Modelo aumentado de árbol de decisión utilizando mapas autoorganizados

Augmented decision tree models using self-organizing maps

Wilson Castillo-Rojas¹ Fernando Medina-Quispe¹ Claudio Meneses-Villegas²

Recibido 25 de febrero de 2014, aceptado 30 de abril de 2014

Received: February 25, 2014 Accepted: April 30, 2014

RESUMEN

Un aspecto clave en el proceso de descubrimiento de conocimiento en bases de datos (KDD) es la comprensibilidad y credibilidad de los modelos generados por los esquemas inductivos de aprendizaje. Este artículo explora la aplicación de la técnica SOM sobre un modelo de árbol de decisión, para lograr una visualización aumentada del modelo. Representaciones de percepción visual del modelo, conjuntamente con datos y patrones, son establecidas basados en un esquema denominado VAM-MD, para soportar eficientemente su exploración y análisis visual durante la etapa de afinamiento del modelo de minería de datos. Con esto se busca responder preguntas genéricas respecto del funcionamiento interno del modelo, y con ello lograr mejorar su comprensión. Esta propuesta se implementó con un *software* prototipo, donde el analista de datos puede aplicar un conjunto de elementos visuales, sobre los datos en cada nodo del árbol, seleccionados apropiadamente para complementar la visualización del modelo generado. Además, el usuario dispone de diversos mecanismos de interacción que le permiten explorar cada componente del modelo. Finalmente, se analizan los resultados obtenidos desde un experimento controlado, llevado a cabo con dos grupos de usuarios, quienes utilizaron el *software* WEKA y el prototipo experimental para realizar una tarea de minería de datos sobre un conjunto de datos previamente preparados. El análisis preliminar de los resultados obtenidos permiten, por un lado, corroborar empíricamente la utilidad del uso de la técnica SOM para aumentar visualmente un árbol de decisión, y por otro, estimar subjetivamente su eficiencia en apoyar la comprensión del modelo generado.

Palabras claves: Minería de datos, minería de datos visual, visualización de modelos de minería de datos, visualización de árboles de decisión, exploración visual de modelos de minería de datos.

ABSTRACT

A key aspect in the process of Knowledge Discovery in Databases (KDD), is the understandability and credibility of models generated by inductive learning schemes. This article explores the application of Self-Organizing Maps (SOM) technique on a model of decision tree, to achieve enhanced visualization of the model. Representations of visual perception model, together with data and patterns are established, based on a schema called VAM-MD, to support exploration and visual analysis efficiently during tuning stage for Data Mining Model. This seeks to answer generic questions about the inner workings of the model and to achieve better understanding. This proposal was implemented through a software prototype, where a set of visual elements may be applied to data from each node in the tree, appropriately selected to complement the visualization of the generated model. Additionally, the user has several mechanisms of interaction that enable for exploration of each component of the model. Finally, the results from a controlled experiment, conducted on two user groups who used the WEKA software and the experimental prototype, for a data mining task on a previously prepared data were analyzed. Preliminary analysis of

¹ Facultad de Ingeniería y Arquitectura. Universidad Arturo Prat. Av. Arturo Prat 2120. Iquique, Chile.
E-mail: wilson.castillo@unap.cl; femedina@unap.cl

² Departamento de Ingeniería de Sistemas y Computación. Universidad Católica del Norte. Av. Angamos 0610. Antofagasta, Chile. E-mail: cmeneses@ucn.cl

the results obtained allow, on one hand empirically corroborate the utility of using the SOM technique to visually enhance a decision tree and on the other subjectively estimate their efficiency in supporting the understanding of the model generated.

Keywords: Data mining, visual data mining, visualization of data mining models, visualization of decision trees, visual exploration of data mining models.

INTRODUCCIÓN

El proceso de descubrimiento de conocimiento en bases de datos (KDD: *Knowledge Discovery in Databases*) es complejo, y muchos obstáculos, preguntas de investigación y problemas requieren ser investigados y clarificados. Un importante aspecto es la comprensibilidad de las entidades involucradas en el proceso de Minería de Datos (MD) en sí. Cuando los usuarios y analistas de datos desean interactuar con estas entidades para mejorar los resultados necesitan más que solo información de entrada/salida, lo que realmente requieren es entender cómo estas entidades trabajan internamente, sus componentes, el proceso llevado a cabo por ellas, y cómo se relacionan unas con otras.

En este contexto, el paradigma de visualización ha sido aplicado en una forma muy limitada en el proceso KDD, y principalmente enfocado hacia la visualización de datos (entrada del proceso) y resultados (salida del proceso). Por lo tanto, visualizaciones integradas en otras tareas y etapas del proceso de MD podrían mejorar este proceso y producir mejores resultados, especialmente cuando solo enfoques algorítmicos y analíticos han sido explorados, o donde la visualización ha sido aplicada débilmente.

La utilidad de un modelo de MD depende principalmente sobre dos factores: la capacidad del modelo para descubrir patrones interesantes, y la facilidad con la que la estructura del modelo puede ser entendida y modificada por usuarios y analistas de datos. Así, junto con la capacidad predictiva y descriptiva de un modelo de MD, su estructura debería ser bien entendida por usuarios y analistas, debido a que la clasificación o descripción de datos, sin una explicación sobre el modelo inducido a partir de los datos, disminuye la credibilidad de los resultados del proceso de MD.

