Estudio de distancias para datos mixtos para análisis visual de datos multidimensionales

Gina Lucia Muñoz Salas

Orientador: Dr Erick Gómez Nieto

Jurado:
Dra. Maria Cristina Ferreira de Oliveira – Universidade de São Paulo – Brasil
Dra. Rosane Minghim– Universidade de São Paulo – Brasil
Dr. José Eduardo Ochoa Luna – Universidad Católica San Pablo – Perú

Tesis presentada al
Departamento de Ciencia de la Computación
como parte de los requisitos para obtener el grado de
Maestro en Ciencia de la Computación.

Universidad Católica San Pablo – UCSP
Abril de 2019 – Arequipa – Perú

Dedicado a Dios, a quien agradezco por todo lo que soy y tengo; y a todos aquellos que, en su momento, me motivaron a salir adelante con una palabra de aliento. Realmente marcaron la diferencia.

Abreviaturas

GPLOM *Generalized Plot Matrix*

GSOM *Generalized Self Organizing Map*

GViSOM *Generalized Visualization-Induced Self Organizing Map*

IDMAP *Interactive Document Map*

LAMP *Local Affine Multidimensional Projection*

LLE *Locally Linear Embedding*

LSP *Least square projection*

MDS *Multi Dimensional Scaling*

PCA *Principal Component Analysis*

PLP *Parallel Coordinate Plot*

SNE *Stochastic Neighbor Embedding*

SMC *Simple Matching Coefficient*

SOM *Self-Organizing Map*

SPLOM *Scatter Plot Matrix*

t-SNE *t-Distributed Stochastic Neighbor Embedding*

ViSOM *Visualization-Induced Self Organizing Map*

Agradecimientos

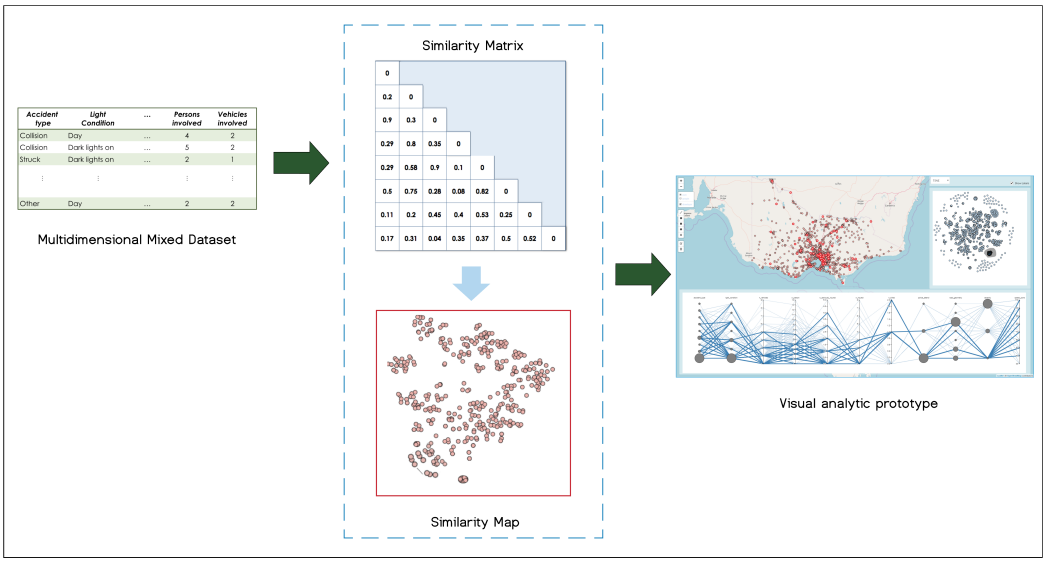
En primer lugar, gracias a Dios por estar presente en cada momento de mi vida, guiando y bendiciendo mi camino.

Gracias a mi familia. A mi madre Gina Salas y a mi padre Dember Muñoz, por todo su amor, sacrificio, apoyo incondicional y confianza. Ustedes han sido mi motor este tiempo, este logro es tan suyo como mío. Gracias a mi abuelos, a mis hermanos Carlos, Guillermo y Ángel, a mis tíos, tías, primos, primas, por sus oraciones, y por todas las muestras de cariño y aliento. Gracias a Rebequita, por su gracia y ternura. Agradezco de forma muy especial a Roger, por inspirarme a ser mejor persona y profesional cada día, por tu amor, paciencia, comprensión y por ser mi *partner in crime* todos estos años. Gracias a Pina, por acompañarme durante las largas noches de trabajo.

Gracias a la Universidad Católica San Pablo (UCSP), y al programa de maestría en Ciencia de la Computación, por permitirme realizar estos estudios, y por todas las oportunidades brindadas durante estos años. Gracias a mis docentes de maestría, por trasmitirme su conocimiento y experiencia, me dieron herramientas para poder continuar con mi vida profesional. Agradezco de forma muy especial al Prof. Dr. Erick Gómez Nieto, por haber elegido ser mi orientador, por su guía, paciencia y disponibilidad para dialogar, buscar nuevas ideas y soluciones durante la investigación. Deseo agradecer también de forma muy especial a la Prof. Dra. Rosane Minghim, de la Universidad de Sao Paulo, por recibirme como alumna visitante y brindarme su apoyo durante estos años de estudio.

Deseo agradecer de manera especial al Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica (CONCYTEC) y al Fondo Nacional de Desarrollo Científico, Tecnológico e Innovación Tecnológica (FONDECYT-CIENCIACTIVA), que mediante Convenio de Gestión 234-2015-FONDECYT, han permitido la subvención y financiamiento de mis estudios de Maestría en Ciencia de la Computación en la UCSP.

Abstract



Real world data may include multiple data types, such as numerical and categorical. Finding ways to handle these different values has become one of the current targets of research in data mining and visualization. In this work, we have studied the consequences of different mixed-type similarity measures on the visual mappings of multidimensional data. Our study focuses on how these measurements will perform when combining them with well-known multidimensional projection techniques, which are frequently the choice for providing a visual mechanism to discover information in multidimensional spaces. We have applied several metrics, namely, silhouette coefficient, neighborhood preservation, and projection stress on the projections of nine different data sets in order to evaluate the different distance measures, in terms of both segregation and similarity preservation. Finally, we show a case study on urban data that illustrates the need for relying on such measures. Based on the analyses we provide recommendations on the application of similarity measures for mixed-type multidimensional data sets in visual analysis tasks.

Keywords: Multidimensional data, mixed data, similarity, multidimensional projection processing.

