Diseño e implementación de una aplicación en *Processing* para la representación visual de datos multidimensionales utilizando técnicas de Minería de Datos

VNIVERSITAT DE VALÈNCIA

Escola Tècnica Superior d'Enginyeria

Departament d´Enginyeria Electrònica

Tesis Doctoral

Septiembre, 2015

Autor:

Miguel Ángel Sáez Ibáñez

Directores:

Emilio Soria Olivas

José David Martín Guerrero

VNIVERSITAT DE VALÈNCIA

Tesis Doctoral

Diseño e implementación de una aplicación en *Processing* para la representación visual de datos multidimensionales utilizando técnicas de Minería de Datos

Autor:

Miguel Ángel Sáez Ibáñez

Directores:

Emilio Soria Olivas

José David Martín Guerrero

Valencia, 2015

Departament d'Enginyeria Electrònica
Universitat de València
Av. Universitat – 46100 Burjassot (Spain)
Phone: +34-9635 44032
FAX: +34-9635 44353
e-mail: dpto.ingenieria.electronica@uv.es
http://www.uv.es/die

D. EMILIO SORIA OLIVAS, Profesor Titular del Departamento de Ingeniería Electrónica de la Escola Técnica Superior d'Enginyeria de la Universitat de València, y

D. JOSÉ DAVID MARTÍN GUERRERO, Profesor Titular del Departamento de Ingeniería Electrónica de la Escola Técnica Superior d'Enginyeria de la Universitat de València

HACEN CONSTAR QUE:

El presente trabajo "Diseño e implementación de una aplicación en *Processing* para la representación visual de datos multidimensionales utilizando técnicas de Minería de Datos" ha sido realizado bajo su dirección en el Departamento de Ingeniería Electrónica de la Universitat de València, por el Licenciado en Ciencias Físicas Miguel Ángel Sáez Ibáñez y que constituye su Tesis para optar al grado de Doctor en Ingeniería Electrónica.

Y para que así conste, en cumplimiento de la normativa vigente, firmamos el presente certificado en Valencia, a _____

_____	_____	_____
Emilio Soria Olivas	José D. Martín Guerrero	Dtor. del Departamento de Ingeniería Electrónica

Tesis Doctoral:	Diseño e implementación de una aplicación en *Processing* para la representación visual de datos multidimensionales utilizando técnicas de Minería de Datos
Autor:	Miguel Ángel Sáez Ibáñez
Directores:	Dr. Emilio Soria Olivas Dr. José David Martín Guerrero

El tribunal nombrado para juzgar la Tesis Doctoral arriba citada, compuesto por los doctores:

Presidente: _____

Vocal: _____

Secretario: _____

Acuerda otorgarle la calificación de _____

Y para que así conste a los efectos oportunos, firmamos el presente certificado.

Valencia, a

Agradecimientos

En primer lugar quiero expresar mi más profundo agradecimiento a mis directores de tesis, Emilio Soria y José David Martín por confiar en mí, por su comprensión y por su continua preocupación en acomodar las exigencias de tan dura labor a una situación personal tan complicada. De ningún otro modo podría haber llevado a cabo una labor de esta magnitud; eso ya lo sabéis. Gracias por vuestras fantásticas ideas, por vuestro esfuerzo y por vuestro aliento que lo han sido todo en este trabajo, pero sobre todo gracias por vuestra consideración.

También quiero expresar un especial agradecimiento a los profesores Enrique Sanchis, Vicente González y Julio Martos su constante preocupación y apoyo en todo momento.

En general, quiero también agradecer a todo el personal del Departamento de Ingeniería Electrónica de la Universidad de Valencia el magnífico clima de trabajo y cooperación creado.

En especial a mi familia. A mis padres por los valores que me han transmitido y todo el esfuerzo y sacrificio dedicado en mi educación y que me han hecho ser quien soy; no sabéis cuánto siento que no hayáis podido disfrutar de este momento. A mi mujer María José por su apoyo y comprensión, por su complicidad, por haber aguantado el insufrible sobreesfuerzo de estos años, por quererme y hacerme sentir querido. A mis hijos Sofía, Andrés y Silvia por ser mi vida y por haber soportado las ausencias con la alegría que los caracteriza. Os prometo recuperar todo este tiempo.

Índice general

Índice de Figuras

Índice de Tablas

Acrónimos y abreviaturas

ACP	Análisis de Componentes Principales
AF	Análisis Factorial
API	*Application Programming Interface*
BCSM	*Between Cluster Scatter Matrix*
CLIQUE	*CLustering In QUEst*
DBCLASD	*Distribution-Based clustering Algorithm for Mining Large Spatial Databases*
DBSCAN	*Density-Based Algorithms for Discovering Clusters in Large Spatial Databases with Noise*
DM	*Data Mining*
DVBSCAN	*Density Based Algorithm for discovering Density Varied Clusters in Large Spatial Databases*
ECV	Error de Conservación de Vecindarios
EM	*Expectation-Maximization*
EM ACP	Análisis de Componentes Principales Probabilístico con algoritmo *Expectation-Maximization*
FCM	*Fuzzy C-means*
GIS	*Geographical Information System*
IDE	*Integrated Development Environment*
Isomap	*Isometric feature Mapping*

KDD	*Knowledge Discovery in Database*
KPCA	*Kernel Principal Component Analysis*
LDA	*Linear Discriminant analysis*
LLE	*Local Linear Embedding*
MDS	*Multidimensional Scaling*
MIT	*Massachusetts Institute of Technology*
OCDE	Organización para la Cooperación y el Desarrollo Económicos
PVC	Promedio de Vecinos Conservados
RD	Reducción de la dimensionalidad
Silh	*Silhouette*
SNE	*Stochastic Neighbor Embedding*
SS	*Sum of Squares*
SSNE	*Symmetric Stochastic Neighbor Embedding*
ST-DBSCAN	*Spatial-Temporal Density Based Clustering*
STING	*STatistical INformation Grid*
SVM	*Support Vector Machines*
t-SNE	*t-Distributed Stochastic Neighbor Embedding*
VDBSCAN	*Varied Density Based Spatial Clustering of Applications with Noise*
VDE	*Visual Data Enviroment*
VDM	*Visual Data Mining*
WCSM	*Within Cluster Scatter Matrix*

Resumen y objetivos

La posibilidad de disponer de representaciones gráficas de los datos es de gran valor a la hora de extraer conocimiento útil. Sus principales ventajas son la visualización de información de una forma sencilla, rápida y directa. No obstante, muchos de los conjuntos de datos contienen numerosos registros, que pueden ser de naturaleza multivariante. En estos casos, la representación visual de datos se convierte en una tarea complicada y las técnicas clásicas que suelen utilizarse obtienen resultados poco intuitivos.

Esta tesis se plantea como objetivo el diseño e implementación de una aplicación versátil capaz de representar visualmente gran cantidad de datos multidimensionales de forma eficaz para su fácil comprensión. Mediante el uso de esta aplicación se pretende que el usuario encuentre un entorno interactivo de uso sencillo con el que poder visualizar su conjunto de datos.

Para conseguir el objetivo de visualizar conjuntos con gran cantidad de datos se propone la utilización de métodos de agrupamiento en la aplicación. Estos procedimientos permiten que el conjunto de datos pueda distribuirse en grupos con características similares y ser representados visualmente de forma unitaria aunque conservando toda la información de los registros individuales que los componen. De este modo se pretende obtener representaciones gráficas de una forma más simple y apta para su inspección visual.

Con el fin de representar visualmente datos multidimensionales se dota a la aplicación de diferentes técnicas de reducción de la dimensionalidad. Mediante estas técnicas, se puede lograr la transformación de los datos de alta dimensión en una representación de menor dimensión que sea significativa y que respete su estructura original. En el caso de la aplicación desarrollada, se utilizan estas técnicas para reducir la dimensión original de los datos a solamente dos para proceder a su representación en el plano.

De igual modo, con esta aplicación no solamente se pretende que pueda representar visualmente conjuntos de datos multidimensionales de manera intuitiva y lógica sino que también ofrezca grandes posibilidades de interacción natural con el usuario. Para ello la aplicación ha de contar con herramientas

mediante las cuales se pueda analizar la representación visual desde distintos puntos de vista y a diferentes niveles de detalle según la voluntad del usuario contando con servicios adecuados para poder manejar y almacenar la información obtenida.

Atendiendo a este objetivo, se estima el entorno de programación *Processing* (*www.processing.org*) como el más adecuado para implementar la aplicación por su sencillez de manejo, velocidad de ejecución y su orientación hacia el desarrollo de aplicaciones visuales. A estas características se le unen la capacidad de producir aplicaciones multiplataforma y un tratamiento interactivo eficaz.

La aplicación, una vez implementada, se evalúa y se utiliza para la extracción de conclusiones del problema analizado. En este sentido, se analizan los resultados obtenidos en diferentes contextos reales. Finalmente, se pone la aplicación a disposición de otros investigadores de manera gratuita.

Resum i objectius

La possibilitat de disposar de representacions gràfiques de les dades és molt valuosa a l'hora d'extraure coneixement útil. Els seus principals avantatges són la visualització d'informació d'una manera senzilla, ràpida i directa. Tanmateix, molts dels conjunts de dades contenen nombrosos registres, que hi poden ser de tipus multivariant. En estos casos, la representació visual esdevé una tasca complicada i les tècniques clàssiques que solen emprar-se obtenen resultats pocs intuïtius.

Esta tesi es planteja com a objectiu el disseny i implementació d'una aplicació versàtil capaç de representar visualment gran quantitat de dades multidimensionals d'una manera eficaç per a la seua fàcil comprensió. Mitjançant la utilització d'esta aplicació, es pretén que l'usuari trobe un entorn interactiu d'ús senzill amb que poder visualitzar conjunts de dades.

Per tal d'assolir l'objectiu de visualitzar conjunts amb gran quantitat de dades, es proposa la utilització de mètodes d'agrupament a l'aplicació. Estos procediments permeten que el conjunt de dades puga distribuir-se en grups amb característiques semblants i ser representats visualment de manera unitària encara que conservant tota la informació dels registres individuals que els composen. D'esta manera, s'obtenen representacions gràfiques d'una manera senzilla i apta per a la seua inspecció visual.

Amb l'objectiu de representar visualment dades multidimensionals, l'aplicació inclou diferents tècniques de reducció de la dimensionalitat. Mitjançant estes tècniques, es pot aconseguir la transformació de les dades d'alta dimensió en una representació de menor dimensió que siga, no obstant això, significativa, tot respectant l'estructura original. L'aplicació desenvolupada només considera la reducció de la dimensió original a estructures bidimensionals per a procedir a la seua representació en un pla.

A més, esta aplicació ha de, no només representar visualment conjunts de dades multidimensionals de manera intuïtiva i lògica, sinó que també ha d'oferir grans possibilitats d'interacció natural amb l'usuari. Per tant, l'aplicació ha de comptar amb eines que permeten analitzar la representació visual des de diferents

punts de vista i a diferents nivells de detall d'acord amb la voluntat de l'usuari, tot comptant amb servicis adequats per a poder gestionar i emmagatzemar la informació obtinguda.

Per tal d'acomplir este últim objectiu, s'ha triat l'entorn de programació *Processing* (*www.processing.org*) com al més adequat per a implementar la aplicació ja que oferix una gestió senzilla, alta velocitat d'execució i està orientat cap al desenvolupament d'aplicacions visuals. A estes característiques se li afegixen la capacitat de produir aplicacions multiplataforma i un tractament interactiu eficaç.

La aplicació, una volta implementada, s'avalua i s'utilitza per a l'extracció de conclusions del problema analitzat. En este sentit, s'analitzen els resultats obtinguts en diferents contexts reals. Finalment, l'aplicació es posa a disposició d'altres investigadors de manera gratuïta.

Summary and goals

The availability of graphical data representations is extremely valuable in order to extract useful knowledge. Its main advantages are related to information visualization in a simple way, fast and straightforward to interpret. Nonetheless, many data sets include numerous records that may be multivariant. In those cases, visual representation becomes an arduous task and classical techniques provide results that usually lack of intuitiveness.

This thesis pursues the goal of designing and implementing a versatile application, able to visually represent large amounts of multidimensional data in a simple way that make data understandable. The use of this application tries to provide a framework in which users can visualize data sets in an interactive and simple way.

Clustering algorithms are proposed to achieve the goal of visualizing large amounts of data. This is because those methods allow the distribution of the data set in different clusters of similar characteristics that can be, in turn, visually represented in a unitary way although keeping all the information from the individual records that form the clusters. This way, graphical representations are obtained easily and can be inspected visually.

In order to carry out a visual representation of multidimensional data, the application includes a number of techniques for dimensionality reduction. Therefore, it is possible to transform the original data into a lower dimensionality structure that must be significant and loyal to the original structure. The application only considers the reduction of the original dimension into bi-dimensional structures so that representation is done in a plane.

Moreover, this application should not only provide a visual representation of multidimensional data in an intuitive and logical way, but also offer a wide and natural user interaction. Therefore, the application must contain tools to carry out visual representations from different points of view as well as different levels of detail that can be selected by the user. Finally, information management and storing must also be guaranteed.

This latter goal suggests the use of *Processing* (*www.processing.org*) as programming framework due to their characteristics, namely, simple management, high execution speed, and its focus on the development of visual applications. Alongside those characteristics, it is also remarkable its capability of producing multi-platform applications and an efficient interactive treatment.

Once the application has been implemented, it is evaluated and used for drawing conclusions from different problems. In particular, results are analyzed in the framework of several actual problems. Finally, the application is offered freely to other researchers.

Capítulo 1

Introducción

Resumen

En este capítulo se presenta la motivación para la elaboración de esta tesis y aspectos generales sobre la Extracción de Conocimiento en Bases de Datos y Minería de Datos. Seguidamente se trata aspectos generales sobre la Representación Visual de Datos y el entorno de programación Processing© , donde se centra el desarrollo de esta tesis.