En este sentido, apropiadas visualizaciones aplicadas sobre modelos de MD pueden transformar estos

en herramientas comprensibles que convierten datos en conocimiento. Así, las visualizaciones pueden transformar modelos inducidos en cajas transparentes tal que usuarios y analistas puedan fácilmente entender la transformación realizada por el modelo, esto es, cómo el modelo transforma datos en patrones. Además, adecuadas visualizaciones de patrones pueden facilitar la tarea de un analista de datos para descubrir conocimiento al interpretar y evaluar estos patrones visualmente.

Este artículo propone crear representaciones gráficas de modelos de MD, conjuntamente con datos y patrones, que soporten eficientemente la exploración visual en la etapa de afinamiento o ajuste del modelo MD generado, combinando técnicas de MD con visualizaciones complementarias, es decir, aplicar una segunda técnica de MD *ad hoc* al modelo generado a partir de una técnica primaria seleccionada.

Además, el usuario puede aplicar elementos visuales sobre los datos, así como utilizar mecanismos de interacción apropiados, que le permitan navegar sobre el modelo y explorar cada uno de sus componentes. Esto busca responder preguntas genéricas de los usuarios/analistas de datos respecto del funcionamiento interno del modelo.

Para lograr lo anterior, esta propuesta se basa en un esquema de visualización aumentada de modelos de minería de datos denominado VAM-MD [18], diseñado en la línea de investigación en la que se encuentra inmerso este estudio.

Este esquema señala un modelo de percepción visual e interacción del usuario centrado en la etapa de ajuste o afinamiento del modelo MD generado previamente, y establece la conveniencia de explorar el modelo original aplicándole visualizaciones con la combinación de una segunda técnica de MD *ad hoc* del tipo descriptiva, junto a un conjunto de artefactos gráficos, todos considerados “aumentadores

visuales”, y diversos mecanismos de interacción disponibles para el usuario/analista. Con esto se logra obtener un “modelo visualmente aumentado”, que permite entender su funcionamiento interno y así aportar, por un lado, a una mejor comprensión del modelo obtenido, y por otro, lograr una mayor credibilidad de este modelo.

Este trabajo contempla la implementación de una parte del esquema VAM-MD sobre un *software* prototipo que permite generar un modelo a partir de un Árbol de Decisión (ÁD) seleccionado como técnica primaria de MD, utilizando un conjunto de datos apropiados y previamente preparados. Luego el usuario debe seleccionar la técnica SOM (*Self-Organizing Map*) como técnica secundaria de MD o aumentador visual, y puede aplicarlo a los distintos componentes o nodos del ÁD generado, incluyendo su nodo raíz.

Adicionalmente, el usuario dispone de un conjunto de elementos visuales que puede aplicar a los datos de los nodos seleccionados, entre los cuales se tienen los siguientes artefactos gráficos tradicionales: diagrama de dispersión, histograma, y mosaico. Además, esta herramienta dispone de diferentes mecanismos de interacción que le permite al usuario navegar y explorar el modelo por todos sus componentes en una misma interfaz sin perder el contexto, logrando un modelo de visualización aumentada del ÁD.

Finalmente, se presenta una evaluación subjetiva del *software* prototipo mediante el desarrollo de un experimento controlado, que consiste en la encuesta a dos grupos (novatos y expertos) de usuarios usuarios/analistas de datos que utilizaron esta herramienta prototipo junto al *software* WEKA (*Waikato Environment for Knowledge Analysis*) comparándolos en una misma tarea de MD diseñada para este propósito, y proporcionaron información respecto de su desempeño, usabilidad, manejo de visualizaciones y apoyo en la comprensión del modelo MD, lo que permitió validar los resultados y contribuciones científicas de esta parte de la investigación.

VISUALIZACIÓN EN MINERÍA DE DATOS

A. Visualización en el proceso KDD

La visualización está siendo incorporada cada vez más en el proceso KDD como una herramienta para

soportar las interacciones entre usuarios/analistas de datos y los componentes involucrados en el desarrollo del proceso de MD. Sin embargo, ninguno de los modelos del proceso de MD existentes (Adriaans & Zantinge; Fayyad; Brachman; John) [1, 3, 5, 7] incorporan y discuten el rol de la visualización dentro del ciclo de MD. Meneses y Grinstein [9] proponen un esquema de proceso de MD con soporte para visualizar cuatro tipos de entidades: datos, espacio de parámetros de algoritmos de MD, modelos inducidos, y patrones.

La visualización de datos apoya el análisis exploratorio de los datos en la etapa previa al proceso de MD. Por ejemplo, visualización puede ser utilizada para obtener un entendimiento preliminar de los datos y refinar los objetivos y tareas definidas inicialmente por el usuario en la fase de formulación del problema.

Varias técnicas han sido propuestas para visualizar datos masivos. Algunas de estas técnicas están limitadas a tratar con conjuntos de datos de baja dimensionalidad (ej.; diagramas de dispersión), mientras que otras se enfocan en conjuntos de datos de alta dimensionalidad (ej., coordenadas paralelas, representaciones iconográficas, visualizaciones radiales, caras de Chernoff [4]).

Keim [8] provee una taxonomía de técnicas visuales para explorar conjuntos de datos masivos, mientras que Hoffman [6] esquematiza una categorización de técnicas visuales para realizar tareas de MD.

La visualización de modelos soporta el entendimiento e interacción con el modelo inducido desde un conjunto de datos de entrenamiento por un algoritmo de MD [10].

