# Métricas para el apoyo de la exploración visual de componentes en modelos de minería de datos

*Metrics for the support of visual exploration of components in data mining models*

Fernando Medina-Quispe[1*] Wilson Castillo-Rojas[2] Claudio Meneses Villegas[3]

Recibido 15 de julio de 2020, aceptado 1 de septiembre de 2020
*Received: July 15, 2020 Accepted: September 1, 2020*

## RESUMEN

La exploración de un modelo de Minería de Datos (MD), mediante el uso de técnicas de representación visual apropiadas y mecanismos de interacción integrados, presentan ventajas para el analista o minero de datos al momento de intentar comprender un modelo datos. Actualmente, existen nuevas propuestas de metodologías y esquemas de visualización para apoyar procesos de MD, que integran características que combinan técnicas de MD y artefactos gráficos ad-hoc con el objetivo de facilitar el análisis y exploración de modelos, mediante el uso de visualización en la entrada (análisis exploratorio de datos) del proceso de MD, luego en el proceso de generación del modelo (visualización y exploración del modelo y de sus componentes internas), y finalmente en la salida de este proceso (visualización de patrones). Sin embargo, esto apunta a un análisis cualitativo y muchas veces subjetivo, que depende directamente de la experiencia y experticia del analista o minero de datos. Para poder complementar este análisis cualitativo, es necesario incorporar en el esquema visual, funciones con métricas que permitan corroborar cuantitativamente. Este trabajo se orienta en esta dirección, y describe la definición, adaptación, e implementación de un conjunto de métricas que permiten validar y complementar el análisis visual de un modelo de MD, mediante el uso de métricas de distancia y similitud, aplicadas sobre las componentes del modelo de MD. Este trabajo utiliza como caso de estudio, un modelo de MD generado a través de la técnica Árbol de Decisión (ÁD), combinada con la técnica mapas de Kohonen o Self-Organizing Map (SOM) aplicada sobre las componentes o nodos del ÁD. Se logra comprobar la validez de las métricas propuestas a partir de su aplicación, sobre un conjunto de datos conocido a partir de una tarea de MD previamente definida.

Palabras clave: Minería de datos, minería de datos visual, visualización de modelos de minería de datos, métricas para clústeres.

## *ABSTRACT*

*The exploration of a Data Mining (DM) model, through the use of appropriate visual representation techniques and integrated interaction mechanisms, present advantages for the analyst or data miner when attempting to understand a data model. Currently, there are new proposals for methodologies and visualization schemes to support DM processes, which integrate features that combine DM techniques and ad-hoc graphic artifacts in order to facilitate the analysis and exploration of models, through the use*

[1] Universidad Arturo Prat. Facultad de Ingeniería y Arquitectura. Iquique, Chile. E-mail: femedina@unap.cl
[2] Universidad de Atacama. Departamento de Ingeniería Informática y Ciencias de la Computación. Copiapó, Chile. E-mail: wilson.castillo@uda.cl
[3] Universidad Católica del Norte. Departamento de Ingeniería de Sistemas y Computación. Antofagasta, Chile. E-mail: cmeneses@ucn.cl
* Autor de correspondencia: femedina@unap.cl

*of visualization in the input (exploratory data analysis) of the DM process, then in the model generation process (visualization and exploration of the model and its internal components), and finally in the output of this process (pattern visualization). However, this points to a qualitative and often subjective analysis, which depends directly on the experience and expertise of the analyst or data miner. In order to be able to complement this qualitative analysis, it is necessary to incorporate functions with metrics in the visual scheme that allow to corroborate it quantitatively. This work is oriented in this direction, and describes the definition, adaptation and implementation of a set of metrics that allow to validate and complement the visual analysis of an DM model, by using distance and similarity metrics, applied to the components of the MD model. This work uses as a case study, an DM model generated through the Decision Tree (DT) technique, combined with the Kohonen maps technique or Self-Organizing Map (SOM), applied to the components or nodes of the DT. It is possible to check the validity of the proposed metrics from their application on a known data set from a previously defined DM task.*

*Keywords: Data mining, visual data mining, visualization of data mining models, metrics for clusters.*

## INTRODUCCIÓN

Uno de los principales desafíos durante el desarrollo de un proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD) es la interpretación y entendimiento de los patrones obtenidos a partir del modelo de Minería de Datos (MD) resultante de la etapa de MD. Por lo general, un proceso de KDD consiste en 5 etapas: selección de datos, preparación o pre-procesamiento de los datos, análisis exploratorio, MD, interpretación de patrones, y evaluación de modelos. De las últimas dos etapas, depende la obtención del modelo de MD final, que en la mayoría de las veces, requiere sucesivas iteraciones sobre el análisis exploratorio de los datos de entrada, y obtención de patrones e interpretación de resultados [1]. Algunos autores, señalan que el uso de la visualización debiera estar incorporada en todo el desarrollo del proceso de KDD, como herramienta para el soporte de las interacciones entre usuarios finales y analistas de datos [1-5]. Sin embargo, pocas propuestas de procesos de KDD incorporan y discuten el rol de la visualización dentro de este proceso [5].

La Visualización es generalmente utilizada para obtener un entendimiento preliminar de los datos al inicio del proceso de KDD, y con esto se logra refinar los objetivos y tareas definidas inicialmente en la fase de formulación del problema. Se han propuesto diversas técnicas para visualizar conjuntos de datos masivos, Keim [8] provee una taxonomía de técnicas visuales para explorar conjuntos de datos masivos, mientras que Hoffman [9, 24] esquematiza una categorización de técnicas visuales para realizar tareas de MD. Mientras que la visualización de modelos de MD, soporta el entendimiento e interacción con el modelo inducido desde un conjunto de datos de entrenamiento por una técnica de MD [10]. Estas visualizaciones deben proveer una forma natural de entender la estructura, componentes, y complejidad del modelo, así como también, las representaciones visuales proveen una forma directa de comparar varios modelos, y permiten el uso de la percepción visual humana para formular hipótesis y conclusiones acerca del modelo y su correlación con los datos [7]. Otras investigaciones, sostienen que la utilización de técnicas de visualización durante la etapa de construcción del modelo, puede mejorar la comprensión del modelo [6], y han abordado, por ejemplo, la formulación de esquemas de visualización aumentada sobre los componentes de modelos de MD, en la cual integran técnicas de visualización y mecanismos interactivos ad-hoc [1, 24].

No obstante, el análisis de modelos de MD a través del uso de visualizaciones en todo el proceso, desde su entrada (análisis visual preliminar de datos), generación del modelo de MD (análisis visual del modelo), y salida (visualización de patrones), corresponde a un análisis cualitativo y subjetivo, y su éxito está directamente ligado a la certera interpretación del analista o minero de datos o su capacidad de percepción visual, lo cual muchas veces depende de su nivel de experiencia y experticia, y al acompañamiento apropiado del usuario final experto en el dominio de los datos [5]. Para complementar este análisis cualitativo, es necesario incorporar elementos que permitan corroborar el análisis visual a través de mediciones

o métricas que calculen el nivel de similitud o diferenciación de las componentes del modelo de MD, por ejemplo, el nivel de homogeneidad, compactación o similitud de clústeres en un modelo de MD generado a través de esta técnica.

El presente trabajo propone y describe un conjunto de métricas que permiten, por un lado, caracterizar las componentes de un modelo de MD, y por otro lado, comparar las componentes del modelo mediante métricas que se basen en la "distancia" entre componentes, por ejemplo: distancia media intra-componente, distancia mínima inter-componente, etc. Para esto, se consideran métricas que se usan generalmente en métodos de clustering para determinar la similitud entre dos objetos, si se ve cada componente como un grupo de instancias asociadas a reglas generadas por la técnica. Por ejemplo: homogeneidad o dispersión asociada a cada componente.

