Visualización Interactiva para Modelos de Clústeres

Interactive Visualization for Clustering Models

Wilson Castillo-Rojas[1]* Juan Vega Damke[2]

Recibido 24 de octubre de 2016, aceptado 25 de septiembre de 2017
Received: October 24, 2016 Accepted: September 25, 2017

RESUMEN

Este artículo presenta el diseño e implementación de un nuevo esquema de visualización interactiva para modelos de clústeres denominado VIMC, en el contexto de un proceso de minería de datos. La visualización de un modelo de clústeres se torna compleja cuando el conjunto de datos es de alto volumen, densidad y dimensionalidad. El diseño del esquema VIMC se basa en cuatro características: visualización interactiva, combinación de modelos, artefactos gráficos ad-hoc, y uso de métricas. El objetivo de este esquema es contribuir al análisis y comprensión de un modelo de clústeres. Las métricas consideradas en este esquema propuesto, permiten comparar componentes de distintos clústeres, lo que a su vez ayuda a entender como se componen los grupos. A través de la implementación de un entorno visual web que reúne las características definidas en VIMC, y una evaluación en línea de 23 usuarios, donde se logran resultados positivos sobre la utilidad de este nuevo esquema de visualización.

Palabras clave: Visualización de clústeres, visualización de modelos para minería de datos, visualización de minería de datos, visualización interactiva de modelos, esquemas de visualización de datos.

ABSTRACT

This paper presents the design and implementation of a new interactive visualization scheme for clustering models called VIMC, in the context of a data-mining process. The visualization models cluster becomes complex when the dataset is high volume, density and dimensionality. The design of VIMC scheme is based on four characteristics: interactive visualization, combination of models, ad-hoc graphics artifacts, and use metrics. The objective of this scheme is to contribute to the analysis and understanding of a clustering model. Metrics considered in this proposed scheme, allow comparison components of different clusters, which in turn helps to understand how groups are composed. Through the implementation of a web visual environment that meets the characteristics defined in VIMC, and an online assessment of 23 users, where positive results on the usefulness of this new visualization scheme are achieved.

Keywords: Cluster Visualization, model visualization for data mining, data mining visualization, interactive visualization of models, data visualization schemes.

[1] Departamento de Ingeniería Informática y Ciencias de la Computación. Universidad de Atacama. Av. Copayapu 485. Copiapó, Chile. E-mail: wilson.castillo@gmail.com
[2] Facultad de Ingeniería y Arquitectura. Universidad Arturo Prat. Av. Arturo Prat 2120. Iquique, Chile. E-mail: jfranciscov7@gmail.com
* Autor de correspondencia

INTRODUCCIÓN

Comprender e interpretar apropiadamente los resultados de un modelo es lo más crítico en un proceso de Minería de Datos (MD). Se sostiene que una de las formas de apoyar esto es utilizando visualización apropiada en la construcción del modelo. Esto establece que al conocer el funcionamiento interno del modelo, permite por un lado, comprender como funciona y por otro lado, interpretar mejor sus resultados. En particular, visualización de modelos de clústeres es difícil de llevar a cabo, principalmente cuando el conjunto de datos es de gran volumen, y el número de dimensiones es alto. Por ejemplo, con 3 dimensiones no se pueden observar todos los atributos simultáneamente, más complejas en el caso de clústeres de alta densidad.

De la revisión del estado de arte en este trabajo, y la evaluación de herramientas de MD existentes, se pudo constatar la dificultad que presenta visualización de modelos de clústeres. También, que existen diversas métricas para comparar clústeres, pero la mayoría de las herramientas utilizan el cálculo de distancia, y muy pocas, implementan comparación de componentes. Además, las herramientas de MD analizadas no realizan combinación de técnicas, y no proporcionan un nivel adecuado de interacción para explorar en profundidad un modelo.

El objetivo de esta investigación justamente aborda el problema de la complejidad para la visualización de modelos de clústeres. Para esto es necesario considerar elementos adecuados para el diseño y desarrollo de un esquema de visualización, que permita explorar un modelo de clústeres, bajo demanda, y para lo cual es fundamental definir mecanismos de interacción apropiados.

Otro elemento es la combinación del modelo de clústeres con un modelo descriptivo que permita establecer relaciones entre los atributos del conjunto de datos. También, es necesario disponer de artefactos gráficos ad-hoc para representar visualmente clústeres, particularmente en conjunto de datos con alto volumen, densidad y dimensionalidad. Finalmente, poder comparar cuantitativamente instancias o componentes de distintos clústeres, así como la homogeneidad del conglomerado, a través de métricas adecuadas para esto.

Este artículo describe un nuevo esquema de visualización para modelos de clústeres denominado VIMC. Se presentan su diseño conceptual y aspectos técnicos, además de la implementación de un entorno visual web, que proporciona las características principales de este esquema VIMC: visualización interactiva para explorar el modelo, combinación de modelos para describir relaciones entre atributos, gráficos para modelo de clústeres, métricas para comparar instancias de distintos clústeres y medir el nivel de compacidad o dispersión de los clústeres.

Finalmente, se discuten los resultados de una evaluación subjetiva obtenida a través de una encuesta en línea con 23 usuarios, todos con experiencia en procesos de MD.

VISUALIZACIÓN EN MINERÍA DE DATOS

El proceso de MD busca obtener, a partir de un conjunto masivo de datos, modelos de datos para describir o predecir nuevas instancias. Esto implica etapas de: preparación de datos (entrada), el procesamiento de los datos para lograr modelos, y la interpretación de los patrones resultantes (salida). Esta salida debe significar un nuevo conocimiento en la organización, útil y comprensible para los usuarios finales, y que se puede integrar a los procesos para apoyar la toma de decisiones. Sin embargo, la dificultad está en identificar modelos en los datos, lo cual es una tarea compleja y, a menudo requiere experiencia, no sólo el analista de datos, sino también el experto en el dominio del problema.

