Universidad de Medellín
Ciencia y Libertad

Visualización de conjunto de datos de múltiples instancias

Jorge Eliecer Valencia Duque

Universidad de Medellín
Facultad de Ingenierías, Maestría en Ingeniería de Software
Medellín, Colombia
2019

Visualización de conjunto de datos de múltiples instancias

Jorge Eliecer Valencia Duque

Trabajo de grado presentado como requisito parcial para optar al título de:
Magister en Ingeniería de Software

Director:
PhD. Carlos Andrés Mera Banguero
Co-Directora:
PhD. Lina María Sepúlveda Cano

Línea Temática:
Visualización y detección de patrones en grandes conjuntos de datos

Universidad de Medellín
Facultad de Ingenierías, Maestría en Ingeniería de Software
Medellín, Colombia
2019

Agradecimientos

Después de largas jornadas estudiando por fin ha llegado el día, este pequeño apartado es el espacio que tengo para agradecer a todas las personas que me ayudaron y apoyaron para finalizar mi trabajo de grado. Ha sido un período de aprendizaje largo tanto a nivel personal como académico como bien muchos lo saben. Además el impacto que ha tenido en mi ha sido gratificante.

Primero que todo quisiera agradecer a mis asesores Carlos Andrés Mera Banguero y Lina María Sepúlveda Cano, por el gran apoyo y ayuda que han sido para darle un enfoque correcto y guiarme para entregar siempre con la máxima calidad. Particularmente Carlos fue un gran apoyo durante la investigación y una fuente de conocimiento que me permitía seguir avanzando.

Además, me gustaría darles las gracias a las personas que validaron la propuesta final, por su valiosa retroalimentación. Definitivamente, esta me ayudó a comprender que este es el inicio y que aún faltan caminos que explorar.

También me gustaría agradecer a mis padres por su apoyo y comprensión. Siempre han estado ahí para darme una voz de aliento en los momentos que desfallecía.

Finalmente, mis amigos y a todas aquellas personas que de una forma u otra permitieron que con sus palabras darme una voz de aliento o simplemente distraerme de la difícil tarea que fue escribir este trabajo de grado.

¡Muchas gracias a todos!

Resumen

Este trabajo de grado aborda la problemática de la visualización de conjuntos de datos de múltiples instancias (MI), en busca de entender las particularidades de estos conjuntos de datos y sus relaciones. Como en la literatura existen pocos trabajos relacionados a este tema, se considera que el resultado puede ser de utilidad para quienes actualmente trabajan con el paradigma de aprendizaje de múltiples instancias (MIL). Así, la intención de este trabajo es desarrollar un método de visualización que permita a los usuarios entender cuales son las relaciones o patrones ocultos en los conjuntos de datos de MI. Con este fin se plantea una pregunta de investigación importante, ¿Que métodos de visualización se pueden adaptar para explorar conjuntos de datos de MI?. La respuesta a la pregunta de investigación se busca mediante la creación de una propuesta de visualización y experimentando con diferentes métodos de visualización en los conjuntos de datos. La propuesta de visualización final se validó mediante encuestas y cuestionarios a expertos en MIL además con pruebas y comparaciones internas. Los experimentos realizados mostraron que usar métodos combinados de visualización permite extraer más información del conjunto de datos. Teniendo esto en cuenta y siguiendo las recomendaciones de los expertos, seria bueno crear herramientas que permitan representar un conjunto de MI en diferentes métodos de visualización y a su vez hacer herramientas más intuitivas, para que el proceso de visualización de datos sea más rápido y efectivo en la detección de patrones.

Palabras clave: aprendizaje de múltiples instancias, Visualización, Representación, Análisis visual, MIL.

Abstract

This degree work addresses the problem of the visualization of data sets of multiple instances (MI), seeking to understand the particularities of these data sets and their relationships. As there are few works related to this topic in the literature, it is considered that the result may be useful for those who currently work with the multi-instance learning paradigm (MIL). Thus, the intention of this work is to develop a visualization method that allows users to understand what are the relationships or hidden patterns in MI data sets. To this end, an important research question is posed, what visualization methods can be adapted to explore MI data sets? The answer to the research question is sought by creating a visualization proposal and experimenting with different visualization methods on the data sets. The final visualization proposal was validated through surveys and questionnaires to MIL experts in addition to internal tests and comparisons. The experiments carried out showed that using combined visualization methods allows extracting more information from the data set. Taking this into account and following the recommendations of the experts, it would be good to create tools that allow representing a set of MI in different visualization methods and in turn make more intuitive tools, so that the data visualization process is faster and more effective in pattern detection.

Keywords: multi-instances learnning, Visualization, Representation, Visual Analysis, MIL.

Contenido

Lista de acrónimos

Tabla 0-1.: Lista de acrónimos

Acrónimo	
MIL	Multi-Instance Learning
MI	Multi-instance
IS	Instance Space
BS	Bag Space
ES	Embedded Space
SVM	Support Vector Machine
MI-SVN	Multi-Instance Support Vector Machine
SMIL	Sparce Multi-Instance Learning
kNN	k-nearest neighbors
MIL-MFS	Multiple-Instance Learningwith Multiple Feature Selection
MILES	Multiple-Instance Learning via Embedded instance Selection
RSL	Revisión sistemática de Literatura
SimpleMIL	Simple Multi-Instance Learning
KPCA	Kernel principal component analysis
FastICA	Fast independent com-ponent analysis
Isomap	Isometric Feature Mapping
LLE	Locally linear em-bedding
MDS	Multidimensional scaling
TSNE	t-distributed Stochastic NeighborEmbedding
PCA	Principal Component Analysis

1. Introducción

La presente investigación trata sobre la problemática de la visualización de conjuntos de datos de múltiples instancias (MI), que son los conjuntos multidimensionales usados en el paradigma de aprendizaje de múltiples instancia (*Multi-Instance Learning*, MIL). Su uso es muy similar a los sistemas de aprendizaje tradicional con la diferencia que este tipo de conjuntos de datos pueden contener más información y más detalles de los objetos reales a los cuales referencia.

Sus características o estructura es más compleja que los conjuntos de datos que tradicionalmente se usan en problemas de clasificación. Un conjunto de datos de MI se basa en bolsas (*bags*) e instancias (*instances*) y cada instancia tiene múltiples características. Las bolsas pueden ser positivas o negativas; una bolsa es negativa cuando todas las instancias de la bolsa son negativas, por el contrario es positiva cuando al menos una instancia es positiva, esto se conoce como el supuesto estándar de MIL [1, 2]. Las bolsas contienen múltiples instancias y el problema radica en que no se conoce la etiqueta de las instancias individuales, solo se conoce la etiqueta de la bolsa que las contiene.

Para analizar el problema de visualización es necesario conocer la estructura de los conjuntos de datos de MI e identificar las relaciones que existen entre los diferentes componentes de los conjuntos de datos; como la relación entre las bolsas e instancias positivas.

La investigación de esta problemática de visualización estuvo motivada por el interés de brindar una herramienta que permita a los investigadores del paradigma MIL a entender cuales son las relaciones o patrones ocultos en los conjuntos de datos de MI. Por otra parte, la tesis doctoral de Carlos Mera titulada "Detección de Defectos en Sistemas de Inspección Visual Automática a través del Aprendizaje de Múltiples Instancias"[2] fue la principal impulsora para evidenciar la problemática de métodos de visualización para este tipo de conjuntos de datos.

En el marco del área temática de la visualización de conjuntos de datos multidimensionales, se siguen una serie de fases para realizar una representación visual de los datos, estos pasos definieron el marco metodológico usado para proponer una visualización de conjuntos de datos de MI. Estas fases consta de etapas de transformación de los datos donde se puede analizar que características de los conjuntos de datos de MI se representaran en la visualización.

Al tener una propuesta de visualización, se continuó con la etapa de validación interna

y externa, mediante juicio de expertos, la utilidad y usabilidad del sistema de visualización. Para la evaluación externa, se realizaron una serie de encuestas y evaluaciones sobre el sistema a un grupo de investigadores que trabajan con el paradigma de MIL. Durante la validación, uno de los principales inconvenientes fue encontrar expertos en el tema dado que pocos investigadores de áreas afines al aprendizaje maquina trabajan con algoritmos MIL. En cuanto a la evaluación interna, se realizaron pruebas con diferentes tipos de conjuntos de datos de MI, cambiando las condiciones de reducción de información y selección de características, para posteriormente compararlos con datos sin procesar y realizando una clasificación con el conjunto resultante.

El documento se encuentra distribuido de la siguiente forma, el Capítulo 1 se encuentra la introducción. El Capítulo 2 tenemos la etapa de contextualización donde se explica y se justificara con mas detalle la problemática abordada. Luego en el Capítulo 3 se analizan los conceptos claves para el desarrollo del proyecto además del estado actual de la visualización de conjuntos de datos de MI. Seguido del Capítulo 4 donde se desarrolla la metodología planteada para lograr los objetivos propuestos. Para sintetizar, el Capítulo 5 presenta los resultados obtenidos de las pruebas realizadas tanto internas como externas. Por ultimo el Capítulo 6 presenta las conclusiones del desarrollo de la investigación.

2. Contextualización

2.1. Introducción

El presente capítulo tiene como objetivo contextualizar en el problema de investigación del trabajo de grado, desde lo global a lo particular, definiendo los aspectos a resolver, la justificación, la pregunta de investigación, las hipótesis y los objetivos.

2.2. Contextualización del problema

El aprendizaje de múltiples instancias es un paradigma de clasificación supervisado en el que, en el contexto del reconocimiento de patrones, los objetos complejos pueden ser representados usando múltiples vectores de características [3]. Esta peculiaridad diferencia a este paradigma del aprendizaje supervisado tradicional, en el que los objetos sólo se pueden representar por un único vector de características. Así, el paradigma de aprendizaje de múltiples instancias, llamado MIL por sus siglas en inglés (*Multiple Instance Learning*), proporciona un marco de trabajo que permite preservar mayor cantidad de información de los objetos que se desean reconocer. Los conjuntos de datos de múltiples instancias poseen una estructura compleja por la cual se representan los objetos, esta estructura esta conformada por conjuntos (denominados bolsas) que contienen múltiples vectores de características (llamados instancias) que representan cada una de sus partes.

La forma de representación de los objetos en MIL genera conjuntos de datos a los que se les denomina conjuntos de datos de múltiples instancias (MI). Este tipo de conjunto de datos tiene una complejidad de exploración mayor, comparados con los conjuntos de datos tradicionales usados en el aprendizaje supervisado. Esto se debe a que además de interpretar las relaciones que existen entre las instancias individuales, se deben representar las relaciones que existen entre las bolsas, es decir, las relaciones entre los conjuntos de instancias. De esta forma se agrega un grado de complejidad adicional cuando se quiere visualizar este tipo de conjuntos de datos.

Por otro lado, la visualización de información es un área de investigación que tiene como objetivo ayudar a los usuarios a explorar y analizar datos a través de representaciones visuales e interactivas de los mismos [4, 5]. Por su potencial para ayudar a descubrir patrones ocultos, la visualización de información ha estado ganando importancia dentro de todo tipo

de organizaciones, convirtiéndose en una herramienta esencial para la toma de decisiones [6, 7]. En general, son diversos los campos de aplicación de la visualización de información, por ejemplo, en el análisis de datos de mercados para detectar anomalías o tendencias de mercados bursátiles [8]; en el análisis de imágenes y vídeos para reconocer personas y detectar patrones en sus comportamientos [9], incluso para ayudar en el diagnóstico de enfermedades mediante el reconocimiento de anomalías en las radiografías; en el análisis de datos deportivos a fin de estimar el rendimiento de los jugadores durante un encuentro determinado [10]; entre otras.

Un aspecto importante que tienen en común todas las aplicaciones en las que se usa la visualización de información es que los datos a explorar o a analizar provienen de conjuntos de datos multidimensionales, es decir, son conjuntos de datos cuya estructura está definida por múltiples características (o dimensiones) que describen aquello que están representando [11, 12].

En la literatura existen diferentes métodos para visualizar conjuntos de datos multidimensionales [4]. Dichos métodos se pueden agrupar de diversas maneras, por ejemplo, de acuerdo a qué tan comunes son estos se pueden clasificar como tradicionales y no tradicionales. Dentro de los métodos tradicionales de visualización se incluyen gráficos de barras [4], histogramas [13, 14], diagramas de pastel [4, 14], diagramas de burbujas [4] y las infografías [4]; mientras que del lado de las técnicas de visualización no tradicionales se consideran estrategias como los diagramas de dispersión [4, 11, 15, 16, 14], las técnicas de visualización basadas en grafos [4, 17], diagramas de coordenadas en estrella [7], las técnicas basadas en orientación de píxeles [4, 13, 15] y los diagramas de coordenadas paralelas [13, 15, 16, 14, 17], entre otros.

2.3. Planteamiento del problema

Si bien las técnicas de visualización mencionadas anteriormente han sido adaptadas para explorar y analizar conjuntos de datos multidimensionales tradicionales, su uso ha sido poco explorado para representar y visualizar conjuntos de datos de múltiples instancias. Esto se debe, primero, a que este tipo de conjuntos son aún relativamente jóvenes comparados con los conjuntos de datos tradicionales y, segundo, porque estos son más complejos de visualizar dado que, como se mencionó antes, no solo deben representar visualmente las relaciones entre las instancias individuales (como se hace en un conjunto de datos multidimensional tradicional), sino que también se deben mostrar las relaciones entre las bolsas que las contienen.

Con base en lo anterior, en este trabajo de investigación se busca proponer o adaptar una estrategia de visualización de información para la visualización y exploración de conjuntos de datos MI.

2.4. Pregunta de investigación

Con base en el problema planteado, en este trabajo se busca dar respuesta a las siguientes preguntas de investigación:

- ¿Qué técnicas de visualización pueden ser usadas y/o adaptadas para visualizar y explorar conjuntos de datos MI?
- ¿Cómo se puede sintetizar la información de las bolsas en los conjuntos de datos de MI a fin de visualizar las relaciones entre estas?

2.5. Hipótesis

Para el trabajo de investigación se planteó las siguientes hipótesis:

H1: La técnica de visualización propuesta será más efectiva para identificar las relaciones o patrones en los conjuntos de datos de múltiples instancias comparado con otros métodos de visualización usados anteriormente.

2.6. Alcance

Dentro de los alcances de este trabajo cabe anotar que se desarrollará o se adaptará una técnica de visualización para representar un conjunto de datos MI. Dicha técnica abarca la definición tanto de un método de proyección de los conjuntos de datos MI a dos o tres dimensiones como el uso de una técnica de visualización para crear la representación visual de estos.

2.7. Justificación

La visualización de información ha cobrado relevancia dentro de las organizaciones y los entornos académicos debido a la masiva generación de datos que día a día se producen. Muchos de esos datos provienen de encuestas, experimentos o de las simples transacciones que a diario hacen las personas en los supermercados y bancos. Otras fuentes de datos incluyen los diferentes sensores y cámaras que ahora hacen parte de nuestro entorno y nuestro diario vivir. A pesar del volumen de datos que se están generando, estos no proporcionan información por sí solos, ya que requieren ser analizados para generar nuevo conocimiento, para tomar decisiones respecto a lo que representan o para darles un significado que permita su comprensión de manera profunda [18].

Actualmente, las propuestas existentes para la visualización de conjuntos de datos no han

considerado los conjuntos de datos MI. Como se mencionó, en la literatura existen diferentes métodos para visualizar datos multidimensionales tradicionales, pero poco se ha estudiado acerca de la visualización de conjuntos de datos MI. Tan solo unos pocos trabajos abordan este tema, pero de manera tangencial, centrándose específicamente en los algoritmos de aprendizaje o en las aplicaciones del paradigma MIL.

La diversidad y complejidad de los datos multidimensionales [17], y en consecuencia los conjuntos de datos MI, abren la posibilidad, para proponer o adaptar las técnicas de visualización existentes de manera que permitan alizar y tomar decisiones sobre conjuntos de datos MI. En este orden de ideas se busca determinar si es posible representar conjuntos de datos de MI mediante la adaptación de los métodos de visualización de datos multidimensionales de los que disponemos hoy en día.

2.8. Objetivos

2.8.1. Objetivo general

Adaptar o desarrollar una técnica para la visualización de conjuntos de datos de múltiples instancias con el fin de permitir su exploración y análisis.

2.8.2. Objetivos específicos

1. Caracterizar y comparar diferentes propuestas de visualización de conjuntos de datos multidimensionales para establecer cuál o cuáles podrían adaptarse mejor a los conjuntos de datos MI y a las características que se desean representar de los mismos.
2. Establecer una estrategia de visualización, con base en la caracterización de los métodos existentes, para desarrollar una técnica que permita la exploración de conjuntos de datos MI.
3. Validar la técnica de visualización propuesta utilizando conjuntos de datos MI para verificar la efectividad de la estrategia de visualización implementada.

3. Fundamentación teórica

Este capítulo se presentarán los conceptos relacionados con el aprendizaje de múltiples instancias (MIL), así como las problemáticas y los usos frecuentes de este paradigma. Además, se podrá entender cómo la visualización de un conjunto de datos de MI puede ayudar a la comprensión más profunda de la relación de los datos qué se analicen y de la estructura interna de los conjuntos de datos de MI.

Por otra parte, se tratará de responder las preguntas de investigación planteadas en el Capítulo 2, y encaminar la investigación para alcanzar los objetivos planteados.

3.1. El paradigma del aprendizaje de múltiples instancias (MIL)

El aprendizaje de múltiples instancias es un nuevo paradigma de aprendizaje supervisado [19] en el qué los objetos son representados por conjuntos (llamados bolsas) de vectores de características (o instancias) [2] lo qué permite describir objetos complejos de manera más precisa.

Formalmente en MIL, una bolsa $B_i = \{x_{i1}, ..., x_{in_i}\}$ contiene n_i instancias, para el espacio de características $\mathbb{R}^d$ el cual se conoce como espacio de instancias [1]. Cada instancia $x_{ij} \in B_i$ con $1 \leq j \leq n_i$, es un vector d-dimensional de múltiples características que se extrae de una parte del objeto [2]. Aunque una instancia contenga todas las características que describen al objeto no necesariamente todas contienen información importante, es decir existen instancias que no proporcionan información sobre el objeto [1]. Basado en esto, el propósito de los algoritmos MIL es aprender una función de conjuntos de instancias (bolsa) al conjunto de etiquetas de clase.

$$F(B) : B \longmapsto \Omega \tag{3-1}$$

Para aprender esta función, es necesario un conjunto de entrenamiento de múltiples instancias con N bolsas: $T = \{(B_1, y_1), ..., (B_n, y_n)\}$, donde cada bolsa B_i esta asociada a una etiqueta de clase $y_i \in \Omega$ indicando la clase a la que pertenece. La mayoría de problemas MIL tienen dos clases, $y_i = -1$ para las bolsas negativas y $y_i = +1$ para las positivas. En el supuesto estándar de MIL una bolsa es positiva si al menos contiene una instancia positiva

[20]. De otra forma, la bolsa es negativa si todas las instancias que contiene son negativas. De acuerdo a lo anterior, los clasificadores MIL toman la siguiente forma:

$$F(B_i) = \begin{cases} +1, & \text{if } \exists \mathbf{x_{ij}} \in B_i | f(\mathbf{x_{ij}}) = +1 \\ -1 & \text{otherwise} \end{cases} \tag{3-2}$$

Note que F depende del clasificador de instancia $f : \mathbb{R}^d \longmapsto \Omega$. De esta forma aunque los datos de entrenamiento son pasados como un conjunto de bolsas el clasificador f debe aprender a nivel de instancia usando las etiquetas de la bolsa [21]. Si se usa una función de probabilidad, F podría definirse sobre la probabilidad posterior de la bolsa como:

$$\begin{aligned} F(B_i) &= \underset{y \in \Omega}{argmax} Pr(y|B_i), \end{aligned} \tag{3-3}$$

Donde $Pr(y|B_i)$ se obtiene combinando las probabilidades posteriores de las instancias en la bolsa [22]. Usando la regla del promedio se obtiene:

$$Pr(y = +1|B_i) = \frac{1}{n_i} \sum_{j=1}^{n_i} Pr(y_{ij} = +1|\mathbf{x_{ij}}) \tag{3-4}$$

$$Pr(y = -1|B_i) = \frac{1}{n_i} \sum_{j=1}^{n_i} (1 - Pr(y_{ij} = +1|\mathbf{x_{ij}})) \tag{3-5}$$

Que representa el supuesto colectivo, donde todas las instancias en la bolsa contribuyen de igual manera a definir su etiqueta [19].

En general, el proceso de aprendizaje de un algoritmo de clasificación MIL consiste en construir un modelo, a partir de un conjunto de bolsas de entrenamiento, para aprender a predecir las etiquetas de clase de nuevas bolsas. La Figura **3-1** ilustra la diferencia entre el aprendizaje supervisado estándar y MIL.

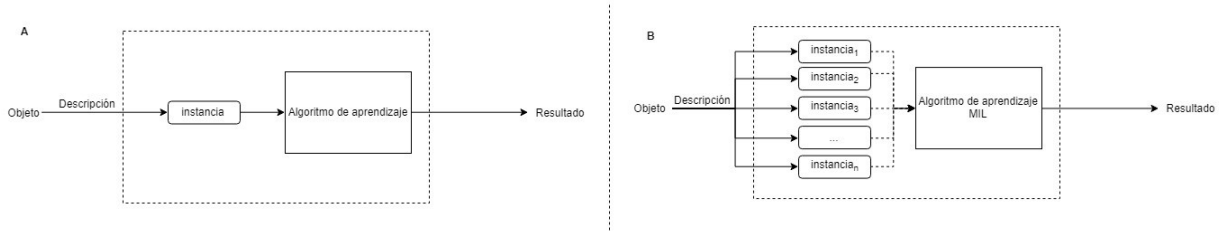


Figura 3-1.: **A.** Escenario tradicional de aprendizaje maquina supervisado. **B.** Aprendizaje de múltiples instancias. Imagen basada en la Figura 1 de [19]

Para ilustrar mejor el concepto se puede usar el problema simple del carcelero descrito por Chevaleyre y Zucker [23]. En este problema se considera una puerta cerrada con N llaveros,

cada uno con un grupo de llaves. Si un llavero (es decir, una bolsa) contiene una llave (es decir, una instancia) qué puede desbloquear la puerta, se considera qué ese llavero es útil. Así, el problema del aprendizaje en MIL consiste en construir un modelo qué pueda predecir si un determinado llavero es útil o no [23].

De acuerdo con Amores [1], los algoritmos MIL pueden ser agrupados en tres familias, cada una de las cuales usa una aproximación diferente, con el fin de generar una función para distinguir entre bolsas positivas y bolsas negativas. Con base en dicha agrupación un algoritmo MIL puede estar basado en instancias (IS), basado en bolsas (BS) o basado en espacios embebidos o transformados (ES) [1, 2].

La familia de algoritmos IS busca predecir las etiquetas de clase de una bolsa a partir de las etiquetas de sus instancias. Para ello, estos algoritmos consideran diferentes supuestos, siendo los más usados el supuesto estándar y el supuesto colectivo [19]. El primero de ellos establece qué una bolsa se etiqueta como positiva cuando al menos una de las instancias es positiva. Por otro lado, una bolsa se etiqueta negativa cuando todas sus instancias son negativas. El segundo supuesto determina qué la etiqueta de una bolsa se define con base en la información combinada de todas sus instancias, por ejemplo, usando algún tipo de ponderación de las etiquetas de clase de todas las instancias [1, 2]. Algunos algoritmos MIL qué hacen uso de este paradigma son *Multiple-Instance Support Vector Machines* (MI-SVM) y *Sparce MIL* (SMIL); qué están basados en máquinas de vectores de soporte (*Support Vector Machine*, SVM) [1, 19].

Los algoritmos qué pertenecen a la familia BS, toman en cuenta las bolsas como un todo y definen medidas de similitud o de distancia entre bolsas, lo qué les permite definir relaciones espaciales entre bolsas y clases [3]. Además, no hay suposiciones sobre las instancias de las bolsas como sucede en IS. Comúnmente, los algoritmos de clasificación de esta familia son clasificadores basados en distancias como el algoritmo de k vecinos más cercanos (k-NN), cuya versión en MIL se denomina *Citation k-NN* [1, 19].

La familia de algoritmos ES usa funciones de transformación donde a partir de las instancias de una bolsa generan un nuevo vector de características, proyectando la bolsa a un nuevo espacio dimensional. Convirtiendo el problema MIL en un problema de aprendizaje supervisado estándar [2]. Por ejemplo, algunos de los algoritmos MIL qué emplean este paradigma son el MIL-MFS (*Multiple-Instance Learning with Multiple Feature Selection*), basado en el aprendizaje de *multi kernel* [19]; otro algoritmo de esta familia es MILES (*Multiple-Instance Learning via Embedded instance Selection*), el cuál convierte el problema MIL en un problema de aprendizaje supervisado estándar [24].

De acuerdo con Amores en [1], entre las tres familias de algoritmos, los basados en instancias presenta una menor precisión comparado con los algoritmos en las familias BS y ES en situaciones donde la clasificación se realiza a nivel de bolsas.

Finalmente, los algoritmos MIL se pueden usar en diferentes industrias y problemas en

donde se maneja gran cantidad de información qué debe ser clasificada. Algunos de sus usos han sido en la clasificación de imágenes médicas, como en la clasificación y detección de regiones cancerígenas en imágenes de histopatología, un uso similar ha sido aplicar el paradigma MIL en imágenes de resonancia magnética para la detección de Alzheimer. Por otro lado, también ha sido usado en problemas de clasificación y de recuperación de imágenes y en el seguimiento de objetos, en el cual MIL genera bolsas positivas en el objeto de interés y bolsas negativas en el resto de objetos en la imagen [2, 25]. Además es usado en la inspección visual automática con el objetivo de mejorar la flexibilidad de este tipo de sistemas, permitiendo una mejora en el etiquetado y clasificación de los objetos de interés [2].

3.2. Métodos de visualización de información

El auge del análisis de grandes cantidades de datos ha provocado un mejor entendimiento de la importancia de la información. Debido a esto, las herramientas qué permiten gestionar y aprovechar estos datos para comprenderlos de manera profunda han ganado relevancia. En consecuencia, las técnicas de visualización de información se han tornado valiosas para las empresas ya qué permiten mejorar procesos impulsando la innovación e incrementando la productividad y obtención de mayores crecimientos, entre otros.

