CEIG - Spanish Computer Graphics Conference (2018)
I. García-Fernández and C. Ureña (Editors)

A Visual Interface for Feature Subset Selection Using Machine Learning Methods

D. Rojo[1], L. Raya[1], M. Rubio-Sánchez[2] and A. Sanchez[2]

[1]Centro Universitario de Tecnología y Arte Digital (U-tad), Spain
[2]Universidad Rey Juan Carlos, Spain

Abstract

Visual representation of information remains a key part of exploratory data analysis. This is due to the high number of features in datasets and their increasing complexity, together with users' ability to visually understand information. One of the most common operations in exploratory data analysis is the selection of relevant features in the available data. In multidimensional scenarios, this task is often done with the help of automatic dimensionality reduction algorithms from the machine learning field. In this paper we develop a visual interface where users are integrated into the feature selection process of several machine learning algorithms. Users can work interactively with the algorithms in order to explore the data, compare the results and make the appropriate decisions about the feature selection process.

CCS Concepts

•Human-centered computing → Visual analytics; Visualization systems and tools; •Computing methodologies → Feature selection;

1. Introducción

La generación de conjuntos de datos masivos que cuentan con un elevado número de observaciones y atributos resulta cada vez mayor y más frecuente en muchos campos como la biología computacional [But99], la informática [GS05], la atención sanitaria [BMSD13] o la genómica [CSH*16]. La gran complejidad de estos conjuntos de datos ha provocado que en muchos casos resulte necesario el uso de algoritmos de aprendizaje automático y técnicas de minería de datos para poder extraer conocimiento de estos.

Cuando se aplican algoritmos de aprendizaje automático a dichos conjuntos de datos ocurre un problema conocido como *curse of dimensionality* [Bel61]. Este fenómeno se debe a que los datos son más dispersos en espacios de muchas dimensiones, lo cual afecta negativamente a algoritmos habituales de aprendizaje automático diseñados para espacios de baja dimensionalidad como, por ejemplo, k-Nearest Neighbours (K-NN) [FHT01].

La principal estrategia para combatir este fenómeno es realizar una reducción de dimensiones como parte del preprocesamiento de los datos. Para ello, existen dos enfoques: la transformación de atributos [MR02] y la selección de un subconjunto de atributos [LCW*17]. La transformación de atributos consiste en la proyección de las variables originales a un nuevo espacio con una menor cantidad de atributos, siendo cada uno de estos nuevos atributos una combinación lineal o no lineal de las variables originales. En cambio, la selección de atributos consiste en la eliminación de aquellas variables originales que son redundantes o irrelevantes (aquellas que no tienen influencia en la salida de nuestro algoritmo de aprendizaje automático).

Para la transformación de atributos se suelen utilizar algoritmos automáticos tanto lineales como no lineales. En cambio, la selección de atributos es un problema de tipo NP-duro [CHW97]. Por ello es frecuente utilizar la visualización de datos para permitir al analista combinar su conocimiento del dominio con su capacidad para entender visualmente las relaciones entre los atributos y actuar en consecuencia. La capacidad del ser humano para analizar y comprender visualmente la información ha convertido a la representación visual de información en un método altamente atractivo para el análisis exploratorio de datos. Si bien es cierto que existen algoritmos automáticos para la selección de atributos (e.g. Fisher Score [DHS12], Mutual Information Maximization [Lew92]), en la mayoría de las ocasiones esta selección es realizada como parte de un proceso de análisis exploratorio de datos.

En este trabajo se propone introducir la visualización de datos en todo el proceso de reducción de dimensiones, utilizando ambos enfoques (selección y transformación de atributos) de forma conjunta y coordinada a través de una herramienta visual e interactiva, que guiará al usuario en la selección de los atributos. Nuestra herramienta permite obtener una imagen global del proceso de reducción de la dimensionalidad así como una mejor comprensión de la importancia de cada una de las variables para el objetivo del algoritmo de aprendizaje automático (clasificación, regresión, detección de clusters,...).

DOI: 10.2312/ceig.20181165

El presente estudio está organizado de la siguiente manera: en la Sección 2 se describen los trabajos más relevantes que guardan relación con nuestra propuesta. En la Sección 3 se detalla la estructura de la interfaz, así como su adecuación para la selección de atributos. En la Sección 4 se comentarán algunos de los resultados obtenidos. Por último, en la Sección 5 se detallan las conclusiones sobre el uso de la interfaz y los trabajos futuros que surgen tras esta investigación.

2. Trabajos relacionados

La Visualización de Información (InfoVis) [Tuf86, CMS99, War04, Spe07] es una disciplina que se centra en el uso de herramientas informáticas para explorar volúmenes de datos abstractos. Estas aplicaciones incluyen operaciones de adquisición, selección, transformación y representación de datos, para facilitar su exploración y entendimiento por parte de analistas humanos. [HCL05] propone una de las librerías más populares para desarrollar aplicaciones de InfoVis, donde además se citan varias técnicas de visualización de información. Con posterioridad, [HBO10] hace un estudio de las técnicas de visualización más utilizadas hasta la fecha. Adicionalmente, existen múltiples trabajos en los que se proponen pautas y recomendaciones para diseñar aplicaciones de InfoVis [AES05, AS05, YKSJ07, Car08, PFG08].

En esta sección nos centramos en las técnicas de visualización de datos multidimensionales, profundizando en aquellas relacionadas con nuestra propuesta. Posteriormente, describimos las distintas interfaces visuales interactivas que hacen uso de métodos de reducción de dimensiones.

2.1. Visualización de datos multidimensionales

En el estado del arte se han propuesto diferentes métodos para la visualización de datos multidimensionales. Las diferentes propuestas se diferencian en la forma de transformar datos en representaciones visuales, así como en sus características de interacción. Mientras algunos investigadores emplean técnicas algorítmicas para la transformación o proyección de los datos, otros confían más en la interacción con el usuario para identificar patrones, agrupaciones o tendencias. El proceso de transformación visual básicamente toma n características de datos y las mapea en k atributos visuales (p.ej. posición, color, tamaño, etc.). Las técnicas de visualización se diferencian en el número de los parámetros de las funciones de transformación, así como en el número de los atributos de los elementos visuales.

Las nuevas herramientas interactivas de visualización de información son capaces de trabajar con datos de diferente naturaleza. Para visualización de datos de carácter numérico se suele recurrir a técnicas provenientes del campo de la estadística, como son los diagramas de dispersión, histogramas, diagrama de cajas, etc. Para visualizar datos de dimensión elevada las técnicas clásicas de visualización incluyen la representación de técnicas automáticas de transformación de atributos cómo *Multidimensional Scaling* (MDS) [CC00], Análisis de Componentes Principales (PCA) [HSM01], Análisis Discriminante Lineal (LDA) [McL04], *Locally Linear Embedding* (LLE) [RS00], o t-SNE [HR03, vdMH08].