Estas visualizaciones deben proveer una forma natural de entender la estructura, componentes, y complejidad de un modelo, como también las representaciones visuales proveen una forma directa de comparar varios modelos, y permiten el uso de la percepción visual humana para formular hipótesis y conclusiones acerca del modelo y su correlación con los datos.

Visualizaciones de modelos han aparecido recientemente en la literatura y han sido incorporadas en algunas herramientas de *software* de análisis de

datos para MD. Becker [2] propone una visualización del clasificador simple de Bayes, para árboles de decisiones y tablas de decisión, y las implementa como parte de la herramienta de SGI-MINESET. Meneses & Grinstein [9] desarrollan visualizaciones para redes neuronales y modelos de mapas de Kohonen, y visualizaciones alternativas para el clasificador simple de Bayes y modelos de ÁD, y las implementan en una plataforma de *software* desarrollada para propósitos experimentales.

La visualización de patrones está referida a la visualización de resultados de aplicar un modelo de MD inducido a un conjunto de datos para validación y/o prueba [10-11]. En el caso de modelos predictivos, estos resultados están dados comúnmente como tasas de error, o una matriz de confusión.

En este caso, visualización puede ser usada para soportar la interpretación de estos patrones, y para proveer retroalimentación visual para correlacionar estos patrones con datos, parámetros, y modelos usados para generarlos.

En el caso de modelos descriptivos (ej., aquellos generados por algoritmos de agrupamiento), la representación visual de los patrones puede ser utilizada para soportar su interpretación y evaluación, como también como un medio de validación e inspección visual de los resultados.

B. Visualizando modelos de ÁD

Los ÁD corresponden a uno de los modelos predictivos más utilizados en MD, los cuales se generan creando particiones del conjunto de datos ordenados por algún valor que los jerarquiza, en subconjuntos, y después dividiendo cada uno de los subconjuntos, hasta que ya no se puedan dividir.

En sintonía con la metáfora del árbol, el conjunto de datos original es el nodo raíz, los subconjuntos son nodos, y los subconjuntos no particionados son las hojas, y estas hojas son definidas por la secuencia de las reglas de partición. Las ramas de un nodo son los subconjuntos creados a partir de este. Una regla de división utiliza típicamente una sola variable al asignar un caso a una rama [16].

Se puede comprender rápidamente una regla y juzgarla si es sensible, sin embargo, el juicio de la sensibilidad de una secuencia de reglas simples es

complicado, y un árbol grande con muchas particiones se torna difícil de comprender. El propósito de construir un ÁD es jerarquizar o repartir un grupo heterogéneo de características (atributos) de cosas o sujetos en grupos más pequeños y homogéneos.

La investigación sobre visualización de ÁD ha recibido gran atención actualmente, debido a su importancia en la práctica, y se han realizado muchos trabajos respecto de visualizadores 3D en MineSet, CATscan para inducir ÁD recubiertos [19], visualización interactiva en la construcción de ÁD [20], visualizador en un mapa de árboles en el sistema CART [21] y Salford.

Se pueden presentar dos problemas con la visualización de ÁD que no han sido investigados de manera profunda [25]: la integración de visualización de ÁD en el proceso KDD, y la visualización de ÁD de gran tamaño.

Existen nuevos enfoques en el campo de la visualización para representar largas estructuras con jerarquía: como los *treemap* [22], *cone-trees* [23], árbol hiperbólico [24], que aún no han sido considerados en el proceso KDD.

MODELOS DE PERCEPCIÓN VISUAL E INTERACCIÓN DE USUARIO EN MD

Presentamos un análisis comparativo de distintos modelos de percepción visual e interacción del usuario, en el contexto del proceso KDD propuesto por diferentes investigadores, y sobre el cual se toman como base elementos que permiten corroborar algunos aspectos relativos a la necesidad de aplicar visualización en todo este proceso.

En primer lugar se considera como referencia el Modelo de Exploración Visual de Datos de Keim [13], el que indica que el proceso de Análisis Visual (AV) se caracteriza por las interacciones entre datos, visualizaciones, modelos acerca de los datos, y los usuarios, con el objetivo de obtener conocimiento.

En este modelo se establecen en general los diferentes estados que se tiene en el proceso KDD, desde la fase de preparación de los datos, pasando por su exploración, modelamiento, visualización para apoyar la comprensión y validación del modelo, hasta la fase de obtención del conocimiento. También

se muestran las diferentes relaciones y acciones realizadas entre los estados, todo esto dentro del marco de un proceso iterativo.

El segundo modelo analizado corresponde a un Modelo Conceptual de Visualización Interactiva de Modelos de Minería de Datos (IVDMM), propuesto por Yan Liu & Gavriel Salvendy [14]. Este modelo es más específico que el anterior, y en él se detallan las acciones realizadas tanto por la máquina como el usuario, y en ambos casos asociados a una lógica de interacción relacionada con la construcción y evaluación del modelo de MD, es decir, no toma las fases previas (preparación y transformación de datos) y posteriores (obtención del conocimiento).

Se centra por tanto en las distintas etapas que se requieren para lograr un modelo apropiado para la tarea de MD. Propone la inclusión de visualización para datos, modelos, y evaluación individual o múltiple de modelos. También establece los mecanismos de interacción para que el usuario pueda lograr una mejor aproximación al modelo apropiado, permitiendo que participe activamente en el proceso de construcción del modelo de MD.