Como caso de estudio, se presenta la definición y adaptación de las métricas de distancia y similitud, las que son implementadas en un software prototipo, aplicándolas sobre componentes de un modelo de MD generado con la técnica ÁD, la que a su vez, es combinada con la técnica SOM, aplicada en cada una de las componentes o nodos del ÁD. Finalmente, se presenta un análisis de los resultados obtenidos a partir de una evaluación subjetiva de las métricas propuestas a partir de la aplicación, a través de la implementación de las métricas definidas y adaptadas para el caso de estudio. Esto, permite comparar las percepciones de los analistas de datos, con los valores obtenidos desde las métricas respecto a la similitud o distancia de las componentes del modelo ÁD. Los resultados, permiten correlacionar y validar apropiadamente los resultados del análisis visual.

## VISUALIZACIÓN DE ÁRBOLES DE DECISIÓN

La visualización está siendo incorporada cada vez más en el proceso de MD como herramienta para soportar interacciones entre usuarios/analistas de datos y las componentes involucradas en el proceso de MD. Sin embargo, ningún modelo existente (Fayyad [1]; Adriaans & Zantinge [2]; Brachman [3]; John [4], incorporan explícitamente y discuten el rol de visualización en el ciclo de MD. Meneses y Grinstein [5] proponen elementos para avanzar hacia un esquema de proceso de MD con soporte para visualizar cuatro tipos de entidades: datos, parámetros de algoritmos, modelos inducidos, y patrones.

Los ÁD corresponden a uno de los modelos predictivos más utilizados en MD. Estos se crean dividiendo el conjunto de datos ordenados por algún valor que los jerarquiza, en subconjuntos, y después dividiendo cada uno de los subconjuntos, hasta que ya no se puedan dividir. En sintonía con la metáfora del árbol, el conjunto de datos original es el nodo raíz, los subconjuntos son nodos, y los subconjuntos no-particionados son hojas, y estas hojas son definidas por la secuencia de reglas de partición. Las ramas de un nodo son subconjuntos creados a partir de un nodo. Una regla de división utiliza típicamente una 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 de características de cosas/sujetos en grupos pequeños y homogéneos. La investigación sobre visualización de ÁD ha recibido gran atención actualmente, debido a su importancia en la práctica, y se han realizado muchos trabajos respecto a visualizadores 3D en MineSet [13], CATScan [13] para inducir ÁD "recubiertos", visualización interactiva en la construcción de ÁD [14], visualizador en un mapa de árboles en el sistema CART [15] y Salford.

Se pueden presentar dos problemas con la visualización de ÁD, y que no han sido investigados de manera profunda; La integración de visualización de ÁD en el proceso de MD, y la visualización de ÁD de gran tamaño. Existen nuevos enfoques en el campo de la visualización para representar estructuras de gran tamaño y jerarquía como los: treemap [16], cone-trees [17], árbol hiperbólico [18]. Un trabajo reciente en el desarrollo de esquemas de visualización de ÁD es el que presenta la implementación de la herramienta llamada EMTree Results Viewer [19]. Está diseñado para ayudar a construir y entender ÁD complejos visualizando la partición de casos, que ayuda al analista de datos a comprender las predicciones de un modelo, y visualizando los diagnósticos del

modelo, que ayuda a determinar la confiabilidad del modelo. Se deduce de este trabajo, que el proceso de análisis de un ÁD implica la conmutación frecuente entre múltiples visualizaciones, ya que ofrecen una variedad de formas de visualizar información relevante, que permite interactuar con la presentación de la información apoyando la comprensión de las relaciones entre las vistas.

Otro trabajo revisado sobre ÁD, 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) [20]. 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 que la estructura de ÁD contiene dos tipos de información: información estructural del árbol, e información contenida en cada nodo. Luego se establece un conjunto de requerimientos para lograr los dos objetivos descritos anteriormente, y estos son:

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

Este visualizador se implementa para el sistema CABRO y despliega un ÁD en dos vistas relacionadas, la global y otra enfocada al detalle. 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 vista ojos de pez (Fish-Eyes View - FEV): permite focalizar la visualización de manera de manejar el espacio para 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.

Con árboles muy grandes el usuario puede encontrar difícil navegar sobre ellos, incluso con la vista pequeña y uso de FEV. Se ha utilizado la técnica llamada T2.5D para resolver este problema. Para explicar esta técnica, se necesita definir el concepto "camino ancho" de un nodo del árbol. Esto se entiende como el conjunto de nodos del camino desde la raíz al nodo y sus hermanos, es decir, contiene a sus familiares directos. Cuando se visualiza un árbol en cada movimiento el camino ancho está activo (active wide path - AWP). En aplicaciones de esta técnica se describen algunos resultados con ÁD grandes:

- T2.5D logra fácilmente manejar ÁD con más de 20000 nodos, y más 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.

Existen varios métodos para resolver los tres problemas de inducción de ÁD en la selección de atributos, poda y discretización. 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 dando como resultado la Tabla 1. Los tres principales criterios

Tabla 1. Comparación de visualizadores de ÁD.

| Criterios | 3D en Mineset [13] | 2D en CART [15] | Árboles Hiperbólicos [18] | T2.5D [20] |
|---|---|---|---|---|
| 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 |

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.

## DEFINICIÓN Y ADAPTACIÓN DE MÉTRICAS

Como se señala anteriormente, en este trabajo se consideran métricas que se utilizan generalmente en los métodos de clustering, las cuales permiten determinar la similitud entre dos instancias, si se ve cada componente de un modelo de MD, como un grupo de instancias asociadas a reglas generadas por la técnica ÁD como, por ejemplo, homogeneidad o dispersión asociada a cada componente. El objetivo de las métricas es tener una forma de comparar cuantitativamente dos componentes de un modelo basado en la información visual. Para esto se definen a continuación, dos tipos de métricas que permitan proveer al analista de datos indicadores calculados en forma automática. La forma de esta comparación cuantitativa se basa en términos de: Medidas de Distancia, y Medidas de Similitud.

Estas métricas se han seleccionado utilizando como criterio el alcance del trabajo de investigación, que considera como caso de estudio un modelo [25] de MD generado con la técnica ÁD. Luego, una vez generado el modelo de MD con esta técnica, es combinada con la técnica SOM aplicada a todo el modelo y por tanto a cada una de sus componentes o nodos. Esta combinación, algunos autores la llaman MD Híbrida [21].

El uso de la técnica SOM aplicado a un ÁD como aumentador visual, se debe a que SOM cumple un doble propósito: partición espacial del subconjunto de datos asociados a cada nodo del árbol, y visualización de esta partición mediante un mapa. Considerando que el árbol por sí sólo, permite establecer reglas de decisión distribuyendo los datos a través de jerarquías representadas por sus nodos, y en cada nodo recopila las instancias que cumplen con estas reglas, sin embargo, no permite visualizar la distribución espacial de las instancias, lo que sí provee SOM. Además, la técnica SOM es compatible al dominio de los datos que maneja el ÁD, y permite describir su distribución en cada nodo. Lo anterior, permite una comparación entre nodos y así poder determinar aquellos con similar distribución o gran diferenciación, a través de la especificación de la "distancia" entre ellos, el número de instancias ubicadas sobre o bajo el centroide de una cuadrícula del mapa.

Se considera calcular las métricas distancia y similitud para dos componentes seleccionadas del modelo de MD generado por la técnica ÁD. Para esto, en primer lugar, se obtiene para cada componente (nodos o reglas) sus respectivos mapas de SOM de la capa de salida, y a cada neurona de estos mapas se le aplican las métricas mencionadas. La definición de estas métricas permite darle soporte a la evaluación que el análisis visual requiere para medir si apoya en la mejora de la comprensión y credibilidad del modelo de MD.