La visualización de información es una respuesta a una serie de retos acerca de cómo representar, interpretar y comparar datos de forma rápida y efectiva. Adicionalmente, permite detectar patrones ocultos en los datos con el fin de brindar información útil para las personas. La visualización de datos se puede presentar de diferentes maneras: estática o interactiva, dependiendo de las necesidades qué se tengan. Esto permite mejorar la toma de decisiones y reducir los tiempos de respuesta a problemas puntuales [6].

Uno de los objetivos de los métodos de visualización es representar los datos en un espacio de pocas dimensiones, usualmente dos o tres dimensiones, con una estructura qué permita conservar la mayor cantidad de información original [18]. Otro de los objetivos es ayudar a los usuarios a explorar, comprender y analizar los datos a través de la exploración visual interactiva [5].

Para qué los datos puedan ser visualizados es esencial contar con una estructura qué permita transformarlos en información para el usuario. Usualmente, esta estructura se constituye en cinco etapas: transformación de datos y análisis, filtración, proyección, representación o renderizado y controles de interfaz de usuario [5]. Aunque no solo estas etapas son suficientes para lograr una visualización también se debe tener en cuenta la estructura interna de los datos.

Según como estén organizados los datos pueden ser clasificados en estructurados y no estructurados. En los datos no estructurados se aplican una serie de técnicas qué permiten

convertirlos en información estructurada qué puede contener menos ruido, debido a qué ha sido filtrada [5]. Posterior a la etapa de filtrado se tiene el módulo qué convierte los datos en figuras geométricas primitivas, como lo son los puntos o líneas, qué luego en la etapa de renderizado se representan en una imagen para qué el usuario pueda interactuar con ella mediante la interfaz de usuario [6]. Lo anterior se puede ilustrar mejor mediante la Figura **3-2**.

Actualmente existen una variedad de métodos y técnicas qué hacen posible representar los datos en un formato qué mejora su entendimiento. Algunas de las técnicas más representativas de visualización de información son:

- Métodos geométricos
  - Diagramas de dispersión
  - Matrices de diagrama de dispersión
  - Gráficos multilínea
  - Curvas de Andrew
  - Coordenadas paralelas
  - Visualizaciones radiales (GridViz, PolyViz, RadViz)
- Iconográficos
  - Caras de Chernoff
  - Diagramas en estrella
- Jerárquicos
  - Apilamientos dimensionales
  - Coordenadas paralelas jerárquicas

Una de las mejores maneras de explorar qué método de visualización puede representar de la mejor forma un conjunto de datos es realizar pruebas con datos cuya estructura sea conocida [18], de esta manera se minimiza el margen de error qué se pueda presentar.

3.3. Estado del arte

Para recopilar el estado del arte del tema de este trabajo de grado se realizó una Revisión Sistemática de Literatura (RSL) sobre la visualización de conjuntos de datos de MI. Después, se analizaron los artículos encontrados para determinar qué métodos de visualización son los más usados en los conjuntos de MI y con qué frecuencia se usan para escoger un método de clasificación MIL o con otro propósito. A continuación, se exploraron las soluciones qué se han propuesto para la visualización de información multidimensional e indagar si existen propuestas para visualizar conjuntos de datos de MI; aunque existen varias técnicas de

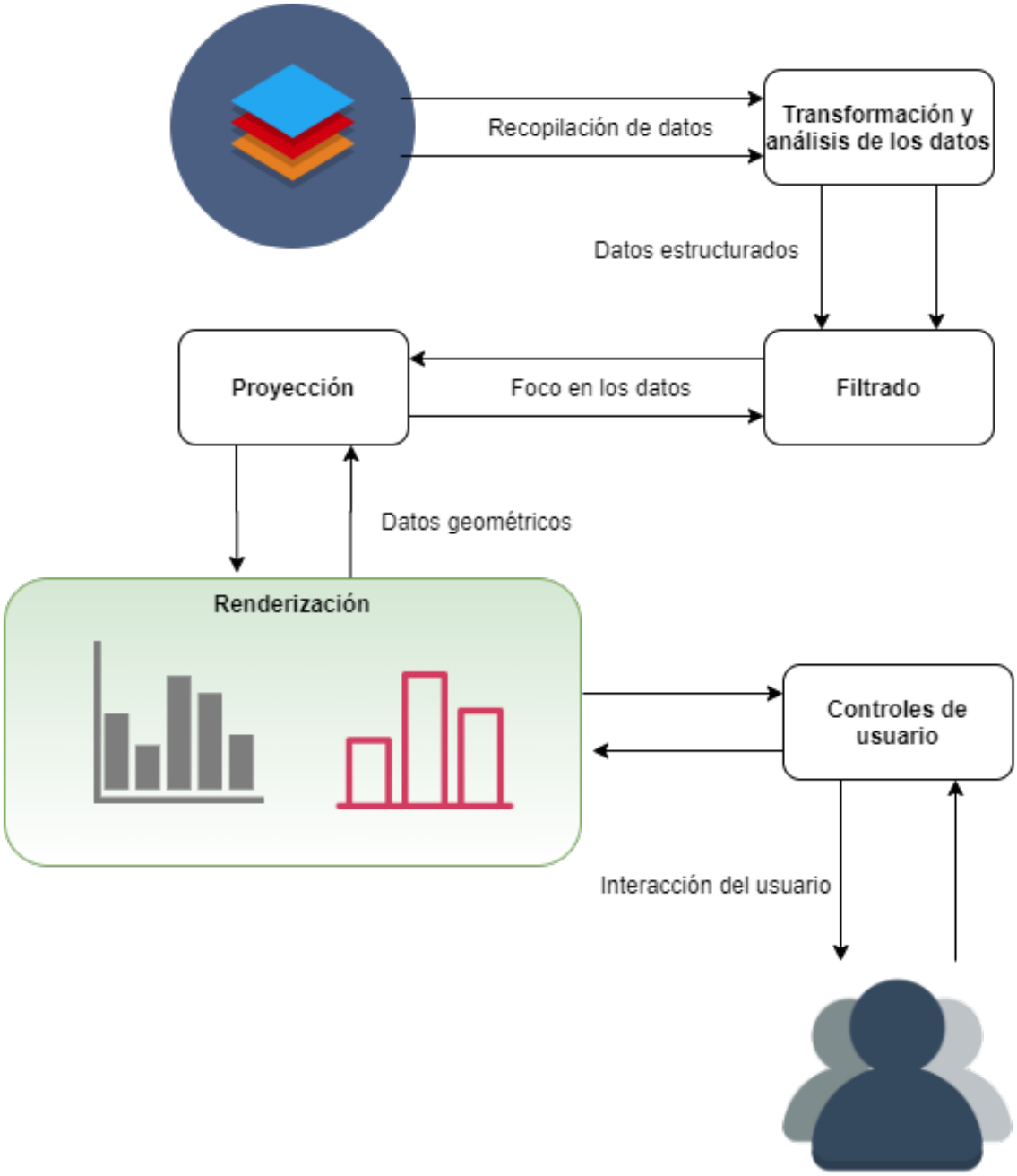


Figura 3-2.: Etapas de la visualización: las principales etapas de la estructura de visualización de información son transformación y análisis, filtrado, proyección, renderización y controles de usuario, tomado de Figura 1 de [5]

visualización de información, muchos de los métodos aún no son suficientes para mostrar la información oculta en los datos [26].

A continuación, se muestra el análisis del estado actual de la visualización de información de MI. El resto del capítulo esta organizado como sigue: exposición de los materiales y métodos usados para intentar responder las preguntas de investigación planteadas. Luego, se usa la información extraída de los artículos seleccionados y se presentan los resultados obtenidos. Por último, se plantea una discusión acerca del tema abordado y las conclusiones correspondientes.

3.3.1. Metodología para el desarrollo de la RSL

Se ha llevado a cabo esta RSL usando la metodología sugerida por [27]. La revisión involucra tres etapas básicas con diferentes actividades.

1. Planeación de la revisión: para definir las necesidades de la RSL, se plantean las preguntas de investigación, y se define el protocolo de revisión; tales como fuentes de datos, estrategia y términos de búsqueda, selección de estudios, extracción y síntesis de datos.
2. Realización de la revisión: selección y revisión de los estudios; para responder las preguntas de investigación; y para presentar los resultados, discusiones y conclusiones.
3. Documentación de la revisión: presentación de los resultados de la revisión en un documento final.

El objetivo de seguir esta metodología es tener una guía clara acerca de los estudios secundarios acerca del tema qué se quiere investigar, además de proveer una serie de actividades fácilmente replicables para obtener los mismos resultados [28].

Planeación de la revisión

Preguntas Se han planteado dos preguntas qué permitirán tener un conocimiento de como se encuentra el estado actual la visualización de conjuntos de datos MI. Igualmente se plantea hacer una revisión de las posibles propuestas nuevas. Dichas preguntas se enumeran a continuación:

- **PI1** ¿Qué propuestas se han realizado para la visualización de datos de MI?
- **PI2** ¿Qué técnicas de visualización pueden ser efectivas para la visualización de conjuntos de datos MI?

Fuentes de datos Las bases de datos electrónicas usadas en esta RSL son Google Scholar para hacer la búsqueda inicial de artículos sobre el tema de visualización de conjuntos de

datos de MI; ScienceDirect qué es una fuente con amplia cantidad de artículos en el tema de visualización de información y ACM, ya qué es la base de datos especializada en ciencias de la computación, por último, se usó IEEE qué es otra fuente con información relevante en temas de visualización de información.

Estrategia de búsqueda y términos Los términos de búsqueda incluyen «visualization» y «multi-instance». También se usaron términos relacionados con los procesos y métodos de visualización, por ejemplo, «chart», «visual analytic» y «visualize». Finalmente, se agregaron términos relacionados con el tratamiento de conjuntos de datos, por ejemplo, «multidimensional». Los detalles de las cadenas de búsqueda usadas se muestran en la Tabla **3-1**.

Tabla 3-1.: Cadenas de búsqueda usadas en las bases de datos científicas.

ID	Cadenas generales de búsqueda
C1	multi-instance AND visual analytics
C2	multi-instance AND information AND (graph OR chart)
C3	multi-instance AND representation
C4	multi-instance AND (visual OR visualization OR visualize)
C5	(multi-instance OR multiple-instance) AND multidimensional AND (visual OR visualization OR visualizing)
C6	(multi-instance OR multiple-instance) AND multidimensional

Criterios de selección de estudios Los criterios de inclusión y exclusión que se definieron son los siguientes:

- Trabajos de investigación.
- Estudios qué contengan mención de técnicas de visualización en conjuntos de datos de MI.
- Estudios qué proporcionen información acerca de alguna aproximación o método de visualización sobre datos complejos.
- Estudios publicados ente 2007 y 2019.
- Trabajos escritos en Inglés o Español.

Los siguientes artículos y trabajos fueron excluidos de la revisión:

- Artículos qué no expliquen como llegar a una representación de los datos o qué no indiquen que técnicas fueron usadas.

- Artículos qué no mencionen o hagan alución al paradigma MIL.
- Artículos qué no mencionen ninguna técnica de visualización en su contenido.

Extracción de información La extracción de información se realizó por medio de un formulario el cual agrupaba y resumía la información de cada artículo encontrado. El formulario contiene la siguiente información de cada artículo:: autores, año de publicación, título, fuente, problema qué se plantea resolver en el artículo, técnicas o métodos de visualización propuestos, algoritmos MIL mencionados, retos de la visualización, así como los problemas o inconvenientes encontrados y por último un breve párrafo sobre el artículo y los resultados obtenidos (Tabla **3-2**).

3.3.2. Desarrollo de la revisión

Usando los criterios de inclusión y exclusión se obtuvieron un total de 45 artículos qué fueron identificados como relevantes para la revisión actual. Algunos artículos se referían al mismo documento o eran basados en investigaciones similares, motivo por el cual fueron descartados. En la Figura **3-3** se provee un diagrama de flujo en el cual se muestran los resultados del proceso de selección.

Después de hacer la búsqueda inicial y aplicar los criterios de inclusión, se redujo significativamente el numero de artículos encontrados, esto muestra también que existen pocos estudios que abarquen el tema de la visualización de conjuntos de datos de MI.

3.3.3. Resultados

Esta sección resume los resultados obtenidos al llevar a cabo la RSL. El análisis de los resultados se centra en responder las preguntas de investigación PI1 y PI2 (Preguntas 3.3.1).

Como ya se había mencionado para la RSL solo se contó con 45 artículos, los cuales se listan en el Anexo E, estos fueron ordenados por año de publicación.

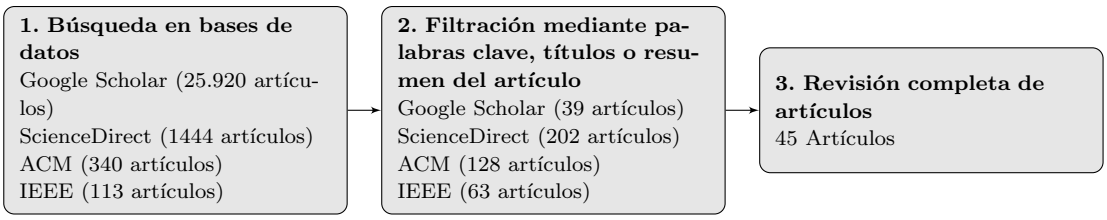


Figura 3-3.: Diagrama de flujo con el resultado del proceso de selección de artículo

Tabla 3-2.: Elementos de extracción para cada estudio, relacionado con las preguntas de investigación

ID	Elemento	Descripción	Pregunta de investigación
E1	Título	Título del documento	Información General
E2	Fuente	Base de datos donde fue extraído el documento	Información General
E3	Autores	Autor(es) del documento	Información General
E4	Año de publicación	En qué año fue publicado el estudio	Información General
E5	Problema qué se pretende resolver	Se quiere conocer la problemática qué se pretende resolver en el artículo	P1, P2
E6	Técnicas o métodos de visualización propuestos o usados	Qué métodos o técnicas de visualización son usados para representar la información extraída	P1, P2
E7	Familias MIL qué se quisieron representar	Qué métodos, frameworks o algoritmos MIL son mencionados en el artículo y a cuáles de las tres divisiones de MIL pertenece	P1
E8	Qué desafíos o retos se presentan en la visualización de conjuntos de datos de MI	Se pretende encontrar los vacíos y retos qué se encuentran dentro de la visualización en MI	P1, P2
E9	Resultados subjetivos	Punto de vista acerca del documento y la información qué contiene	Información adicional

El análisis de los artículos seleccionados permitió caracterizar los problemas asociados a los conjuntos de datos MIL. En general, se identificaron 4 categorías de problemas relacionadas con el. En la Figura **3-4** se muestra cada uno de los problemas encontrados.

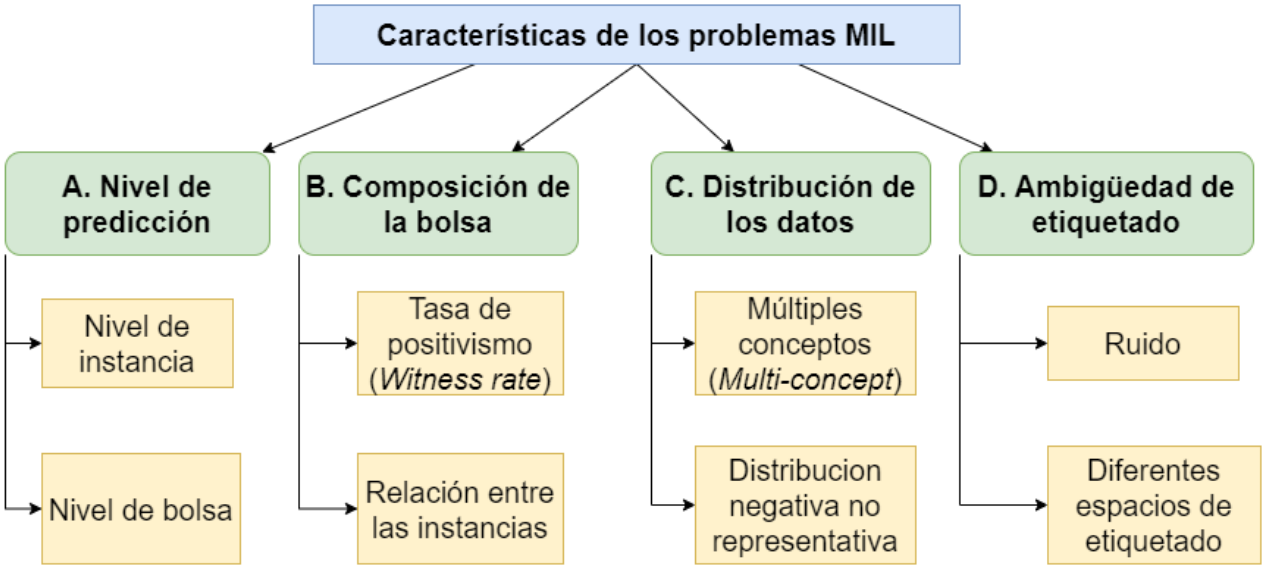


Figura 3-4.: Características de los problemas MIL

Nivel de predicción, de la composición de la bolsa, de la distribución de los datos y de la ambigüedad en el etiquetado [29, 1].

A. Nivel de predicción

Las características de los conjuntos de datos de MI están relacionadas con el algoritmo qué se ajusta mejor con el problema de clasificación. En ese contexto existen don niveles: nivel de instancia (*Instance-Level*) y nivel de bolsa (*Bag-Level*) [29, 1].

Esto tiene además relación con los dos supuestos en MIL; el supuesto estándar y el supuesto colectivo. El primero define la etiqueta de una bolsa enunciando qué muy pocas instancias aportan a la etiqueta de la bolsa [30]; en otras palabras, para qué la bolsa sea positiva por lo menos debe tener una instancia positiva [1]. El supuesto colectivo, por otra parte, considera qué todas las instancias contribuyen de igual forma a la etiqueta de la bolsa [30, 1], es decir, para qué una bolsa sea positiva se asume qué todas las instancias tienen características positivas [29].

En la literatura existe una larga lista de algoritmos MIL qué bajo el supuesto estándar, realizan las clasificación y obtienen un nivel de precisión diferente en ambos niveles [29, 31]; esto dice mucho en cuanto al conjunto en sí, porque permite saber bajo qué nivel (instancias o bolsas) de clasificación podría funcionar mejor una visualización. Conocer bajo que nivel de clasificación el conjunto de datos arroja mejores resultados nos da información acerca de

la distribución de las instancias positivas en el conjunto de datos, siendo útil para usarlo como base en la propuesta de visualización.

B. Composición de la bolsa

La composición de la bolsa puede afectar el rendimiento en los algoritmos de clasificación MIL. La métrica qué se usa para medir la composición de las bolsas se conoce como *witness rate* o la proporción de instancias positivas entre las bolsas positivas [29]. Es por eso qué los métodos de clasificación qué analizan la distribución de las instancias en una bolsa se ven seriamente afectados, debido a qué la distribución de las bolsas positivas y negativas cuando el *witness rate* es bajo son muy similares.

Por otra parte, muchos métodos de clasificación MIL asumen qué la distribución de las bolsas positivas es independiente de la distribución de las bolsas negativas. Este supuesto no considera la relación con las bolsas. Sin embargo, en el mundo real este tipo de situaciones puede variar, es por ello qué es importante tener en cuenta que las distribuciones con conjuntos de datos mas complejos puede no ser la esperada [32, 10].

Lo anterior resulta importante para analizar un conjunto de datos y determinar cuál podría ser el mejor algoritmo de clasificación qué se puede usar con dicho conjunto de datos. Adicionalmente, puede ser de gran ayuda qué un método de visualización de datos MI permita determinar el *witness rate* para facilitar la toma de una decisión en cuanto a qué algoritmo MIL aplicar en la clasificación.

C. Distribución de los datos

Muchos algoritmos MIL hacen suposiciones sobre la forma de las distribuciones o sobre que tan bien la distribución negativa se encuentra representada en el conjunto de entrenamiento. Algunos algoritmos MIL por ejemplo suponen que las instancias positivas se encuentran en una sola región en el espacio de características, esto es conocido como el supuesto de cluster único. Lo anterior puede ser razonable en algunos contextos pero perjudiciales en otros, como cuando el conjunto de datos tiene características variadas dado que es poco probable que cluster único funcione correctamente [29].

Otro factor qué puede afectar la distribución de los datos es la distribución negativa no representativa, es decir, en un conjunto de datos de MI el conjunto de entrenamiento no representa realmente los datos negativos en un conjunto objetivo. Hay varios métodos MIL que pueden funcionar con diferentes distribuciones negativas en el conjunto de entrenamiento, para ello en la mayoría de los casos estos métodos buscan una región que represente el concepto positivo en los datos y todo aquello que se encuentre lejos de esta región se considera negativo [29]. Esto aunque puede afectar la clasificación, puede ser mitigado mediante la selección de un conjuntos de entrenamiento que tomen en cuenta la composiciones de las bolsas [33].

Al conocer mejor la distribución de los datos en el conjunto de datos, se puede proponer

métodos de visualización no convencionales y qué proporcionen una ayuda para distinguir cómo puede ser tratado el conjunto de datos de entrenamiento.

D. Ambigüedad del etiquetado

Esta característica la posee la mayoría de conjuntos de datos. Sin embargo, en MIL puede llegar a ser un poco más compleja debido a la estructura de los datos [34]. Por ejemplo, la poca claridad en las etiquetas qué se puede dar en diferentes niveles dependiendo si se mira el supuesto estándar o el supuesto colectivo. En el supuesto estándar se mitiga este problema puesto qué se asegura qué las instancias negativas solo contienen información negativa no obstante, en conjuntos de datos más complejos como imágenes, no es posible sostener este tipo de afirmaciones [29].

Otro factor qué puede afectar el etiquetado es la similitud de las características en el espacio de instancias, es decir, las instancias pueden presentar características similares al de otro grupo de instancias qué se etiqueten como positivas y por su parecido sean etiquetadas erróneamente como negativas [35, 36].

La ambigüedad en el etiquetado y la complejidad de los datos de MI, pueden dificultar la propuesta de un método de visualización qué permita representar este tipo de problemas en un conjunto de datos de MI. Sin embargo, sería ideal encontrar un método qué permita evidenciarlo en algún nivel.

Análisis de información de la RSL

De acuerdo al análisis de la RSL se encontraron varios algoritmos los cuales se abordan en este documento. Estos algoritmos se pueden agrupar en tres grandes familias [25, 8], como se mencionó en el Capítulo 2. La Figura **3-5**, muestra la agrupación de los algoritmos mencionados en la RSL y su tipo, donde convencional significa que es un algoritmo comúnmente usado por los investigadores y es bien conocido su funcionamiento, indeterminados son aquellos que no se encontró información detallada acerca de su funcionamiento y propuesta son los que en el articulo usan el paradigma MIL desde otro enfoque diferente al tradicional.

En su mayoría, se evidencia los algoritmos MIL mencionados en los artículos de la RSL pertenecen a la familia de algoritmos basados en instancias, en contraste, los algoritmos embebidos no son demasiados ya qué es una clasificación relativamente nueva. Por último, la revisión evidencia que hay un crecimiento en el número de algoritmos MIL basados en bolsas. Esto quizá se debe a que tienden a ser más precisos por mantener el contexto de la relación entre las bolsas, comparados con los algoritmos basados en instancias [29].

Por otro lado, entre los algoritmos más usados se encuentran *mi-SVM* y *MI-SVM*, así como *SimpleMIL* y *Citation-kNN*, es importante conocerlos para poder tener claro con qué algoritmos se pueden realizar validaciones considerando qué son los más usados por los investigadores.

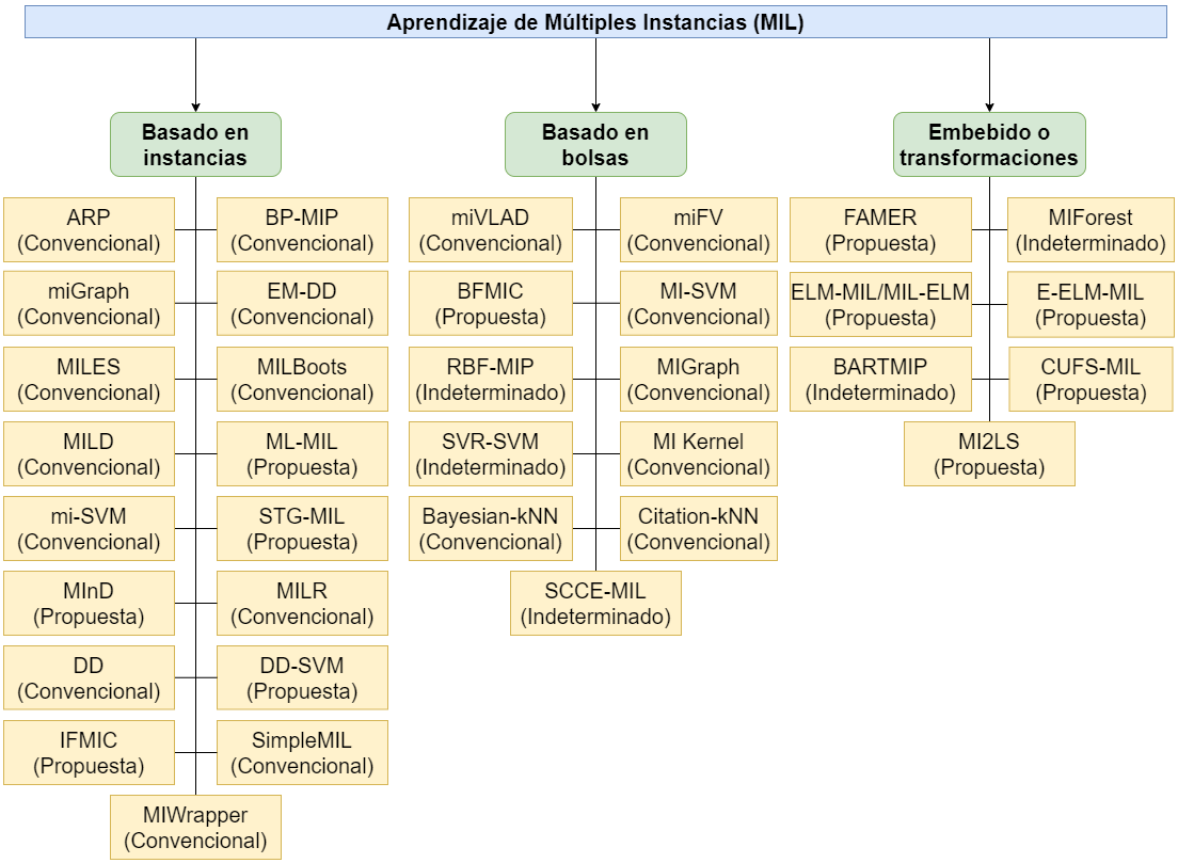


Figura 3-5.: Algoritmos MIL agrupados por familia.

