Metodología para procesos de inteligencia de negocios con mejoras en la extracción y transformación de fuentes de datos

Autores:
Santiago Leonardo Morales (smorales@uce.edu.ec)
Mario Raúl Morales (mmoralesm@uce.edu.ec)
Dr. Ramon Rizo A. (ramon.rizo@gcloud.ua.es)

Institución: Universidad Central del Ecuador
Universidad de Alicante

Temática: Sistemas

Resumen

El hecho de no tener claro un método para ser eficiente en la búsqueda de decisiones en inteligencia de negocios (IN) generan altos costos en el uso de recursos, por lo cual, la ciencia de la computación llegó a solventar esta problemática mediante una rama de la misma que es la inteligencia artificial y en un sub campo como es el Aprendizaje Automático (A-A). En el presente trabajo, se define una metodología llamada M3S de IN, en la cual, se determina un procedimiento adecuado a seguir para tener los mejores resultados en la extracción de información, para obtener mayor conocimiento y valor en la información; a su vez, la metodología se sustenta en el algoritmo ID3 de árboles de decisión, el cual se fundamenta de conceptos matemáticos como son la entropía y ganancia de información dentro del análisis de los mejores atributos que se pueden considerar para toma de decisiones.

Palabras clave: En Memoria, Aprendizaje automático, árboles de decisión, Proceso ETC

Abstract

The fact of not having a clear method to be efficient in the search of decisions in business intelligence (IN) generate high costs in the use of resources, reason why, the computer science got to solve this problem through a branch of the same that is the artificial intelligence and in a subfield as is the Automatic Learning (AA). In the present work, a methodology called M3S of IN is defined, in which, an adequate procedure is determined to be followed to have the best results in the extraction of information, to obtain greater knowledge and value in the information; in turn, the methodology is based on the decision tree algorithm ID3, which is based on mathematical concepts such as entropy and information gain within the analysis of the best attributes that can be considered for decision making.

Keywords: In memory, Automatic learning, decision tree, ETL Process.

Introducción

Los datos de origen para las aplicaciones de inteligencia de negocios (IN) provienen de varias plataformas heterogéneas, que son gestionadas por una variedad de sistemas operacionales y aplicaciones. Partiendo de la premisa de que los sistemas deben aprender y adaptarse a su entorno, el propósito de los procesos de extracción,

transformación y carga de datos (ETC) en esta investigación, es unir los datos de estas plataformas y transformarlos a un formato que contenga todas las clasificaciones de información necesarias, de tal manera, que a partir de estos se tenga capacidad de decidir sobre todo el volumen de información actual y predecirla para llegar a un objetivo determinado.

Por esto, la investigación propone como objetivo general lo siguiente:

- Desarrollar una metodología para la ejecución de procesos de inteligencia de negocios (IN), que sirvan como estrategia para la implantación de soluciones tecnológicas con éxito en empresas tanto del sector público como privado, mejorándose principalmente los procesos de extracción y transformación de fuentes de datos heterogéneos con la ayuda de árboles de decisión.

Desarrollo

Estado del Arte

Dentro del estudio existen conceptos teóricos que fortalecen a esta propuesta como son:

Aprendizaje automático: desarrollar técnicas que permitan a las computadoras aprender. De forma más concreta, se trata de crear programas capaces de generalizar comportamientos a partir de una información suministrada en forma de ejemplos. Es, por lo tanto, un proceso de inducción del conocimiento.

El aprendizaje automático tiene una amplia gama de aplicaciones, incluyendo motores de búsqueda, diagnósticos médicos, detección de fraude en el uso de tarjetas de crédito, análisis del mercado de valores, clasificación de secuencias de ADN, reconocimiento del habla y del lenguaje escrito, juegos y robótica.

Los diferentes algoritmos de Aprendizaje Automático, se agrupan en función de la salida de los mismos, y uno es:

Aprendizaje supervisado: es el algoritmo que produce una función que establece una correspondencia entre las entradas y las salidas deseadas del sistema. Un ejemplo de este tipo de algoritmo es el problema de clasificación, donde el sistema de aprendizaje trata de etiquetar (clasificar) una serie de vectores utilizando una entre varias categorías (clases).