Un tercer modelo analizado propuesto por Vitiello & Kalawsky [12], que integra el AV basado en un flujo de trabajo para la elaboración de sentidos (*sense-making*), evolucionada a partir de toma de decisiones naturalistas desde la investigación de Klein [15].

Los aspectos claves en este esquema relacionan el reconocimiento de patrón, también referidos como marcos o cuadros. Estos pueden verse como un mapa mental de la situación informando la decisión. Los cuadros pueden también verse como análogos a la hipótesis de “fluidos” en que estos pueden desarrollarse, cuanto más información lleguen a tener disponible.

En este contexto, el objetivo del proceso *sense-making* es alcanzar la congruencia entre los datos y el marco, enfocándose más en la percepción visual, y para ello propone un marco de trabajo para la elaboración de sentido para sistemas de pensamientos en ingeniería.

En resumen, estos modelos de percepción visual e interacción analizados señalan como base común:

- El rol relevante de la visualización en todo el proceso KDD, confirmando lo necesario que es aplicar visualizaciones desde la exploración de los datos, en la etapa de MD, como apoyo en la construcción y visualización de modelos, hasta la obtención de conocimiento aplicando visualización a los patrones resultantes, todo esto dentro de un proceso iterativo.
- La combinación de técnicas MD se recomienda por parte de algunos autores como herramienta útil en el análisis del modelo en su fase de construcción y ajuste, para conocer mejor su funcionamiento y desempeño.
- Proveen estos modelos un marco referencial de trabajo para la aplicación de visualizaciones en la etapa de construcción del modelo de MD, no obstante, no establecen una especificación detallada respecto de los modelos, y no dan énfasis en la utilización de mecanismos de interacción para el usuario.

EL ENFOQUE VAM-MD

El enfoque VAM-MD propuesto por Castillo [18] considera las características de los modelos de percepción analizados y descritos en el punto anterior, y recoge los aspectos más relevantes de cada uno. Aporta el concepto de “Visualización Aumentada” para modelos de MD, y lo que propone es que dada una técnica de MD a visualizar, denominada Técnica Primaria de Minería de Datos (TP-MD), en este esquema, permite al usuario incorporar a esta visualización diferentes elementos visuales *ad hoc* al modelo y al dominio de los datos, y a su vez aplicar otra técnica de MD, denominada Técnica Secundaria de Minería de Datos (TS-MD), como aumentador visual que permita explorar la TP-MD.

La técnica TS-MD a seleccionar debe cumplir tres requisitos: ser una técnica descriptiva, luego apropiada al dominio de los datos sobre el cual está trabajando la TP-MD, y tercero, aportar información complementaria a la entregada por el modelo generado por la TP-MD.

Adicionalmente, dispone el uso de un conjunto apropiado de mecanismos de interacción para el usuario. Todo este conjunto de visualizaciones aplicadas apuntan a lograr aumentar el AV del modelo en su etapa de refinamiento o ajuste. Esta

combinación de técnicas de MD es señalada por diversos autores [17] como mecanismos para poder lograr un mejor entendimiento no solamente de los datos, sino también sobre los modelos generados.

En la Figura 1 se puede observar la arquitectura del esquema VAM-MD, donde se presentan además sus componentes de acciones e interacciones, partiendo de un conjunto de técnicas MD, que según sea el caso serán TP o TS. Luego se establece un conjunto de aumentadores visuales, entre los cuales se disponen artefactos gráficos tradicionales, como también técnicas MD seleccionadas para actuar como explorador visual de otra técnica.

Se agrega al esquema VAM-MD las diferentes acciones que el usuario puede realizar en este proceso, y que representan los diferentes mecanismos de interacción que dispone el usuario.

Figura 1. Esquema de visualización aumentada de modelos de minería de datos, VAM-MD [18].

En el centro se establece la componente principal que permite visualizar el modelo en etapa de refinamiento o ajuste, partiendo por la selección de una TP-MD del conjunto disponible, para aplicarle los diferentes aumentadores visuales seleccionados, ya sean artefactos gráficos (elementos visuales) como TS-MD, e incorporando los mecanismos de interacción requeridos para explorar y navegar el modelo.

Finalmente, con este esquema se logra obtener un modelo aumentado visualmente con las vistas o visualizaciones complementarias aplicadas al modelo original, permitiendo conocer mejor su funcionamiento interno, y de este modo apoyar la comprensión del modelo.

AUMENTANDO MODELOS DE ÁRBOLES DE DECISIÓN CON SOM

Uno de los principales problemas de las técnicas de MD es su representación visual, y la comprensión para el analista de datos del funcionamiento interno del modelo. Para el caso de ÁD esto es mucho más complejo cuando se trata de árboles de gran tamaño, el conjunto de datos a analizar tiene una alta dimensionalidad, y debido a su característica de estructura jerárquica [16].

De la revisión comparativa de esquemas de visualización para técnicas MD, que realiza Castillo & Meneses [17], entre las cuales se incluyen los ÁD, se concluye que:

- La mayoría de las investigaciones recomiendan utilizar una combinación apropiada de esquemas de visualización con diversas técnicas de MD, dependiendo de la tarea a realizar y las características de los datos.
- Es fundamental considerar la incorporación de mecanismos de interacción para el usuario en el diseño de nuevas visualizaciones.
- El rol de la visualización en el proceso KDD debe ser entendido y extendido en todas sus etapas, de modo que permita explorar los datos, modelos y patrones obtenidos.

