Una Revisión Comparativa de Esquemas de Visualización Multidimensional para Técnicas de Minería de Datos

A Comparative Review of Schemes of Multidimensional Visualization for Data Mining Techniques

Wilson Castillo Rojas
Área Computación, Departamento de Ingeniería
Universidad Arturo Prat, UNAP
Iquique, Chile
wilson.castillo@unap.cl

Claudio Meneses Villegas
Departamento de Ingeniería de Sistemas y Computación
Universidad Católica del Norte, UCN
Antofagasta, Chile
cmeneses@ucn.cl

***Resumen*— Este artículo presenta un estudio y análisis comparativo de diversos esquemas de visualización multidimensional aplicados a modelos generados por algunas técnicas de minería de datos. La principal motivación es analizar las distintas formas que actualmente están utilizando los esquemas de visualización, para abordar el problema de la alta dimensionalidad de los datos y su manejo para evitar la obstrucción visual o reducción de esta dimensionalidad, como también conocer los mecanismos o grado de interacción que incluyen. Por otro lado, cómo la visualización permite lograr una mejor comprensión de éstas técnicas de minería de datos, y su facilidad de interpretación de los patrones o reglas obtenidas. Esta base de estudio permitirá plantear nuevos esquemas de visualización o combinación de estos esquemas, en distintas técnicas de minería de datos, considerando la incorporación de mecanismos de interacción que permitan al usuario navegar por los modelos, con el objetivo final de detectar patrones y entender como los algoritmos transforman datos en patrones que pueden constituir conocimiento nuevo y útil.**

Palabas claves; Esquemas de Visualización, Visualización Multidimensional, Visualización de Modelos para Minería de Datos, Visualización para Minería de Datos, Minería de Datos.

***Abstract*— This article presents a study and comparative analysis of various schemes of multidimensional visualization applied to models generated by some data mining techniques. The main motivation is to analyse the different forms which are currently using visualization schemes to address the problem of the high dimensionality of the data and its management to avoid visual obstruction or reduction of this dimensionality, as well as to know the mechanisms or degree of interaction that include. On the other hand, how visualization allow to achieve a better understanding of these techniques of data mining, and ease of interpretation of patterns or learned rules. This basis of study will allow to raise new schemes of visualization or combination of these schemes in different data mining techniques, whereas the incorporation of interaction mechanisms that allow the user to browse models, with the ultimate goal of detecting patterns and understanding how algoritmos transforms data into patterns that it can be novel and useful knowledge.**

Keywords-component; Schemes of Visualization, Multidimensional Visualization, Visualization of models for Data Mining, Visualization for Data Mining, Data Mining.

I. INTRODUCCIÓN

La Minería de Datos (MD) utiliza un conjunto de métodos analíticos para descubrir relaciones previamente desconocidas en datos históricos de una organización, relaciones que a menudo son complejas, y que representan patrones de comportamiento en los datos analizados, lo que se traduce en un nuevo conocimiento, patrones que a simple vista no son simples de identificar, y para lo cual es necesario apoyarse en un análisis visual. El análisis visual es una capacidad del ser humano que le permite percibir e identificar estos patrones, siempre y cuando disponga de una visualización apropiada.

Y es aquí donde cobra relevancia el desarrollo de esquemas de visualización con dos características fundamentales:

- Que permitan comprender fácilmente los modelos y patrones:

  En términos generales los esquemas de visualización deben ser: apropiados para representar los patrones obtenidos, simples e intuitivos para el manejo del usuario final, con capacidad de manejo de gran cantidad de patrones descubiertos (evitando la oclusión o obstrucción en la interfaz).
- Que permitan interactuar con modelos y patrones:

  En lo que respecta a la característica de interacción, se deben considerar los trabajos de interacción humano computador (HCI) que se centran en el enfoque de búsqueda en Información visual ilustrado por Shneiderman: "*Overview first, zoom and filter, then details on demand*" [1], según esto, en primer lugar las herramientas gráficas deben proporcionar una vista global del sistema para deducir los puntos principales, el usuario debe poder determinar el punto de partida de su análisis gracias a esta vista global, luego el usuario debe poder explorar áreas particulares profundizando si desea.

Puesto que las técnicas de visualización han demostrado actualmente, ser una solución eficiente para alcanzar estas dos características fundamentales, el desarrollo de herramientas gráficas eficientes y fáciles de utilizar para los usuarios es hoy un desafío prometedor de la MD, las cuales deben ser capaz de desplegar conocimiento explícito y

restituirlo, a partir de los modelos y sus parámetros de entrada, todo en una forma interactiva, gráfica que fomente a menudo nuevos entendimientos, y de esta manera motiven al usuario a formar y validar nuevas hipótesis para una mejor solución del problema y ganando un conocimiento más profundo del dominio.

Sin embargo, se debe insistir en que no solamente la visualización debe servir para el usuario final, sino que además también debe facilitar la comprensión de los modelos complejos que se presentan durante el proceso de MD.

Considerando estas premisas, se presenta a continuación una revisión de trabajos de relacionados con visualización multidimensional sobre diversos esquemas de visualización aplicados a técnicas de MD, que permiten en algunos, comprender mucho mejor el funcionamiento del modelo.

Se considera en este estudio, las técnicas de ; *Reglas de Asociación*, *Árboles de Decisión*, *Clasificador Bayesiano*, y *Redes Neuronales Artificiales*. Algunas de estos nuevos esquemas han sido ya implementados por conocidos sistemas comerciales de MD.

Para cumplir con el objetivo propuesto, este artículo se organiza de la siguiente manera: en la sección II se explican las características de Visualización en MD y algunas clasificaciones, en la sección III se analizan implementaciones de nuevos esquemas de visualización para las técnicas de; *Reglas de Asociación*, *Árboles de Decisión*, *Clasificador Simple de Bayes*, y para *Redes Neuronales Artificiales*. En la sección IV se establecen y discuten algunas evaluaciones de visualizaciones en MD. Finalmente en la sección V se plantean las conclusiones y el trabajo futuro proyectado sobre la base de éste.

II. VISUALIZACIÓN EN MINERÍA DE DATOS

La visualización actualmente cumple un rol promisorio en el ámbito de la ciencia de la computación, ya que se enfoca en una de las características más avanzadas de los humanos, la visión. Sin embargo, la visualización sólo ha sido desarrollada de manera parcial para los procesos de MD, limitada solamente a la visualización de resultados obtenidos en este proceso. Muchos autores coinciden actualmente en la necesidad de que la visualización sea extendida a otras etapas claves del proceso análisis de datos KDD (*Knowledge Discovery in Databases*), como la exploración de datos y la visualización de modelos.

Al explorar datos, los seres humanos buscan estructuras, patrones y relaciones entre los datos. Este análisis es más fácil si los datos se presentan en forma gráfica, lo que se conoce como visualización gráfica.

Una visualización satisface varios propósitos: proporciona una descripción de conjuntos de datos complejos y grandes, muestra un resumen de los datos y ayuda en la identificación de patrones y estructuras en los datos.

Según Fayyad [60] ; "*El objetivo de la visualización es reducir la complejidad de un conjunto de datos dado y, al mismo tiempo, perder la menor cantidad de información*".

Grinstein en [61] define Visualización como : "*la comunicación gráfica (como opuesto a lo textual o verbal) de información (datos, documentos, estructura)*".

2.1 Visualización y Exploración de Datos

La visualización es un proceso que transformar la información en una forma visual permitiendo al espectador observar, navegar, interactuar, entender el sentido y la información resultante. Utiliza las computadoras para procesar datos y su visualización, con métodos gráficos interactivos, de proyección de imagen, y diseño visual, y requiere la capacidad visual del ser humano para percibir y procesar esta información.

La exploración de datos consiste en descubrir información o patrones, en un gran volumen de datos, a través de un proceso de MD. Para ello, es necesario realizar una secuencia de transformaciones y depuración, sobre el conjunto de datos, para poder descubrir la información o patrón de conocimiento resultante. Este proceso de exploración es variado y complejo, requiriendo un alto grado de habilidades técnicas y perceptivas por parte del analista. El objetivo de una exploración de datos es lograr una descripción de estos y descubrir hechos interesantes en ellos, que debe incluir: calidad de los datos, búsqueda de estructuras o patrones, obtención de un conocimiento básico de los datos y las relaciones existentes entre las variables.

La interacción es un componente fundamental de la visualización que permite al usuario modificar sus parámetros de cofiguración, y puede interactuar recíprocamente con los datos presentados de diversas formas, tales como ; navegando, muestreando, a través de consultas, con la manipulación de parámetros gráficos, especificando las fuentes de datos a ser desplegadas, creando la salida para el análisis adicional o del despliegue de otra información disponible sobre los datos [61].

Los algoritmos automáticos para MD son capaces de analizar y extraer estructuras cada vez más complejas de datos. Una consecuencia del uso de estos algoritmos es que el usuario humano ha sido alejado del proceso de exploración de datos. Este proceso también ha llegado a ser más difícil de comprender para el usuario, que tiene que entender la estructura de los datos y los detalles del proceso de exploración. La visualización ayuda en el proceso MD de dos maneras ; puede proporcionar la comprensión visual de complejas aproximaciones computacionales, puede ser utilizada para descubrir relaciones complejas entre los datos que no son detectables por los métodos de cómputos actuales, pero que puede ser trazado por el sistema visual humano.

La hablidad humana de percepción visual permite al usuario analizar complejos eventos en un intervalo corto de tiempo, reconocer patrones importantes y tomar decisiones mucho más efectivas que el computador. Para lograr esto, el usuario debe tener los datos presentados de una manera lógica con una buena descripción de toda la información. Ésta es la razón que el desarrollo de los métodos eficaces de visualización están llegando a ser cada vez más importante [36] [60] [62].