Árboles de decisiones: uno de los enfoques que tiene el aprendizaje automático son los árboles de decisiones como modelo predictivo. Se mapean observaciones sobre un objeto con conclusiones sobre el valor final de dicho objeto.

Una de las dos formas principales de árboles de decisión es la desarrollada por Quinlan al medir la entropía en cada rama y su vez que son más adecuados cuando:

- Las instancias del concepto son representadas por pares atributo-valor.
- La función objetivo tiene valores de salida discretos.
- Las descripciones del objeto para analizar son disyuntivas.

Modelo de clasificación ID3: significa "inducción mediante árboles de decisión". Es un sistema de aprendizaje supervisado que aplica la estrategia "divide y vencerás" para hacer la clasificación, implementando métodos y técnicas para la realización de procesos inteligentes, representando así el conocimiento y el aprendizaje, con el propósito de automatizar tareas. Podemos decir que su filosofía es:

- Determinar el árbol de decisión mínimo, con la menor entropía para todas las clases.
- El árbol debe estar organizado y entendible para cualquier persona, seleccionando en cada nodo el mejor atributo.
- Realiza una búsqueda de Arriba hacia Abajo, por medio de preguntas
- Está basado en criterio estadístico
- Selecciona los atributos en base a Ganancia de información

Entropía: se considera a la incertidumbre, impureza, desorden de un conjunto de datos relativo a una clasificación binaria.

$Entropia(S) = \sum_{i=1}^{n} -p_i \log_2 p_i$

Donde:

S: es una colección de objetos

P_i : es la probabilidad de los posibles valores

i: las posibles respuestas de los objetos

Si todos los datos pertenecen a una sola categoría la Entropía es cero $(0 * Log(0) = 0)$

Si los datos están mezclados (positivos = negativos = 0.5), la Entropía tendrá un valor máximo de 1[i].

Ganancia de Información: ocurre por el menor valor de entropía esperado (S), al ordenar todas las S basándose en cada atributo.

$$Ganancia(S, A) \equiv Entropia(S) - \sum_{v \in Valores(A)} \frac{|S_v|}{|S|} Entropia(S_v)$$

Dónde:

S : es una colección de objetos

A : son los atributos de los objetos

$V(A)$: Conjunto de valores que A puede tomar[ii]

Algoritmo ID3

La implementación de ID3, supone el siguiente código:

Atributo_salida: Atributo a predecir por el árbol.

Atributos: Lista de atributos a comprobar por el árbol.

1. Crear una raíz para el árbol.
2. Si todos los ejemplos son positivos Retorne (raíz, +)
3. Si t d l j l ti todos los ejemplos son negativos R t e urn (raíz, -)
4. Si Atributos=∅ Retorne (raíz, l)

(l es el máximo valor común de Atributo_salida en Ejemplos)

Si ninguna de las anteriores condiciones se cumple

Inicio

1. Seleccionar el atributo A con mayor Ganancia(Ejemplos,A)
2. El atributo de decisión para raiz es A
3. Para cada posible valor vi de A

3.1. Añadir una rama a raiz con el test A=vi

3.2. Ejemplos_vi es el subconjunto de Ejemplos con valor vi para A

3.3. Si Ejemplos_vi =∅ entonces añadir un nodo (n,l) a partir de la rama creada. (l es el

máximo valor común de Atributo_salida en Ejemplos).

Sino añadir a la rama creada el subárbol ID3 (Ejemplos vi, Atributo_salida, Atributos-{A}) **Fín**

Retornar (raíz)

Descripción

Dentro de la metodología en la fase de diseño y construcción ETC, se busca la opción de crear Arboles de Decisión, generados en base a experiencias concretas con los mejores atributos en lo que se refiere a ganancia de información que permitan testear nuevos datos y procesarlos para obtener resultados sobre hipótesis propuestas.

El criterio principal, es mejorar los resultados con la experiencia de información que contamos infiriendo suposiciones en hechos concretos (aprendizaje), lo cual permite realizar predicciones.

- Para esto utilizaremos el tipo de Aprendizaje Supervisado.
- Tenemos un conjunto de datos para los cuales buscamos resultados.
- Queremos una o varias reglas que establezcan todas las opciones posibles.

