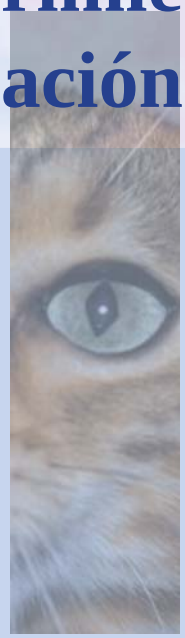
Uso adecuado de los mapas autoorganizados para el descubrimiento de información



Por Lic. Aldo L. Guerra Pulido, Especialista B en Telemática, Ing. Mario L. Basulto Núñez, Especialista Principal, Ing. Erasmo Govín Tejera, J' Grupo, Dpto. Sistema Operativo, CDNT, ETECSA, y Lic. Anabel Basulto Casas, Instructora, Moderadora de EcuRed, Palacio Central de Computación, MES.
aldo.guerra@etecsa.cu, mario.basulto@etecsa.cu, erasmo.govin@etecsa.cu, anabel.basulto@jovenclub.cu

Introducción

La humanidad siempre ha tenido especial interés por las representaciones visuales gráficas debido a la claridad espacial que brindan y al volumen de información que contienen, lo que permite que la mente procese con mayor rapidez, reconozca las características esenciales de la información y realice importantes procesos de inferencia. En la actualidad, este campo está muy relacionado con la computación gráfica.

En este marco podrían ubicarse los mapas autoorganizados, los cuales brindan excepcionales posibilidades para el descubrimiento de información no trivial. No obstante, en ocasiones el desconocimiento tácito de sus características lleva a un uso inadecuado y, en consecuencia, a la obtención de resultados incongruentes, deformados e insuficientes, mayoritariamente triviales que demeritan las verdaderas potencialidades de estas técnicas para el descubrimiento de información.

Minería de datos

El total de datos acumulados diariamente a nivel mundial es del orden de miles de terabytes, por lo que se puede inferir la cantidad almacenada en soportes informáticos por la humanidad. Estos datos pueden tener un mejor o peor destino según cómo se almacenen y/o procesen, teniendo en cuenta las disímiles tecnologías y medios de almacenamiento existentes.

En la última década, han ganado auge nuevas corrientes del pensamiento respecto al análisis de los datos y la producción de herramientas digitales en correspondencia con esas necesidades, de ahí han surgido técnicas especialmente diseñadas para el descubrimiento del conocimiento dentro de mareas de datos que aparentemente no expresan nada.

Una de estas corrientes es la minería de datos (MD) que aplica una gran variedad de metodologías y herramientas para "exprimir" y entender los datos, además de aportar resultados en casi todos los campos del saber humano.

Prestigiosas empresas en el mundo, en su afán de mejorar e incrementar sus producciones y servicios, han optado por el "riesgo" de poner sus datos en manos de especialistas en minería de datos e información, obteniendo resultados espectaculares en cuanto al descubrimiento de información no trivial implícita en la madeja de bases de datos, páginas Web, imágenes, manuscritos, secuencias de tiempo, tendencias financieras, etc. Este hecho les ha permitido reorientar, en muchos casos, el curso de los acontecimientos o viabilizar la obtención de ganancias netas millonarias al poder entender mejor los datos.

Redes neuronales artificiales

Dentro de la MD se agrupan diversos tipos de algoritmos como los modelos de redes neuronales, las que tienen entre sus virtudes la capacidad del aprendizaje y, por tanto, de mejorar su funcionamiento. A partir de los estudios realizados acerca del sistema nervioso del cerebro, el hombre ha creado modelos de redes neuronales artificiales (RNA) para resolver problemas de cierta complejidad, donde cada neurona sintoniza o aprende por sí misma a reconocer un determinado tipo de patrón de entrada. Las RNA constituyen modelos de aprendizaje y procesamiento automático. Conceptualmente se definen como un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida. En general, consisten en la simulación de las propiedades observadas en los sistemas neuronales biológicos a través de modelos matemáticos recreados mediante mecanismos artificiales, cuyo objetivo es conseguir que las máquinas den respuestas similares a las que puede dar el cerebro, que se caracterizan por su generalización y robustez.