Resumen

Los datos encontrados en conjuntos reales pueden incluir múltiples tipos de datos, como numéricos y categóricos. Encontrar formas de manejar estos diferentes valores se ha convertido en uno de los objetivos actuales de la investigación en minería y visualización de datos. En este trabajo, se ha estudiado las consecuencias de diferentes medidas de similitud de tipo mixto en mapas visuales de datos multidimensionales. El estudio se centra en analizar el impacto de estas medidas combinándolas con técnicas de proyección multidimensionales conocidas, que con frecuencia son la opción al proporcionar un mecanismo visual para descubrir información en espacios multidimensionales. Se aplicó las métricas coeficiente de silueta, preservación del vecindad y coeficiente de estrés en las proyecciones de nueve conjuntos de datos para evaluar las diferentes medidas de distancia, tanto en términos de segregación como de preservación de la similitud. Además, se presenta un estudio de caso sobre datos urbanos que ilustra la necesidad de confiar en tales medidas. Sobre la base de los análisis, proporcionamos recomendaciones sobre la aplicación de medidas de similitud para conjuntos de datos multidimensionales de tipo mixto en tareas de análisis visual.

Palabras clave: Datos mixtos, datos multidimensionales, similitud, proyecciones multidimensionales.

Índice general

Índice de cuadros

Índice de figuras

Capítulo 1

Introducción

1.1. Motivación y Contexto

Con los avances en la tecnología, es posible la recolección extensiva de datos. Por lo que, es evidente el gran potencial en el análisis de datos para la extracción de información útil, facilitando asi, la toma de decisiones. Sin embargo, la integración de los datos es un problema frecuente debido a la necesidad de analizar diversas fuentes en una variedad de formatos. Otro problema recurrente al realizar la integración de datos es la alta dimensionalidad, ya que cuando la dimensión aumenta, el volumen del espacio también es incrementado, multiplicando la complejidad de análisis y organización. El análisis de estos datos multidimensionales es uno de los temas más abordados en la visualización.

Generalmente, la investigación de datos multidimensionales se basa en una representación de los mismos mediante un vector de características, que es utilizado para construir diseños visuales y explorar por similitud diferentes tipos de datos, como imágenes (Li y Yu, 2014) (Joia et al., 2012) (Wang, 2009), texto (Paulovich et al., 2008) (Minghim et al., 2006), datos multimedia (Mutchima y Sanguansat, 2010) (Meghdadi y Irani, 2013) e incluso campos vectoriales (Motta et al., 2015).

Durante el proceso de extracción de características se debería producir un vector de un solo tipo que contenga valores de datos numéricos o categóricos. Estos dos tipos presentan estructuras diferentes. Por ejemplo, los datos categóricos pueden incluir datos nominales que no siguen ningún criterio de orden y resultan difíciles de manejar, sobre todo para el cálculo de la similitud (o distancia), que es una de las partes centrales de esta investigación. Para el tipo de datos continuo, existen algunas medidas conocidas como la familia de distancias de Minkowski (Cha, 2010), que son la forma más popular para calcular la distancia entre dos puntos multidimensionales. A diferencia de los datos continuos, los datos categóricos, pueden estar no ordenados, dificultando la comparación entre dos puntos de datos; no obstante, también se ha presentado medidas de similitud para datos categóricos como el coeficiente *Simple Matching Coefficient* (SMC) (Sneath et al., 1963). Sin embargo, en muchos casos no es posible separar estas características, y la tabla de atributos final contiene características

que combinan tipos numéricos y categóricos, también conocidos como datos *mixtos*. Esta estructura compuesta trae consigo un nuevo conjunto de desafíos y requisitos para la tarea de extracción de información, por ejemplo, existe la necesidad de imponer una medida de similitud precisa para comparar diferentes instancias y de un mecanismo visual para apoyar la comprensión de las razones de estas relaciones de similitud.

Tradicionalmente, se utiliza algún tipo de transformación de variables para evitar el problema de tratar con datos de tipo mixto. Estas conversiones establecen un conjunto de valores requeridos para medir la similitud, por ejemplo, utilizando la distancia euclidiana solo para datos numéricos, y la información mutua o el coeficiente de similitud de *Jaccard* para datos categóricos. Las transformaciones se pueden dar, en general, de dos diferentes formas:

- Atributos numéricos a categóricos, discretizando el rango total de datos en subintervalos; por ejemplo: 1 a 100 en tres categorías $A = [1,40]$, $B = [41,60]$ y $C = [61,100]$. Usualmente este proceso tiene como desventaja la pérdida de información de los valores numéricos.
- Atributos categóricos a numéricos, mediante la codificación con valores numéricos para cada categoría, por ejemplo, la categoría A recibe un valor de 10, la categoría B recibe 20 y así sucesivamente. Es común que las categorías codificadas de esta manera pierdan la proximidad original, estableciendo un nuevo criterio de orden, dificultando la interpretación de resultados.

Las desventajas de transformar los datos mixtos, han motivado la investigación centrada en descubrir nuevos métodos que tengan en cuenta su naturaleza mixta conservando los valores originales en los cálculos.

El procesamiento de datos de tipo mixto ha sido abordado mayormente para propósitos de *clustering*. Hsu (2006) y Hsu y Lin (2011) muestran el uso de redes neuronales *Self-Organizing Map* (SOM) para la visualización de datos multidimensionales mixtos, presentando una medida unificada para datos numéricos y categóricos, basada en árboles de jerarquía para cada atributo. Li y Biswas (2002) presentan un algoritmo de aprendizaje no supervisado para datos mixtos, mediante la distancia de *Goodall*, que basa su cálculo tanto en la frecuencia como unicidad de los atributos de los datos.

En el contexto de visualización existen algunos métodos para la exploración de datos de tipo mixto, por ejemplo DOMINO (Gratzl et al., 2014) y HEDA (Loorak et al., 2017). Estos métodos siguen un enfoque híbrido, combinando conocidas técnicas de visualización de datos numéricos o categóricos, para generar una nueva visualización que represente datos mixtos. Sin embargo, la mayoría de ellos están dedicados a mapear explícitamentecada atributo individualmente, evitando la exploración visual mediante un diseño basado en la similitud de los datos.

En este trabajo, se discute y evalúa el uso de medidas de distancia existentes para datos de tipo mixto y su impacto en las técnicas de visualización de datos multidimensionales, específicamente, en métodos de proyecciones multidimensionales, que son una opción convencional cuando el número de atributos excede la capacidad para el mapeo visual o

cuando una medida de similitud es fundamental en la tarea de análisis. Esta investigación está desarrollada en base a varias métricas para evaluar la precisión de las proyecciones proporcionadas por dichas medidas de similitud en términos de segregación y preservación de la similitud. También se presenta un estudio de caso que explora un conjunto de datos de accidentes de tráfico para validar la eficiencia de dichas medidas en tareas de análisis visual.

1.2. Objetivos

Los objetivos de la presente investigación son:

- Evaluar medidas de similitud en datos de tipo mixto para un análisis de datos multidimensionales mixtos basado en proyecciones.
- Elaborar una guía sobre la aplicación de medidas de similitud en datos de tipo mixto en alta dimensión en tareas de análisis visual.
- Implementar un prototipo para visualizar datos de tipo mixto asociados a coordenadas georeferenciadas con vistas enlanzadas a una proyección multidimensional y coordenadas paralelas.