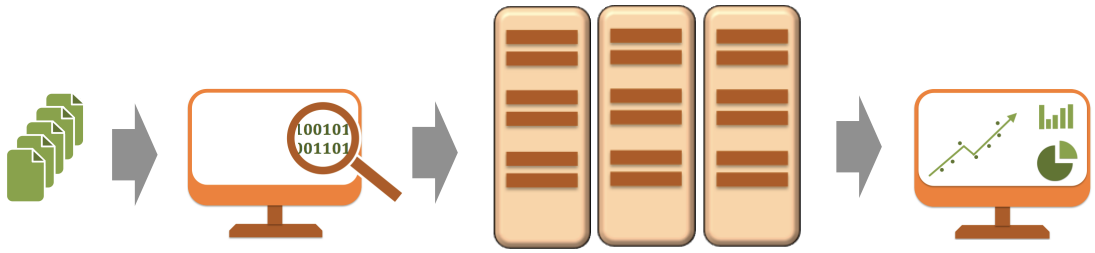
Finalmente, se exponen los objetivos y la estructura del presente trabajo.

1.1. Motivación

Nunca antes se ha generado una cantidad de datos tan grande como se genera en nuestros días. Según (Gantz & Reinsel, 2012) desde el año 2012 al 2020, el volumen de datos digital será aproximadamente el doble cada dos años. De la misma manera, las herramientas para su uso en los diversos campos del conocimiento (adquisición, gestión del almacenamiento, seguridad, recuperación, mantenimiento, etc.) deben desarrollarse para hacer frente a este crecimiento. La exploración y el análisis de grandes volúmenes de datos se ha convertido en una labor cada vez más difícil y encontrar información valiosa oculta en los datos que son almacenados digitalmente es una tarea complicada. Si no se tiene posibilidad

para explorar adecuadamente grandes cantidades de datos, éstos se recopilan debido a su potencial utilidad convirtiéndose en algo inútil. Las corporaciones, entidades gubernamentales y científicos, por nombrar sólo algunos, se están dando cuenta de los desafíos y, por otra parte, las oportunidades que existen en la utilización efectiva de esos datos. Sin embargo, para desbloquear el potencial contenido se requiere la aplicación de técnicas para explorar y transmitir las ideas clave (Kirk, 2012) (ver figura 1-1).

Existe una necesidad urgente de disponer de herramientas y teorías que ayuden en la selección de información útil (conocimiento) a partir de los crecientes volúmenes de datos digitales. Estas teorías y herramientas son el tema del campo de Extracción de Conocimiento en Bases de Datos o *KDD* (del inglés *Knowledge Discovery in Database*) (Usama Fayyad, 1996; Kurgan & Musilek, 2006).



(a) Datos. *(b) Digitalización.* *(c) Almacenamiento* *(d) Conocimiento*

Figura 1-1: *Grandes volúmenes de datos producidos hoy en día (a) son digitalizados (b) para poder recopilarse debido a su potencial valor informativo. Si no se dispone de herramientas efectivas para la extracción de información, estos datos quedan almacenados convirtiéndose en algo inservible (c). Solamente a través de técnicas especializadas en la extracción y selección de información útil se puede desbloquear ese potencial dando acceso al conocimiento (d).*

La extracción del conocimiento oculto a partir de datos es posible a través de la Minería de Datos o *DM* (del inglés *Data Mining*). *KDD* se refiere al proceso global de descubrir conocimiento útil a partir de datos, mientras que la Minería de Datos se refiere a una determinada fase en este proceso. Minería de Datos es la extracción automatizada o conveniente de los patrones que representan el conocimiento implícitamente almacenado o que se halla capturado en grandes bases de datos, almacenes de datos, la web, otros repositorios de información masivos o flujos de datos (Han & Kamber, 2006).

Para que la extracción de la información en Minería de Datos sea eficaz, es importante incluir el ser humano en el proceso de exploración de datos de tal modo que se combine la flexibilidad, la creatividad y el conocimiento general que éste posee con la enorme capacidad de almacenamiento y la potencia de cálculo de los ordenadores de hoy en día. Un enfoque orientado hacia la exploración visual de

los datos permitiría al usuario entender más fácilmente el proceso y reconocer los eventos importantes. Esta estrategia de Minería de Datos basada en la exploración visual de los datos se denomina Minería de Datos Visual o *VDM* (del inglés *Visual Data Mining*) (Keim, 2002). *VDM* tiene como objetivo facilitar el proceso de exploración de datos mediante la presentación de la información de una forma mucho más intuitiva que la simple exposición de los datos en estado crudo facilitando la extracción de conocimiento. La idea básica es la de presentar los datos de una forma visual, de tal modo que el usuario pueda obtener una perspectiva adecuada, sacar conclusiones e interactuar directamente con ellos. A través de la visualización, se pueden representar los datos de manera que permiten al analista observar sus datos con una nueva perspectiva, advertir patrones, tendencias, excepciones y contemplar las posibles historias que permanecen ocultas. Por ello, puede considerarse la visualización como una herramienta para la extracción de conocimiento (Kirk, 2012).

Una vez centrados en lo complicado de obtener información relevante a partir de grandes cantidades de datos y los indudables beneficios que aporta la visualización de los mismos, nos encontramos con tres problemas:

- El primer problema se refiere a la representación visual de grandes cantidades de datos. Cuando el número de datos a representar es muy grande los mismos datos representados impiden percibir la situación de forma adecuada y es complicado obtener información relevante. En estos casos las técnicas de agrupamiento o conglomerados (en inglés, *clustering*) son adecuadas. El análisis de conglomerados consiste en la organización de un conjunto de datos en grupos basándose en su similitud (Xu & Wunsch, 2009). De este modo, si varios datos con características similares son representados como uno solo la labor de exploración visual y extracción de conocimiento se ven beneficiadas.
- El segundo problema que se presenta es la alta dimensionalidad. Los datos del mundo real, tales como señales de voz, fotografías digitales, … suelen tener una alta dimensionalidad, es decir, se componen de una alta cantidad de características o atributos. Con el fin de manejar adecuadamente tales datos del mundo real y para proceder a su representación visual, su dimensionalidad necesita ser reducida. La reducción de dimensionalidad consiste en la transformación de los datos de alta dimensión en una representación significativa con una dimensión menor. Como resultado, la reducción de dimensionalidad facilita la clasificación, la visualización, y la compresión de datos de alta dimensión (van der Maaten , Postma, & van den Herik , 2009).

- El tercer problema es encontrar un entorno adecuado para la representación visual de los datos. Este entorno ha de ser lo suficientemente flexible como para permitir explorar e interactuar con los datos directamente. Asimismo, ha de permitir tanto facilitar información relevante del conjunto de datos como permitir al usuario examinar características particulares y datos de forma individual. Este aspecto será examinado en el punto 1.5. donde se aborda el entorno de programación *Processing* (*www.processing.org*) para el desarrollo de aplicaciones con esas condiciones.

1.2. Extracción de Conocimiento en Bases de Datos

La Extracción de Conocimiento en Bases de Datos es el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos (Usama Fayyad, 1996).

Las propiedades deseables del conocimiento extraído pueden resumirse en:

- **Válido:** Hace referencia a que los patrones deben seguir siendo precisos para datos nuevos, con cierto margen de certidumbre, y no sólo para aquellos que han sido usados en su obtención.
- **Novedoso:** Que aporte algo inédito tanto para el sistema como para el usuario.
- **Potencialmente útil:** La información debe conducir a acciones que reporten algún beneficio para el usuario.
- **Comprensible:** Que facilite la interpretación, revisión, validación y uso en la toma de decisiones.

Es importante tener en cuenta que la Minería de Datos no abarca todo el proceso de *KDD*, sino que es una parte del mismo. La Minería de Datos comprende todo un conjunto de técnicas encaminadas a la extracción de conocimiento implícito en las bases de datos. Sin embargo, aunque el procesado de datos previo a la aplicación de dichas técnicas así como su procesado posterior son una parte importante de la Minería de Datos, ésta no llega a abarcar todo aquello relacionado

con lo que es la preparación de los datos, así como la evaluación y la interpretación de resultados.

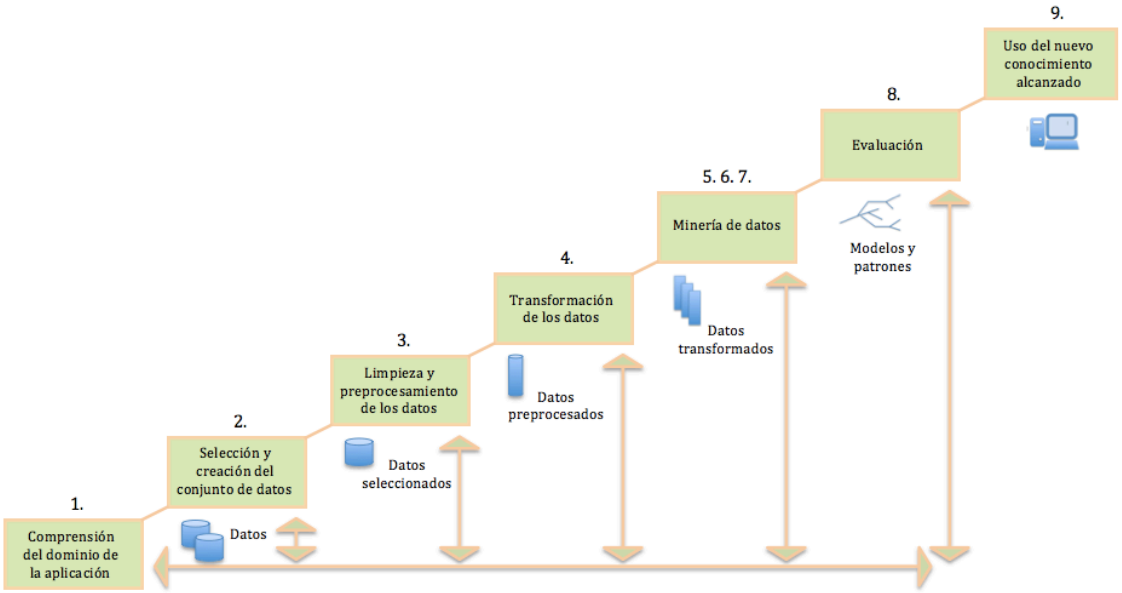


Figura 1-2: *Proceso de Extracción de Conocimiento en Bases de Datos o KDD (del inglés Knowledge Discovery in Database).*

Existen diferentes formas de ordenar el proceso de *KDD* todas ellas similares. Una forma de organizar el proceso *KDD* aparece en la figura 1-2 y las diferentes fases que ahí aparecen son las siguientes (Sumathi & Sivanandam, 2006; Oded & Rokach, 2010):

1. **Comprensión del dominio de la aplicación**: En esta fase se necesita entender y definir los objetivos del usuario final. A medida que el proceso *KDD* avanza, se pueden realizar revisiones de este paso.
2. **Selección y creación del conjunto de datos:** En esta fase se indaga acerca de los datos disponibles sobre los cuales se quiere obtener información, se obtienen datos adicionales si fuera necesario y finalmente se procede a la integración de todos ellos incluyendo los atributos que se considerarán para el proceso. Este paso es fundamental ya que si faltan algunos datos importantes, el resultado del estudio entero puede fallar.
3. **Limpieza y preprocesamiento de los datos:** Esta fase se encarga del manejo de carencia de datos, datos anómalos, datos incompletos y

eliminación de ruido. Este paso puede involucrar métodos estadísticos y empleo de algoritmos complejos en función del problema (selección de características, modelos de predicción, …). Parte de estas labores son competencia de la Minería de datos.

4. **Transformación de los datos:** Este paso es crucial y suele ser específico para cada proyecto *KDD*. Aquí se preparan los datos para ser tratados por las técnicas para la extracción de conocimiento en las siguientes fases, incluyendo tareas de normalización y codificación de los datos para que sean procesados de manera adecuada posteriormente.
5. **Elección de la tarea de análisis apropiada:** Esta tarea es totalmente competencia de la Minería de Datos. En este apartado se decide qué tipo de tarea utilizar en función de los objetivos de *KDD*.
6. **Elección de algoritmos:** En esta etapa se selecciona el método específico que se utilizará para lograr el fin marcado.
7. **Empleo del algoritmo:** Esta fase puede considerarse la última de la Minería de Datos y es donde se utiliza el algoritmo escogido. Este proceso puede requerir emplearlo varias veces ajustando los parámetros del mismo de forma adecuada hasta obtener resultados satisfactorios.
8. **Evaluación:** Este paso se centra en comprender el resultado obtenido y estimar su utilidad. La interpretación de la información adquirida puede hacer reconsiderar los pasos de preprocesamiento con respecto a sus efectos en los resultados del algoritmo de Minería de Datos. Esta fase también incluye la documentación del conocimiento logrado para su uso posterior.
9. **Uso del nuevo conocimiento alcanzado:** Este conocimiento puede ser utilizado para entrenar el modelo con otros parámetros, extraer o incluir patrones en el conjunto de datos, o incluso formar otro modelo general para mejorarlo. En este sentido, *KDD* es un proceso tanto interactivo como iterativo.

Como puede observarse en la figura 1-2 este proceso no siempre discurre de forma secuencial. Si en alguna fase el resultado no es satisfactorio puede regresarse a cualquiera de las etapas anteriores y replantearse. Este esquema puede repetirse tan a menudo como se considere necesario hasta obtener un modelo satisfactorio. Una vez comprobado el modelo, si es válido para los propósitos planteados (proporcionando salidas apropiadas y con márgenes de error aceptables) está listo para su uso.