Los modelos de redes neuronales artificiales se pueden clasificar de la siguiente forma [1]:

Figura 1 Clasificación de modelos neuronales (Fuente: [1]).

La importancia del desarrollo de la técnica de las redes neuronales artificiales radica en su comportamiento útil al aprender, reconocer y aplicar relaciones entre objetos y tramas de objetos propios del mundo real.

Mapas autoorganizados

Los mapas autoorganizados —también conocidos por las siglas SOM, del inglés *Self-Organizing Maps*— son redes neuronales no supervisadas con una técnica de proyección no lineal que muestra datos de alta dimensión en una rejilla de dos dimensiones con dos capas de neuronas, una de entrada y otra de procesamiento. Las neuronas de la primera capa se limitan a recoger y canalizar la información. La segunda capa se conecta a la primera a través de pesos sinápticos, preservando las características esenciales de los datos en forma de relaciones de vecindad.

Una red neuronal no supervisada evoluciona durante su entrenamiento. El SOM es capaz de identificar patrones en nuevos datos una vez entrenado.

Es oportuno destacar que la información que se representa ha sido sometida a un proceso de filtraje previo, donde se eliminan los ruidos, datos erróneos o puntos demasiado alejados de la media, de manera que resulte un proceso coherente.

Dentro del contexto de los procesos cognitivos del cerebro, la forma de situar al aprendizaje no supervisado es considerándolo semejante a los procesos inconscientes, en los cuales ciertas neuronas del cerebro aprenden a responder a un conjunto específico y recurrente de estímulos provenientes del medio externo, de esta forma se construyen los llamados "mapas sensoriales" en el cerebro.

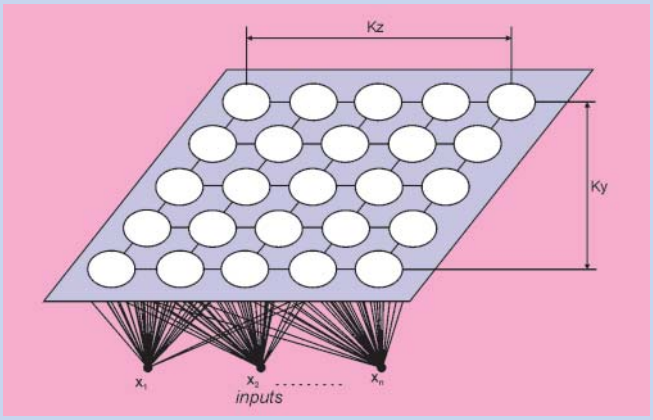


Figura 2 Representación de una red neuronal SOM de dos dimensiones (topología rectangular) (Fuente: [2]).

Cada neurona está representada por un vector prototipo, es decir, un vector de peso n-dimensional.

El proceso de aprendizaje de un mapa autoorganizado se puede dividir en tres fases:

- 1ra fase: Competitiva, en la que se determina la neurona ganadora, es decir, la mejor emparejada con el patrón de entrada según la distancia euclidiana.
- 2da fase: Cooperativa, establecida en términos de una función de vecindad entre la neurona ganadora y las demás.
- 3ra fase: Adaptativa, que actualiza los vectores sinápticos de la unidad ganadora y sus vecinas según la regla de aprendizaje.

El resultado del proceso de formación es una agrupación visual que muestra las similitudes y diferencias en los datos.

