PROYECTO MINERÍA DE DATOS

Nelson Cruz G , Trujillo García Raúl, Ricardo Vanegas G.
nacruzg@unal.edu.co, rtrujillog@unal.edu.co, rvanegasg@unal.edu.co
Universidad Nacional De Colombia

ABSTRACT

The project consists of the use of you hoist of decision to solve classification problems using the knowledge and approaches of the mining of data.

RESUMEN

El proyecto consta de la utilización de arboles de decisión para resolver problemas de clasificación usando los conocimientos y criterios de la minería de datos.

INTRODUCCIÓN

La gran cantidad de datos almacenados actualmente en las organizaciones, unido al gran desarrollo tecnológico de las computadoras, ha supuesto la aparición de nuevas posibilidades, agrupadas bajo el término generalmente conocido como "data mining". El aprovechamiento de estos datos requiere el desarrollo de proyectos con características específicas.

Los proyectos de Data Mining tienen por objetivo extraer información útil a partir de grandes cantidades de datos y se aplican a todos los sectores y en todos los campos. Así existen proyectos de este tipo en sectores tan dispares como el comercio electrónico, la banca, las empresas industriales o la exploración petrolífera. La extracción de esta información útil es un proceso complejo, que requiere la aplicación de una metodología estructurada para la utilización ordenada y eficiente de las técnicas y herramientas disponibles.

Usando la metodología CRISP-DM para la realización de nuestro proyecto tenemos:

Modelo De Referencia

El objetivo de nuestro proyecto es lograr la clasificación de datos en varias clases para lo cual necesitaremos un conjunto grande de datos los cuales serán la base de la creación de un modelo, y un conjunto de datos de prueba para analizar, mejorar y optimizar dicho modelo.

El proyecto debe lograr extraer la información suficiente del conjunto grande de datos, para poder hacer una buena clasificación del conjunto de prueba.

El proyecto se basa en crear un árbol de decisiciones con el conjunto grande de datos el cual permitirá analizar, mejorar, y confrontar el modelo con el conjunto de prueba, esperando obtener un porcentaje alto de aciertos en la clasificación de los mismos.

Todos los datos serán sacados de una base de datos confiables, de manera tal que los mismos no puedan interferir con los resultados reales de las pruebas.

Primera Fase

Diseñar un modelo de arboles de decisión para la clasificación de datos en diferentes categorías, dependiendo del área de búsqueda, el modelo se desarrolla teniendo como finalidad el reconocimiento y clasificación de datos en la era actual en que vivimos donde el tiempo se hace tan escaso y los datos son absurdamente grande, es indispensable obtener la mayor información trascedente en poco tiempo.

Segunda Fase

En la segunda fase se realizará la recolección inicial de datos, en orden a que sea posible establecer un primer contacto con el problema, identificando la calidad de los datos y estableciendo las relaciones más evidentes que permitan establecer las primeras hipótesis.

Tercera Fase

Se prepararán los datos, de tal forma que puedan ser tratados por las técnicas del modelado de arboles de decisión y los cuáles sean compatibles con java. La preparación de datos incluye las tareas generales de selección de datos a los que se va a aplicar la técnica de arboles de decisión (variables y muestras), limpieza de los datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato.

Los datos serán obtenidos teniendo en cuenta los siguientes parámetros:

1. Ser apropiados al problema.
2. Disponer de datos adecuados.
3. Cumplir los requerimientos del problema.
4. Tiempo necesario para obtener un modelo.
5. Conocimiento de la técnica.

Cuarta Fase

Establecer un diseño del método de evaluación de los modelos, que permita establecer el grado de bondad del modelo de árbol de decisión que permita comparar la realidad de los resultados modelados con los resultados físicos..

Quinta Fase

Se evalúa el modelo, no desde el punto de vista de los datos, sino del cumplimiento de los criterios de éxito del problema, teniendo el cuenta que el éxito del mismo es la clasificación de datos para obtener la mayor información de los mismos en un margen de alta calidad. Se debe revisar el proceso seguido, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso en el que, a la vista del desarrollo posterior del proceso, se hayan podido cometer errores. Si el modelo generado es válido en función de los criterios de éxito establecidos en la primera fase, se procede a la explotación del modelo

CONCLUSIONES

El proyecto esta diseñado para poner en practica todos los conocimientos que se han adquirido con aprendizaje de maquinas, inteligencia artificial y que se van adquirir en el transcurso del curso de minería de datos.

REFERENCIAS

[1] http://www.crisp-dm.org/ (18 Septiembre 2008)