



Metodología para Procesos de Inteligencia de Negocios con mejoras en la extracción y transformación de fuentes de Datos

**Santiago Leonardo Morales Cardoso ¹, Mario Raúl Morales Morales ²,
Ramón Rizo Adelgher ³**

1 Universidad Central del Ecuador, smorales@uce.edu.ec

2 Universidad Central del Ecuador, mmoralesm@uce.edu.ec

3 Universidad Central del Ecuador, ramon.rizo@ua.es

Resumen

El hecho de no tener claro un método para ser eficiente en la búsqueda de decisiones en inteligencia de negocios 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 va a definir 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 dicha metodología se sustentara 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: In memory, Aprendizaje automático, árboles de decisión



Methodology for Business Intelligence Processes with improvements in the extraction and transformation of data sources

Abstract.

The fact of not having a clear method to be efficient in the research of decisions in business intelligence generates high costs in the use of resources. Computer Science solve this problem through Artificial intelligence and in the Automatic Learning (AA). In the present paper, a methodology called M3S of IN will be defined, in which an appropriate procedure is determined to be followed to have the best results in the extraction of Information; to obtain greater knowledge and value of the information. This methodology will be based on the ID3 algorithm of decision trees which is based on mathematical concepts such as entropy and information gain within the analysis of the best attributes that can be considered for making a decision.

Keywords: In memory, Automatic learning, decision trees



1 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:

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

Y algunos de sus objetivos específicos son los siguientes:

- Desarrollar una metodología que permita establecer de manera óptima tanto en tiempo, como en recursos soluciones de inteligencia de negocios (IN) con un alto grado de confiabilidad en el conocimiento obtenido.
- Contribuir con una nueva metodología formal en inteligencia de negocios (IN), para la generación de conocimiento en las organizaciones, particularmente en lo concerniente a manejo de indicadores relevantes de un determinado negocio o actividad de la vida.



Metodología para Procesos de Inteligencia de Negocios con mejoras en la extracción y transformación de fuentes de Datos

Revista Publicando, 4 No 11. (2). 2017, 107-119. ISSN 1390-9304

- Usar aprendizaje automático con el objetivo ID3 de árboles de decisión dentro del proceso de extracción, transformación y carga de información . (ETC).

2 Preliminares

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

Aprendizaje automático: El aprendizaje automático es 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. ¹

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



Metodología para Procesos de Inteligencia de Negocios con mejoras en la extracción y transformación de fuentes de Datos

Revista Publicando, 4 No 11. (2). 2017, 107-119. ISSN 1390-9304

mapean observaciones sobre un objeto con conclusiones sobre el valor final de dicho objeto.ⁱⁱ

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.

$$H(S) = - \sum_{i=1}^L P_i \log_2 P_i$$

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 ($0 * L - 0 = 0$)

Si los datos están mezclados (positivos = negativos = 0.5),

la Entropía tendrá un valor máximo de 1ⁱⁱⁱ

Ganancia de Información:

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

$$Ganancia(S, A) \equiv Entropía(S) - \sum_{v \in \text{valores}(A)} \frac{|S_v|}{|S|} Entropía(S_v)$$

Dónde: S es una colección de objetos

A : son los atributos de los objetos

$V(A)$: Conjunto de valores que A puede tomar^{iv}



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(raiz, +)	
(3) Si todos los ejemplos son negativos Retorne(raiz, -)	
(4) Si Atributos= \emptyset Retorne (raiz,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 v_i de A	
(3.1) Añadir una rama a raiz con el test $A=v_i$	
(3.2) Ejemplos_ v_i es el subconjunto de Ejemplos con valor v_i para A	
(3.3) Si Ejemplos_ v_i = \emptyset 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_ v_i , Atributo_salida, Atributos-{A})	
FIN Retornar(raiz)	

3 Descripción

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



Metodología para Procesos de Inteligencia de Negocios con mejoras en la extracción y transformación de fuentes de Datos

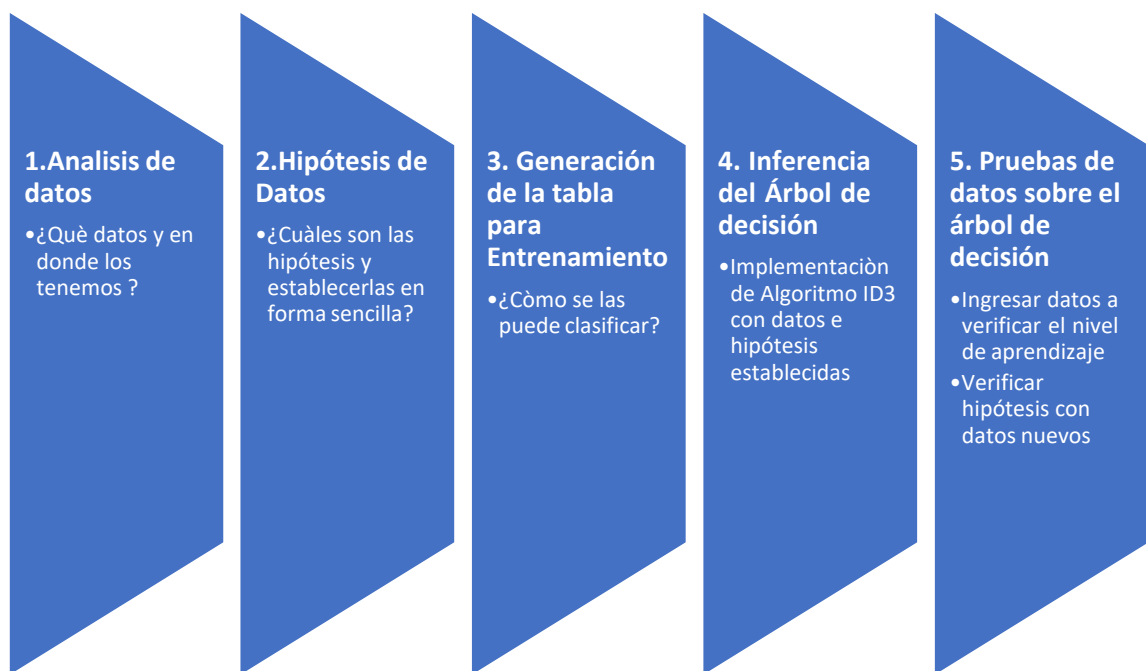
Revista Publicando, 4 No 11. (2). 2017, 107-119. ISSN 1390-9304