2.2 Visual Data Mining

La minería de datos visual o Visual Data Mining (VDM), se basa en la integración de conceptos desde computación gráfica, métodos de visualización de información, y percepción visual en análisis de datos exploratorio. El concepto integra la capacidad humana de percepción con el poder de cálculo y capacidad de almacenamiento del computador [64].

Una vez obtenidos los resultados de un proceso de MD surge una nueva interrogante, ¿cómo interpretar estos datos?. Una opción es a través de tablas con resultados numéricos, valores estadísticos, etcétera, opción si bien viable, muchas veces impracticable por el masivo caudal de datos que es necesario procesar. Sin embargo, visualmente es mucho más provechoso llevar los datos a representaciones gráficas que proporcionen una idea del comportamiento y evolución que van teniendo estos datos, tomando diferentes variables como parámetros, es así como tienen relevancias los aspectos de recuperación de información, representación (meta-datos) y visualización de resultados enun proceso de MD.

Ankerst define la minería de datos visual como: "*... un paso en el proceso de descubrimiento del conocimiento que utiliza la visualización como canal de comunicaciones entre la computadora y el usuario para obtener patrones nuevos e interpretables*" [36].

Se tiene otra definición formal para VDM a partir de Souk [55], que afirma: "*la llave que ayuda al negocio y al análisis de datos a descubrir nuevos patrones y tendencias. La visualización provee métodos para comunicar estos descubrimientos a quienes deben tomar decisiones*".

Las ventajas principales de la exploración visual de los datos sobre la MD automática, son que la exploración visual permite una interacción directa con el usuario y proporciona una retroalimentación inmediata, que es intuitiva y que no requiere ninguno o solamente poca comprensión de algoritmos matemáticos o estadísticas complejas. Como resultado de esto, el usuario obtiene una mayor confianza en los patrones resultantes y en el proceso. La investigación reciente ha probado que una visualización apropiada reduce tiempo para reconocer la información y sentido que toma esto. El usuario humano también proporciona conocimiento adicional del dominio del problema y aplica más efectivamente restricciones a la MD automática.

Otra ventaja de la exploración visual de los datos sobre los métodos automáticos de MD, es que pueden manejar fácilmente los datos no homogéneos y los que generan ruidos (datos espurios), que son difícil de analizar con métodos puramente automáticos.

Las herramientas VDM actuales, se encargan de entregar representaciones visuales que es posible interpretar, sin embargo, muchas veces estas herramientas resultan tan complejas que sólo especialistas en análisis de datos o estadísticos, pueden manejarlas, alejándolas de quienes realmente deberían tener el control de la visualización e interpretación de los resultados de un proceso de MD.

2.3 Clasificación de Métodos de Visualización

Cuando se dispone de datos en gran volumen, y antes de abordar análisis más complejos, un primer paso consiste en presentar esta información de forma que se pueda visualizar de una manera sistemática y resumida [63]. Los datos que interesan dependen, en cada caso, del tipo de variables que se manejan. Se pueden separar las visualizaciones en dos grandes grupos [55]:

i. Visualizaciones Multidimensionales: Son representaciones comúnmente utilizadas en el campo de visualización de un proceso de MD. Este tipo de visualización, permite al usuario comparar visualmente las dimensiones de un dato respecto a otra dimensión, para lo cual ocupa un sistema de ejes de coordenadas.
ii. Representaciones Jerárquicas y Visualización de Mapas: Corresponden a formas complejas de agrupación de datos, los cuales no son visibles a simple vista y que requieren de procedimientos y visualizaciones no convencionales. Para ejemplificar esto se puede pensar en la forma habitual de representar los lazos familiares, un árbol de familia presenta de forma jerárquica a sus integrantes y existen ocasiones en que los datos tienen ocultas estas jerarquías, que a través de representación de árbol se pueden visualizar.

De acuerdo a la secuencia de pasos automáticos de la MD y la parte visual interactiva, en [36] se propone la siguiente clasificación:

- Visualización del resultado de MD: los patrones producidos por un algoritmo automático de MD son visualizados en un paso de post-procesamiento para hacerlos interpretables.
- Visualización de un resultado intermedio: un resultado intermedio del algoritmo de MD es visualizado, del cual el usuario detecta relaciones y patrones, que él entonces utiliza para el análisis adicional con el mismo u otro algoritmo automático de MD.
- Visualización de los datos: los datos se visualizan como un paso de pre-procesamiento antes de cualquier algoritmo de MD sea aplicado sobre ellos, para descubrir y quitar cualquier registro inconsistente o espurio desde el conjunto de datos.

Una tercera clasificación es la que se presenta en [65] y que considera tres criterios, el tipo de ; datos a ser visualizados, método de visualización, y el grado de interacción y distorsión.

Estos tres criterios son ortogonales entre sí, y forman un espacio tridimensional de visualización. Esto significa que algunas visualizaciones pueden ser utilizadas para los tipos de datos y con algunos de los métodos de interacción. Keim [9] define los tipos de datos siguientes: datos unidimensionales, datos de dos dimensiones, datos multi-dimensionales, texto e hipertexto, jerarquías y gráficos, y finalmente algoritmos y software. Los tipos de visualización son los gráficos: estándar 2D/3D, geométricamente

transformadas, basadas en iconos, densidad de pixel y jerárquicos.

Las interacciones y métodos de distorsión son: proyección interactiva, filtración interactiva, enfoque interactivo, distorsión interactiva, y cepillados y lazos interactivos.

Otras dos clasificaciones dividen visualizaciones en ; geométrico o simbólico, o en estático o dinámico.

Los datos en un método de visualización geométrico se representan usando líneas, superficies o volúmenes, y por lo general son de tipo numéricos. La visualización simbólica representa datos no numéricos usando pixeles, iconos, arreglos o grafos. Tanto para geométrico y simbólico la representación visual en pantalla puede tener dos o tres dimensión o ser estereoscópica. La división entre los métodos estáticos y dinámicos se basa en el tipo de datos representados, si los datos en sí mismo son estáticos o cambian en un cierto tiempo [60].

2.4 Representación de un Modelo de Minería de Datos

La representación de los modelos de un proceso de MD puede ser separado en dos áreas claves: la comprensión y la confianza [24].

La comprensión es indudablemente el motivo fundamental para representar un modelo, aunque la manera más sencilla de tratar con un modelo de MD es considerar este proceso en forma de una caja negra, en la cual el usuario no ganará necesariamente una comprensión de la conducta fundamental en la que ellos están interesados.

Una manera interesante de utilizar un modelo de MD, es donde el usuario entiende realmente lo que pasa, para que puedan tomar medidas directamente. Representar un modelo de MD, debe permitir a un usuario discutir y explicar la lógica detrás del modelo con los demás. La obtención de la lógica o la base, forma parte de la construcción de la confianza de los usuarios en los resultados y en el proceso en sí. A menos que el resultado de un proceso de MD se pueda entender cualitativamente, no será de gran utilidad. Además, las necesidades del modelo deben ser entendidas para que las acciones que se tomen puedan ser justificadas como resultado.

Dentro de los medios de comprensión de un problema juega también un rol importante, si el usuario puede entender lo que se ha descubierto en el contexto de su negocio, ellos confiarán en el resultado y lo pondrán en uso. Hay dos partes en este punto ; la representación gráfica de los datos deben influir en significativamente al entendimiento de un fenómeno, y debe permitir al usuario interactuar con la gráfica, para que las preguntas sencillas se puedan responder.

2.5 Visualización en el Proceso de KDD

En [66] se establece la visualización como concepto en el proceso KDD, como una herramienta que ayuda a entender de mejor manera a los usuarios, en las distintas etapas de este proceso.

En general, la visualización puede jugar un rol fundamental en las tareas de descubrir conocimiento a partir de datos, y puede ser utilizada principalmente para:

- Resumir grandes volumenes de datos de una manera compacta.
- Apoyar la tarea humana de entender y descubrir hechos interesantes en los datos.
- Entender y explorar modelos (representación de patrones) derivados desde los datos por la aplicación de algoritmos de MD.
- Reportar resultados del proceso KDD.

Algunas aplicaciones de visualización en diferentes etapas del proceso KDD incluyen [12]:

- Visualización de Datos Multidimensionales
- Visualización del Espacio de Parámetros de un algoritmo de MD.
- Visualización de Modelos y Patrones generados por un algoritmo de MD.

2.6 Visualización Multidimensional

La visualización de datos multidimensionales tiene un tremendo efecto en ciencia, ingeniería, y en la toma de decisiones en los negocios, sin embargo presentan problemáticas multidimensionales, donde los conjuntos de datos por lo general contienen mas de tres columnas de datos. El uso de visualización para dar sentido a los datos es una técnica poderosa para los usuarios en la toma de decisiones.

Hay diversos enfoques para visualizar conjunto de datos multidimensionales, entre los pioneros se encuentran las matrices de permutaciones de Bertin's [2] donde los datos son representados en filas y columnas, y las celdas contienen descripciones gráficas. Con el arreglo de estas filas y columnas, los usuarios pueden formar cluster de datos, generalmente en la diagonal de la matriz, para entender la distribución de los datos.

Otras propuestas que también han planteado técnicas para visualización multidimensional, son las caras de Chernoff's [3] que hace uso de iconos para representar las multi-dimensiones y se encuentra entre los trabajos más conocidos del área. Tufte [4] también provee muchos ejemplos de visualizaciones multidimensionales. Los trabajos más recientes incluyen las coordenadas paralelas de Inselberg's [5-6], mundos dentro de mundos [7], la tabla lentes [8], VisDB [9], consultas dinámicas [10], y explorador de atributo [11]. Finalmente, hay que destacar la clasificación que realiza Wong en [13], donde establece cuatro etapas del desarrollo de la visualización multidimensional.