Particularmente para la técnica ÁD, la mayoría de las representaciones visuales analizadas en [17] proponen a un ÁD en su forma normal jerárquica de dendograma en una visualización estática, sin la posibilidad de que el usuario/analista pueda interactuar con cada uno de los nodos del árbol.

La mayoría de las herramientas visuales de MD revisadas en [17], aunque entregan información general del árbol, junto a la matriz de confusión asociada, no permiten combinar técnicas de MD que aporten más allá de la información de las reglas del modelo y las instancias en cada nodo. Por ejemplo información de dispersión de los datos y su distribución espacial en cada nodo, mucho menos proveen al usuario mecanismos de interacción para que pueda navegar, seleccionar y explorar cada componente o nodo del árbol.

El uso de la técnica SOM aplicado a un ÁD como aumentador visual se debe a que SOM cumple un

doble propósito: partición espacial del subconjunto de datos asociados a cada nodo del árbol, y visualización de esta partición mediante un mapa. Se considera que el árbol por sí solo permite establecer reglas de decisión distribuyendo los datos por medio de jerarquías representadas por sus nodos, y en cada nodo recopila las instancias que cumplen con estas reglas, sin embargo, no permite visualizar la distribución espacial de las instancias, lo que sí provee SOM. Además, la técnica SOM es compatible al dominio de los datos que maneja el ÁD, y permite describir su distribución en cada nodo.

Lo anterior permite una comparación entre nodos y así poder determinar aquellos con similar distribución o gran diferenciación, mediante la especificación de la “distancia” entre ellos, el número de instancias ubicadas sobre o bajo el centroide de una cuadrícula del mapa. Por lo tanto, la selección de la TS-MD debe ser apropiada al dominio y tipo de atributos, y complementarla visualmente a lo que ofrece la TP-MD, es decir, que aporta elementos adicionales de análisis de los datos y del modelo, para obtener una visualización aumentada, y con ello apoyar su comprensión y lograr una mayor credibilidad del modelo.

Con todos estos elementos se puede implementar la metáfora de visualización aumentada del modelo de un ÁD en el proceso de MD, y permitir una mejor comprensión del funcionamiento del modelo generado.

Prototipo experimental

El prototipo del ambiente visual desarrollado tiene como alcance para este análisis experimental la implementación del esquema VAM-MD, para la TP-MD de modelo jerárquico, particularmente ÁD en combinación con la técnica SOM como TS-MD o aumentador visual para su exploración y análisis. Además, incorpora un set de elementos visuales o artefactos gráficos (tabla de datos, histograma, diagrama de dispersión y mosaico), y diferentes mecanismos de interacción (acercamiento aplicado a elementos visuales y al aumentador visual, manejo de transparencia para mantener el contexto del árbol, selección de nodos en cada nivel, y explorador del árbol).

En la Figura 2 se puede observar la vista general de la interfaz principal del *software* prototipo, donde

se visualiza un ÁD en la parte central, junto a vistas complementarias y elementos visuales del lado derecho. En esta interfaz se pueden observar todas las componentes que la herramienta visual contiene:

Figura 2. Vista general de interfaz principal del prototipo experimental.

- Selección, configuración de parámetros de la TP-MD a visualizar, y generación del modelo de MD: esta sección cuenta con dos pestañas, la primera “Datos”: permite al usuario seleccionar el conjunto de datos a ser analizados, y muestra características de estos datos y sus atributos, como por ejemplo, número de instancias, tipo y cantidad de atributos. En la segunda pestaña “T. Primaria MD”: el usuario tiene la posibilidad de seleccionar y configurar la TP-MD. Posteriormente se ejecuta el algoritmo de la TP-MD generando el modelo de MD.
- Área de visualización de la TP-MD: es el área de trabajo principal donde se presenta la vista general de la TP-MD seleccionada, y donde el usuario puede interactuar (recorrer, explorar y seleccionar) con cada elemento o nodo del árbol. Cuenta con dos pestañas, la primera llamada “Datos”, donde se muestran los atributos y datos originales del conjunto de datos a ser analizados. La segunda pestaña “Técnica Primaria MD” muestra la técnica seleccionada que para este caso es un ÁD.
- Técnica secundaria MD como aumentador visual: en esta área se presenta la vista minimizada de la TS-MD, que el usuario selecciona y aplica a la TP-MD. También el usuario puede maximizar esta vista, permitiendo modificar sus parámetros de configuración. Para este trabajo se considera la técnica SOM como aumentador visual o TS-MD.

- d) Elemento visual: en esta zona se despliegan los elementos visuales seleccionados desde la sección g), y en esta imagen se puede apreciar un diagrama de dispersión del nodo seleccionado.
- e) Técnica de navegación: se dispone de un navegador o explorador de la TP-MD, y permite al usuario tener la orientación de la zona o nivel donde se encuentra recorriendo el árbol.
- f) Barra de aumentadores visuales: se ubica en la parte inferior del área de trabajo de la herramienta y presenta un set de aumentadores visuales o TS-MD que permiten al usuario seleccionar uno apropiado a la TP-MD para aplicarlo, y con esto poder explorar cada componente del modelo generado. Para este trabajo se muestra la selección y arrastre del ícono de SOM sobre el ÁD.
- g) Selección de elemento visual: en esta sección el usuario dispone de un conjunto apropiado de artefactos gráficos que puede seleccionar, configurar y aplicar a la TP-MD, y su gráfica se presenta en el cuadro d). Dentro del conjunto de elementos visuales dispone de histograma, mosaico, y diagrama de dispersión.
- h) Información general de la TP-MD: muestra el número de nodos y hojas del árbol.