Una forma de apoyar el análisis de datos, modelos y patrones, es a través de su representación visual. El uso de las capacidades de la percepción visual humana, que puede detectar patrones más fácilmente. Bajo este enfoque, visualización en minería-datos ha sido utilizada mayormente en el análisis descriptivo de los datos (entrada) y en la presentación de los patrones (salida), dejando limitado este paradigma para el análisis de modelos [1].

Meneses y Grinstein discuten en [2], elementos para avanzar hacia un esquema para el proceso de MD, con soporte para visualizar cuatro tipos de entidades: datos, parámetros de algoritmos, modelos inducidos, y validación para los patrones.

Un factor clave en MD, para mejorar la capacidad predictiva o descriptiva de un modelo, es entender cómo el modelo inducido está trabajando. Entre otros aspectos, por ejemplo; es importante entender cómo el modelo realiza una distribución del conjunto de datos de acuerdo a diferentes atributos, cómo las componentes del modelo están correlacionadas con un subconjunto de observaciones, y cómo los valores de los atributos es particionado por el modelo.

Algunas técnicas de aprendizaje automático funcionan como sistemas cerrados, y es difícil lograr una interpretación de los patrones obtenidos, y responder preguntas del usuario acerca de la transformación que realiza el modelo. Un ejemplo notable son las redes neuronales artificiales, las cuales convergen a un conjunto de pesos numéricos que no tienen una interpretación directa en el dominio del problema.

Las técnicas como árboles de decisión son fácilmente comprensibles cuando el modelo es pequeño, árboles de mayor tamaño (más de tres niveles) su interpretación es compleja. Esto mismo ocurre para un modelo de clúster, ya que con datos de gran volumen y alta dimensionalidad, no es simple visualizar todos los atributos simultáneamente. Otras técnicas de aprendizaje, como reglas de asociación, también el tamaño del modelo presenta problemas (largas listas de reglas), y requiere el desarrollo de nuevas representaciones gráficas, que proporcionen una mejor visualización e interacción con el modelo, para facilitar su interpretación.

Se considera que la integración de visualización en el proceso de MD puede ser de importancia significativa, ya que puede ayudar de dos maneras:

- Proporcionar comprensión visual de complejas aproximaciones computacionales y;
- Descubrir relaciones complejas entre los datos que no son detectables por los métodos automáticos de análisis, pero que si pueden ser captados por el sistema visual humano.

De lo anterior, se deduce que la iteración alternada entre visualización y minería de datos automática provee al analista de datos, soporte en la tarea de reconocimiento de patrones. Incluyendo al ser humano en este proceso, se logra una buena combinación de su percepción visual, con el poder de cálculo y almacenamiento del computador [2].

Hoy en día, existen pocas herramientas que incorporan visualización de modelos [3, 4]. Las que existen son muy limitadas en su capacidad de exploración, y en responder preguntas acerca de las transformaciones, que realiza el modelo en los datos, para generar patrones.

VISUALIZACIÓN DE CLÚSTERES

El análisis de clúster o segmentación es una tarea de MD, que tiene por objetivo determinar y describir grupos en un conjunto de datos, de tal manera que los elementos asignados a un mismo grupo sean similares entre sí, mientras que los elementos pertenecientes a grupos distintos sean disímiles entre sí. En general, la medida de distancia es la más utilizada para realizar comparaciones entre los atributos de las instancias. Esta tarea es utilizada para realizar levantamiento de perfiles en: clientes, trabajadores, pacientes de hospitales, etc. Al finalizar el proceso de clustering, se logra como resultado etiquetar cada instancia del conjunto de datos, asociándola con uno de los grupos generados [5].

Existen diversos métodos o técnicas para generar agrupamientos, y cada uno utiliza un principio distinto de inducción. Fraley en [6], propone 2 tipos de métodos de clustering: jerárquicos y particiones. Por otro lado, Han y Kamber proponen [7], 3 tipos de clustering basados en: densidad, modelos y rejillas.

Existen diversos tipos de visualizaciones para representar modelos de clústeres:

- *Gráfico de Dispersión:* es el gráfico más utilizado, y permite observar la relación que existe entre dos variables numéricas. Cada punto corresponde a los valores en las coordenadas *(x, y)*. A cada grupo se le asigna un color y las instancias pertenecientes a dicho grupo toman ese color, de esta forma es fácil identificar a que grupo pertenecen. Las instancias del conjunto de datos son graficadas con un máximo de 3 dimensiones. A través de este gráfico, se pueden identificar correlaciones entre las variables.
- *Diagrama de Voronói:* Es aplicable a técnicas que generan centroides, y consiste en una partición del espacio euclidiano. Para cada par de centroides, se crean bisectrices perpendiculares, las cuales generan segmentos paralelos entre

ellos. Este tipo de visualización para clúster genera polígonos para cada grupo, donde las instancias se distribuyen en su interior.