1.3. Organización de la tesis

Este trabajo está estructurado de la siguiente manera. En el Capítulo 2 se describen medidas para determinar la similitud entre datos mixtos; también se presentan métodos de proyección y visualización de datos multidimensionales. En el Capítulo 3 se describen las métricas y los conjuntos de datos utilizados en el estudio comparativo realizado, cuyos resultados se pueden observar en el Capítulo 4. A continuación, el Capítulo 5 detalla el prototipo realizado en la investigación, analizando un caso de estudio real. En el Capítulo 6 se presentan las conclusiones y recomendaciones para trabajos futuros resultantes de esta investigación. Para finalizar, en el Apéndice A, se explican las medidas estadísticas empleadas en el estudio realizado.

Capítulo 2

Una revisión de métodos para el procesamiento y visualización de datos multidimensionales mixtos

2.1. Consideraciones iniciales

La representación de datos multidimensionales en un espacio adecuado para su visualización resulta importante para el análisis y extracción de información. En este capítulo, se describen diferentes métodos para el procesamiento y visualización de datos. Primero, se describe cuatro medidas para calcular la similitud de los datos de tipo mixto, incluyendo la distancia euclidiana con una estrategia tradicional de codificación. Luego, se detalla diferentes métodos de proyección multidimensional, utilizados para mapear datos multidimensionales en el espacio 2D. Por último, se describen técnicas de visualización conocidas para datos multidimensionales.

2.2. Cálculo de disimilitud para datos mixtos

La naturaleza de los datos (numérica, categórica o mixta) es fundamental para decidir que enfoque se elegirá para el procesamiento de datos y visualización. A pesar que muchos conjuntos de datos tienen una naturaleza mixta, gran parte de los algoritmos de procesamiento están orientados a datos solamente numéricos o categóricos.

Para datos numéricos, se tiene diversas familias de distancias utilizadas en diferentes campos de la ciencia, siendo la más común la familia de distancias de Minkowski que engloba a distancias como la distancia Euclidiana y Manhattan. Cha (2010) presenta una taxonomía con distancias y la relación de similitud entre ellas aplicadas a la comparación de histogramas. Para datos categóricos, se utiliza varios coeficientes, entre ellos, SMC, que está definido por el número de atributos no coincidentes en proporción al número total de atributos de un objeto.

Para datos mixtos, es posible utilizar distancia Euclidiana mediante un proceso de transformación de datos. Un método común para el cálculo de distancia es por medio de Gower (Gower, 1971). Hsu (2006) propone una distancia basada en la creación de árboles de jerarquía preservando la relación semántica de las categorías. Otra estrategia es utilizar la distancia de Goodall (Li y Biswas, 2002), la cual tiene un enfoque estadístico donde los elementos menos comunes en todo el conjunto de datos tienen mayor contribución al calcular la disimilitud. Todas estas medidas de similitud son descritas a continuación.

2.2.1. Distancia Euclidiana

Para la aplicación de distancia euclidiana es necesaria la transformación de datos categóricos a numéricos. Esta se puede dar a través de dos procedimientos comunes ilustrados en la Figura 2.1. El primero es el método de codificación *1-of-k*. Este método transforma cada una de las categorías de un atributo categórico en un nuevo atributo numérico, que toma el valor de 1, si el atributo corresponde al valor del atributo categórico original y 0 en caso contrario (Hsu et al., 2016). Como desventajas de este procedimiento, se encuentran el aumento de la dimensionalidad de los datos y que esta codificación no mantiene la información de similitud semántica del atributo. El segundo método reemplaza los valores categóricos por una lista ordenada de valores numéricos, imponiendo un orden si este no existía en la codificación original.

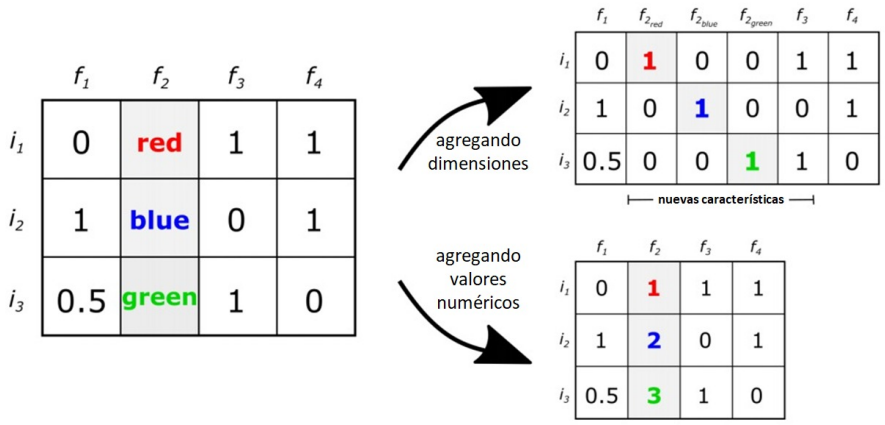


Figura 2.1: Dos formas comunes para transformar categórico a numérico para el cálculo de distancia euclidiana.

Después de la transformación de datos, se obtiene un vector multidimensional con valores numéricos, donde se puede realizar el cálculo con la distancia Euclidiana:

$$D(p,q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}, \tag{2.1}$$

donde p, q son dos puntos arbitrarios en el conjunto de datos y n es el número de dimensiones.

2.2.2. Distancia de Gower

Gower (1971) permite el cálculo de una matriz de distancias para un conjunto de datos mixtos. Para un atributo categórico k de los puntos p y q del conjunto de datos, la distancia es obtenida mediante:

$$\delta(p_k, q_k) = \begin{cases} 0, & p_k = q_k \\ 1, & p_k \neq q_k \end{cases} \tag{2.2}$$

Para un atributo numérico u, la distancia es calculada con la ecuación:

$$\delta(p_u, q_u) = \frac{|p_u - q_u|}{R_u}, \tag{2.3}$$

donde R_u es el rango del atributo u en todas sus instancias. Después de calcular las distancias parciales para cada atributo, estas son agregadas con la siguiente ecuación:

$$D(p,q) = \frac{1}{n}\sum_{i=1}^{n} \delta(p_i, q_i), \tag{2.4}$$

donde n es el número total de dimensiones.

Una ventaja de la distancia de Gower es que es realizada por un cálculo directo. Como desventaja, el cálculo requiere el cálculo de una matriz $n * n$ lo que supone un uso de memoria intensivo para grandes muestras de datos.

2.2.3. Distancia basada en jerarquías

Esta medida de distancia es propuesto en (Hsu y Lin, 2011),(Hsu, 2006) y (Hsu et al., 2016). Es un cálculo que permite encontrar la disimilitud entre los datos preservando la relación semántica de las diferentes categorías que poseen sus atributos categóricos.