En cuanto a las técnicas de visualización son muy pocos los artículos qué usan una técnica para analizar los conjuntos de datos, también se encontró qué los pocos documentos qué usan alguna técnica, usualmente lo hacen para ilustrar mediante conjuntos de datos sintéticos (en su mayoría conjuntos Gaussianos) cuál es el comportamiento esperado de la clasificación [37, 38].

La Figura **3-6** muestra lo poco qué se trata el tema de visualización de los conjuntos de datos de MI, ya qué solo el 25 % de los artículos usan alguna técnica de visualización; el restante 75 % no presenta ninguna técnica de visualización, aunque en algunos casos se intenta describir la estructura interna y complejidad de los conjuntos de datos MI que usan. Lo anterior permite concluir que en muchos de los artículos seleccionados no se presenta una técnica de visualización clara para tratar la problemática de la representación de conjuntos de datos de MI, aunque se puede resaltar qué algunos de ellos lo intentan mediante técnicas poco convencionales como los grafos, e intentan representar las relaciones qué hay en los datos. Además, queda claro qué la visualización de conjuntos de datos de MI aún es un tema qué se debe explorar más a profundidad.

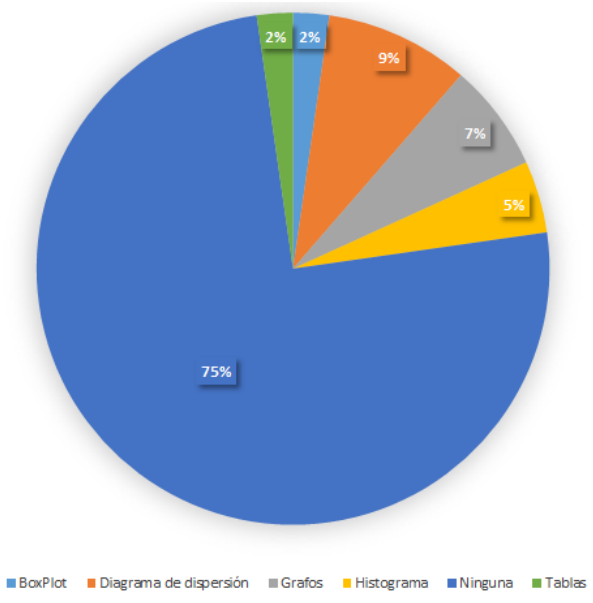


Figura 3-6.: Técnicas de visualización usadas en conjuntos de datos de MI.

3.3.4. Discusión problemas abiertos

En esta revisión de literatura permitió identificar cuáles son los métodos de clasificación MIL más frecuentemente usados. Adicionalmente, se pudo observar qué actualmente se trabaja en el mejoramiento de estos métodos, bien ya sea mediante la unión con otros algoritmos de aprendizaje o tratando de explorar nuevas soluciones. Igualmente, se encontró que para

comprobar la eficiencia de los métodos propuestos por los investigadores sobre mejoras en los algoritmos MIL, son comunes las comparaciones contra los algoritmos de clasificación tradicionales del paradigma MIL.

Por otra parte, se evidenció que es recurrente el uso de los mismos conjuntos de datos para validar los nuevos métodos de clasificación que se proponen en el paradigma MIL. Estos son: *Musk1* [20], *Musk2* [20], *Tiger* [39] y *Corel* [40], entre otros. Esta situación puede llevar a un sesgo la validación porque la distribución y la estructura de estos conjuntos de datos ya es bien conocida.

Adicionalmente, se pudo observar qué en la exploración de los artículos seleccionados se encuentran muy pocas técnicas de visualización qué permitan representar los conjuntos de datos de MI, sin embargo, es satisfactorio ver qué en algunos casos se explora la representación de los datos de MI mediante grafos y diagramas de dispersión qué ayudan a la comprensión de información multidimensional que en general presentan una alta complejidad. Cabe decir qué las pocas técnicas de visualización usadas son aplicadas sobre conjuntos de datos artificiales o semi-artificiales creados por los investigadores.

Otro tema a resaltar es qué muchas veces los resultados obtenidos para la visualización son dados por la combinación de diferentes técnicas qué permiten mejorar la forma en la qué se muestran los datos, así se logran cubrir las deficiencias de un método de visualización acoplándolo con alguno qué lo complemente. De igual manera, se puede observar qué con los algoritmos MIL pasa algo similar, teniendo en cuenta qué muchos de ellos han surgido gracias a la combinación de varias técnicas qué por sí solas presentarían algunas debilidades, pero en conjunto se complementan.

Para finalizar, se puede ver una relación entre las problemáticas MIL y los algoritmos propuestos para solucionarlos. Esto nos lleva a considerar que igualmente puede haber una relación entre estos problemas y las técnica de visualización. Es decir, se podrán explorar diferentes técnicas de visualización qué puedan mostrar de mejor forma el *witness rate* de un conjunto de datos y evidenciar de otra manera el ruido o la distribución de los datos, de esta forma tener una visión general del conjunto de datos de MI.

3.3.5. Conclusión

Los resultados de esta revisión sistemática de literatura muestran qué los métodos más usados en MIL son los llamados "tradicionales" por algunos autores, entre ellos se encuentran *mi-SVM*, *Citation-kNN*, *MILES*, entre otros. Estos métodos son usados en la mayoría de casos de estudio de MIL para contrastar con los métodos propuestos recientemente, dado qué son la base para verificar qué se hacen mejoras y avances sobre el paradigma MIL.

A partir de la RSL es posible concluir qué los métodos de visualización multidimensionales en conjuntos de datos de MI no han tenido un foco de atención en la mayoría de las

investigaciones. Esto evidencia qué hay una línea de investigación abierta en este tema.

4. Marco Metodológico

En este capítulo se detalla la metodología usada para el desarrollo de una propuesta de visualización de conjuntos de datos de MI. Para ello, en cada sección se describen las actividades implementadas en cada fase y se explica como estas ayudaron a alcanzar el objetivo general propuesto en este trabajo de maestría.

4.1. Descripción del área de estudio

El área de estudio principal es la visualización de conjuntos de datos multidimensionales, específicamente la visualización de conjuntos de datos de estructuras complejas, como los son los conjunto de datos del paradigma MIL. En general, las herramientas de visualización son ampliamente usadas para extraer información de de un conjunto de datos; sin embargo, la construcción de esas herramientas es un reto en lo que respecta a como representar efectiva y visualmente, en un plano 2D o 3D, un conjunto de datos n-dimensional [4].

4.2. Metodología de investigación

Durante el desarrollo de este trabajo se realizaron numerosas pruebas de concepto, no obstante, siempre se consideraron las fases que Liu *et al.* propusieron para la construcción de sistemas de visualización de datos multidimensionales [5]. En cada prueba de concepto se desarrollaron diferentes actividades para dar cumplimiento a los objetivos específicos planteados. Así, la adopción de la metodología permitió seguir unos pasos claros que fueron la hoja de ruta principal para la implementación de la herramienta de visualización que se propone en este trabajo.

4.3. Desarrollo del marco metodológico

Para el cumplimiento de los objetivos específicos se consideraron 5 fases, cada una de las cuales tiene sus propias actividades específicas a desarrollar. La Figura **4-1** ilustra, de manera general, la metodología que se siguió para el desarrollo de este trabajo. Cómo puede observarse, la Fase 4 se dividió en 4 sub-fases; cada una con sus propias actividades específicas.

En el resto del Capítulo, al iniciar cada fase o sub-fase se listan las actividades desarrolladas y se describe el proceso que se adoptó en cada una.

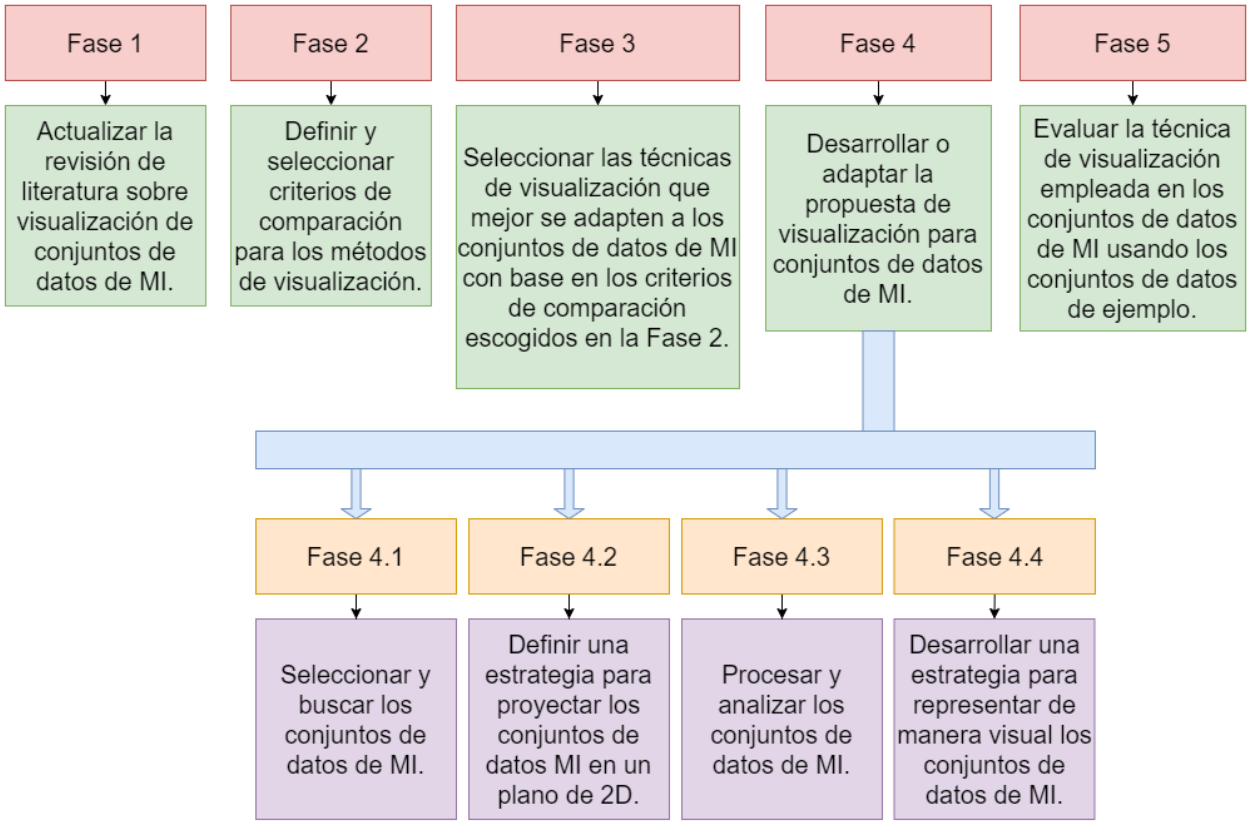


Figura 4-1.: Fases de la metodología usada.

4.3.1. Fase 1: Revisión de literatura sobre visualización de conjuntos de datos de MI

Esta fase consideró las siguientes actividades:

- Actualizar y evaluar las cadenas de búsqueda
- Seleccionar estudios primarios actuales
- Identificar artículos relevantes acerca de la visualización de conjuntos de datos de MI
- Extraer información relevante de los nuevos estudios primarios seleccionados
- Sintetizar y presentar los datos obtenidos

El objetivo de estas actividades fue conocer el estado del arte y las diferentes propuestas existentes para la visualización de conjuntos de datos de MI, así como obtener sugerencias

de qué técnicas y métodos de visualización podrían ser adaptados para usarse con conjuntos de MI. En el Capítulo 3 se desarrollaron estas actividades y se describió la información encontrada tras la revisión.

4.3.2. Fase 2: Definir y seleccionar los criterios de comparación para los métodos de visualización

En esta fase se desarrollaron las siguientes actividades:

- Definir la lista de métodos y estrategias de visualización que puedan ser usadas para visualizar conjuntos de datos multidimensionales
- Seleccionar y establecer los criterios que debe tener un método visualización de conjuntos de datos de MI para facilitar su exploración

Estas actividades fueron ejecutadas de manera simultanea con la revisión de la literatura. No obstante, al finalizar la Fase 4.3.1 se lograron identificar las estrategias que podían ser empleadas para crear una representación efectiva de conjuntos de datos multidimensionales complejos, como lo son los conjuntos de datos MI.

Si bien la lista de métodos de visualización para representar conjuntos multidimensionales es bastante amplia, estos se pueden agrupar de acuerdo con sus características relevantes. A modo de resumen, se listan los métodos que son usados en la literatura. Dicha lista agrupa los métodos de acuerdo a su características predominantes. Algunos de estos métodos son explicados más adelante en la Sección 4.3.3.

- Proyecciones geométricas [4].
  - Diagramas de dispersión (*Scatterplot*) [41]
  - Coordenadas Paralelas (*Parallel Coordinates*) [42]
  - Curva de Andrews (*Andrews Curve*) [43]
  - Visualización de coordenadas radiales (*Radical Coordinates Visualization*) [4]
  - Coordenadas de estrella (*Star Coordinates*) [44]
  - *Table Lens* [45]
- Técnicas orientadas a los pixeles (*Pixel-Oriented Techniques*) [4].
  - Espacio de llenado de la curva (*Space filling curve*) [46]
  - Patrón Recursivo (*Recursive Pattern*) [47]
  - Técnica de espiral y ejes (*Spiral and Axes Techniques*) [48]
  - Segmento Círculo (*Circle Segment*)[49]
  - Gráfico de barras de pixeles (*Pixel Bar Chart*) [50]

- Visualización Jerárquica [4].
  - Eje jerárquico (*Hierarchical Axis*) [51]
  - Apilamiento Dimensional (*Dimensional Stacking*) [52]
  - *Worlds Within Worlds* [53]
  - *Treemap* [54]
- Iconografía [4].
  - Caras de Chernoff (*Chernoff Faces*) [55]
  - *Star Glyph* [56]
  - *Stick Figure* [57]
  - Codificación de Formas (*Shape Coding*) [58]
  - Icono de Color (*Color Icon*) [59]
  - Texturas [60]

En muchos casos, la selección de un método de visualización u otro suele ser algo subjetivo. Si bien existen diferentes factores que influyen esa decisión [14], uno de los más importantes es la estructura propia del conjunto de datos a visualizar [4, 61, 62]. Adicionalmente, hay otros criterios que permiten establecer que método escoger para la visualización de un conjunto de datos. Para el caso específico de los conjuntos de datos MI, uno de los principales criterios es que el método de visualización debe permitir representar las relaciones que existen, no solo entre las instancias, sino también entre las bolsas del conjunto de datos. El visualizar estas relaciones puede ayudar a entender la estructura de composición de las bolsas y cómo estas se interrelacionan.

Otra característica deseable para un método de visualización de datos MI es que permita analizar más de una característica al tiempo. Esto con el fin de comparar la distribución de cada característica y determinar cuál o cuales pueden tener una mayor incidencia en el proceso de clasificación de las bolsas. También, resulta lógico pensar que la interactividad es otro elemento importante en una representación visual puesto que proporcionan al usuario la capacidad de manipular la visualización para obtener diferentes puntos de vista del conjunto de datos. La interacción, como característica de una visualización, puede mejorar la comprensión de la información en los datos.

La Figura **4-2**[1] muestra sugerencias de los posibles métodos de visualización qué pueden ser de utilidad para crear una representación visual de los datos, de acuerdo a la información que se quiere extraer con esta.

[1]Figura extraída de: ©A. Abela, 2010. `www.ExtremePresentation.com` [63]

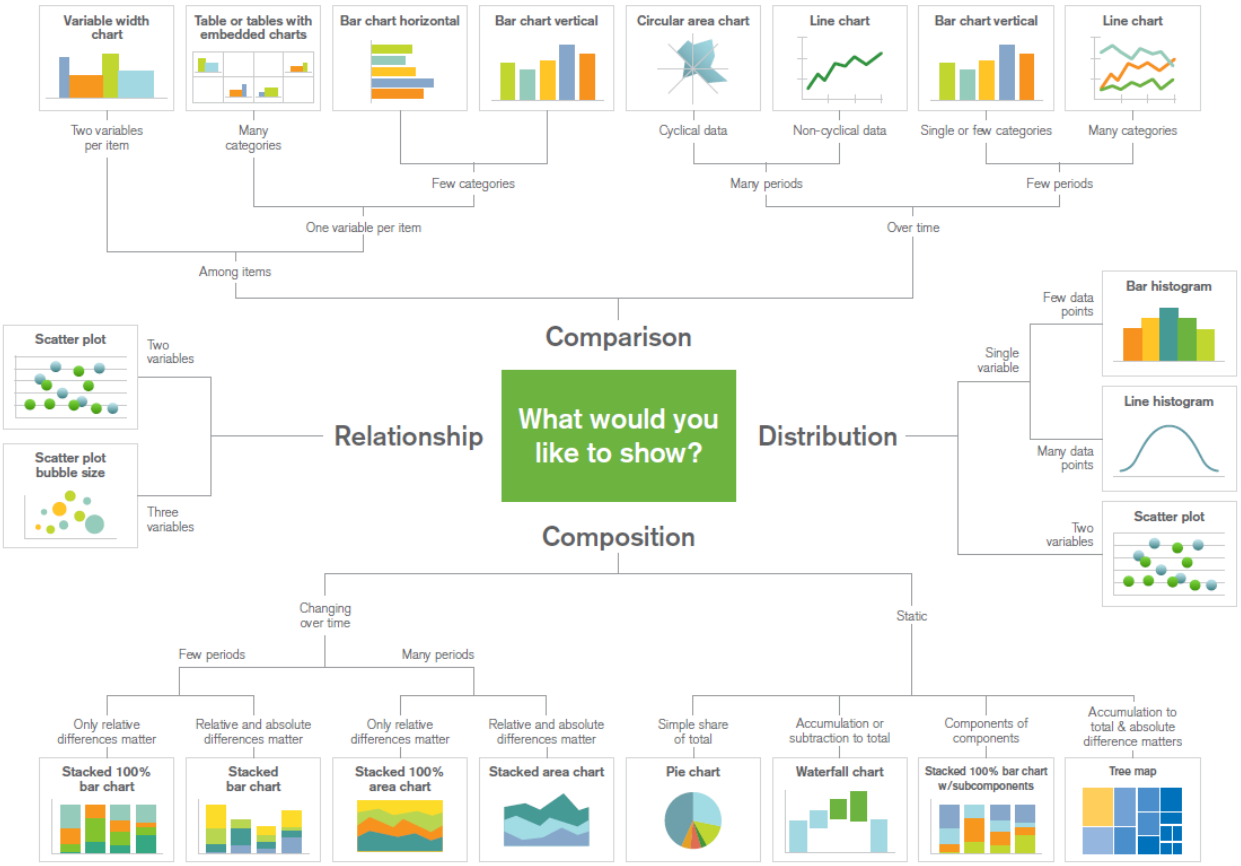


Figura 4-2.: Algunas técnicas de visualización de información organizadas en función de su utilidad.

4.3.3. Fase 3: Seleccionar las técnicas de visualización que mejor se adapten a los conjuntos de datos de MI con base en los criterios de comparación escogidos en la Fase 2

En esta fase se realizaron las siguientes actividades:

- Comparar los métodos de visualización de información
- Evaluar los resultados de la comparación mediante tablas o diagramas
- Elegir las técnicas de visualización con mejores resultados

Con base en la RSL del Capítulo 3 y de acuerdo con los criterios básicos definidos en la Sección 4.3.2, se generó una lista con los métodos de visualización que pueden resultar útiles para representar conjuntos de datos de MI. A continuación se describen, de manera muy breve, esos métodos de visualización.

Diagrama de Dispersión

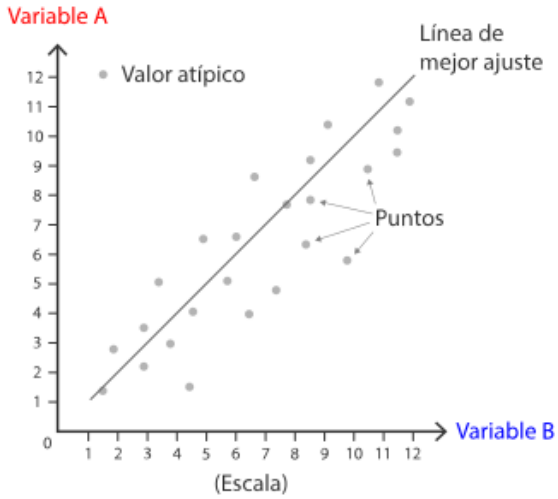


Figura 4-3.: Ejemplo de diagrama de dispersión.

Son diagramas que usan una colección de puntos colocados en un plano cartesiano para mostrar valores de dos o tres variables. Al mostrar una variable en cada eje, se puede detectar si existe una relación o correlación entre las dos variables. Es usado usualmente para realizar estimaciones mediante interpolaciones al trazar la línea de tendencias [64, 4, 65].

Las correlaciones que se muestran en el gráfico no son causales y es posible que otras variables puedan estar influyendo en los resultados del diagrama, esta es una desventaja puesto que se puede generar ruido en la visualización que conlleve a malas lecturas en el mismo [65].

En algunos casos, los diagramas de dispersión son enriquecidos usando diferentes figuras geométricas, colores y tamaños para permitir visualizar 3 dimensiones adicionales al gráfico. A modo de ejemplo, la Figura **4-3** muestra un diagrama de dispersión en 2D para un conjunto de datos.

Gráfico de Radar

En este gráfico a cada variable se le proporciona un eje que empieza en el centro del plano. Todos los ejes se disponen radialmente, manteniendo la misma escala entre todos los ejes. Cada valor de variable se traza a lo largo de su eje individual y todas las variables en un conjunto de datos se conectan para formar un polígono. Un ejemplo de este tipo de diagrama se puede ver en la Figura **4-4**.

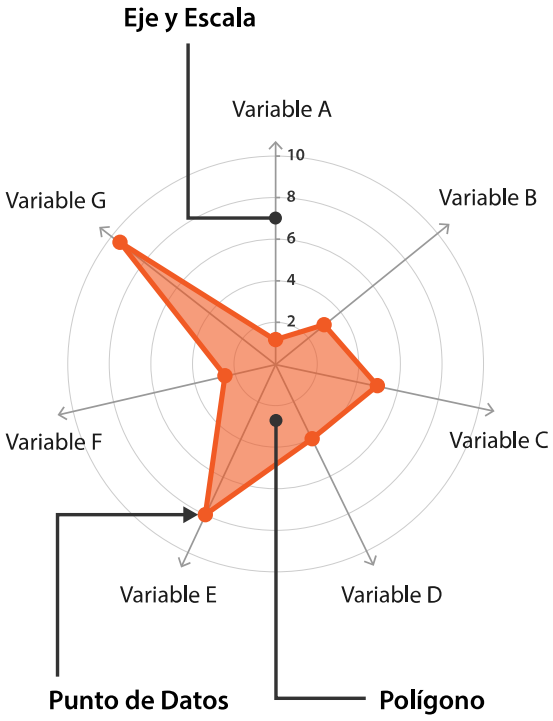


Figura 4-4.: Ejemplo de gráfico de radar.

Los gráficos de radar o de estrella son útiles para visualizar qué variables tienen valores similares o cuales variables tienen valores atípicos. También son útiles para determinar qué dimensiones tienen valores más altos en los datos. Es parte de las buenas prácticas mantener los gráficos radiales simples y limitar el número de variables utilizadas debido a que se pueden volver difíciles de comprender [66, 64, 65].

Una desventaja de este tipo de gráfico es que no son buenos para comparar valores entre las variables, debido al desorden presentado cuando se usan demasiadas dimensiones [65].

Gráficos de Barras

Este gráfico es usado usualmente para mostrar la cantidad de repeticiones de los valores de una variable que suele ser discreta. Para mostrar esto, en uno de los ejes se muestran las categorías y en el otro se muestra una escala de valor, cada barra se ajusta a esta escala para poder realizar comparaciones [4]. La Figura **4-5** ilustra un diagrama de barras típico.

Los gráficos de barras ayudan a estimar, cuando se usan a modo de histogramas, en dónde se concentran los valores, cuáles son los extremos y si hay valores atípicos. También, son útiles para dar una visión aproximada de la distribución de probabilidad [62, 4, 65] de la variable que estos representan.

Una limitación de los gráficos de barras está relacionada con el etiquetado puesto que al

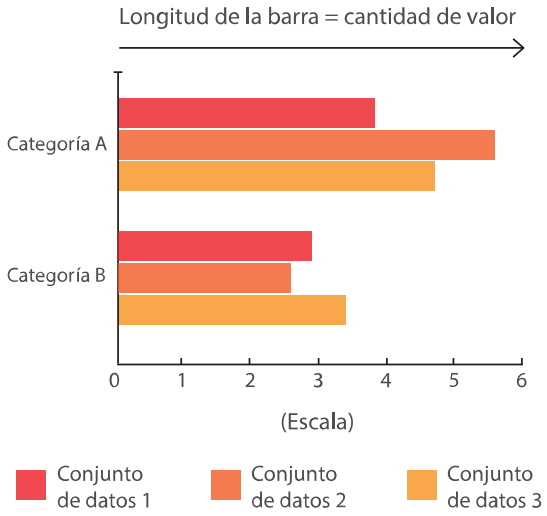


Figura 4-5.: Ejemplo de gráfico de barras.

crecer el número de barras se hacen menos legibles las etiquetas de las barras [12].

Gráfico de Burbujas

Al igual que un diagrama de dispersión, los gráficos de burbujas utilizan un sistema de coordenadas cartesianas para trazar puntos. No obstante, a diferencia de un gráfico de dispersión, a cada punto se le asigna una etiqueta o una categoría y, además, se suelen utilizar colores, tamaños y formas diversas para representar variables de datos adicionales [65].

Un inconveniente con este gráfico es que demasiadas burbujas pueden hacer que la representación sea difícil de leer, por lo que tienen una capacidad limitada respecto al conjunto de datos a representar [65]. Otra desventaja es que en algunos casos el tamaño de los círculos puede cambiar de maneras poco coherentes, esto daría lugar a interpretaciones erróneas. A modo de ejemplo, la Figura **4-6** muestra un gráfico de burbujas.