Desde la perspectiva de la segmentación, los SOM tienen varias ventajas en comparación con los métodos de optimización matemáticos y los enfoques más estadísticos. La principal es que al ser un método altamente visual, facilita la explicación de los resultados a los que toman las decisiones. De hecho, resulta más intuitivo para un público sin inclinaciones matemáticas, ya que es posible descubrir patrones inesperados por su capacidad de presentar automáticamente un mapa en el cual se puede observar una descripción de la similitud entre los datos**.**

Las técnicas SOMine presentan y combinan diferentes enfoques de la minería de datos por lo que son capaces de leer formatos de textos, de Excel, SPSS, etc., con la habilidad de procesar centenares de miles de registros con cientos de variables.

Los SOM poseen una gran cantidad de aplicaciones técnicas, por ejemplo:

- Reducción de dimensiones
- Clasificación de patrones
- Cuantificación vectorial
- Minería de datos exploratorios
- Minería de textos
- Minería Web
- Extracción de rasgos
- Tratamiento y reconocimiento de imágenes
- Interpolación de una función
- Análisis estadístico
- Clustering (Agrupamiento)
- Metáforas visuales

Por lo que tienen un amplio abanico de aplicaciones prácticas en diversas áreas:

- Perfiles de mercado y clientes
- Análisis de marketing y finanzas
- Investigaciones biológicas
- Investigaciones astronómicas
- Investigaciones clínicas y médicas
- Análisis sociales
- Inteligencia de negocios
- Procesos de optimización
- Procesos de ingeniería
- Análisis Web
- Categorización de documentos
- Control de movimientos robóticos
- Monitorización de procesos
- Metáforas geográficas

Probablemente éstas sean, junto con el Perceptrón Multicapa, las redes neuronales más usadas.

Este trabajo aborda, en particular, el empleo del sistema de software Viscovery SOMine (VSOM), desarrollado por la firma Viscovery Software GMBH. SOMine es muy fácil de usar e incluye una serie de herramientas avanzadas de procesamiento de datos y de pre-análisis, tales como la agrupación automatizada del mapa basada en el método de agrupamiento jerárquico de Ward.

Entrenamiento del modelo

Viscovery SOMine se basa en algoritmos de entrenamiento y utiliza un progresivo aumento del tamaño del mapa durante el proceso de formación, lo que hace que su implementación sea muy eficiente. A pesar de que VSOM es bastante tolerante con los datos ruidosos o perdidos, el pre-procesamiento de los datos es una parte importante de la tarea de minería de datos, encargado de la tarea de garantizar la calidad de los datos y resolver problemas como datos vacíos, erróneos o atípicos.

Para enfrentar los datos atípicos, VSOM usa la transformación sigmoide que hace énfasis en los valores de tendencia central, con la consiguiente reducción de la influencia de los valores extremos de entrada. También se usan adaptaciones de varianza para lograr que las variables sean comparables (método de Ward).

En general, el tamaño del mapa dependerá del propósito de la aplicación. Un mapa hexagonal grande es bueno para lograr una visualización más precisa del nivel de registro individual, mientras que uno pequeño es más adecuado para el agrupamiento debido a su capacidad de comprimir los datos en una cantidad menor de grupos.

Por su parte, el tamaño del mapa de nodos se selecciona como un equilibrio entre la agrupación y la visualización ya que los grupos pueden no ser muy homogéneos lo que posibilita juzgar con precisión las diferencias intra-agrupación.

Este sistema requiere pocos parámetros de configuración. En un inicio, se puede ajustar la tensión pues esencialmente es un valor para el radio de la vecindad en la etapa final del entrenamiento, donde los resultados de tensión baja ofrecen gran detalle local (precisión), mientras que los valores altos de tensión tiene un efecto promedio (suavizado) sobre el mapa. El valor de tensión típico es 0,5 y, normalmente, se usa la función de vecindad gaussiana. En la agrupación de dos etapas, las neuronas en el mapa basan el agrupado en sus distancias euclidianas, utilizando un algoritmo de agrupamiento adecuado. En muchos casos, se utiliza el método jerárquico de Ward, incluido en el software, para identificar los grupos en el mapa final.