1.3. Minería de Datos

El campo de la Minería de Datos ha sido objeto de intenso estudio en las dos últimas décadas. La primera conferencia internacional de *KDD* y Minería de Datos, se llevó a cabo en 1995 y hay una variedad de definiciones para este concepto. Entre ellas se plantean las siguientes (Shmueli, Patel, & Bruce, 2005):

- Extracción de información útil desde grandes conjuntos de datos.
- Proceso de exploración y análisis, por medios automáticos o semiautomáticos, de grandes cantidades de datos con el fin de descubrir patrones significativos y reglas.
- Proceso de descubrir nuevas correlaciones significativas, patrones y tendencias mediante depuración de grandes cantidades de datos almacenados en repositorios, utilizando tecnologías de reconocimiento de patrones, así como técnicas estadísticas y matemáticas.

Las tareas y técnicas que son competencia de la Minería de Datos pueden describirse como una amalgama de enfoques del Aprendizaje Automático (en inglés, *Machine Learning*) (Alpaydin, 2010) y la Estadística (Coenen, 2011). Desde esta perspectiva la Minería de Datos se puede decir que se ha desarrollado a partir de ambas. De hecho, la comunidad de la Minería de Datos está dominada por una mezcla de informáticos y estadísticos. Hay, sin embargo, una distinción entre la Minería de Datos y el Aprendizaje Automático. La Minería de Datos se centra en los datos (en todos sus formatos) y, como tal, puede ser vista como una aplicación; mientras que el Aprendizaje Automático, al menos en su forma tradicional, se centra en los mecanismos mediante los cuales las computadoras pueden aprender. De este modo, el Aprendizaje Automático puede interpretarse como una tecnología, mientras que la Minería de Datos, y por extensión *KDD*, como una aplicación.

Dentro de la Minería de Datos podemos distinguir dos enfoques principales (Fürnkranz, Gamberger, & Lavrač, 2012):

- **Aprendizaje supervisado:** Este enfoque asume que los datos o bien están catalogados en un cierto número de clases (problema de clasificación), es decir, cada uno de ellos dispone de una etiqueta que indica la clase a la que pertenece, o bien existe una variable de salida que

debe modelarse pero de la que se conocen sus valores correctos para el entrenamiento de los modelos (problema de regresión o modelado).

- **Aprendizaje no supervisado:** En este caso, el análisis se realiza sobre datos en los que no se conoce la salida deseada que debe proporcionar el modelo.

En ambos casos, el objetivo es inducir un modelo para todo el conjunto de datos, o bien descubrir uno o más patrones que encajen con alguna parte del conjunto de datos.

Dentro del aprendizaje no supervisado se encuentran las técnicas de agrupamiento y de reducción de la dimensión. Éstas ya fueron mencionadas en el apartado 1.1. como de gran utilidad a la hora de representar visualmente grandes cantidades de datos con alta dimensionalidad.

1.4. Representación Visual de Datos

1.4.1. Introducción y descripción básica

Ciertamente, la presentación gráfica tiene numerosas ventajas sobre la presentación de resultados en tablas numéricas ya que es más atractiva y suscita mayor interés sobre la atención del lector (Everitt & Hothorn, 2011). Algunas de las ventajas de los métodos gráficos de presentación son las siguientes:

- Las relaciones visuales descritas mediante gráficos se captan más fácilmente y son más fáciles de recordar.
- El uso de gráficos ahorra tiempo ya que el significado esencial de grandes cantidades de datos se puede obtener a partir de una simple observación.
- Los gráficos proporcionan una visión global del problema que proporciona una comprensión más completa y mejor equilibrada que la que podría derivarse de otras formas de presentación mediante tablas o texto.
- Los gráficos pueden poner de manifiesto los hechos y las relaciones ocultas pudiendo estimular, así como ayudar, al pensamiento analítico.

A la anterior lista hemos de añadir que uno de los beneficios más importantes de la presentación gráfica es que permite el acceso a enormes cantidades de datos que no sería posible de otro modo.

La Representación Visual de Datos (en inglés, *Data Visualization*) es más que traducir una tabla de datos en una forma apta para su visualización debiendo presentar los datos de la manera más eficaz con el fin de revelar información de forma rápida, precisa y potente. La creación de un entorno visual puede resumir y comunicar fácilmente los datos a otras personas haciendo incluso las más grandes, o más complicadas, series de datos comprensibles.

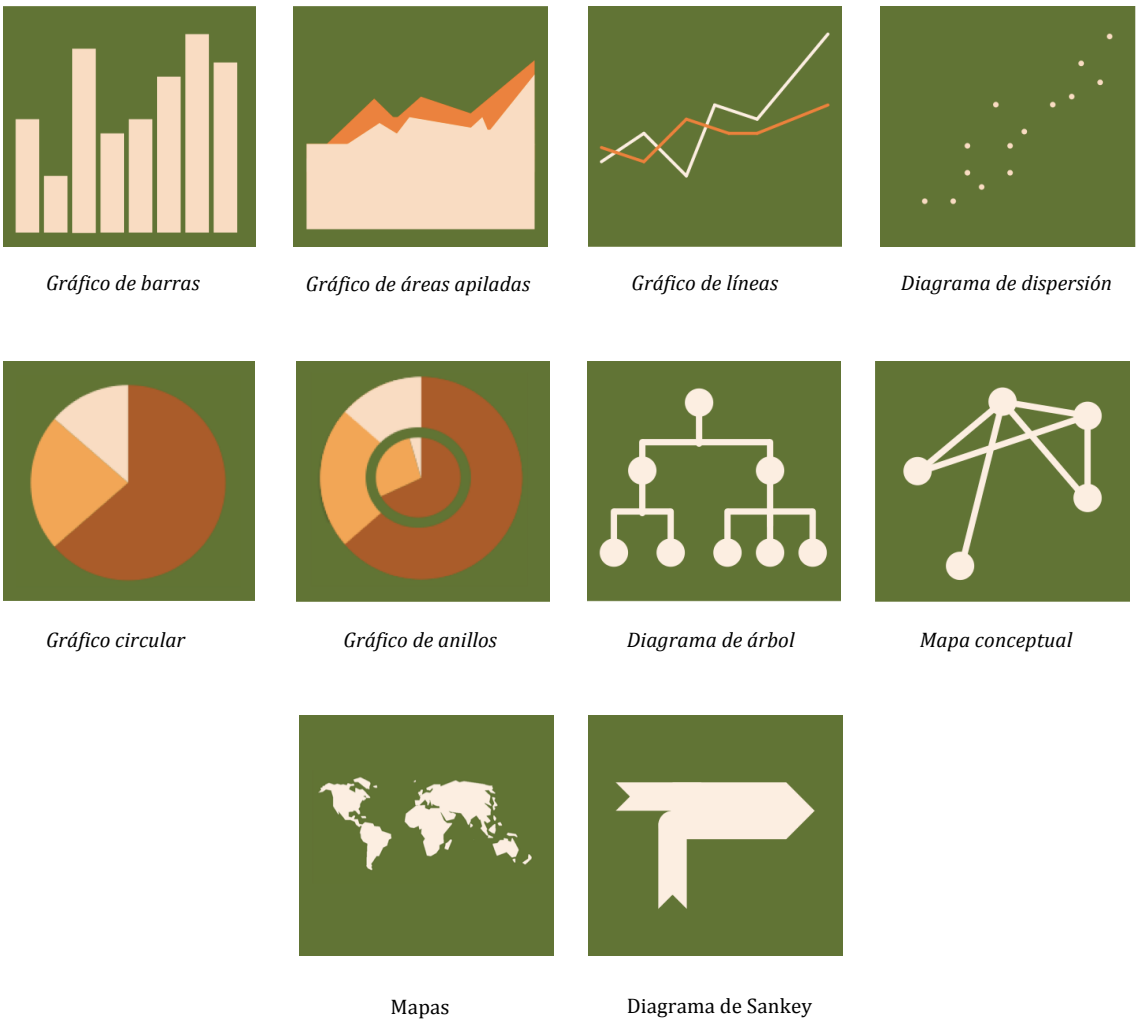


Figura 1-3: *Modelos clásicos de Representación Visual de Datos.*

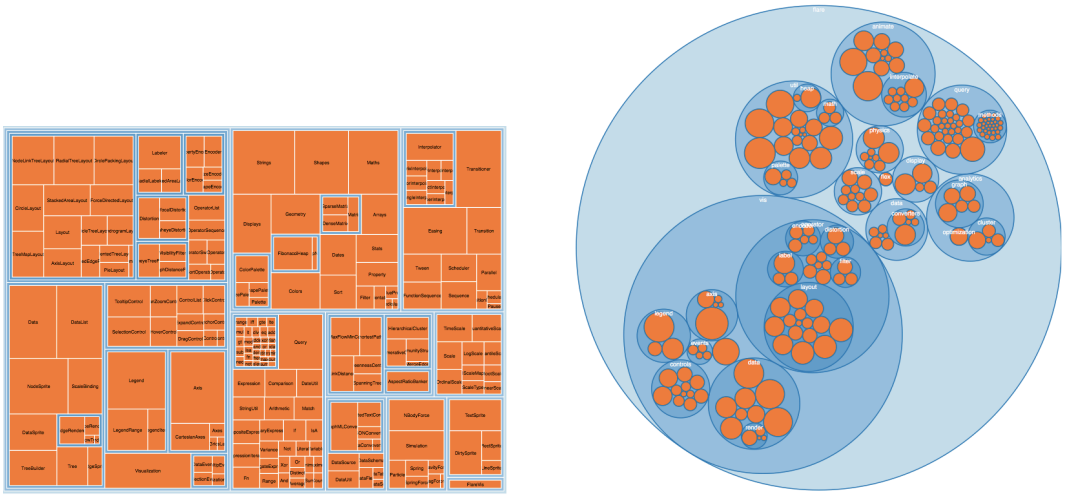
La Representación Visual de Datos puede adoptar muchas formas diferentes, dependiendo de la información que se está comunicando así como del

propósito de la visualización. Las formas más populares que pueden encontrarse se muestran en la figura 1-3. Algunas de ellas son ejemplos simples de Representación Visual de Datos que vienen utilizándose desde hace décadas (Harris, 1999).

Hay muchas ventajas en utilizar esos sencillos modelos. Por una parte, han sido probados para trabajar con eficacia en diferentes conjuntos de datos y, por otra, su amplio y habitual uso facilitan su comprensión. Sin embargo, muchos conjuntos de datos tienen características únicas que obligan a utilizar nuevas formas de comunicar los datos de una manera más eficaz. Cuando únicamente un modelo de visualización de datos no es adecuado, la solución muchas veces es crear una combinación de varios. En este caso cada modelo se puede usar para representar variables específicas en los datos.

En cuanto a las herramientas utilizadas en la Representación Visual de Datos, éstas pueden ser clasificadas en una de las dos grandes categorías siguientes o incluso estar en ambas (Bansal & Sood, 2011):

- **Especializadas en visualizaciones jerárquicas y espaciales:** Son utilizadas especialmente para estructuras determinadas de datos. Por un lado, algunos conjuntos de datos poseen una estructura jerárquica inherente. En estos casos las visualizaciones en árbol pueden ser útiles para la exploración de las relaciones entre los niveles de jerarquía. Ejemplos de este tipo de herramientas son los círculos anidados y los *treemaps* (Heer, Bostock, & Ogievetsky , 2010) (figura 1-4). Otros conjuntos de datos tienen una estructura geográfica o espacial inherente. En estos otros casos, una visualización del mapa o del contexto espacial puede ser útil para la exploración de las relaciones en el conjunto de datos.
- **Especializadas en visualizaciones multidimensionales:** Las herramientas más comúnmente utilizadas en la Representación Visual de Datos son aquellas destinadas a conjuntos de datos multidimensionales. Éstas permiten a los usuarios comparar visualmente algunas dimensiones de los datos (atributos), con otras utilizando un sistema de coordenadas espaciales. También se utilizan para investigar las relaciones entre dos o más atributos continuos o discretos en el conjunto de datos. En la figura 1-5 se muestra un ejemplo de mapa con diagrama de burbujas que, al mismo tiempo que da referencia de la situación espacial, permite la visualización de varias dimensiones a través del tamaño, color o forma de la burbuja.



(a) Treemap. *(b) Círculos anidados.*

Figura 1-4: *Representaciones visuales jerárquicas (Heer, Bostock, & Ogievetsky, 2010). En (a) treemap divide el área recursivamente en rectángulos para mostrar la jerarquía. En (b) los círculos anidados, por el contrario, utilizan círculos para representarla.*

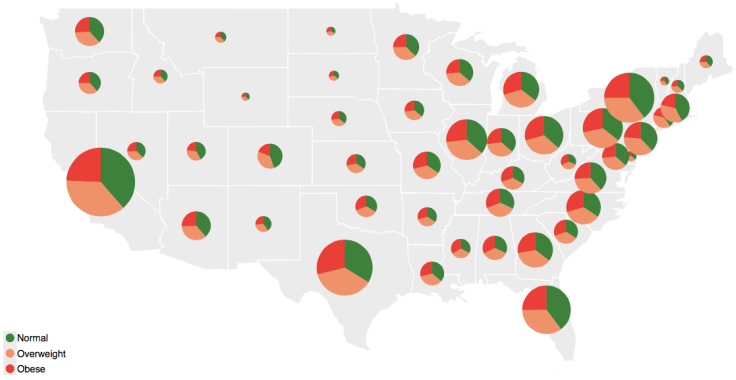


Figura 1-5: *Mapa con diagrama de burbujas. Elaborado por el National Center for Chronic Disease Prevention and Health Promotion, representa gráficamente para cada estado de los Estados Unidos una burbuja de tamaño proporcional a su número de habitantes y en la que se ha dispuesto un gráfico circular donde, a su vez, se expresa el porcentaje de la población dentro de tres categorías (normal, sobrepeso y obesidad). Obtenido el día 27 de febrero de 2015, de la dirección http://hci.stanford.edu/jheer/files/zoo/ex/maps/symbol.html.*

Sea cual sea la forma que se adopte, lo más importante en la Representación Visual de Datos es mantener la excelencia gráfica que consiste en comunicar ideas complejas con claridad, precisión y eficiencia. Ello no quiere decir que la visualización de datos tenga que parecer aburrida para ser funcional o extremadamente sofisticada para que resulte atractiva. Para transmitir ideas de forma efectiva, la estética y la funcionalidad deben ir juntas, facilitando información sobre complejos conjuntos de datos y proporcionando los aspectos clave de una forma más intuitiva.