Para cada atributo un árbol de jerarquía es creado. En el caso de atributos categóricos nominales, cada nodo hoja representa una categoría del atributo. Para datos categóricos ordinales o datos numéricos los puntos pueden encontrarse en las aristas del árbol. Los valores bajo el mismo nodo padre son más similares a otros puntos bajo otro padre. Cada una de las jerarquías puede ser construída manualmente o mediante técnicas de agrupación jerárquicas.

La distancia entre dos puntos es calculada según la función:

$$\delta(p_i, q_i) = d_{p_i} + d_{q_i} - 2d_{LCP(p_i,q_i)}, \tag{2.5}$$

donde d_{p_i} y d_{q_i} son las distancias de los atributos de los puntos p_i y q_i a la raíz del árbol respectivo. *LCP* es el punto en común más cercano entre p_i y q_i en el árbol de la jerarquía y $d_{LCP}(p_i, q_i)$ es la distancia del punto *LCP* a la raíz.

Las distancias $D(p_i, q_i)$ son calculadas por separado para cada atributo i para cada par de objetos (p_i, q_i) en el conjunto de datos con su estructura jerárquica correspondiente. La agregación de distancias parciales se calcula con la ecuación:

$$D(p,q) = \left(\sum_{i=1}^{n} w_i (\delta(p_i, q_i))^L \right)^{1/L}, \tag{2.6}$$

donde n es el número total de dimensiones , w_i es un peso asociado para el atributo i y L es un valor entero constante.

2.2.4. Distancia de Goodall

Goodall (1966) propuso este cálculo para la realización de taxonomías biológicas que incluyen datos mixtos. Li y Biswas (2002) generalizó el concepto para medir la similitud entre objetos cualquier dominio. La ventaja de esta técnica es considerar la frecuencia y la unicidad de los valores de los atributos. En caso de atributos categóricos, valores poco comunes en el conjunto de datos aportan mayor contribución a la similitud global de dos objetos que valores comunes entre los datos. Para el atributo categórico k, la disimilitud se calcula mediante:

$$\delta(p_k, q_k) = \begin{cases} 1, & p_k \neq q_k \\ \sum\limits_{l \in MSFVS(p_k, p_k)} \frac{(f_l)_k \cdot ((f_l)_k - 1)}{m(m-1)}, & p_k = q_k \end{cases}, \tag{2.7}$$

donde m es el número total de instancias en el dataset. $(f_l)_k$ es la frecuencia de ocurrencia del valor l con respecto al atributo k, y $MSFVS(p_k, p_k)$ es el *More Similar Feature Value Set*, es decir, el conjunto de todos los pares que tienen igual o menor frecuencia de ocurrencia que el par actual.

Para los atributos numéricos, la similitud está basada en la magnitud del intervalo y la distribución de los datos. Se calcula mediante:

$$D(p_u, q_u) = \sum_{l \in MSFSS(p_u, q_u)} (\alpha_{pq})_u, \tag{2.8}$$

donde $MSFSS(p_u, q_u)$ es el *More Similar Feature Segment Set*,es decir, el conjunto de todos los pares con menor magnitud que el par (p_u, q_u), junto con los pares con igual magnitud pero con menor número de elementos incluidos en el intervalo del par (p_u, q_u). $(\alpha_{pq})_u$ es la probabilidad de ocurrencia del par (p,q) en los datos con respecto al atributo u, dado por:

$$(\alpha_{pq})_u = \begin{cases} \frac{2\cdot(f_p)_u(f_q)_u}{m(m-1)}, & (\alpha_p)_u \neq (\alpha_p)_u \\ \frac{(f_p)_u((f_p)_u-1)}{m(m-1)}, & (\alpha_p)_u = (\alpha_p)_u \end{cases} \tag{2.9}$$

Al igual que las distancias anteriores, las distancias parciales son calculadas para cada atributo y después agregadas con la ecuación:

$$X(p,q) = 2\sum_{i=1}^{t_{CAT}} 1 - \frac{\delta(p_i,q_i)\cdot\ln(\delta(p_i,q_i)) - (\delta(p_i,q_i))'\cdot\ln(\delta(p_i,q_i))'}{\delta(p_i,q_i) - \delta(p_i,q_i)'} - 2\sum_{i=1}^{t_{NUM}} \ln(\delta(p_i,q_i)) \tag{2.10}$$

$$D(p,q) = \exp(-\frac{X(p,q)}{2}) \sum_{i=0}^{t_{NUM}+t_{CAT}-1} \frac{(\frac{1}{2})X(p,q))^i}{i!}, \tag{2.11}$$

donde t_{CAT} y t_{NUM} representan el total de atributos categóricos y numéricos respectivamente, $\delta(p_i,q_i)$ es la disimilitud de las instancias p,q en relación con el atributo i y $\delta(p_i,q_i)'$ es la siguiente diferencia menor con respecto al atributo categórico i.

2.3. Métodos de proyección multidimensional

Los métodos de proyección multidimensional reducen los datos de un espacio en alta dimensión, extrayendo una estructura con información similar, a un espacio de menor dimensión, que preserva la topología de los datos. Al reducir el espacio a 2 o 3 dimensiones, la proyección se puede visualizar como un mapa de similitud, donde se puede observar patrones con mayor facilidad, permitiendo encontrar información relevante en un sistema visual. Se ha seguido la taxonomía propuesta por Liu et al. (2014) para técnicas de reducción de dimensionalidad, clasificando los métodos en proyecciones lineales, proyecciones no lineales y proyecciones basadas en puntos de control.

2.3.1. Proyecciones lineales

Los métodos de proyección lineales utilizan combinaciones lineales para realizar la reducción al espacio de menor dimensión. A pesar de tener limitaciones comparadas con las proyecciones no lineales, presentan ventajas como mostrar propiedades genuinas de los datos y poseer ejes significativos, al ser resultado de una combinanción lineal; además, presentan de la facilidad de agregar nuevos datos sin recalcular la proyección, y en general presentan una complejidad computacional baja. (Koren y Carmel, 2003). Las proyecciones tales como *Principal Component Analysis* (PCA) (Jolliffe, 2005) y *Multi Dimensional Scaling* (MDS) clásico (Koren y Carmel, 2003) son ejemplos característicos de proyecciones lineales.

- PCA determina la proyección a un espacio de baja dimensión y se calcula encontrando una transformación ortogonal que maximice la varianza de la base del espacio resultante. Como resultado, PCA proyecta los datos en direcciones de mayor covarianza.

- Las técnica de MDS recibe una matriz de disimilitud entre los datos y genera un diseño de los datos que preserva al máximo las distancias del espacio multidimensional. Mediante la minimización de una función de costos busca generar coordenadas para el nuevo espacio buscando la mejor configuración que represente las distancias dadas. MDS no utiliza las coordenadas de los datos, a diferencia de PCA. Se utiliza el método de MDS clásico o métrico para disimilitudes cuantitativas. Para disimilitudes cualitativas, se puede emplear MDS no métrico una técnica, que aproxima un transformación monotónica no lineal de los datos, donde las distancias de los datos en el espacio proyectado corresponden a la disimilitud en el espacio original, en términos de orden, mas no intenta preservar la distancia. Para ello, se busca reducir una función de estrés mediante algoritmos de optimización iterativos como el algoritmo de *Shepard-Kruskal* (B Kruskal, 1964).