No obstante, en ocasiones, la minería de datos, por su novedad, tiene en contra el uso inadecuado de sus técnicas y posibilidades asociado al desconocimiento acerca del tema y a la ausencia de las metodologías apropiadas para su implementación.

Uso inadecuado de la herramienta VSOM

A veces, a causa de métodos inadecuados pueden obtenerse mapas autoorganizados que, en realidad, no aportan conocimiento alguno. Estos mapas suelen crearse mediante algún procedimiento que ha sido ejecutado en forma de trucaje, lo cual desvirtúa la esencia de la herramienta VSOM. Tales manejos traen como consecuencia su empleo en cuestiones triviales y en la generación de resultados pobres en calidad, lo que provoca la falta de interés e incomprensión por parte de los que toman decisiones, el rechazo de los escépticos y, en general, la subestimación de las herramientas SOM.

A continuación se representa un mal procedimiento de ordenamiento y presentación de los datos, donde se evidencia que no era necesario el uso de VSOM.

La siguiente tabla presenta, a modo de ejemplo, un pequeño conjunto de datos con valores numéricos arbitrarios:

Pinar del Río	20 000
Mayabeque	15 000
Artemisa	17 000
La Habana	200 000
Matanzas	24 000
Cienfuegos	18 000
Villa Clara	40 000
Sancti Spíritus	53 000
Ciego de Avila	27 000
Camagüey	21 000
Las Tunas	14 000
Holguín	15 000
Granma	39 000
Santiago de Cuba	160 000
Guantánamo	13 000
Isla de la Juventud	5000

Tabla 1 Ejemplo con los valores numéricos desorganizados (Fuente: elaboración propia).

Como se observa en la tabla 1, hay un ordenamiento geográfico por provincias en el sentido de oeste a este de Cuba, mientras que el municipio especial Isla de la Juventud aparece al final.

Estos datos pudieron haberse representado mediante un gráfico de columnas donde se desplegaran las provincias en el eje X, lo que sería comprensible para cualquier lector. Sin embargo, se utilizaron para generar un mapa autoorganizado usando Viscovery SOMine que, independientemente de su variedad de colores, aparece ahora con un ordenamiento por valores de acuerdo a cierta escala, sin que esto ofrezca claridad alguna sobre la información contenida en los datos.

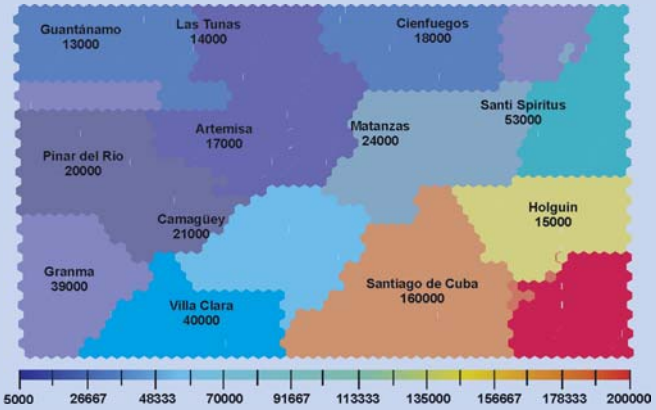


Figura 3. Ejemplo de mapa autoorganizado (Fuente: elaboración propia).

En la figura 3 se hace muy difícil descubrir conocimiento, pues se trata de una representación inadecuada a partir de datos que habían sido etiquetados de antemano, donde no se revelan patrones interesantes ni se aprecian relaciones que salten a la vista; además, sólo se manejan 16 registros. Como se expondrá más adelante, en este caso la estadística y minería de datos de VSOM han sido notablemente subutilizadas.