Como se señala en la sección c) de la interfaz principal, el usuario puede maximizar la imagen de la TS-MD como se puede observar en la Figura 3, y al hacerlo se abre una ventana donde se presenta una vista detallada de esta técnica, y el usuario puede reconfigurar los parámetros iniciales de esa técnica, que para este caso corresponde a la técnica SOM.

Figura 3. Vista detallada de la TS-MD aplicada a la TP-MD.

Una característica de la visualización de la técnica SOM es que maneja un nivel de transparencia que el usuario puede manipular, de modo de mantener el contexto de la visualización del ÁD.

Sobre cada nodo seleccionado, incluyendo el nodo raíz, el usuario puede aplicar SOM para visualizar la distribución de los ítems en los nodos del árbol, en la que además se proporciona al usuario elementos de interacción para cambiar el tipo de color del fondo del mapa y el tipo de clase, el tipo de gráfico de cada ítem del mapa, elegir el tipo de clase a colorear en la representación, seleccionar el conjunto de prueba o entrenamiento, etc. La técnica SOM provee una visualización complementaria al usuario, ya que cumple un doble propósito: partición espacial del subconjunto de datos asociados a un nodo del árbol, y visualización de esta partición mediante un mapa.

VALIDACIÓN EMPÍRICA Y ANÁLISIS DE RESULTADOS

A. Experimento controlado

El experimento controlado contempla la realización de una tarea de MD, definida para este propósito para generar un modelo de ÁD con un conjunto de datos previamente preparados. Se trata de generar y analizar este modelo de ÁD tanto en el *software* prototipo como en el *software* WEKA.

Luego los usuarios deben evaluar subjetivamente su percepción visual en ambas herramientas, utilizando todos los elementos de análisis que cada *software* les provee, para finalmente entregar su opinión por medio de una encuesta en línea [26], respecto del apoyo que el *software* prototipo le entrega en la comprensión del modelo generado.

Este experimento se realiza con un universo de 13 personas con distintos niveles de conocimientos acerca de procesos de MD, el uso de herramientas de MD, y en particular del *software* WEKA. A *posteriori* se clasificó a los participantes en grupos de “expertos” (7) y “novatos” (6), de acuerdo con la información captada en la encuesta [26].

Con ambas herramientas los usuarios participantes en este experimento deben generar un modelo de ÁD, visualizarlo, interactuar con el modelo, e interpretar los patrones o reglas obtenidas, utilizando

las opciones que los dos *softwares* ofrecen. Se busca que los participantes realicen una tarea genérica de clasificación y puedan responder preguntas respecto del modelo, sus componentes, y relacionar el modelo con las características de los datos desde el cual el modelo fue generado por el algoritmo de ÁD utilizado. Las preguntas de esta tarea son:

- a) Búsqueda de patrón en los datos mediante el modelo.
- b) Pregunta relacionada a la exploración del modelo y sus componentes.
- c) Pregunta asociada a detectar una relación dato-modelo.

La finalidad de este experimento es:

- Comprobar la utilidad que provee la combinación y aplicación de la técnica SOM sobre las componentes de un ÁD, para aumentar visualmente el modelo generado.
- Obtener una caracterización del nivel de efectividad del *software* prototipo respecto del aporte que ofrece al aplicar visualizaciones complementarias a un modelo de MD previamente generado, comparándolo con el *software* WEKA.
- Verificar si los usuarios logran comprender mejor el modelo del ÁD, aplicándole a cada uno de sus nodos vistas complementarias con la técnica SOM, junto a elementos visuales o artefactos gráficos provistos en el prototipo *software*.

Posteriormente, una vez realizada la tarea de MD, los usuarios/analistas de datos deben proporcionar sus apreciaciones subjetivas, por una encuesta diseñada con este propósito, respecto del desempeño de ambas herramientas acerca de manejo de visualización del modelo del ÁD generado, usabilidad, utilidad de los elementos visuales proporcionados, conveniencia de combinar y aplicar la técnica SOM a un ÁD logrando un modelo visualmente aumentado, y la eficiencia en la comprensión del modelo generado.

La encuesta diseñada contiene 28 preguntas y está dividida en tres secciones: a) Aspectos generales en el análisis de datos, b) Uso de herramientas de MD, y c) Uso de la herramienta prototipo y el *software* WEKA [26].

B. Análisis de resultados

La totalidad de los usuarios/analistas de datos encuestados concuerdan que el uso de técnicas descriptivas de datos, como también la naturaleza y comprensión de estos datos, es clave para realizar un exitoso proceso de MD.

Un factor importante a considerar es la experiencia previa que tienen los usuarios con las distintas herramientas de MD existentes en la actualidad. Los resultados arrojaron que el 100% de los participantes tienen experiencia previa con el *software* WEKA, seguido de otras como Clementine (67,5%) y Knime (58,3%). De esto se puede inferir que los usuarios tienen experiencia en herramientas de MD, con distintos niveles de experiencias en proyectos de MD.

Tanto para el grupo de usuarios novatos como expertos la experiencia de desarrollar una tarea de MD utilizando el *software* prototipo tuvo una gran aceptación, desde el punto de vista de la usabilidad y desempeño de esta herramienta, manifestando en el 100% que esta herramienta les permite encontrar algún tipo de relaciones entre los atributos del conjunto de datos, por ejemplo los gráficos de dispersión aplicados entre las diferentes variables seleccionadas por el usuario.