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

4 Resultados

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



- 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 obtendrá mediante uso de diccionarios de datos, estadísticas y conocimiento de expertos, para poder escoger los atributos más relevantes para el estudio.



Metodología para Procesos de Inteligencia de Negocios con mejoras en la extracción y transformación de fuentes de Datos

Revista Publicando, 4 No 11. (2). 2017, 107-119. ISSN 1390-9304

- 2. Hipótesis de Datos:** El objetivo que tendrá 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. Es producto también 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 podemos 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 utilizará inferencia inductiva que es aprendizaje supervisado. Este árbol es la aplicación del algoritmo ID3, el cual se basa en las reglas del negocio formalmente elegidas; 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 procederá 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 nos genera la mayor ganancia de información; elaboramos un test para verificar que siempre se obtendrán 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



Metodología para Procesos de Inteligencia de Negocios con mejoras en la extracción y transformación de fuentes de Datos

Revista Publicando, 4 No 11. (2). 2017, 107-119. ISSN 1390-9304

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:

$$IN = IN + \text{APRENDIZAJE AUTOMÁTICO}$$

Formalizando esta metodología matemáticamente será:

1. Análisis de Calidad de Datos:

Sea \mathcal{X} : $\mathcal{X} \rightarrow \mathcal{Y}$ en donde:

\mathcal{X} son todos los casos posibles de entradas

\mathcal{Y} espacio de salidas

2. Hipótesis de Datos (H)

Serán todas las $h \in \mathcal{H}$

3. Generación de la tabla para Entrenamiento:

Si $x \in \mathcal{X}$ entonces es Aprendizaje considerado

Si $x \notin \mathcal{X}$ entonces no será considerado

4. Inferencia de Árbol de decisión:

Usar conocimiento a priori para asegurar que $\mathcal{X} \subset \mathcal{Y}$ 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.

$$H(\mathcal{X}) = \sum_i p_i \log_2 \left(\frac{1}{p_i} \right)$$

- Y también mediante la ganancia que se obtendrá

$$G(A) = H(\mathcal{X}) - \sum_{v \in \mathcal{V}} \frac{|S_v|}{|S|} H(S_v)$$



5. Pruebas de datos sobre el árbol de decisión:

Utilizaremos medidas comunes de evaluación que son:

$$\begin{aligned} & \frac{\frac{\frac{TP}{TP+FN}}{\frac{TP}{TP+FN} + \frac{FP}{FP+FN}}}{\frac{TP}{TP+FN}} = \frac{\frac{TP}{TP+FN}}{\frac{TP}{TP+FN} + \frac{FP}{FP+FN}} \\ & \frac{\frac{TP}{TP+FN}}{\frac{TP}{TP+FN} + \frac{FP}{FP+FN}} = \frac{TP}{TP+FN} \\ & \frac{TP}{TP+FN} = \frac{TP}{TP+FN} \\ & \frac{TP}{TP+FN} = \frac{TP}{TP+FN} \end{aligned}$$



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)

5 Conclusiones

- La eficiencia de los algoritmos de árboles de decisión dependen en gran medida del conocimiento que nos entreguen los expertos de negocio, para generar una óptima tabla de entrenamiento (TE).
- La implementación de ID3 sobre TE, nos entrega un árbol de decisión que mantiene la mayor ganancia de información posible, que 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.
- Finalmente, basados en resultados iniciales, que deberán ser nuevamente validados, considero que los objetivos planteados pueden ser alcanzados en el tiempo previsto.

Referencias bibliográficas

Sempere, J. (2014). Aprendizaje de árboles de decisión. *Universidad Politécnica de*



J.R. Quinlan (1973, 1986): Método ID3

García López, D. A. (2007). Algoritmo de discretización de series tiempo basado en entropía y su aplicación en datos colposcópicos.

Mitchell, T. (1997). Machine Learning, McGraw Hill.

Pyle, D. (2003). *Business modeling and data mining*. Morgan Kaufmann Publishers.

Rafael, M. (2015). 11. Últimas aportaciones del marketing | marketing- xxi.com. Retrieved August 10, 2016, from <http://www.marketing-xxi.com/ultimos-25-anos-de-marketing-en-espana-11.htm>

Rivadera, G. R. (2010). La metodología de Kimball para el diseño de almacenes de datos (Data warehouses), 56–71.

Rodallegas Ramos Erika, Torres González Areli ,. Gaona Couto Beatriz B, Gastelloú Hernández Erick, Lezama Morale Rafael As, V. O. S. (2010). *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.*

Retrieved

from

http://ccita2010.utmetropolitana.edu.mx/recursos/Recursos_digitales.pdf#page=34

Lior Rokach and Oded Maimon (2008). Data mining with decision trees: theory and applications. World Scientific.