III. VISUALIZACIÓN DE MODELOS Y PATRONES GENERADOS POR TÉCNICAS DE MINERÍA DE DATOS

En esta sección se describen trabajos relacionados con esquemas de visualización de patrones y modelos, aplicadas en los últimos años en técnicas de MD, particularmente se presentan ejemplos en: Reglas de Asociaciones, Modelos Jerárquicos (Árboles de Desición), Clasificador Simple de Bayes (Clasificador Probabilísticos), y Redes Nueronales Artficiales.

3.1 Visualización de Reglas de Asociaciones

Las Reglas de Asociaciones (RA) representan las relaciones entre varias variables unas con otras, es decir, se debe considerar que una RA es una implicación de la forma $X \rightarrow Y$, donde X es un conjunto de ítemes llamados antecedentes, e Y es el ítem consecuente. Al menos cinco parámetros debe ser considerados en una visualización de las RA: el conjunto de ítemes antecedentes, los ítemes consecuentes, las asociaciones entre antecedentes y consecuentes, el soporte de las reglas, y la confianza de las reglas.

La investigación de visualización de RA se puede categorizar en tres grupos principales basados en : Tablas, Matrices, y en Grafos.

Las técnicas basadas en Tablas son el enfoque más común y tradicional para representar RA. Las columnas de una tabla representan generalmente los ítemes, los antecedentes, los consecuentes, el soporte, y la confianza de las RA, y cada fila representa una RA. Algunos ejemplos de técnicas basadas en tablas se pueden encontrar en los sistemas comerciales SAS Enterprise Miner [14] y DBMiner [15].

Las técnicas basadas en Matriz tales como las implementadas en MineSet [16] (matriz 2D), matriz 3D en INFOVIS [17], donde la grilla representan los antecedentes, y los consecuentes son presentados en una grilla cuadrada basada en ejes de coordenadas.

El último grupo son las técnicas basadas en Grafos, y utilizan nodos para representar los ítemes, y los bordes para representar las asociaciones de los ítemes en las reglas.

Algunos de estos trabajos, han propuesto varios tipos de representaciones conocidas para estudiar un gran conjunto de datos como árboles hiperbólicos [18], paredes de perspectiva [19], vistas de ojos de pez (Fish Eyes View - FEV) [20] o superposición de vistas transparentes [21].

De estas, la vista de ojos de pez (FEV) muestra el punto de interés detalladamente y la descripción del despliegue en la misma representación. Un primer acercamiento utilizando FEV en un método de visualización de RA fue propuesto en [22]. En este trabajo, los autores explotaron una mezcla de representación para desplegar un conjunto de reglas, y actualmente varias herramientas implementan FEV como AISEE [23], INFOVIS [65] y LARM [25].

Hay otras técnicas, no especificadas en esta categorización, estas son: Gráficos de mosaico interactivos [26], que aplica mosaicos para representar las relaciones entre ítemes en cada RA desde una tabla de contingencia en vez de visualizar sus resultados de minería, CrystalClear [27] que utiliza una técnica integrada de grilla con árbol basado en el despliegue para ver el número de ítemes y las listas de antecedentes/consecuentes, y una técnica propuesta por Zhao y Liu [28] que utiliza una línea para representar cada RA, el eje x representa datos del tiempo, el eje y representa el valor del soporte o confianza, su visualización utiliza una técnica similar a la técnica de coordenadas paralelas [29]. Finalmente, una herramienta visual integrada de DM en [30].

Aunque estas tres técnicas están bien implementadas en algunos sistemas comerciales de visualización, son capaces de desplegar RA minadas, pero las visualizan todas de una sola vez en una misma pantalla, presentando demasiada información, lo que suele también generar distorsión en la interfaz, haciendo difícil su análisis para los usuarios.

A. VisAR

En uno de los estudios revisados aparece una nueva técnica llamada VisAR, la cual se centra en la reducción de la complejidad de visualizar un gran número de RA en un sola interfaz [31].

La figura 1 muestra un ejemplo de visualización de ítemes seleccionados en VisAR (para este caso: Cd y arroz), donde todas las RA que contienen los ítemes especificados se visualizan en el panel derecho, y todos los antecedentes y consecuentes de las RA calificadas se despliegan a lo largo del eje y. Los antecedentes se colocan en la parte superior y los consecuentes debajo del eje x que se muestra como línea en negrilla y color negro. Los ítemes de interés seleccionados se despliegan sobre otros ítemes no seleccionados en los antecedentes a lo largo del eje y. Cada regla es visualizada por una línea vertical paralela al eje y con puntos circulares que representan ítemes en cada RA.

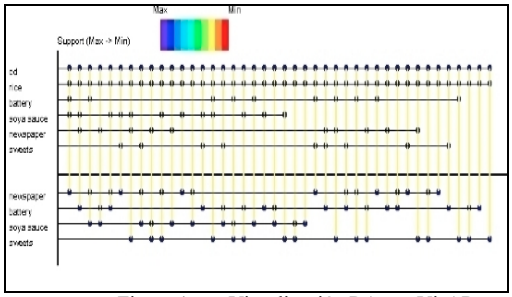


Figura 1: Visualización RAs en VisAR.

B. Visualización de RA Basada en Matriz

Otro estudio revisado sobre técnicas de visualización de RA es el que se presenta en [32], y que permite visualizar RA del tipo uno-a-mucho. En vez de usar celdas de una matriz 2D para mostrar las RA de ítem-a-ítem, utiliza la matriz para representar la relación regla-a-ítem. En la figura 2, se puede observar que las filas del piso de la matriz representan los ítemes, y las columnas representan las asociaciones del ítem. Los bloques azules y rojos de cada columna (regla) representan el antecedente y consecuente de la regla.

Cambiando desde un diseño de un arreglo ítem-a-ítem a un diseño de una matriz de dos dimensiones de regla-a-ítem, deja efectivamente desplegar RA con tópicos antecedentes múltiples.

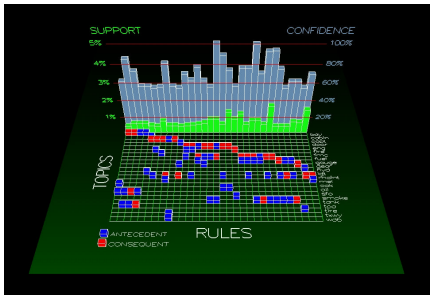


Figura 2: Visualización de asociación de ítem.

C. CBVAR

Una tercera revisión de esquema de visualización de RA, es el que se plantea en [33] y presenta un prototipo de visualización escalable llamado CBVAR. Su originalidad radica en que utiliza un meta-conocimiento final para guiar RA con agrupamientos.

Esto último se logra porque utiliza un Framework integrado para la extracción y visualización de RA, con esto evita modificar el formato de salida de los algoritmos de MD existentes para fijar los requisitos de entrada del módulo de visualización, por lo que sólo se debe proporcionar un archivo de datos con las RA. Se compone de dos partes independientes: la generación de clusters y la visualización de RA agrupadas:

- Para generar los clusters, utiliza el algoritmo K-media que es aplicado sobre los datos, y se obtienen clusters disjuntos cada uno con un conjunto reducido de RA.
- Para la visualización, combina representaciones 2D y 3D. Con el despliegue 2D permite tener una vista global de las RA y una regla por cada cluster, regla que se selecciona de acuerdo a una medida de elevación, soporte o confianza. Los usuarios pueden variar las métricas soporte y confianza. La información y reglas principales del clusters se dan en áreas de texto.

Cada área representa una regla del correspondiente cluster y utiliza la técnica FEV para detallarlo. Para mostrar el detalle de lo enfocado, despliega una representación de proyección 3D para aborda aspectos de obstrucciones y la representación 2D, respectivamente para los contextos escasos y densos. Con esto proporciona una guía de exploración basada en un cluster de RA, este segmento implementa la recomendación de Shneiderman [1]. Un muestra de representación de RA a través de este esquema de visualización, es el que se puede observar en la figura 3.

El objetivo de este trabajo no es detallar la información del cluster actual pero si visualizarlos todos al mismo tiempo, luego de seleccionar un cluster, se puede ejecutar LARM [34] que es capaz de obtener la información detallada.

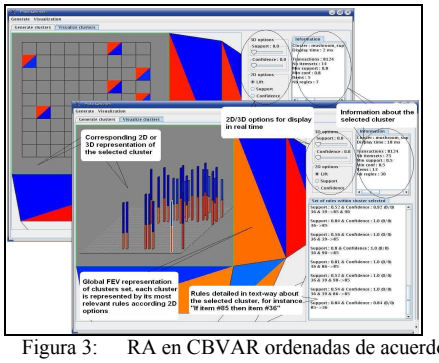


Figura 3: RA en CBVAR ordenadas de acuerdo a los consecuentes.

Los experimentos documentados demuestran que se pueden desplegar muchas reglas en pocos segundos y que la visualización propuesta basada en cluster, puede ser una herramienta eficiente para manejar una gran cantidad de RA, gracias a una mezcla de representación que explota representación 2D y FEV.

3.2 Visualización de Árboles de Decisión

Los Árboles de Decisión (ÁD) corresponden a uno de los modelos predictivos más utilizados en MD. Estos se crean particionando el conjunto de datos ordenados por algún valor que los jerarquiza, en subconjuntos, y después dividiendo cada uno de los subconjuntos, hasta que los subconjuntos 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 reglas de particionamiento. Las ramas de un nodo son los subconjuntos creados a partir de un nodo. Una regla de división utiliza típicamente una sola variable al asignar un caso a una rama.

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, 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 a visualizadores 3D en MineSet, CATscan para inducir ÁD "recubiertos" [35], visualización interactiva en la construcción de ÁD [36], visualizador en una mapa de árboles en el sistema CART [37] y Salford. Se pueden presentar dos problemas con la visualización de ÁD que no han sido investigados de manera profunda; 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 [38], cone trees [39], árbol hiperbólico [40], que aun no han sido consideradas en KDD.