Sin embargo, se debe destacar que las fórmulas de estas métricas han sido adaptadas o extendidas para el cálculo de instancias representadas en matrices, ya que originalmente, vienen expresadas para comparar instancias representadas en vectores unidimensionales. Las matrices, permiten representar todas las instancias de un ÁD alojadas en una componente o nodo del árbol, y que una vez aplicada la técnica SOM sobre el modelo y sus componentes, se tiene como resultado una matriz de neuronas, donde en cada celda (neurona) de la matriz representa a cada componente o nodo del ÁD.

**Métrica de Distancia para Mapas SOM**

Esta métrica representa una adaptación de la medida de distancia entre vectores de datos. Es útil denotar la distancia entre dos instancias $X_i$ y $X_j$ como d($X_i$ y $X_j$). Una medida de distancia válida debe ser simétrica y su valor mínimo es cero dada la condición que para dos instancias comparables resultan iguales si $x = y$. Dada la naturaleza del dominio de los datos se utiliza la métrica de Minkowski[11] como medida de distancia.

La expresión general de distancia de Minkowski entre dos instancias dadas de $p$ dimensiones $x_i = (x_{i1}, x_{i2}, \ldots x_{ip})$ y $x_j = (x_{j1}, x_{j2}, \ldots x_{jp})$ es:

$$d(x_i, x_j) = \left( \left| x_{i1} - x_{j1} \right|^g + \left| x_{i2} - x_{j2} \right|^g + \cdots + \left| x_{ip} - x_{jp} \right|^g \right) \frac{1}{g} \quad (1)$$

Dependiendo del valor particular de $g$ utilizado se pueden obtener distintas métricas de distancia (por ejemplo: Euclides con $g = 2$, Manhattan con $g = 1$, y Chebychev con $g = \infty$). Una modificación a este cálculo considera la asignación del peso de cada atributo en la instancia de acuerdo con su importancia. Entonces la distancia ponderada por el peso de cada atributo queda como:

$$d(x_i, x_j) = \left( w_1 \left| x_{i1} - x_{j1} \right|^g + w_2 \left| x_{i2} - x_{j2} \right|^g + \cdots + w_p \left| x_{ip} - x_{jp} \right|^g \right) \frac{1}{g} \quad (2)$$

Donde: $w_i$ e [0, ∞]

Para el caso de un mapa de Kohonen (SOM), la distancia puede ser medida en términos de los pesos de la red entrenada y en términos de la capa de salida (mapa). Para esto se puede medir la distancia de los vectores de pesos de dos mapas de entrada representados por los vectores *X* e *Y* respectivamente. Similarmente, se puede comparar los mapas de salida de dos redes de Kohonen asociadas a distintas componentes de un modelo de MD, considerando que en este caso el vector *X* representa los valores del mapa 1 y el vector *Y* representa los valores del mapa 2, una vez entrenada la red mediante el algoritmo SOM.

Es necesario considerar una instancia representativa de cada neurona, ya que el cálculo de distancia de Minkowski se refiere a instancias y no a un conjunto de ellas como se agrupan en cada neurona de un mapa SOM. Para esto, se utilizan como criterios: para el caso de atributos con valores numéricos el promedio aritmético de los valores de las instancias para cada neurona de ambos mapas ($x_v$ *e* $Y_v$), y para atributos con valores no-numéricos se utiliza el valor que más se repite es decir la moda. Para el caso de los pesos ($w$) de los atributos se utiliza un arreglo $w$ aleatorio que posteriormente es mejorado mediante el cálculo de *Best Matching Units (BMU)* [22].

Por tanto, la adaptación de la métrica de distancia de Minkowski considerando el cálculo de Euclides ($g = 2$), queda de la siguiente forma:

Dados dos vectores *Xv* e *Yv* de ($n \bullet m$) dimensiones (o neuronas de las capas de salida de dos redes de Kohonen), en que cada neurona $x_i \in Xv$ con $t$ instancias y $p$ atributos, y cada neurona $y_i \in Xv$ con $k$ instancias y $p$ atributos, la expresión general de la distancia entre dos neuronas $Xv_i$ e $Yv_i$ es:

$$d(X_{vi}, Y_{vi}) = \frac{\sum_{i=1}^{t} \left( \sum_{j=1}^{k} \left( w_1 \left| x_{i1} - y_{j1} \right|^2 + w_2 \left| x_{i2} - y_{j2} \right|^2 + \cdots + w_p \left| x_{ip} - y_{jp} \right|^2 \right)^{\frac{1}{2}} \right)}{t \cdot k} \quad (3)$$

**Alcances y situaciones de borde:**

i. Este cálculo se debe replicar para las $(n \bullet m)$ pares neuronas de cada mapa de salida *Wv* e *Yv* respectivamente. Se debe tener en cuenta que cada mapa de SOM representan a dos componentes del modelo que se comparan.
ii. Antes del cálculo de la distancia entre dos neuronas se debe generar la instancia representativa de cada una, basado en los criterios establecidos previamente (la media para atributos numéricos y la moda para atributos no-numéricos).
iii. En caso de que uno de los pares de neuronas ($Xv_i$, $Yv_i$) no agrupe o clasifique instancias, al cálculo de la distancia $d(Xv_i, Yv_i)$ se le asigna el valor *null*. La justificación para utilizar este valor se debe a que la distancia proporciona un valor absoluto (positivo).
iv. Condiciones de borde:
    - $t, k > 0,0$.

- $d(Xv_i, Yv_i) > 0 \; \forall i \in \{1, \ldots, n\}, n \in \mathrm{N}$
- $d(Xv_i, Xv_i) = 0 \; \forall i \in \{1, \ldots, n\}, n \in \mathrm{N}$

v. Como salida se debe obtener una matriz con valores de las respectivas distancias de los pares de neuronas. Esta matriz se puede visualizar utilizando un color gris en degradé sobre cuadrículas. Así se puede observar con un tono más oscuro la representación de lejanía entre las componentes (valor de distancia mayor), y un color más claro para representar la cercanía (valor de distancia menor o tendiendo a cero). Para los casos señalados en el punto iii, es decir distancias no calculadas y asignadas con *null*, se coloca la cuadrícula sin color.

**Métrica de Similitud para Mapas SOM**

Similitud es un concepto alternativo al de distancia, y que sirve para comparar dos componentes representados como vectores $x_i$ y $y_j$, en función de su nivel de similitud denotada por *S(Xi, Yi)* (Maimon y Rokach [11]). Esta función debe ser simétrica, es decir; *S(Xi, Yj)* |= *S(Xj, Yi)*, y tienen un gran valor cuando *Xv* y *Yv)* son "similares" y constituyen el valor máximo cuando los vectores son idénticos.

Generalizando, si *Xv* e *Yv* son dos vectores de instancias *p* dimensionales, entonces la función *S(Xv, Yv)* mide la similitud entre estos dos vectores. Se establece que $-1 \leq S(Xv, Yv) \leq 1$ (función de similitud dicotómica), y se puede calcular utilizando diferentes métodos: medida del coseno, correlación de Pearson, extensión de Jaccard, y coeficiente de Sorensen-Dice. Para este estudio se utiliza el cálculo de la métrica utilizando la correlación de Pearson. La medida de correlación de Pearson para dos instancias en su forma normalizada se define como:

$$S\left(Xv, Yv\right) = \frac{\left(Xv - \underline{x}\right) * \left(Yv - \underline{y}\right)}{\left\|Xv - \underline{x}\right\| \cdot \left\|Yv - \underline{y}\right\|} \qquad (4)$$

Donde: $\underline{x}$ e $\underline{y}$ son los valores promedio de los vectores *Xv* e *Yv*, respectivamente.

En el contexto de un mapa de Kohonen, *Xv* e *Yv* pueden representar ya sea la capa de los pesos o bien la capa de salida de la red. En este caso, *S(Xv, Yv)* entrega una medida cuantitativa de la similitud ya sea la matriz de pesos de ambas redes o bien de la capa de salida de las redes. También es necesario considerar una instancia representativa de cada neurona en el cálculo de similitud utilizando la medida de correlación de Pearson. Ya que esta medida se refiere a instancias y no a un conjunto de ellas como se agrupan en cada neurona de un mapa SOM. Para esto, se consideran los mismos criterios utilizados para el cálculo de distancia (media para atributos numéricos y moda para atributos no-numéricos). Para el caso de los pesos *(w)* de los atributos se considera el valor que proporciona un algoritmo de ranking.