En los últimos años han ido aparecido diferentes técnicas interactivas de visualización multidimensional. Dichas técnicas de visualización multivariante se pueden categorizar según diferentes criterios, incluyendo el tipo de datos, las formas de interacción [Kei02], los objetos gráficos y las disposiciones que componen las gráficas. Algunas son capaces de mostrar datos de elevada dimensión sin pérdida de información (por ejemplo, *Parallel Coordinates* [ID90, SR06, Ins09, YGX*09], o *Table Lens* [RC94]), mostrando valores de atributos exactos directamente al representar los elementos como polilíneas. Por otro lado, existen métodos basados en ejes radiales [DLR09, DBB10, RSSL17, SSRMJ*18] que generan transformaciones de datos a un espacio en 2 o 3 dimensiones, donde se puede observar con mayor facilidad. Entre ellos destacan *Star Coordinates* [Kan00, Kan01, RSRDS16] y RadViz [HGM*97, DGRG12]. Sin embargo, hay información que se pierde inevitablemente en el proceso de reducción de la dimensionalidad.

Para la interfaz visual aquí presentada, se utiliza *Star Coordinates* (SC) como método de visualización de datos multidimensionales sobre el que añadiremos la capacidad de selección de atributos. SC genera proyecciones de un espacio n-dimensional a un espacio m-dimensional ($m \leq 3$) para poder representar los datos gráficamente. La proyección queda definida por un conjunto de n vectores m-dimensionales $\mathbf{v}_i$ con un origen común, donde $\mathbf{v}_i$ está asociado con el atributo i-ésimo de los datos. La proyección $\mathbf{p} \in \mathbb{R}^m$ de una muestra $\mathbf{x} \in \mathbb{R}^n$ de los datos, viene dada por la combinación lineal de los vectores $\mathbf{v_i}$, siendo los coeficientes los valores de los atributos de $\mathbf{x}$. Es decir,

$$\mathbf{p} = x_1\mathbf{v}_1 + x_2\mathbf{v}_2 + \cdots + x_n\mathbf{v}_n = \mathbf{V}^{\mathrm{T}}\mathbf{x}, \tag{1}$$

donde $\mathbf{V}$ es la matriz de dimensiones $n \times m$ cuyas filas son los vectores $\mathbf{v_i}$.

2.2. Interfaces visuales que hacen uso de métodos de reducción de dimensiones

En las disciplinas de InfoVis y Visualización Analítica (VA) se han desarrollado múltiples interfaces visuales interactivas que incorporan diferentes métodos de reducción de dimensiones para analizar datos de elevada dimensión. La creación de estas interfaces visuales interactivas es una de las soluciones más comunes a la dificultad de seleccionar algoritmos de reducción de dimensiones por parte de los analistas, así como de configurarlos e interpretarlos correctamente [LMW*15].

La mayoría de interfaces visuales que hacen uso de métodos de reducción de dimensiones se centran en un único método [JZF*09, ML14] y raramente permiten seleccionar entre los distintos métodos existentes. Un reciente revisión del estado del arte [SZS*17] detecta, de hecho, únicamente cuatro interfaces visuales [RL15, LWBP14, MDL07, NM13] hasta la fecha que posibilitan la selección de varios métodos de reducción de dimensiones. En dicho trabajo, los autores destacan que la investigación de nuevas interfaces visuales que guíen al analista en la selección del método de reducción de dimensiones es interesante para la realización de futuros trabajos.

En concreto, *Persistent Homology* [RL15] presenta una interfaz

que permite comparar las configuraciones de diversos métodos de reducción de dimensiones y en la que se calculan varias medidas de calidad que permiten validar y ordenar cada una de las configuraciones de los diversos métodos. Al igual que la interfaz que nosotros presentamos, permite comparar diversos algoritmos de reducción de dimensiones como PCA, t-SNE o LLE; sin embargo, no provee al usuario de herramientas para realizar selección de atributos, siendo este uno de los objetivos principales de la interfaz visual que se describe en este trabajo. Esto mismo ocurre con múltiples interfaces analizadas [LWBP14, MDL07, CLL*13, BNH14, ML14].

En cambio, *Tripadvisor*$^{N\text{-}D}$ [NM13] y *GGobi* [CS07] permiten seleccionar atributos y al igual que nuestra propuesta utilizan SC como método de visualización base. Sin embargo, *Tripadvisor*$^{N\text{-}D}$ sólo permite analizar proyecciones arbitrarias obtenidas directamente por el analista en SC para ir clasificando los datos en clusteres y *GGobi* solo las proyecciones lineales especificadas por el *Grand Tour* [Asi85], en lugar de las proyecciones generadas por métodos lineales y no lineales de reducción de dimensiones de nuestra propuesta. Además, como se describe en las conclusiones de [NM13] las proyecciones lineales definidas por la continua interacción del usuario en SC (y el uso de *Grand Tour*) dificulta la interacción con la interfaz cuando se quieren visualizar datos de dimensión significativamente elevada (más de 20 dimensiones). En cambio, la utilización de algoritmos de reducción de dimensiones permite trabajar con datos de dimensión más elevada.

3. Interfaz

El objetivo de nuestra interfaz es proporcionar una herramienta visual e interactiva para permitir la selección de atributos sobre conjuntos de datos multidimensionales. Con el fin de que la visualización sea efectiva y aporte el mayor conocimiento posible para su exploración, la interfaz guía a los analistas en la selección de un conjunto de atributos significativo. Una vez indicados los atributos con los que trabajar, la herramienta permite la elección de métodos automáticos de reducción de dimensionalidad lineales y no lineales. Tras el preprocesamiento de los datos, éstos se utilizan para crear un modelo de predicción basado en algún algoritmo de clasificación supervisada.

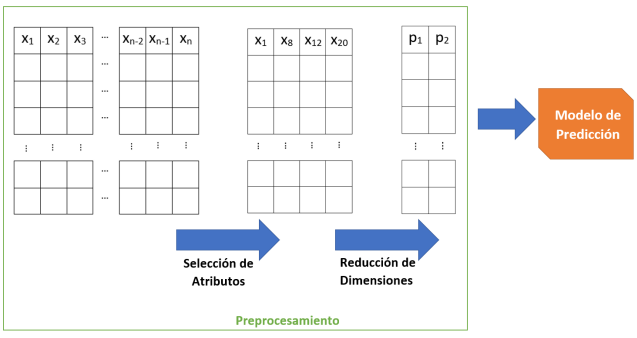

Figura 1: *Flujo de los datos hasta la generación del modelo de predicción.*

En los problemas de clasificación supervisada, cada una de las muestras está compuesta por un vector de atributos y la clase, que suele denominarse variable predictiva. El objetivo es conseguir un modelo que prediga lo mejor posible la variable predictiva en función del vector de atributos correspondiente. El flujo de los datos hasta la generación del modelo de predicción se detalla en la Fig. 1.

Adicionalmente a la selección de atributos, el uso de la interfaz proporciona a los usuarios una visión global de todo el proceso iterativo de selección de variables, permitiendo entender mejor la importancia de las distintas variables en el algoritmo de clasificación.