Gráfico de Velas

Los gráficos de velas son ideales para detectar y predecir las tendencias del mercado a lo largo del tiempo y son útiles para interpretar los flujos del mercado, a través del color y la forma de cada símbolo de candelabro. Estos patrones encontrados en los gráficos de velas son útiles para mostrar las relaciones de precios y pueden usarse para predecir el posible movimiento futuro de un mercado. La limitaciones en el análisis del mercado se dan porque no muestran lo ocurrido entre el valor inicial y final en su representación [65]. La Figura **4-7** ilustra un gráfico de velas y sus partes.

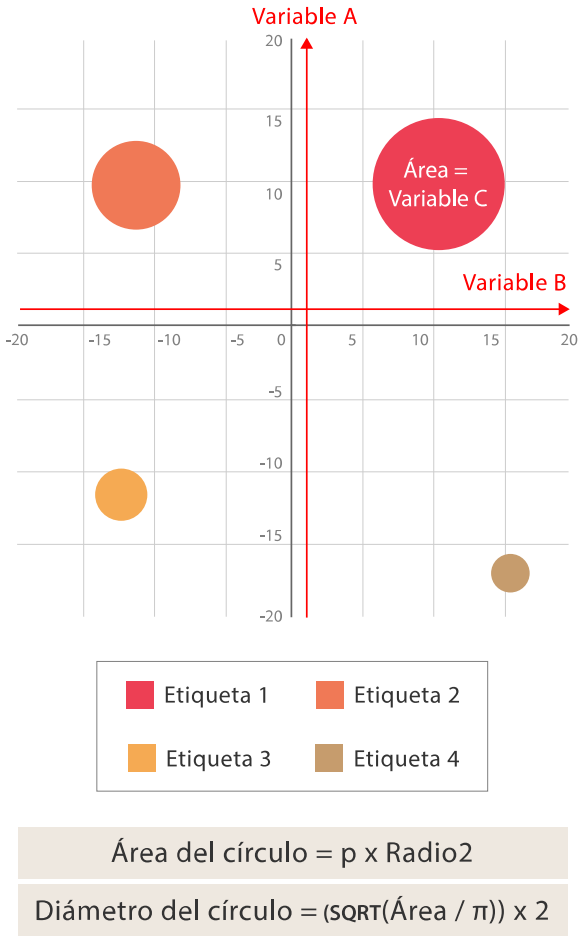


Figura 4-6.: Ejemplo de gráfico de burbujas.

Mapa de Calor

Los mapas de calor son buenos para mostrar la varianza a través de múltiples variables, revelando cualquier patrón asociado a el conjunto de datos. Esto permite ver si las variables son similares entre sí y para detectar si existen correlaciones entre ellas, también son usados para mostrar los cambios en los datos con el tiempo, entre otras utilidades. Este gráfico requiere una leyenda junto a un mapa de calor para que se pueda leer de manera correcta [65, 67]. La Figura **4-8** ejemplifica un mapa de calor.

Debido a su dependencia del color para comunicar valores, los mapas de calor son un gráfico adecuado para mostrar una visión generalizada de datos numéricos, ya que es más difícil distinguir con precisión las diferencias entre tonos de color y extraer puntos de datos específicos [12].

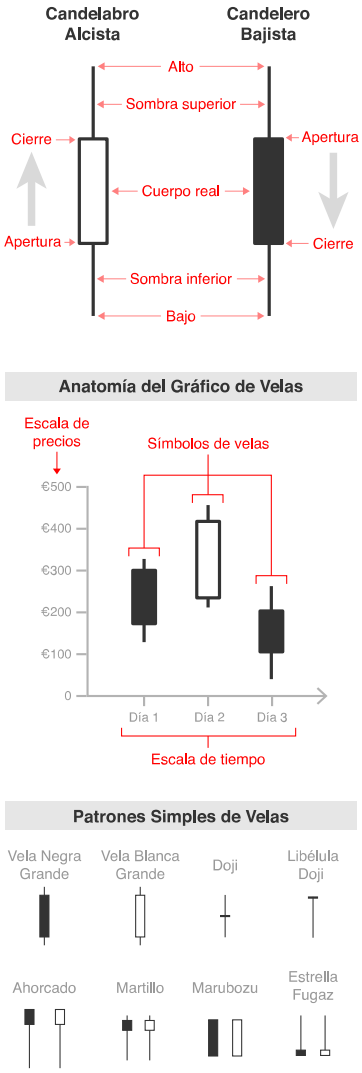


Figura 4-7.: Ejemplo de gráfico de velas.

Diagrama de Red

Este tipo de visualización muestra cómo las variables están interconectadas a través del uso de nodos y líneas que los enlazan para representar sus conexiones. Este tipo de diagrama ayuda a percibir el tipo de relación que existe entre grupos de entidades [5]. Normalmente, los nodos se dibujan cómo pequeños puntos o círculos, pero también se pueden usar iconos o figuras geométricas que ayudan a representar dimensiones adicionales de los datos. Asimismo, se pueden utilizar para interpretar la estructura de una red a través de la búsqueda de agrupaciones en los nodos [65].

Los diagramas de red tienen una capacidad de datos limitada y son difíciles de leer cuando hay demasiados nodos o cuando la agrupación de los nodos no es significativa [12]. A modo de ejemplo, la Figura **4-9** muestra un diagrama de red representado como un grafo.

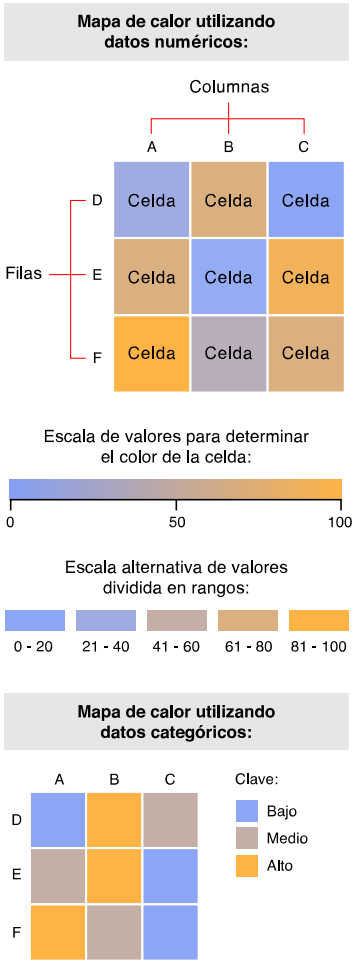


Figura 4-8.: Ejemplo de mapa de calor.

Gráfico de Coordenadas Paralelas

Este gráfico se utiliza para trazar datos numéricos multivariados. Los gráficos de coordenadas paralelas son ideales para comparar muchas variables y ver las relaciones entre ellas, similar a los diagramas radiales [4, 17, 66]. Un ejemplo típico de un gráfico de coordenadas paralelas se ilustra en la Figura **4-10**.

Uno de los inconvenientes de estos gráficos puede ser el orden en el que se disponen los datos, ya que variables que se encuentren más cerca serán más fáciles de comparar que las que estén más alejadas. Este problema se hace mas evidente al incrementarse la cantidad de datos [14, 12].

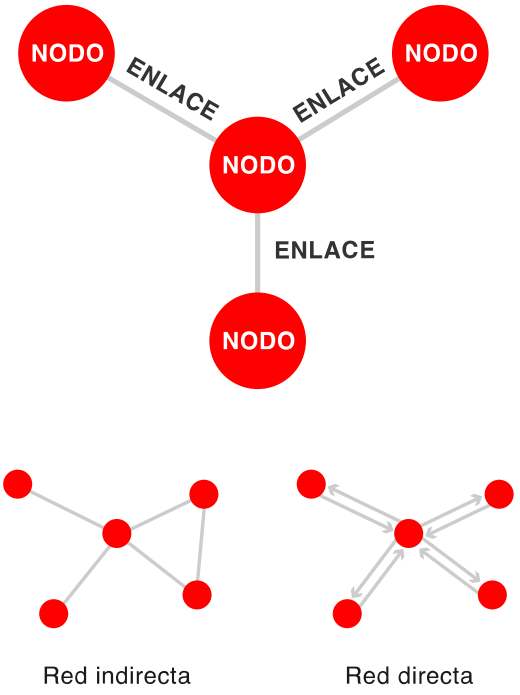
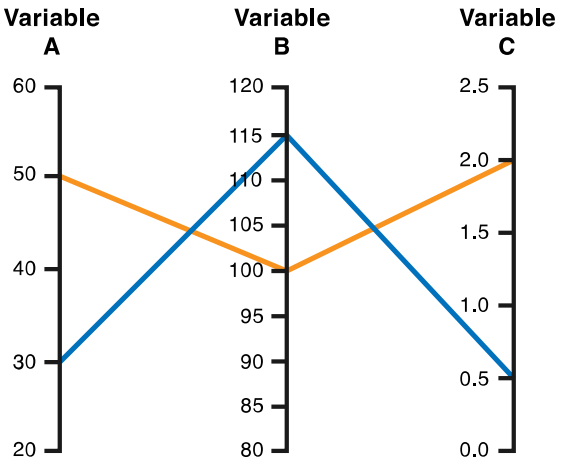
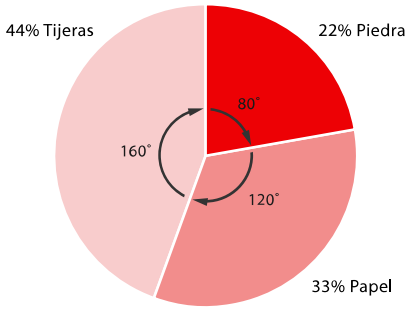


Figura 4-9.: Ejemplo de diagrama de red.



Datos			
	Variable A	Variable B	Variable C
Artículo 1	50	100	2.0
Artículo 2	30	115	0.5

Figura 4-10.: Ejemplo de gráfico de coordenadas paralelas.



Datos			
Piedra	Papel	Tijeras	TOTAL
2	3	4	9
Para calcular porcentajes			
2/9=22%	3/9=33%	4/9=44%	100%
Grados para cada «porción de tarta»			
(2/9) x 360 = 80°	(3/9) x 360 = 120°	(4/9) x 360 = 160°	360°

Figura 4-11.: Ejemplo de gráfico de pastel.

Gráficos de Pastel

Este gráfico es ideal para dar una idea rápida de la distribución proporcional de los datos [14], como muestra la Figura **4-11**. una limitación de este tipo de gráficos es que no se puede mostrar sino unos pocos valores porque a medida que estos aumentan el tamaño de cada segmento se hace más pequeño lo que dificulta su lectura y comprensión. A causa de esto, no son buenos para hacer comparaciones exactas, ya que usualmente los valores que se muestran son proporciones [65, 62, 66].

Gráfico de Columna Radial

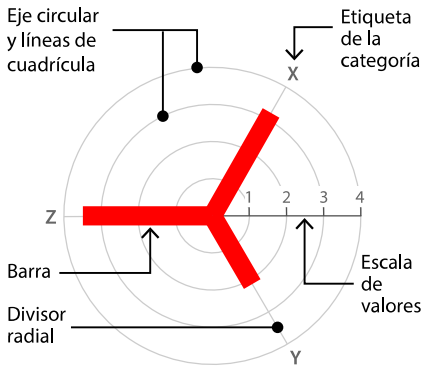


Figura 4-12.: Ejemplo de gráfico de columna radial.

Este gráfico es un híbrido entre un gráfico de barras y uno radial, en ese sentido preserva

las ventajas de ambos gráficos. Como ventaja, este tipo de gráficos es útil cuando se pretende comparar pocas variables, aunque al crecer los datos se hacen difíciles de comparar [65]. La Figura **4-12** presenta un ejemplo de este tipo de gráficos.

Mapa de Árbol (o Treemap)

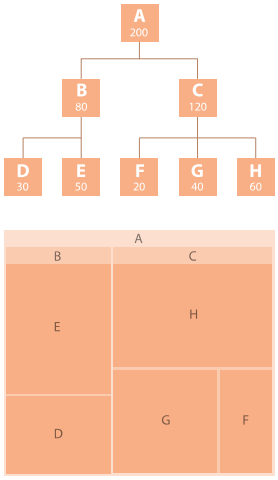


Figura 4-13.: Ejemplo de mapa de árbol.

Los mapas de árboles son una opción compacta y eficiente para mostrar jerarquías, dando una visión general y rápida de la estructura de los datos [4] en la jerarquía. También son ideales para comparar las proporciones entre categorías a través de su tamaño. Sin embargo, cuando el conjunto de datos tiene muchas variables y la jerarquía es demasiado profunda, puede ser confusa su lectura [62, 65]. Un ejemplo de como se representa una jerarquía con un gráfico de este tipo se muestra en la Figura **4-13**.

Todos los diagramas de ejemplo anteriores fueron extraídos del Catálogo de Visualización de Datos[2], que describe como cada método de visualización puede ser usado y sus características principales.

En la Tabla **4-1** se listan estos métodos de visualización con la información adicional que ayudó a decidir qué métodos pueden ser los más apropiados para alcanzar los objetivos propuestos.

Tabla 4-1.: Métodos de visualización de información

Nombre Gráfica	**Categoría**	**Características**	**Tipo variables**
Diagrama de Dispersión (*Scatterplot*)	Patrones Relaciones	2 o 3 dimensiones	Numéricas

[2]Extraído de: `https://datavizcatalogue.com/`

Nombre Gráfica	**Categoría**	**Características**	**Tipo variables**
Gráfico Radial (Gráfico de Estrella)	Comparaciones Relaciones Patrones	múltiples dimensiones	Numéricas
Gráficos de Barras	Comparaciones Patrones	2 dimensiones	Numéricas y categóricas
Gráfico de Burbujas	Comparaciones Dimensiones Distribución Patrones Relaciones	3 a 4 dimensiones	Numéricas y categóricas
Gráfico de Velas	Patrones Rangos Relaciones	7 o más dimensiones	Numéricas y categóricas
Mapa de Calor (Matriz)	Comparaciones Patrones Relaciones	múltiples dimensiones	
Histograma	Comparaciones Distribución Patrones Rangos	2 dimensiones	Numéricas y categóricas
Diagrama de Red (Grafo)	Relaciones	Múltiples dimensiones (Dependiendo cómo se use)	
Gráfico de Coordenadas Paralelas	Comparaciones Relaciones Patrones	Múltiples dimensiones	Numéricos
Gráficos de Pastel	Comparaciones parte de la totalidad Dimensiones	1 dimensión	Numéricos y categóricas
Gráfico de Columna Radial	Comparaciones	Múltiples dimensiones	Numéricas
Mapa de Árbol (*Treemap*)	Comparaciones Jerarquía Parte de la totalidad Dimensiones	Múltiples dimensiones	Categorías

4.3.4. Fase 4: Desarrollar o adaptar la propuesta de visualización para conjuntos de datos de MI

Esta fase se dividió en 4 sub-fases a fin de permitir el desarrollo de la propuesta de visualización y representación de conjuntos de datos de MI.

Fase 4.1: Seleccionar y buscar los conjuntos de datos de MI

Las actividades en esta sub-fase son:

- Recopilar, de bases de datos científicas, los conjuntos de datos de MI disponibles
- Elegir un grupo de conjuntos de datos de MI para hacer la validación

En la literatura existen varios conjuntos de datos de MI que se usan en la etapa de experimentación de múltiples investigaciones. Las áreas de aplicación de estos conjuntos de datos son diversas, siendo las más comunes la predicción de la actividad molecular, el etiquetado de imágenes, la categorización de texto, la clasificación de páginas web y la clasificación de grabaciones de audio. En la Tabla **4-2** se describen los conjuntos de datos de MI más comunes entre los investigadores.

Adicionalmente, es común encontrar conjuntos de datos MI artificiales construidos a partir de distribuciones Gaussianas [29]. Los autores suelen usar conjuntos de datos artificiales para mostrar ciertas propiedades de sus algoritmos. Esto se debe a que los conjuntos de datos artificiales ofrecen un mayor control sobre las características de los datos. En nuestro caso se creó un algoritmo que genera conjuntos de datos artificiales basados en Gaussinas 3D. El conjunto artificial creado tiene un total de 20 bolsas: 10 negativas y 10 positivas.

Fase 4.2: Definir una estrategia para proyectar los conjuntos de datos MI en un plano 2D

- Establecer y evaluar algunas técnicas de transformación de datos para reducir las dimensiones de la información obtenida del análisis de los conjuntos de datos de MI
- Relacionar los datos con figuras 2D de tal forma que permita ser consistente con la información que se quiere presentar

Debido a la alta complejidad de los conjuntos de datos de MI se hace necesario tratar de simplificar los datos sin perder información valiosa en el proceso. Es por ello que para transformar los datos se estudiaron diversas estrategias de reducción de dimensión.

En general, el uso de estrategias de reducción de la dimensionalidad permite remover las características irrelevantes en un conjunto de datos [71]. Esto a su vez ayuda a mejorar el rendimiento de los algoritmos de clasificación, además que pueden ayudar a incrementar la

Tabla 4-2.: Conjuntos de datos de MI

Datasets	**Bolsas**			**Instancias**			**Caract.**
	Pos	**Neg**	**Total**	**Pos**	**Neg**	**Total**	
Musk 1 [20]	47	45	92	207	269	476	166
Musk 2 [20]	39	63	102	1017	5581	6598	166
Mutagenesis 1 [68]	125	63	188	7790	2696	10486	7
Mutagenesis 2 [68]	13	29	42	660	1472	2132	7
Tiger [39]	100	100	200	544	676	1220	230
Fox [39]	100	100	200	647	673	1320	230
Elephant [39]	100	100	200	762	629	1391	230
Corel, African [40]	100	1900	2000	484	7463	7947	9
Corel, Antique [40]	100	1900	2000	338	7609	7947	9
Corel, Battleships [40]	100	1900	2000	280	7667	7947	9
Newsgroups 1, alt.atheism [32]	100	50	50	2667	2776	5443	200
Web recommendation 1 [69]	17	58	75	579	1633	2212	5863
Birds, Brown creeper [70]	197	351	548	4759	5473	10232	38
Birds, Winter Wren [70]	109	439	548	1824	8408	10232	38

comprensión sobre los datos [72].

Durante el proceso de exploración de los conjuntos de datos de MI se encontró que las técnicas de reducción se pueden aplicar de dos formas distintas: bien sea en el espacio de las instancias o en el espacio de las bolsas. Esto está relacionado con las familias de los algoritmos MIL explicados en el Capítulo 3 y contextualizado en la Figura **3-5**. Lo anterior quiere decir que es posible aplicar casi cualquier método de reducción de dos formas diferentes. La primera a nivel de instancias; esto quiere decir que se aplica en todo el conjunto de datos sin tener en cuenta a que bolsa pertenecen las instancias. La segunda a nivel de bolsa, es decir se aplica el método de reducción a las instancias de cada bolsa por separado y al final se forma un nuevo conjunto de datos con los resultados. Durante los experimentos se encontró que esta última forma de aplicar la reducción no funciona muy bien y puede ser contraproducente debido a que se forman subconjuntos de información que probablemente no guarde relación entre si. Por lo anterior, los métodos de reducción se aplicaron en menor medida en el espacio de las bolsas.

Actualmente existen diferentes técnicas para la reducción de la dimensión de un conjunto de datos. Entre ellas, las más conocidas son *Kernel Principal Component Analysis* (KPCA) [73], *Fast Independent Component Analysis* (FastICA) [74], *Isometric Feature Mapping* (Isomap) [75], *Locally Linear Embedding* (LLE) [76], *Multidimensional Scaling* (MDS) [77] y *T-Distributed Stochastic Neighbor Embedding* (TSNE) [78]. A continuación se describen brevemente estas estrategias.

KPCA: Es una modificación del algoritmo *Principal Component Analysis* (PCA) [73, 79] que incluye el uso de funciones tipo kernel. Estas últimas dotan al algoritmo de transformación con la capacidad de realizar proyecciones no lineales. En este sentido, KPCA calcula los vectores principales a partir de la matriz del kernel, en lugar de hacerlo sobre la matriz de covarianzas, como lo hace PCA [80].

FastICA: Es un algoritmo eficiente para el análisis de componentes independientes. Se basa en un esquema de iteración de punto fijo para encontrar un máximo de la "no Gaussianidad". Contrariamente a los algoritmos basados en gradiente, no hay parámetros de tamaño de paso para elegir. Esto significa que el algoritmo es fácil de usar. Los componentes independientes se pueden estimar uno por uno, lo que equivale a realizar una búsqueda de proyección. Esto es útil en el análisis exploratorio de datos. FastICA tiene la mayoría de las ventajas de los algoritmos neuronales: es paralelo, distribuido, computacionalmente simple y requiere poco espacio de memoria [74].

MDS: Es un conjunto de técnicas matemáticas que permiten descubrir la estructura oculta de los datos [77, 71]. Su principal objetivo es representar los datos en pocas dimensiones mientras trata de preservar la distancia entre los puntos. MDS es similar a PCA cuando se

usa distancia Euclidiana. Existen varios métodos MDS que se diferencian en la métrica de distancia usada así cómo en el calculo del rendimiento [81].

Isomap: Es una técnica no lineal de reducción de dimensiones que puede clasificarse cómo una variación de *Multidimensional scaling* (MDS), con distancias geodésicas entre puntos, en vez de distancias Euclidianas. Las distancias geodésicas es representada por los caminos más cortos a lo largo de la superficie curva [75, 81]. Es un método global que produce un espacio de bajas dimensiones al preservar las distancias por pares entre los puntos de datos [82].

LLE: Es una técnica no lineal de reducción de dimensiones que produce espacios de bajas dimensiones preservando los datos embebidos en las altas dimensiones [82, 78]. LLE al compararlo con Isomap es más eficiente [76, 81].

TSNE: Usado ampliamente para visualización de información, convierte las similitudes entre los puntos de datos en probabilidades e intenta minimizar la divergencia entre las probabilidades en los espacios de altas y bajas dimensiones [78].

Todos los algoritmos anteriormente descritos fueron implementados dentro de la propuesta de visualización, la razón de incluirlos fue basada en dar flexibilidad y brindar un abanico amplio de posibilidades a los usuarios. También se hizo pensando en realizar visualizaciones con diferentes transformaciones de datos y comparar con cual de los algoritmos de reducción funciona mejor determinado conjunto de datos.

Fase 4.3: Procesar y analizar los conjuntos de datos de MI

- Seleccionar las características relevantes de un conjunto de MI y analizar cómo relacionarlas a nivel de bolsas e instancias
- Construir un modelo de visualización que permita analizar las diferentes características de un conjunto de datos de MI

Como se ha mencionado, los conjuntos de datos de MI tienen una estructura que hace complejo su análisis y entendimiento. Así mismo, seleccionar las características adecuadas para representar el conjunto de datos en un gráfico con limitaciones espaciales es una tarea difícil y, además, subjetiva. Básicamente esto depende del uso que se le esté dando a los datos y de la comprensión que tiene el usuario sobre los datos que está intentando analizar.

A pesar de lo anterior, es posible visualizar cierta información particular de todos los conjuntos de datos de MI a fin de ayudar al usuario a comprender mejor su estructura. Entre esta información está la distribución de las bolsas y la distribución de las instancias que cada bolsa contienen. Así mismo, es posible usar los métodos de reducción de la dimensionalidad

para crear una representación gráfica que ayude al análisis del conjunto de datos. La Figura **4-14** muestra la estrategia que se utilizó para transformar los datos a partir de las técnicas de reducción descritas anteriormente.

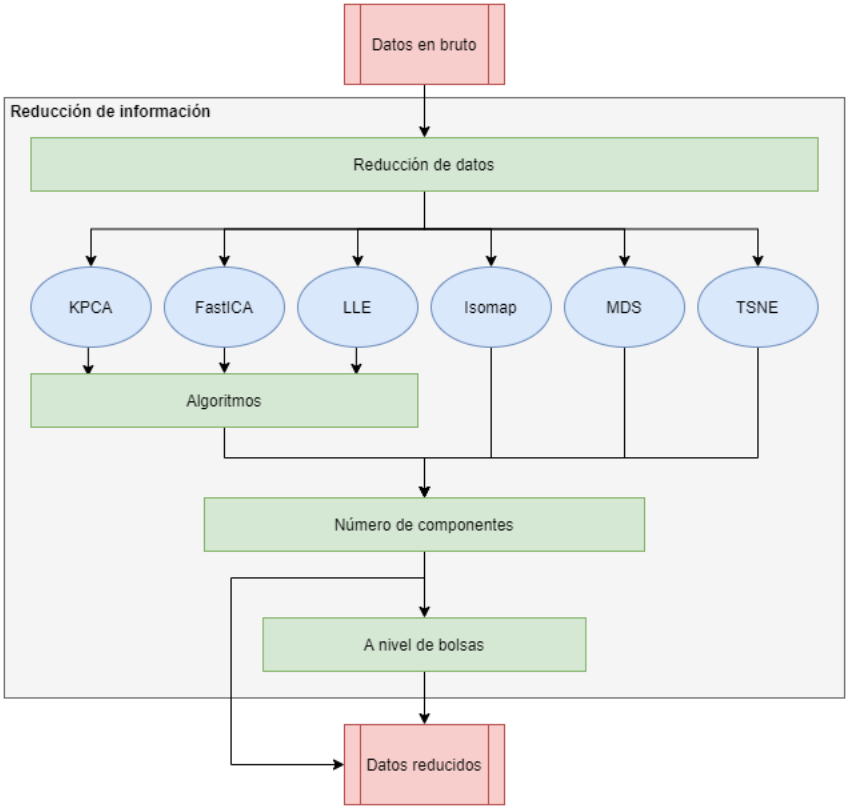


Figura 4-14.: Estrategia de reducción de información.

Por otra parte, existen problemas de aplicación en los que se pueden generar conjuntos de datos de MI cuyas bolsas positivas y negativas pueden ser muy densas. Esto pasa cuando cada bolsa del conjunto de datos contiene una cantidad enorme de instancias. Esto de alguna forma puede sesgar los clasificadores MIL [83] y además dificultad la compresión del conjunto de datos durante su visualización.

Con el fin de abordar ese problema, en el desarrollo de la estrategia de visualización se exploraron diferentes estrategias para reducir el número de instancias que contiene cada bolsa. Entre estas estrategias se estudiaron métodos estadísticos de primer orden para extraer los máximos, mínimos, promedios y extremos de cada bolsa.

Además, se implementó una estrategia de reducción basada en *Kernel Density Estimator* (KDE), como lo hicieron Mera *et al.* en [83]. KDE es un estimador no paramétrico de densidades univariadas o multivariadas [84, 85] que puede ser usado para disminuir el número de instancias en las bolsas positivas y negativas. El uso de este algoritmo depende del parámetro de suavizado, comúnmente conocido como el ancho de banda (*bandwidth*) [86, 85]. En

nuestro caso se usó como parámetro de ancho de banda una densidad Gaussiana con media -1 y desviación estándar 1.