La adaptación de la métrica de correlación de Pearson queda:

- Dados dos vectores *Xv* e *Yv* de $(n \cdot m)$ dimensiones (o neuronas de las capas de salida de dos redes de Kohonen), en que cada neurona $X_{vi} \in X_v$ con *t* instancias y *p* atributos, y cada neurona $Y_{vi} \in y_v$ con *k* instancias y *p* atributos, la expresión general de la similitud a través del cálculo de correlación entre dos neuronas $X_{vi}$ e $Y_{vi}$ es:

$$S\left(X_{vi}, Y_{vi}\right) = \frac{\left(X_{vi}, \underline{X}\right) * \left(Y_{vi}, \underline{Y}\right)}{\left\|X_{vi}, \underline{X}\right\| \cdot \left\|Y_{vi}, \underline{Y}\right\|} \qquad (5)$$

**Alcances y situaciones de borde:**

i. Este cálculo se debe replicar para las $(n \cdot m)$ pares neuronas de cada mapa de salida $X_v$ e $Y_v$ respectivamente. Se debe considerar que cada mapa de SOM representan a dos componentes del modelo que se comparan.

ii. Antes del cálculo de la similitud entre dos neuronas se debe determinar la instancia representativa de cada neurona, basado en los criterios establecidos previamente (la media para atributos numéricos y la moda para atributos no-numéricos).

iii. En caso de que uno de los pares de neuronas $(X_{vi}, Y_{vi})$ no agrupe o clasifique instancias, al cálculo de la similitud $S(X_{vi}, Y_{vi})$ se le asigna el valor *null*.

iv. Condiciones de borde:
- $S(X_{vi}, Y_{vi}) \in [0,1] \subset R, \; \forall \; i \in \{1,\ldots, n\}, n \in N$
- $S(X_{vi}, Y_{vi}) = 1 \forall \; i \in \{1,.., n\}, n \in N$

Como salida se debe obtener una matriz con valores de las respectivas similitudes de todos los pares de neuronas. Esta matriz se puede visualizar utilizando un color rojo en degradé sobre cuadrículas diferente al utilizado en la distancia. Así se puede observar con un tono más oscuro del color rojo la representación de mayor similitud entre las componentes (valor de Pearson cercano a 1,0),

y un tono más claro para representar la menor similitud (valor de Pearson cercano o tendiendo a cero). Para los casos señalados en el punto iii, es decir similitudes no calculadas y asignadas con *null*, se coloca la cuadrícula sin color.

## HEURÍSTICA PARA EL CÁLCULO DE MÉTRICAS

Basada en la definición adaptada de las métricas descrita en la sección anterior, se presenta la Figura 1, donde se muestran las métricas propuestas en notación abstracta (entrada/salida) donde se define el algoritmo ***SOM_Metrics.*** Este algoritmo implementa una lógica general de cálculo de distancias para los mapas de SOM. Se consideran dos componentes seleccionados desde un modelo ÁD, donde los mapas son representados por vectores de dimensiones *N_ROWS* por *N_COLUMNS*. El algoritmo recibe como parámetros de entrada dos componentes seleccionados a partir un modelo ÁD. La salida proporciona dos cuadrículas que contienen valores resultantes para cada nodo de ambos mapas comparados. Estos pueden ser visualizados para ambas cuadrículas.

Una descripción detallada del algoritmo ***SOM_Metrics*** se presenta a continuación:

a. Las líneas 1-2 establecen el estado inicial para dos componentes seleccionadas desde el modelo de MD.
b. La línea 3 establece el ciclo general para el proceso de generación de las métricas.
c. La línea 4 controla el recorrido y procesamiento de una determinada fila de neuronas en ambas matrices de mapas SOM.
d. En las líneas 5-6 se realiza el llamado a la función generadora de la instancia representativa de cada instancia a partir de su correspondiente mapa SOM.
e. La línea 7 controla si la instancia representativa de alguna instancia procesada se encuentra vacía. Se asigna el valor *null*.

```
Algorithm SOM_metrics:  Proceso general de cálculo de métricas para mapas
                        SOM de dos componentes de modelos ÁD o RA. Cada
                        mapa SOM tiene dimensión N_ROWS por N_COLUMNS.

Input:  c1, c2 ∈ model_component
Output: g1, g2 ∈ grid_metric
Types:
    s1, s2 ∈ SOM_map
    i1, i2 ∈ rep_instance
    g1, g2 ∈ grid_metric
    d, s ∈ real
    f, c ∈ integer

Functions:
    create_SOM(c ∈ model_component) return SOM_map
    generate_rep_instance(s ∈ SOM_map, f ∈ integer, c ∈ integer) return
                                                          rep_instance
    isempty(i ∈ rep_instance) return boolean
    fdistance(i1 ∈ rep_instance, i2 ∈ rep_instance) return real
    fsimilar(i1 ∈ rep_instance, i2 ∈ rep_instance) return real
    add_gridMetric(m ∈ real  f ∈ integer, c ∈ integer, g ∈ grid_metric)
    visualize_metric(g ∈ grid_metric, v ∈ visual_color)

Algorithm:
     1. s1 ← create_SOM(c1)
     2. s2 ← create_SOM(c2)
     3. for f ← 1 to N_ROWS do
     4.      for c ← 1 to N_COLUMNS do
     5.           i1 ← generate_instance(s1, f, c)
     6.           i2 ← generate_instance(s2, f, c)
     7.           if isempty(i1) or isempty(i2) then
     8.                d ← -1
     9.                s ← -1
    10.           else
    11.                d ← fdistance(i1, i2)
    12.                s ← fsimilar(i1, i2)
    13.           end if
    14.           add_gridMetric(d, f, c, g1)
    15.           add_gridMetric(s, f, c, g2)
    16.      end for
    17. end for
    18. visualize_metric(g1, BLACK)
    19. visualize_metric(g2, RED)
```

Figura 1. Algoritmo *SOM_Metrics*.

f. Las líneas 11-12 establecen la semántica para el llamado de las funciones que calculan y retornan el resultado de ambas métricas (distancia y similitud) para las instancias representativas de los mapas SOM.
g. En las líneas 14-15, se agregan los resultados de las métricas en las respectivas matrices de cuadrículas en las mismas posiciones [*f, c*].
h. En las líneas 16-17 se controlan los ciclos iterativos de las columnas y filas para las matrices de los mapas de SOM.
i. Finalmente, en las líneas 18-19 se visualizan las matrices de cuadrículas con los resultados de ambas métricas. Se utilizan colores para diferenciar cada cuadrícula. Negro para distancia, y Rojo para similitud.

