Modelo Híbrido Predictivo y de Recomendaciones con Técnicas de Minería de Datos e Inteligencia Artificial

Hybrid Predictive Model and Recommendations with Techniques of Data Mining and Artificial Intelligence

René Cruz Guerrero 1*, Ma. De los Ángeles Alonso Lavernia2, Anilú Franco Árcega31, 2, 3, Universidad Autónoma del Estado de Hidalgo
Carretera Pachuca-Tulancingo, Km 4.5, Mineral de la Reforma, Hidalgo, México, CP. 42186.
{cr061289, marial, afranco}@uaeh.edu.mx

1 Instituto Tecnológico Superior del Oriente del Estado de Hidalgo
Carretera Apan-Tepeapulco Km 3.5, Colonia Las Peñitas, Apan Hidalgo, C.P. 43900.
Correo-e: 1*rencrug@gmail.com

PALABRAS CLAVE:

Modelos predictivos, Minería de Datos, Sistemas Basados en Conocimiento, Ontologías, Reglas de Asociación.

RESUMEN

El presente trabajo propone un modelo híbrido predictivo capaz de utilizar datos y conocimiento para brindar los resultados, enriqueciéndolo, en el caso que así lo requiera, con recomendaciones que faciliten la toma de decisiones. Se utilizaron técnicas de Inteligencia Artificial para representar en un esquema ontológico el conocimiento obtenido al aplicar reglas de asociación.

KEYWORDS:

Predictive Models, Data Mining, Knowledge Based Systems, Ontologies, Association Rules.

ABSTRACT

The present work proposes a hybrid predictive model capable of using data and knowledge to provide the results, enriching it, in the case that requires it, with recommendations that facilitate decision making. Artificial Intelligence techniques are used to represent in an ontological schema the knowledge obtained by applying rules of association.

Recibido: 30 de junio del 2017 Aceptado: 31 de agosto de 2017 Publicado: 15 de diciembre de 2017

1 INTRODUCCIÓN

Actualmente, cada vez son más las organizaciones que tratan de aprovechar su información disponible con el objetivo de identificar riesgos y oportunidades con anticipación, cuya solución requiere del uso de técnicas de predicción. Son un área de la Minería del Datos que pretende extraer conocimiento para predecir tendencias y patrones de comportamiento.

La construcción de modelos predictivos ayuda a prevenir situaciones, lo cual ayuda en la toma de decisiones proactivas y facilita la solución de problemas que tradicionalmente ocupan demasiado tiempo en resolverse mediante el análisis de patrones ocultos en grandes bases de datos.

Este tipo de modelos permiten estimar valores futuros o desconocidos de variables objetivo o dependientes usando otras variables independientes o predictivas. Para lograr esta capacidad requieren de un conjunto de pruebas y de interacciones de entrenamiento para la generación del modelo.

En el presente artículo se expone un modelo híbrido predictivo como una solución al problema de predicción, debido a que además de usar técnicas predictivas y descriptivas de Minería de Datos, se incorporan técnicas de Inteligencia Artificial sobre el conocimiento extraído de los datos. Esta solución permitirá enriquecer los resultados que proporciona un sistema predictivo convencional, debido a que además de brindar el resultado de la predicción, en los casos que sea posible podrá identificar comportamientos particulares en los datos, y así a través del uso de técnicas de recomendación, proporcionárselos al usuario.

ANTECEDENTES

En los últimos años se han propuesto diversos modelos predictivos, los cuales aplican diferentes técnicas de solución. En un primer tipo están los que hacen uso de técnicas Estadísticas utilizando métodos que solo se enfocan a modelos de regresión como el análisis multivariante [1] o regresión [2]. Otro tipo son los que emplean técnicas de Minería de Datos aplicando distintos métodos como Redes Neuronales [3], Árboles de Decisión [4], Máquinas de Soporte Vectorial [5], Redes Bayesianas [6], [7], entre otros. Finalmente, también se han realizado trabajos donde se ocupan técnicas de Inteligencia Artificial haciendo uso del conocimiento brindado por expertos [8] [9].