La interfaz se ha desarrollado íntegramente en Python utilizando pandas para la gestión de las estructuras de datos, Plotly para la creación de los distintos gráficos y sus métodos de interacción, scikit-learn para la aplicación de los distintos algoritmos de aprendizaje automático y Dash para la coordinación de todos los elementos y la creación de la aplicación e interfaz web.

La interfaz propuesta se divide en distintas zonas. En la zona 1 de la Fig. 2, la herramienta permite la visualización simultánea de múltiples métodos de reducción de dimensiones tanto lineales como no lineales en función de la selección indicada por el usuario. Será en la siguiente sección del artículo (ver Sección 3.1) donde se describa el proceso de selección de visualización para cada tipo de proyección asociada a estos métodos de reducción de dimensiones. En los mismos gráficos de esta primera zona se muestran los aciertos y errores de predicción. Al lado derecho de cada gráfico, la interfaz muestra una tabla describiendo los valores de precisión de clasificación para cada método de reducción de dimensionalidad empleado. La selección de estos valores y su uso se describe en la Sección 3.3. Debajo de esta zona 1, se representa un diagrama de barras que muestra los atributos ordenados según una medida de influencia de cada atributo (ver zona 2 de Fig. 2). En la Sección 3.2 se detalla cómo este diagrama de barras guía al usuario en la selección de atributos. Para configurar estas visualizaciones el usuario dispone de un sencillo menú en la zona 3 de Fig. 2. En este menú, el usuario puede seleccionar el conjunto de datos, los métodos de reducción de dimensiones, los hiperparámetros relevantes de estos métodos y del algoritmo de predicción, así como modificar la selección de atributos actual.

La posibilidad de interactuar con las visualizaciones resulta clave para realizar un análisis efectivo, debido a que puede ayudar al analista a formular distintas hipótesis según los resultados, sacar conclusiones sobre las visualizaciones o tomar decisiones de las mejores configuraciones a realizar. Por ello, nuestra interfaz es interactiva e intuitiva, proporcionando las tareas habituales del conocido Information Seeking Mantra [Shn96, HS12]. En concreto se permite realizar, *zoom*, *overview*, filtrado de elementos usando la leyenda, recuperación de valores y detalles mediante HoverTool, selección y resalte de elementos, y *linking* [YVMF06] o coordinación de todas las vistas. Adicionalmente cuenta con un historial de cambios para recuperar las últimas acciones realizadas, permite la exportación de gráficos y guía al usuario en el proceso de selección de atributos mediante la presentación de la influencia de estos en la visualización.

Con el objeto de optimizar el proceso de exploración, la interfaz es reactiva, permitiendo de esta manera al usuario ver rápidamente los cambios que tienen lugar al variar cualquiera de las distintas

Figura 2: *Interfaz visual propuesta. En la zona 1 se pueden ver las representaciones de los distintos métodos automáticos de reducción de dimensiones. En la zona 2, el diagrama de barras con distintas medidas sobre los atributos originales para guiar al usuario en la selección de atributos. En la zona 3, el menú de carga, configuración y selección de parámetros.*

opciones. Todo ello facilita el análisis y el ajuste de los distintos parámetros.

3.1. Visualización de métodos de reducción de dimensiones

Los algoritmos de reducción de dimensiones (tanto lineales como no lineales) tratan de optimizar algún objetivo. Por ejemplo, el método PCA busca obtener nuevas variables que capturan la máxima varianza posible, LDA busca aquellas direcciones que optimizan la separación de las clases y MDS trata de preservar de la mejor manera posible las distancias entre puntos.

Es por esto que la información que obtenemos de la estructura de los datos originales a través de cada uno de los métodos de reducción de dimensiones es diferente en cada caso. La herramienta permite al usuario observar varios de estos métodos simultáneamente, así como observar el comportamiento del predictor sobre cada uno de ellos, permitiendo al analista entender qué características de los datos es la que facilita más la separación de las clases.

A la hora de visualizar los métodos de reducción de dimensiones, la interfaz utiliza un gráfico de dispersión en el que las coordenadas de los puntos corresponden a las coordenadas de los datos reducidas por cada algoritmo de reducción de dimensiones utilizado a un espacio bidimensional. Por ejemplo, en la Fig. 3 se puede ver el gráfico de dispersión asociado al método de reducción de dimensiones MDS sobre el conjunto de datos *Breast Cancer Wisconsin (Diagnostic) Data Set* obtenido de UCI Machine Learning Repository [DKT17].

Nótese que, tanto en el caso del conjunto de datos de *Breast Cancer Wisconsin (Diagnostic)* como en el resto del documento, antes de aplicar cualquier método de reducción de dimensiones, nuestra herramienta estandariza los datos. Este proceso hace que cada atributo del conjunto de datos pase a tener media 0 y desviación típica 1, provocando que los algoritmos de reducción de dimensiones no den mayor importancia a variables con mayor rango de valores.

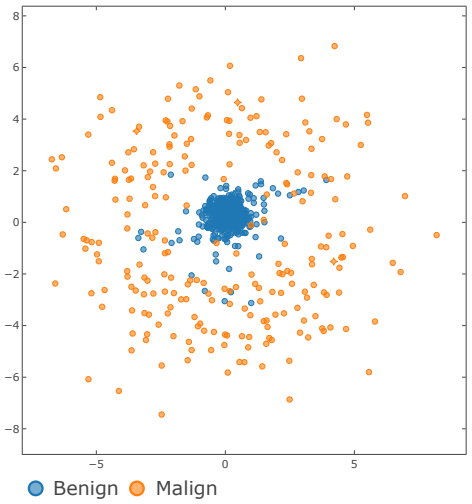

Figura 3: *Gráfico MDS para representar el Breast Cancer Wisconsin (Diagnostic) Data Set.*

Adicionalmente, se representa la influencia de las variables en el gráfico. En el caso de los métodos lineales, se utiliza con este fin SC. Esto se debe a que SC es capaz de representar cualquier proyección lineal sobre los datos en 2-3 dimensiones, permitiendo visualizar, sobre el gráfico de dispersión, el conjunto de ejes que representa el método de reducción de dimensionalidad lineal empleado. En base a esos ejes se puede observar la influencia de las variables en el gráfico. Así, en los últimos años, algunos trabajos de investigación han tratado de encontrar automáticamente las configuraciones de los vectores que representan los ejes en SC para poder realizar distintas tareas de análisis de datos [SY06, TC08, SYHX08, RSRDS16, SSRMJ*18].

Los distintos métodos lineales de reducción de dimensiones (PCA, LDA, ...) tienen una matriz asociada que transforma cada muestra original $\mathbf{x}$ en su proyección $\mathbf{p}$ del espacio de llegada (en nuestro caso bidimensional). Es decir,

$$\mathbf{p} = \mathbf{A}\,\mathbf{x}. \tag{2}$$

Se puede construir un modelo de SC que genere un gráfico en el que la representación de cada muestra original $\mathbf{x}$ venga dada por la proyección (2) del correspondiente método lineal seleccionado y, por tanto, como consecuencia de (1), $\mathbf{V} = \mathbf{A}^{\mathrm{T}}$. En la Fig. 4 podemos ver la configuración de los ejes de SC asociada al método de reducción de dimensiones PCA sobre el conjunto de datos *Iris Data Set* obtenido de UCI Machine Learning Repository [DKT17]. Este conjunto de datos tiene atributos de 3 especies de lirio: setosa (azul), versicolor (naranja) y virginica (verde). Los ejes de gráfico SC ($\mathbf{V}$) se han obtenido automáticamente como $\mathbf{V} = \mathbf{A}^{\mathrm{T}}$, donde $\mathbf{A}$ es la matriz que tiene en sus dos filas los dos autovectores principales que se obtienen del problema de autovectores que plantea el algoritmo PCA.