- *Densidad de Grupos:* corresponde a una visualización donde se pueden observar las densidades de cada grupo. Se muestra la región que cubre cada clúster, y se puede realizar análisis de las instancias.
- *Matriz de Diagramas de Dispersión:* es una generalización de un diagrama de dispersión que permite observar de manera simultánea, el comportamiento de variables numéricas en en varias dimensiones. Es una herramienta de exploración de datos, que permite comparar un conjunto de instancias con respecto a todos sus atributos. Este gráfico muestra todas las combinaciones de gráficos de dispersión entre dos variables, en una sola vista bajo una estructura matricial. Para *n* dimensiones se muestran *n* filas y columnas.
- *Coordenadas Paralelas:* este tipo de diagrama se compone de un eje horizontal y varios ejes verticales. Cada eje vertical representa a una variable, cuyos valores son representados a lo largo de su eje horizontal. Para el caso de variables numéricas, los valores se ordenan de menor a mayor (de abajo hacia arriba). Cada línea poligonal de este diagrama corresponde a una instancia (fila) del conjunto de datos, y se mueve entre los ejes verticales dependiendo de los valores que le corresponde en cada variable. Este tipo de gráfico, acompañado de elementos de interacción, como transparencia y filtros ofrece una visualización muy potente para objetos multidimensionales.
- *RadViz:* es una técnica de visualización de coordenadas radiales. Muestra todos los atributos como puntos anclados al perímetro de una circunferencia, y separados en forma equidistante dependiendo de la cantidad de atributos. Dentro del círculo, se muestran las instancias en forma de puntos, los cuales están dispuestos en el gráfico. Toma como base al paradigma del tensor proveniente de la física de partículas, puntos de la misma clase se atraen entre sí, los de diferente clase se repelen entre sí, y las fuerzas resultantes se ejercen sobre los puntos de anclaje. Una ventaja de RadViz es que conserva ciertas simetrías de los datos, y su principal desventaja es la superposición de puntos [8].
- *Dendrograma:* es la forma más simple en la cual una estructura jerárquica de datos puede ser representada, y corresponde a un gráfico con forma de árbol (dendro; significa árbol). Los ítems de datos son representados por las hojas en el nivel final de la estructura del árbol, mientras que los nodos en el nivel más alto, son los que representan a los grupos o clústeres de los ítems de datos, con diversos niveles de semejanza (Jain y otros [9]). La manera clásica de representar el dendrograma es dibujarlo como árbol enraizado, con la raíz anclada centralmente en el tope de la imagen, y las ramas de los nodos-hijos hacia abajo utilizando líneas derechas o diagonales.
- *Gráfico de Radar:* en este gráfico cada instancia es graficada como un polígono en un espacio radial, y sus aristas son generadas en relación a la magnitud de sus dimensiones, las instancias pertenecientes a mismos grupos presentan los mismos colores. Las figuras geométricas que presenten similitudes en su estructura corresponden a objetos similares.

TRABAJOS RELACIONADOS

Long propone en [10], un trabajo consistente en una solución de visualización para modelos de clústeres jerárquicos. Su hipótesis principal es que existen dos grandes problemas para visualizarlos:

a) La gran cantidad de datos limita la visualización de un modelo de clústeres, ya que el espacio en la pantalla no permite observar todas las instancias de manera simultánea.
b) Cuando el número de dimensiones es mayor a 3, no se pueden visualizar todos los atributos simultáneamente.

La propuesta consiste en dos pasos: clustering y visualización. Ayuda a los analistas de datos, a entender la distribución de un conjunto de datos con alta dimensionalidad. En el primer paso, se utilizan técnicas de algoritmos jerárquicos para generar clústeres. En el segundo paso, se proporcionan dos métodos para visualizar los resultados. El primer método utiliza gráfico optimizado de estrellas, que minimiza el solapamiento. En el segundo método, se combinan las técnicas de visualización; coordenadas paralelas, disposición radial para estructuras jerárquicas, y coordenadas paralelas circulares.

En la Figura 1, se puede observar la utilización de coordenadas paralelas enlazadas con visualización radial de árboles jerárquicos. El diagrama de coordenadas paralelas ubicados en la parte derecha, es desplegado interactivamente de la selección de clústeres del lado izquierdo.

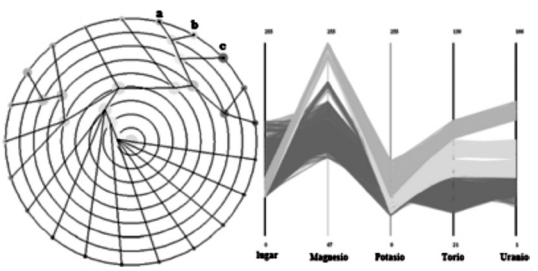

Figura 1. Coordenadas paralelas enlazadas a visualización radial de clústeres jerárquicos.

Lee y otros proponen en [11], un entorno visual interactivo para modelos de clustering, en un contexto de análisis de documentos de textos, y que genera grupos con base en las palabras que más se repiten.

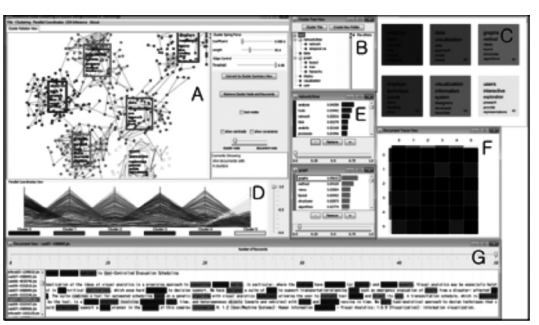

Figura 2. Entorno visual propuesto por Lee y Otros [11].

La Figura 2, muestra la interfaz principal, en la cual existen diversos elementos para el análisis de clúster: coordenadas paralelas, vistas de relación para clústeres, vistas de árboles, entre otros. Este esquema de visualización provee diversos mecanismos de interacción para refinar los resultados de un proceso de clustering, también permite la filtración de datos fuera de rango, y el re-agrupamiento de datos.

En lo que se refiere al uso de métricas, para la comparación de instancias de clústeres, Grabusts en [12] presenta un trabajo donde realiza un análisis de los resultados obtenidos por el algoritmo K-medias, con un conjunto de datos determinado. Para la generación de los clústeres, utiliza las métricas: distancia Euclidiana, distancia de Manhattan y coeficiente correlación de Pearson. Se concluye en este estudio, que los resultados obtenidos son muy similares con cada métrica utilizada, y que la correlación de Pearson, le entrega mejores valores para todos los clústeres.