Resultados

En base a lo descrito anteriormente sobre el Diseño ETC, se consideran los siguientes pasos para la aplicación de ID3 dentro de la metodología M3S

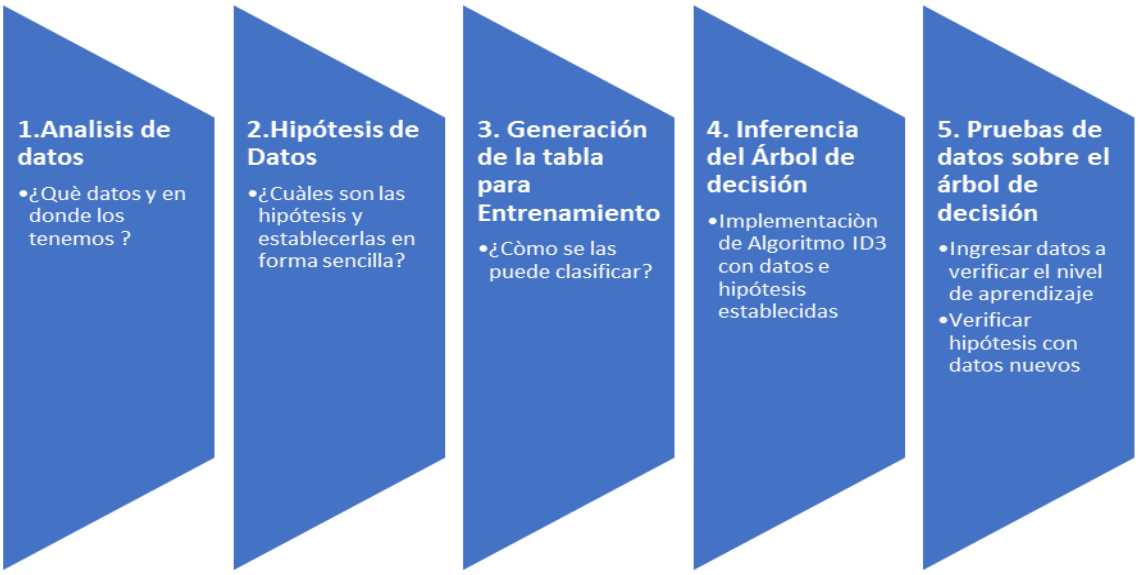

Ilustración 1. Fases de la metodología M3S, 2017

1. **Análisis de Calidad de Datos:** se procede con el análisis de los mejores datos que podamos tomar para un proceso de decisiones. Esto se obtien mediante uso de diccionarios de datos, estadísticas y conocimiento de expertos, para poder escoger los atributos más relevantes para el estudio.
2. **Hipótesis de Datos:** el objetivo que tiene esta parte metodológica es justamente encontrar una descripción del concepto objetivo mediante suposiciones generales a partir de hechos concretos de los datos en estudio y, a raíz de estos, clasificar la información para generalizar los casos nuevos.
3. **Generación de la tabla para Entrenamiento:** esta tabla es el producto final de los datos limpios y lo más refinados posible. A su vez, de la formalización de la o las reglas de negocio previamente establecidas con las cuales se han logrado clasificar la información. A raíz de esta tabla, se puede generar el proceso matemático y estadístico para la creación del árbol de decisión óptimo de acuerdo al algoritmo ID3 y a la objetividad que se haya dado en el análisis.
4. **Inferencia del Árbol de decisión:** para generar el árbol de decisión se utiliza la inferencia inductiva que es aprendizaje supervisado. Este árbol es la aplicación del algoritmo ID3, el cual se basa en las reglas del negocio formalmente elegido; del cual se prevé tener la mayor ganancia de información, y el mismo se encuentra detallado en la etapa de construcción. En forma simple para crear el indicado árbol se procede de forma recursiva de arriba abajo. Utilizar en cada nodo el atributo “más importante”, esto quiere decir, que es aquel que mejor discrimina los ejemplos que han llegado hasta este nodo y por último clasificando los casos que siguen en el siguiente nivel.
5. **Pruebas de datos sobre el árbol de decisión:** una vez que contamos con el árbol de decisión, el cual con los cálculos de entropía genera la mayor ganancia de información; se elabora un test para verificar que siempre se obtienen los mejores resultados con nuevos datos que ingresen a evaluarse en el árbol. Se debe tomar en cuenta que, para algunos casos al probar nuevos datos, estos podrían hacer que cambie la clasificación y reglas de la tabla de entrenamiento, lo cual produciría una alteración en el árbol de decisión generado. Es por esto que al usar M3S, se debe tener claro que Inteligencia de Negocios lo podemos definir matemáticamente como:
   IN = IN + Aprendizaje Automático