Respecto a trabajos donde se combinan técnicas para realizar predicción, uno de ellos es el propuesto por Hudaib, Dannoun & Harfoushi [10], el cual es considerado híbrido porque combina dos técnicas de Minería de Datos (agrupamiento y clasificación). Otro trabajo propuesto por Doreswamy & Hemanth [11] considerado también híbrido debido a que combina algoritmos de dos enfoques, uno de Minería de Datos (clasificador Naive Bayes) y otro de Máquinas de Aprendizaje (método de Correlación de Pearson).

PROBLEMÁTICA

Respecto a los trabajos de investigación desarrollados sobre los modelos predictivos, se han detectado los siguientes inconvenientes: La mayoría de los trabajos se enfocan a la solución de una situación específica, no se recurre a otras formas de obtener conocimiento y en la mayoría de los casos, el resultado tiene que ser interpretado por el usuario. Estos inconvenientes conllevan a que los usuarios que utilizan este tipo de modelos cuentan con pocos elementos para la toma de decisiones.

En respuesta a la problemática planteada, se propone una solución donde por una parte se hace uso tanto de técnicas predictivas como descriptivas de Minería de Datos y por otra, se utilizan técnicas de Inteligencia Artificial, con ello se tienen las siguientes ventajas:

- Se puede aplicar en cualquier área.
- La adquisición de la mayor información posible de los datos fuente, obteniendo conocimiento utilizando técnicas descriptivas de Minería de Datos.
- Se brinda un conjunto de recomendaciones además de la categoría de clase, lo cual facilitará la toma de decisiones.
- Selección de un método de clasificación considerando diferentes enfoques para obtener aquel que brinde mejores resultados según la naturaleza de los datos.

MODELO PROPUESTO

El modelo propuesto es considerado híbrido porque incluye tanto técnicas de Minería de Datos como de Inteligencia Artificial, incluso dentro de la Minería de Datos también es híbrido porque involucra tanto la clasificación como la generación de reglas de asociación. En la Figura 1, se muestran los procesos y componentes que incluye dicho modelo.

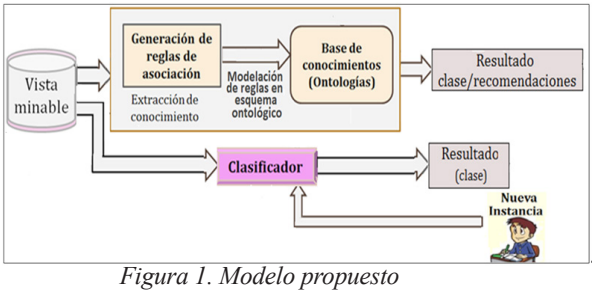

Figura 1. Modelo propuesto

Generación de reglas de asociación.- Se hace uso de esta técnica para extraer los patrones más frecuentes en conjuntos de datos, cuyo conocimiento sirve de base para brindarle al usuario en los casos que sea posible una serie de recomendaciones.

Debido a que el conocimiento obtenido se utiliza en la tarea de predicción, se genera un tipo particular de reglas denominadas Reglas de Asociación de Clase (RAC) que se caracterizan por contener en el sucedente un solo ítem correspondiente a la categoría de la clase. Para esto, se utilizó el método TNR debido a que después de realizar un análisis comparativo entre diferentes algoritmos, resultó ser el que brinda los mejores resultados respecto al problema de redundancia.

Base de conocimientos.- Contiene el conocimiento extraído con técnicas de Minería de Datos, es un componente en el cual se modelan los patrones identificados bajo un esquema ontológico, lo que permitirá de manera natural, realizar recomendaciones pertinentes al usuario. Los axiomas son los componentes que permiten aplicar razonamiento sobre los datos de una nueva instancia y extraer las recomendaciones en los casos que sea posible.

Clasificador.- Esta técnica se utiliza para clasificar una instancia a partir de sus características y con la clase obtenida ayudar al modelo a emitir una serie de recomendaciones en caso de que se requiera.

El proceso de modelación de las RAC en formato ontológico inicia con la preparación de la Base de Datos fuente y termina con la modelación de la Base de Conocimientos. En la Figura 2 se muestran las diversas tareas involucradas en este proceso.