Un objetivo prioritario para la Representación Visual de Datos es encontrar representaciones generales de conjuntos de datos de más de tres variables. Hay gran variedad de técnicas que abordan este tema, en (Heer, Bostock, & Ogievetsky , 2010) se plantean algunas como *Parallel Coordinate Visualization* (ver figura 1-6) o la matriz de diagramas de dispersión (ver figura 1-7) que utiliza múltiples diagramas de dispersión mostrando las relaciones y correlación entre cada par de variables. Otras técnicas de visualización para conjuntos de datos de alta dimensión se plantean en (Chen, Härdle, & Unwin, 2008) como *Grand Tour* (ver figura 1-8) que permite una visualización dinámica de los datos mediante proyecciones desde diferentes ángulos y también en (Keim, 2002) como *Dense Pixel Displays* (ver figura 1-9) cuya idea básica es representar visualmente cada dimensión en un pixel coloreado y agrupar los píxeles pertenecientes a cada dimensión en áreas adyacentes. Más información sobre herramientas y técnicas sobre visualización de datos puede encontrarse en (Simoff, Böhlen , & Mazeika, 2008; Heer, Bostock, & Ogievetsky , 2010).

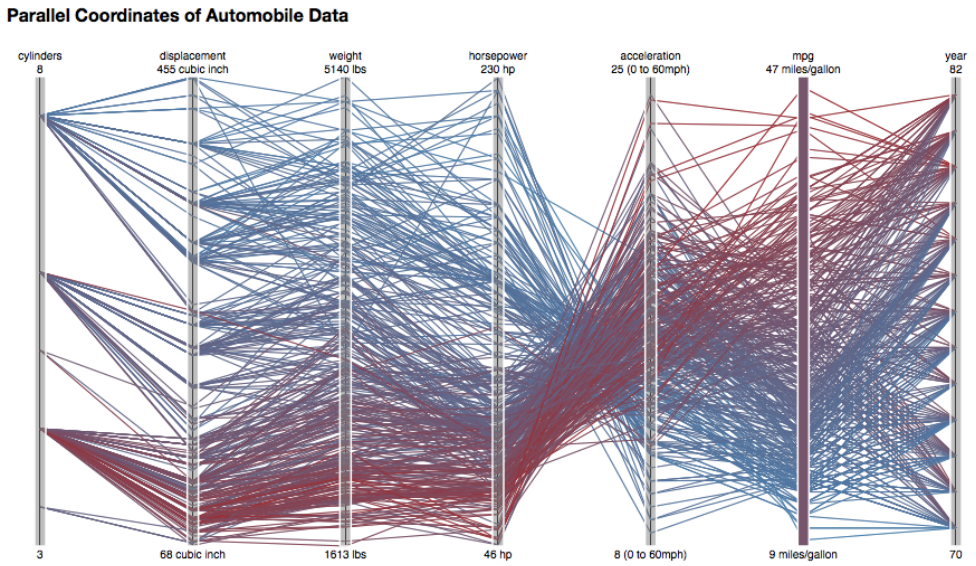


Figura 1-6: *Parallel Coordinate Visualization (Heer, Bostock, & Ogievetsky, 2010). Esta herramienta dispone todos las atributos en ejes paralelos. Cada línea (polilínea) representa un dato, de tal modo que conecta los valores de sus atributos a través de los distintos ejes. El ejemplo muestra de forma gráfica y conjunta siete atributos de distintos automóviles.*

Aunque estas técnicas de visualización de datos de alta dimensión son de probada validez, adolecen de ciertos problemas como, por ejemplo, no ser muy intuitivas, requerir cierto conocimiento previo para poderlas interpretar o no reducir el tamaño ni la cantidad de datos. Por tanto, no son de suficiente ayuda para analizar conjuntos de datos de gran tamaño y/o alta dimensionalidad donde los resultados serían de difícil comprensión.

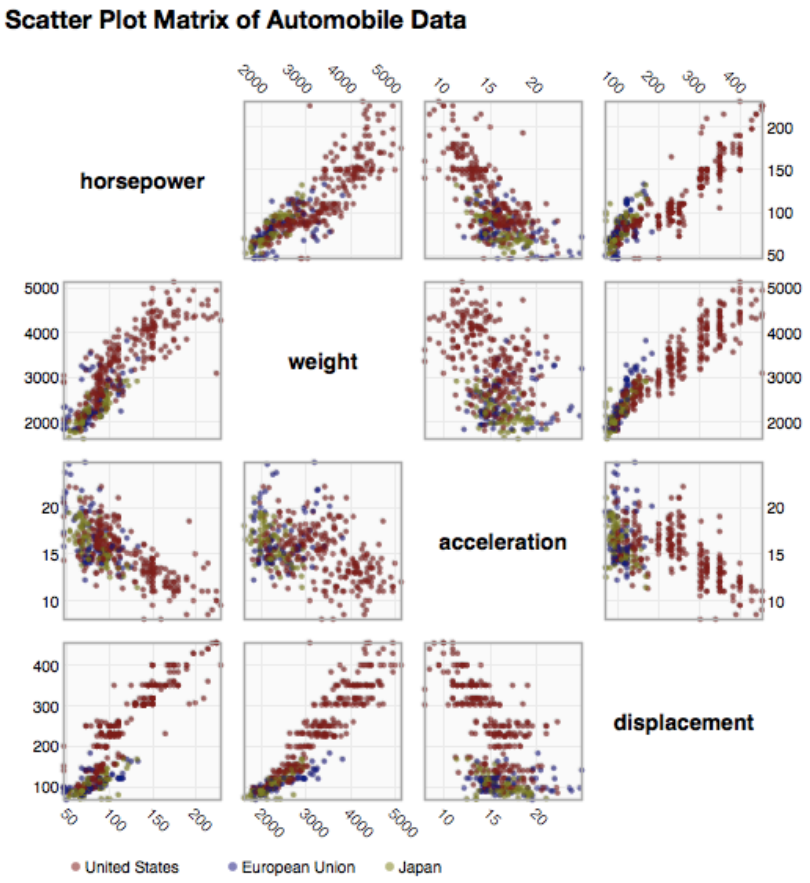


Figura 1-7: *Matriz de diagramas de dispersión (Heer, Bostock, & Ogievetsky, 2010). En* el ejemplo se representa una base de datos de automóviles en los que se ha observado cuatro atributos. Estos atributos se representan gráficamente por parejas señalando cada dato con un color diferente para indicar el país de origen.

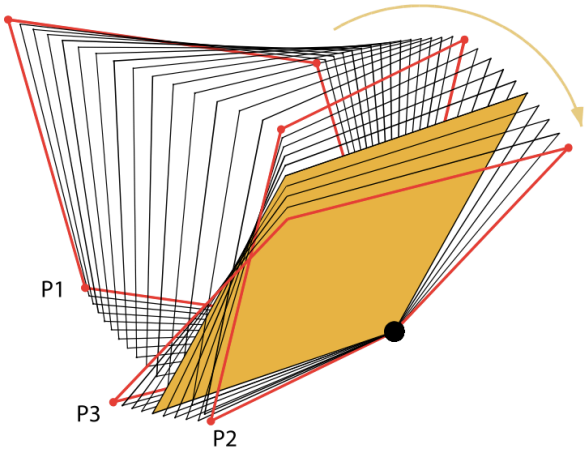
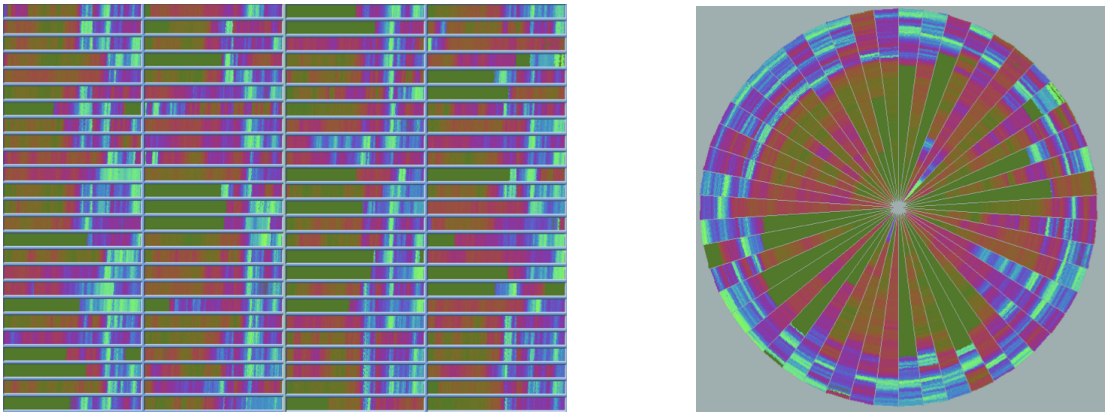


Figura 1-8: *Grand Tour (Chen, Härdle, & Unwin, 2008) es una herramienta dinámica de visualización de datos que permite al analista ver los datos desde todos los ángulos posibles. La idea es proyectar los datos multidimensionales en una línea o un plano mediante rotación mostrando imágenes en una o dos dimensiones de las proyecciones obtenidas.*



(a) Técnica de patrones recursivos. *(b) Técnica de segmentos circulares.*

Figura 1-9: *Dense Pixel Displays (Keim, 2002). Esta herramienta representa cada valor de dimensión en un color de píxel y agrupa los píxeles de cada dimensión en zonas adyacentes. Normalmente, se utiliza un píxel por cada valor de los datos permitiendo que grandes cantidades puedan mostrarse.*

Otro aspecto destacado a tener en cuenta es la posibilidad de interactuar con la representación visual de los datos. Cuando se toma una representación tradicional y se le agrega la combinación de la interacción del usuario y la manipulación directa, se crea una fórmula de gran alcance para el análisis de datos y la extracción de información relevante. Una presentación estática no es particularmente interesante, especialmente cuando existe la posibilidad de un entorno interactivo (Fry, 2008).

Hay muchas formas de interacción del usuario con el entorno de representación de los datos. Estas formas pueden englobarse en dos categorías:

- **Selección y filtrado de datos**, donde el usuario puede controlar qué datos se visualizan. La selección y filtrado de datos ayuda a los usuarios a encontrar y centrarse en aquellos datos que le resulten interesantes en función de que lo se está buscando y ayuda a evitar la sobrecarga de información.
- **Distribución de los datos y navegación por el entorno**, donde el usuario puede controlar cómo se visualizan los datos. Este tipo de interacción puede ayudar a encontrar un nuevo significado en los datos. Simplemente una nueva perspectiva puede ayudar a obtener nuevas conclusiones y ver diferentes relaciones entre los datos.

Permitir a los usuarios controlar estas dos características al instante hace la visualización de datos más potente, simplemente porque la hace más particular y única para el usuario.

Un ejemplo de herramienta interactiva (Stefaner & GmbH) puede verse en la figura 1-10. Esta figura muestra una imagen de "*Better Life Index*" de la Organización para la Cooperación y el Desarrollo Económicos (OCDE), donde se compara el bienestar entre diferentes países. El gráfico se basa en la forma de una flor para representar a cada país. Cada pétalo simboliza cada una de las diferentes variables encontrándose dimensionado de acuerdo con el valor cuantitativo asociado y distinguido a través del color.

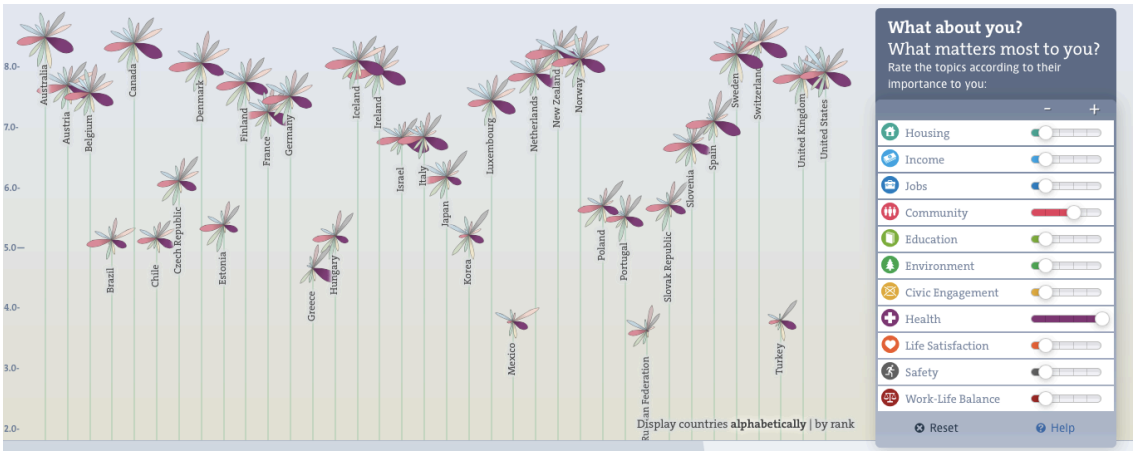


Figura 1-10: *"Better Life Index" de la OCDE (obtenido el 29 de enero de 2015 de http://oecdbetterlifeindex.org).*

En resumen, en este apartado se ha mostrado la importancia de la Representación Visual de Datos para su análisis y se ha realizado un breve repaso a alguna de las técnicas más utilizadas. También se ha subrayado la insuficiencia de esas técnicas para abordar la representación visual de conjuntos de datos de gran tamaño y/o alta dimensionalidad a pesar de su probada validez. Finalmente, se ha destacado la interacción humana en la Representación Visual de Datos como aspecto de gran importancia en la extracción de información. De acuerdo con (Fry, 2008), cada conjunto de datos tiene particulares necesidades de visualización y el propósito para el cual se está utilizando el conjunto de datos tiene tanto efecto sobre esas necesidades como los propios datos.