A. *TreeMap*

Los treemaps fueron diseñados originalmente para visualizar archivos de disco duro de un computador, pero se han aplicado, posteriormente, a un amplio rango de colecciones de datos, como análisis geo-referenciales o resultados deportivos. También es utilizada esta técnica para permitir la visualización y exploración en línea de datos financieros de alta dimensionalidad. Un ejemplo de esta herramienta de visualización es el mapa del mercado, utilizado para observar los cambios en las acciones casi en tiempo real.

Existe una amplia familia de algoritmos [59] para particionar el área de visualización de un treemap. En todos estos casos, la entrada es un conjunto de n números y un rectángulo. Por cada número n, se crea una partición del rectángulo original. El algoritmo garantiza que toda el área disponible se rellena, y que los rectángulos tienen un área proporcional a los valores del conjunto de datos.

En particular, los treemap por su naturaleza han sido utilizados ampliamente para representar Á-D, de modo de establecer una visualización plana de su estructura jerárquica, y donde cada nodo del árbol, partiendo desde su raíz tienen una representación cuadricular.

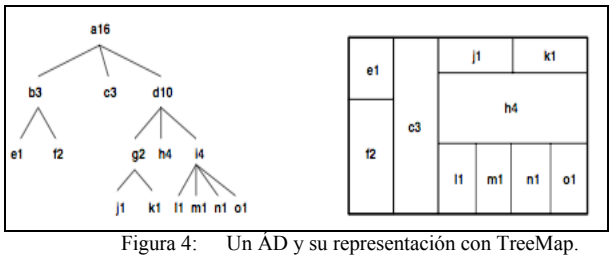


Figura 4: Un ÁD y su representación con TreeMap.

Como se puede observar en la figura 4, la naturaleza jerárquica de los Á-D para establecer reglas de decisión generadas a través de un modelo de análisis de datos, puede ser fácilmente representado e interpretado a través de los cuadrantes definidos en un treemap, tomando los niveles del Á-D como base la raíz (a16) y sus nodos hijos (b3, c3 y d10) a nivel vertical, luego en el segundo nivel, la partición y lectura es a nivel horizontal dentro de cada nodo, logrando representar hasta un tercer nivel (hojas del Á-D). Utilizando estas particiones verticales y horizontales combinadas en cada nivel en un treemap, se logra representar jerárquicamente los datos de manera similar a las ramas de un Á-D.

B. *EMTree Results Viewer*

Un trabajo reciente en el desarrollo de esquemas de visualización de ÁD es el que se presenta en [41], y se han implementado en una herramienta llamada EMTree Results Viewer. Está diseñado para ayudar a construir y entender ÁD complejos visualizando la partición de casos, que ayuda al analista a comprender las predicciones de un modelo, y visualizando los diagnósticos del modelo, que ayuda a determinar la confiabilidad del modelo. La creación y evaluación de ÁD es un proceso iterativo, requiere visualizaciones de árboles en los niveles con variación de detalle, diagramas de diagnóstico, y una variedad de tablas. La figura 5 muestra una interfaz de la aplicación con 4 vistas ligadas disponibles para el usuario.

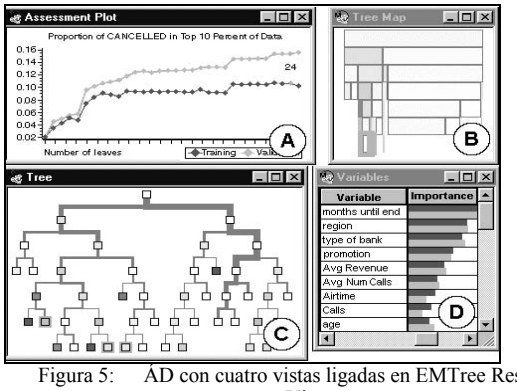


Figura 5: ÁD con cuatro vistas ligadas en EMTree Results Viewer.

Se deduce de éste trabajo que el proceso de análisis de un ÁD, implica la conmutación frecuente entre las vistas y su modificación. Además, que las vistas ligadas de un ÁD ofrecen una variedad de maneras de visualizar la información relevante, que permite también interactuar con la presentación de la información apoyando la comprensión de las relaciones entre las vistas.

C. *CABRO con T2.5D*

Otro trabajo revisado sobre ÁD y documentado en [42], presenta un esquema de visualización para el aprendizaje interactivo en ÁD de gran tamaño, donde se incluye una nueva técnica llamada T2.5D (relacionado con árboles de 2.5 Dimensiones). Este trabajo presenta un desarrollo de métodos que puedan cumplir con dos objetivos principales; (1) visualizar de manera eficiente ÁD grandes; y además (2) visualizar su proceso de aprendizaje.

Para enfrentar este desafío, se considera por un lado que la estructura de los ÁD contienen dos tipos de información, la información estructural del árbol, y la información contenida en cada nodo. Luego se establece un conjunto de requerimientos para lograr los dos objetivos descritos anteriormente, y estos son:

1. Embebido en el proceso de inducción: La información estructural y de contenido debe estar disponible no sólo después de la inducción del árbol, sino también durante el proceso de inducción. El visualizador de ÁD debe proveer la habilidad de acceder rápidamente y manipular fácilmente un gran árbol de acuerdo a las operaciones del proceso de aprendizaje del ÁD.
2. Comprensión: El visualizador de árboles debe facilitar el entendimiento de su estructura y la relación con sus nodos.
3. Eficiencia: Se debe tener un uso eficiente del espacio visual de las herramientas al desplegar árboles grandes.
4. Interacción: Debe existir un control interactivo sobre la estructura con la habilidad para los usuarios de poder modificar diferentes diseños.

5. Estética: Debe contar con una presentación amigable y confortable para el usuario.

Este visualizador fue implementado para el sistema CABRO, y despliega un árbol en dos vistas relacionadas, la global y otra enfocada al detalle (figura 6), cada una de estas vistas pueden ser expandidas o replegadas y soportan los siguientes modos:

- Vista de paneles ajustados: La estructura global de un ÁD se muestra en un panel y su detalle en otro.
- Vista personalizada: El árbol puede ser desplegado de manera completa o sólo su raíz y sus hijos directos, por lo que el usuario puede usar vistas interactivas del árbol de acuerdo a sus necesidades.
- Vista de decisión/nodos hojas: El usuario puede abrir un nodo y ver su información: asociando el atributo y su valor correspondiente, número de casos cubiertos, la clase más grande, el porcentaje asociado, el camino recorrido desde el nodo raíz, etc.
- Modo pequeño con FEV: permite focalizar la visualización de manera de manejar mejor el espacio para visualizar la estructura del árbol, el usuario puede determinar rápidamente el punto de vista y tomar una vista panorámica del área de interés, esto le permite ver la estructura del árbol mientras presta atención a diferentes partes de sus relaciones.

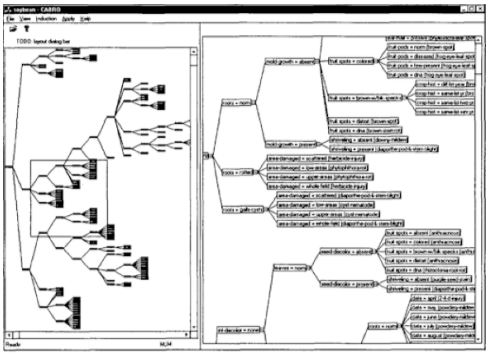


Figura 6: Visualización de AD en CABRO con vistas relacionadas.

Con árboles muy grandes, el usuario puede encontrar difícil navegar sobre ellos, incluso con la vista pequeña y el uso de FEV, se ha utilizado la nueva técnica llamada T2.5D para resolver este problema. Para explicar esta técnica, se necesita definir el concepto de camino ancho de un nodo del árbol, se entiende como el conjunto de todos los nodos en el camino desde la raíz al nodo y sus hermanos, en otras palabras contiene a todos sus familiares directos (figura 7). Cuando se visualiza un árbol, en cada movimiento el camino ancho está activo (active wide path - AWP).

En las aplicaciones de esta nueva técnica se describe algunos resultados con ÁD grandes:

- T2.5D logra fácilmente manejar ÁD con más de 20000 nodos, y mas de 1000 nodos pueden ser desplegados juntos en una sola pantalla.
- Esto es útil para tener una vista clara de un AWP de un nodo en la imagen del árbol y tener la vista completa del árbol al mismo tiempo.
- La navegación incluso en árboles enormes es rápida y fácil.

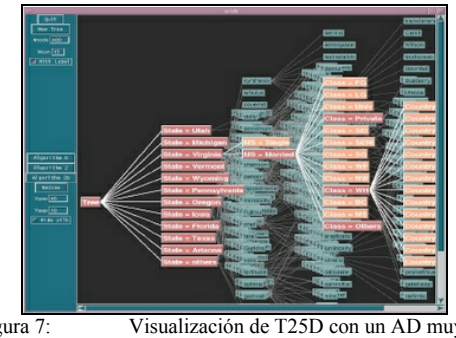


Figura 7: Visualización de T25D con un AD muy grande.

Existen varios métodos para resolver los tres principales problemas de inducción de ÁD en la selección de atributos, poda, y discretización. Se sabe que ninguno de ellos es superior a otro, esto genera un problema práctico en la selección del modelo, como escoger el método más apropiado para una tarea de MD. Este problema requiere un meta-conocimiento y evaluación comparativa entre métodos y modelos.

En lo que específicamente se refiere al apoyo de esta técnica en la visualización del proceso de aprendizaje de un ÁD, se presenta una tabla de comparación considerando los 5 requisitos descritos al comienzo, dando como resultado lo mostrado en la Tabla 1.