- Para grupos de documentos, la técnica *Interactive Document Map* (IDMAP) tiene un mejor rendimiento en terminos de usabilidad, generación, separación y exploración de texto (Minghim et al., 2006). La idea principal es la de generar mapas que ayuden a la exploración en las colecciones de documentos, con el fin de permitir la extracción de información de grupos de texto sin la necesidad de hacer un escaneo individual a cada documento.

2.3.2. Proyecciones no lineales

Las proyecciones no lineales son utilizadas para analizar estructuras no lineales indetectables para métodos de proyecciones lineales. Usualmente, sus salidas son difíciles de interpretar y suelen ser mucho más complejos de analizar que los algoritmos lineales. Ejemplos de estos métodos son IsoMap (Tenenbaum et al., 2000) y *Locally Linear Embedding* (LLE) (Roweis y Saul, 2000).

- *IsoMap* es uno de los métodos representativos de este conjunto , el cual esta basado en el MDS clásico pero intentando preservar la geometría intrínseca de los datos en el conjunto original para utilizar distancias de grafo en lugar de distancias euclidianas.

- Otro método no-lineal es LLE, el cual tiene un enfoque local que intenta proyectar los datos cercanos en el espacio de entrada a puntos cercanos en el espacio de salida con el uso de proyecciones lineales locales. Además, recupera la estructura global de los datos, por lo que evita cálculos de distancia de puntos separados y no es necesario el uso de una matriz de distancia.

- Por último, *t-Distributed Stochastic Neighbor Embedding* (t-SNE) (Van der Maaten y Hinton, 2008), una variación de *Stochastic Neighbor Embedding* (SNE)(Hinton y Roweis, 2003), calcula la proyección a partir de una matriz de distancia. Primero, modela una distribución de probabilidad gausiana para codificar la información de

vecindad entre los puntos originales y una distribución *t*. A continuación, intenta establecer distribución de incompatibilidad para minimizar la diferencia entre ambas distribuciones y eliminar fuerzas de atracción no deseadas, con el fin de resolver el problema de oclusión presente en SNE. El método está regulado por el parámetro de perplejidad, que se puede entender como el número de vecinos cercanos considerados al encajar las distribuciones; asi, se puede regular un enfoque local con baja perplejidad o global con alta perplejidad. En Hsu et al. (2016) se presenta una extensión de t-SNE para facilitar el uso de datos mixtos mediante una función de distancia de jerarquía.

2.3.3. Proyecciones basadas en puntos de control

Estas proyecciones resultan óptimas para el manejo de conjuntos de gran cantidad de datos complejos, donde los métodos anteriormente descritos están limitados por la eficiencia computacional. Métodos como *Least square projection* (LSP) (Paulovich et al., 2008) y *Local Affine Multidimensional Projection* (LAMP) (Joia et al., 2011) siguen un enfoque de dos etapas: primero se proyectan un conjunto de puntos de control y a continuación, se proyecta el resto de puntos teniendo las ubicaciones de los puntos del conjunto inicial como base preservando características locales, obteniendo un método de proyección escalable, que permite modificar los puntos de control seleccionados para obtener el resultado esperado.

- LSP es un método no lineal que proyecta los puntos basados en su vecindad. En el primer paso, LSP selecciona usualmente $\sqrt{n}$ puntos de control(nc), elegidos después de realizar un algorimo de clustering (*k-medoids*), agrupando los datos en tantos grupos como nc y eligiendo al punto más cercano de cada centroide. Después de la selección, los nc son proyectados en el espacio visual mediante MDS clásico.

  Como siguiente paso, se realiza la proyección del restante de puntos. Para ello, se toma en cuenta tanto las coordenadas cartesianas de los puntos nc ya proyectados, y la vecindad (relaciones locales) de cada punto en el espacio original. Esto se logra utilizando los grupos ya definidos en el *clustering* del paso anterior. Primero, se definen los k grupos más cercanos para cada grupo. Luego, para cada punto p a proyectar, se examinan el grupo al que pertenece y los grupos más cercanos, optimizando el cálculo.

- LAMP es un método robusto en cuanto al número de puntos de control necesita, pues presenta poca distorsión incluso con pocos puntos de control. Se utiliza un modelo de fuerza para colocar los puntos de control aleatoriamente en el espacio visual.

  La información de los puntos de control es utilizada para construir mapeo ortogonal afín para cada punto del resto de instancia. Este mapeo afín asegura una transformación rígida, evitando escalamiento o recortes en las distancias originales. El hecho de tener un mapeo ortogonal evita la propagación excesiva de los errores en la posición de los puntos de control. El método es interactivo ya que el mapeo sigue el diseño de los puntos de control.

  LAMP presenta ventajas sobre LSP por ser un método interactivo y requerir una menor cantidad de puntos de control. Sin embargo, LSP se puede realizar por medio de una matriz de disimilitud, por lo que se ajusta mejor a nuestro problema.

2.3.4. Proyecciones basadas en algoritmos de agrupamiento

El agrupamiento (*clustering*) de datos es una de las tareas más importantes para la minería de datos complejos. El objetivo de estas técnicas es agrupar datos no etiquetados. Para nuestro estudio, las técnicas de agrupamiento visuales como SOM (Kohonen, 1990) nos resultan relevantes, ya que permiten mapear puntos multidimensionales a un espacio 2-D preservando la topología de los mismos. A continuación se detalla la implementación de SOM y variaciones del método para análisis visual de datos multidimensionales.

2.3.4.1. *Self Organizing Map*

Un mapa SOM consiste usualmente de una grilla en 2-D regular de neuronas las que cada una cuenta con una relacion de vecindad estando conectadas adjacentemente con sus vecinas. La cantidad de neuronas determinan la precisión y la capacidad de generalización de la SOM. Durante el entrenamiento iterativo, el mapa SOM forma una red elástica que se ajusta a los datos de entrada, por lo que los puntos cercanos en los datos originales son mapeados a neuronas cercanas, preservando asi la topología (Vesanto y Alhoniemi, 2000). A pesar de esta propiedad que convierte a los mapas SOM en herramientas adecuadas para la exploración de datos multidimensionales, presenta dos limitaciones, la primera que la distancia en el mapa 2-D entrenado no refleja la distancia del espacio original de los datos, a pesar de preservar la topología, y este método solo puede ser utilizado para datos de tipo numérico.