1.4.2. Representación Visual de Datos y *Data Mining*

Según (Zaixian, 2011), en los últimos años los investigadores de Minería de Datos han empezado a considerar las técnicas de Representación Visual de Datos como un aspecto crítico del proceso de toma de decisiones y el análisis de datos.

Ambos campos son a menudo coalescentes ya que su objetivo común es acceder a las potentes capacidades de procesamiento de imágenes del cerebro humano con el fin de facilitar la extracción de información significativa a partir de datos complejos y, posiblemente, de gran tamaño (Stahl, Gabrys, Medhat Gaber, & Berendsen, 2013).

La visualización de datos tiene dos aplicaciones principales (Chen, Härdle, & Unwin, 2008):

- **Exploración:** Para este tipo de propósito, el analista de datos utilizará muchos gráficos que son, en su mayoría, inadecuados para fines de representación pero que pueden revelar características muy interesantes e importantes.
- **Presentación:** En este tipo de aplicación, los resultados son desplegados a un público más amplio interesado en el conjunto de datos. A modo de ejemplo, en la figura 1-11 se muestra un infográfico en el que se presenta información acerca de un análisis de mercado por parte de la empresa *Panasonic*©. Un infográfico se define como una visualización de datos o ideas que trata de transmitir información compleja de manera que pueda ser interpretada de forma rápida y sencilla (Smiciklas, 2012).

Como resultado, la Representación Visual de Datos se utiliza en diferentes lugares dentro de la Minería de Datos (Bansal & Sood, 2011):

- Como un primer paso en el que se proporciona al usuario una idea de por dónde empezar la minería (aplicación de exploración anterior a la Minería de Datos).
- Como una forma de mostrar los resultados de Minería de Datos y el modelo predictivo de una manera que sea comprensible para el usuario final que no es un experto (aplicación de presentación).
- Como una forma de proporcionar una confirmación de que la Minería de Datos que se llevó a cabo fue correcta (aplicación de exploración posterior a la Minería de Datos).
- Como una forma de realizar Minería de Datos directamente a través de un análisis exploratorio, permitiendo al usuario final buscar y encontrar patrones de manera tan eficiente que se puede hacer en tiempo real sin necesidad de utilizar otras técnicas automatizadas (aplicación de exploración dentro de la Minería de Datos).

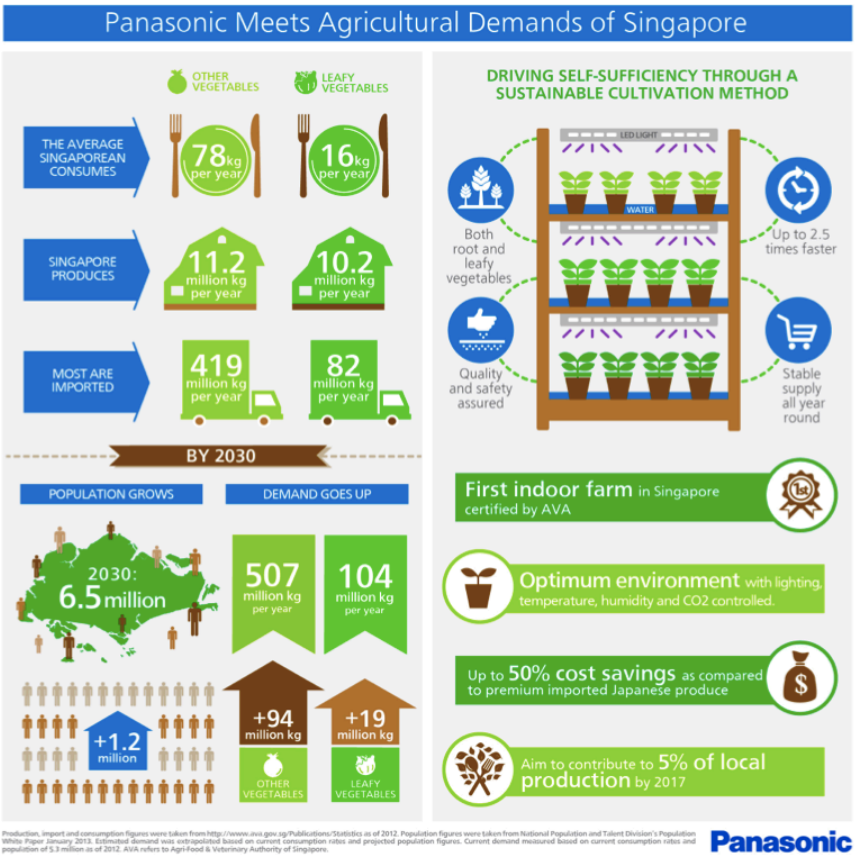


Figura 1-11: *Infográfico de la empresa Panasonic©. En esta representación visual se analizan los factores más significativos sobre población, consumo, producción y demanda en Singapur de productos vegetales en referencia al desarrollo de soluciones para el desarrollo y automatización de la agricultura de interior. Obtenido el 2 de marzo de 2015, de http://www.panasonic.com/sg/corporate/news/article/singaporeindoorfarm.html.*

Para aplicaciones de presentación, en (Fry, 2008) se sugiere un proceso de diseño en siete etapas para la Representación Visual de Datos:

1. Adquirir los datos, ya sea a partir de un documento de *Excel*©, una fuente *XML*, etcétera.
2. Estructurar adecuadamente los datos.
3. Filtrar los datos, eliminando aquellos de menos interés con el objetivo de evitar la sobrecarga de información.
4. Aplicar métodos analíticos como una forma de encontrar patrones o significado en el conjunto de datos.
5. Proceder a la representación visual mediante un modelo básico.
6. Perfeccionar la representación básica para que sea más clara y más atractiva visualmente.

7. Añadir métodos para manipular e interactuar con los datos. Esto permite a los usuarios controlar lo que ven o, incluso, posiblemente cómo lo ven.

Especialmente importante para el desarrollo de esta tesis es la concepción de la Representación Visual de Datos dentro de la Minería de Datos como una fase final en la que el usuario forma parte de ella controlando qué es lo que está viendo con el objetivo de extraer información personalizada y de relevancia.

1.5. *Processing*

Processing© (*https://processing.org*) es un lenguaje de programación y un entorno de desarrollo integrado de código abierto basado en *Java*®. Fue iniciado por Ben Fry y Casey Reas en el *Aesthetics and Computation Group* del *Massachusetts Institute of Technology* (*MIT*) dirigido por John Maeda. Se distribuye bajo la licencia *GNU GPL*. *Processing* es de descarga gratuita y de código abierto. Ha estado en desarrollo desde 2001 siendo utilizado por decenas de miles de personas para el desarrollo de toda clase de trabajos (Fry, 2008; Noble, 2009; Bohnacker, Gross, Laub, & Lazzeroni, 2012).

El entorno de programación de *Processing* hace que sea fácil producir código para obtener imágenes visuales rápidamente. Una vez se controla este escenario, es posible utilizar un habitual entorno integrado de desarrollo o *IDE* (del inglés, *Integrated Development Environment*) *Java* para escribir código de *Processing* ya que la interfaz de programación de aplicaciones o *API* (del inglés, *Application Programming Interface*) se basa en ese lenguaje. Este entorno se le conoce como *PDE* (del inglés, *Processing Development Environment*) (ver figura 1-12).

Según (Fry, 2008), es un lenguaje que combina de forma adecuada el costo, la facilidad de uso y la velocidad de ejecución sobre otros lenguajes de programación como *Flash*® cuya adquisición es cara y tiende a tener problemas con grandes conjuntos de datos o *Python*® y *Ruby*© que, a pesar de su utilidad, tienen velocidades de ejecución inferiores a *Java*.

Este entorno de programación puede producir aplicaciones listas para ser ejecutadas en las tres principales plataformas: *Mac OS*®, *Linux*© *y Windows*®. De igual modo, también se contempla la posibilidad de desarrollar aplicaciones de *Processing*:

- Para su ejecución en Internet (como un *applet* de *Java*).
- Para dispositivos móviles.

- En conexión con dispositivos y prototipos electrónicos como los proyectos *Arduino® y Wiring©*.

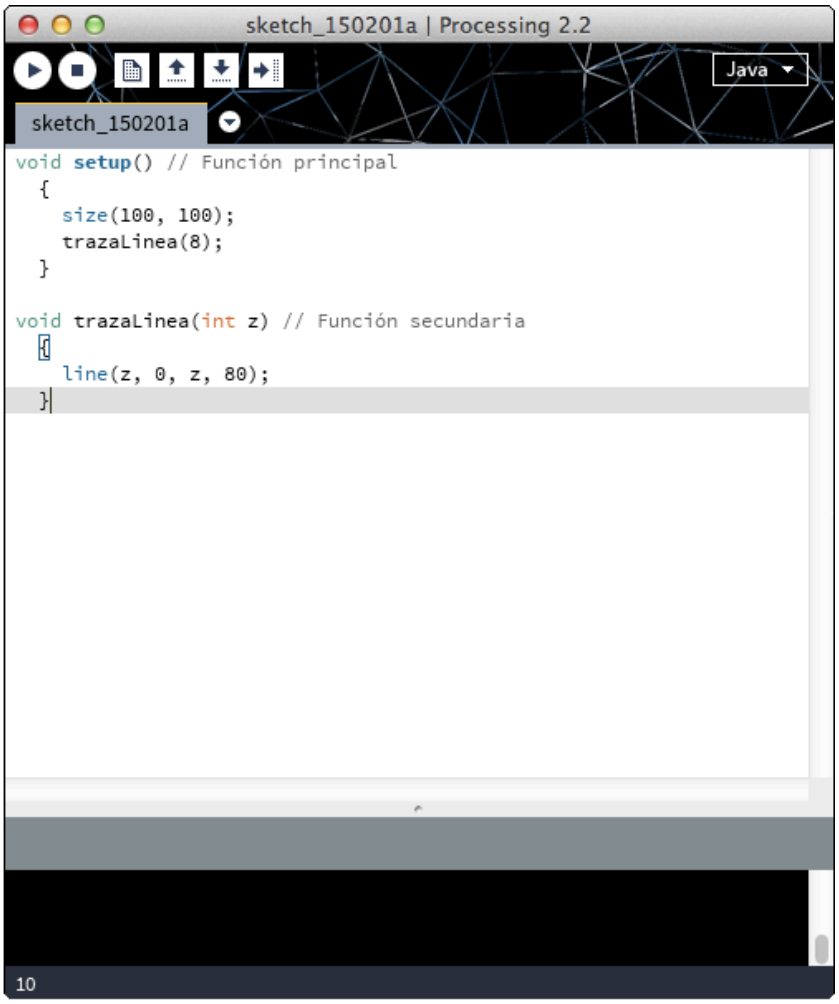


Figura 1-12: *Entorno de programación de Processing©.*

Otra propiedad importante de este entorno es la posibilidad de incorporar animaciones e interactividad de una forma cómoda y sencilla en las aplicaciones generadas.

Una muestra de las posibilidades que ofrece *Processing* lo encontramos en (Morán Álvarez , 2012). En este trabajo, se elabora una aplicación *Processing* interactiva de visualización de datos correspondientes al consumo energético en edificios públicos (ver figura 1-13).

Todas estas características de *Processing* hacen que sea un entorno óptimo para la Representación Visual de Datos, siendo éste el objetivo principal de esta tesis.