En comparación con el *software* WEKA, prevalece una gran valoración del *software* prototipo, respecto del grado de utilidad de las opciones y parámetros que ofrece para poder entender mejor el modelo del ÁD. En el caso de los novatos, 83,3% lo considera alto y 16,7% muy alto, mientras que 80% en los usuarios expertos, y solamente 20% de estos últimos lo consideraron de manera regular.

El gráfico de la Figura 4.a) mide la utilidad del conjunto de visualizaciones entregadas por el prototipo experimental para comprender las particiones realizadas por el algoritmo de ÁD, y se puede observar en este gráfico la alta valoración tanto para usuarios expertos como novatos. Solo 15,4% de los usuarios expertos lo considera de bajo a regular en esta medición.

La capacidad para describir los datos en un nodo mediante el uso de la técnica SOM, para entregar una visualización aumentada del ÁD, considerando a ambos grupos de usuarios, es muy alta y alta,

a) Nivel de utilidad de las visualizaciones entregadas por el *software* prototipo.

b) Nivel de capacidad para describir los datos en un nodo mediante el uso de la técnica SOM.

c) Grado de aceptación del *software* prototipo, para la comprensión del modelo del ÁD.

Figura 4. Gráficos de mediciones del prototipo.

con 38,5% y 61,5%, respectivamente, lo que se ilustra en el gráfico de la Figura 4.b). Con esto se puede deducir que la combinación y aplicación de la técnica SOM sobre un ÁD permite, por un lado, complementar la visualización del modelo generado a partir del ÁD, y por otro, ayudar a mejorar la comprensión de este modelo.

El grado de aceptación del *software* prototipo, según experiencia de los usuarios con otras herramientas de MD, es alto, considerando los dos grupos de usuarios, según se observa del gráfico en la Figura 4.c). Solo el 8% de los usuarios expertos considera suficiente este aspecto. Se puede concluir desde aquí, que la aplicación de la técnica SOM sobre un ÁD permite apoyar la comprensión del modelo generado.

CONCLUSIONES Y TRABAJO FUTURO

Las conclusiones preliminares que se han logrado obtener en este trabajo son las siguientes:

- El esquema VAM-MD y sus lineamientos sobre combinar técnicas de MD *ad hoc* para aumentar visualmente un modelo generado, y con la aplicación de vistas complementarias de artefactos gráficos para ser aplicados a los datos de los distintos componentes de un modelo, sirven como guía para construir visualizaciones que apoyen la tarea de análisis y exploración de modelos de MD, en su etapa de refinamiento o ajuste.
- La percepción positiva obtenida en la encuesta da pie para sostener la conveniencia y utilidad de combinar la técnica sobre un modelo de ÁD generado previamente como técnica complementaria para visualizar y describir las instancias en cada nodo del árbol, aportando con su visualización espacial de los datos por medio de un mapa.
- Considerando la medición subjetiva lograda en el experimento controlado y su encuesta, se pudo observar como resultado una tendencia tanto de los usuarios expertos como novatos, que el uso de la técnica secundaria SOM permite describir complementariamente los datos de los distintos componentes de un ÁD.
- La aplicación de SOM sobre un ÁD, según lo manifestado por los usuarios del experimento en la encuesta, al parecer permite mejorar el entendimiento de modelos de MD respecto de otras herramientas.
- También que la disposición de elementos visuales o artefactos gráficos provistos en el *software* prototipo, aplicados a los datos en cada nodo del ÁD, cumplen con apoyar en el análisis y exploración del modelo generado.

En relación con el trabajo futuro, actualmente se está trabajando en los siguientes aspectos:

- Se están evaluando técnicas descriptivas de MD, que aporten visualizaciones adicionales que provee la técnica SOM para aumentar visualmente un modelo de ÁD.
- Así como también seleccionar artefactos gráficos adicionales que puedan ser de mayor utilidad en la exploración de los datos de cada nodo de

un árbol, preferentemente en conjunto de datos con alta dimensionalidad.

- Implementar en el *software* prototipo mecanismos de interacción que permitan comparar nodos de un ÁD mediante el mapa provisto por la técnica SOM, que mida cuantitativamente el grado de similitud entre estos nodos, por ejemplo con el uso de gráficos de distancias u otro.
- Se tiene previsto explorar y evaluar la posibilidad de combinar otras técnicas de MD, como por ejemplo aumentar visualmente modelos generados por reglas de asociaciones con redes bayesianas.