TABLA I. COMPARACIÓN DE ALGUNOS VISUALIZADORES DE ÁD.

Criterios	3D en Mineset	2D en CART	Árboles Hiperbólicos	T2.5D
Embebido en Proceso de Inducción	No	No	N/A	Si
Comprensión	Medio	Medio	Alto	Alto
Eficiencia	Medio	Medio	Alto	Alto
Interactividad	Medio	Alto	Alto	Medio
Estética	Alto	Medio	Muy alto	Alto

Los tres principales criterios para seleccionar el modelo son; tamaño, exactitud, y entendimiento en árboles grandes.

El tamaño del árbol y la exactitud pueden ser evaluados de manera cuantitativa, entre estos últimos la exactitud es considerada la más importante. El entendimiento de los árboles es difícil de ser cuantificada a medida, y la idea del visualizador es dar ayuda al entendimiento de los árboles por parte de los usuarios.

3.3 Visualización de Clasificador Bayesiano

El Clasificador Bayesiano (CB) es una técnica en DM que permite realizar clasificación probabilística basado en el teorema de Bayes y algunas hipótesis simplificadoras adicionales. Es a causa de estas simplificaciones, que se suelen resumir en la hipótesis de independencia entre las variables predictoras, que recibe el apelativo de ingenuo (Naive).

El modelo de probabilidad para un clasificador es: $p(C, F_1, \ldots F_n)$, sobre una variable dependiente C, con un

pequeño número de resultados o clases. Esta variable está condicionada por varias variables independientes desde F_1 a F_n. El problema es que si el número n de variables independientes es grande o cuando éstas pueden tomar muchos valores, entonces basar éste modelo en tablas de probabilidad se vuelve imposible. Una ventaja del CB es que utiliza un modelo bastante simple y permite que los usuarios puedan fácilmente entenderlo.

A. Visualizador de Evidencia

En [43] se describe una técnica de representación visual del modelo CB que se ha implementado con éxito en MineSet, con el nombre de Visualizador de Evidencia (VE), que despliega la estructura del CB y permite a los usuarios interactuar con ésta, examina valores específicos, muestra probabilidades de objetos escogidos, y hace preguntas del tipo what-if. Hay dos paneles en el VE, como se puede observar en la figura 8, el derecho muestra un gran gráfico circular con las probabilidades previas de los valores posibles etiquetados. Como el usuario puede interactuar eligiendo los valores para los atributos, la actualización de las rodajas muestra las probabilidades posteriores. El panel izquierdo contiene filas de gráficos circulares, uno para cada atributo, y se clasifican en orden de importancia calculada como la entropía condicional [44] de cada etiqueta del atributo seleccionado. Este panel izquierdo cambiará a una representación de barras de evidencia de una etiqueta que se selecciona.

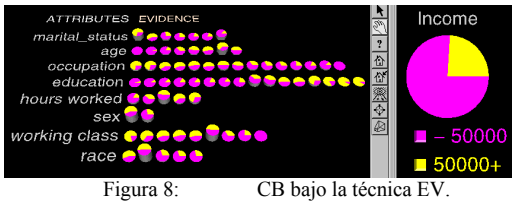


Figura 8: CB bajo la técnica EV.

Hay una gráfico de sectores o barra para cada valor discreto o rango del atributo (figura 9), donde la altura de cada circulo representa el número de casos para el valor.

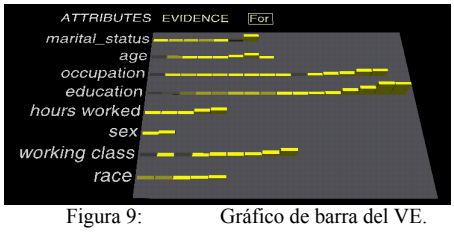


Figura 9: Gráfico de barra del VE.

En esta figura, la altura muestra la evidencia de la clase seleccionada, las barras se saturan menos cuando el número de casos disminuye, lo que implica un intervalo de confianza más amplio. Es posible un acercamiento para lograr un mejor detalle sobre los valores de atributos, y la selección de rangos específicos del valor para ver la probabilidad como se puede observar en la figura 10.

Este tipo de representación es buena para caracterizar una clase de interés, si se selecciona una clase diferente las alturas y los colores de las barras cambian para representar la nueva clase, los colores se saturan menos si el intervalo de confianza para la evidencia seleccionada es muy grande, indicando que lo estimado está basado en una pequeña cantidad de instancias. Esta visualización es interactiva y los usuarios pueden destacar una barra moviendo el cursor sobre ella.

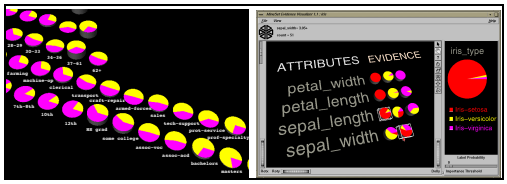


Figura 10: Acercamiento en el VE.

Cuando hay valores que agregan evidencia similar a todas las clases, las alturas de la barra serán bajas. Este método acentúa la importancia de diferencias en evidencia en comparación con los valores absolutos. El visualizador pide los valores (o rangos) dependiendo del tipo del atributo, si el atributo es ordenado (los rangos continuos son comúnmente discretizados), son presentados en el orden natural definido. Si el atributo es nominal, hay una variedad de maneras que el visualizador proporciona para pedir los valores: los valores se pueden clasificar alfabéticamente, pueden ser ordenados disminuyendo el número de instancias con los valores extremos izquierdos que tienen la altura más grande y que disminuyen a la derecha. También los valores pueden ser ordenados disminuyendo el tamaño del valor condicional de la probabilidad, para una etiqueta de valor especificado.

Para atributos nominales con muchos valores, esto genera problemas porque inundan la visualización con una gran cantidad de circulos que puedan ser difíciles de interpretar o comparar, para solucionar esto, la técnica proporciona dos mecanismos primero, ordena los valores basados en la preferencia del usuario (alfabético, cuentas, probabilidades), en segundo lugar, da la opción para remover valores basados en un pequeño porcentaje definido, típicamente 1% o menos, así elimina porciones menos significativas.

Es muy importante desplegar la confiabilidad de las estimaciones de probabilidad en la visualización. Los gráficos circulares logran desplegar esto mostrando el número de registros detrás de cada distribución como la altura; sectores más altas son más confiables. También es importante el despliegue de la confianza en el modo de las barras porque las distribuciones de probabilidad estimadas de pocos datos tienden a ser sesgadas, dando por resultado barras grandes que las cuales quedan fuera de la visualización. Esta visualización también maneja los valores nulos, y para ello dispone sectores o circulos de la distribución de probabilidad que representan valores de atributos nulos, en una columna especial. Este visualizador es muy útil en la ayuda para el descubrimiento de conocimiento y entendimiento de patrones.

3.4 Visualización de Redes Neuronales Artificiales

Las Redes Neuronales Artificiales (ANNs) han alcanzado madurez y se han establecido como herramienta útil para el tratamiento de la información, especialmente para los datos complejos, donde están limitados el conocimiento y modelos anteriores. Sin embargo, las ANNs son en sí complejas, consistiendo generalmente en muchas unidades centrales no lineales interconectadas o neuronas, qué las hacen difíciles de examinar y analizar.

Hoy, las ANNs son herramientas reconocidas para el análisis de datos, cuando los datos son complejos y el conocimiento sobre él es escaso. Contrario a los métodos estadísticos, ANNs no confía en un modelo específico para los datos que son analizados, aunque también hagan ciertas asunciones débiles. En la mayoría de los tipos de ANNs, especialmente se distribuyen los perceptrones de múltiples capas, el conocimiento: no puede localizar una parte distinta de él cuál es responsable de cierto comportamiento. En lugar, la red entera participa en acciones. Aunque mucho del mencionado atractivo de las ANNs sea una consecuencia directa de ese hecho, es también su drenaje importante detrás: es extremadamente difícil analizar el conocimiento almacenado en una red y ganar la compenetración qué está sucediendo en ella y porqué.

La alta interconexión da lugar a una alta dimensionalidad del espacio de parámetro, y compuesto con las no-lineales implicadas en cada neurona, hace el análisis matemático muy difícil. Es decir es muy duro entender una ANN, para describir su conocimiento en términos simples.

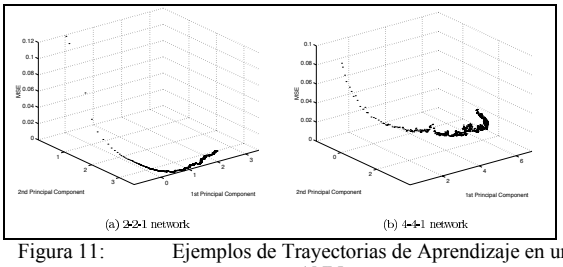
Sin embargo, el área de visualización en ANNs es una de las que ha recibido poca atención [45], la mayoría de los métodos propuestos en esta área solamente llegan a escalar a redes con pocos pesos. Por ejemplo el diagrama de Hinton, proporciona solamente la información referente a la magnitud de pesos en la red, y así revela poco sobre cómo se comporta en el entrenamiento de Turing [46].

Otros autores han proporcionado visualización de la superficie del error eligiendo dos de los pesos en la red, y trazando éstos contra la función de costo, y de tal modo revelan una porción o rebanada en 2D de la superficie multidimensional del error [47], [48], [49], pero cómo una podría elegir pares de pesos para la visualización en cualquier situación dada, esto sigue siendo problemático. Además, el análisis ha revelado que la superficie del error puede ser muy complejo, incluso para problemas simples tales como XOR, y que simplemente la visualización de rebanadas a través de la superficie del error puede proporcionar información engañosa [50].