Una vez realizado el proceso de clustering, es necesario evaluar el resultado que se obtiene para validar la calidad del modelo. El análisis de clúster es considerado como un proceso de aprendizaje no supervisado, es decir el resultado no puede ser comparado con un valor conocido previamente. Por lo tanto las matrices de confusión y las métricas de precisión utilizadas en tareas predictivas, no son aplicables para estos casos. Sin embargo, un método para evaluar el proceso de clustering es a través de criterios de evaluación, los que se dividen en internos y externos. En la literatura existen diversos criterios para la evaluación de clústeres, no obstante, sólo algunos de ellos son aplicados. Entre estos criterios destacan: la suma del error cuadrático (SSE), el índice de Dunn, y el índice de Davies-Bouldin [5].

Por otro lado, se revisaron 5 herramientas de MD, en relación a los modelos de clústeres, y particularmente: el nivel de visualización, si provee métricas, y el nivel de interacción en tres tipos de gráficos seleccionados. Los resultados se pueden observar en la Tabla 1.

Se califica el nivel de visualización, considerando la cantidad de elementos visuales que la herramienta tiene. Si es menor a 2, su calificación es baja, si es igual a 2 se califica regular, y mayor a 2 es buena. Para medir el nivel de interacción en los gráficos, se tiene en cuenta el número de interacciones disponibles, por cada elemento visual: menor a 3 es bajo, entre 3 y 4 regular, y mayor a 4 es alto.

Se puede observar desde la Tabla 1, que R es la herramienta más completa, aunque no dispone del gráfico radar, y que junto a Weka son las únicas que proporcionan métricas para el análisis de clústeres. No obstante, R es un entorno de programación que maneja la generación de modelos y visualización, a través de código con la librería llamada Ggobi [13].

Tabla 1. Características de Herramientas Analizadas.

Herramienta	Nivel de Visualización	Mecanismos de Interacción	Provee Métricas
Orange	Regular	• Coordenadas paralelas: **Regular** • Gráfico de radar: **N/A** • Gráfico de dispersión: **Regular**	No
Weka	Bajo	• Coordenadas paralelas: **N/A** • Gráfico de radar: **N/A** • Gráfico de dispersión: **Regular**	Sí
R	Alto	• Coordenadas paralelas: **Alto** • Gráfico de radar: **Bajo** • Gráfico de dispersión: **Alto**	Sí
Knime	Alto	• Coordenadas paralelas: **Regular** • Gráfico de radar: **Bajo** • Gráfico de dispersión: **Regular**	No
SPSS Modeler	Regular	• Coordenadas paralelas: **Bajo** • Gráfico de radar: **N/A** • Gráfico de dispersión: **Bajo**	No

En lo concerniente al diseño de la propuesta de este trabajo, se basa en dos enfoques:

i) El campo de Analítica Visual descrita por Keim y otros [14], se centra en el proceso de análisis y manejo de grandes volúmenes de datos (heterogéneos y dinámicos), mediante integración del juicio humano sobre representaciones visuales, y mecanismos de interacción. Combina áreas de: visualización, MD, y estadística.

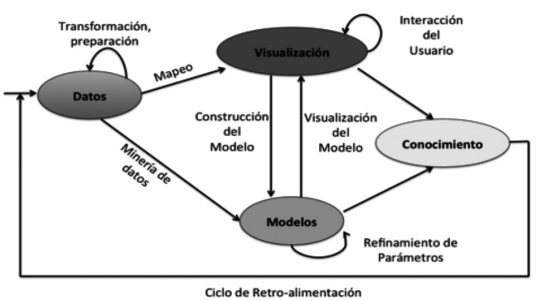

Figura 3. El proceso de AV [14].

Es un ciclo de retro-alimentación que comprende diferentes etapas representadas mediante óvalos, y transiciones (flechas) como se observa en la Figura 3. En este enfoque, visualización es una etapa central del proceso, que se orienta no sólo a la descripción visual de datos (entradas), sino que es fundamental para la construcción del modelo (proceso), y para representar el conocimiento obtenido a través de patrones (salida).

ii) El segundo enfoque es el esquema de Visualización Aumentada para Modelos (VAM), en un proceso de MD, propuesto por Castillo y otros [1].

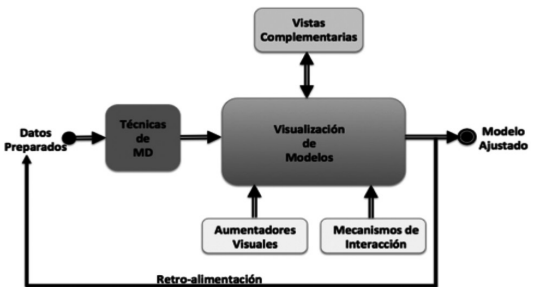

Figura 4. Arquitectura general del esquema VAM [1].

Se puede observar en la Figura 4, que plantea un modelo de percepción visual e interacción del usuario, centrado en la etapa de ajuste del modelo en un proceso de MD, y establece una forma de explorar el modelo original y sus componentes, considerando las características o ejes indicados recientemente.

Este esquema considera como ejes centrales: visualización de modelos para el análisis exploratorio, modelos de percepción visual, mecanismos de interacción, combinación de técnicas de MD, y métricas para comparar las componentes del modelo.

PROPUESTA VIMC

El objetivo de este trabajo de investigación, es diseñar y desarrollar un esquema de Visualización Interactiva para Modelos de Clustering (VIMC), en el contexto de un proceso de MD. Este esquema permite al analista: explorar y analizar visualmente un modelo, en la etapa de su revisión. Está orientado principalmente a tratar con conjuntos de datos de alto volumen y dimensionalidad. Esta basado en dos enfoques: analítica visual [14] y visualización aumentada de modelos [1].