Aplicando los anteriores métodos de simplificación se busca reducir la cantidad de instancias tratando de mantener la información importante de la bolsa en el en el proceso. Esta estrategia se ilustra en la Figura **4-15**.

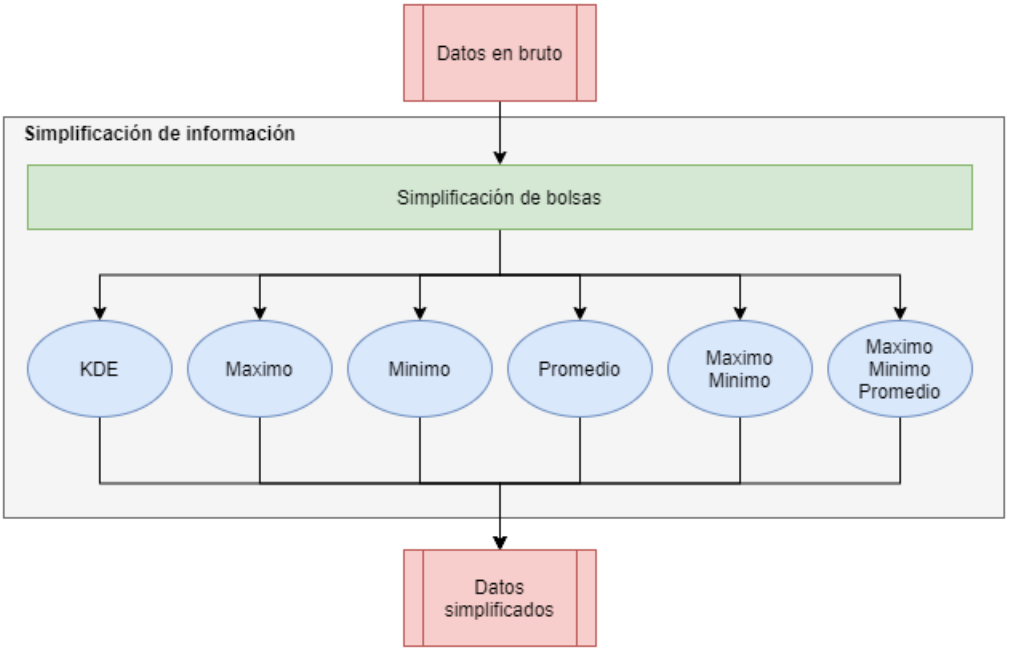


Figura 4-15.: Estrategia de simplificación de información.

Adicionalmente, como se muestra en la Figura **4-16**, las estrategias de reducción y simplificación de datos se combinaron con el fin de aumentar las posibilidades de encontrar patrones en los conjuntos de datos de MI.

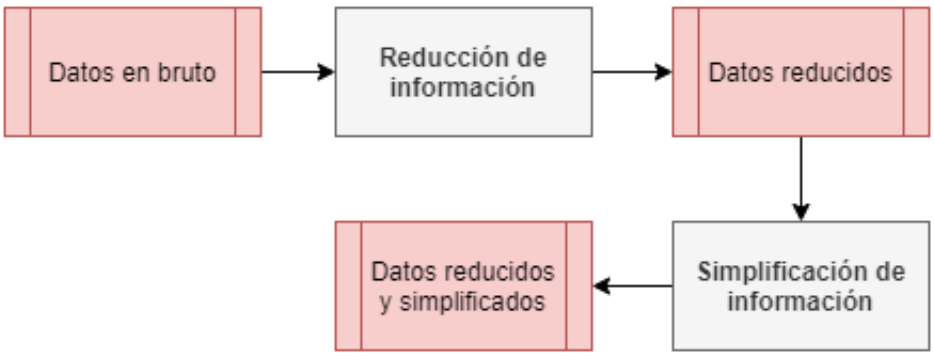


Figura 4-16.: Estrategia de reducción y simplificación de información.

En conjunto, las estrategias mencionadas constituyen un modelo de transformación de

datos que permite, de manera sólida, mejorar su procesamiento para efectos de visualización; al menos en lo que respecta a los conjuntos de datos de MI con altas dimensiones y muchas instancias por bolsa.

Fase 4.4: Desarrollar una estrategia para representar de manera visual los conjuntos de datos de MI

- Transformar la información del modelo en figuras geométricas básicas para su posterior proyección
- Realizar una representación de la información en un plano 2D en el cuál se relacione la información de las bolsas e instancias

En esta fase se utilizaron los elementos descritos en las fases anteriores con el fin de crear una representación visual de un conjunto de datos de MI en un plano 2D. Para lograrlo se hicieron algunas aproximaciones que se describen a continuación. Primero se utilizaron algunos de los métodos de visualización descritos para evaluar su efectividad en la representación visual de los conjuntos de datos. En este caso la evaluación buscó identificar, subjetivamente, si cada método de visualización permitía extraer información relevante del conjunto de datos.

Tras terminar las pruebas de concepto se pudo observar que algunos métodos, como los diagramas de burbujas, los gráficos de barras y los diagramas de coordenadas paralelas, no son útiles para la representación debido a la enorme cantidad de datos a visualizar. Por otra parte los diagramas de dispersión funcionan con casi todos los conjuntos de datos relativamente bien, pese a eso son difíciles de interpretar y cuando la reducción de las dimensiones se realiza con conjuntos de datos que poseen muchas características, como *Musk 1*, la distribución mostrada no es clara y se pierde demasiada información en el proceso de reducción.

Así, para poder trasformar los datos y proyectarlos en un plano fue necesario conocer las particularidades de cada uno de los métodos de visualización. Esto nos permitió identificar que preprocesamiento se debía realizar a los datos con el fin de proyectarlos en el plano, evitando generar representaciones visuales confusas.

Como ejemplo de las pruebas de concepto, en una de las primera aproximaciones consistió en realizar una reducción de dimensiones usando KPCA con un kernel lineal sobre el conjunto de datos *Musk 1*. A partir de esta transformación se utilizó un diagrama de dispersión para proyectar las componentes principales. El resultado de la visualización en 2 y 3 dimensiones se presentan en las Figuras **4-18** y **4-17**, respectivamente.

Cómo puede apreciarse en ambos diagramas, no existe mucha diferencia entre la representación con KPCA usando 2 o 3 componentes principales para *Musk1*. En parte esto se debe a que por la naturaleza del propio conjunto de datos es difícil ver la relación entre las instancias, además, que la transformación no es la más efectiva para preservar la información

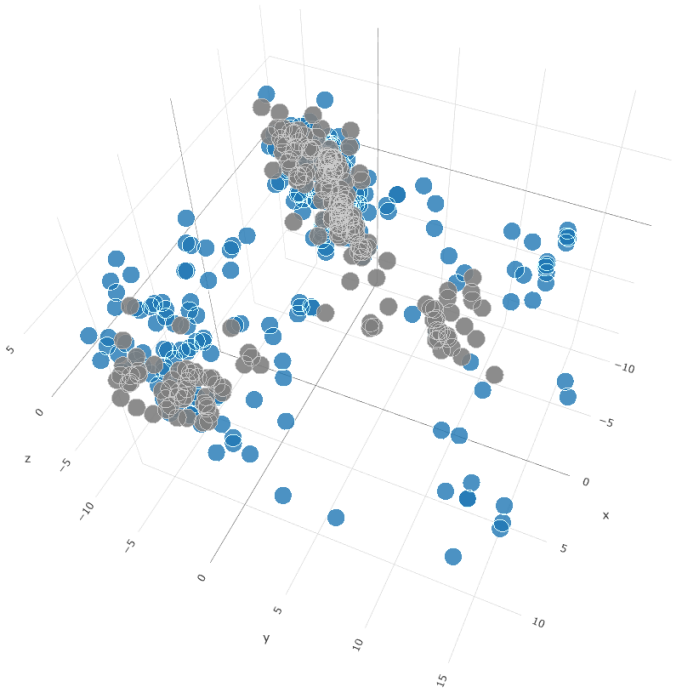


Figura 4-17.: Visualización *Musk1* después de reducir 166 dimensiones a 3 usando KPCA (instancias positivas en gris y negativas en azul).

de las 166 características iniciales, en las 2 o 3 componentes principales de KPCA.

Además del diagrama de dispersión, en las primeras pruebas se optó por proyectar el conjunto de datos sobre un diagrama de coordenadas paralelas, el cual se muestra en la Figura **4-19**. La idea de de usar este de diagrama era evaluar su capacidad para permitir analizar si existía una tendencia entre las instancias positivas y negativas con respecto a los valores de las característica en el conjunto de datos. Sin embargo, se detectó que, por la cantidad de características, la visualización queda muy saturada de información. Esto evidenció la debilidad de este tipo de diagramas para realizar una comparación entre instancias de manera efectiva.

Posteriormente, se intentaron combinar diferentes tipos de diagramas para tratar de representar las diversas características de los conjuntos de datos de MI. Al hacer esto se pretendía mitigar las falencias individuales de cada método de visualización. Como resultado se encontró que la combinación entre un diagrama de radar con un diagrama de barras permitía representar la información completa de una bolsa, incluida la información de las instancias de la misma.

Un ejemplo de esta visualización se presenta en la Figura **4-20**. En el gráfico de esta Figura se observa una bolsa sencilla que sólo contiene dos instancias, cada instancia está separada por un radio que forma una sección del gráfico de radar. La escala de valores de

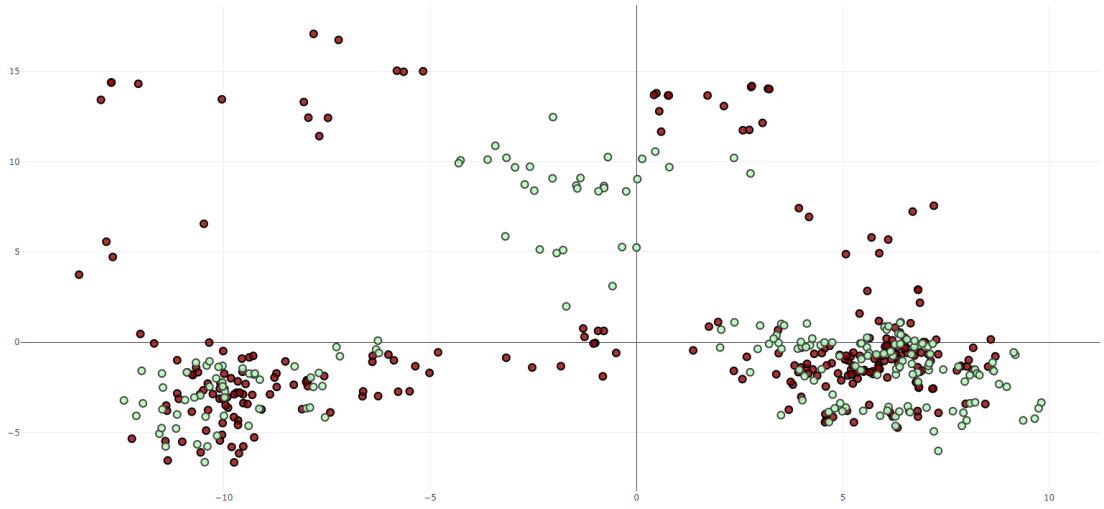


Figura 4-18.: Visualización *Musk1* después de reducir 166 dimensiones a 2 usando KPCA (instancias positivas en verde y negativas en rojo).

las características se muestra en el eje principal y está en el rango [0.01, 100]. Estos valores son obtenidos después de normalizar los datos del conjunto. En cada uno de las secciones se trazan barras qué corresponden al valor normalizado de la característica en cada instancia. En este caso solo se representan 5 características a manera de ejemplo, sin embargo, se podría incrementar el número de características sacrificando legibilidad y saturando un poco la visualización.

Esta representación es útil para comparar valores de cada atributo y conocer en un vistazo cuales características afectan de manera mas significativa una bolsa positiva o negativa.

Con base en lo anterior, se procedió a utilizar este diagrama con todos los conjuntos de prueba y se encontró que su mayor limitación está definida por el número de características que se pueden visualizar a la vez. Además, cuando crece el número de instancias en las bolsas se hace aún más difícil la lectura del gráfico. Para ilustrar esto se representó el conjunto de datos *Musk 1* en esta propuesta de visualización, la cual se muestra en la Figura **4-21**. Como se observa, la visualización completa es una matriz donde cada celda representa una bolsa del conjunto de datos y, cómo se indicó, en cada diagrama individual se ve la distribución de las instancias respecto a las características seleccionadas para la representación.

Esta visualización permite observar la distribución de las bolsas positivas y negativas en términos de sus instancias. El gráfico nos permite identificar rápidamente que las bolsas negativas tienen una mayor cantidad de instancias, comparadas con las bolsas positivas. En cuanto a las características que se representan hay una de ellas (la característica en color magenta) que posee valores mas altos, en comparación con el resto de características representadas. Esto puede indicar que este atributo particular tiene un peso significativo en el conjunto de datos.

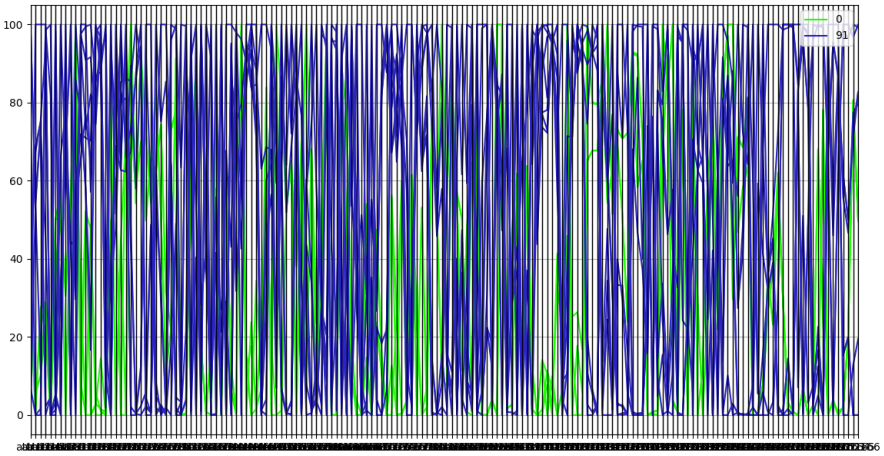


Figura 4-19.: Visualización *Musk1* en un diagrama de coordenadas paralelas.

Ahora, si bien el gráfico da información rápida sobre la composición de las bolsas, es difícil analizarlas de manera individual, principalmente cuando la bolsa tiene muchas instancias, como es el caso de la primera bolsa en la segunda fila, de abajo a arriba. Adicionalmente, la cantidad de atributos que se pueden visualizar es limitada y a medida que se quieren mostrar más, el diagrama pierde legibilidad y en consecuencia la idea inicial de comparar las características entre las instancias no es efectiva. Una última limitación de este tipo de diagramas es que muestra la información de cada bolsa pero el diagrama en sí es incapaz de mostrar las relaciones entre las bolsas del conjunto de datos.

Por la razón anterior, se realizaron pruebas concepto con otros métodos de visualización hasta llegar a una representación satisfactoria para los conjuntos de datos de MI. En este caso, la visualización propuesta consta de un grafo; en el cuál se representan las bolsas y sus relaciones mediante las líneas que conectan cada nodo. Además, la visualización propuesta se dotó de un método de interacción que permite al usuario seleccionar un conjunto de nodos (que representan a las bolsas) y mostrar en un diagrama de radar la distribución de las instancias de ese grupo de bolsas. Esto permite al usuario analizar grupos de bolsas, con base en los valores de sus instancias. Un ejemplo e la propuesta de visualización se presenta en la Figura **4-22**.

Esta propuesta involucra dos tipos de gráficos. El primero consiste de una visualización del conjunto de datos como un grafo. En este hay dos tipos de nodos: los círculos rojos, que representa las bolsas positivas, y los triángulos azules que representan las bolsas negativas. El tamaño de cada nodo está definido con base en la cantidad de instancias, así entre más instancias tiene una bolsa, mayor es el tamaño de la figura que la representa. Para el caso de la Figura **4-22**, se puede observar rápidamente que las bolsas negativas tienen mayor

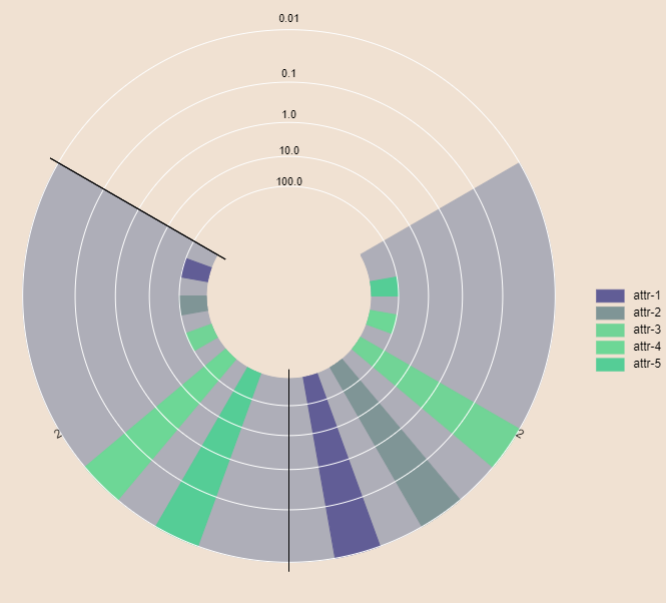


Figura 4-20.: Visualización de una bolsa en un diagrama radial de barras

mayor tamaño y por tanto las instancias negativas son las que predominan en el conjunto de datos. Adicionalmente, en el lateral de la visualización se encuentran dos barras que indican el número de conexiones de las bolsas en una escala de intensidad por color. La escala se planteó para ser independiente para las bolsas positivas y negativas a fin de facilitar el análisis de la predominancia de conexiones entre ambos dos tipos de bolsas.

La disposición de los nodos en el grafo está definida por diferentes algoritmos, los cuales pueden ser cambiados interactivamente. Estos son: *spring*, *kamada kawai*, *circular*, *random*, *planar* y *spectral* [3].

Por defecto, en la visualización se usa el algoritmo *Fruchterman-Reingold force-directed* (*spring*). La idea de este algoritmo es básicamente minimizar la energía del sistema moviendo los nodos y cambiando las fuerzas entre dos nodos cualquiera. En este sentido, la suma de los vectores de fuerza determina en qué dirección debe moverse un nodo para posicionarlo automáticamente de manera que se minimice la energía total del sistema. El grafo se estabiliza cuando el sistema alcanza su estado de equilibrio, es decir cuando se minimiza por completo la energía del sistema [87, 88].

El uso de este algoritmo en grafos de tamaño medio, de 50 a 500 vértices, tiene, generalmente, muy buenos resultados. Sin embargo, una desventaja es que su tiempo de procesamiento es alto. Esto se debe a que la complejidad del algoritmo es exponencial dependiendo de la cantidad de nodos que existan [87].

[3]Extraído de: Drawing — NetworkX 2.5. `https://networkx.github.io/documentation/latest/reference/drawing.html#module-networkx.drawing.layout`

Figura 4-21.: Visualización *Musk 1* en propuesta de visualización combinada (bolsas negativas en gris y positivas en rosado).

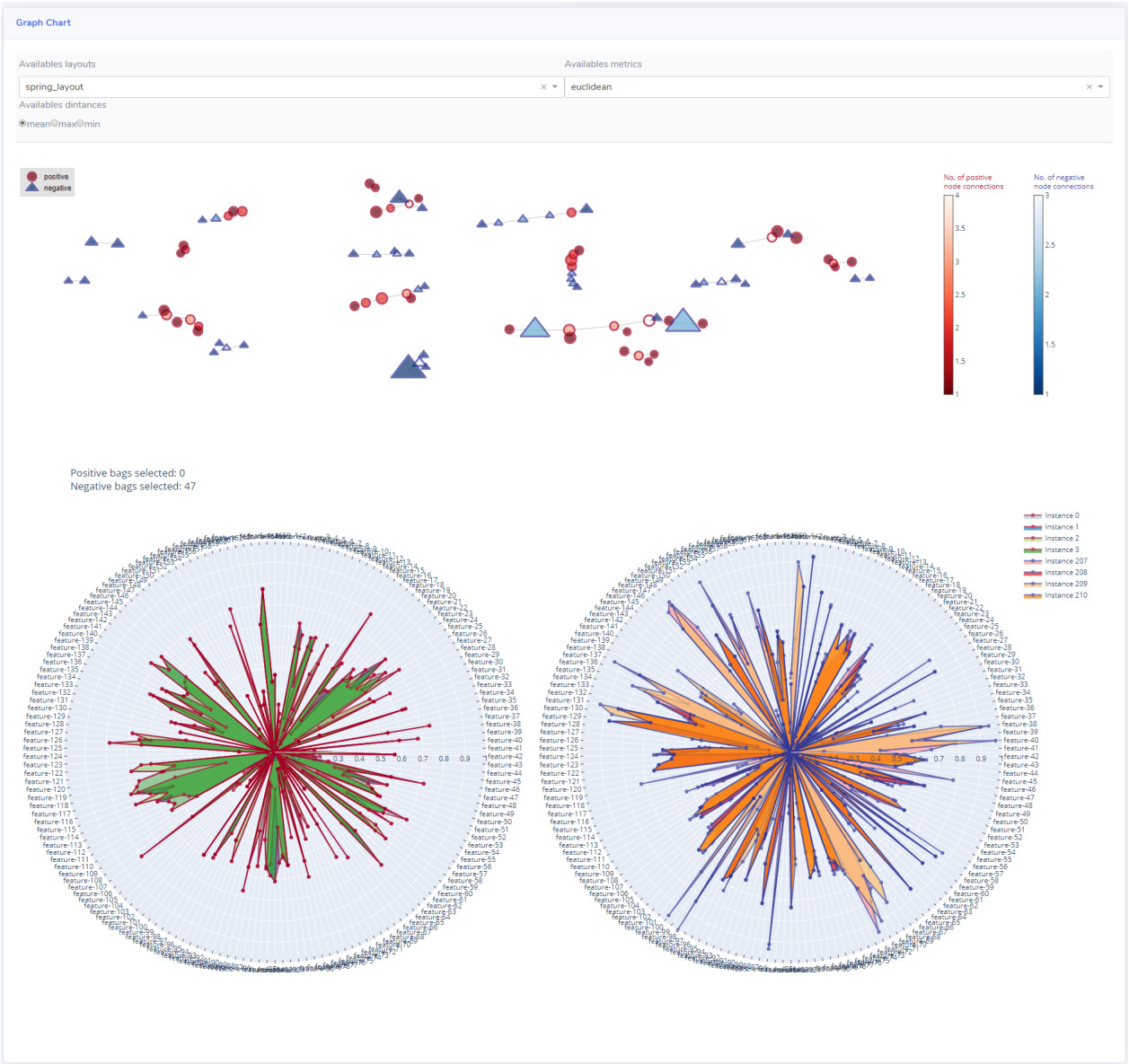


Figura 4-22.: Método de visualización propuesto para representar conjuntos de datos de MI.

Como el peso de los vértices está definido con base en la distancia entre las bolsas, y dicha distancia se calcula con base en las instancias entre las bolsas [1], entonces se puede escoger si el peso de un vértice se asigna con base en la distancia mínima, máxima o la mediana entre las distancias de las instancias en dos bolsas. Por otro lado, la métrica de distancia usada es la distancia Euclidiana, aunque se tienen implementados diferentes métricas en caso de que el usario considere más cnveniente usar otra. Estas métricas son: *braycurtis*, *canberra*, *chebyshev*, *cityblock*, *correlation*, *cosine*, *dice*, *euclidean*, *hamming*, *jaccard*, *jensenshannon*, *kulsinski*, *mahalanobis*, *matching*, *minkowski*, *rogerstanimoto*, *russellrao*, *sokalmichener*, *sokalsneath*, *yule* [4].

El segundo gráfico de la propuesta de visualización corresponde a dos diagramas de radar. Uno para visualizar las instancias de las bolsas positivas y el otro para visualizar las instancias de las bolsas negativas. Cada uno de estos diagramas muestra la distribución de las instancias respecto a las características seleccionadas. Este gráfico cambia de manera interactiva conforme cambian las bolsas que se seleccionan en el grafo.

Los diagramas de radar implementados permiten, rápidamente, determinar qué características son predominantes entre las instancias. Esto es útil para determinar que atributos pueden tener mayor influencia en la definición y la distinción de las bolsas positivas de las negativas. No obstante, una de las debilidades de esta representación se presenta cuando se seleccionan bolsas con demasiadas instancias, puesto que debido a la naturaleza del gráfico este se torna más difícil de leer conforme aumenta el número de instancias.

La explicación en detalle del funcionamiento y los controles de la propuesta de visualización de conjuntos de datos de MI que se presenta aquí, se puede ver en el Anexo A.

Herramientas usadas en el desarrollo de la propuesta de visualización

Durante el desarrollos de la propuesta se realizaron algunas pruebas de conceptos previas usando diferentes tipos de herramientas y librerías que facilitaron el proceso de evaluación rápida y nos permitieron gestionar los conjuntos de datos de MI. Como herramienta principal y como lenguaje de desarrollo se uso *Python 3.7*[5], debido a su versatilidad y gran cantidad de documentación asociada al manejo de datos. Por otro lado, para la herramienta web se uso el Framework *Django*[6] en su versión 2.2.4, este se uso sobre *Flask*[7] debido a que este ultimo esta mas enfocado en microservicios y *Django* posee muchos módulos que agilizan la tarea de construir una herramienta web.

Ya teniendo las anteriores herramientas establecidas como la base, se empezaron a explo-

[4]Extraído de: SciPy v1.4.1 `https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.pdist.html`
[5]`https://www.python.org/`
[6]`https://www.djangoproject.com/`
[7]`https://flask.palletsprojects.com/en/1.1.x/`

rar diferentes módulos que ayudaran al manejo de los datos, entre las mas usadas por la comunidad se encuentra la librería *pandas*[8], que es una de las mas usadas para el manejo y análisis de datos, también se uso *numpy*[9] para el manejo de algunas operaciones con los datos. Entre las librerías para la clasificación se opto por usar *scikit-learn*[10] y *scipy*[11] para hacer uso del clasificador *KNeighborsClassifier* que ya tiene la librería implementada y el cual fue usado en el algoritmo de *SimpleMIL*. Adicionalmente se usaron estas librerías para implementar los métodos de reducción explicados anteriormente.