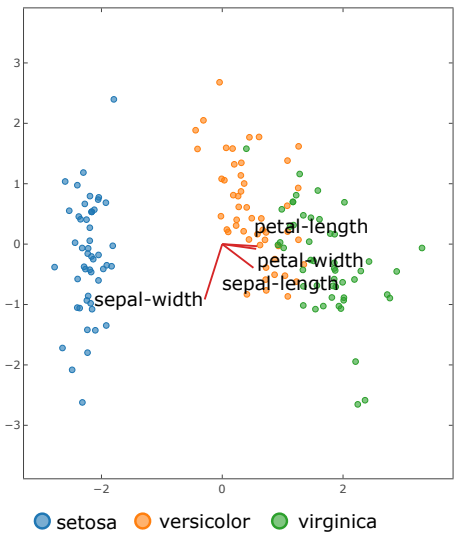

Figura 4: *Gráfico PCA para representar el Iris Data Set. Se puede observar el eje de SC para cada uno de los atributos de los datos.*

La representación visual en SC de estos algoritmos de reducción de dimensiones ha permitido a los investigadores analizar la influencia de los atributos originales en el gráfico realizado. En general las propuestas a tener en cuenta se basan en longitud [RSRDS16, WLN*17] y orientación [SSRMJ*18] de los ejes asociados. En base a dicha influencia, es posible ir descartando variables. Esta información puede servir de apoyo en los algoritmos no lineales de reducción de dimensiones que no poseen ejes con los que representar esta influencia. Para poder visualizar y trabajar de forma sencilla con la influencia de las variables se ha añadido a la interfaz el diagrama de barras interactivo descrito en la Sección 3.2.

Por último, la clase de cada uno de los datos se representa visualmente a través del color, permitiendo obtener una imagen global de la estructura de cada clase, así como la influencia de los atributos originales.

Por ejemplo, en la zona 1 de la Fig. 2 puede verse la disposición cuando se muestran cuatro gráficos correspondientes a los métodos LDA, MDS, t-SNE y LLE. Todos los gráficos representan el mismo conjunto de datos sobre características de vino, *Wine Data Set*, obtenido de UCI Machine Learning Repository [DKT17]. Si se interactúa con cualquiera de ellos mediante la selección de elementos, dicha operación afectará también al resto de gráficos para poder realizar así un análisis efectivo de los elementos. El *zooming* y *panning* son independientes para cada uno de los gráficos ya que las escalas de cada uno de los ejes son dispares, puesto que dependen del método de reducción de dimensiones. Para optimizar el espacio disponible se dispone de una única leyenda común a los cuatro gráficos. Dicha leyenda es interactiva, permitiendo al usuario indicar a través de ella si desea mostrar u ocultar los elementos de cada una de las clases disponibles. Además, al hacer *hover* sobre cualquiera de los puntos aparecerá un *tooltip* (véase Fig. 5) con información sobre el valor del subconjunto de atributos originales seleccionados de dicha observación.

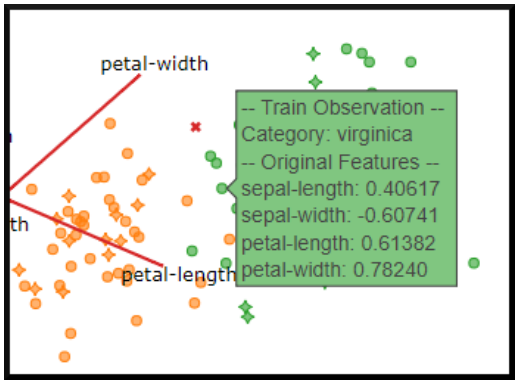

Figura 5: *Tooltip de un punto del Iris Data Set. En él se puede ver el valor de los atributos originales.*

3.2. Diagrama de barras para guiar la selección de variables

La interfaz aquí presentada utiliza un diagrama de barras para guiar al usuario en el proceso de selección de atributos. El diagrama de barras muestra, para cada variable, el valor de una medida que trata de capturar la influencia de ese atributo en el algoritmo de reducción de dimensiones seleccionado (ver Fig. 6). La medida

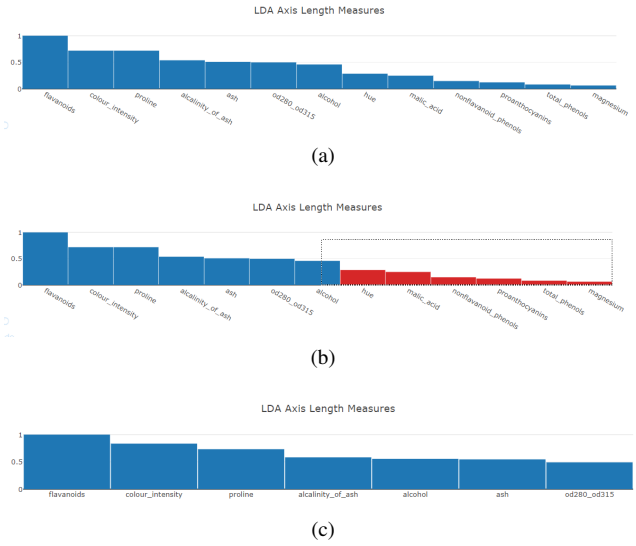

Figura 6: *Diagrama de barras para guiar la selección de las variables. En (b) el usuario ha seleccionado interactivamente las 6 variables con menor valor para su eliminación. En (c), el diagrama de barras una vez se han eliminado las 6 variables.*

mostrada en el diagrama corresponde a valores relativos a los ejes obtenidos en la representación mediante SC de los métodos lineales de reducción de dimensiones. Por ejemplo, [RSRDS16, WLN*17] plantean que los ejes más cortos de un gráfico SC son posibles candidatos a ser descartados por un método de selección de atributos interactivo. En este sentido, se puede representar como medida de influencia la longitud de cada eje de forma ordenada en el diagrama de barras. Así, por ejemplo, la representación mostrada en la Fig. 6 corresponde con la longitud de los ejes del método lineal LDA. Adicionalmente, se puede seleccionar qué método lineal (PCA, etc.) se desea utilizar como base para la obtención de dichos valores de influencia.

El usuario puede interactuar sobre el diagrama de barras para descartar aquellos atributos que considere poco significativos en función del orden del diagrama de barras. Dicha interacción se realiza directamente mediante mecanismos de selección (véase figura Fig. 6b) o *point-and-click*. Este proceso permite al usuario comprobar la influencia de los distintos atributos, decidir cuáles se quieren descartar en función de la influencia y su conocimiento de dominio, y descartarlos fácilmente sin necesidad de utilizar el menú de configuración descrito en la Sección 3.4.