El proceso de generación de la instancia representativa (***generate_rep_instance***) para una componente se muestra en la Figura 2. Se utiliza el criterio establecido usando el cálculo de la media para atributos numéricos, y la moda para atributos no numéricos. Esta función recibe como parámetro de entrada un componente del modelo que puede ser indistintamente de un mapa SOM o nodo de ÁD. Como salida se entrega un vector con la instancia representativa de la componente analizada. Una descripción detallada del funcionamiento del algoritmo *generate_rep_instance* se presenta a continuación:

a. La línea 1 establece el estado inicial del ciclo del procesamiento de la componente que se recibe como parámetro de entrada. Este objeto tiene N_ATTR atributos y N_INST instancias. El ciclo controla el total de atributos en la componente.
b. En las líneas 2-5 se verifica el tipo de dato del atributo indexado con la variable i. Si se trata de un atributo numérico se invoca a la función que calcula el valor promedio de todas las instancias. En caso de que el atributo sea no-numérico se llama a la función que obtiene la moda (el valor más repetido) de todas las instancias asociadas. El resultado es almacenado en el vector de instancia representativa IR.
c. En la línea 7 cierra el ciclo general volviendo a su control en la línea 1.
d. El algoritmo termina en la línea 8 retornando la instancia representativa.

```
Algorithm generate_rep_instance: Genera instancia representativa de una
                                 componente de modelo de MD. La
                                 componente contiene N_ATTR atributos y
                                 N_INST instancias.

Input:  c ∈ model_component
Output: IR ∈ rep_instance
Types:
    IR ∈ rep_instance
    i ∈ integer
Functions:
    isnumeric(c.attribute[] ∈ model_component) return boolean
    faverage(numeric_attribute[] ∈ real, n ∈ integer) return real
    fmoda(no_numeric_attribute[] ∈ string, n ∈ integer) return string

Algorithm:
     1. for i ← 1 to N_ATTR do
     2.      if isnumeric(c.attribute[i]) then
     3.            IR[i] ← faverage(c.attribute[i][], N_INST)
     4.      else
     5.            IR[i] ← fmoda(c.attribute[i][], N_INST)

     6.      end if
     7. end for
     8. return IR
```

Figura 2. Heurística del proceso de generación de la instancia representativa.

## IMPLEMENTACIÓN DE MÉTRICAS

Las dos métricas definidas anteriormente, permiten comparar de manera cuantitativa dos componentes de un modelo de ÁD, las cuales son implementadas sobre una instancia representativa del conjunto de datos agrupadas en cada instancia de los mapas SOM. Estos mapas son generados a partir de dos componentes seleccionados. Adicionalmente, para el modelo de ÁD, ambas métricas se implementan directamente en los datos ya clasificados que pertenecen a los diferentes nodos del ÁD. Esto permite visualizar los valores de distancia y similitud entre las demás componentes del modelo. Para validar la implementación y uso de ambas métricas, se utiliza como ejemplo el modelo de MD generado a través de la técnica ÁD utilizando el conjunto de datos Iris.

En la Figura 3 se encuentran destacados los dos nodos seleccionados del modelo de ÁD (ícono hexagonal) que tienen generado su mapa SOM respectivo.

En la Figura 4 se presentan dos mapas de SOM que muestran los detalles internos que resulta de explorar los nodos seleccionados del modelo ÁD (Figura 3), y su distribución de instancias.

Según se ha establecido en la definición de las métricas para cada componente se encuentran representados por matrices que almacenan las instancias agrupadas en los componentes del mapa. El software prototipo ha limitado estas matrices a una dimensión de 5 x 5 para simplificar el despliegue visual en pantallas con menor resolución motivos de simplificación. Se debe generar la instancia representativa con los criterios establecidos, esto es: el valor promedio para atributos numéricos y la moda para atributos no-numéricos. Luego a estas componentes se le aplican las métricas de distancia y similitud en forma simétrica, es decir para la instancia del primer mapa ubicada en la posición [*fila, columna*] = [1, 1] se calculan las métricas respecto a la otra instancia del segundo mapa ubicada en la misma posición [1, 1]. En caso de que la instancia no tenga asignada datos se asigna el valor *null* en ambas métricas. El orden de interpretación visual de las matrices SOM es vertical, es decir la fila 1 corresponde a la primera columna del lado izquierdo, y la última fila (5) es la del extremo derecho.

### Implementación Métrica Distancia

Se utiliza como métrica de distancia la fórmula de Minkowski parametrizada para el método de distancia Euclidiana. Una vez generados los respectivos

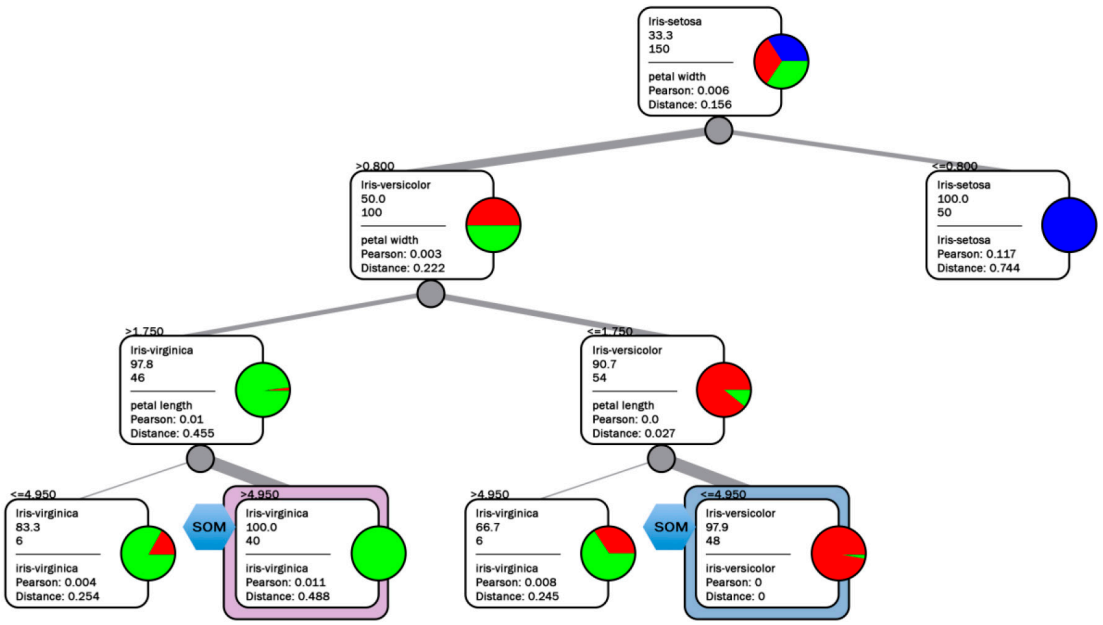


Figura 3. Técnica ÁD utilizando el conjunto de datos Iris.

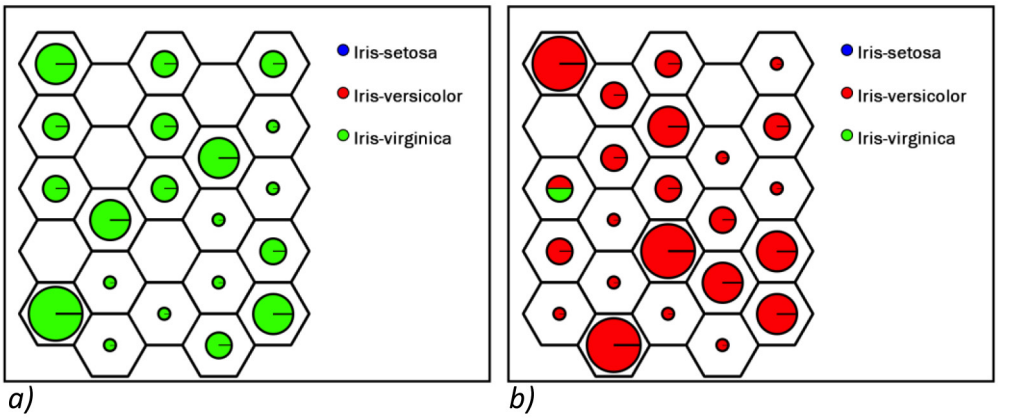


Figura 4. Técnica ÁD utilizando el conjunto de datos Iris.

mapas SOM para los dos nodos seleccionados, se genera una matriz de cuadrículas también de orden 5 x 5 (Figura 5). Estas cuadrículas se ubican dentro de pequeños círculos y representan las distancias Euclidianas de todos los pares de instancias de ambos mapas SOM. Se puede observar visualmente el uso del color negro con degradación para representar la distancia entre las correspondientes instancias de los mapas de SOM.