En cuanto a las librerías usadas para la visualización de los datos en un principio se uso *matplotlib*[12] para la etapa exploratoria de los datos, posteriormente y debido a las necesidades que se requerían para la visualización se exploro el uso de *bokeh*[13], una librería especializada en visualización la cual posee mucha versatilidad aunque la curva de aprendizaje es algo alta con respecto a otras librerías de este tipo, la ventaja que posee *bokeh* es que se puede personalizar y crear gráficos desde las figuras geométricas básicas y enlazar eso con los datos que se poseen, es muy versátil para crear gráficos desde cero o crear propuestas totalmente nuevas. Esta librería también posee una función que permite interactuar con los diferentes aspectos del gráfico sin embargo, es engorrosa de implementar y no funciona bien cuando se usa sobre un Framework como *Django*.

Por la razones anteriores la implementación final de la propuesta se opto por usar *plotly*[14], una de las librerías para la visualización de información mas usadas por los científicos de datos debido a su versatilidad y facilidad de uso. *Plotly*, es una librería con una curva de aprendizaje mucho menor que *bokeh* aunque pierde en algunos aspectos de flexibilidad ya que se debe seguir una estructura impuesta por la librería para realizar la visualización en las gráficas que se deseen, pero para nuestros propósitos era suficientemente adaptable además que la integración con las herramientas que teníamos era mucho mas fácil y trasparente, *plotly* también posee un sistema de interacción que funciona muy bien en casi cualquier caso, ya sea actualización de datos, ocultar información, resaltar o opacar zonas del gráfico entre muchas otras, lo único que se puede ver como inconveniente es que requiere un buen uso y gestión de los datos para no consumir demasiada memoria y que las actualizaciones e interacciones sean fluidas y no presente tiempos de carga muy largos (sin embargo al usar grandes cantidades de datos esto sera inevitable).

Para terminar se usaron algunos otros módulos como *networkx*[15] que fue fundamental para la construcción del grafo, *jupyter notebook*[16] para la exploración y análisis de datos

[8] `https://pandas.pydata.org/`
[9] `https://numpy.org/`
[10] `https://scikit-learn.org/`
[11] `https://www.scipy.org/`
[12] `https://matplotlib.org/`
[13] `https://bokeh.org/`
[14] `https://plotly.com/`
[15] `https://networkx.github.io/`
[16] `https://jupyter.org/`

inicial así como algunas librerías del lado del frontend para el manejo de las tablas y algunas interacciones en la la herramienta, la mas destacada es *datatable*[17].

Cabe resaltar que todas las herramientas usadas son software libre y tienen un gran soporte de la comunidad y cuentan con grandes empresas que contribuyen a su desarrollo.

4.3.5. Fase 5: Evaluar la técnica de visualización empleada en los conjuntos de datos de MI usando los conjuntos de datos de ejemplo

El desarrollo de esta fase incluyó las siguientes actividades:

- Elaborar encuestas y cuestionarios que permitan una evaluación objetiva sobre los resultados
- Buscar expertos en la temática para evaluar los resultados obtenidos
- Tabular la información obtenida mediante el uso de las encuestas y cuestionarios
- Concluir con base en los resultados obtenidos mediante el juicio de expertos

En esta fase se realizó la validación del sistema que implementa la propuesta de visualización. En este sentido se realizaron una serie de documentos para verificar la utilidad de las gráficas y obtener resultados objetivos para evaluar posibles mejoras y trabajos futuros.

Uno de los primeros documentos realizados es el protocolo de investigación de usuarios (Anexo B), el cuál es un documento que permite establecer las pautas para que los usuarios, en nuestro casos expertos en el temas de MIL, puedan realizar las pruebas pertinentes en el sistema.

En el documento se describe la metodología y logística de la prueba, también se establece un arquetipo de la persona que debe realizar la evaluación de la visualización. El documento se empleó para obtener la muestra de usuarios, establecer el flujo de ejecución de la prueba y la hipótesis de la misma, las tareas a desarrollar por los usuarios y las preguntas con las cuales se buscó extraer información objetiva acerca del sistema desarrollado.

Otro de los documentos presentados a los expertos es una pequeña encuesta (Anexo C), que consiste en 7 preguntas con las cuales se buscó ahondar más en la utilidad del sistema de visualización propuesto e identificar posibles puntos de mejora.

Debido a que la cantidad de usuarios que ejecutaron las pruebas fueron muy pocos, las preguntas de la encuesta fueron abiertas ya que no presenta ninguna dificultad de tabulación o extracción de información, esto nos permitió tener más claridad acerca de la opinión subjetiva de cada experto y darle relevancia para el objetivo principal que se quería cumplir.

[17] `https://datatables.net/`

Adicionalmente, se realizó una prueba de usabilidad (Anexo D), la cuál permite medir el nivel de usabilidad de la visualización de manera cuantitativa; para ello se usó una escala en la que el usuario podía calificar cada ítem de usabilidad como se cumple de manera muy pobre hasta es excelente.

Así la prueba usabilidad consta de 10 apartados y 41 preguntas. Los resultados de los usuarios permitieron medir de manera cuantitativa la apreciación sobre que tan bien fue construido el sistema y nos ayudó a la identificación de falencias y puntos críticos que se deben mejorar en cuanto a experiencia de usuario.

Como parte de la validación interna, se implementó el algoritmo *SimpleMIL*. En este algoritmo, usa una función de transformación para simplificar en una, todas las instancias de una bolsa [1], y convertir el problema de clasificación MIL en un problema de clasificación tradicional. Las funciones de transformación que comúnmente se usan corresponden a calcular el promedio de las instancias de cada bolsa u obtener los valores máximos y mínimos entre ellas [2, 19]. Para nuestra implementación de *SimpleMIL*, se decidió usar el clasificador *k-nearest neighbors* (k-NN) [89] y como función de transformación se empleó el promedio de las instancias. Los resultados de clasificación de este algoritmo se compararon realizando diferentes cambios al conjunto de datos con base en la visualización propuesta. De esta forma se verificó si la visualización daba información al usuario que permitiera reducir la dimensionalidad del conjunto de datos sin perder información importante del mismo.

Se ha decidido usar el algoritmo *SimpleMIL* sobre otros algoritmos de clasificación MIL como son *mi-SVM* o *Citation-kNN*, debido a la facilidad de implementación y, ya que la herramienta no pretende ser usada como un clasificador de conjuntos de datos de MI se optó por la opción más sencilla de implementar. Además, *SimpleMIL* es solo usado para comprobar el cambio en el rendimiento al eliminar algunas características basados en los resultados de la visualización pese a que se sabe que no es el mejor clasificador para todos los conjuntos de datos usados aunque puede servir como referencia para que el usuario implemente otro clasificador que se ajuste mejor a sus expectativas.

Al finalizar, se ha recolectado esta información dada por los usuarios y la tomada al realizar las pruebas internas y se ha podido llegar a conclusiones interesantes acerca de la propuesta de visualización, estos resultados serán expuestos en el Capítulo 5 de resultados.

5. Resultados

5.1. Introducción

En este capítulo se expondrán los resultados obtenidos de las pruebas y encuestas realizadas sobre el sistema de visualización propuesto. El capítulo está divido en dos: resultados internos, que son los obtenidos en las pruebas realizadas sobre la herramienta; y resultados externos, que presentarán las evaluaciones realizadas por expertos sobre la propuesta de visualización.

5.2. Resultados Internos

En esta sección se evalúan los diferentes aspectos de la propuesta de visualización de conjuntos de datos de MI. Específicamente, nos centramos en evaluar una posible mejora en la comprensión del conjunto de datos midiendo el impacto de las decisiones sobre esta en el desempeño de la clasificación.

En conjunto, los algoritmos de transformación y visualización implementados proporcionan un amplio abanico de posibilidades para que el usuario pueda explorar los datos; sin embargo, dado el número posible de combinaciones posibles para realizar las pruebas, se optó por usar un número limitado de trasformaciones para mantener la consistencia de las mismas sobre los conjuntos de datos de MI utilizados. Estos últimos son: un conjunto de datos artificial creado a partir de dos Gaussianas, una de la que se generaron las instancias positivas que hacen que una bolsa sea positiva y otra de la que se generaron las instancias negativas, las cuales comparten las bolsas negativas y positivas. Este conjunto se seleccionó puesto que su simplicidad permite mejorar el entendimiento del funcionamiento de la herramienta. El segundo conjunto de datos es *Musk 1*. Este es quizá el primer conjunto de datos MIL que se propuso en la literatura y es ampliamente reconocido por los investigadores en el área. El tercero conjunto de datos es *Fox*, un conjunto de datos extraído a partir de la base de datos de imágenes de Corel en el contexto de reconocimiento de objetos en imágenes. La Tabla **5-1** resume la información de estos conjuntos de datos de prueba, estos conjuntos fueron escogidos debido a que son unos de los mas usados en las validaciones de clasificadores MIL, además, los datos de cada uno de ellos presentan características diferentes para explorar de manera visual y poder comparar como se comporta la propuesta de visualización con

diferentes tipos de conjuntos de datos de MI. La variedad de los conjuntos de datos usados permitió obtener múltiples resultados para evaluar la eficiencia de la herramienta.

Tabla 5-1.: Conjuntos de datos usados en las pruebas.

Dataset	Bolsas			Instancias			Atributos
	Positiva	Negativa	Total	Positiva	Negativa	Total	
Custom Data Gauss	10	10	20	30	30	60	3
Musk 1 [20]	47	45	92	207	269	476	166
Fox [39]	100	100	200	647	673	1320	230

Por otra parte, para no visualizar los conjuntos de datos en bruto, se emplearon tres métodos de reducción los cuales son KPCA (con un kernel lineal), Isomap y LLE. Adicionalmente, a se aplicó un método para la reducción de instancias en cada bolsa considerando la instancia más positiva y más negativa calculadas a partir de la estimación de densidad de la clase negativa basada usando el método propuesto en [2].

Como punto de referencia para la comparación en las las pruebas, se utilizó la herramienta con los datos en bruto, es decir, sin aplicar a estos ningún tipo de reducción. Además, al realizar la clasificación posterior, se ejecuto de dos maneras distintas, una de ellas excluyendo las características que según mostraba la gráfica radial tenían mayor valor en las bolsas tanto positivas como negativas que se encontraban mas cercas unas de otras, en la segunda ejecución no se hizo ningún tipo de exclusión de características. Lo anterior se realizo con el fin de tener un punto de referencia con el cual comparar y verificar si la precisión se veía afectada por estos cambios.

Después de recolectar los datos de las pruebas (Anexo F) se evidencia que no todos los métodos de reducción se pueden usar en el espacio de instancias o el espacio de bolsas de igual forma, muchas veces los datos resultantes al aplicar un método de reducción en el espacio de bolsas generan datos vacíos. En la Figura **5-1**, se muestra cual fue la proporción de las pruebas realizada en cada uno de los espacios.

Con este primer gráfico podemos inferir que los conjuntos de datos en los cuales la visualización se realiza a nivel de bolsas no es fácil de realizar, como se explico anteriormente es debido a qué se generan datos vacíos qué a su vez provocan errores en el modelo de visualización planteado por falta de datos.

Por otra parte, en las pruebas realizadas se hicieron comparaciones contra la precisión del algoritmo *SimpleMIL* cuando se excluían algunos atributos de los conjuntos de datos de MI. Esto permitió ver que en algunos casos se puede mejorar la clasificación si los atributos excluidos son los adecuados, en caso contrario la precisión del algoritmo disminuye. Esto se evidencio para los tres algoritmos usados, sin embargo es más notorio en los conjuntos *Musk 1* y *Fox* que en el *Gaussiano*.

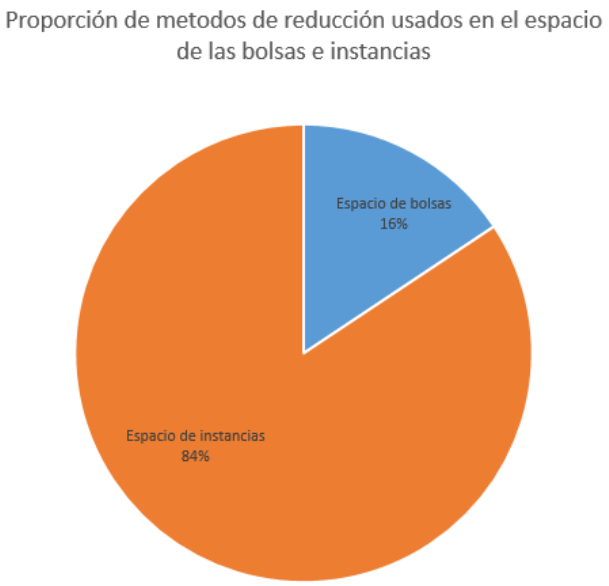


Figura 5-1.: Proporción métodos aplicados en los diferentes espacios de los conjuntos de datos de MI (*Instance Space* y *Bag Space*)

La exclusión de ciertas características en la visualización se hizo con base en la visualización obtenida de cada uno de ellos. Para ello, primero se utilizó la visualización basada en grafos la cual permitió detectar los grupos de bolsas que estaban relacionadas, de acuerdo a una métrica de distancia, similar a como lo hace el algoritmo de clasificación MIL, *Citation K-NN*. Por ejemplo, para el caso del conjunto de datos *Musk 1*, la Figura **5-2** muestra su representación basada en grafos después de reducir sus dimensiones a 83 usando el método de KPCA, se selecciono la mitad de las características que posee el conjunto de datos porque de esta manera al probar con otros conjuntos de datos se hizo lo mismo y se mantiene una consistencia en todas las pruebas realizadas, además por temas de rendimiento de la aplicación usar las 166 características no resultaba óptimo puesto que ya se contaba con los valores en bruto con todas las características para realizar la comparación. Después, a partir del grafo, se seleccionaron aquellas bolsas positivas y negativas que se encuentran conectadas y cercanas unas a otras. A partir de esta selección, se actualizó el gráfico de radar y se obtuvo la imagen de la Figura **5-3**.

En el gráfico de radar de la Figura **5-3**, hay dos imágenes, la de la izquierda muestra los valores de las características de las instancias en las bolsas positivas, mientras que el de la derecha visualiza los valores de las características de las instancias en las bolsas negativas. Continuando con la exclusión de algunas características se tomó como base esta representación, la cual permite identificar aquellas características con valores altos en cada uno de los gráficos. Tomamos esas características con valores cercanos a 100 como las que pueden ser las más representativas dentro del conjunto de datos y excluimos esas características del conjunto de datos para hacer una prueba de clasificación con el fin de observar como se ve

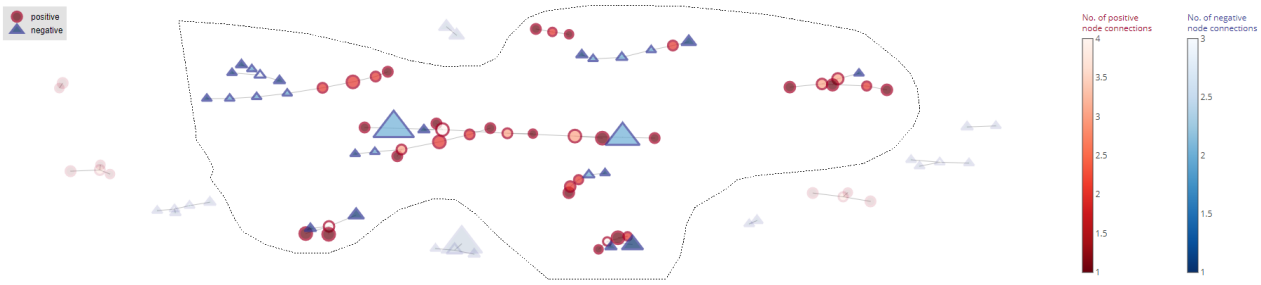


Figura 5-2.: Visualización de las bolsas de *Musk 1* por medio de un grafo

afectada la clasificación a nivel de precisión sin estas características.

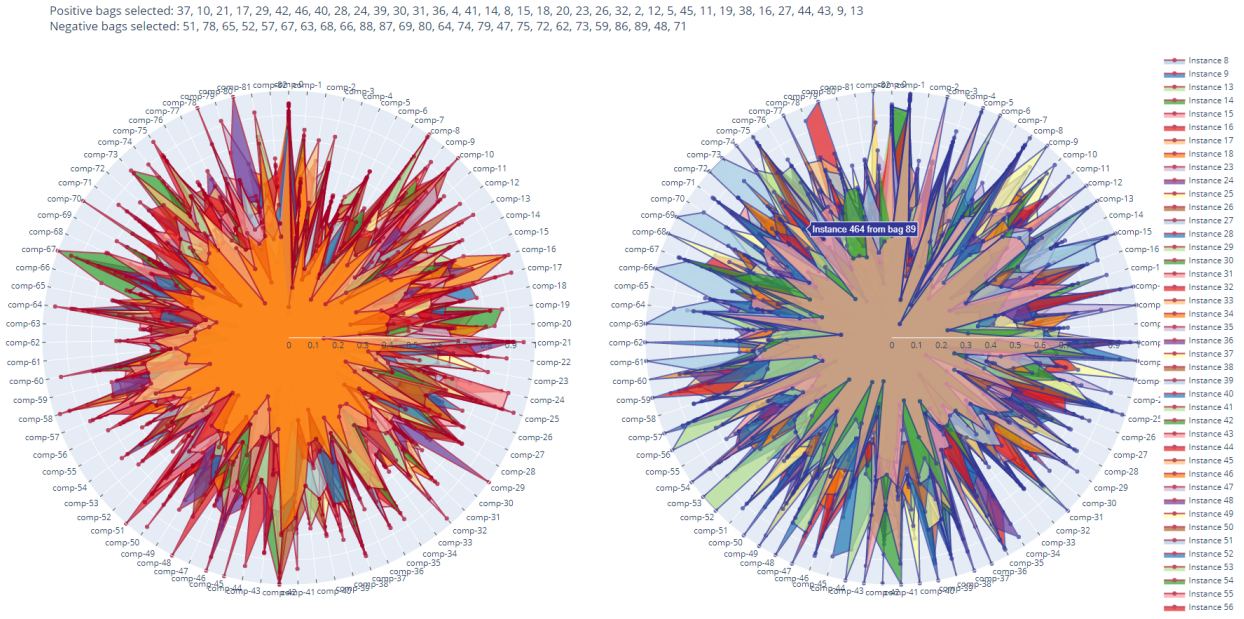


Figura 5-3.: Visualización de las instancias de *Musk 1* por medio de gráficos de radar

Al realizar lo anterior encontramos que si se retiran las características negativas que tienen valores altos se puede obtener una mejora leve en la precisión del algoritmo. Por el contrario, si se retiran las características positivas en la mayoría de los casos la precisión del algoritmo baja; estos datos se pueden ver en el Anexo F. La mejora y la disminución en la precisión en cualquier caso esta en un rango de 0 a 10 puntos, no obstante esto depende de el conjunto de datos usados, se noto que la precisión es mejor entre el conjunto de datos sea mas complejo, por el contrario en conjuntos simples como los conjuntos artificiales la mejora no es significativa, como se vera en las gráficas mas adelante.

En las gráficas de resultados de las pruebas se tiene una serie de aclaraciones para po-

der entenderlas mejor. Lo primero el clasificador usado para obtener la precisión sobre todos los conjuntos de datos fue *SimpleMIL*. Adicional, en los gráficos cada linea representa el resultado de la precisión dada por las diferentes variaciones (máximos, mínimos, promedio, extremos) del clasificador en el conjunto de datos. La nomenclatura en el eje x de los gráficos siguen el siguiente patrón *[método de reducción-espacio en el que se aplica (bolsas, BS o instancias, IS)-algoritmo simplificador (características del conjunto excluidas)]*, ejemplo de estas categorías son: KPCA-BS (F0), que indica que el método de reducción es KPCA y se aplico en el espacio de las bolsas excluyendo la característica inicial, otros ejemplos de esto son Isomap-IS-KDE (F0), LLE-IS-KDE (F23-27-35-37-47-52-60) y KPCA-IS (F0-1-3-4-6-13-19-24-26). Para terminar, la linea roja en estos gráficos representa el punto de comparación entre las diferentes precesiones obtenidas, este punto rojo es el valor obtenido de la precisión de los tres conjuntos de datos (*Gaussiano*, *Musk 1* y *Fox*) al aplicar el clasificador sin ninguna clase de transformación en los datos.

En la Tabla **5-2**[1] se muestra cómo varía la precisión promedio del clasificador *SimpleMIL* cuando se aplican los métodos de reducción de dimensiones comparado con la línea base de referencia. Se nota qué para el conjunto de datos *Gausiano*, no es muy eficiente aplicar métodos de reducción en especial Isomap y KPCA, que arrojan una precisión baja en comparación con los valores de referencia de la ultima columna y la primera fila. En cuanto a los conjuntos de datos *Musk 1* y *Fox* los métodos de reducción son un poco mas efectivos, aunque no siempre es así con las diferentes variaciones del algoritmo de clasificación. En *Musk 1* KPCA es el método con mejores resultados, mientras en *Fox* LLE e Isomap muestran buenos resultados aunque no son los mejores.

Lo anterior nos permite concluir que la reducción en conjuntos de datos de MI de pocas dimensiones puede resultar contraproducente en la clasificación, también que no todos los métodos de reducción de datos son efectivos o funcionan de igual manera en los diferentes tipos de conjuntos de datos de MI.

Continuando con los hallazgos internos, en la Figura **5-4** se presenta los resultados de la clasificación en el conjunto de datos *Gaussiano*; como se muestra en la Tabla **5-1** esta compuesto de tres dimensiones, usando el algoritmo *SimpleMIL* después de pasar el conjunto de datos por diferentes transformaciones de datos; entre ellas se usaron las técnicas de reducción KPCA, Isomap y LLE a nivel de instancia y de bolsa, asimismo en algunas pruebas se usaron estos mismos algoritmos combinados con el método de simplificación de instancias KDE. En este caso los cambios de precisión en comparación a la precisión base; que es tomado de la ejecución sin aplicar las transformaciones mencionadas anteriormente, resulta en un empeoramiento de la precisión, muchas veces bajando mas de 40 puntos. Esto nos deja que al excluir características en conjuntos de datos con pocas dimensiones la posibilidad que la precisión empeore alta, es mas, no resulta valioso aplicar técnicas de reducción de dimensiones dado que el beneficio que se obtiene no es significativo.

[1]Valores de comparación extraídos de `http://homepage.tudelft.nl/n9d04/milweb/index.html`

Tabla 5-2.: Comparación de la precisión promedio del clasificador *SimpleMIL* usando los diferentes métodos de reducción de dimensiones en los conjuntos de MI, frente a la línea base de clasificación.

Método de reducción	**SIMPLE MIL Max (Precisión)**	**SIMPLE MIL Min (Precisión)**	**SIMPLE MIL promedio (Precisión)**	**SIMPLE MIL extreme(Precisión)**	**Simple MIL comparación (Precisión)**
Gaussiano					
Base	100,00	100,00	100,00	100,00	100,00
Isomap	40,00	80,00	70,00	80,00	100,00
KPCA	72,73	72,73	63,64	68,18	100,00
LLE	94,44	100,00	83,33	88,89	100,00
Musk 1					
Base	81,67	75,00	80,00	76,67	73,70
Isomap	60,00	65,00	53,33	55,00	73,70
KPCA	73,33	80,00	80,00	75,00	73,70
LLE	68,89	70,00	54,44	62,22	73,70
Fox					
Base	50,83	53,33	56,67	55,83	63,80
Isomap	40,83	55,00	57,50	58,33	63,80
KPCA	50,00	53,33	55,00	50,83	63,80
LLE	56,67	52,78	56,11	58,33	63,80

En la Figura **5-5** la clasificación se realiza sobre el conjunto *Musk 1* y se muestra los resultados de la precisión obtenidos después de ejecutar el clasificador excluyendo algunas características de las instancias; las características excluidas son las que presentaban un valor mas cercano a 100 en la gráficas de radar explicadas anteriormente. En este conjunto de datos se ve aún más variaciones en la precisión cuando se excluyen los atributos de la clasificación y se aprecia como la inclusión o exclusión de características en las instancias afecta la precisión mejorándola o empeorándola dependiendo si se excluyen características con valores cercanos a 100 o con valores menores de 10.

Esto mismo pasa en la Figura **5-6** el cual muestra los resultados de la clasificación sobre el conjunto de datos de MI *Fox*. En este último se nota una mejora en la precisión con respecto al punto de comparación (la linea roja en la figura), dejando evidencia que realizar transformaciones de datos (reducciones de dimensiones y simplificaciones de instancias) dentro de conjuntos de datos de MI que representan imágenes es beneficioso para obtener una mejora en la precisión.