VIMC consta de dos estados: visualización de datos, y visualización del modelo de clústeres, como se representa en la Figura 5.

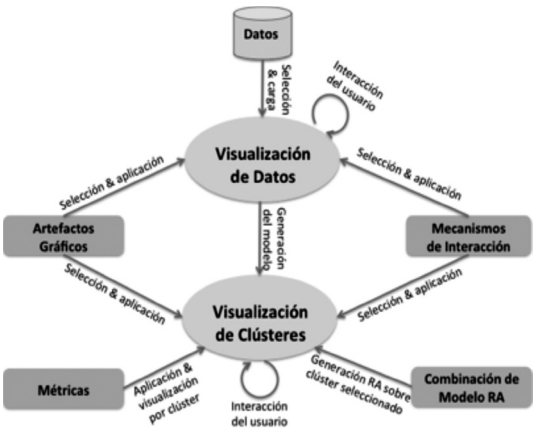

Figura 5. Esquema VIMC (elaboración propia).

Estado 1 (visualización de datos): permite al analista de datos: seleccionar, cargar y presentar el conjunto de datos de entrada o vista minable (archivos CVS). El usuario puede realizar análisis exploratorio sobre los datos, utilizando un conjunto de artefactos gráficos y mecanismos de interacción adecuados. El conjunto de datos debe ser previamente preparado, ya que este esquema VIMC no considera la preparación preliminar de la vista minable. La preparación del conjunto de datos se puede realizar utilizando herramientas de extracción, transformación y carga, conocidas como ETL (Extract, Transform, Load).

Estado 2 (visualización de clústeres): la transición al segundo estado ocurre, cuando el analista de datos requiere la generación del modelo. Para esto, el usuario debe indicar el número de clústeres como parámetro. Además de los artefactos gráficos, y mecanismos de interacción, el usuario puede utilizar métricas para comparar instancias de distintos clústeres, y también para medir la compacidad o dispersión de cada clúster. Adicionalmente, el analista de datos puede generar un modelo de Reglas de Asociación (RA), sobre un clúster seleccionado, con el fin de lograr una descripción que pueda determinar relaciones existentes entre atributos de las instancias, y con esto, puede conocer mejor la conformación de los clústeres.

Para esto, considera en su diseño las siguientes características:

- *Artefactos gráficos:* es necesario contar con un conjunto de artefactos gráficos vinculados, y adecuados para representar modelos de clústeres. Principalmente, orientados a la manipulación de conjunto de datos con gran volumen y alta dimensionalidad. Como resultado del estudio de arte y trabajos relacionados, se logran determinar los elementos visuales más apropiados, para abordar las complejidades señaladas anteriormente, estos son: matriz de diagramas de dispersión, coordenadas paralelas, gráfico de radar y circular.

- *Mecanismos de interacción:* el analista debe poder moverse a través del modelo de clústeres para explorarlo, así como en sus componentes (instancias). También, debe tener la posibilidad de seleccionar instancias, para compararlas entre sí. Para esto, es necesario que el analista pueda interactuar con los artefactos gráficos. Se consideran los mecanismos de interacción que se indican en la Tabla 2.

Tabla 2. Mecanismo de Interacción en VIMC.

Gráficos y sus mecanismos de interacción	Selección y realce	Selección y enlace	Zoom y paneo	Filtro	Reordenar	Detalle
Coordenadas Paralelas	No	No	No	Si	Si	No
Gráfico de Radar	Si	No	No	No	No	Si
Diagrama de Dispersión	Si	Si	Si	Si	No	Si
Gráfico Circular	No	No	No	No	No	No

Combinación de modelo RA: sobre el modelo de clústeres, es posible aplicar la técnica RA, para ampliar la descripción del modelo, y determinar relaciones existentes entre atributos. La razón por la cual se ha seleccionada la técnica RA, radica en que se trata de una técnica descriptiva, que a través de reglas permite tener una mirada adicional de la conformación de los clústeres. Se utiliza para la generación de los clústeres el algoritmo “K-medias”, y el algoritmo “A priori” para generar las RA.

Métricas: los resultados de los procesos de clustering se comparan a través de criterios de evaluación internos, los cuales consideran las distancias: intra-clúster e inter-clúster. Esto permite comparar cuantitativamente instancias de distintos clústeres, y medir el nivel de homogeneidad de los clústeres. Con esto, se puede lograr conocer la conformación

de los grupos generados por el proceso. Desde el estudio realizado sobre métricas para modelos de clústeres, se han seleccionado tres medidas: suma del error cuadrático medio (SSE), índice de Davies-Bouldin (DB-Index), y la distancia euclidiana. Las dos primeras, se implementan para medir el nivel de compacidad o dispersión de los clústeres, y la tercera se utiliza para la generación de los clústeres

ENTORNO VISUAL WEB

El entorno visual implementa el esquema VIMC, bajo una arquitectura cliente-servidor. El lado del cliente, se encarga de generar las visualizaciones e interacciones, y es en el servidor, donde se ejecutan los algoritmos de clustering y RA, y también se calculan las métricas. El entorno visual está desarrollado principalmente utilizando JavaScript tanto en el cliente, como el servidor. Su motor gráfico y de interacciones es D3.js. Para calcular las métricas y realizar las llamadas a los algoritmos de MD (K-medias y Apriori), utiliza Node.js.

En la Figura 6, se presenta la interfaz principal del entorno visual una vez que el usuario ha seleccionado y cargado el conjunto de datos a analizar. Esta parte representa el primer estado del esquema VIMC (visualización de datos). En este estado, el analista puede llevar a cabo un análisis multidimensional de los datos, pudiendo seleccionar múltiples instancias para ser comparadas. El diagrama de dispersión y las coordenadas paralelas están vinculados.