El valor mínimo de distancia (cercanía de las componentes) que tiende a cero se representa por la tonalidad más clara. El valor máximo de distancia (lejanía de las componentes) se representa por una tonalidad que va de gris a negro. Para el caso de instancias que no agrupan valores el cálculo de distancia no es aplicable y se asigna un valor *null*. Estos últimos se representan en la matriz de cuadrículas por el símbolo ∅ que significa vacío. La matriz de cuadrículas mantiene el mismo orden de recorrido visual descritos para los mapas de SOM, igual que la Tabla 2 que registra los valores que se obtienen en esta métrica.

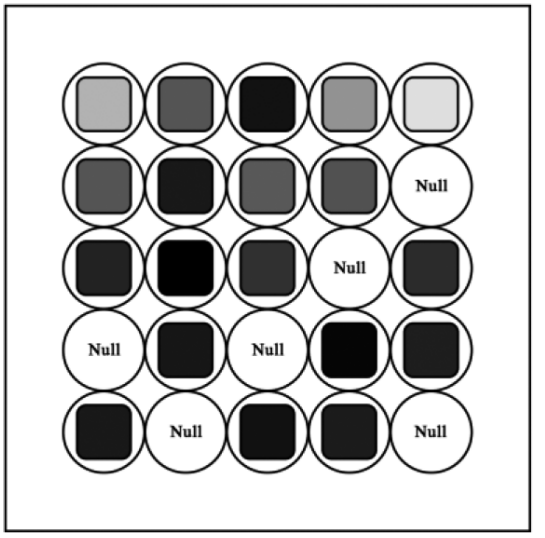


Figura 5. Cuadrículas de distancia entre dos mapas SOM.

**Evaluación de Métrica Distancia**

Se establecieron criterios para evaluar el nivel de cercanía entre dos componentes. Se considera el rango de valores para la distancia representada por $d(X_{vi}, Y_{vi}) \in R > 0$.

*Se establecen los siguientes criterios:*

a. Valor de instancias cercanas si el valor resultante $d(X_{vi}, Y_{vi}) < 0\ d(X, Y)$.
b. Valor de instancias lejanas si el valor resultante $d(X_{vi}, Y_{vi}) \geq d(X, Y)$.
c. Mapas o componentes cercanas si el 70% de su $d(X_{vi}, Y_{vi})\ es < d(X, Y)$.
d. Mapas o componentes lejanas si el 70% de su $d(X_{vi}, Y_{vi})\ es \geq d(X, Y)$.

Donde $d(X, Y)$ corresponde a la media general de los valores de distancia. En este cálculo no se consideran las celdas con valor *null*.

Basado en los criterios establecidos y al evaluar los datos que se obtienen desde la Tabla 2, se pueden corroborar algunos valores que hacen coherente la matriz de cuadrículas de distancia:

Tabla 2. Distancia de dos mapas de SOM.

| | **Fila 1** | **Fila 2** | **Fila 3** | **Fila 4** | **Fila 5** |
|---|---|---|---|---|---|
| Col 1 | 0,8515 | *null* | 0,6386 | *null* | 0,8335 |
| Col 2 | *null* | *null* | 0,6504 | 0,5663 | 0,6797 |
| Col 3 | 0,4958 | 0,4451 | 0,5621 | 0,4523 | 0,6982 |
| Col 4 | *null* | 0,3334 | *null* | 0,5577 | 0,6740 |
| Col 5 | 0,3100 | 0,2891 | 0,4837 | 0,4424 | 0,5069 |
| Mín | 0,8515 | 0,2891 | 0,4837 | 0,4424 | 0,5069 |
| Max | 0,8515 | 0,4451 | 0,6504 | 0,5663 | 0,8335 |
| Media | 0,5500 | 0,3500 | 0,5800 | 0,5000 | 0,6700 |
| Media General | | | 0,5300 | | |

- En la [*fila, columna*] = [1,1] se encuentra el mayor valor de distancia que se obtiene (0,8515) y en la misma posición de la matriz de cuadrículas de distancias aparece con la tonalidad más oscura (negra). Además, este valor es mayor a la media general (0,5300). Esto indica que las instancias representativas de las neuronas de ambos mapas SOM ubicadas en esta posición (Figuras 4 a) y b)) son *distantes* respecto a las demás.
- Diferente a lo anterior, el valor mínimo (0,2891) ubicado en la posición [*fila, columna*] = [2,5] obtiene una cuadrícula con la tonalidad más clara, y menor a la media general, lo que representa *cercanía* entre las dos neuronas de los mapas SOM que están en la misma posición.
- También en todas aquellas celdas donde aparece el valor *null* significa que una de las neuronas o ambas, no agrupan instancias por lo que el cálculo de distancia entre ellas no es posible. Por ejemplo, la posición [*fila, columna*] = [3,4] tiene este valor porque la neurona del mapa SOM del nodo izquierdo (Figura 4 a)) no agrupa instancias en esta posición. En cambio, la posición [*fila, columna*] = [4,1] de ambos mapas SOM no agrupa instancias y por tanto el valor *null* también es asignado.

Finalmente, en la Figura 6 se muestra el desarrollo general del modelo incorporando la métrica de distancia sobre la técnica ÁD.

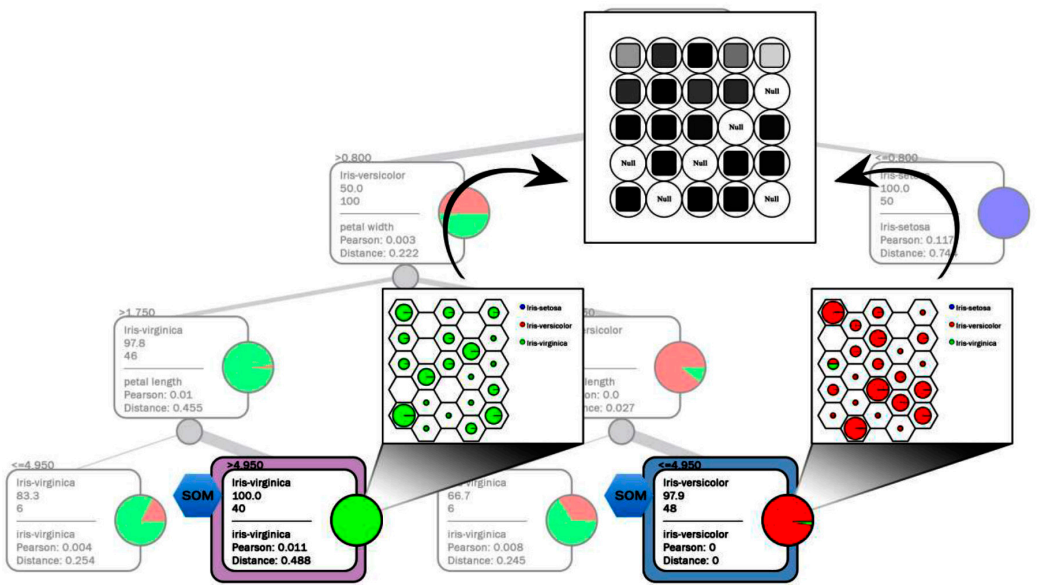


Figura 6. Modelo ÁD y métrica de distancia.

**Implementación Métrica Similitud**

La segunda métrica implementada es la similitud entre componentes, basada en la fórmula de correlación normalizada de Pearson [12]. Una vez generados los respectivos mapas SOM de los dos nodos ÁD seleccionados, se genera una matriz de cuadrículas de orden 5 x 5 (Figura 7). De forma similar a la matriz de cuadrículas de distancias, se ubican dentro de pequeños círculos que representan el nivel de similitud de todos los pares de instancias de ambos mapas SOM. Se puede observar visualmente para cada cuadrícula la utilización del color rojo en degradación para representar la distancia entre las correspondientes instancias de los mapas de ambos nodos ÁD. Los valores que se obtienen desde la correlación de Pearson oscilan en el rango [0,1].