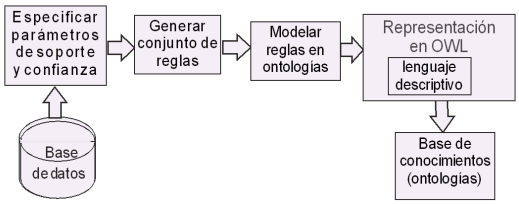

Figura 2. Submodelo para representar reglas en ontologías

Para obtener las RAC se utiliza la base de datos fuente previamente preprocesada especificando los valores de soporte y confianza. La cantidad de reglas depende directamente de estos parámetros, pues la frecuencia de aparición de un conjunto de reglas debe ser mayor o igual a estos valores.

Una vez generadas las RAC, el siguiente paso consiste en representarlas en esquema ontológico. Debido a que no existe un método que efectué esta tarea, se desarrolló un nuevo procedimiento para hacerlo. Este se implementó haciendo uso de los componentes de las reglas de asociación de clase (pares atributo-valor y nombres de clases) y de las ontologías (conceptos, propiedades de datos y axiomas), el procedimiento para realizar la modelación se compone de los siguientes pasos:

Los pasos para modelar el conocimiento de un conjunto de RAC a un esquema ontológico son los siguientes:

- Paso 1: Identificar en las RAC las categorías de clase y atributos
- Paso 2: Modelar conceptos y propiedades de datos
- Paso 3: Modelar axiomas
- Paso 4: Crear la ontología

Después de efectuar la modelación se lleva a cabo la representación en lenguaje OWL, debido a que es de los más robustos. Una vez que el conocimiento se ha representado en lenguaje formal fácil de interpretar por el ordenador, es posible ejecutar los axiomas con algún razonador.

EVALUACIÓN DEL MODELO

Para esta etapa se desarrolló en java un framework que contiene cada uno de los componentes del modelo propuesto. Para programar los métodos de minería de datos se utilizaron librerías de Weka [12] y SPMF [13], para la parte de ontologías se ocuparon librerías de Jena y el razonador Pellet.

Primer caso de estudio

Para este caso se utilizó una base de datos de alumnos de una institución educativa de nivel superior que contiene datos de tipo académico y socioeconómico de los alumnos con el objetivo de pronosticar con anticipación qué estudiantes se encuentran en riesgo de deserción.

En el Cuadro 1 se muestra el conjunto de reglas obtenidas a partir de la base de datos fuente después de aplicar el método TNR.

Pro_Prog >= 90 Est_Padr=Prep ==> class=Egresa #SUP: 30 #CONF: 1.0
Pro_Prog=<70 ==> class=No_Egresa #SUP: 36 #CONF: 0.972
Especial=Comp Pro_Mate >=90 ==> class=Egresa #SUP: 39 #CONF: 1.0
Pro_Mate >=90 ==> class=Egresa #SUP: 42 #CONF: 0.933
Pro_Mate=<70 ==> class=No_Egresa #SUP: 37 #CONF: 0.973
Pro_Mate>=90 Pro_Prog>=90 ==> class=Egresa #SUP: 33 #CONF: 1.0
Pro_Mate>=90 Est_Padr=Prep ==> class=Egresa #SUP: 33 #CONF: 0.916
Pro_Prog >=90 ==> class=Egresa #SUP: 42 #CONF: 0.933
Tip_Bach=CBTA Pro_Prog=<70 ==> class=No_Egresa #SUP: 30 #CONF: 1.0
Est_Padr=Sec Pro_Mate=<70 ==> class=No_Egresa #SUP: 25 #CONF: 1.0

Cuadro 1. Reglas generadas con la base de datos de estudiantes

Al conjunto de reglas generadas se les aplicó el método de conversión a esquema ontológico, donde a cada una de las clases encontradas en el conjunto de reglas le corresponde una clase de equivalencia de la ontología (egresa, no egresa) y donde cada antecedente de cada regla es representado con un axioma. En la Figura 3 se muestra parte de los resultados obtenidos, para lo cual se utilizó la herramienta Protegé.

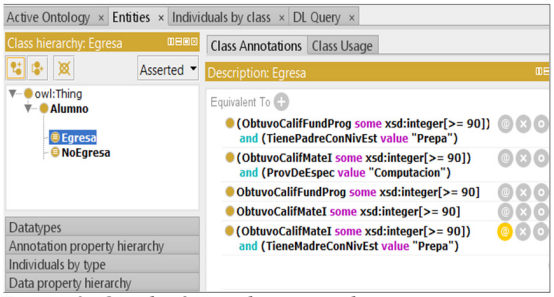

Figura 3. Ontología resultante en el primer caso

Una vez representada la ontología, es posible aplicar razonamiento sobre los valores de los datos de la nueva instancia para obtener tanto la clase a la que pertenece como la serie de recomendaciones encontradas.