El Análisis de Componente Principal (PCA) es una técnica estadística que se puede utilizar para trazar datos altamente dimensional sobre un espacio de una más baja dimensionalidad [51]. El PCA intenta capturar las mejores direcciones de variación de un conjunto de datos, realizando el equivalente a una rotación del espacio original de los datos. Después de que se hace esto, el PCA corresponde al valor principal más grande en la dirección de variación de los datos. La ignorancia de los componentes principales con los valores principales más pequeños corresponde a realizar una reducción de la dimensionalidad, con los datos que son proyectados con eficacia sobre un espacio dimensional más bajo. En un sentido lineal un PCA es óptimo, en que la suma de los cuadrados de las diferencias entre la distancia original entre los puntos y la distancia proyectadas es mínima. El PCA no se ha utilizado previamente como mecanismo para la reducción de la dimensionalidad en espacio del peso.

A. Visualizador de Proceso de Aprendizaje de ANNs

En [52] describe un método para visualizar el proceso de aprendizaje en una red neuronal que permite observar las características de la superficie del error a lo largo de la trayectoria tomada por la red durante el entrenamiento. Los resultados experimentales indican que éste método es aplicable a redes de un tamaño práctico, diferente a métodos anteriores de visualización. En la figura 11 se presenta un ejemplo de visualización proporcionada por este método, y que muestra el Error Medio Cuadrático (EMC) trazado a través de los primeros 2 componentes principales para una configuración de red 2-2-1.



(a) 2-2-1 network (b) 4-4-1 network

Figura 11: Ejemplos de Trayectorias de Aprendizaje en una ANN.

Para este ejemplo en particular, los primeros dos componentes principales capturan 83.00% y 14.12% de la variación en los datos respectivamente. Los valores del punto del trazado de esta manera dan una cierta idea de la velocidad de la trayectoria, es decir, eliminación rápida en EMC e inicialmente localización, progresando el pequeño movimiento hacia el final del aprendizaje.

B. Visualizador Simbrain de ANNs

En [53] se presenta un framework visual para el análisis de ANNs, llamado Simbrain, el cual apunta a la gráfica de redes neuronales, buscando una forma simple de diseñar ANNs de manera arbitraria, y con cualquier topología, considerando los tipos de neuronas mas comunes como el perceptron y las neuronas spiked, como también distintos tipos de neuronas.

La visualización de las neuronas se realiza mediante colores que representan a los tipos (nodos, o neuronas), los números dentro de las neuronas representan el estado de activación, donde esta activación depende el tipo de neurona. El número corresponde a un valor de redondeo, el valor exacto se puede obtener seleccionando la neurona. Las tonalidades cambian de intensidad mientras las neuronas se acercan a su máximo en color rojo, mientras que con azul se acercan a su nivel mínimo (figura 12).

Las líneas que conectan a las neuronas corresponden a las conexiones entre ellas, los discos pequeños de colores que están al final de estas conexiones representan a las sinapsis, el tamaño del disco es proporcional a como la neurona se acerca a su limite superior o inferior, esto simboliza que tan fuerte es la conexión de la neurona.

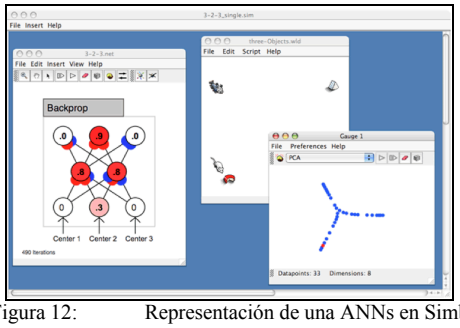


Figura 12: Representación de una ANNs en Simbrain.

El proceso de aprendizaje de una red neuronal se puede apreciar mediante el cambio en los pesos de las conexiones, y su visualización puede ser analizada inmediatamente cuando comienza el entrenamiento de la red, se pueden ingresar valores a los pesos o bien dejarlos aleatorios. Las operaciones son compartidas tanto para las neuronas como para las sinapsis.

Existen problemas también con este enfoque, dado que se pueden modificar los valores de los pesos, activaciones, etc. En el proceso de entrenamiento, las neuronas locales generalmente se ven afectadas, esto es un problema cuando se usa back-propagation ya que este necesita que los errores se calculen fuera de la red, usando los valores calculados con los inmediatamente anteriores.

Se presenta el cuadro más a la derecha de la figura 12, las visualizaciones del tipo alta dimensionalidad, donde cada punto azul presenta un punto en un espacio dimensional *n*, un vector de *n* componentes es proyectado en dos dimensiones. Los componentes de estos vectores provienen de activaciones neuronales, de las sinapsis en cualquier combinación. Esta visualización se puede ligar con una red ya diseñada, de manera que puedan trabajar en conjunto.

IV. Algunos Trabajos de Evaluación de Visualizaciones en Tareas de Minería de Datos

Se pudo constatar en este estudio, que no existen actualmente muchos trabajos que proporcionen herramientas o métodos de evaluación de técnicas de visualización, y así lograr establecer empíricamente las capacidades de cada técnica para responder preguntas del negocio y las tareas de MD.

En esta sección se discuten tres trabajos de evaluación de esquemas de visualización aplicados a distintas técnicas asociadas a tareas de MD; la compración subjetiva de Keim [58], el trabajo empírico de Hoffman [56] y la evaluación realizada por Marghescu [54].

A. *Comparación de Técnicas de Visualización para Tareas de Data Mining (Keim, 2001)*

Una de las primeras evaluaciones referenciadas es la comparación sobre distintas técnicas de visualización que realiza y describe Keim en [58] sobre proceso de MD en grandes base de datos. Esta comparación tiene como criterio basado sobre la idoneidad de la técnica para ciertas;

- Características de los datos ➪ (por ejemplo; nro. de variantes, nro. elementos, datos categóricos, etc.).
- Características de la tarea ➪ (por ejemplo; clustering, multi-variable, curva de aprendizaje, etc.).
- Características de visualización ➪ (por ejemplo; la superposición visual, curva de aprendizaje, etc.).

La tabla II expresa la opinión personal y subjetiva del autor, obtenida de la literatura y experimentación con varias de las técnicas descritas.

TABLA II. Comparación de Técnicas de Visualización para Realizar Tareas de Data Mining [58],

		cluster-ing	multi-variate hot spot	no. of variates	no. of data items	cate-gorical data	visual overlap	learning curve
Geometric Techniques	Scatterplot Matrices	++	++	+	+	-	o	++
	Landscapes	+	+	-	o	o	+	+
	Prosection Views	++	++	+	+	-	o	+
	Hyperslice	+	+	+	+	-	o	o
	Parallel Coordinates	o	++	++	-	o	--	o
Icon-based Techniques	Stick Figure	o	o	+	-	-	-	o
	Shape Coding	o	-	++	+	-	+	-
	Color Icon	o	-	++	+	-	+	-
Pixel-oriented Techniques	Query-Independent	+	+	++	++	-	++	+
	Query-Dependent	+	+	++	++	-	++	-
Hierarchical Techniques	Dimensional Stacking	+	+	o	o	++	o	o
	Worlds-within-Worlds	o	o	o	+	o	o	o
	Treemap	+	o	+	o	++	+	o
	Cone Trees	+	+	o	+	o	+	+
	InfoCube	o	o	-	-	o	o	+
Graph-based Techniques	Basic Graphs	o	o	-	+	o	o	+
	Specific Graphs	++	+	-	+	o	+	+

Donde; ++ = excelente, + = bueno, o = medio, - = pobre, y -- = muy pobre.

Muchas de las calificaciones son discutible y dependen en gran medida de los datos considerados, la tarea de exploración, la experiencia del usuario, como otros factores. Además, las implementaciones de las técnicas en sistemas reales suelen evitar los inconvenientes de una técnica única al combinarlas con otras técnicas, lo que no se refleja en las evaluaciones.

B. *Evaluación Empírica de Visualizaciones (Hoffman, 1999)*

Este autor define en [56] una cuadrícula o rejilla, que corresponde a un área de despliegue de una una visualización (el conjunto de pixeles), y a un número de medidas estadísticas para evaluar las técnicas de visualización, estas medidas consideradas son; utilización de la pantalla, el tamaño efectivo del registro de datos, la superposición, y la utilización del registro de datos que pueden caracterizar una visualización para un conjunto de datos específico. Se basa en la hipótesis de que estas medidas están muy bien relacionadas con visualizaciones que pueden detectar valores fuera de borde, cluster o patrones interesantes. Con estos elementos genera un tabla con las estadísticas básicas y principales logrando una evaluación empírica de cada una de las técnicas que considera en su análisis, asociándolas a tareas de MD. Estas

técnicas son; Radviz, Coordenadas Paralelas, Diagrama de Survey, Segmentos de Circulos, y Matriz de Diagramas de Dispersión. La tabla III muestra un ejemplo aplicado al conjunto de datos IrisFlower.