El proceso en cuestión arrojó un mapa temático (Figura 4), donde los datos relacionados con el color fueron tomados manualmente: se llevó la imagen del mapa autoorganizado hasta un software de captura que brinda el RGB de cada color; luego, se incorporaron a un sistema de gestión de mapas temáticos, en específico el GIS Mapinfo; y, posteriormente, fueron insertados en un mapa de Cuba. Por último, se insertó en la parte inferior del mapa una copia de la escala de colores de la figura 3.

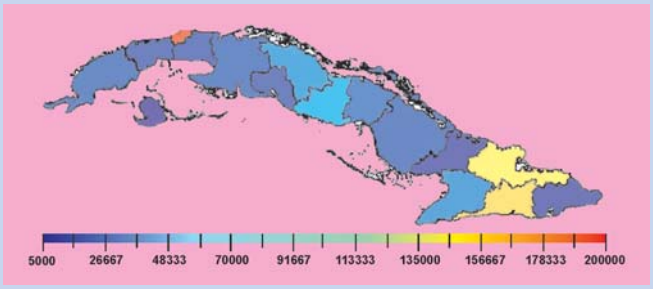


Figura 4 Un mapa temático de Cuba (Fuente: elaboración propia).

Los mapas temáticos están basados en mapas topográficos que representan cualquier fenómeno geográfico de la superficie terrestre. Hacen referencia a la representación de ciertas características de distribución, relación, densidad o regionalización de objetos reales —vegetación, suelos, geología, etc.— o de conceptos abstractos —indicadores de violencia,

desarrollo económico, calidad de vida, etc.—, pero, evidentemente, no son producto de la aplicación de la minería de datos.

Las herramientas de MD del tipo SOMine no deben utilizarse para representar mapas temáticos de forma directa, sería un error de concepto y una aguda subutilización de recursos tan potentes. Estos casos siempre contienen problemas mal planteados.

No obstante, sería factible que, después de hacer minería con estas herramientas, los resultados sean representados mediante Sistemas de Información Geográfica, lo que implicaría un correcto procedimiento y el consecuente esclarecimiento del proceso realizado.

En el ejemplo expuesto anteriormente, una solución más simple, lógica y esclarecedora podría ser ordenar los valores de mayor a menor y "semaforizarlos" según cierto criterio de límites, lo cual resulta mucho más visual y comprensible (Tabla 2).

La Habana	200 000
Santiago de Cuba	160 000
Holguín	150 000
Sancti Spíritus	53 000
Villa Clara	40 000
Granma	39 000
Ciego de Avila	27 000
Matanzas	24 000
Camagüey	21 000
Pinar del Río	20 000
Cienfuegos	18 000
Artemisa	17 000
Mayabeque	15 000
Las Tunas	14 000
Guantánamo	13 000
Isla de la Juventud	5000

Tabla 2 Ejemplo con los valores numéricos organizados en orden descendentemente y "semaforizados" (Fuente: elaboración propia).

Uso adecuado de la herramienta VSOM

A continuación se propone un ejemplo donde se utilizan los datos demográficos de un grupo de empresas —factor de riesgo, edad, solvencia, volumen de negocios, cambio en % del volumen de negocio, balance general y ROE (*Return on Equity)*— para crear un mapa. Los resultados de la segmentación demográfica se emparejaron con la información de ventas para cada producto [3].

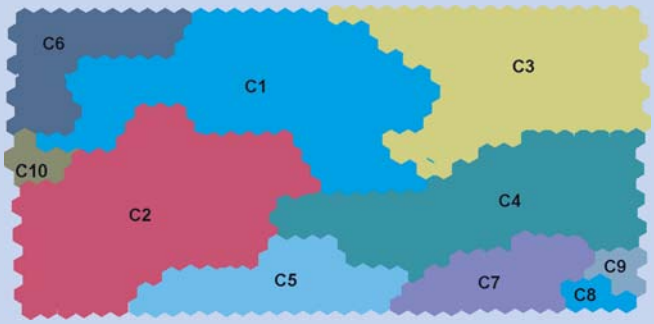


Figura 5 Planos componentes del mapa (Fuente: [3]).