Cuando el usuario elimina algunos de los atributos, se calculan los distintos métodos de reducción de dimensiones para el nuevo subconjunto de atributos y las nuevas medidas del diagrama de barras. El nuevo diagrama de barras solo muestra los atributos seleccionados (véase Fig. 6c). Del mismo modo, los gráficos de dispersión se actualizan para mostrar tanto la nueva proyección como los nuevos ejes de los atributos seleccionados.

Por último, este diagrama de barras puede servir también como guía para seleccionar atributos sobre las otras representaciones no lineales en las cuales no se conoce la influencia de los atributos originales al no poder representarse mediante ejes en SC. El usuario puede ir descartando atributos sobre el diagrama obtenido teniendo en cuenta la influencia sobre un método lineal y observar el efecto que tiene sobre las proyecciones no lineales. En caso de que el efecto no sea el esperado, pueden añadirse de nuevo atributos utilizando la zona inferior del menú de configuración descrito en la Sección 3.4.

3.3. Representación visual de clasificación

En el caso de nuestra interfaz, el preprocesamiento de los datos precede a un algoritmo de clasificación. Este preprocesamiento puede dar lugar a clasificadores más precisos [KZP06]. Además, el conocimiento sobre la importancia de los atributos y su contribución a la separación de clases permite tener una mejor idea de cómo funciona el clasificador resultante.

Para poder evaluar la predicción, suele ser habitual contar con dos conjuntos de datos distintos: entrenamiento y test. El conjunto de datos de entrenamiento se utiliza para ajustar el comportamiento del algoritmo intentando que aprenda a generalizar de qué depende la variable predictiva a partir de las muestras disponibles. El conjunto de test se utiliza para poder analizar cómo se comportaría el modelo ante datos nunca vistos por el algoritmo. Este último conjunto contiene una serie de muestras que no se usan hasta que se ha seleccionado el modelo final. Una de las métricas principales que se utilizan para medir la calidad del modelo es la precisión, ratio de muestras cuya clase se ha predicho correctamente sobre el total de muestras predichas.

Para comparar el comportamiento de varios modelos, parte de los datos del conjunto de datos de entrenamiento se reservan para la validación a través de la realización de pruebas intermedias. Existen otros métodos alternativos para la selección de modelos como bootstrapping [Sha96] o cross-validation [AC10], pero en nuestra herramienta se ha implementado la división de los datos en conjunto de entrenamiento y de validación por porcentaje, siendo el usuario el que deberá reservar parte de los datos u obtener nuevos datos para el conjunto de test. Los resultados sobre el conjunto de validación permiten al usuario seleccionar a través de la interfaz visual tanto el subconjunto de atributos como el método de reducción de dimensiones e hiperparámetros adecuados. Para poder analizar cómo se comportaría el modelo de predicción ante datos nunca vistos, debería utilizar el conjunto de test.

Los resultados obtenidos por el algoritmo de clasificación se pueden ver en la interfaz de dos formas complementarias (véase Fig. 7). Por un lado, junto a cada gráfico correspondiente a un modelo se muestra una tabla compuesta de tres valores que indican la precisión del gráfico en tres situaciones: i) la precisión obtenida con todos los atributos posibles del conjunto de datos; ii) la precisión obtenida con los atributos seleccionados en ese momento y iii) la precisión obtenida en la situación inmediatamente anterior. Este último valor de precisión se almacena y es mostrado con el fin de que los analistas puedan evaluar la última modificación realizada (por ejemplo, un descarte de nuevos atributos) y puedan deshacerla si lo consideran conveniente. Con el objetivo de permitir al analista conocer rápidamente si los cambios realizados mejoran (respectivamente, empeoran) la predicción actual se marca en verde

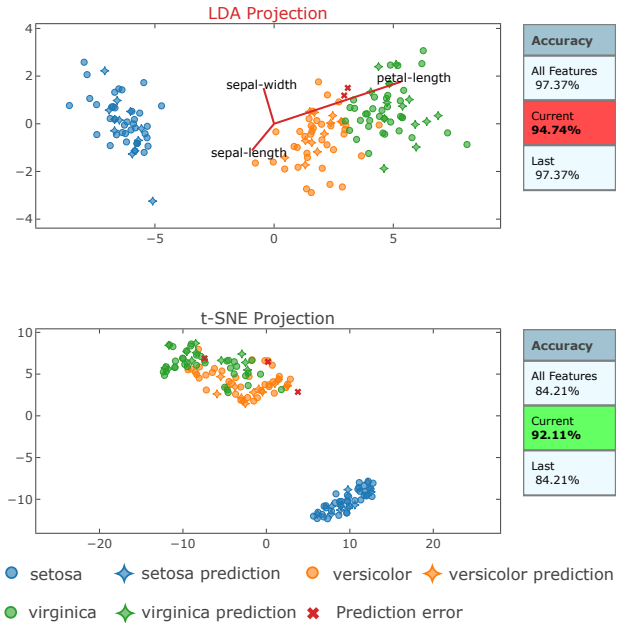

Figura 7: *Resultados de predicción para los métodos de reducción de dimensiones LDA (lineal, donde se observan los ejes de SC) y t-SNE (no lineal) sobre el Iris Data Set en el que se ha descartado el atributo 'petal-width'.*

(respectivamente, en rojo) el fondo de la casilla con el dato actual. Por otro, en los gráficos correspondientes a los distintos modelos, se diferencian visualmente los puntos correspondientes al conjunto de entrenamiento frente a los puntos de validación. Los elementos correspondientes al conjunto de datos de entrenamiento se representan con un círculo, las predicciones erróneas del conjunto de validación se representan mediante cruces de color rojo y aquellas predicciones correctas son representadas mediante diamantes. En este último caso, el color será el mismo que el de la clase correspondiente. La posibilidad de observar los datos de validación en el gráfico permite al usuario saber la estructura de los errores de predicción para entender mejor el modelo obtenido mediante el algoritmo de reducción de dimensiones. Además, el usuario tendrá acceso a la clase real y la predicha por el algoritmo para cada error de predicción a través del *tooltip* que aparece al realizar *hover* sobre el punto.

3.4. Configuración de la interfaz

Con el fin de seleccionar el conjunto de datos a visualizar, así como configurar la visualización, la interfaz dispone de un menú de configuración que puede verse en la Fig. 8.

Para cargar el conjunto de datos el usuario selecciona el fichero de texto o bien abriendo el explorador de ficheros haciendo clic en el botón "Upload" o bien haciendo uso de *drag-and-drop*. El usuario debe indicar el separador utilizado para separar los diferentes atributos (coma, tabulador, punto y coma, etc.)

Una vez cargado el conjunto de datos, el usuario debe seleccionar la columna correspondiente a la variable predictiva de entre el conjunto de variables disponibles en el desplegable "Target feature".