TABLA III. ESTADÍSTICAS PRINCPALES Y EVALUACIÓN DE VISUALIZACIÓN DE HOFFMAN [56]

		Iris Flower Dataset			
Grid size = 640x480	Radviz	Par. Coord.	Survey Plot	Circle Segment	Scatter Plot Matrix
Data Records	150	150	150	150	150
Dimensions Used	4	5	5	5	5
SCR. Util (SU)	0.006299	0.131543	0.263776	0.535612	0.056898
Avg. Data Rec Size (ADRS)	16	542.4067	540.2133	1103.56	400
Effective Data Rec. Size (EDRS)	12.9	269.4	540.2133	1096.933	116.5267
Space Overlap (SO)	0.160417	0.13155	0	0.006005	0.19295
Object Overlap (OO)	0.033333	0.371763	0	0	0.515733
Total Overlap (TO)	0.19375	0.503312	0	0.006005	0.708683
Data Record Utilization (DRU)	1	1	1	1	1
Avg. Obj. size	16	135.6017	54.02133	110.356	16
Marks	150	0	0	0	3750
Mark Duplicates	5	0	0	0	1934
Lines	0	600	1500	0	0
Line Duplicates	0	233	0	0	0
Polygons	0	0	0	1500	0
See Outliers	Y	Y	Y		Y
See Clusters	Y	Y	Y		Y
Find Class Clusters	Y	Y	Y		Y
See All Important Features		Y	Y	Y	
See Some Important Features		Y	Y	Y	Y
See Possible/Rule Model	Y	Y	Y		Y
See Exact Rule Model					

Concluye este autor que ninguna visualización satisface todas las necesidades de minería de datos visual y herramientas de aprendizaje de máquinas, tales somo clustering y clasificadores. Muchas de estas visualizaciones pueden ayudar a encontrar valores fuera de borde, clústeres e incluso detectar o sugerir reglas o patrones en los datos.

C. *Evaluación de Visualización en Tareas de Data Mining (Marghescu, 2008)*

Marghescu en [54], realiza una evaluación de nueve técnicas de visualización, en el contexto de un problema de análisis financiero comparativo entre distintas empresas, las técnicas evaluadas son; Gráficos de Líneas Múltiples, Matriz de Permutación, Diagrama de Estudio, Matriz de Diagrama de Dispersión, Coordenadas Paralela, Mapas de Calor, PCA, Trazado de Sammon y SOM.

En este estudio, se evalúan y comparan las técnicas de visualización con respecto a cuatro criterios establecidos:
1) Eficacia visual: la capacidad de las técnicas para responder a las preguntas y tareas de minería de datos, formuladas para el problema.
2) Información de alto/bajo nivel: la capacidad de las técnicas para mostrar ítems de datos o modelos de datos.
3) Tipo de datos: el tipo de datos usados como entrada para la técnica de visualización (datos originales o normalizados).
4) La expresividad de la técnica: la extensión para lo cual las variables son representadas gráficamente.

La tabla 4 resume la evaluación de las nueve técnicas consideradas en [16], donde se puede observar la capacidad de cada una respecto a las preguntas del negocio derivadas en tareas de minería de datos, tomando como base los cuatro criterios indicados anteriormente.

Las nueve técnicas evaluadas proporcionan una descripción general del conjunto de datos analizados, y estas visualizaciones obtenidas pueden desplegar simultáneamente todas las clases y la mayor parte de las variables de la tabla de datos.

TABLA IV. RESULTADOS DE EVALUACIÓN DE MARGHESCU.

Técnica / Tareas de Data Mining	Detección de Valores Extremos	Análisis de Dependencias (Relaciones)	Clústeres	Descripción de Clústeres	Clases	Descripción de Clases	Comparación entre Itemes
Gráfico de Línea Múltiples	X	X			X	X	X
Matriz de Permutaciones	X	X					X
Survey Plot (Diagrama de Medición)	X	X			X	X	X
Matriz de Diagrama de Dispersión	X	X					X
Coordenadas Paralelas	X	X					X
Treemap (Mapa de Árbol)	X				X	X	X
PCA (Principal Components Analysis)	X	X	X	X			X
Sammon's Mapping					X	*	
SOM (Todas las Vistas)	X	X	X	X	X	X	X

También nos dan información sobre la estructura de los datos y sus particularidades. Por ejemplo, se puede observar si hay clústeres en los datos, sin importar si las variables están correlacionadas, si hay valores extremos o fuera de borde, y si hay semejanzas y diferencias entre las clases de interés. Sin embargo, no todas las técnicas ofrecen la misma información sobre los datos, por lo que se puede confirmar que: "diferentes técnicas descubren distintos patrones en los datos".

La evaluación utiliza el método de inspección y de cuestionario, y está basada en la experiencia del autor en las técnicas de visualización consideradas y del conjunto de datos utilizado, a través de un experimento controlado con un conjunto reducido de usuarios. Una limitación de este estudio es que los usuarios evaluaron solamente visualizaciones estáticas de los datos, mientras que agregando interacción en las visualizaciones, estas podrían influenciar las opiniones de los usuarios sobre las capacidades de las técnicas.

El marco de trabajo de la evaluación de este estudio fue basado en el modelo descrito por Soukup y Davidson [55] de mapeo de un problema de negocio a las preguntas del negocio, a las tareas de minería de datos, y a las técnicas de visualización.

Esta evaluación también confirma que no hay una técnica de visualización que sea capaz, por sí sola, de contestar todas las tareas de minería de datos, y por lo tanto, sugiere que las visualizaciones múltiples o combinadas son más convenientes para ser implementadas en una herramienta de evaluación comparativa. La necesidad de usar visualizaciones múltiples debido a las limitaciones de una sola técnica, también es enfatizada por otras evaluaciones en [56] y [57], realizadas sobre conjuntos de datos artificiales o de prueba para diversos tipos de tareas.

V. CONCLUSIONES Y TRABAJO FUTURO

De la revisión realizada, se desprende que no hay una visualización que sea mejor para la exploración de datos de alta dimensionalidad en MD, algunas visualizaciones son mejores para mostrar clústeres o detectar valores fuera de borde, otras pueden mostrar dos y tres maneras de correlaciones. La mayoría de las investigaciones

recomiendan utilizar una combinación apropiada de esquemas visualización con diversas técnicas de MD, dependiendo de la tarea a realizar y las características de los datos. Es fundamental considerar en el diseño de nuevas visualizaciones, la incorporación de mecanismos de interacción para el usuario. Las visualizaciones para datos de alta dimensionalidad requiere cierto entrenamiento y experiencia del usuario.

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. Mientras que las técnicas de visualización de datos tienen algunas limitaciones inherentes para conjuntos de datos de gran volumen, tienen poderosas capacidades para el análisis de datos interactivos. Respecto a las evaluaciones, se deduce que es necesario abordar más técnicas de MD no consideradas en las evaluaciones revisadas, utilizando y estableciendo métodos subjetivos y empíricos para su medición.

Para el trabajo futuro, este estudio establece una base que permitirá proponer y desarrollar un diseño preliminar de un esquema de visualización multidimensional, combinando un conjunto de técnicas de MD, agregando mecanismos de interacción para que el usuario pueda recorrer el modelo generado en la etapa de exploración de datos de un proceso de minería de datos, permitiendo lograr un mejor entendimiento del comportamiento interno de este modelo.

REFERENCIAS

[1] B. Shneiderman and C. Plaisant. Designing the user interface. Boston: Addison-Wesley, 2005.

[2] J. Bertin, Graphics and Graphic Information Processing, Walter de Gruyer & Co., Berlin, 24-31, 1981.

[3] Chernoff, H. “The Use of Faces to Represent Points in k-Dimensional Space Graphically”, Journal of American Statistical Association, 68, 361-368.

[4] Tufte, E. R. “The Visual Display of Quantitative Information”, Graphics Press, Cheshire, Connecticut, 1983.

[5] Inselberg, A. “Parallel Coordinates: A guide for the Perplexed, in Hot Topics”. Proc. of IEEE Conference on Visualization, pp. 35-38. IEEE Computer Society, Los Alamitos, CA, 1996.

[6] Inselberg, A. “Multidimensional Detective”. Proc. of IEEE Information Visualization ’97, pp. 100-107, 1997.

[7] S. Feiner and C. Beshers. Worlds within Worlds: Metaphors for Exploring n-Dimensional Virtual Worlds. Proc. UIST ’90, pp. 76-83,1990.

[8] R. Rao, and S. K. Card. The Table Lens: Merging Graphical and Symbolic Representations in an Interactive Focus + Context Visualization for Tabular Information. Proc. CHI ’94, pp. 318-322, 1994.

[9] Keim, D. A., and Kriegel, H. P. VisDB: Database Explorations Using Multidimensional Visualization. IEEE Computer Graphics and Applications, 40-49, September 1994.

[10] Shneiderman, B. “Dynamic Queries for Visual Information Seeking”. IEEE Software, 11(6), 70-77, 1994.

[11] Spence, R., and Tweedie, L., The Attribute Explorer: Information synthesis viaexploration, Interacting with Computers, 11, 137-146, 1998.

[12] Meneses, C. “Visual Support for Enhancing the Data Mining Process”, Doctoral Dissertation, University of Massachusetts, 2002.

[13] Pak Chung Wong, P. Ch., and Bergeron R. D. "30 Years of Multidimensional Multivariate Visualization". Department of Computer Science University of New Hampshire Durham, New Hampshire 03824, USA.

[14] http://www.sas.com/technologies/analytics/datamining/miner/, S.I.I.

[15] Han, J., Kamber, M.: Data Mining Concepts and Techniques. Morgan Kaufmann (2001).

[16] http://www.sgi.com/software/mineset.html, S.)Brunk, C., Kelly, J. and Kohavi, R.: MineSet: An In- tegrated System for Data Mining. Proc. of Third In- tel: Knowledge Discovery and Data Mining (1997) 135-138.

[17] Wong, P.C., Whitney, P., Thomas, J.: Visualizing association rules for text mining. In: INFOVIS. (1999) 120–123.

[18] Lamping, J., Rao, R., and P. Pirolli. A focus+context technique based on hyperbolic geometry for visual- izing large hierarchies. In Proceedings of the ACM conference on Human Factors in Computing Systems (CHI’95), ACM Press, Denver, Colorado, USA, pages 401–408, 1995.

[19] J. D. Mackinlay, G. G. Robertson, and S. K. Card. The perspective wall: Detail and context smoothly integrated. In Proceedings of the ACM conference on Human Factors in Computing Systems (CHI’91) , ACM Press, pages 173–179, 1991.