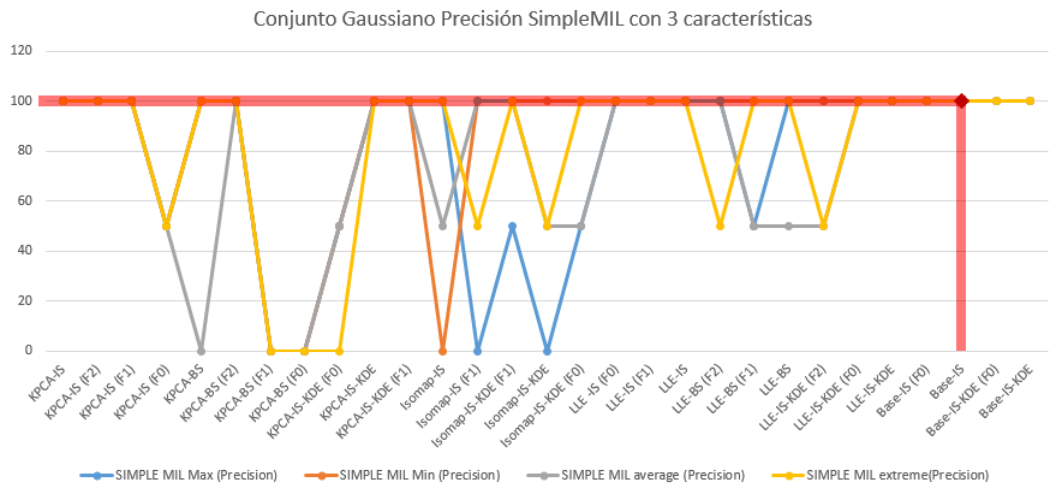


Figura 5-4.: Precisión del conjunto *Gaussiano* (La linea roja representa la precisión base con la cual se compara)

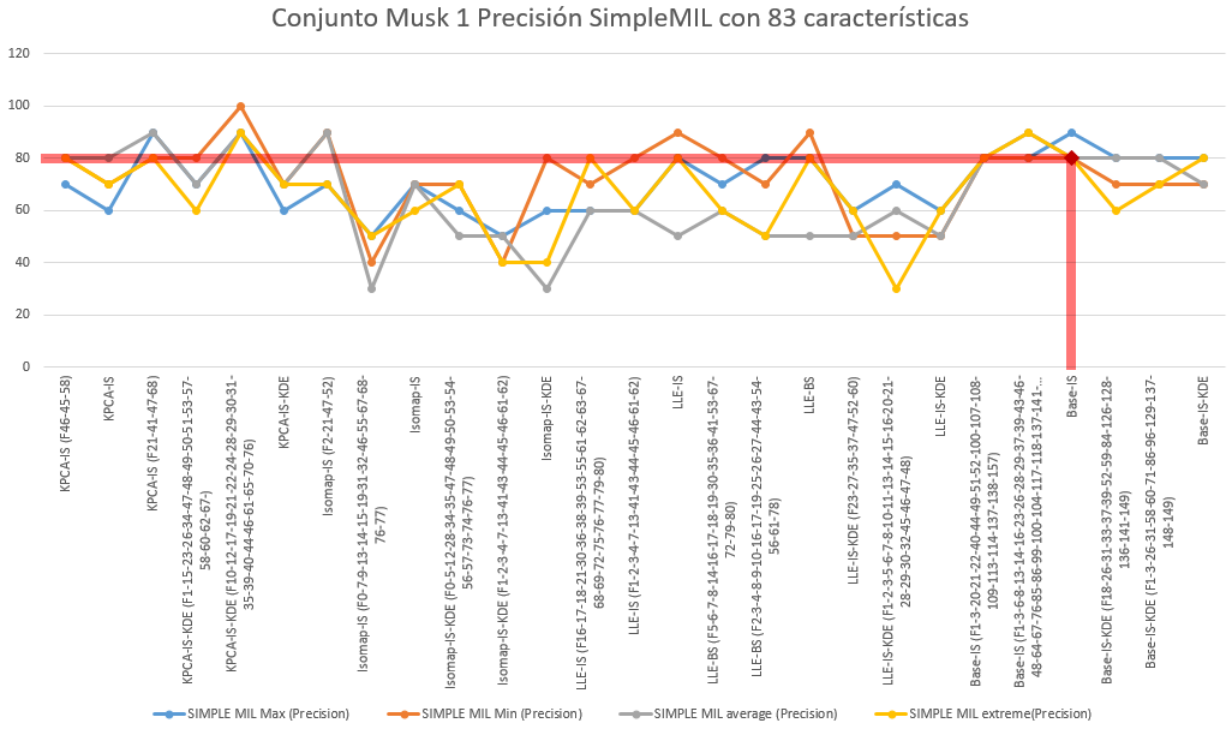


Figura 5-5.: Precisión del conjunto *Musk 1* (La linea roja representa la precisión base con la cual se compara)

En el apartado del rendimiento al aplicar cualquier método de reducción ya sea KPCA, Isomap o LLE no hay una mejora en cuanto a tiempos de ejecución promedios del clasificador *SimpleMIL*, con la excepción de la variación *extreame* del clasificador, que posiblemente se deba a la disminución de los datos en el momento de su ejecución.

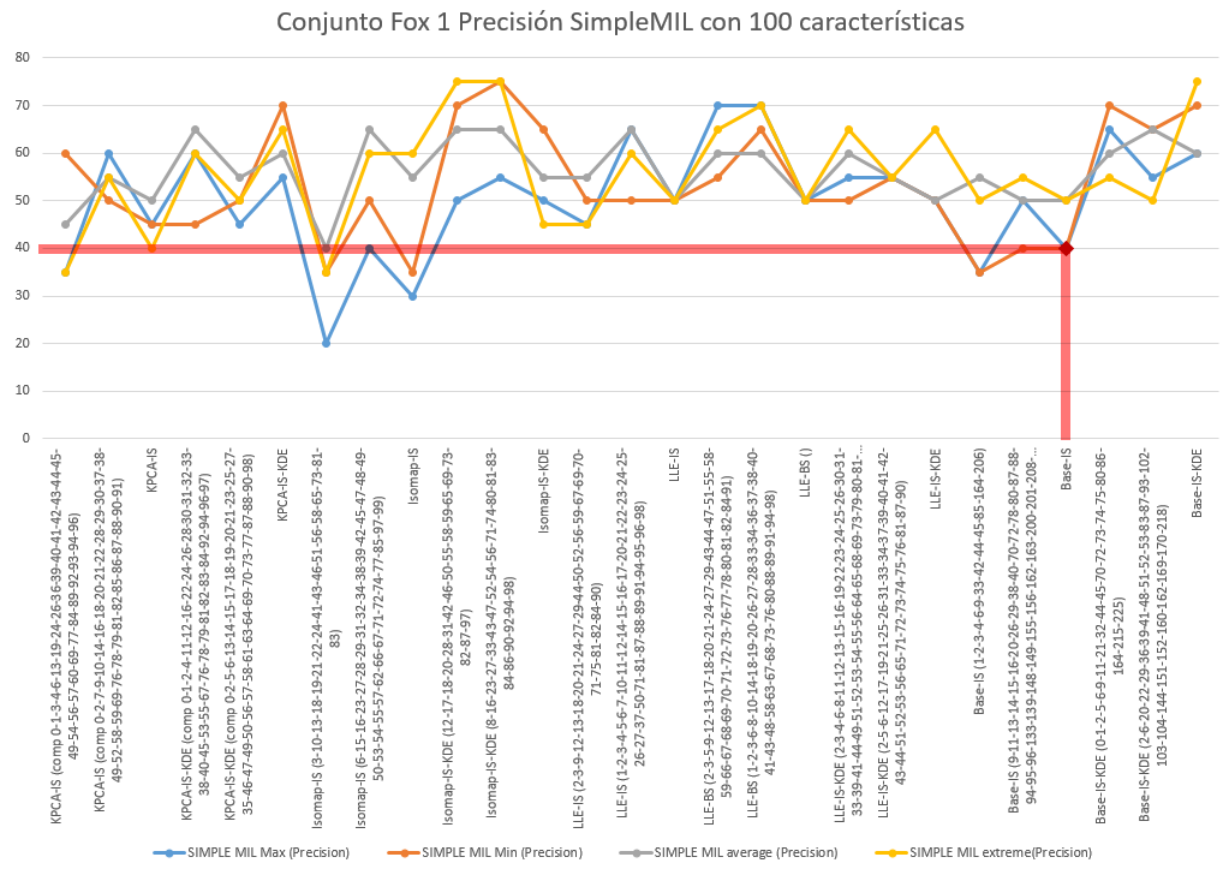


Figura 5-6.: Precisión del conjunto *Fox* (La linea roja representa la precisión base con la cual se compara)

5.3. Resultados Externos

En esta sección se presentarán las respuestas de los expertos a la encuesta y a la evaluación de usabilidad de la herramienta de visualización desarrollada. Como se mencionó, la muestra de usuarios es bastante pequeña debido a las limitaciones en cuanto a la cantidad de personas que conocen de la temática de MIL. Es por ello que los resultados de esta sección podrán ser subjetivos y puede que no refleje el uso habitual de la mayoría de expertos en MIL.

En total, los expertos que atendieron el llamado a evaluar la propuesta de visualización fueron 3, los cuales son académicos y conocedores del paradigma MIL. Una de las evaluaciones de los expertos no se completó satisfactoriamente debido a la inestabilidad en la aplicación en el momento de aplicar la prueba, pese a esto se pudo extraer información del usuario y tomar las sugerencias para realizar las mejoras correspondientes en la herramienta de visualización. En la Tabla **5-3** aparecen los resultados consolidados del puntaje de usabilidad asignado por los expertos a la herramienta. Esta evaluación se basó en los 45 principios de usabilidad de mejores prácticas descritos en [90], aunque para adaptarla mejor a la herramienta propuesta

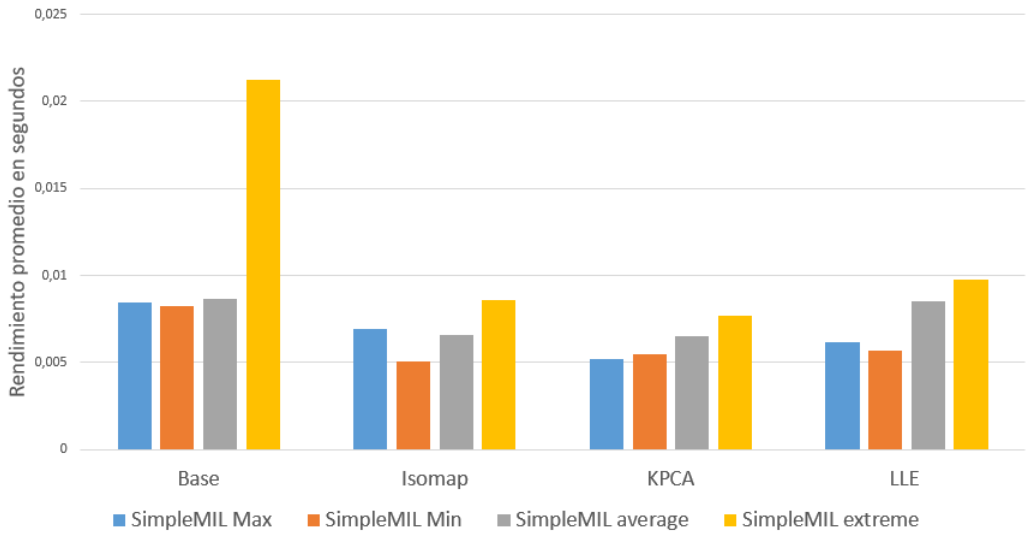


Figura 5-7.: Rendimiento promedio con los método de reducción usados en todos los conjuntos de datos

se descartaron algunas preguntas que no tenían relevancia para nuestro caso concreto.

Tabla 5-3.: Resultados revisión de usabilidad para cada experto que participó en la validación.

Experto #	Puntaje general de usabilidad	Rango del puntaje	Profesión Experto
Experto 1	80	Bueno	PhD. Ingeniería de Sistemas
Experto 2	62	Moderado	Ingeniero de Sistemas
Experto 3	69	Moderado	Profesor

Estos resultados dejan claro que aún se debe mejorar la usabilidad de la herramienta, bien sea creando instructivos más fáciles de entender o proporcionando herramientas de ayuda que vayan guiando al usuario durante el uso de la herramienta. En la Figura **5-8**, se muestra un mapa de calor en la cual se puede ver más claramente cómo cada una de las secciones evaluadas en la experiencia de usuario no lograron satisfacer por completo a los expertos. Sin embargo, hay secciones importantes como *Características y funcionalidad* (5 preguntas) que tuvieron en general una evaluación positiva, lo mismo pasa con la sección de *Navegación* (7 preguntas).

En cuanto a las secciones de *Buscar* (4 preguntas), *Formularios* (5 preguntas) y *Errores* (4 preguntas) se debe realizar una mejor implementación de estas características en la propuesta

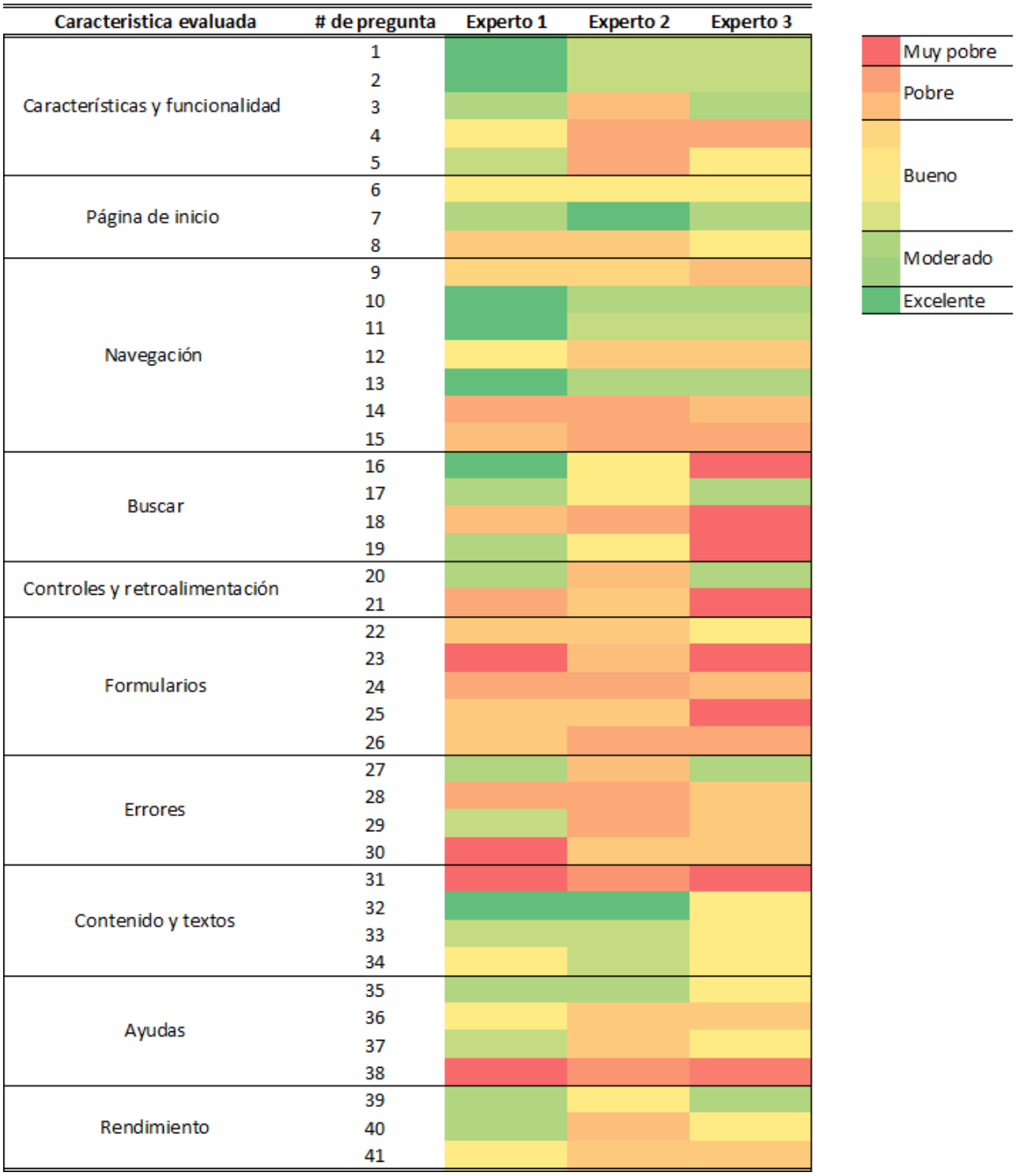


Figura 5-8.: Mapa de calor de los resultados de la experiencia de usuario

de visualización para que genere una mejor experiencia de usuario y permita hacer un mejor uso de los resultados de la visualización de los conjuntos de datos de MI. Esto deja claro que el sistema deberá mejorar en cada uno de los aspectos de usabilidad evaluados.

Para tener un mejor entendimiento de lo que los expertos opinaban de la herramienta se realizó una encuesta con 7 preguntas (Anexo C). A continuación se presentan las respuestas obtenidas en cada una de ellas por los expertos consultados.

5.3.1. Encuesta

¿Qué nivel de experticia considera que tiene en el manejo de conjuntos de datos de MI?

R1: Medio-Bajo

R2: Intermedio

R3: Medio

¿Considera herramientas como esta útiles para el proceso de toma de decisiones en el paradigma de aprendizaje de múltiples instancias (MIL)?

R1: Claro, siempre será importante la exploración inicial y visualización de los datos. Herramientas como esta ayudan a tener una idea general de nuestros datos y a realizar transformaciones que son muy útiles en el aprendizaje de MIL.

R2: Si

R3: Si

¿Los datos mostrados son precisos y útiles para el entendimiento en general del conjunto de datos de MI?

R1: Si, pero falta mayor información descriptiva y visual acerca de los datos. En las columnas de features, donde se muestra la imagen del histograma, se podría visualizar la distribución real de los datos. Para una versión posterior, la aplicación podría tener una herramienta que permita transformar los datos a partir de pequeños códigos en python programados por el usuario.

R2: Es un buen inicio

R3: Si

¿La transformación de los datos (Simplificación de bolsas, reducción de características y demás formas de transformación) ayudan a un entendimiento mejor de los datos y proporcionan un contexto adecuado?

R1: Si

R2: Si

R3: Si

¿Es necesario una forma de visualización de los datos diferentes para mejorar el entendimiento de los datos?

R1: Creo que usaron las métricas y tipos de visualización más, se podría ir alimentando con otras más especificas

R2: Si

R3: Más que una nueva forma de visualización, considero que sería de ayuda tener más información descriptiva (texto) sobre las visualizaciones y los algoritmos en la aplicación web

¿Qué revela la visualización de datos? (Búsqueda de patrones o tendencias emergentes que puedan contar una historia)

R1: No lo veo muy claro.

R2: Sin respuesta

R3: La distribución de instancias por número de bolsas, los valores de las instancias

¿La visualización responde claramente las preguntas planteadas inicialmente o plantea nuevas preguntas sobre los datos?

R1: Si, en general es una muy buena herramienta de visualización de datos MIL. La interfaz es agradable y limpia. Les sugiero que revisen la herramienta DataPrep de Google-Trifacta, aunque no es para la visualización de datos de MIL, se podría replicar algunas ideas de visualización. Probando la herramienta me sale un error, pudo haber sido por el mal manejo de la herramienta, sin embargo, no es agradable que no muestre el porqué.

R2: Nuevas preguntas dependiendo de cada conjunto de parámetros para visualizar

R3: Las visualizaciones no me parecieron claras. Es necesario acompañar las gráficas en la interfaz con texto adicional que facilite su interpretación. Las explicaciones no deberían estar únicamente en el manual de usuario

En resumen las respuestas dadas por los expertos, se puede deducir que aún hay un camino de mejoras que se deben hacer en la visualización de conjuntos de datos de MI, esto muestra también que esta primera aproximación es una buena idea y un campo de investigación que podría ser de ayuda significante en la exploración de conjuntos de datos complejos y ayudaría en un entendimiento mejor de los conjuntos de MI y los algoritmos MIL para su clasificación.

6. Conclusiones y recomendaciones

6.1. Conclusiones

Lo expuesto a lo largo del trabajo nos permite realizar diferentes conclusiones acerca del estado actual de la visualización de conjuntos de datos de MI y de su utilidad para los investigadores.

En cuanto a los métodos de visualización estudiados se encontró que existe una gran variedad de estos, los cuales son usados en diferentes ámbitos, con diversos objetivos. Esto permitió buscar métodos que se alinearan con el objetivo de realizar una propuesta para visualizar la compleja estructura de los conjuntos de datos de múltiples instancias.

Para llegar a una propuesta efectiva se realizaron diferentes etapas en la investigación y se tomó como base el ciclo que se sigue para realizar visualizaciones en conjuntos de datos multidimensionales tradicionales. Lo anterior porque estos últimos presentan estructuras similares, aunque menos complejas, que las de los conjuntos de datos de múltiples instancias. Otro punto importante que se desarrolló en la propuesta consistió en considerar la estructura interna de los conjuntos de datos (la composición de las bolsas y la relación entre sus instancias). Antes de hacer esta consideración, muchos de los métodos de visualización adaptados no resultaban ser adecuados, lo que motivó a realizar modificaciones a estos para poder tener una representación que reflejara los diferentes elementos dentro de los conjuntos de datos.

La comparación de los diferentes métodos de visualización y la caracterización de los conjuntos de datos de MI permitió encontrar formas de representar estos datos y extraer información relevante para mejorar la precisión en el entrenamiento de un algoritmo MIL. Además, posibilitó el desarrollo de una herramienta de visualización que, pese a sus debilidades y algunas opciones de mejora, puede ser de utilidad a los investigadores del área.

Como respuesta a las preguntas de investigación planteadas, se encontró que usar métodos combinados de visualización permite extraer más información del conjunto de datos, comparado a cuando sólo se utiliza un solo método de visualización. Esto motivo a que la propuesta de visualización que se implementó combinara las ventajas de una visualización de grafos, que permite ver las relaciones entre las bolsas del conjunto de datos, y un diagrama de radar, que muestra de manera resumida la distribución de las instancias en las bolsas del conjunto de datos. Además, para complementar la visualización, se agregó cierta información general del conjunto que incluye el número de bolsas e instancias y una la distribución de valores de

cada atributo.

Por otro lado, dando respuesta a la segunda pregunta de investigación, encontramos que el uso de métodos de transformación como KPCA, LLE e Isomap pueden ser de gran ayuda para disminuir efectivamente las las características del conjunto de datos a visualizar. No obstante, también se evidenció que el uso de estos métodos de transformación, dependiendo del conjunto de datos, puede tener resultados adversos en la precisión del clasificador. Esto plantea algunas incógnitas, en especial la relacionada con qué métodos de reducción podrían resultar más efectivos en conjuntos de datos de múltiples instancias. Esto es relevante para mejorar o adaptar mejor el método de visualización a un conjunto de datos específico.

Por último, se resalta que en la etapa de validación se obtuvieron resultados que estaban alineados al objetivo de esta investigación. En esta etapa se logró obtener una visión subjetiva por parte de los expertos acerca de la utilidad de contar con una herramienta que permita la visualización de conjuntos de datos de múltiples instancias, además de resaltar esta primera aproximación para lograr un sistema que permita explorar los datos de forma interactiva. Sin embargo, aún queda mucho por hacer en cuanto a mejoras en la representación de este tipo de conjuntos de datos y se deben hacer esfuerzos en encontrar métodos de visualización que representen la estructura y relaciones de los datos de mejor manera, además de investigar más acerca de las particularidades de la estructura de datos que manejan los conjuntos de datos de múltiples instancias.

6.2. Recomendaciones

Como recomendaciones, a partir de la valoración de los expertos y de las pruebas autónomas, es necesario explorar e implementar otros métodos reducción de dimensiones y proponer técnicas para disminuir el número de instancias dentro de las bolsas de los conjuntos de datos, preservando su información esencial. También, se recomienda ahondar en nuevos métodos de visualización que permitan evidenciar información de los conjuntos de datos que quizá aún está oculta.

Adicionalmente, cómo resultado de la evaluación de usabilidad, quedan por mejorar las ayudas de la herramienta para mejorar no solo el entendimiento de la misma, sino también de la visualización y sus componentes.

6.3. Trabajos futuros

Como trabajos futuros se tiene la posibilidad de mejorar el sistema de visualización propuesto siguiendo las recomendaciones dadas por los expertos, una de ellas sería hacer más intuitiva la sección de transformación de datos ya que debido a tantas opciones presentes

puede ser algo confuso de usar, además otra de las recomendaciones sería permitir ejecutar *scripts* creados por los propios usuarios para el análisis de los conjuntos de datos dentro de la herramienta y de esa forma mejorar el análisis de los datos previo a la visualización.

Otras propuestas futuras podría ser usar técnicas de visualización diferentes o adaptarlas según el tipo de conjunto de datos (imágenes, moléculas, texto, ...), ya que se demostró que dependiendo del conjunto de datos usado la representación puede verse afectada, entonces, creando métodos que se adapten al tipo o estructura del conjunto de datos de MI se podrá obtener mejores resultados y será de mucha más ayuda para los investigadores que la usen para extraer información relevante de ellos, estos cambios podrían llevarse a cabo junto con experimentos con otras técnicas de reducción o con propuestas diversas en el tratamiento de los datos.