El único otro parámetro que el usuario debe seleccionar obligatoriamente para comenzar a visualizar y trabajar con los datos es el algoritmo de reducción de dimensiones. El usuario puede seleccionar aquellos con los que desee trabajar, pudiéndose visualizar simultáneamente. Entre los algoritmos de reducción de dimensiones disponibles se encuentran: LDA, PCA, t-SNE, MDS o LLE. A partir de ese momento, cualquier cambio realizado sobre el resto del menú afecta de forma reactiva a las visualizaciones.

Además de dichos parámetros, nuestra interfaz cuenta con un control deslizante para determinar el porcentaje del conjunto de datos que se destinará a entrenamiento, quedando el resto para validación. Adicionalmente, un desplegable permite seleccionar el método lineal de reducción de dimensiones utilizado para mostrar la influencia de los atributos en el diagrama de barras descrito en la Sección 3.2, junto a una serie de campos de entrada que permiten introducir los hiperparámetros tanto de los métodos de reducción de dimensiones (e.g. perplejidad del método t-SNE, o vecinos en el algoritmo LLE) como los parámetros del algoritmo de predicción utilizado.

En la zona inferior del menú hay un desplegable que permite ver

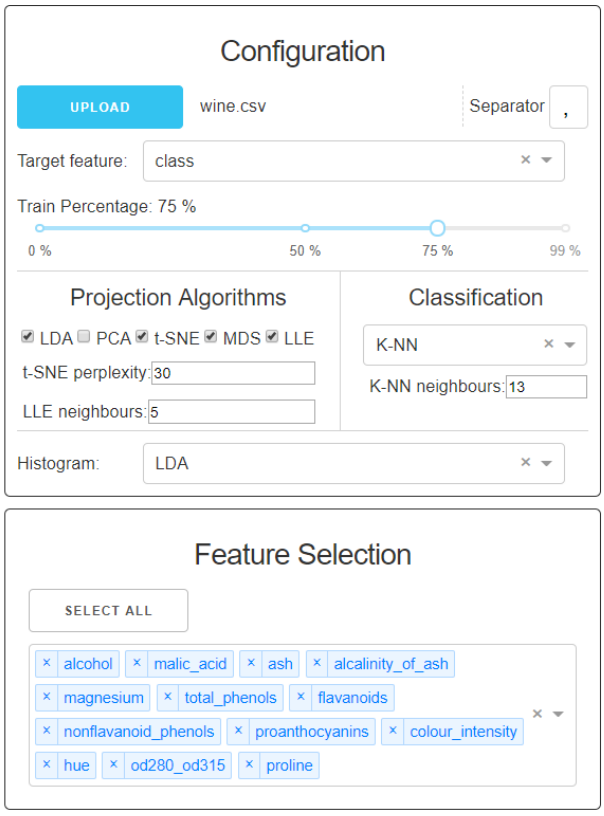

Figura 8: *Menú de configuración de la interfaz.*

los atributos seleccionados. Este desplegable se puede abrir para ver todos los atributos que no se encuentran seleccionados, filtrarlos por nombre y añadirlos (si se desea).

4. Resultados

En esta sección se describen los resultados obtenidos al utilizar la interfaz visual con algunos conjuntos de datos.

En un primer experimento se ha utilizado el *Wine Data Set*, obtenido de UCI Machine Learning Repository [DKT17]. Este conjunto de datos consta de 178 muestras de distintos vinos con 13 atributos numéricos y una variable predictiva que indica la variedad del vino (3 categorías posibles). Utilizando todos los atributos y destinando un 75 % de los datos a entrenamiento (15 % validación), la predicción es del 100 % tanto para LDA como para t-SNE (véase Fig. 2). Tras eliminar la mitad (6) de los atributos con menor longitud de eje LDA, se mantiene el 100 % de precisión en t-SNE y en LDA se disminuye a 97,68 %.

El tiempo necesario para recalcular ambos métodos (LDA y t-SNE) y las visualizaciones correspondientes para el nuevo subconjunto de 7 atributos del *Wine Data Set* es de 0,98 segundos lo que hace que la variación y prueba de múltiples configuraciones carezca de tiempos de espera largos para los usuarios. La mayoría de este tiempo es consumido por el cálculo de cada uno de los métodos de reducción de dimensiones. En la interfaz aquí presentada todos los métodos aplicados provienen de la librería scikit-learn.

En un segundo experimento se ha utilizado un dataset de mayor dimensionalidad, el *Weight Lifting Exercises Dataset* [VBG*13]. Este conjunto de datos consta de 4024 muestras de ejercicios de levantamiento de peso monitorizados con 53 atributos numéricos. La variable predictiva indica cómo se ha ejecutado el ejercicio, con 5 categorías posibles.

Utilizando todos los atributos y destinando un 80 % de los datos a entrenamiento y un 20 % a validación, los métodos de reducción de dimensiones con los que se obtiene mejor precisión son LDA (95,40 %) y t-SNE (89,69 %). Tras eliminar los 43 atributos con menor longitud de eje LDA, se obtienen precisiones de 94,16 % (LDA) y 94,04 %(t-SNE). Esto permite clasificar con casi la misma precisión utilizando sólo 10 de las 53 medidas originales. Estas precisiones se pueden mejorar mediante la variación de los distintos hiperparámetros.

En el caso del *Weight Lifting Exercises Dataset*, un conjunto con un mayor número de muestras y de dimensiones que *Wine Data Set*, el tiempo necesario para recalcular ambos métodos (LDA y t-SNE) y las visualizaciones correspondientes para el nuevo subconjunto de 10 atributos es de 23,95 segundos. Un tiempo prudencial que permite seguir utilizando la interfaz de forma interactiva, aunque con una usabilidad no tan buena.

En estos y otros experimentos realizados la interfaz ha cumplido su objetivo principal: guiar al analista en la selección de un conjunto de atributos significativo permitiendo además la selección del método de reducción de dimensiones lineal o no lineal más adecuado para dicha selección de atributos.

5. Conclusiones

La selección de un subconjunto de atributos es un problema de gran interés en el campo del aprendizaje automático. Disponer de un conjunto reducido de variables sin perder información permite acelerar los distintos algoritmos y obtener modelos más comprensibles para los analistas. Sin embargo, encontrar una selección adecuada de un menor número de atributos es un problema complejo de tipo NP-duro.

En este trabajo hemos propuesto una herramienta visual e interactiva para guiar la selección de un subconjunto de atributos utilizando métodos de reducción de dimensiones para representar el problema de clasificación a resolver. El uso de esta interfaz permite al usuario aportar su conocimiento sobre el conjunto de datos de interés iterando y configurando los distintos algoritmos disponibles. De esta forma puede observar la influencia de los distintos atributos en la representación y seleccionar un subconjunto de ellos.

Una de las principales ventajas del uso de la interfaz se encuentra en la sencillez de iterar múltiples veces en la selección de distintos subconjuntos de atributos. Adicionalmente la interfaz presentada cumple con la mayor parte de las tareas habituales del conocido Information Seeking Mantra.