En la Figura 4, se muestra un ejemplo de predicción donde se puede observar que el sistema además de proporcionar la clase, le muestra al usuario la serie de recomendaciones que logró encontrar de acuerdo a los valores introducidos en los atributos de la nueva instancia, dichas recomendaciones se obtienen a partir de la lista de axiomas existentes.

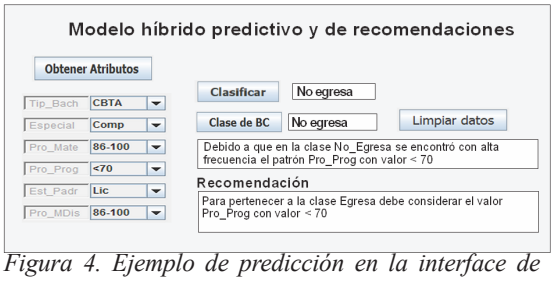

Figura 4. Ejemplo de predicción en la interface de aplicación

En este ejemplo se puede observar que el modelo propuesto, además de dar a conocer la clase a la que pertenece la nueva instancia, le proporciona al usuario las recomendaciones que se encontraron en los axiomas de la base de conocimientos de acuerdo a las características de dicha instancia.

Segundo caso de estudio

Para este caso se utilizó el conjunto de datos Pime Diabetes del repositorio UCI, teniendo como objetivo predecir si un paciente padecerá o no diabetes. Las RAC generadas de este conjunto de datos se muestran en el Cuadro 2.

HbA1c-pa=Mayor-6-4 Prec-Art=Mayor-122 ==> class=Positivo #SUP: 16 #CONF: 1.0
Ant-Fami=Ambos HbA1c-pa=Mayor-6-4 ==> class=Positivo #SUP: 20 #CONF: 1.0
Gluc-Bas=Mayor-110 HbA1c-pa=Mayor-6-4 ==> class=Positivo #SUP: 24 #CONF: 1.0
Gluc-Bas=Mayor-110 ==> class=Positivo #SUP: 28 #CONF: 1.0
Gluc-Bas=80-95 ==> class=Negativo #SUP: 20 #CONF: 0.95
Prec-Art=110-116 Per-cint=80-90 ==> class=Negativo #SUP: 16 #CONF: 0.94
Per-cint=80-90 Gluc-Bas=80-95 ==> class=Negativo #SUP: 16 #CONF: 0.94

Cuadro 2. Reglas generadas con la Base de datos Pime Diabetes

Después de modelar las RAC anteriores, se obtiene una ontología con los axiomas correspondientes a las dos categorías (positivo y negativo) que se manejan en el conjunto de datos. En la Figura 5, se muestran los axiomas obtenidos correspondientes a la categoría positivo.

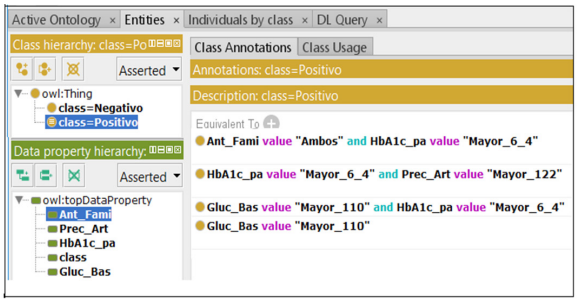

Figura 5. Ontología resultante en el segundo caso

De manera similar que en el caso anterior, la base de conocimientos integrada sirve para proporcionarle al usuario la categoría de clase y de ser posible las recomendaciones necesarias en relación a los valores de los datos de la nueva instancia.

CONCLUSIONES

La incorporación de una técnica descriptiva de Minería de Datos en un modelo predictivo, permitió obtener correlaciones interesantes entre valores de atributos, cuyo conocimiento es de suma importancia para el usuario.