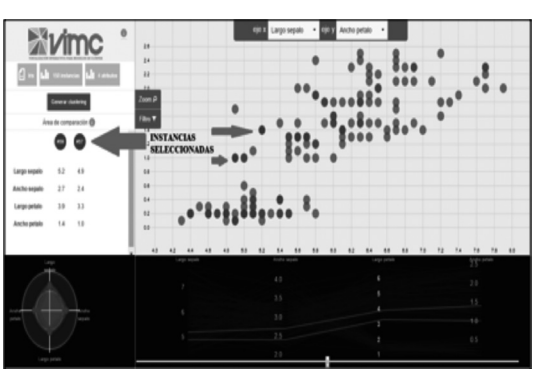

Figura 6. Visualización de datos (estado 1).

Además, en estos gráficos, el usuario puede realizar las interacciones descritas en la Tabla 2. La matriz de dispersión se presenta de a un diagrama, el usuario puede seleccionar los atributos que requiere para su análisis, a través de las coordenadas *(x, y)*, y los gráficos enlazados, se actualizan de manera simultánea en la pantalla.

Se puede observar además en esta misma figura, dos instancias seleccionadas, y que son indicadas con flechas de color verde, en el diagrama de dispersión. También, son destacadas simultáneamente en el gráfico de coordenadas paralelas, que aparece en la parte baja de la pantalla. Los valores de los atributos de las instancias seleccionadas, son agregadas al área de comparación de la parte izquierda de la interfaz, como se indica también con una flecha verde, y son visualizadas en el gráfico de radar ubicado en el vértice inferior izquierdo. De esta forma, es posible comparar cuantitativamente dos o más instancias, con el objetivo de encontrar diferencias o similitudes en sus características.

Una vez que el usuario hace clic en el botón de opción para generar el modelo de clústeres, se pasa al segundo estado en el esquema VIMC (visualización de clústeres), y la interfaz es la que se presenta en la Figura 7. Se puede observar, que aparece una nueva sección en el lado derecho de la pantalla. En esta sección aparecen vistas complementarias que son provistas por: la descripción del clúster junto al valor de sus métricas (SSE y DB-Index), el gráfico de radar, y RA generadas sobre los clústeres.

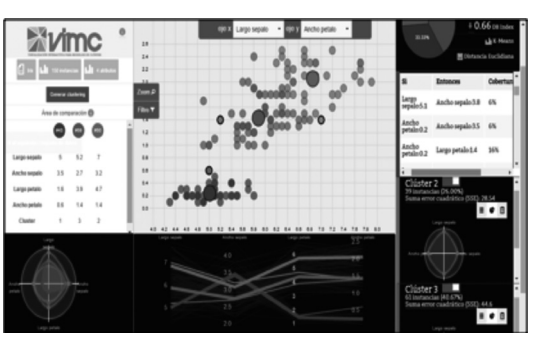

Figura 7. Visualización de clústeres (estado 2).

En este segundo estado, además de los gráficos y mecanismos de interacción, el analista dispone de una vista multidimensional del modelo de clústeres, a través de una matriz de diagramas de dispersión que puede configurar según los atributos que requiera analizar. Como en el primer estado, éste diagrama sigue aún enlazado con el gráfico coordenadas paralelas que es presentado en la parte inferior de la pantalla.

Adicionalmente, el usuario puede obtener vistas complementarias del modelo de clústeres, con la aplicación de una técnica de RA sobre cada clúster, la cual permite describir relaciones existente entre atributos de las instancias. La técnica RA está disponible para cualquier conjunto de datos, y el usuario puede proporcionar como parámetro el porcentaje de confianza y cobertura requerido. También puede agregar un gráfico de radar para cada clúster y con esto, puede complementar el análisis exploratorio de las RA, por ejemplo puede determinar que atributos tienen más peso en un clúster.

VALIDACIÓN Y ANÁLISIS DE RESULTADOS

Experimento Controlado

Para estimar el valor práctico de las características definidas en el esquema VIMC, se ha llevado a cabo un experimento para obtener una percepción subjetiva sobre su utilidad. Para esto, se ha diseñado una tarea de MD, que consiste en generar un modelo de clústeres a partir de un conjunto de datos determinado.

Este experimento se realizó con un universo de 23 personas, los cuales tienen distintos niveles de conocimientos acerca de: procesos de MD, y en el uso de herramientas de MD. La totalidad de los participantes tienen formación en informática a nivel de pre-grado y postgrado.

La experimentación se ha llevado a cabo en 3 etapas:

- En primer lugar, los usuarios acceden a la siguiente URL: http://vimc.inf.unap.cl. En esta dirección, ellos encuentran un tutorial con animación que explica el trabajo de investigación, el acceso al entorno visual, y también la encuesta en línea que deben responder, una vez que analicen el modelo de clústeres presentado.
- *Primera Etapa - Tutorial:* En esta etapa, se explica el trabajo de investigación, y las características principales del esquema VIMC. También, el usuario conoce los aspectos generales del entorno visual.
- *Segunda Etapa - Entrenamiento:* El usuario trabaja con el entorno visual y un conjunto de datos de entrenamiento, donde él puede interactuar libremente generando un modelo de clústeres, comparar instancias, y aplicar RA en el modelo. El objetivo de esta etapa, es familiarizar al usuario con el entorno visual, y la usabilidad que proporciona para: generar, explorar y analizar un modelo de clústeres.
- *Tercera Etapa - Evaluación:* en esta tercera etapa se presenta al usuario, el resultado del proceso de clustering con un segundo conjunto de datos, en el cual se han generado 2 clústeres. Además, ha sido aplicada la técnica RA al primer clúster, y dos elementos han sido seleccionados (uno de cada clúster). Esta etapa tiene como objetivo, que el usuario debe observar y analizar el modelo de clústeres presentado, para luego responder una encuesta en línea, la cual permite evaluar el esquema VIMC y el entorno visual.