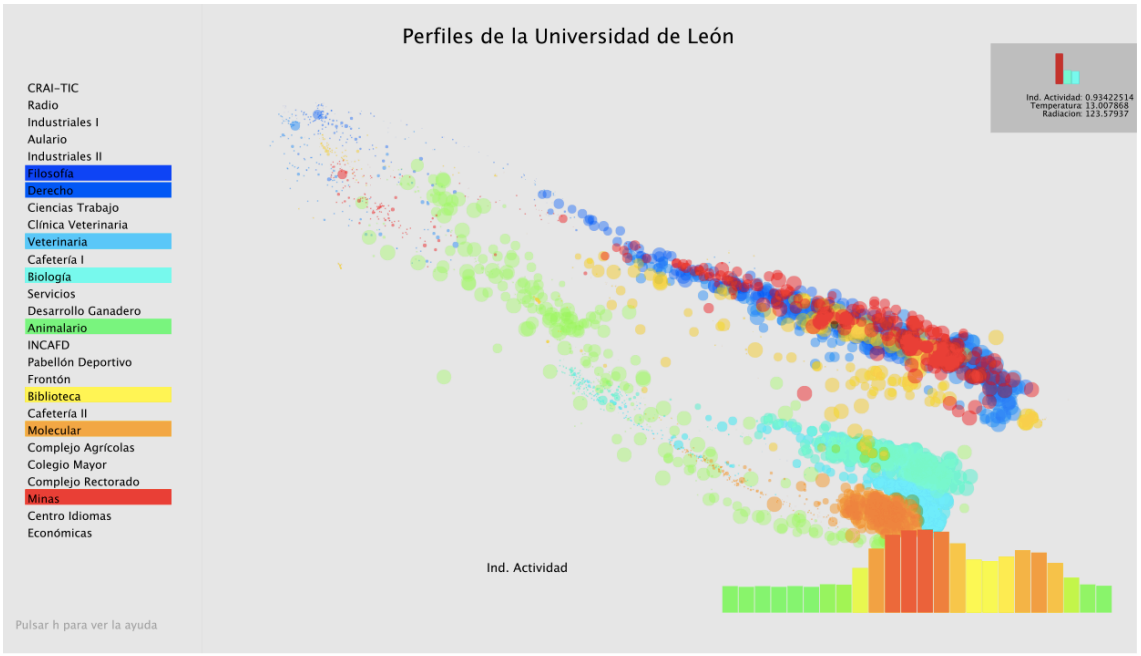


Figura 1-13: *Imagen de aplicación realizada en Processing (Morán Álvarez, 2012). La aplicación desarrollada muestra una mapa de diferencias para la comparación de los perfiles de consumo energético entre distintos edificios de la Universidad de León. Los perfiles son proyectados para su visualización con el fin de marcar la similitud en el comportamiento de estos edificios.*

1.6. Objetivos

El objetivo de esta tesis es desarrollar una aplicación en el entorno *Processing* para la representación visual de conjuntos de datos en dos dimensiones. A tenor de la problemática argumentada en este capítulo en la representación de datos, esta aplicación deberá ser capaz de:

- **Representar gran cantidad de datos**. Cuando el conjunto de datos es de gran tamaño, se hace difícil la exploración visual. En estos casos, la aplicación utilizará técnicas de agrupamiento (en inglés, *clustering*) mediante las cuales se podrá representar varios datos con características similares en uno solo. De este modo, se facilita la exploración visual y el análisis.
- **Representar datos multidimensionales.** Ya se ha comentado en este capítulo que normalmente los datos tienen naturaleza multidimensional por lo que éstos han de ser proyectados a una dimensión inferior para poder ser representados visualmente. En esta aplicación se utilizarán

técnicas de reducción de la dimensión para poder representar el conjunto de datos en dos dimensiones, ideal para la exploración visual.

- **Ofrecer información detallada y global.** Esta aplicación proporcionará información de conjunto, de partes del conjunto (aplicación de técnicas de agrupamiento por parte del usuario) o bien detallada de cada uno de los datos, permitiendo al usuario visualizar en todo momento aquella información que estime relevante. Y todo ello siempre en un entorno gráfico.
- **Interactuar con el usuario.** Permitiendo en todo momento al usuario tener el control de aquello que le interese visualizar. Podrá controlar *zoom*, movimiento, búsqueda de elementos, técnica apropiada de reducción de la dimensión, agrupamientos, …

Y todo ello en un entorno atractivo, intuitivo y muy sencillo de utilizar además de ser versátil.

1.7. Estructura del trabajo

El resto de la memoria de esta tesis se estructura de la siguiente forma:

El capítulo 2 se dedica a la presentación de conceptos básicos del análisis de agrupamiento, ya mencionado como una de las piedras angulares en el desarrollo de la aplicación elaborada en esta tesis, haciendo especial hincapié en aquellos métodos y medidas de proximidad utilizados así como en la forma de validar los resultados obtenidos.

El capítulo 3 trata sobre del segundo factor esencial para la aplicación realizada: la reducción de la dimensionalidad. Aquí se expondrán sus aspectos fundamentales y se realizará un análisis de las técnicas utilizadas así como de la evaluación de la calidad en el empleo de las mismas.

El capítulo 4 presenta la aplicación desarrollada para la visualización de datos describiéndose sus características y los aspectos relacionados con su utilización y gestión.

En el capítulo 5 se ilustra las posibilidades de la aplicación elaborada empleándola en el análisis de cinco conjuntos de datos reales.

Finalmente, en el capítulo 6 se exponen las oportunas conclusiones así como las propuestas de futuro.

Capítulo 2

Análisis de agrupamiento

Resumen

Como ya se ha comentado en el capítulo introductorio, uno de los puntales principales en los que se apoya la aplicación desarrollada en esta tesis para la representación visual de datos es el análisis de agrupamiento. En este capítulo se analizan los conceptos básicos relacionados con esta disciplina.

De este modo, los conceptos de proximidad entre datos y su medida así como los métodos de agrupamiento son examinados haciendo especial hincapié en aquellos relacionados con la aplicación como son los algoritmos K-medias y Fuzzy C-means.

Finalmente, se examinan los criterios de validación utilizados en la aplicación con el fin de proporcionar información acerca de los resultados obtenidos al realizar un determinado agrupamiento.

2.1. Introducción

Una de las actividades más importantes de entre todas las relacionadas con el análisis de datos es la de clasificarlos o agruparlos dentro de una variedad de categorías o grupos. La idea que se persigue es que los datos que se clasifiquen en un mismo grupo muestren propiedades similares en base a algunos criterios

establecidos (Xu & Wunsch, 2009).

La clasificación, además de ser una actividad conceptual básica del ser humano, es también fundamental en la mayoría de las ramas de la ciencia (Everitt, Landau, Leese, & Stahl, 2011). Con el objetivo de comprender un nuevo objeto o fenómeno, éste se examina tratando de identificar sus características descriptivas y compararlas con las de otros objetos o fenómenos conocidos para establecer grados de similitud o desemejanza de acuerdo con ciertas normas y reglas con el fin de poderlo clasificar. Con la clasificación, se pueden inferir propiedades de objetos específicos en función de la categoría a la que pertenecen y disponer de una organización mediante la cual poderlos entender más fácilmente.

La clasificación puede ser de dos tipos: supervisada y no supervisada. En la clasificación supervisada (Kotsiantis, 2007), para la tarea de clasificar un objeto dentro de una categoría o clase se cuenta con modelos ya agrupados que tienen características comunes. Podemos diferenciar dos fases generales dentro de este tipo de clasificación:

- Una primera fase donde se tiene un conjunto de entrenamiento o de aprendizaje (utilizado en el diseño del clasificador en base a un algoritmo de selección) y otro llamado de test o de validación (para la evaluación del clasificador). En esta fase se construye el modelo o regla general para la clasificación.
- Una segunda fase donde se hace uso del modelo diseñado anteriormente para clasificar objetos de los que se desconoce la clase a la que pertenecen.

Por otro lado, en la clasificación no supervisada, también llamada análisis de agrupamiento (en inglés, *clustering*), los datos disponibles no se encuentran clasificados. El objetivo del agrupamiento es separar un conjunto de datos sin clasificación previa en un número determinado de grupos (en inglés, *clusters*), de tal modo que aquellos que pertenezcan a un grupo deberán ser similares entre sí, mientras que aquellos que están en diferentes grupos habrán de ser lo más disimilares posible. Esto implica la necesidad de precisar para un conjunto de datos los conceptos de semejanza y desemejanza de forma clara y significativa (ver figura 2-1).

Los resultados obtenidos en el análisis de agrupamiento de un conjunto de datos dependen de:

- La medida de proximidad utilizada para comparar los datos. Este aspecto es tratado en la sección 2.2.
- El método de agrupamiento seleccionado. Que se revisa dentro de la sección 2.3.

Es por ello que un criterio o una técnica de agrupamiento diferente, e incluso una misma técnica pero configurada con distintos parámetros, puede provocar resultados completamente diferentes en el agrupamiento de un mismo conjunto de datos. Sin embargo, no hay ninguna forma de determinar qué criterio es mejor en general ya que ello depende de la situación particular del problema a analizar.

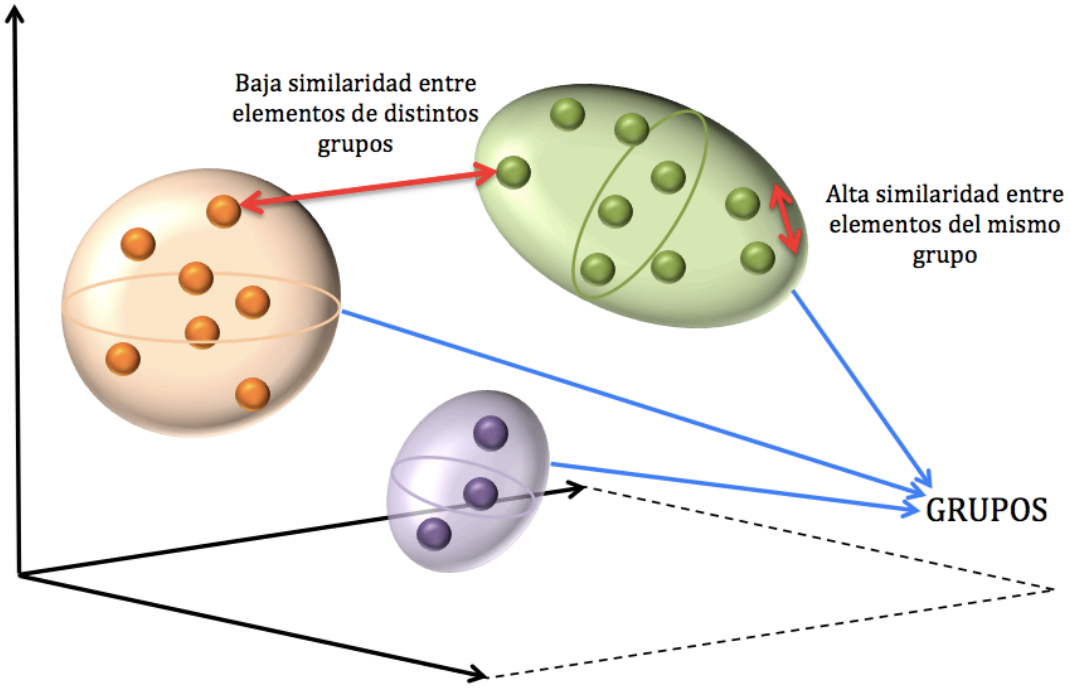


Figura 2-1: *Objetivo del agrupamiento: En la figura se representa un agrupamiento interpretando la desemejanza entre datos en términos de distancia. El objetivo del agrupamiento es clasificar los distintos datos en grupos intentando maximizar la semejanza entre elementos del mismo grupo y la desemejanza entre datos de grupos distintos.*

Casi siempre se da el caso en el que existen diversas clasificaciones alternativas para el mismo conjunto de datos. La figura 2-2 ilustra un ejemplo del efecto de la subjetividad en las agrupaciones resultantes en un conjunto formado por veinte datos representados gráficamente (figura 2-2(a)). En la figura 2-2(b) una partición gruesa divide el conjunto de datos en dos grupos principales,

mientras que una mas fina sugiere que los datos pueden agruparse en cuatro o seis (figura 2-2(c) y figura 2-2(d) respectivamente). El hecho de adoptar un tipo de agrupamiento u otro depende de las exigencias específicas del problema y, en general, no puede determinarse si los resultados son mejores o peores.

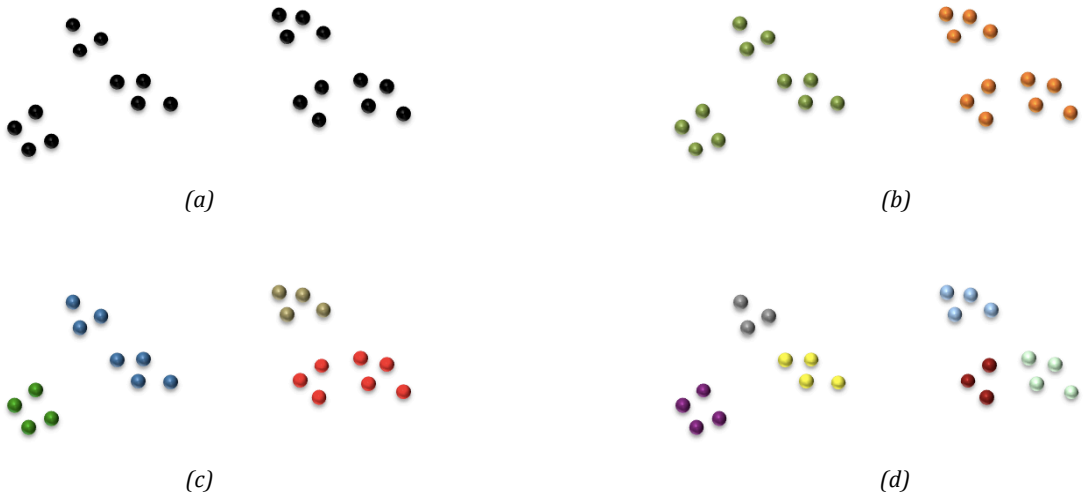


Figura 2-2: *Subjetividad respecto del número de agrupaciones para un mismo conjunto de datos*

El análisis de agrupamiento ha sido y es aplicado en una amplia variedad de campos como Ingeniería, Ciencias de la Computación, Salud y Medicina, Astronomía, Ciencias Sociales o Economía. Ejemplos de tareas realizadas son reconocimiento biométrico (Shotton, et al., 2013), segmentación (Müller & Hamm, 2014), análisis de archivos de registro en *Internet* (patrones de acceso similares) (Sudhamathy & Venkateswaran, 2011), elaboración de mapas *GIS* (*Geographical Information System*) (Gupta, Kumar, Singh, & Kumar, 2015) y un largo etc.

El análisis de agrupamiento también puede utilizarse como paso previo a otras técnicas de Minería de Datos en tareas de reducción de datos o detección de datos anómalos (Koupaie, Ibrahim, & Hosseinkhani, 2013), entre otras. El aspecto relacionado con la reducción de datos es muy importante para tareas de visualización de grandes conjuntos y es utilizado en el desarrollo de la aplicación de esta tesis.

En los siguientes apartados de este capítulo se admitirá disponer de un conjunto de n datos respecto de los cuales se han observado D variables o atributos y que se hallan dispuestos en una matriz $\boldsymbol{X}$ de dimensiones $n \times D$ ($\boldsymbol{X}_{n \times D}$) donde cada dato se representa por un vector $\boldsymbol{X}_i$ ($i \in \{1,2,\ldots,n\}$) y los valores de cada variable observados en el conjunto se recogen en el vector $\boldsymbol{X}^j$ ($j \in \{1,2,\ldots,D\}$).