2.3.4.2. *Visualization Induced Self Organizing Map*

Visualization-Induced Self Organizing Map (ViSOM) (Yin, 2002) es una variante del mapa SOM, la diferencia se encuentra en solucionar el problema de la preservación de la distancia original. ViSOM captura la estructura de los datos y la conserva en el mapa entrenado junto con la topología. Esto se logra tomando en consideración la distancia entre los puntos en ambos espacios y multiplicando el ratio entre la distancia en el espacio de los datos y espacio visual con un parámetro de resolución λ. El parámetro λ indica la distancia en el mapa deseada entre dos neuronas vecinas en el espacio de datos (Hsu y Lin, 2011). A menor valor de λ, la resolución del mapa será mayor. Para este mapa, se puede utilizar la distancia entre neuronas se puede usar para medir la distancia entre los puntos mapeados.

2.3.4.3. *Generalized Self Organizing Map*

Como se ha detallado anteriormente, un mapa SOM está limitado a datos numéricos. En *Generalized Self Organizing Map* (GSOM) (Hsu, 2006) se plantea una variación para permitir el uso de datos del tipo categóricos o mixtos. La única modificación necesaria para añadir esta nueva caracteristica, es la de adquirir una medida de distancia para datos categóricos. GSOM implementa distancia de jerarquía, que permite el cálculo de distancia tanto de atributos

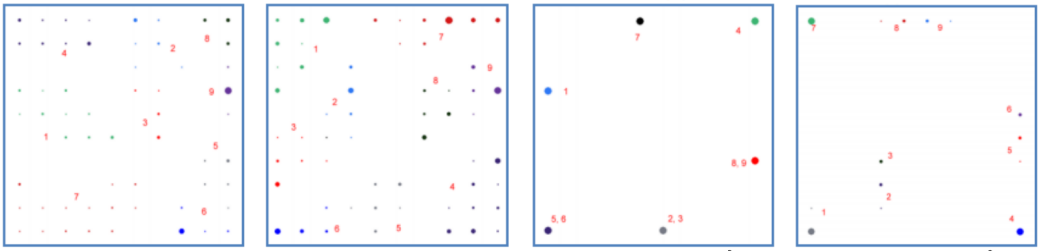


Figura 2.2: Resultados obtenidos del entrenamiento con SOM, GSOM, ViSOM y GViSOM para un conjunto de datos mixtos. Extraído de (Hsu y Lin, 2011).

númericos y categóricos. Esta distancia es integrada a SOM para facilitar el cálculo de la misma. Los detalles de esta distancia están explicados en la sección 2.1.3.

2.3.4.4. *Generalized Visualization Induced Self Organized Map*

Generalized Visualization-Induced Self Organizing Map (GViSOM) es un algoritmo de agrupación visual que mantiene la distancia del conjunto de datos multidimensional original y es posible su utilización en datos de naturaleza mixta. GViSOM es una variación del mapa SOM tradicional que hereda las caracteristicas brindadas por GSOM y ViSOM.

Los resultados obtenidos por las diferentes variantes de SOM, pueden ser comparadas en la Figura 2.2. Se puede observar que el método GViSOM obtiene un mejor agrupamiento que el resto de métodos.

Es posible procesar datos mixtos, sin alterar la naturaleza de los mismos, mediante el uso de medidas de distancia como distancia de Gower o distancia de jerarquía. En el caso de las técnicas de reducción de dimensionalidad, las técnicas que se basan en el uso de una matriz de disimilitud, pueden ser calculadas para datos mixtos mediante el uso de una matriz de distancia. De igual manera, hemos mostrado la variación del algoritmo clásico SOM con distancia de jerarquía para su funcionamiento con datos mixtos. En el Cuadro 2.1 se puede observar los dististos métodos descritos, asi como la factibilidad de aplicación sobre datos numéricos, categóricos o mixtos.

2.4. Visualización de datos

La visualización de datos tiene un rol importante al convertir el resultado del procesamiento de datos a una estructura visual para su renderizado. Los métodos más comunes para la visualización son los métodos basados en ejes donde las relaciones entre objetos se expresan a través de ejes que representan dimensiones de los datos, dimensiones proyectadas

Cuadro 2.1: Comparación de métodos de procesamiento de datos multidimensionales por tipo de dato soportado

Categoría	Subcategoría	Método	Numérico	Categórico	Matriz de disimilitud
RD	Lineales	PCA	✓	✗	✗
		MDS	✓	✗	✓
	No lineales	IsoMap	✓	✗	✓
		LLE	✓	✗	✓
		t-SNE	✓	✗	✓
	Puntos de control	LAMP	✓	✗	✗
		LSP	✓	✗	✓
Proyecciones basadas en agrupamiento		SOM	✓	✗	✓
		ViSOM	✓	✗	✓
		GSOM	✓	✓	✓
		GViSOM	✓	✓	✓

Cuadro 2.2: Comparación de métodos de visualización por tipo de dato soportado

Categoría	Método	Numérico	Categórico
Matriz de diagrama de dispersión	SPLOM	✓	✗
	GPLOM	✓	✓
Coordenadas paralelas	PCP	✓	✗
	Parallel sets	✗	✓
Radiales	*Star coordinates*	✓	✗
	Star coordinates- Orthographic constraint	✓	✗
	iStar coordinates	✓	✗
	Radviz	✓	✗
	Concentric Radviz	✗	✓
Híbrido	DOMINO	✓	✓
	HEDA	✓	✓

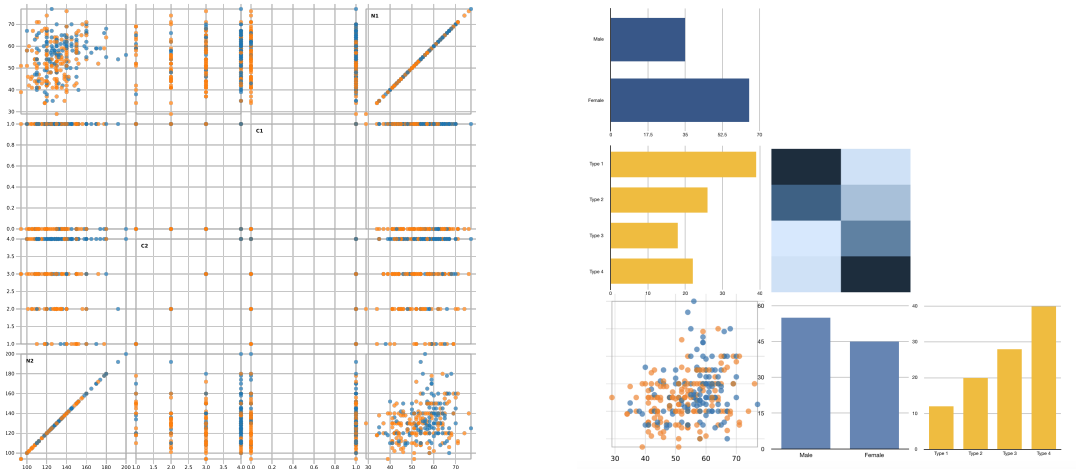
o un híbrido entre ambas (Liu et al., 2017).