La figura 5 muestra el mapa final. La agrupación está compuesta por diez grupos de diversos tamaños, de C1 a C10. El color de la agrupación sólo significa la pertenencia al clúster, sin implicar ningún valor. El mapa fue creado de acuerdo a las variables de las categorías descriptivas. Con el objetivo de interpretar el mapa y, en particular, de las características de cada grupo deben utilizarse los planos de los componentes de la figura 6. Los planos componentes muestran la distribución de los valores de las variables en el mapa. Los valores de cada variable se representan por el color de la neurona: los colores cálidos —rojo, naranja y amarillo— muestran valores altos y los colores fríos —azul— muestran los valores bajos. Los valores aproximados se indican mediante la escala en cada plano componente. El mapa es interpretado por la lectura de los planos de los componentes para cada grupo. Por ejemplo, el Grupo 6 muestra valores de medio a altos en la solvencia y el ROE, y valores bajos en la edad, el volumen de negocios y el total del balance. El Grupo 6 también muestra diversos factores de riesgo en un rango de bajo a alto, por lo que puede concluirse que son pequeñas empresas relativamente jóvenes, aunque muy rentables. Asimismo, puede verse que el factor de riesgo es elevado en los segmentos C2 y C5, lo que significa que estos contienen las compañías menos fiables. Las empresas más antiguas se encuentran en los segmentos C4, C5 y C7.

La figura 6 muestra los datos demográficos por separado de las empresas en estudio.

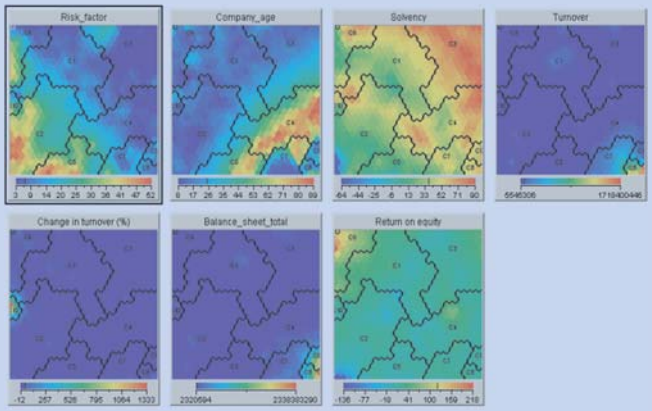


Figura 6 Planos componentes del mapa (Fuente: [3]).

Análisis de los resultados

Los resultados de la segmentación de clientes se resumen en la tabla 3, donde aparecen las dimensiones y las características distintivas de cada uno de los conglomerados. Las agrupaciones C1, C2 y C3 son las más grandes.

Los grupos identificados son los siguientes:

- **Grupo 1**: El factor de riesgo, la edad, el volumen de negocios y el balance general son bajos, mientras que la solvencia oscila de medio a alto. El rendimiento sobre el capital es bueno en promedio. Este grupo también incluye a los clientes

con menor rentabilidad sobre los recursos propios, así como los de las empresas con menor solvencia.