2.2. Medidas de proximidad

El primer factor fundamental del cual depende el análisis de agrupamiento es cuantificar la proximidad, es decir, medir la semejanza o desemejanza respecto de dos datos, un dato y un grupo de datos o entre dos grupos.

De este modo, el análisis de agrupamiento para un conjunto de n datos tiene como punto de partida una matriz de dimensiones $n \times n$ cuyos elementos reflejan, en cierto sentido, una medida cuantitativa de la semejanza o desemejanza entre los datos. Dos individuos están próximos cuando el valor de su semejanza es grande o el de su desemejanza pequeño.

Para discutir las medidas de proximidad se van a considerar aquellas que adecuadas para datos con atributos categóricos, para datos cuyos atributos son continuos y, finalmente, se hará referencia al caso en el que los datos contienen ambos tipos de variable.

2.2.1. Medidas de proximidad para datos categóricos

La medida de proximidad entre datos categóricos (también conocidos como datos nominales o cualitativos) no es algo sencillo debido a que no existe noción explícita para poder ordenar sus valores. En aquellos conjuntos de datos en los que todas las variables son categóricas, las medidas de proximidad más comúnmente utilizadas son las de similitud. En el conjunto de datos representado por la matriz $\boldsymbol{X}$, una medida de similitud entre sus elementos se puede definir como una función $S: \boldsymbol{X} \times \boldsymbol{X} \rightarrow \mathbb{R}$ que verifica las siguientes propiedades (Lesot , Rifqi, & Benhadda, 2009):

- $\forall\, \boldsymbol{X}_i,\ \boldsymbol{X}_j \ / \ i,j \in \{1,2,\ldots,n\},\ S(\boldsymbol{X}_i, \boldsymbol{X}_j) \geq 0$ (es positiva) (2-1)
- $\forall\, \boldsymbol{X}_i,\ \boldsymbol{X}_j \ / \ i,j \in \{1,2,\ldots,n\},\ S(\boldsymbol{X}_i, \boldsymbol{X}_j) = S(\boldsymbol{X}_j, \boldsymbol{X}_i)$ (es simétrica) (2-2)
- $\forall\, \boldsymbol{X}_i,\ \boldsymbol{X}_j \ / \ i\,,j \in \{1,2,\ldots,n\},\ S(\boldsymbol{X}_i,\ \boldsymbol{X}_i) \geq S(\boldsymbol{X}_i,\ \boldsymbol{X}_j)$ (2-3)

aunque puede ocurrir que alguna de estas propiedades como la exigencia de simetría (propiedad (2-2)) se relaje dando lugar a definiciones de similitud más generales. Normalmente, las medidas quedan restringidas al intervalo $[0,1]$, aunque en ocasiones pueden expresarse en forma de porcentaje siendo su rango de 0 a 100.

Con esa definición, si dos datos $\boldsymbol{X}_i$ y $\boldsymbol{X}_j$ tienen un valor de similitud uno indica que tienen valores idénticos para todas las variables mientras que si el valor es cero, muestra que los valores en todas las variables difieren al máximo para ambos.

El caso más común de datos categóricos multivariantes es aquel donde todas las variables son binarias, es decir, los valores que toman todas las variables o atributos son 0 (atributo no presente) ó 1 (atributo presente). En este caso, para cada dos datos $\boldsymbol{X}_i$ y $\boldsymbol{X}_j$ se definen los siguientes valores:

- a: número de atributos presentes en ambos datos.
- b: número de atributos presentes en $\boldsymbol{X}_i$ pero no en $\boldsymbol{X}_j$.
- c: número de atributos presentes en $\boldsymbol{X}_j$ pero no en $\boldsymbol{X}_i$.
- d: número de atributos que no se hayan presentes en ninguno de los dos datos.

A partir de estos valores, se han elaborado una gran cantidad de medidas de similitud algunas de las cuales se muestran en la tabla 2-1. Otras más pueden encontrarse en (Lesot , Rifqi, & Benhadda, 2009).

Tabla 2-1: *Medidas de similitud para datos binarios*

Medida de Similitud	*Definición*
Jaccard	$\frac{a}{a+b+c}$
Dice	$\frac{2a}{2a+b+c}$
Sorensen	$\frac{4a}{4a+b+c}$
Ochiai	$\frac{a}{\sqrt{a+b}+\sqrt{a+c}}$
Russel and Rao	$\frac{a}{a+b+c+d}$
Sokal and Michener	$\frac{a+d}{a+b+c+d}$
Rogers and Tanimoto	$\frac{a+d}{a+2(b+c)+d}$

Para aquellos datos categóricos con variables que tienen más de dos estados, una estrategia sencilla y directa para poder medir su similitud es definir para cada estado de estas variables una nueva variable binaria que indique la presencia o ausencia de ese estado en la variable. La desventaja de este método es que existe la posibilidad de introducir demasiadas variables binarias (Xu & Wunsch, 2009).

Otro método más eficaz y de uso común es asignar una similitud $w \geq 0$ si los valores en una determinada variable categórica coinciden y una similitud de 0 en caso contrario. Dado que los datos son categóricos multivariados, la similitud entre ellos será directamente proporcional al número de atributos en los que coincidan. Esta medida es también conocida como la medida de solapamiento (Boriah, Chandola, & Kumar, 2008). Por lo general $w = 1$, sin embargo, el valor de w puede variar en base a las propiedades y los requisitos de los datos. Por ejemplo, si existe un gran número de posibles valores para una característica, es razonable dar a w un valor superior a uno con el fin de dar más peso a la coincidencia en esta característica.

Más información sobre medidas para datos categóricos puede encontrarse en (Boriah, Chandola, & Kumar, 2008).

2.2.2. Medidas de proximidad para datos continuos

Cuando todas las variables registradas son continuas, la proximidad entre los datos suele ser cuantificada por medidas de disimilitud o medidas de distancia. Una medida de disimilitud definida en el conjunto representado mediante la matriz $\boldsymbol{X}$ es una distancia si es una función de la forma $Dist: \boldsymbol{X} \times \boldsymbol{X} \rightarrow \mathbb{R}$ que cumple las siguientes propiedades:

- $\forall\, \boldsymbol{X}_i,\ \boldsymbol{X}_j\ /\ i,j \in \{1,2,\ldots.,n\},\ Dist(\boldsymbol{X}_i, \boldsymbol{X}_j) \geq 0$

  $$\text{y } Dist(\boldsymbol{X}_i, \boldsymbol{X}_j) = 0 \leftrightarrow \boldsymbol{X}_i = \boldsymbol{X}_j \qquad (2\text{-}4)$$

- $\forall\, \boldsymbol{X}_i,\ \boldsymbol{X}_j\ /\ i,j \in \{1,2,\ldots.,n\},\ Dist(\boldsymbol{X}_i, \boldsymbol{X}_j) = Dist(\boldsymbol{X}_j, \boldsymbol{X}_i) \qquad (2\text{-}5)$

- $\forall\, \boldsymbol{X}_i,\ \boldsymbol{X}_j,\ \boldsymbol{X}_k /\ i\ , j, k \in \{1,2,..,n\},\ S(\boldsymbol{X}_i, \boldsymbol{X}_k) \leq S(\boldsymbol{X}_i, \boldsymbol{X}_j) + S(\boldsymbol{X}_j, \boldsymbol{X}_k) \qquad (2\text{-}6)$

A la propiedad (2-6) se le suele conocer con el nombre de desigualdad triangular.

Hay una amplia variedad de medidas propuestas con el fin de obtener las medidas de disimilitud en un conjunto de observaciones multivariantes continuas. Algunas de las medidas de disimilitud más utilizadas se resumen en la tabla 2-2 (Abdesselam & Zighed, 2012).

De todas las medidas expuestas, se han implementado en la aplicación desarrollada en esta tesis la *distancia Euclídea* y la *distancia de Mahalanobis* (Xiang, Nie, & Zhang, 2008) por ser comúnmente utilizadas en el ámbito del análisis de agrupamiento.

Tabla 2-2: *Medidas de disimilitud para datos continuos. Donde* $(\boldsymbol{\alpha_k})_{\boldsymbol{k=1,\dots,D}} \geq \mathbf{0}$, $\boldsymbol{\Sigma^{-1}}$ *es la inversa de la matriz de covarianzas,* $(\boldsymbol{\sigma_k})_{\boldsymbol{k=1,\dots,D}}$ *es la desviación típica en la característica k, p > 0,* $\boldsymbol{m_{ij}^k} = \frac{X_i^k + X_j^k}{2}$ *y* $\boldsymbol{\rho}(\boldsymbol{X_i}, \boldsymbol{X_j})$ *es el coeficiente de correlación lineal de Bravais-Pearson.*

Medida	*Definición*
Eucídea	$Dist(\boldsymbol{X}_i, \boldsymbol{X}_j) = \sqrt{\sum_{k=1}^{D} (\boldsymbol{X}_i^k - \boldsymbol{X}_j^k)^2}$
Mahalanobis	$Dist(\boldsymbol{X}_i, \boldsymbol{X}_j) = \sqrt{(\boldsymbol{X}_i - \boldsymbol{X}_j)^t \, \Sigma^{-1} \, (\boldsymbol{X}_i - \boldsymbol{X}_j)}$
Manhattan	$Dist(\boldsymbol{X}_i, \boldsymbol{X}_j) = \sum_{k=1}^{D} \left\lvert \boldsymbol{X}_i^k - \boldsymbol{X}_j^k \right\rvert$
Minkowski	$Dist(\boldsymbol{X}_i, \boldsymbol{X}_j) = \left(\sum_{k=1}^{D} (\boldsymbol{X}_i^k - \boldsymbol{X}_j^k)^p \right)^{\frac{1}{p}}$
Tchebytchev	$Dist(\boldsymbol{X}_i, \boldsymbol{X}_j) = max_{1 \leq k \leq D} \left\lvert \boldsymbol{X}_i^k - \boldsymbol{X}_j^k \right\rvert$
Disimilitud del coseno	$Dist(\boldsymbol{X}_i, \boldsymbol{X}_j) = 1 - \frac{\langle \boldsymbol{X}_i, \boldsymbol{X}_j \rangle}{\lVert \boldsymbol{X}_i \rVert \lVert \boldsymbol{X}_j \rVert}$
Canberra	$Dist(\boldsymbol{X}_i, \boldsymbol{X}_j) = \sum_{k=1}^{D} \frac{\left\lvert \boldsymbol{X}_i^k - \boldsymbol{X}_j^k \right\rvert}{\left\lvert \boldsymbol{X}_i^k \right\rvert + \left\lvert \boldsymbol{X}_j^k \right\rvert}$
Distancia de cuerda	$Dist(\boldsymbol{X}_i, \boldsymbol{X}_j) = \sum_{k=1}^{D} \left(\sqrt{\boldsymbol{X}_i^k} - \sqrt{\boldsymbol{X}_j^k} \right)^2$
Euclídea ponderada	$Dist(\boldsymbol{X}_i, \boldsymbol{X}_j) = \sqrt{\sum_{k=1}^{D} \alpha_k (\boldsymbol{X}_i^k - \boldsymbol{X}_j^k)^2}$
Ji-cuadrado	$Dist(\boldsymbol{X}_i, \boldsymbol{X}_j) = \sqrt{\sum_{k=1}^{D} \frac{(\boldsymbol{X}_i^k - \boldsymbol{m}_{ij}^k)^2}{\boldsymbol{m}_{ij}^k}}$
Divergencia de Jeffrey	$Dist(\boldsymbol{X}_i, \boldsymbol{X}_j) = \sum_{k=1}^{D} \left(\boldsymbol{X}_i^k \, log \frac{\boldsymbol{X}_i^k}{\boldsymbol{m}_{ij}^k} + \boldsymbol{X}_j^k \, log \frac{\boldsymbol{X}_j^k}{\boldsymbol{m}_{ij}^k} \right)$
Correlación de Pearson	$Dist(\boldsymbol{X}_i, \boldsymbol{X}_j) = 1 - \left\lvert \rho(\boldsymbol{X}_i, \boldsymbol{X}_j) \right\rvert$
Eucídea normalizada	$Dist(\boldsymbol{X}_i, \boldsymbol{X}_j) = \sqrt{\sum_{k=1}^{D} \left(\frac{\boldsymbol{X}_i^k - \boldsymbol{X}_j^k}{\sigma_k} \right)^2}$

2.2.3. Medidas de proximidad para datos mixtos

Cuando los datos están formados por atributos de tipo mixto, es decir, categóricos y continuos, una forma de calcular la proximidad es combinando métodos como los mencionados en los apartados anteriores. La disimilitud $Disim(\boldsymbol{X}_i, \boldsymbol{X}_j)$ entre dos datos conteniendo D atributos de tipo mixto puede definirse como (Rokach & Maimon, 2005):

$$Disim(\boldsymbol{X}_i, \boldsymbol{X}_j) = \sum_{k=1}^{D} w_{ijk} \cdot d_{ijk} \Big/ \sum_{k=1}^{D} w_{ijk} \tag{2-7}$$

donde $w_{ijk} = 0$ si alguno de los valores no posee medida en el atributo *k-ésimo*. La disimilitud d_{ijk} se calcula para cada atributo del siguiente modo:

- Si el atributo es binario o categórico $d_{ijk} = 0$ si $\boldsymbol{X}_i^k = \boldsymbol{X}_j^k$ y $d_{ijk} = 1$ en otro caso.
- Si el atributo es continuo $d_{ijk} = \left|\boldsymbol{X}_i^k - \boldsymbol{X}_j^k\right| / R_k$ donde R_k es el rango de observaciones para la el atributo *k-ésimo.*

Otras formas de abordar las medidas de proximidad para datos mixtos se proponen en (Everitt, Landau, Leese, & Stahl, 2011).