Como trabajo futuro, se pretende mejorar la interfaz incluyendo la incorporación de nuevas medidas que guíen de forma más precisa al usuario en la selección de atributos, la inclusión de nuevos algoritmos de reducción de dimensiones o el uso de métodos adicionales de entrenamiento y validación. También se pretenden implementar versiones paralelas de los distintos algoritmos de reducción de dimensiones que optimicen los tiempos de ejecución y, por tanto, mejoren la interactividad y usabilidad de la interfaz.

Además, se pretende desarrollar una librería que permita a los usuarios no sólo usar de base la interfaz aquí propuesta sino extender la misma añadiendo nuevos métodos de reducción de dimensiones, nuevos clasificadores o nuevas medidas.

References

[AC10] ARLOT S., CELISSE A.: A survey of cross-validation procedures for model selection. *Statist. Surv. 4* (2010), 40–79. 6

[AES05] AMAR R., EAGAN J., STASKO J.: Low-level components of analytic activity in information visualization. In *Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization* (Washington, DC, USA, 2005), INFOVIS '05, IEEE Computer Society, pp. 15–. 2

[AS05] AMAR R. A., STASKO J. T.: Knowledge precepts for design and evaluation of information visualizations. *IEEE Transactions on Visualization and Computer Graphics 11*, 4 (July 2005), 432–442. 2

[Asi85] ASIMOV D.: The grand tour: a tool for viewing multidimensional data. *SIAM journal on scientific and statistical computing 6*, 1 (1985), 128–143. 3

[Bel61] BELLMAN R.: Adaptive control processes: a guided tour. 1

[BMSD13] B. MURDOCH T., S. DETSKY A.: The inevitable application of big data to health care. *JAMA : the journal of the American Medical Association 309* (04 2013), 1351–2. 1

[BNH14] BRADEL L., NORTH C., HOUSE L.: Multi-model semantic interaction for text analytics. In *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on* (2014), IEEE, pp. 163–172. 3

[But99] BUTLER D.: Computing 2010: From black holes to biology. *Nature* (1999), 67–70. 1

[Car08] CARPENDALE S.: Information visualization. Springer-Verlag, Berlin, Heidelberg, 2008, ch. Evaluating Information Visualizations, pp. 19–45. 2

[CC00] COX T. F., COX M.: *Multidimensional Scaling, Second Edition*, 2 ed. Chapman and Hall/CRC, 2000. 2

[CHW97] CHEN B., HONG J., WANG Y.: The minimum feature subset selection problem. *Journal of Computer Science and Technology 12*, 2 (Mar 1997), 145–153. 1

[CLL*13] CHOO J., LEE H., LIU Z., STASKO J., PARK H.: An interactive visual testbed system for dimension reduction and clustering of large-scale high-dimensional data. In *Visualization and Data Analysis 2013* (2013), vol. 8654, International Society for Optics and Photonics, p. 865402. 3

[CMS99] CARD S. K., MACKINLAY J. D., SHNEIDERMAN B. (Eds.): *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999. 2

[CS07] COOK D., SWAYNE D. F.: *Interactive and dynamic graphics for data analysis: with R and GGobi*. Springer Science & Business Media, 2007. 3

[CSH*16] CHEN R., SHI L., HAKENBERG J., NAUGHTON B., SKLAR P., ZHANG J., ZHOU H., TIAN L., PRAKASH O., LEMIRE M., SLEIMAN P., YI CHENG W., CHEN W., SHAH H., SHEN Y., FROMER M., OMBERG L., DEARDORFF M. A., ZACKAI E., BOBE J. R., LEVIN E., HUDSON T. J., GROOP L., WANG J., HAKONARSON H., WOJCICKI A., DIAZ G. A., EDELMANN L., SCHADT E. E., FRIEND S. H.: Analysis of 589,306 genomes identifies individuals resilient to severe mendelian childhood diseases. *Nature biotechnology 34* (2016), 531–538. 1

[DBB10] DIEHL S., BECK F., BURCH M.: Uncovering strengths and weaknesses of radial visualizations—an empirical approach. *IEEE Transactions on Visualization and Computer Graphics 16* (November 2010), 935–942. 2

[DGRG12] DANIELS K. M., GRINSTEIN G. G., RUSSELL A., GLIDDEN M.: Properties of normalized radial visualizations. *Information Visualization 11*, 4 (2012), 273–300. 2

[DHS12] DUDA R. O., HART P. E., STORK D. G.: *Pattern classification*. John Wiley & Sons, 2012. 1

[DKT17] DHEERU D., KARRA TANISKIDOU E.: UCI machine learning repository, 2017. 4, 5, 8

[DLR09] DRAPER G. M., LIVNAT Y., RIESENFELD R. F.: A survey of radial methods for information visualization. *IEEE Transactions on Visualization and Computer Graphics 15* (September 2009), 759–776. 2

[FHT01] FRIEDMAN J., HASTIE T., TIBSHIRANI R.: *The elements of statistical learning*, vol. 1. Springer series in statistics New York, 2001. 1

[GS05] GOMES C. P., SELMAN B.: Computational science: Can get satisfaction. *Nature 435*, 7043 (June 2005), 751–752. 1

[HBO10] HEER J., BOSTOCK M., OGIEVETSKY V.: A tour through the visualization zoo. *Commun. ACM 53*, 6 (June 2010), 59–67. 2

[HCL05] HEER J., CARD S. K., LANDAY J. A.: prefuse: a toolkit for interactive information visualization. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (New York, NY, USA, 2005), CHI'05, ACM, pp. 421–430. 2

[HGM*97] HOFFMAN P., GRINSTEIN G., MARX K., GROSSE I., STANLEY E.: DNA visual and analytic data mining. In *Proceedings of the 8th conference on Visualization '97* (Los Alamitos, CA, USA, 1997), VIS '97, IEEE Computer Society Press, pp. 437–441. 2

[HR03] HINTON G., ROWEIS S.: Stochastic neighbor embedding. *Advances in neural information processing systems 15* (2003), 833–840. 2

[HS12] HEER J., SHNEIDERMAN B.: Interactive dynamics for visual analysis. *Commun. ACM 55*, 4 (Apr. 2012), 45–54. 3

[HSM01] HAND D. J., SMYTH P., MANNILA H.: *Principles of Data Mining*. MIT Press, Cambridge, MA, USA, 2001. 2

[ID90] INSELBERG A., DIMSDALE B.: Parallel coordinates: a tool for visualizing multi-dimensional geometry. In *Proceedings of the 1st conference on Visualization* (Los Alamitos, CA, USA, 1990), VIS'90, IEEE Computer Society Press, pp. 361–378. 2

[Ins09] INSELBERG A.: *Parallel Coordinates: Visual Multidimensional Geometry and Its Applications*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2009. 2

[JZF*09] JEONG D. H., ZIEMKIEWICZ C., FISHER B., RIBARSKY W., CHANG R.: ipca: An interactive system for pca-based visual analytics. In *Computer Graphics Forum* (2009), vol. 28, Wiley Online Library, pp. 767–774. 2

[Kan00] KANDOGAN E.: Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions. In *Proceedings of the IEEE Information Visualization Symposium* (2000), vol. 650, Citeseer, p. 22. 2