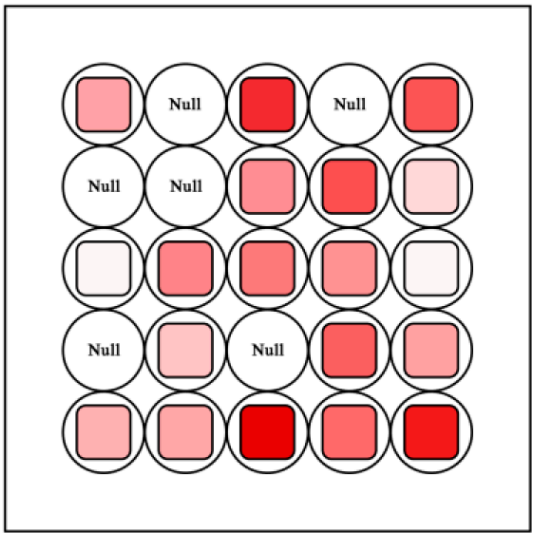


Figura 7. Cuadrículas de similitud entre dos mapas SOM.

El valor mínimo de similitud (diferencia de las componentes) tendiendo a cero se representa por la tonalidad más clara. El valor máximo de similitud (igualdad de las componentes) se representa por una tonalidad en degradación a rojo oscuro. Para el caso de instancias que no agrupan valores el cálculo de similitud no es aplicable y se asigna un valor *null*. La representación de vacío se hace mediante el símbolo ∅. La Tabla 3 registra los valores obtenidos en esta métrica.

**Evaluación Métrica de Similitud**

Igual que en la métrica de distancia, se deben establecer criterios de evaluación para los dos componentes, cuyo rango de valores es acotado: $S(X_{vi}, Y_{vi} = [0,1]$:

Instancias con similitud:

a. "alta" si: $0,7 < S(X_{vi}, Y_{vi}) \leq 1$
b. "media" si: $0,5 < S(X_{vi}, Y_{vi}) \leq 0.7$
c. "baja" si: $S(X_{vi}, Y_{vi}) \leq 0.5$

Mapas o componentes con similitud:

a. "alta" si el 70% de sus $S(X_{vi}, Y_{vi})$ *es alta.*
b. "media" si el 70% de sus $S(X_{vi}, Y_{vi})$ *es media.*
c. "baja" si el 70% de sus $S(X_{vi}, Y_{vi})$ *es baja.*

Con estos criterios de evaluación, al someter los valores resultantes de la métrica de similitud obtenida desde la Tabla 3, también es posible corroborar algunos valores que hacen coherente la matriz de cuadrículas:

- Aunque los valores de esta tabla son más cercanos al mínimo (cero), se pueden observar en las posiciones: [*fila, columna*] = [3,1], [3,5], [5,5] los tres valores más altos que se obtienen respectivamente (0,0237, 0,0311, 0,0262), y en

Tabla 3. Similitud para dos mapas de SOM.

| | **Fila 1** | **Fila 2** | **Fila 3** | **Fila 4** | **Fila 5** |
|---|---|---|---|---|---|
| Col 1 | 0,0101 | *null* | 0,02237 | *null* | –0,0191 |
| Col 2 | *null* | *null* | 0,0124 | 0,198 | 0,0045 |
| Col 3 | 0,0008 | 0,0135 | 0,0152 | 0,0120 | 0,0009 |
| Col 4 | *null* | 0,0102 | 0,0152 | 0,0180 | 0,0066 |
| Col 5 | 0,0088 | 0,0097 | 0,0124 | 0,0120 | 0,0262 |
| Mín | 0,0008 | 0,0097 | 0,0311 | 0,0198 | 0,0009 |
| Max | 0,0101 | 0,0135 | 0,0311 | 0,0198 | 0,0262 |
| Media | 0,0000 | 0,0100 | 0,0200 | 0,0100 | 0,0100 |

la mismas posiciones de la matriz de cuadrículas de similitud aparecen con tonalidad rojiza más oscura. Esto indica que las instancias representativas de las neuronas de ambos mapas SOM ubicadas en estas posiciones (Figuras: 4 a) y b)) tienen un nivel mayor de similitud respecto a las demás. Si se evalúan bajo los criterios definidos, todos los valores están bajo el 0,5, por lo que el nivel de similitud de todas las neuronas es bajo.

- Los valores mínimos (0,0008 y 0,0009) están ubicados en las posiciones: [*fila, columna*] = [1,3], [5,3] respectivamente. Obtienen estos valores las cuadrículas con tonalidad más clara del color rojo, incluso tomado el color blanco. Esto también representa un nivel bajo de similitud entre las dos neuronas de los mapas SOM que están en la misma posición.
- En las celdas donde aparece el valor *null* significa que una de las neuronas o ambas, no agrupan instancias y el cálculo de esta métrica no es posible.
- Una visión general de la matriz de cuadrículas de similitud permite afirmar que todas las neuronas (100%) de ambos mapas contienen una tonalidad más cercana al rango rojo claro o valor mínimo (cero), lo que se puede deducir que son componentes con nivel de similitud bajo.

Finalmente, en la Figura 8 se muestra el desarrollo general del modelo incorporando la métrica de distancia sobre la técnica ÁD.

## CONCLUSIONES Y TRABAJO FUTURO

Se ha logrado establecer en este trabajo el rol relevante que cumple la visualización en el proceso de MD, principalmente en el análisis exploratorio de los modelos en la etapa de ajuste y construcción. Lo anterior, dejando de manifiesto lo crucial que es la comprensión del modelo de MD que se construye, para lograr una interpretación adecuada de los patrones que se obtienen. Todo esto en el marco de un proceso iterativo e interactivo y con la participación activa tanto del usuario final como el analista de datos.

Una de las complejidades que conlleva la obtención de un adecuado modelo de MD, es la comprensión de su funcionamiento interno. Y en esto, la mayoría de las actuales herramientas de MD no aportan mucho, ya que por lo general sólo proporcionan visualización estática de la estructura del modelo sin permitir explorar sus componentes. Esto puede impactar la credibilidad del modelo debido a la falta de entendimiento y dificultad, en algunos casos, en la interpretación y difusión de los patrones generados a partir del modelo de MD.

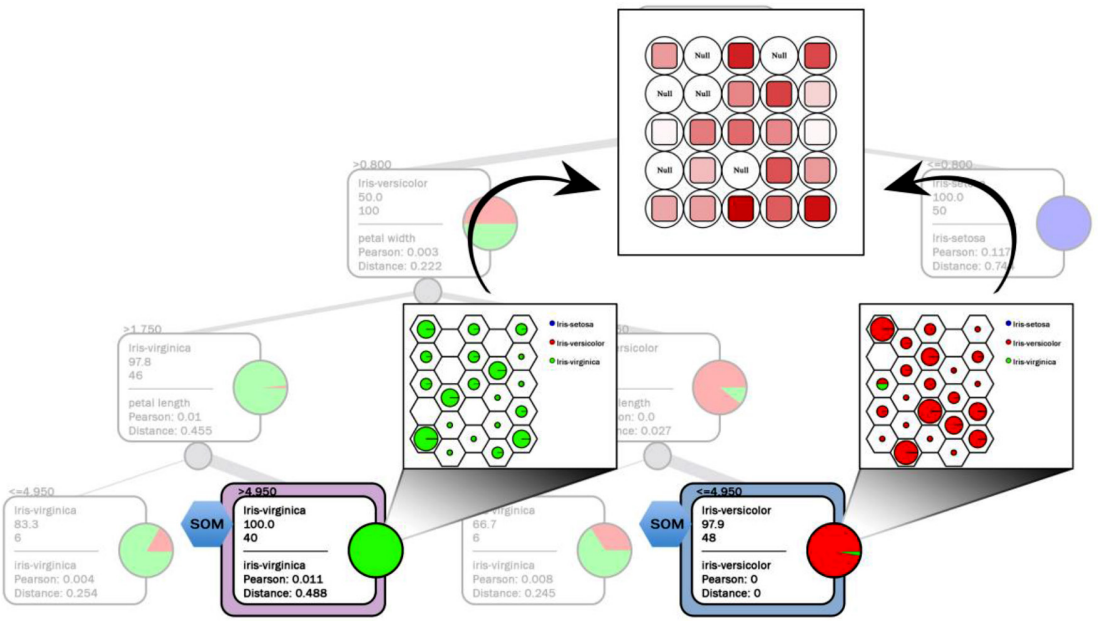


Figura 8. Modelo ÁD y métrica de similitud.