A continuación se muestra ejemplos de mapeamientos visuales con métodos basado en ejes. Además, un análisis del tipo de datos que los métodos manejan puede apreciarse en el Cuadro 2.2.

2.4.1. Matriz de diagramas de dispersión

Un diagrama de dispersión permite observar la relación bidimensional entre dos atributos, originalmente numéricos. Una matriz de diagramas de dispersión o *Scatter Plot Matrix* (SPLOM) (Hartigan, 1975) es una técnica de visualización que permite observar varias relaciones bidimensionales en forma simultánea de forma simple y clara.

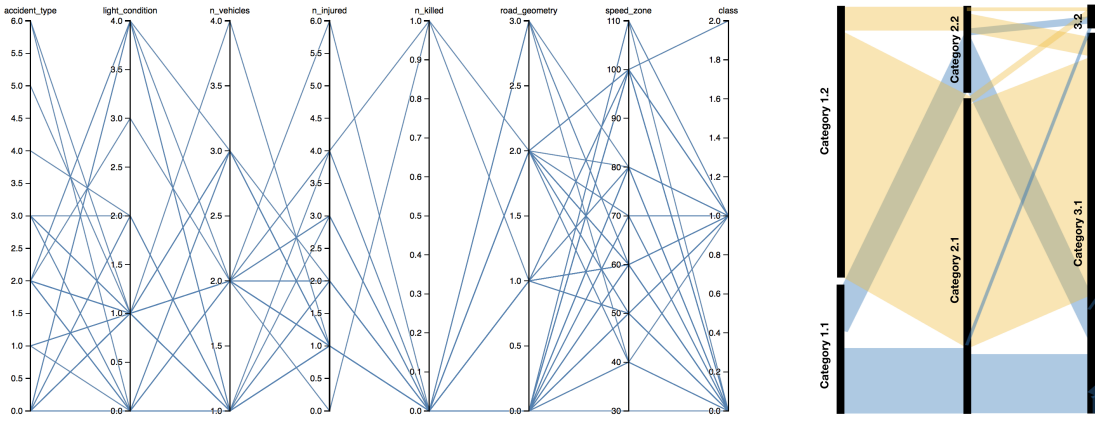
SPLOM fue extendido en Im et al. (2013) con la creación de GPLOM para lidiar con datos mixtos mediante el uso de gráficos de barras y mapas de calor en reemplazo de los diagramas de dispersión en los casos de atributos categóricos. Se puede observar la diferencia



(a) SPLOM (b) GPLOM

Figura 2.3: Comparación de visualizaciones SPLOM y GPLOM. (a) SPLOM[1]. Representación de las variables mediante *scatterplots*, (b)GPLOM. Los gráficos de barras y mapa de calor muestran datos agregados por suma para representar las relaciones de atributos categóricos y mixtos.

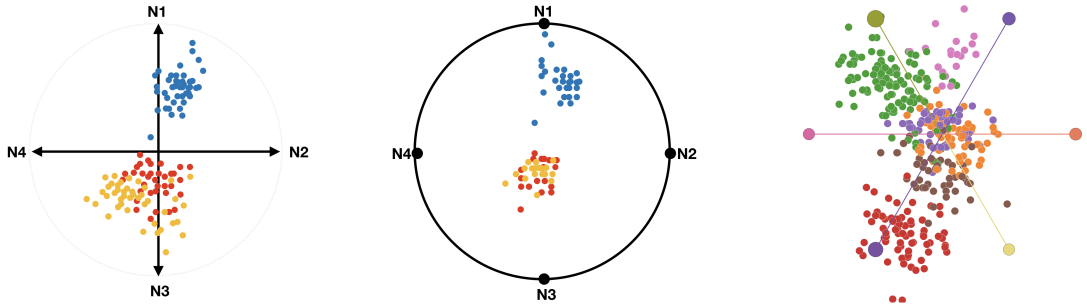
entre SPLOM y GPLOM, aplicado para el mismo conjunto de datos mixto en la Figura 2.3 , donde se puede notar en GPLOM una mejor comprensión de las comparaciones que incluyen datos categóricos, gracias a los mapas de calor y gráficos de barras aprovechando mejor el espacio visual.



(a) PLP (b) *Parallel Sets*

Figura 2.4: Comparación de visualizaciones (a) PLP[2]y (b) *Parallel Sets*.

[1]bl.ocks.org/mbostock/4063663
[1]bl.ocks.org/jasondavies/1341281



(a) *Star Coordinates* (b) *RadViz* (c) *iStar Coordinates*

Figura 2.5: Comparación de visualizaciones (a) *Star Coordinates*, (b) *Radviz* y (c) *iStar Coordinates*[3]

.

2.4.2. Coordenadas paralelas

Las coordenadas paralelas o PLP (Inselberg y Dimsdale, 1990) son una representación visual para datos multidimensionales numéricos, donde cada atributo del conjunto de datos está representado por un eje vertical, cada uno con una escala propia. Los datos se representan trazando líneas, uniendo los ejes a la altura del valor correspondiente para dicho atributo. Se puede diferenciar con los métodos de SPLOM debido a que permite observar la relación multidimensional de los datos, en lugar de relaciones bidimensionales.

El orden de los ejes influye en la obtención de información, por lo que interactuar con la visualización, reordenando los ejes, resulta importante para la exploración y descubrimiento de patrones. Sin embargo, al estar diseñado para atributos numéricos, no es común la reordenación de los valores de un mismo eje, salvo siguiendo un orden ascendente o descendente, no presente en un dato categórico.

Otra interacción importante se conoce como *brushing* que permite resaltar una o más líneas, separándolas del resto para analizar una sección específica. Esto es especialmente útil para conjuntos de datos extensos, que presentan sobreposición de líneas al visualizarlos.

Bendix et al. (2005) propuso *Parallel Sets* que adopta el diseño de PLP para manejar la naturaleza discreta de datos categóricos. En este método se sustituye los puntos individuales que representan los datos en la visualización original, por una representación basada en frecuencia, logrando asi mostrar la relación entre atributos y la frecuencia de sus categorías.

En la Figura 2.4 se puede observar la comparación entre PLP y *Parallel Sets*. En la Figura 2.4a se ilustra la visualización para un conjunto de datos mixto, cuyos atributos categóricos han sido convertidos a numéricos y asignados a los ejes en un orden específico. Se puede notar en la Figura 2.4b que en *Parallel Sets* se tiene mayor información sobre el conjunto de datos, sin embargo; no se aplica para manejar datos numéricos y categóricos simultáneamente.

[3] rics.ucsp.edu.pe/publicaciones/recursos/iStar/code/index.html

2.4.3. Métodos radiales

En esta subsección se detalla las visualizaciones que representan los datos multidimensionales numéricos como puntos en un arreglo radial, siendo las más conocidas *Star Coordinates* y *RadViz*.