El conjunto de datos utilizado en las etapas 1 y 2, es conocido como Iris, y contiene 150 registros de 3 tipos de flores. En donde los primeros 50 registros son del tipo Setosa, los siguientes 50 registros son Virgínica, y los últimos 50 registros son Versicolor. Cada registro tiene 4 atributos (longitud del pétalo, longitud del sépalo, anchura del pétalo, y anchura del sépalo).

Para la etapa 3, se utilizó el conjunto de datos llamado Vino. Contiene 1000 registros de vinos con 12 atributos (acidez fija, acidez volátil, ácido cítrico, azúcar residual, cloruros, dióxido de azufre libre, dióxido de azufre total, densidad, pH, sulfatos, y grado de alcohol). No obstante, el entorno visual permite manejar cualquier conjunto de datos, y sólo se requiere que esté en el formato abierto CSV.

La encuesta diseñada para la validación del trabajo, consiste en 22 preguntas divididas en 6 secciones: información del participante, visualizaciones, análisis sobre el modelo de clústeres, análisis de RA, comparación de objetos, y sobre el esquema VIMC.

Evaluación de Resultados

La totalidad de los participantes en la experimentación, expresan una alta valoración del esquema VIMC y su entorno visual. Respecto a las tres características del esquema propuesto, se puede observar desde el gráfico de la Figura 8, que la mayoría ha evaluado a la herramienta como muy buena con un 70%, y buena con un 30%, en relación al nivel de interacciones que proporcionan sus elementos visuales. La combinación de modelos de MD es bien considerada por los usuarios, con un 30%

muy buena y 57% buena, y esta misma valoración obtienen las métricas que permiten comparar la compacidad y dispersión de los clústeres.

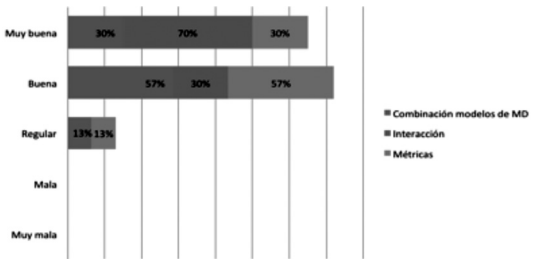

Figura 8. Valoración de características de VIMC.

En relación al nivel de utilidad de los elementos visuales disponibles, para describir tanto datos como instancias en los clústeres, se puede observar desde el gráfico de la Figura 9, que los usuarios entregan la más alta valoración para el diagrama de dispersión, con un 87% de apreciación positiva (61% muy buena y 26% buena), en segundo lugar, al gráfico radar, con un 73% (dividida en 43% muy buena y 30% buena). Y en tercer lugar, las coordenadas paralelas, con un 61% de valoración positiva (9% muy buena y 52% buena), y que además, obtienen la mayor valoración negativa (26% regular y 13% mala).

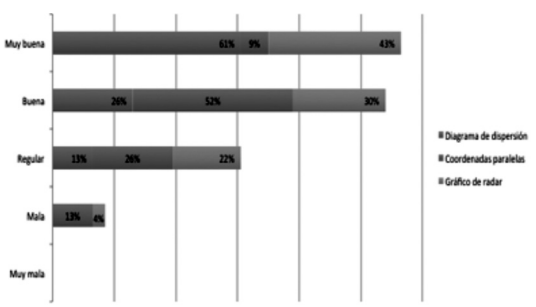

Figura 9. Nivel de utilidad de los elementos visuales.

Al ser consultados los participantes, si la generación de un modelo de RA sobre un clúster, ayuda a mejorar la comprensión de éstos, un 95% afirma que sí, mientras que sólo un 4% señala lo contrario. Cuando se les pregunta; si la comparación de instancias ayuda a comprender de mejor forma como se componen los clústeres, un 35% responde estar totalmente de acuerdo, un 57% está de acuerdo, y sólo un 9% no está de acuerdo. En cuanto a las métricas disponibles, un 30% de los usuarios considera que su utilidad es muy buena, un 57% cree que es buena, sólo un 13% piensa que es regular, y ninguno piensa que es mala o muy mala.

Una de las tareas del experimento tiene por objetivo, validar el análisis visual que consigue el usuario con la herramienta. Y para esto, el usuario debe observar dos clústeres, y luego determinar cuál de los dos es más disperso. Su respuesta es comparada con la métrica (SSE), que determina empíricamente la dispersión de los clústeres. Se obtiene como resultado que el 91% de los participantes, ha podido responder correctamente que clúster tiene mayor dispersión, y que corresponde al segundo clúster presentado. Esto es corroborado con su valor para SSE de 15.057, frente al primer clúster que sólo tiene un valor para SSE de 10.806.

Finalmente, frente a la pregunta: si el esquema VIMC, apoya la comprensión de un modelo de clústeres; el 48% de los participantes responde estar totalmente de acuerdo, y un 52% está de acuerdo. Esto significa que la totalidad de los participantes, evalúan positivamente el esquema visual propuesto.