Todas las componentes de la ontología modeladas a partir de las reglas de asociación de clases permiten abordar la predicción con técnicas de Inteligencia Artificial y en especial los axiomas, debido a que son los elementos que permiten inferir conocimiento que no está indicado explícitamente en la taxonomía de conceptos.

El modelo propuesto le ofrece al usuario más recursos para la toma de decisiones, en la etapa de validación se comprobó que el modelo es capaz de personalizar las sugerencias proporcionadas al usuario. Esto debido a que para brindarle las recomendaciones, del total de patrones frecuentes que se detectan en los datos, solo se utilizan aquellos que coinciden o se relacionan con las características de la nueva instancia a clasificar.

REFERENCIAS

1. Hanafy, O. A Multivariate Model for Predicting the Efficiency of Financial Performance for Property and Liability Egyptian Insurance Companies. Sadat Academy for Management Sciences. 2008, 1(1), 54-78.
2. Miguéis, V., Camanho, A., Falcao, J. Customer attrition in retailing: An application of Multivariate Adaptive Regression Splines. Expert Systems with Applications. 2013, 40(16), 6225-6232.
3. Hobson, R., Alkhasawneh, R. Modeling Student Retention in Science and and engineering disciplines using neural networks Global Engineering Education Conference IEEE, 2011, 86-97.
4. Aiswarya, I., Jeyalatha, S. Diagnosis of Diabetes using Classification Mining Techniques. International Journal of Data Mining & Knowledge Management Process. 2014, 2(1), 5-19.
5. Anbuselvan, S., Balamuralithara, B. Holistic Prediction of Student Attrition in Higher Learning Institutions in Malaysia Using Support Vector Machine Model. International Journal of Research Studies in Computer Science and Engineering. 2014, 29-35.
6. Lowd, D., Domingos, P. Naive Bayes Models for Probability Estimation. Department of Computer Science and Engineering University of Washington. 2008. 529-536.
7. Tahi, N. A Comparative Analysis of Techniques for Predicting Academic Performance. IEEE Xploree Frontiers in Education Conference. 2007, 1(1), T2G-7-T2G-12.
8. Espín, V., Hurtado, M., Noguera, M. Nutrition for Elder Care: a nutritional semantic recommender system for the elderly. Expert Systems. 2016, 201-210.
9. Gopalachari, V., Sammulal, P. Personalized Web Page Recommender System using integrated Usage and Content Knowledge. IEEE Advanced Communication Control and Computing Technologies, 2014, 1066-1071.
10. Hudaib, A., Dannoun, R., Harfoushi, O. Hybrid Data Mining Models for Predicting Customer Churn. Communications Network and System Sciences. 2015, 1, 91-96.
11. Doreswamy, L., Hemanth, D. Hybrid Data Mining Technique for Knowledge Discovery from Discovery from Engineering Materials Datasets. Computer Science. 2011, 1-12.
12. 12 V. Fournie, P., Soltani, A. SPMF: a Java Open-Source Pattern Mining Library. Journal of Machine Learning Research – JMLR. 2014, 1,1-5.

Acerca de los autores

René Cruz Guerrero es un estudiante de doctorado en el Centro de Investigación en Tecnologías de Información y Sistemas en la Universidad Autónoma del Estado de Hidalgo, México. Obtuvo su grado de maestría en Ciencias Computacionales en la Universidad Autónoma del Estado de Hidalgo. Sus intereses de investigación incluyen Minería de Datos y Sistemas Basados en Conocimiento.

María de los Ángeles Alonso Lavernia es profesora investigadora en el Centro de Investigación en Tecnologías de Información y Sistemas en la Universidad Autónoma del Estado de Hidalgo, México. Obtuvo su doctorado en Ciencias Computacionales en el Instituto Politécnico Nacional (IPN), México, DF. Sus intereses de investigación incluyen Reconocimiento de Patrones, Minería de Datos y Sistemas Basados en Conocimiento.

Anilú Franco Árcega es Doctora en Ciencias Computacionales por el Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Puebla, México en 2010. Actualmente es investigadora de tiempo completo en la Universidad Autónoma del Estado de Hidalgo, México, sus intereses de investigación incluyen Minería de Datos, Inteligencia Computacional, Reconocimiento de Patrones, Selección de Variables, Clasificación y Agrupamiento.