2.4.3.1. *Star coordinates*

Star coordinates, presentado por Kandogan (2000), genera un mapeo lineal del espacio multidimensional a un espacio visual. Para esta visualización cada atributo numérico se representa mediante un eje con un origen común en el centro de la visualización, como se muestra en la Figura 2.5a. El rango de valores del atributo se escala a la longitud del eje, con el mínimo para el origen y el máximo para el otro extremo. Los puntos son calculados como la combinación lineal de los valores del atributo para cada instancia del conjunto de datos, es decir: $P_x = x_1v_1 + x_2v_2 + ... + x_nv_n$, donde P_x es el punto correspondiente a la instancia x en la visualización, x_i es el valor del atributo i en la instancia x y v_i representa al eje del atributo i.

Una ventaja del método es la posibilidad de variación de tamaño y orientación de los ejes, lo que le permite interactividad, pero a la vez puede ocasionar problemas de distorsión. Otro problema detectado es la oclusión de datos cuando existe una gran cantidad de dimensiones.

Una variante del método es *Ortographic Star Coordinates* (Lehmann y Theisel, 2013) que restringe el método a una proyección ortográfica para mostrar una visualización libre de distorsiones, realizando optimizaciones no lineales para cada modificación de los ejes.

Otra variante del mismo es *iStar Coordinates* (Zanabria et al., 2016), que se enfoca en solucionar los problemas de oclusión en la visualización de conjuntos con alta dimensionalidad. Para esto, *iStar Coordinates* realiza agrupamiento de atributos, reordenamiento de ejes y permite diversas interacciones para la exploración de los datos. En la Figura 2.5c se puede observar la visualización *iStar Coordinates*, donde se muestra iun conjunto de datos de 34 atributos, agrupados en 6 grupos, evitando asi la oclusión en la visualización.

2.4.3.2. *RadViz*

RadViz (Hoffman et al., 1997) como *Star Coordinates*, sigue un patrón circular. En este caso cada una los atributos estan distribuidos a lo largo del perímetro de un círculo (Figura 2.5(b)). Se sigue la metáfora de que cada punto multidimensional está unido a cada uno de los atributos en el perímetro del círculo con una fuerza igual al valor que el punto para ese determinado atributo. Cada punto se ubica donde la suma de las fuerzas es igual a cero. *Radviz* está implementado para datos numéricos. Las ventajas de este método es la capacidad de mapear un conjunto de datos de alta dimensión de una forma robusta y su capacidad de interacción al poder mover libremente los atributos a lo largo del perímetro del círculo. Una limitación del método es la oclusión de puntos en el espacio visual y que solo utiliza datos de tipo numérico.

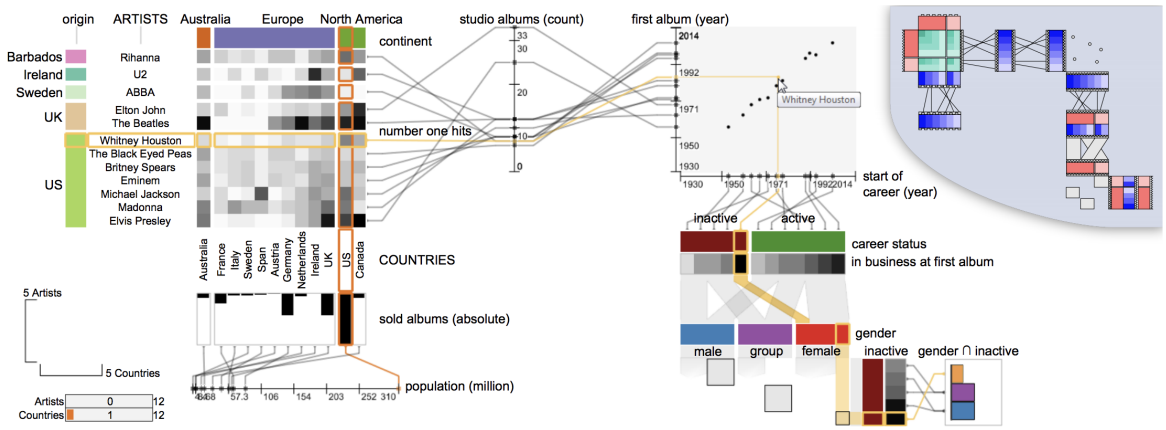


Figura 2.6: Método DOMINO para visualización subconjuntos de datos mixtos, utilizando visualizaciones como . Extraído de (Gratzl et al., 2014).

En (Ono et al., 2015) se propone una variación de *RadViz* que permite la utilización de datos categóricos. *Concentric Radviz* utiliza diferentes círculos concéntricos, uno para cada dimensión que se desee analizar. Cada círculo contiene las diferente categorías de una dimensión y permite la combinación de categorías de diferentes dimensiones para explorar los datos interactivamente. Su principal desventaja es que el número de dimensiones (Círculos concéntricos) máximo es bajo.

2.4.4. Métodos híbridos

Los métodos híbridos combinan técnicas de visualización conocidas para crear nuevas visualizaciones. Por ejemplo, Gratzl et al. (2014) proponen una técnica de meta-visualización llamada DOMINO que permite la creación de nuevas visualizaciones conectadas. El método es interactivo y permite ordenar y manipular los datos para explorar nuevas relaciones entre los subconjuntos. Permite es trabajo con datos tanto numéricos como categóricos. Se puede observar una construcción del método DOMINO en la Figura 2.6, donde se aprecia el uso de diferentes visualizaciones para un conjunto de datos mixtos. Se utilizan mapas de calor o *parallel sets* para la comparación la comparación entre atributos categóricos, diagramas de dispersión para comparar atributos numéricos. De igual forma es posible la comparación entre datos numéricos y categóricos mediante líneas.

Recientemente, Loorak et al. (2017) propuso HEDA, un componente que extiende las visualizaciones ya mencionadas a través de la manipulación de datos multidimensionales mediante la integración de visualizaciones tabulares. Sin embargo, ambos métodos mencionados anteriormente se ven fuertemente afectados cuando se trata de conjuntos de datos de alta dimensión, ya que exigen un espacio más grande de área de diseño y se dificulta la creación y análisis de la visualización al manejar gran cantidad de información, y la superposición de líneas entre los bloques, que dificulta la lectura.

2.5. Consideraciones finales

En este capítulo fueron presentadas medidas para el cálculo de disimilitud entre datos multidimensionales de tipo mixto. Generalmente, se realiza transformaciones sobre los datos para que distancias para datos numéricos, por ejemplo, la distancia euclidiana, puedan aplicarse sobre datos mixtos. Para distancias como *Gower*, jerárquica y *Goodall* no es necesario realizar dichas transformaciones para el cálculo de similitud. Además, se presentaron diferentes métodos para la reducción de dimensionalidad y visualización de datos, donde se observó cuales son adecuados para su uso con datos de tipo mixto. En el próximo capítulo se detalla la metodología seguida para la experimentación con estos métodos para datos de tipo mixto.