Las métricas de similitud y distancia, adaptadas y aplicadas a la técnica SOM, entregan una herramienta complementaria de validación de las componentes o nodos de un modelo generado con la técnica ÁD. Una visión general de esta matriz de cuadrículas de distancias permite afirmar que la mayoría de las neuronas de ambos mapas contienen una tonalidad más cercana al rango gris a oscuro, y observando sus valores 17 de 19 neuronas sus distancias son mayor e igual a la media general, esto es el 89% de las neuronas de ambos mapas son lejanas. Esto permite deducir que las componentes analizadas son mayormente distantes ya que más del 70% de sus neuronas lo son. Contrastando las matrices de cuadrículas de ambas métricas (distancia y similitud), se puede afirmar que las dos componentes comparadas reúnen instancias mayormente distintas y distantes. Esto permite tener una variable adicional al momento de analizar, interpretar y comprender el modelo de MD.

En relación al trabajo futuro, se está trabajando en los siguientes aspectos:

- Se requiere una mejora en la implementación de la heurística para mapas de SOM con objetivo de mejorar los tiempos de cálculos en mapas con matrices de orden superior a 5 x 5.
- Se tiene previsto evaluar la integración de nuevas técnicas de evaluación de similitud para nuevos modelos: Reglas de Asociación y Clústering.
- Se debe avanzar en la implementación y fundamentación de la aplicación de este tipo de métricas con el objetivo de generalizarse para un dominio de datos más diverso.
- Es necesario integrar las técnicas de visualización y similitud en la exploración general del modelo ÁD y SOM.

## REFERENCIAS


[1] U. Fayyad, G. Piatestky-Shapiro and P. Smyth. "The KDD Process for Extracting Useful Knowledge from Volumes of Data". Comm. Of the ACM. Vol. 39, Issue 11, pp. 27-34. 1996.

[2] P. Adriaans and D. Zantinge. "Data Mining". Addison-Wesley Pub. Reading, Mass, USA. 1996. ISBN: 0201403803.

[3] E. Brachman and T. Anad. "The Process of Knowledge Discovery in Databases". In Advances in Knowledge Discovery and Data Mining, AAAI Press/ The MIT Press, pp. 37-57. 1996.

[4] G. John. "Enhancements to the Data Mining Process". Doctoral Dissertation, Depart. of Comp. Science, Stanford University. 1997.

[5] C. Meneses and G.G. Grinstein. "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, Orlando, Florida, USA. April 16-20, 2001.

[6] W. Castillo-Rojas and C. Meneses. "Graphical Representation and Exploratory Visualization for Decision Trees in the KDD Process". XXXVIII Conferencia Latinoamericana en Informática - CLEI 2012, Medellín, Colombia. Libro de Resúmenes página 137. 2012. ISBN IEEE Xplore: 978-1-4673-0793-2, ISBN CD: 978-1-4673-0792-5. Print ISBN: 978-1-4673-0794-9.

[7] W. Castillo-Rojas, C. Meneses and F. Medina-Quispe. "Augmented Data Mining Models Using Visualization". Artículo presentado en 6th Latin American Conference on Human Computer Interaction-CLIHC 2013. Guanacaste, Costa Rica, pp. 148-155. 2013. Springer LNCS 8278. ISSN: 0302-9743. ISBN: 978-3-319-03067-8. DOI: 10.1007/978-3-319-03068-5.

[8] D. Keim. "Visual Techniques for Exploring Databases". Tutorial Notes in the Third International Conference on Knowledge Discovery and Data Mining, KDD-97. Newport Beach, CA. 1997.

[9] P. Hoffman. "Table Visualizations: A Formal Model and its Applications". Sc.D. Thesis, Dept. of Comp. Science, University of Massachusetts at Lowell. 1999.

[10] K. Thearling, B. Becker, D. DeCoste, B. Mawby, M. Pilote and D. Sommerfield. "Visualizing Data Mining Models". In Proceedings of the Integration of Data Mining and Data Visualization Workshop, Springer-Verlag. 1998.

[11] O. Maimon and L. Rokach. "Data Mining and Knowledge Discovery Handbook, 2nd ed". Springer Science+Business Media, Inc. Edited by Oded Maimon and Lior Rokach

Tel-Aviv University. Israel. 2010. ISBN 978-0-387-09822-7.

[12] K. Pearson. "Notes on regression and inheritance in the case of two parents". Proceedings of the Royal Society of London. Vol. 58, pp. 240-242. 1895.

[13] J.S. Rao and W.J.E Potts. "Visualizing Bagged Decision Trees. Proc. of Third International Congress on Knowledge Discovery and Data Mining". AAA1 Press, pp. 243-246. 1997.

[14] M. Ankerst, C. Elsen, M. Ester and H.P. Kriegel. "Visual Classification: An Interactive Approach to Decision Tree Construction". Proc. of V International on Knowledge Discovery and Data Mining, pp. 392-397. 1997.

[15] L. Breiman, J. Friedman, R. Olshen and C. Stone. "Classification and Regression Trees". Belmont, CA: Wadsworth. 1997.

[16] B. Johnson and B. Shneiderman. "Treemaps: A Space-Filling Approach to the Visualization of Hierarchical Information Structures". Proc. of IEEE Information Visualization, pp. 275-282. 199.

[17] G.G. Robertson, J.D. Mackinlay and S.K. Card. "Cone Trees: Animated 3D Visualization of Hierarchical Information". Proc. of the ACM Congress on Human Factors in Computing Systems, pp. 189-194. 1991.

[18] J. Lamping and R. Rao. "The Hyperbolic Browser: A Focus + Context Techniques for Visualizing Large Hierarchies". Journal of Visual Languages and Computing, pp. 33-55. 1997.

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

[20] T.D. Nguyen, T.B. Ho and H. 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. Vol. 1805/2000, pp. 345-348. 2000. DOI: 10.1007/3-540-45571-X_40.

[21] J. Hernández, M. Ramírez y C. Ferri. "Introducción a la Minería de Datos". Pearson Prentice Hall. Madrid, España. 2004. ISBN: 9788420540917.

[22] M. Sjöberg and J. Laaksonen. "Optimal Combination of SOM Search in Best-Matching Units and Map Neighborhood". In: Príncipe J.C., Miikkulainen R. (eds) Advances in Self-Organizing Maps. WSOM 2009. Lecture Notes in Computer Science. Vol 5629. Springer, Berlin, Heidelberg. 2009.

[23] J. Han, J. Pei and M. Kamber. "Data Mining: Concepts and Techniques". The Morgan Kaufmann Series in Data Management Systems. Elsevier Science. 2011.

[24] X. Zhang, T. Simpson, M. Frecker and G. Lesieutre. "Supporting knowledge exploration and discovery in multi-dimensional data with interactive multiscale visualisation". Journal of Engineering Design. Vol. 23, Issue 1, pp. 23-47. 2012.

[25] W. Castillo-Rojas, F. Medina-Quispe y C. Meneses-Villegas. "Modelo aumentado de árbol de decisión utilizando mapas autoorganizados". Ingeniare. Revista chilena de ingeniería. Vol. 22 Nº 3, pp. 351-362. 2014. URL: https://dx.doi.org/10.4067/S0718-33052014000300006