REFERENCIAS

- [1] P. Adriaans and D. Zantinge. "Data Mining". Addison-Wesley Pub. Reading, Mass, USA. ISBN: 0201403803. 1996.
- [2] B. Becker, R. Kohavi and D. Sommerfield. "Visualizing the Simple Bayesian Classifier". In Proceedings of the KDD-97 Workshop on Issues in the Integration of Data Mining and Data Visualization, Newport Beach, CA, August 17, 1997. Springer-Verlag. 1998.
- [3] E. Brachman and T. Anad. "The Process of Knowledge Discovery in Databases". In Advances in Knowledge Discovery and Data Mining, AAAI Press/ The MIT Press, pp. 37-57. 1996.
- [4] H. Chernoff. "The Use of Faces to Represent Points in k-dimensional Space Graphically". Journal of the American Statistical Association, Vol. 68, Issue 342, pp. 361-368. June, 1973.
- [5] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth. "The KDD Process for Extracting Useful Knowledge from Volumes of Data". Comm. Of the ACM, Vol. 39, Issue 11, pp. 27-34. 1996.
- [6] P. Hoffman. "Table Visualizations: A Formal Model and its Applications". Sc.D. Thesis, Dept. of Comp. Science, University of Massachusetts at Lowell. 1999.
- [7] G. John. "Enhancements to the Data Mining Process". Doctoral Dissertation, Depart. Of Comp. Science, Stanford University. 1997.
- [8] D. Keim. "Visual Techniques for Exploring Databases". Tutorial Notes in the Third International Conference on Knowledge Discovery and Data Mining, KDD-97. Newport Beach, CA. 1997.
- [9] C. Meneses and G. Grinstein. "Visualization for Enhancing the Data Mining Process". In Proceedings of the Data Mining and Knowledge Discovery: Theory, Tools, and Technology III Conference. Orlando, FL. 2001.
- [10] K. Thearling, B. Becker, D. DeCoste, B. Mawby, M. Pilote and D. Sommerfield. "Visualizing Data Mining Models". In Proceedings of the Integration of Data Mining and Data Visualization Workshop, Springer-Verlag. 1998.
- [11] M. Humphrey, S. Cunningham and I. Witten. "Knowledge Visualization Techniques for Machine Learning". Intelligent Data Analysis N° 2, pp. 333-347. 1998.
- [12] P. Vitiello and R. Kalawsky. "Visual Analytics: A Sensemaking Framework for Systems Thinking in Systems Engineering". Systems Conference (SysCon). IEEE International Conference Publications, pp. 1-6. ISBN: 978-1-4673-0748-2. 2012.
- [13] D. Keim, J. Kohlhammer, G. Ellis and F. Mansmann. "Mastering the Information Age Solving Problems with Visual Analytics". Edited by the authors Published by the Eurographics Association Postfach 8043, 38621 Goslar, Printed in Germany, Druckhaus Thomas Müntzer GmbH, Bad Langensalza. Theoretical Issues in Ergonomics Science. Vol. 8, Issue 1, ISBN: 978-3-905673-77-7. 2010.
- [14] Y. Liu and G. Salvendy. "Visualization support to better comprehend and improve decision tree classification modelling process: a survey and appraisal". Theoretical Issues in Ergonomics Science. Vol. 8, Issue 1, pp. 63-92. ISSN: 1463-922X (print). ISSN: 1464-536X (online). 2007.
- [15] G. Klein. "A Recognition-Primed Decision (RPD) Model of Rapid Decision Making". Decision making in action: Models and methods. Vol. 5, Issue 4, pp. 138-147. 1993.
- [16] W. Castillo-Rojas and C. Meneses. "Graphical Representation and Exploratory Visualization for Decision Trees in the KDD Process". XXXVIII Conferencia Latinoamericana en Informática-CLEI 2012, Medellín, Colombia.

- Libro de Resúmenes página 137, ISBN IEEE Xplore: 978-1-4673-0793-2, ISBN CD: 978-1-4673-0792-5. Print ISBN: 978-1-4673-0794-9. 2012.
- [17] W. Castillo-Rojas and C. Meneses. "A Comparative Review of Schemes of Multidimensional Visualization for Data Mining Techniques". III Congreso Internacional de Computación e Informática del Norte de Chile (INFONOR-CHILE). Arica-Chile. 2012.
 - [18] W. Castillo-Rojas, C. Meneses and F. Medina. "Augmented Data Mining Models Using Visualization". Artículo presentado en 6th Latin American Conference on Human Computer Interaction-CLIHIC 2013, Guanacaste, Costa Rica Proceedings pp. 148-155. Springer LNCS 8278. ISSN: 0302-9743. ISBN: 978-3-319-03067-8. DOI: 10.1007/978-3-319-03068-5. 2013.
 - [19] J. Rao and W. Potts. "Visualizing Bagged Decision Trees". Proc. of Third International Congress on Knowledge Discovery and Data Mining, AAAI Press, pp. 243-246. 1997.
 - [20] M. Ankerst, C. Elsen, M. Ester and H. Kriegel. "Visual Classification: An Interactive Approach to Decision Tree Construction". Proc. of Fifth International on Knowledge Discovery and Data Mining, pp. 392-397. 1999.
 - [21] L. Breiman, J. Friedman, R. Olshen and C. Stone. "Classification and Regression Trees". Belmont, CA: Wadsworth. 1984.
 - [22] B. Johnson and B. Shneiderman. "Treemaps: A Space-Filling Approach to the Visualization of Hierarchical Information Structures". Proc. of IEEE Information Visualization, pp. 275-282. 1991.
 - [23] G. Robertson, J. Mackinlay and S. Card. "Cone Trees: Animated 3D Visualization of Hierarchical Information". Proc. of the ACM Congress on Human Factors in Computing Systems, pp. 189-194. 1991.
 - [24] J. Lamping and R. Rao. "The Hyperbolic Browser: A Focus + Context Techniques for Visualizing Large Hierarchies". Journal of Visual Languages and Computing, pp. 33-55. 1997.
 - [25] T. Barlow. "Case study: visualization for decision tree analysis in data". Visualization of decision trees. 2001.
 - [26] W. Castillo-Rojas, C. Meneses and F. Medina. "Encuesta en línea *software* prototipo VOLAM-VDM". URL: <https://es.surveymonkey.com/s/EncuestaVolam>. Fecha de Consulta: 12 de mayo de 2013.