A. Anexo: 1

Manual de usuario del sistema de visualización propuesto

En el siguiente enlace se encuentra el Manual de usuario.pdf, que ilustra el funcionamiento del sistema `http://mil-visualization.com/`

B. Anexo: 2

Protocolo de investigación de usuarios

Protocolo de investigación de usuarios.pdf

C. Anexo: 3

Encuesta

Encuesta.pdf

D. Anexo: 4

Revisión de Usabilidad

Usabilidad.xlsx

E. Anexo: 5

Artículos seleccionados para la RSL

Tabla E-1.: Artículos seleccionados para RSL

#-Ref	Artículo	Autores	Año
1-[91]	*Clustering-based multiple instance learning with multi-view feature*	He, Chengkun Shao, Jie Zhang, Jiasheng Zhou, Xiangmin	2019
2-[38]	*Incremental learning of concept drift in Multiple Instance Learning for industrial visual inspection*	Mera, Carlos Orozco-Alzate, Mauricio Branch, John	2019
3-[92]	*MIRSVM: Multi-instance support vector machine with bag representatives*	Melki, Gabriella Cano, Alberto Ventura, Sebastián	2018
4-[29]	*Multiple instance learning: A survey of problem characteristics and applications*	Carbonneau, Marc André Cheplygina, Veronika Granger, Eric Gagnon, Ghyslain	2018
5-[33]	*Multiple instance learning for credit risk assessment with transaction data*	ZHANG, Tao ZHANG, Wei XU, Wei HAO, Haijing	2018
6-[93]	*SALE: Self-adaptive LSH encoding for multi-instance learning*	Xu, Dongkuan Wu, Jia Li, Dewei Tian, Yingjie Zhu, Xingquan Wu, Xindong	2017
7-[26]	*A visual analytical approach for transfer learning in classification*	Ma, Yuxin Xu, Jiayi Wu, Xiangyang Wang, Fei Chen, Wei	2017
8-[94]	*Ensemble Extreme Learning Machine for Multi-instance Learning*	Sastrawaha, Songpon Horata, Punyaphol	2017

#-Ref	Artículo	Autores	Año
9-[95]	*Functional Annotation of Human Protein Coding Isoforms via Non-convex Multi-Instance Learning*	Luo, Tingjin Zhang, Weizhong Qiu, Shang Yang, Yang Yi, Dongyun Wang, Guangtao Ye, Jieping Wang, Jie	2017
10-[96]	*Diversified dictionaries for multi-instance learning*	Qiao, Maoying Liu, Liu Yu, Jun Xu, Chang Tao, Dacheng	2017
11-[31]	*Scalable algorithms for multi-instance learning*	Wei, Xiu Shen Wu, Jianxin Zhou, Zhi Hua	2017
12-[30]	*Supervised aggregated feature learning for multiple instance classification*	Langone, Rocco Suykens, Johan A.K.	2017
13-[97]	*Visual exploration of machine learning results using data cube analysis*	Kahng, Minsuk Fang, Dezhi Chau, Duen Horng (Polo)	2016
14-[9]	*Human action recognition with graph-based multiple-instance learning*	Yi, Yang Lin, Maoqing	2016
15-[98]	*MI-ELM: Highly efficient multi-instance learning based on hierarchical extreme learning machine*	Liu, Qiang Zhou, Sihang Zhu, Chengzhang Liu, Xinwang Yin, Jianping	2016
16-[99]	*Deep Metric Learning for Visual Tracking*	Hu, Junlin Lu, Jiwen Tan, Yap-Peng	2016
17-[25]	*Fuzzy multi-instance classifiers*	Vluymans, Sarah Tarrago, Danel Sanchez Saeys, Yvan Cornelis, Chris Herrera, Francisco	2016
18-[34]	*A multi-instance multi-label learning algorithm based on instance correlations*	Liu, Chanjuan Chen, Tongtong Ding, Xinmiao Zou, Hailin Tong, Yan	2016
19-[100]	*Instance-level accuracy versus bag-level accuracy in multi-instance learning*	Vanwinckelen, Gitte Tragante do O, Vinicius Fierens, Daan Blockeel, Hendrik	2016

#-Ref	Artículo	Autores	Año
20-[101]	*Weakly supervised activity analysis with spatio-temporal localisation*	Gu, Feng Sridhar, Muralikrishna Cohn, Anthony Hogg, David Flórez-Revuelta, Francisco Monekosso, Dorothy Remagnino, Paolo	2016
21-[10]	*Multiple instance learning with bag dissimilarities*	Cheplygina, Veronika Tax, David M.J. Loog, Marco	2015
22-[102]	*Characterizing multiple instance datasets*	Cheplygina, Veronika Tax, David M.J.	2015
23-[103]	*Robust visual tracking based on interactive multiple model particle filter by integrating multiple cues*	Dou, Jianfang Li, Jianxun	2014
24-[35]	*Multi-label image categorization with sparse factor representation*	Sun, Fuming Tang, Jinhui Li, Haojie Qi, Guo Jun Huang, Thomas S.	2014
25-[104]	*Action recognition using ensemble weighted multi-instance learning*	Chen, Guang Giuliani, Manuel Clarke, Daniel Gaschler, Andre Knoll, Alois	2014
26-[105]	*Exploring Features for Complicated Objects: Cross-View Feature Selection for Multi-Instance Learning*	Wu, Jia Hong, Zhibin Pan, Shirui Zhu, Xingquan Cai, Zhihua Zhang, Chengqi	2014
27-[8]	*TRASMIL: A local anomaly detection framework based on trajectory segmentation and multi-instance learning*	Yang, Wanqi Gao, Yang Cao, Longbing	2013
28-[106]	*Identifying user attributes through non-i.i.d. multi-instance learning*	Song, Hyun-Je Son, Jeong-Woo Park, Seong-Bae	2013
29-[107]	*MI2LS: multi-instance learning from multiple informationsources*	Zhang, Dan He, Jingrui Lawrence, Richard	2013
30-[108]	*Gaussian multiple instance learning approach for mapping the slums of the world using very high resolution imagery*	Vatsavai, Ranga Raju	2013

#-Ref	Artículo	Autores	Año
31-[109]	*MIL-SKDE: Multiple-instance learning with supervised kernel density estimation*	Du, Ruo Wu, Qiang He, Xiangjian Yang, Jie	2013
32-[110]	*Multi-instance multi-graph dual embedding learning*	Wu, Jia Zhu, Xingquan Zhang, Chengqi Cai, Zhihua	2013
33-[36]	*Multi-label multi-instance learning with missing object tags*	Shen, Yi Peng, Jinye Feng, Xiaoyi Fan, Jianping	2013
34-[8]	*TRASMIL: A local anomaly detection framework based on trajectory segmentation and multi-instance learning*	Yang, Wanqi Gao, Yang Cao, Longbing	2013
35-[111]	*Multi-instance learning with any hypothesis class*	Sabato, Sivan Tishby, Naftali	2012
36-[112]	*Multiple instance learning with missing object tags*	Shen, Yi Fan, Jianping	2011
37-[113]	*Beyond tag relevance*	Feng, Songhe Lang, Congyan Xu, De	2010
38-[114]	*An automatic feature generation approach to multiple instance learning and its applications to image databases*	Cheng, Hao Hua, Kien A. Yu, Ning	2010
39-[115]	*G3P-MI: A genetic programming algorithm for multiple instance learning*	Zafra, Amelia Ventura, Sebastián	2010
40-[37]	*Multi-instance dimensionality reduction*	Sun, Yu-Yin Ng, Michael K Zhou, Zhi-Hua	2010
41-[116]	*Generalized multi-instance learning: Problems, algorithms and data sets*	Zhang, Min Ling	2009
42-[32]	*Multi-instance learning by treating instances as non-I.I.D. samples*	Zhou, Zhi Hua Sun, Yu Yin Li, Yu Feng	2009
43-[117]	*Multi-instance Multi-label Learning for Relation Extraction*	Coronado, Eugenio Giménez-Saiz, Carlos Gómez-García, Carlos J. Romero, Francisco M.	2008
44-[118]	*Two new bag generators with multi-instance learning for image retrieval*	Liu, Wei Xu, Weidong Li, Hua Li, Guoliang	2008

F. Anexo: 6

Resultado pruebas internas

PruebasInternas.xlsx

Bibliografía

[1] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," aug 2013.

[2] C. Mera, M. Orozco-Alzate, J. Branch, and D. Mery, "Automatic visual inspection: An approach with multi-instance learning," *Computers in Industry*, vol. 83, pp. 46–54, dec 2016.

[3] F. Herrera, S. Ventura, R. Bello, C. Cornelis, A. Zafra, D. Sánchez-Tarragó, and S. Vluymans, *Multiple instance learning: Foundations and algorithms*. Cham: Springer International Publishing, 2016.

[4] W. W.-y. Chan, "A Survey on Multivariate Data Visualization," *Science And Technology*, no. June, pp. 1–29, 2006.

[5] S. Liu, W. Cui, Y. Wu, and M. Liu, "A survey on information visualization: recent advances and challenges," *The Visual Computer*, vol. 30, pp. 1373–1393, dec 2014.

[6] D. J. Janvrin, R. L. Raschke, and W. N. Dilla, "Making sense of complex data using interactive data visualization," *Journal of Accounting Education*, vol. 32, pp. 31–48, dec 2014.

[7] A. P. H. Kiyadeh, A. Zamiri, H. S. Yazdi, and H. Ghaemi, "Discernible visualization of high dimensional data using label information," *Applied Soft Computing Journal*, vol. 27, pp. 474–486, feb 2015.

[8] W. Yang, Y. Gao, and L. Cao, "TRASMIL: A local anomaly detection framework based on trajectory segmentation and multi-instance learning," *Computer Vision and Image Understanding*, vol. 117, pp. 1273–1286, oct 2013.

[9] Y. Yi and M. Lin, "Human action recognition with graph-based multiple-instance learning," *Pattern Recognition*, vol. 53, pp. 148–162, may 2016.

[10] V. Cheplygina and D. M. J. Tax, "Characterizing Multiple Instance Datasets," pp. 15–27, 2015.

[11] N. Elmqvist, P. Dragicevic, and J.-D. Fekete, "Rolling the Dice: Multidimensional Visual Exploration using Scatterplot Matrix Navigation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, pp. 1539–1148, nov 2008.

[12] N. C. Hkust, "A Survey on Multidimensional Visual Analysis Techniques Introduction – Motivation • Real world data contain multiple dimensions," 2011.

[13] S. Barlowe, T. Zhang, Y. Liu, J. Yang, and D. Jacobs, "Multivariate visual explanation for high dimensional datasets," in *VAST'08 - IEEE Symposium on Visual Analytics Science and Technology, Proceedings*, pp. 147–154, IEEE, oct 2008.

[14] T. Muhammad and Z. Halim, "Employing artificial neural networks for constructing metadata-based model to automatically select an appropriate data visualization technique," *Applied Soft Computing*, vol. 49, pp. 365–384, dec 2016.

[15] S. M. Kocherlakota and C. G. Healey, "Interactive visual summarization of multidimensional data," in *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, pp. 362–369, IEEE, oct 2009.

[16] Q. Li, L. Chen, H. Liao, and J. Yong, "PatternTrack: A Visual Pattern Detection Technique for Multidimensional Data," *2012 International Conference on Computer Science and Service System*, pp. 1360–1365, aug 2012.

[17] J. Kehrer and H. Hauser, "Visualization and visual analysis of multifaceted scientific data: A survey," mar 2013.

[18] T. Jirka, *Multidimensional Data Visualization*, vol. 34. Springer, 2003.

[19] J. Foulds and E. Frank, "A review of multi-instance learning assumptions," 2010.

[20] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1-2, pp. 31–71, 1997.

[21] N. Weidmann, E. Frank, and B. Pfahringer, "A two-level learning method for generalized multi-instance problems," in *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, vol. 2837, pp. 468–479, Springer, Berlin, Heidelberg, 2003.

[22] L. Dong, "A Comparison of Multi-instance Learning Algorithms," tech. rep., 2006.

[23] J.-D. Zucker and Y. Chevaleyre, "Solving multiple-instance and multiple-part learning problems with decision trees and decision rules. Application to the mutagenesis problem,"

[24] Y. Chen, J. Bi, and J. Z. Wang, "MILES: Multiple-Instance Learning via Embedded Instance Selection,"

[25] S. Vluymans, D. S. Tarrago, Y. Saeys, C. Cornelis, and F. Herrera, "Fuzzy multi-instance classifiers," *IEEE Transactions on Fuzzy Systems*, vol. 24, pp. 1395–1409, dec 2016.

[26] Y. Ma, J. Xu, X. Wu, F. Wang, and W. Chen, "A visual analytical approach for transfer learning in classification," *Information Sciences*, vol. 390, pp. 54–69, jun 2017.

[27] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering - A systematic literature

review," jan 2009.

[28] D. Quiñones and C. Rusu, "How to develop usability heuristics: A systematic literature review," *Computer Standards and Interfaces*, vol. 53, pp. 89–122, aug 2017.

[29] M. A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognition*, vol. 77, pp. 329–353, may 2018.

[30] R. Langone and J. A. Suykens, "Supervised aggregated feature learning for multiple instance classification," *Information Sciences*, vol. 375, pp. 234–245, 2017.

[31] X. S. Wei, J. Wu, and Z. H. Zhou, "Scalable algorithms for multi-instance learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 4, pp. 975–987, 2017.

[32] Z. H. Zhou, Y. Y. Sun, and Y. F. Li, "Multi-instance learning by treating instances as non-I.I.D. samples," *Proceedings of the 26th International Conference On Machine Learning, ICML 2009*, pp. 1249–1256, 2009.

[33] T. ZHANG, W. ZHANG, W. XU, and H. HAO, "Multiple instance learning for credit risk assessment with transaction data," *Knowledge-Based Systems*, vol. 161, no. November, pp. 65–77, 2018.

[34] C. Liu, T. Chen, X. Ding, H. Zou, and Y. Tong, "A multi-instance multi-label learning algorithm based on instance correlations," *Multimedia Tools and Applications*, vol. 75, no. 19, pp. 12263–12284, 2016.

[35] F. Sun, J. Tang, H. Li, G. J. Qi, and T. S. Huang, "Multi-label image categorization with sparse factor representation," *IEEE Transactions on Image Processing*, vol. 23, no. 3, pp. 1028–1037, 2014.

[36] Y. Shen, J. Peng, X. Feng, and J. Fan, "Multi-label multi-instance learning with missing object tags," *Multimedia Systems*, vol. 19, no. 1, pp. 17–36, 2013.

[37] Y.-Y. Sun, M. K. Ng, and Z.-H. Zhou, "Multi-Instance Dimensionality Reduction," pp. 587–592.

[38] C. Mera, M. Orozco-Alzate, and J. Branch, "Incremental learning of concept drift in Multiple Instance Learning for industrial visual inspection," *Computers in Industry*, vol. 109, pp. 153–164, 2019.

[39] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support Vector Machines for Multi ple-Instance Learning," tech. rep., 2003.

[40] Y. Chen, "Multiple-Instance Learning via Embedded Instance Selection," tech. rep.

[41] W. S. Cleveland, R. Mcgill, and S. Cleveland, "The Many Faces of a Scafferplot," *Faces*, vol. 79, no. 388, pp. 807– 822, 2011.

[42] A. Inselberg, "The plane with parallel coordinates," *The Visual Computer*, vol. 1,

pp. 69–91, dec 1985.

[43] P. Hoffman, "Table Visualization: A formal model and its applications," 1999.

[44] E. Kandogan, "Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions," *In Proceedings of the IEEE Information Visualization Symposium, Late Breaking Hot Topics*, vol. 650, pp. 9—-12, 2000.

[45] R. Rao and S. K. Card, "The table lens," pp. 318–322, Association for Computing Machinery (ACM), 1994.

[46] D. A. Keim and H. P. Kriegel, "Visualization techniques for mining large databases: A comparison," *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, no. 6, pp. 923–938, 1996.

[47] D. A. Keim, H. P. Kriegel, and M. Ankerst, "Recursive pattern: a technique for visualizing very large amounts of data," in *Proceedings of the IEEE Visualization Conference*, pp. 279–286, 1995.

[48] D. A. Keim and H.-P. Kriegel, "VisDB: Database Exploration Using Multidimensional Visualization," tech. rep., 1994.

[49] M. Ankerst, D. Keim, and H. Kriegel, "'Circle Segments': A Technique for Visually Exploring Large Multidimensional Data Sets," *Proc. IEEE Visualization '96, Hot Topic Session*, pp. 5–8, 1996.

[50] D. Keim, M. C. Hao, J. Ladisch, M. Hsu, and U. Dayal, "Pixel Bar Charts : A New Technique for Visualizing Large Multi-Attribute Data Sets without Aggregation," tech. rep.

[51] T. Mihalisin, J. Timlin, and J. Schwegler, "Visualization and analysis of multi-variate data: A technique for all fields," in *Proceedings of the 2nd Conference on Visualization 1991, VIS 1991*, pp. 171–178, 1991.

[52] J. LeBlanc, M. Ward, N. W. o. t. F. I. C. on ..., and undefined 1990, "Exploring n-dimensional databases," *ieeexplore.ieee.org*.

[53] S. Feiner and C. Beshers, "Visualizing n-dimensional virtual worlds with n-vision," in *Proceedings of the 1990 Symposium on Interactive 3D Graphics, I3D 1990*, pp. 37–38, Association for Computing Machinery, Inc, feb 1990.

[54] W. Wang, H. Wang, G. Dai, and H. Wang, "Visualization of large hierarchical data by circle packing," in *Conference on Human Factors in Computing Systems - Proceedings*, vol. 1, pp. 517–520, 2006.

[55] H. Chernoff, "The use of faces to represent points in k-dimensional space graphically," *Journal of the American Statistical Association*, vol. 68, no. 342, pp. 361–368, 1973.

[56] W. S. Cleveland and R. McGill, "Graphical perception: Theory, experimentation, and application to the development of graphical methods," *Journal of the American Sta-*

tistical Association, vol. 79, no. 387, pp. 531–554, 1984.

[57] R. M. Pickett, "Iconographic Displays For Visualizing Multidimensional Data Computational geometry View project Information Visualization View project," *researchgate.net*.

[58] J. B. P. o. t. F. I. C. On and undefined 1990, "Shape coding of multidimensional data on a microcomputer display," *ieeexplore.ieee.org*.

[59] H. L. P. o. t. N. C. on Visualization' and undefined 1991, "Color icons: Merging color and texture perception for integrated visualization of multiple parameters," *dl.acm.org*.

[60] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2001*, pp. 341–346, Association for Computing Machinery, 2001.

[61] Y. Xiao, N. Rodriguez, and O. Strauss, "Proceedings of the IADIS International Conference Computer Graphics, Visualization, Computer Vision and Image Processing 2013, CGVCVIP 2013," 2013.

[62] S. Liu, D. Maljovec, B. Wang, P. T. Bremer, and V. Pascucci, "Visualizing High-Dimensional Data: Advances in the Past Decade," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 3, pp. 1249–1268, 2017.

[63] E. Diday, "An Introduction to Symbolic data Analysis and its Application to the Sodas Project," *Revista de Matemática: Teoría y Aplicaciones*, vol. 7, no. 1-2, p. 1, 2012.

[64] A. Maalej, N. Rodriguez, A. Maalej, N. Rodriguez, and R. Nancy, "Survey of multidimensional visualization techniques To cite this version :," *CGVCVIP'12: Computer Graphics, Visualization, Computer Vision and Image Processing Conference*, p. 11, 2012.

[65] S. Ribecca, "The Data Visualisation Catalogue," pp. 1–4, 2015.

[66] G. M. Draper, Y. Livnat, and R. F. Riesenfeld, "A survey of radial methods for information visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 5, pp. 759–776, 2009.

[67] B. Filipič and T. Tušar, "A taxonomy of methods for visualizing pareto front approximations," in *GECCO 2018 - Proceedings of the 2018 Genetic and Evolutionary Computation Conference*, pp. 649–656, Association for Computing Machinery, Inc, jul 2018.

[68] A. Srinivasan, S. Muggleton, and R. D. King, "Comparing the use of background knowledge by inductive logic programming systems," in *Proceedings of the 5th International Workshop on Inductive Logic Programming*, pp. 199–230, 1995.

[69] Z. H. Zhou, K. Jiang, and M. Li, "Multi-instance learning based web mining," *Applied Intelligence*, vol. 22, no. 2, pp. 135–147, 2005.

[70] Z. H. Zhou, K. Jiang, and M. Li, "Multi-instance learning based web mining," *Applied Intelligence*, vol. 22, no. 2, pp. 135–147, 2005.

[71] M. A. Carreira-Perpiñán, "A Review of Dimension Reduction Techniques *," tech. rep., 1997.

[72] X. Huang, L. Wu, and Y. Ye, "A Review on Dimensionality Reduction Techniques," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 33, no. 10, pp. 975–8887, 2019.

[73] B. Schölkopf, A. Smola, and K. R. Müller, "Kernel principal component analysis," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 1327, no. 3, pp. 583–588, 1997.

[74] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications.," *Neural networks : the official journal of the International Neural Network Society*, vol. 13, pp. 411–30, jun 2000.

[75] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[76] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, dec 2000.

[77] J. B. Kruskal, "Nonmetric multidimensional scaling: A numerical method," *Psychometrika*, vol. 29, no. 2, pp. 115–129, 1964.

[78] L. Van Der Maaten and G. Hinton, "Visualizing Data using t-SNE," tech. rep., 2008.

[79] M. E. Tipping ME and C. M. Bishop CMBishop, "Probabilistic Principal Component Analysis," tech. rep., 1997.

[80] P. O. Box, L. Van Der Maaten, E. Postma, and J. Van Den Herik, "Tilburg centre for Creative Computing Dimensionality Reduction: A Comparative Review Dimensionality Reduction: A Comparative Review," tech. rep., 2009.

[81] F. S. Tsai and K. L. Chan, "Dimensionality reduction techniques for data exploration," in *2007 6th International Conference on Information, Communications and Signal Processing, ICICS*, 2007.

[82] S. Surendran Associate, "A Review of Various Linear and Non Linear Dimensionality Reduction Techniques," tech. rep.

[83] C. Mera, M. Orozco-Alzate, and J. Branch, "Improving representation of the positive class in imbalanced multiple–instance learning," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8814, pp. 266–273, Springer Verlag, 2014.

[84] on statistics, B. S. M. applied Probability, and undefined 1986, "Kernel density estimation technique for statistics and data analysis,"

[85] J. Kim and C. D. Scott, "Robust Kernel Density Estimation," tech. rep., 2012.

[86] S. J. Sheather, "Density Estimation," *Statistical Science*, vol. 19, no. 4, pp. 588–597, 2004.

[87] S. G. Kobourov, "Spring Embedders and Force Directed Graph Drawing Algorithms," 2012.

[88] P. Gajdoš, T. Ježowicz, V. Uher, and P. Dohnálek, "A parallel Fruchterman-Reingold algorithm optimized for fast visualization of large graphs and swarms of data," *Swarm and Evolutionary Computation*, vol. 26, pp. 56–63, feb 2016.

[89] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood Components Analysis," tech. rep.

[90] N. Turner, "A guide to carrying out usability reviews -," 2011.

[91] C. He, J. Shao, J. Zhang, and X. Zhou, "Clustering-based multiple instance learning with multi-view feature," *Expert Systems with Applications*, no. xxxx, 2019.

[92] G. Melki, A. Cano, and S. Ventura, "MIRSVM: Multi-instance support vector machine with bag representatives," *Pattern Recognition*, vol. 79, pp. 228–241, 2018.

[93] D. Xu, J. Wu, D. Li, Y. Tian, X. Zhu, and X. Wu, "SALE: Self-adaptive LSH encoding for multi-instance learning," *Pattern Recognition*, vol. 71, pp. 460–482, apr 2017.

[94] S. Sastrawaha and P. Horata, "Ensemble extreme learning machine for multi-instance learning," *ACM International Conference Proceeding Series*, vol. Part F1283, pp. 56–60, 2017.

[95] T. Luo, W. Zhang, S. Qiu, Y. Yang, D. Yi, G. Wang, J. Ye, and J. Wang, "Functional annotation of human protein coding isoforms via non-convex multi-instance learning," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. Part F1296, pp. 345–354, 2017.

[96] M. Qiao, L. Liu, J. Yu, C. Xu, and D. Tao, "Diversified dictionaries for multi-instance learning," *Pattern Recognition*, vol. 64, pp. 407–416, 2017.

[97] M. Kahng, D. Fang, and D. H. P. Chau, "Visual exploration of machine learning results using data cube analysis," in *Proceedings of the Workshop on Human-In-the-Loop Data Analytics - HILDA '16*, (New York, New York, USA), pp. 1–6, ACM Press, 2016.

[98] Q. Liu, S. Zhou, C. Zhu, X. Liu, and J. Yin, "MI-ELM: Highly efficient multi-instance learning based on hierarchical extreme learning machine," *Neurocomputing*, vol. 173, pp. 1044–1053, 2016.

[99] J. Hu, J. Lu, and Y. P. Tan, "Deep Metric Learning for Visual Tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 11, pp. 2056–2068, 2016.

[100] G. Vanwinckelen, V. Tragante do O, D. Fierens, and H. Blockeel, "Instance-level accu-

racy versus bag-level accuracy in multi-instance learning," *Data Mining and Knowledge Discovery*, vol. 30, no. 2, pp. 313–341, 2016.

[101] F. Gu, M. Sridhar, A. Cohn, D. Hogg, F. Flórez-Revuelta, D. Monekosso, and P. Remagnino, "Weakly supervised activity analysis with spatio-temporal localisation," *Neurocomputing*, vol. 216, pp. 778–789, 2016.

[102] V. Cheplygina and D. M. J. Tax, "Characterizing Multiple Instance Datasets,"

[103] J. Dou and J. Li, "Robust visual tracking based on interactive multiple model particle filter by integrating multiple cues," *Neurocomputing*, vol. 135, pp. 118–129, 2014.

[104] G. Chen, M. Giuliani, D. Clarke, A. Gaschler, and A. Knoll, "Action recognition using ensemble weighted multi-instance learning," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4520–4525, IEEE, may 2014.

[105] J. Wu, Z. Hong, S. Pan, X. Zhu, Z. Cai, and C. Zhang, "Exploring features for complicated objects: Cross-view feature selection for multi-instance learning," *CIKM 2014 - Proceedings of the 2014 ACM International Conference on Information and Knowledge Management*, pp. 1699–1708, 2014.

[106] H. J. Song, J. W. Son, and S. B. Park, "Identifying user attributes through non-i.i.d. multi-instance learning," *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2013*, no. Mil, pp. 1467–1468, 2013.

[107] D. Zhang, J. He, and R. Lawrence, "MI2LS: multi-instance learning from multiple informationsources," in *KDD '13: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, (New York, New York, USA), pp. 149–157, ACM Press, 2013.

[108] R. R. Vatsavai, "Gaussian multiple instance learning approach for mapping the slums of the world using very high resolution imagery," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. Part F1288, pp. 1419–1426, 2013.

[109] R. Du, Q. Wu, X. He, and J. Yang, "MIL-SKDE: Multiple-instance learning with supervised kernel density estimation," *Signal Processing*, vol. 93, no. 6, pp. 1471–1484, 2013.

[110] J. Wu, X. Zhu, C. Zhang, and Z. Cai, "Multi-instance multi-graph dual embedding learning," *Proceedings - IEEE International Conference on Data Mining, ICDM*, pp. 827–836, 2013.

[111] S. Sabato and N. Tishby, "Multi-instance learning with any hypothesis class," *Journal of Machine Learning Research*, vol. 13, pp. 2999–3039, 2012.

[112] Y. Shen and J. Fan, "Multiple instance learning with missing object tags," *ACM International Conference Proceeding Series*, pp. 9–12, 2011.

[113] S. Feng, C. Lang, and D. Xu, "Beyond tag relevance: Integrating visual attention model and multi-instance learning for tag saliency ranking," *CIVR 2010 - 2010 ACM International Conference on Image and Video Retrieval*, pp. 288–295, 2010.

[114] H. Cheng, K. A. Hua, and N. Yu, "An automatic feature generation approach to multiple instance learning and its applications to image databases," *Multimedia Tools and Applications*, vol. 47, no. 3, pp. 507–524, 2010.

[115] A. Zafra and S. Ventura, "G3P-MI: A genetic programming algorithm for multiple instance learning," *Information Sciences*, vol. 180, no. 23, pp. 4496–4513, 2010.

[116] M. L. Zhang, "Generalized multi-instance learning: Problems, algorithms and data sets," *Proceedings of the 2009 WRI Global Congress on Intelligent Systems, GCIS 2009*, vol. 3, pp. 539–543, 2009.

[117] E. Coronado, C. Giménez-Saiz, C. J. Gómez-García, and F. M. Romero, "Multi-instance Multi-label Learning for Relation Extraction," *Solid State Sciences*, vol. 10, no. 12, pp. 1794–1799, 2008.

[118] W. Liu, W. Xu, H. Li, and G. Li, "Two new bag generators with multi-instance learning for image retrieval," *2008 3rd IEEE Conference on Industrial Electronics and Applications, ICIEA 2008*, pp. 255–259, 2008.