CONCLUSIONES Y TRABAJO FUTURO

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

- En general, respecto a los resultados del trabajo de investigación realizado, estos son muy alentadores ya que la mayoría de los participantes en el experimento, dan una valoración positiva sobre las características del esquema VIMC.
- El entorno visual logra una evaluación positiva en cuanto a: usabilidad, mecanismos de interacción, y gráficos implementados. Estos últimos, permiten un análisis exploratorio apropiado del conjunto de datos, y también para un modelo de clúster con datos multidimensionales. Destaca la alta valoración de las interacciones implementadas en cada elemento visual, y también entre los artefactos gráficos, el diagrama de dispersión es el que obtiene mejor evaluación.
- Respecto a la métricas seleccionadas e implementadas, si bien sus evaluaciones son buenas, no son las que se esperaban, por lo que se considera una oportunidad de mejora para una próxima versión del entorno visual,

principalmente en lo que se relaciona con visualización de estas métricas.

- En lo concerniente a, combinar técnicas de MD, aunque es valorada positivamente la aplicación de RA sobre cada clúster, y que sirven para mejorar la descripción de los grupos, la idea es mejorar y evaluar su representación visual, de modo que sean más entendibles por parte de los usuarios.
- Finalmente, el esquema VIMC a través de su entorno visual, logra el objetivo de apoyar en la comprensión de un modelo de clústeres. El análisis visual del usuario en la experimentación, fue corroborado con métricas, y ambas medidas coinciden, por ejemplo en reconocer si un clúster es más compacto que otro.

En relación al trabajo futuro, este tiene relación con los siguientes aspectos:

- Evaluar otras técnicas descriptivas de MD, que aporten visualizaciones adicionales a las RA. También, visualizar modelos generados por la técnica RA.
- Seleccionar y aplicar nuevos artefactos gráficos, que puedan ser de mayor utilidad en la exploración de los datos, y de los clústeres, preferentemente para conjunto de datos con alta dimensionalidad.
- Por el momento, VIMC utiliza el algoritmo K-medias con la distancia euclidiana. Este algoritmo trabaja sólo con conjuntos de datos numéricos. Se propone incorporar otros algoritmos de clustering, y con otras medidas de distancia.
- Una mejora interesante, es integrar todos los algoritmos de MD que proporciona Weka.
- Profundizar en aspectos de percepción visual, explorando otras técnicas como: análisis de texturas, utilización de glifos, etc.
- Evaluar la implementación de otras métricas, que permitan comparar clústeres y sus componentes.
- Se considera explorar el nivel de rendimiento de los algortimos de MD y de visualización, utilizados en la implementación del entorno visual web.

REFERENCIAS

[1] W. Castillo-Rojas, C. Meneses and F. Medina. Augmented visualization for data-mining models. Journal: Elsevier Procedia Computer Science. ISSN: 1877-0509, DOI: 10.1016/j.procs.2015.07.063, volumen 55, pp. 650-659, July 2015.

[2] C.J. Meneses and G.G. Grinstein. Visualization for Enhancing the Data Mining Process. In Proceedings of the Data Mining and Knowledge Discovery: Theory, Tools, and Technology III Conference. Orlando, FL., 2001.

[3] D. Keim, J. Kohlhammer, E. Geoffrey and F. Mansmann. Mastering the Information Age Solving Problems with Visual Analytics. Edited by the authors Published by the Eurographics Association Postfach 8043, 38621 Goslar, Printed in Germany, Druckhaus Thomas Müntzer GmbH, Bad Langensalza. Theoretical Issues in Ergonomics Science. Vol. 8, Nº 1, ISBN 978-3-905673-77-7, 2010.

[4] W. Castillo-Rojas and C. Meneses. A Comparative Review of Schemes of Multidimensional Visualization for Data Mining Techniques. III Congreso Internacional de Computación e Informática del Norte de Chile (INFONOR-CHILE). Arica - Chile, 2012.

[5] 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. ISBN 978-0-387-09822-7, 2010.

[6] C. Fraley and A. Raftery. Model-based methods of classification: Using the mclust Software in Chemometrics. University of Washington Seattle, United States. DOI: 10.18637/jss.v018.i06, 2007.

[7] J. Han and M. Kamber. Data Mining: Concepts and Techniques, 2a ed., Estados Unidos, Elsevier. ISBN-10: 1558609016. ISBN-13: 9781558609013, 2006.

[8] P.E. Hoffman and G.G. Grinstein. A Survey of Visualizations for High-Dimensional Data Mining, in: Fayyad U., Grinstein G.G., Wierse A. (eds.), Information Visualisation in Data Mining and Knowledge Discovery, Morgan Kaufmann Publishers, San Francisco, pp. 47-85, 2002.

[9] A.K. Jain, M.N. Murty and P.J. Flynn. Data Clustering: A Review, ACM Computing Surveys, Vol. 31., Nº 3., pp. 264-323, 1999.

[10] T.V. Long. Visualizing High-density Clusters in Multidimensional Data. Bremen, Germany. DOI 10.1007/s00180-011-0271-3, 2011.

[11] H. Lee, J. Kihm, J. Choo, J. Stasko and H. Park. iVisClustering: An interactive visual document clustering via topic modeling. In Computer Graphics Forum (Vol. 31, Nº 3 pt 3, pp. 1155-1164). Blackwell Publishing Ltd., June 2012.

[12] P. Grabusts The choice of metrics for clustering algorithms. Letonia. ISSN 1691-5402, ISBN 978-9984-44-071-2, 2011.

[13] C. Dianne and F.S Deborah. Interactive and Dynamic Graphics for Data Analysis: With R and GGobi, ISBN 978-0-387-71762-3, 2007.

[14] D. Keim, J. Kohlhammer, E. Geoffrey F. and Mansmann. Mastering the Information Age Solving Problems with Visual Analytics. Edited by the authors Published by the Eurographics Association Postfach 8043, 38621 Goslar, Printed in Germany, Druckhaus Thomas Müntzer GmbH, Bad Langensalza. Theoretical Issues in Ergonomics Science. Vol. 8 Nº 1, ISBN 978-3-905673-77-7, 2010.