[20] G.W. Furnas. A fisheyes follow-up: Further reflec- tions on focus+context. In Proceedings of the ACM conference on Human Factors in Computing Sys- tems (CHI’06), ACM Press, Montréal, Canada, pages 999–1008, 2006.

[21] B. L. Harrison and K. J. Vicente. An experimental evaluation of transparent menu usage. In Proceedings of the ACM conference on Human factors in comput- ing systems (CHI’96), ACM Press, New York, USA, pages 391–398, 1996.

[22] O. Couturier, J. Rouillard, and V. Chevrin. An in- teractive approach to display large sets of association rules. In Proceedings of the 12th International Conference on Human-Computer Interaction (HCI’ 07), Beijing, China, 2007.

[23] http://www.aisee.com (visitado diciembre, 2011).

[24] Thearling, K., Becker, B., DeCoste, D., (2001). Visualizing Data Mining Models. In Proceeedings of the Integration of Data Mining and Data Visualization Workshop, Springer-Verlag.

[25] O. Couturier, E. Mephu Nguifo, and B. Noiret. A formal approach to occlusion and optimization in as- sociation rules visualization. In Proceedings of Inter- national Symposium of Visual Data Mining (VDM) of IEEE 9th International Conference on Information Visualization (IV@VDM’ 05), Poster, Londres, UK, July 2005.

[26] Hofmann, H., Siebes, A.P.J.M., Wilhelm, A.F.X.: Visualizing association rules with interactive mosaic plots. In: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM Press (2000) 227–235.

[27] Ong, K.H., Ong, K. L., Ng, W. K., Lim, E. P.: Crystalclear: Active visualization of association rules. In: International Workshop on Active Mining (AM-2002), in conjunction with IEEE International Conference On Data Mining. (2002).

[28] Zhao, K., Liu, B.: Visual analysis of the behavior of discovered rules. In: Workshop Notesin ACM SIGKDD-2001 Workshop on Visual Data Mining. (2001).

[29] Inselberg, A., Dimsdale, B.: Parallel coordinates for visualizing multidimensional geometry. Computer Graphics (Proceedings of CG International) (1987) 25–44.

[30] Techapichetvanich, K., Datta, A.: Visual mining of market basket association rules. In: Computational Science and Its Applications - ICCSA 2004, International Conference, Assisi, Italy, May 14-17, 2004, Proceedings, Part IV. Volume 3046 of Lecture Notes in Computer Science., Springer (2004) 479–488.

[31] Kesaraporn T., and Datta A. Advanced Data Mining And Applications Lecture Notes in Computer Science, 2005, Volume 3584/2005, 88-95, DOI: 10.1007/11527503_12. VisAR : A New Technique for Visualizing Mined Association Rules.

[32] Pak Chung Wong, Paul Whitney, Jim Thomas.Visualizing Association Rules for Text Mining. Pacific Northwest National Laboratory, P. O. Box 999, Richland, WA 99352, USA.

[33] Olivier Couturier, Tarek Hamrouni, Sadok Ben Yahpia, Engelbert Mephu Nguifo. A scalable association rule visualization towards displaying large amounts of knowledge. 2007.

[34] O. Couturier, E. Mephu Nguifo, and B. Noiret. A formal approach to occlusion and optimization in association rules visualization. In Proceedings of International Symposium of Visual Data Mining (VDM) of IEEE 9th International Conference on Information Visualization (IV@VDM' 05), Poster, Londres, UK, July 2005.

[35] Rao, J. S., Potts, W. J. E.: Visualizing Bagged Decision Trees. Proc. of Third International Congres on Knowledge Discovery and Data Mining, AAA1 Press (1997) 243- 246.

[36] Ankerst, M., Elsen, C., Ester, M., Kriegel, H. P.: Visual Classification: An Interactive Approach to Decision Tree Construction. Proc. of Fifrh International on Knowledge Discovery and Data Mining (1999), 392- 397.

[37] Breiman, L., Friedman, J., Olshen, R. and Stone, C.: Classification and Regression Trees. Belmont, CA: Wadsworth (1984).

[38] Johnson, B., Shneiderman, B.: Treemaps: A Space-Filling Approach to the Visualization of Hierarchical Information Structures.Proc. of IEEEInfomation Visualization (1991) 275-282.

[39] Robertson, G. G., Mackinlay, J. D., Card, S. K.: Cone Trees: Animated 3D Visualization of Hierarchical Information. Proc. of the ACM Congres on Human Factors in Computing Systems (1991) 189-194].

[40] Lamping, J., Rao, R.: The Hyperbolic Browser: A Focus + Context Techniques for Visualizing Large Hierarchies. Joumal of Visual Languages and Computing, 7(1) (1997) 33-55.

[41] T., Barlow. Case study: visualization for decision tree analysis in data, 2001. Visualization of decision trees.

[42] Trong Dung Nguyen, Tu Bao Ho and Hiroshi Shimodaira. A visualization tool for interactive learning of large decision trees. Knowledge Discovery And Data Mining. Current Issues And New Applications Lecture Notes in Computer Science, 2000, Volume 1805/2000, 345-348, DOI: 10.1007/3-540-45571-X_40.

[43] Becker B., Kohavi R., Sommerfield D. Visualizing the Simple Bayesian Classifier (1997), http://pegasus.ece.utexas.edu/course/ee380l/1998fa.

[44] Cover, T. M. & Thomas, J. A. (1991), Elements of Information Theory, John Wiley & Sons.

[45] Mark W. Craven and Jude W. Shavlik. "Visualizing learning and computation an artificial neural networks", Tech. Rep. 91-5, University of Wisconsin Computer Sciences Department, 1991.

[46] J. Wejchert and G. Tesauro, "Visualizing processes in neural networks", IBM Journal of Research and Development, vol. 35, no. 1/2, pp. 244-253, 1991.

[47] Don R. Hush, John M. Salas, and Bill Horne, "Error surfaces for multi-layer perceptrons", in Proc. International Joint Conference on Neural Networks, Seattle, 1991, vol. I, pp. 759-764.

[48] Frederic Jordan and Guillaume Clement, "Using the symmetries of a multi-layered network to reduce the weight space", in International Joint Conference on Neural Networks, Seattle, 1991, vol. II, pp. 391-396.

[49] Yi Shang and Banjamin W. Wah, "Global optimization for neural network training", IEEE Computer, vol. 29, no. 3, pp. 45-54, 1996.

[50] Leonard G. C. Hamey, "The structure of neural network error surfaces", in Proc. Sixth Australian Conference on Neural Networks, Sydney, 1995, pp. 197-200.

[51] Edward J. Jackson, A User's Guide to Principal Components, Wiley, 1991.

[52] Gallagher M. IEEE Xplore - Visualization of learning in multilayer perceptron, 2003. R. Rojas, M. Verleysen, "Visualizing the learning process for neural networks", Proc. ESANN, pp. 211 - 216, 1994.

[53] Yoshimi J. "Simbrain: A visual framework for neural network analysis and education". C Merced, School of Social Sciences, Humanities, and Arts; Center for Computational Biology, P.O. Box 2039 Merced, CA 95344, USA.

[54] Marghescu D., (2008). Evaluating Multidimensional Visualization Techniques in Data Mining Tasks. ISBN 978-952-12-2152-1, ISSN 1239-1883, Painosalama Oy – Turku, Finland.

[55] Soukup, T., and Davidson, I., (2002). Visual Data Mining: Techniques and Tools for Data Visualization and Mining, John Wiley & Sons.

[56] Hoffman, P. E. "Table Visualizations: A Formal Model and Its Applications." PhD Thesis, University of Massachusetts Lowell, 1999.

[57] Keim, D. A., "Pixel-oriented visualization techniques for exploring very large databases". In Journal of Computational and Graphical Statistics, pp. 58-77, 1996.

[58] Keim, D. A., and Kriege H. P., (1996). "Visualization Techniques for Mining Large Databases: A Comparison". IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, pp. 923-938. Dec, 1996.

[59] Bederson, B. B., (2001). "Quantum Treemaps and Bubblemaps for a Zoomable Image Browser". ACM Conference on User Interface and Software Technology (UIST 2001), pp. 71-80. HCIL-2001-10 , CS-TR-4256 , UMIACS-TR-2001-39.

[60] Fayyad U. and Grinstein G. G. (2002), Introduction, in: Fayyad U., Grinstein G. G., Wierse A. (eds.), Information Visualisation in Data Mining and Knowledge Discovery, Morgan Kaufmann Publishers, San Francisco, pp. 1-17.

[61] Grinstein G. G. and Ward M. O. (2002), Introduction to Data Visualization, in: Fayyad U., Grinstein G. G., Wierse A. (eds.), Information Visualisation in Data Mining and Knowledge Discovery, Morgan Kaufmann Publishers, San Francisco, pp. 21-45.

[62] Wierse A. (2002), "What can Visualization do for Data Mining". Information Visualisation in Data Mining and Knowledge Discovery, Morgan Kaufmann Publishers, San Francisco, pp. 123-124.

[63] Hernandez, J., Ramírez, M. J., Ferri, C., (2004). Introducción a la Minería de Datos.

[64] Hsu J. (2003), "Critical and Future Trends in Data Mining: A Review of Key Data Mining Technologies/Applications", in: Wang J. (ed.), Data Mining, Opportunities and Challenges, Idea Group Inc., Palo Alto, California.

[65] Keim D. A. (2001), "Visual Exploration of Large Data Sets", Communications on ACM, Vol. 44, No. 8, pp. 38-44.

[66] Meneses, C. J., Grinstein, G. G., (2001). Visualization for Enhanching the Data Mining Process. Published in the proceedings of the Data Mining and Knowledge Discovery : Theory, Tools, and Technology III Conference, to be held at the SPIE's 15th Annual International Symposium AEROSENSE, April 16-20, 2001, Orlando, Florida, USA.