- **Grupo 2**: Presenta un factor de riesgo mucho mayor, menor solvencia y menor rentabilidad sobre los recursos propios que el Grupo 1.
- **Grupo 3**: Similar al Grupo 1, a excepción de una solvencia mucho mayor. La edad media de la compañía es algo mayor, aunque el factor de riesgo parece ser similar al del Grupo 1. Es uno de los tres grupos más grandes en términos de número de clientes.
- **Grupo 4**: Contiene las empresas más antiguas del conjunto de datos y, en general, presenta una elevada solvencia y buena rentabilidad.
- **Grupo 5**: Es un conjunto de tamaño medio de mayor alcance que las empresas promedio. La solvencia es buena y las empresas son relativamente nuevas. La rentabilidad es media.
- **Grupo 6**: Es un grupo de tamaño medio que incluye las empresas más rentables del conjunto de datos. Representa las pequeñas empresas en crecimiento. Casi la mitad de la agrupación muestra un factor de riesgo muy alto, pero algunas empresas son también muy solventes.
- **Grupo 7**: El grupo es de tamaño medio y abarca las empresas suficientemente grandes en términos de volumen de negocio y activos totales. La solvencia es buena en promedio. El cúmulo comprende una mezcla de viejas y nuevas empresas.
- **Grupo 8**: Es uno de los tres grupos de pequeñas empresas y contiene las mayores empresas en términos de activos y volumen de negocios. Las empresas son solventes y rentables de manera justa y su factor de riesgo es muy bajo. La antigüedad es superior a la media.
- **Grupo 9**: Es otro pequeño grupo de grandes empresas. Difiere del Grupo 8 en que las empresas son más recientes y la rotación es menor.
- **Grupo 10**: Es el grupo más pequeño y último identificado. Contiene compañías en crecimiento que son bastante rentables y solventes, y tiene un factor de riesgo muy bajo.

Clusters	Clientes	%	Clientes distintivo(s)
Cluster1	1,659	20,68	Atributo no especificado
Cluster2	1,689	21,05	Alto facto de riesgo
Cluster3	1,699	21,17	La más alta solvencia
Cluster4	1,139	14,19	Las compañía más antiguas y de gran solvencia
Cluster5	640	7,98	Las grandes empresas de alta solvencia
Cluster6	565	7,04	Factor de alto riesgo, buena solvencia con la más alta solvencia
Cluster7	398	4,96	Empresas tanto viejas como jóvenes, de buena rotación, mayor volumen de negocios
Cluster8	91	1,13	De gran balance
Cluster9	80	1,00	Gran total del balance
Cluster10	64	0,80	El más grande de cambio en el volumen de negocios(%)

Tabla 3 Dimensiones y características distintivas de cada uno de los conglomerados (Fuente: [3]).

Una vez identificados los grupos, el próximo paso fue comparar la información de ventas para cada categoría de productos en los segmentos creados. Si la cantidad de trabajo empleado en las ventas para cada uno de los segmentos es la misma, la división de los segmentos de clientes puede extenderse para describir a los clientes rentables, a los promedio y a lo que sean sin fines lucrativos.

Conclusiones

Los modelos neuronales ofrecen un nuevo enfoque de los datos y una nueva forma de representar la información multidimensional y compleja a fin de que el ser humano pueda entender las relaciones subyacentes, permitiéndole descubrir conocimiento no observable a simple vista.

Mediante el empleo de la minería de datos, en particular de la herramienta VSOMine, se contrastó un caso distorsionado típico del uso inadecuado de los mapas autoorganizados que en ciertas ocasiones se ha utilizado, en oposición a un ejemplo bien concebido.

El análisis permitió confirmar lo fructífero que resulta el uso adecuado de los conceptos, las metodologías y el conocimiento acerca de los mapas autoorganizados. De esta manera, se contribuye a mejorar considerablemente su aceptación y uso como herramienta de descubrimiento de información y apoyo a la toma de decisiones en las empresas y los centros de investigación.

Referencias bibliográficas

[1] Arias S., Francisco Javier. "Redes Neuronales Artificiales". 2009. http://es.scribd.com/doc/78811085/Map-Ask-Oh-on-En (acceso octubre 15, 2012).

[2] Olga Kurasova, Pavel Stefanovi. "Visual analysis of self-organizing maps". *Nonlinear Analysis: Modelling and Control,* vol. 16, no. 4 (diciembre, 2011): 488–504.

[3] Holmbom, Annika H.; Eklund, Tomas; and Back, Barbro, "Customer Portfolio Analysis Using the SOM". *ACIS 2008 Proceedings.* 19th Australasian Conference on Information Systems, Finlandia, 2008. http://aisel.aisnet.org/acis2008/5 (acceso octubre 15, 2012).