[Kan01] KANDOGAN E.: Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (2001), ACM, pp. 107–116. 2

[Kei02] KEIM D. A.: Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics 8* (January 2002), 1–8. 2

[KZP06] KOTSIANTIS S. B., ZAHARAKIS I. D., PINTELAS P. E.: Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review 26*, 3 (Nov 2006), 159–190. 6

[LCW*17] LI J., CHENG K., WANG S., MORSTATTER F., TREVINO R. P., TANG J., LIU H.: Feature selection: A data perspective. *ACM Computing Surveys (CSUR) 50*, 6 (2017), 94. 1

[Lew92] LEWIS D. D.: Feature selection and feature extraction for text categorization. In *Proceedings of the workshop on Speech and Natural Language* (1992), Association for Computational Linguistics, pp. 212–217. 1

[LMW*15] LIU S., MALJOVEC D., WANG B., BREMER P.-T., PASCUCCI V.: Visualizing high-dimensional data: Advances in the past decade. In *Proc. Eurographics Conf. Visualization* (2015), pp. 20151115–127. 2

[LWBP14] LIU S., WANG B., BREMER P.-T., PASCUCCI V.: Distortion-guided structure-driven interactive exploration of high-dimensional data. In *Computer Graphics Forum* (2014), vol. 33, Wiley Online Library, pp. 101–110. 2, 3

[McL04] MCLACHLAN G. J.: *Discriminant analysis and statistical pattern recognition*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley-Interscience, 2004. 2

[MDL07] MAO Y., DILLON J., LEBANON G.: Sequential document visualization. *IEEE transactions on visualization and computer graphics 13*, 6 (2007). 2, 3

[ML14] MOLCHANOV V., LINSEN L.: Interactive Design of Multidimensional Data Projection Layout. In *EuroVis - Short Papers* (2014), Elmqvist N., Hlawitschka M., Kennedy J., (Eds.), The Eurographics Association. 2, 3

[MR02] MARKOVITCH S., ROSENSTEIN D.: Feature generation using general constructor functions. *Machine Learning 49*, 1 (2002), 59–98. 1

[NM13] NAM J. E., MUELLER K.: Tripadvisorˆ{ND}: A tourism-inspired high-dimensional space exploration framework with overview and detail. *IEEE transactions on visualization and computer graphics 19*, 2 (2013), 291–305. 2, 3

[PFG08] PLAISANT C., FEKETE J.-D., GRINSTEIN. G.: Promoting insight-based evaluation of visualizations: From contest to benchmark repositor. *IEEE Transactions on Visualization and Computer Graphics 14*, 1 (2008), 120–4134. 2

[RC94] RAO R., CARD S. K.: The table lens: merging graphical and symbolic representations in an interactive focus+context visualization for tabular information. In *Conference companion on Human factors in computing systems* (New York, NY, USA, 1994), CHI'94, ACM, pp. 318–322. 2

[RL15] RIECK B., LEITTE H.: Persistent homology for the evaluation of dimensionality reduction schemes. In *Computer Graphics Forum* (2015), vol. 34, Wiley Online Library, pp. 431–440. 2

[RS00] ROWEIS S. T., SAUL L. K.: Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science 290*, 5500 (2000), 2323–2326. 2

[RSRDS16] RUBIO-SÁNCHEZ M., RAYA L., DÍAZ F., SANCHEZ A.: A comparative study between radviz and star coordinates. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (January 2016), 619–628. 2, 5, 6

[RSSL17] RUBIO-SÁNCHEZ M., SANCHEZ A., LEHMANN D. J.: Adaptable radial axes plots for improved multivariate data visualization. *Computer Graphics Forum (Proc. EuroVis)* (2017). 2

[Sha96] SHAO J.: Bootstrap model selection. *Journal of the American Statistical Association 91*, 434 (1996), 655–665. 6

[Shn96] SHNEIDERMAN B.: The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages* (Washington, DC, USA, 1996), IEEE Computer Society, pp. 336–343. 3

[Spe07] SPENCE R.: *Information Visualization: Design for Interaction (2Nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2007. 2

[SR06] SIIRTOLA H., RÄIHÄ K.-J.: Interacting with parallel coordinates. *Interacting with Computers 18*, 6 (2006), 1278 – 1309. 2

[SSRMJ*18] SANCHEZ A., SOGUERO-RUIZ C., MORA-JIMENEZ I., RIVAS-FLORES F., LEHMANN D., RUBIO-SANCHEZ M.: Scaled radial axes for interactive visual feature selection: A case study for analyzing chronic conditions. *ExpertSystemsWithApplications* (2018). 2, 5

[SY06] SHAIK J. S., YEASIN M.: Visualization of high dimensional data using an automated 3d star coordinate system. In *International joint conference on neural networks* (2006), pp. 1339–1346. 5

[SYHX08] SUN Y., YUAN J., HU Y., XIAO W.: An improved multivariate data visualization technique. In *International Conference on Information and Automation, ICIA'08.* (june 2008), pp. 1525–1530. 5

[SZS*17] SACHA D., ZHANG L., SEDLMAIR M., LEE J. A., PELTONEN J., WEISKOPF D., NORTH S. C., KEIM D. A.: Visual interaction with dimensionality reduction: A structured literature analysis. *IEEE transactions on visualization and computer graphics 23*, 1 (2017), 241–250. 2

[TC08] TSAI C.-Y., CHIU C.-C.: A clustering-oriented star coordinate translation method for reliable clustering parameterization. In *Proceedings of the 12th Pacific-Asia conference on Advances in knowledge discovery and data mining* (Berlin, Heidelberg, 2008), PAKDD'08, Springer-Verlag, pp. 749–758. 5

[Tuf86] TUFTE E. R.: *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, USA, 1986. 2

[VBG*13] VELLOSO E., BULLING A., GELLERSEN H., UGULINO W., FUKS H.: Qualitative activity recognition of weight lifting exercises. In *Proceedings of the 4th Augmented Human International Conference* (2013), ACM, pp. 116–123. 8

[vdMH08] VAN DER MAATEN L., HINTON G.: Visualizing data using t-SNE. *Journal of Machine Learning Research 9* (2008), 2579–2605. 2

[War04] WARE C.: *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2004. 2

[WLN*17] WANG Y., LI J., NIE F., THEISEL H., GONG M., LEHMANN D. J.: Linear discriminative star coordinates for exploring class and cluster separation of high dimensional data. *Computer Graphics Forum (Proc. EuroVis)* (2017). 5, 6

[YGX*09] YUAN X., GUO P., XIAO H., ZHOU H., QU H.: Scattering points in parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics 15* (November 2009), 1001–1008. 2

[YKSJ07] YI J. S., KANG Y. A., STASKO J., JACKO J.: Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (Nov. 2007), 1224–1231. 2

[YVMF06] YOUNG F. W., VALERO-MORA P. M., FRIENDLY M.: *Visual Statistics: Seeing Data with Dynamic Interactive Graphics*. Wiley-Interscience, 2006. 3