde 8, 9 o 10 estudiantes se habían registrado, en algunos casos el paralelo no tenía estudiantes y permanecía abierto esperando llenarse. En otro caso, se necesitaba abrir otro paralelo debido a la gran cantidad de estudiantes que deseaban inscribirse en el curso.

Todos estos son hechos controlables y sin mucha trascendencia. Sin embargo el peor caso que ha ocurrido es cuando el paralelo no cumple el mínimo de estudiantes inscritos y por ende se lo cierra; eso genera una desviación en el desarrollo académico del estudiante que se encontraba registrado en aquel paralelo, porque los retrasa en el flujo de su carrera o porque descoordinan el horario que en un comienzo habían planificado para su semestre.

De acuerdo a lo mencionado, el problema se encuentra en la planificación de apertura de paralelos, ya que no existe una herramienta que permita equilibrar la predicción sobre datos pasados contra la demanda real.

CONTENIDO

Análisis del Problema

Para el problema en la planificación de apertura de paralelos en un semestre académico, la toma de decisión sobre "¿Cuántos paralelos deben abrirse?" se verifica mediante datos del semestres anteriores, tales como:

- Alumnos que aprobaron la(s) materia(s) que es (son) requisito(s).
- Alumnos que cumplen con otros requisitos, como el nivel de la carrera u aprobación de cursos especiales.

Sin embargo, los resultados que aparecen después de este pequeño análisis de variables, no tienen un nivel de certeza alto. Esto se debe a la inconsistencia de los datos como campos en blanco o datos mal ingresados, etc.

También se usan datos que ocurrieron en el mismo semestre pero en años pasados, para comparar y ver cual es la tendencia que los estudiantes tienen al registrarse en una materia

Como alternativa pueden medirse otras variables y calcular a partir de ellas mediante un modelo adecuado la predicción deseada. No obstante, en muchas ocasiones, no se dispone de dicho modelo porque se conocen sólo las principales variables de entrada, pero no las relaciones existentes entre ellas. Para este tipo de problemas, la minería de datos ayuda a encontrar un modelo que represente una aproximación de las relaciones con un grado de probabilidad.

Selección de Modelo

Las redes bayesianas proveen una forma compacta de representar el conocimiento y métodos flexibles de razonamiento basados en las teorías probabilísticas (teorema de Bayes) capaces de predecir el valor de variables no observadas y explicar las observadas.

Entre las características que poseen las redes bayesianas, se puede destacar que permiten aprender sobre relaciones de dependencia y causalidad, permiten combinar conocimiento con datos, evitan el sobreajuste de los datos y pueden manejar bases de datos incompletas. Existen muchas modelos de redes bayesianas, para este trabajo se usará un modelo Naive Bayes (NB)

Naïve Bayes es el modelo más simple de clasificación con redes bayesianas, ya que asume independencia entre todos los atributos dada una clase. NB es, por tanto, un modelo de atributos independientes. En este caso, la estructura de la red es fija y sólo es necesario aprender los parámetros (distribución de probabilidades). El fundamento principal del clasificador NB es la suposición de que todos los atributos son independientes conocido el valor de la variable clase.

A pesar de que asumir esta suposición en el clasificador NB es sin duda bastante fuerte y poco realista en la mayoría de los casos, se trata de uno de los clasificadores más utilizados. Además sus resultados son altamente competitivos contra los resultados de otras técnicas

Diseño de la Solución

El método bayesiano persigue relacionar las variables influyentes sobre un evento. A través de esta característica, es posible construir varias redes bayesianas que representan las situaciones deseadas para la evaluación de la variable clase, el evento del registro en la materia.

Inicialmente, probamos plantear una pregunta sin condiciones, tan solo especificando la materia: ¿Cuál es la probabilidad de que alguien se registre en la materia X? . Gráficamente podemos verlo así:

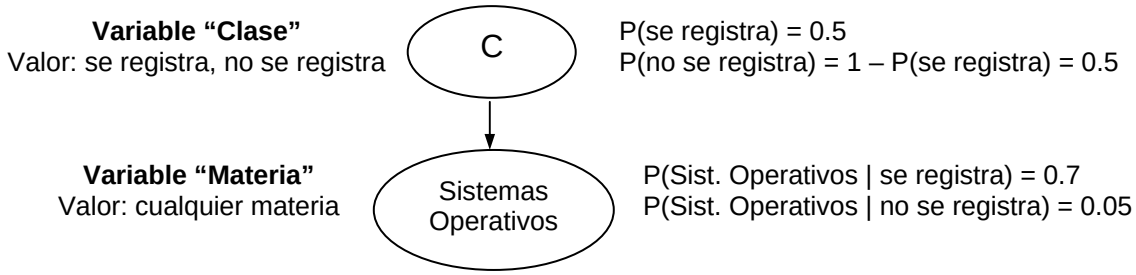

FIG 1. MODELO NB DE UNA VARIABLE

Ahora, podríamos realizar la misma pregunta agregando una condición: ¿Cuál es la probabilidad de que alguien se registre en la materia X, dado que este alguien pertenece a la carrera Y? Gráficamente podemos verlo así:

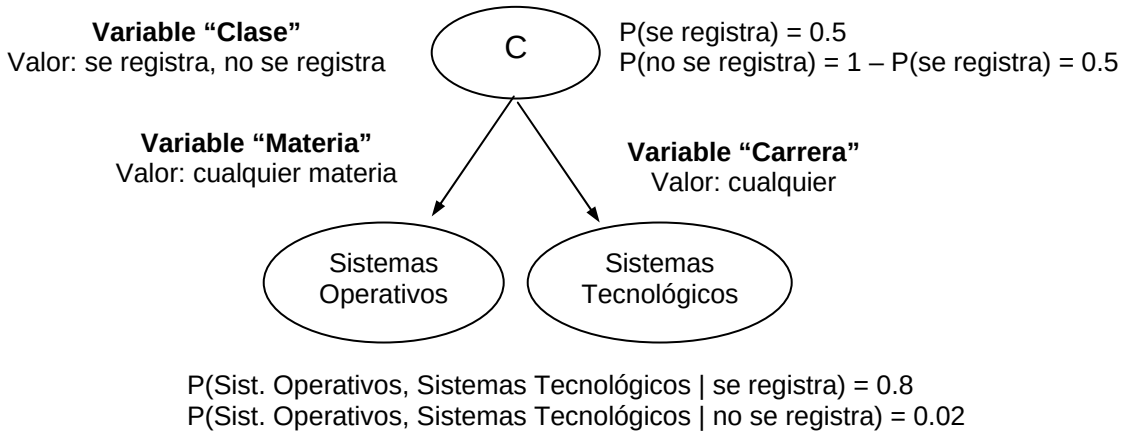

FIG 2. MODELO NB CON DE 2 VARIABLES

De esta forma, podemos agregar algunas otras condiciones. A medida que agregamos más condiciones nuestra pregunta se hace más específica. Es así que podemos "filtrar" las circunstancias bajo las cuales la variable "registro" muestra comportamientos diferentes.

Y, usando el Teorema de Bayes llegamos a obtener “¿Cuál es la probabilidad de que un estudiante se registre dado un conjunto de variables (de 1 a n variables)?” este resultado es directamente relacionado con la solución al problema de ¿Cuántos paralelos de una materia específica abrir?

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

FORMULA 1. TEOREMA DE BAYES

CONCLUSIÓN

Podemos decir que las redes bayesianas tienen un gran potencial en muchos campos y el académico no es la excepción, al proveer información importante y necesaria sobre la apertura de paralelos de una o mas materias.

Este método se basa en valores estimados, en nuestro caso, con valores de registros académicos pasados, de los cuales se analiza el comportamiento de los estudiantes frente a la decisión de registrarse o no en una materia.

Con la aplicación de un software producto del análisis y diseño de este proyecto de tesis, se puede dar soporte para la toma de decisión de ¿cuántos paralelos abrir? y así reducir uno de los problemas más comunes al momento del registro de un estudiante en un paralelo, como lo es la falta de cupo por estar lleno o el cierre del mismo por falta de estudiantes.

REFERENCIAS

1. A. Bonilla, M. Ojeda, “Implementación de Minería De Datos Basada En Redes Bayesianas Para La Toma De Decisiones En Los Registros Académicos” (Tesis, Facultad de Ingeniería Electrica y Computación, Escuela Superior Politécnica del Litoral, 2006)

2. J. Hernández, C. Ferri, Introducción a la Minería de Datos (Valencia, Prentice Hall, 19XX), 257p.

3. R. Jonson, Probabilidad y Estadística para Ingenieros (6ª. Edición, Wisconsin, Prentice Hall, 19XX) 76p.