Formalizando esta metodología matemáticamente, será:

1. **Análisis de Calidad de Datos:**

   Sea $f : X -> Y$, en donde:

X son todos los casos posibles de entradas

Y espacio de salidas

2. **Hipótesis de datos (H)**

   Serán todas las $h \in H$

3. **Generación de la tabla para entrenamiento:**

Si $f \in H$ entonces es Aprendizaje considerado

Si $f \notin H$ entonces no será considerado

4. **Inferencia de Árbol de Decisión:**

Usar conocimiento a priori para asegurar que $H \subset f$ en donde la inferencia se lo realiza mediante:

- Cálculo de la entropía

Dados:

- Un problema con dos clases (positiva y negativa)
- S, el conjunto de ejemplos.

$$Entropia(S) = \sum_{i}^{k} p_i \log_2 \left(\frac{1}{p_i}\right) \; con \; k \; clases$$

- Y también mediante la ganancia que se obtendrá

$$Ganancia(S, A) = Entropia(S) - \sum_{v \, \epsilon \, valores \, de \, A} \frac{|S_v|}{|S|} Entropia(S_v)$$

5. Pruebas de datos sobre el árbol de decisión:

Utilizaremos medidas comunes de evaluación que son:

$$promedio\ positivo\ verdadero = \frac{TP}{TP + FN}$$

$$promedio\ positivo\ falso = \frac{FP}{FP + TN}$$

$$presiciòn = \frac{TP}{TP + FP}$$

$$correcciòn = \frac{TP + TN}{TP + FP + TN + FN}$$

Donde:

TP: Verdaderos positivos (clasificado: positivo, realidad: positivo)

TN: Verdaderos negativos (clasificado: negativo, realidad: negativo)

FP: falsos positivos (clasificado: positivo, realidad: negativo)

FN: falsos negativos (clasificado: negativo, realidad: positivo)

Conclusiones

- La eficiencia de los algoritmos de árboles de decisión dependen en gran medida del conocimiento que entreguen los expertos de negocio, para generar una óptima tabla de entrenamiento (TE).
- La implementación de ID3 sobre TE, entrega un árbol de decisión que mantiene la mayor ganancia de información posible; sin embargo, está sujeto a cambios en base a nueva información que puede ingresar al sistema.
- Como resultados iniciales, se creó una metodología dentro de ETC, en la cual, se implementó ID3 para extraer información con mayor calidad de cualquier almacenamiento.

Referencias Bibliográficas

- *Algoritmo de discretización de series tiempo basado en entropía y su aplicación en datos colposcópicos.* García López, D. A. (2007).
- *Aprendizaje de árboles de decisión.* Sempere, J. (2014). Valencia. Universidad Politécnica de Valencia.
- *Business modeling and data mining*. Pyle, D. (2003). Morgan Kaufmann Publishers.
- *Data mining with decision trees: theory and applications.* Lior Rokach and Oded Maimon (2008). World Scientific
- *La metodología de Kimball para el diseño de almacenes de datos (Data warehouses).* Rivadera, G. R. (2010) 56–71.
- *Machine Learning,* Mitchell, T. (1997). McGraw Hill.
- *Método ID3.* J.R. Quinlan (1973, 1986)
- *Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos*. *Recursos Digitales*. Rodallegas Ramos Erika, Torres González Areli ,. Gaona Couto Beatriz B, Gastelloú Hernández Erick, Lezama Morale Rafael As, V. O. S. (2010). Retrieved from http://ccita2010.utmetropolitana.edu.mx/recursos/Recursos_digitales.pdf#page=34
- Rafael Muñiz. (2015). Últimas aportaciones del marketing. agosto 10, 2016, de Marketing XXI Sitio web: http://www.marketing-xxi.com/ultimos-25-anos-de-marketing-en-espana-11.htm