2.3. Métodos de agrupamiento

Una vez determinada la medida de proximidad apropiada, el siguiente paso es utilizar un algoritmo de agrupamiento que permita determinar los datos en diferentes grupos de acuerdo con esa medida de proximidad. Este paso se realiza en base a la optimización de algún criterio específico expresado en forma de función.

No existe ningún algoritmo de agrupamiento universal para resolver todos los problemas, es por ello que es importante conocer las características del problema particular con el fin de seleccionar el método que mejor pueda adaptarse.

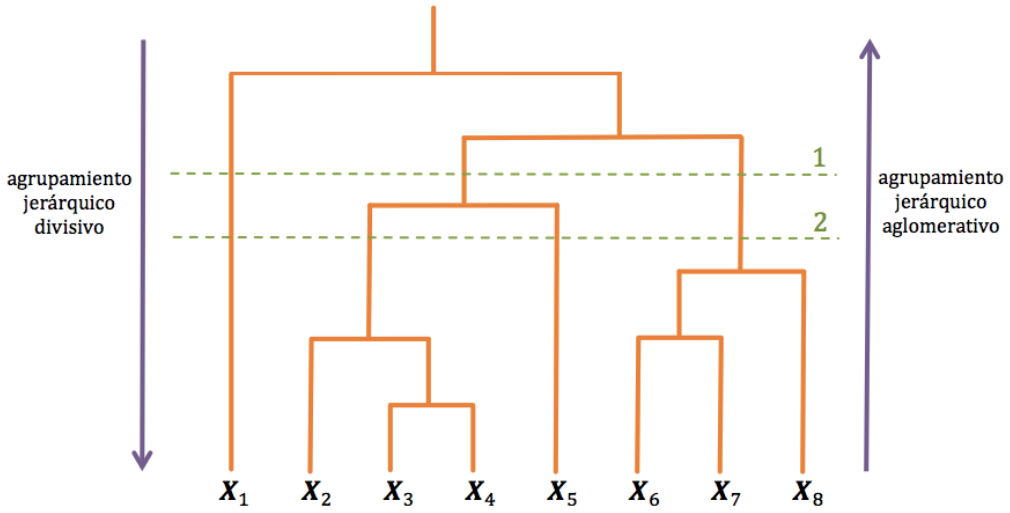


Figura 2-3: *Ejemplo de dendrograma para agrupamiento jerárquico. La dirección para el agrupamiento en el agrupamiento jerárquico divisivo es la opuesta del agrupamiento jerárquico aglomerativo. Según el corte del dendrograma (líneas discontínuas) se obtienen los distintos agrupamientos: en la línea 1 se obtienen tres grupos: $\{X_1\}$, $\{X_2, X_3, X_4, X_5\}$ y $\{X_6, X_7, X_8\}$; mientras que utilizando la línea 2 se obtienen cuatro:$\{X_1\}$, $\{X_2, X_3, X_4\}$, $\{X_5\}$ y $\{X_6, X_7, X_8\}$.*

Los métodos más utilizados en análisis de agrupamiento pueden clasificarse en cuatro tipos (Mann & Kaur, 2013):

- **Métodos jerárquicos**: Estos métodos realizan los agrupamientos de los datos basándose en una estructura jerárquica a partir de una secuencia de particiones anidadas, bien sea desde grupos aislados a un solo grupo que incluye todos los individuos (métodos aglomerativos) o viceversa (métodos divisivos). Los resultados del agrupamiento jerárquico se representan generalmente mediante un árbol binario o dendrograma, como se representa en la figura 2-3. El nodo raíz del dendrograma representa todo el conjunto de datos, y cada nodo hoja es considerado como un grupo de un solo dato. Las diferentes agrupaciones pueden obtenerse mediante el corte del dendrograma a diferentes niveles (líneas discontinuas 1 y 2 de la figura 2-3).
- **Métodos de partición**: En estos métodos los datos se dividen directamente en algún número de grupos previamente especificado, sin ninguna estructura jerárquica. Dentro de este tipo pueden distinguirse dos clases:
    - **Métodos de partición estricta**: En los que cada dato pertenece exclusivamente a un único grupo. Los grupos en los que se divide el conjunto de datos son disjuntos dos a dos, no

vacíos y su unión abarca a todo el conjunto de datos. El algoritmo *K-medias* es un método de esta clase implementado en la aplicación desarrollada en esta tesis y será analizado en la sección 2.3.1.

- **Métodos de partición difusa**: En estos métodos se permite a un dato pertenecer a todos los grupos con un cierto grado de pertenencia. El algoritmo *Fuzzy C-Means* implementado en la aplicación es un método de esta clase que se analiza en la sección 2.3.2.

- **Métodos basados en la densidad**: En estos métodos, las agrupaciones se definen como zonas donde la densidad de datos es mayor que en el resto del conjunto de datos. Los datos situados en áreas poco densas por lo general se consideran como ruido o datos de frontera. Algoritmos representativos son *Density-Based Algorithms for Discovering Clusters in Large Spatial Databases with Noise* (*DBSCAN*), *Varied Density Based Spatial Clustering of Applications with Noise* (*VDBSCAN*), *Density Based Algorithm for discovering Density Varied Clusters in Large Spatial Databases* (*DVBSCAN*), *Distribution-Based clustering Algorithm for Mining Large Spatial Databases* (*DBCLASD*) *y Spatial-Temporal Density Based Clustering* (*ST-DBSCAN*) (Parimala, Lopez, & Senthilkumar, 2011).

- **Métodos basados en cuadrícula**: Estos métodos utilizan una estructura de datos en cuadrícula de varias resoluciones. El espacio de los datos se divide en un número finito de celdas que forma una estructura de rejilla donde se lleva a cabo todas las operaciones de análisis de agrupamiento. La principal ventaja de este enfoque es su tiempo de procesamiento, que normalmente es independiente del número de datos, dependiendo únicamente del número de celdas en las que el espacio se subdivide. Ejemplos de algoritmos de agrupamiento basados en cuadrícula son *STatistical INformation Grid* (*STING*), *CLustering In QUEst* (*CLIQUE*) o *Wave Cluster* entre otros (Ilango & Mohan, 2010).

2.3.1. Algoritmo *K-medias*

El algoritmo *K-medias* es uno de los algoritmos de agrupamiento más conocidos y populares (Xu & Wunsch, 2009). Dado el conjunto de datos de elementos $\boldsymbol{X}_i$ ($i \in \{1,2,\ldots,n\}$) y un número de grupos *K*, *K-medias* busca una partición óptima de los datos en esos *K* grupos que minimice la expresión:

$$J = \sum_{j=1}^{K} \sum_{i=1}^{n} v_{ij} \left\| \boldsymbol{X}_i - \boldsymbol{m}_j \right\|^2 \tag{2-8}$$

donde $v_{ij} = \begin{cases} 1 & \text{si } \boldsymbol{X}_i \text{ pertenece al grupo } j \\ 0 & \text{en otro caso} \end{cases}$ con lo que $\sum_{j=1}^{K} v_{ij} = 1 \ \ \forall i$ y siendo $\boldsymbol{m}_j = \frac{1}{n_j} \sum_{i=1}^{n} v_{ij} \boldsymbol{X}_i$ la media muestral o centroide del *j-ésimo* grupo con n_j datos.

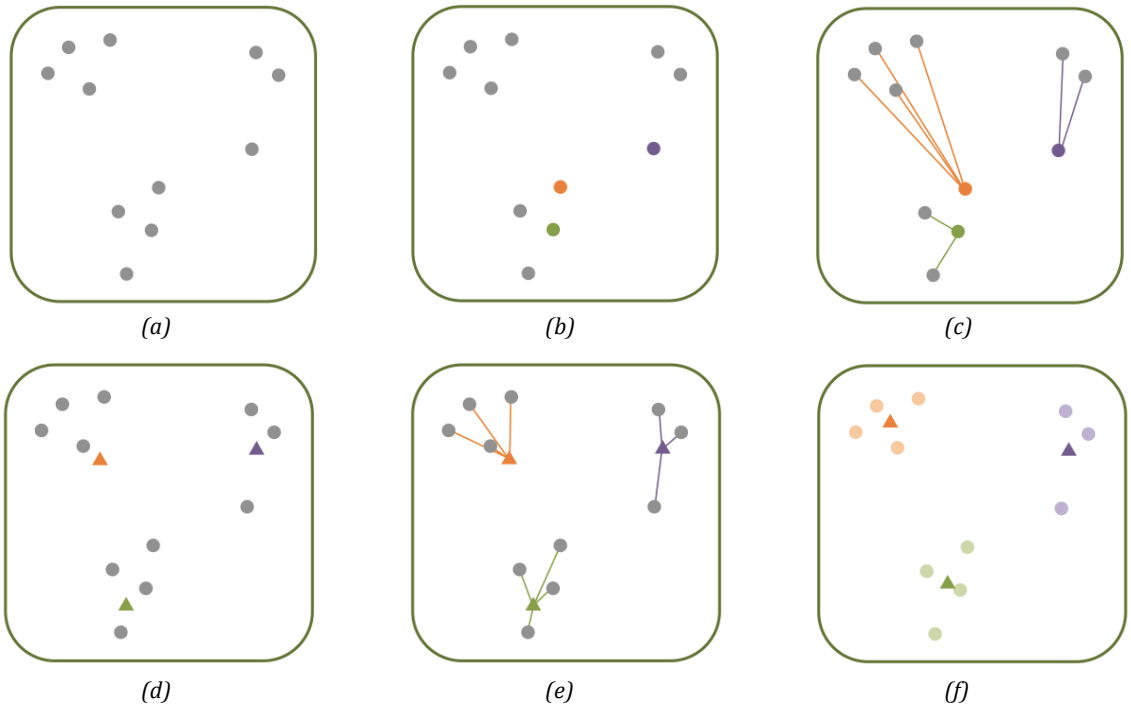


Figura 2-4: *Algoritmo K-medias. En (a) se dispone de un conjunto de once datos para distribuir en tres grupos (K=3). En primer lugar se escoge tres datos al azar (b) como centroides iniciales de los grupos, seguidamente se calcula para cada dato el centroide más cercano formándose las primeras agrupaciones (c). Para cada agrupación se calculan los nuevos centroides (triángulos en d). El proceso de formación de los grupos (e) y cálculo de nuevos centroides (f) se repite hasta que no hay cambios en ningún grupo. En el ejemplo, el algoritmo termina en la segunda iteración.*

El procedimiento básico de *K-medias* es iterativo y puede resumirse del modo siguiente (ver figura 2-4):

1. Dado el conjunto de datos (figura 2-4(a)), escoger al azar o en base a un conocimiento previo K datos para ser los K centroides iniciales $\boldsymbol{m}_j$ $j \in \{1,2,\ldots,K\}$ (figura 2-4(b)).

2. Formar la partición $C = \{C_1, C_2, \ldots, C_K\}$ asignando cada dato a la agrupación más cercana C_l (figuras 2-4(c) y 2-4(e)), es decir,

$$\boldsymbol{X}_i \in C_l \text{ si } \|\boldsymbol{X}_i - \boldsymbol{m}_l\| < \left\|\boldsymbol{X}_i - \boldsymbol{m}_j\right\|$$
$$\forall i \in \{1,2,\ldots,n\} \text{ y } \forall j \in \{1,2,\ldots,K\} \text{ con } j \neq l \qquad (2\text{-}9)$$

3. Calcular para cada grupo el nuevo centroide mediante la media muestral de sus elementos (figuras 2-4(d) y 2-4(f)):

$$\boldsymbol{m}_j = \frac{1}{n_j}\sum_{i=1}^{n} \boldsymbol{X}_i \quad \text{donde} \quad \boldsymbol{X}_i \in C_j \quad \text{y} \quad n_j = \left|C_j\right|, \quad j \in \{1,2,\ldots,K\} \qquad (2\text{-}10)$$

4. Repetir los pasos 2 y 3 hasta que no haya ningún cambio en ningún grupo. En la figura 2-4 el proceso termina en la segunda iteración (figuras 2-4(e) y 2-4(f)).

Los principales inconvenientes de *K-medias* son (Xu & Wunsch, 2009):

- Dificultad para reconocer grupos que no tienen forma hiperesférica. Este problema es abordado en el desarrollo de la aplicación utilizando la *distancia de Mahalanobis* además de la *distancia Euclídea*.
- No garantiza la convergencia a una solución óptima global.
- Asume que el número inicial de grupos K es previamente conocido.
- Puede verse afectado negativamente por valores atípicos.
- Funciona de manera eficiente sólo con atributos numéricos.

No obstante, *K-medias* es considerado como un elemento básico de los métodos de agrupamiento, debido a su facilidad de implementación. Funciona bien para muchos problemas prácticos, en particular cuando las agrupaciones resultantes son compactas y tienen forma hiperesférica. El tiempo de ejecución de *K-medias* es $O(nKDT)$, donde T es el número de iteraciones y como K, D y T son generalmente mucho menores de n, el tiempo de ejecución de *K-medias* es aproximadamente lineal. Por lo tanto, *K-medias* es una buena elección para

agrupar conjuntos de datos a gran escala. Ejemplos de aplicaciones de *K-medias* los hay en imagen médica (de Oliveira Martins, Junior, Silva, de Paiva, & Gattass, 2009), segmentación (Dubey, Dixit, Singh, & Gupta, 2013; Jumb, Sohani, & Shrivas, 2014), servicios en *Internet* (Vijayan & Balasundaram, 2013), etc…

2.3.2. Algoritmo *Fuzzy C-means (FCM)*

Fuzzy C-means (*FCM*) es un método de agrupamiento de datos de partición difusa, por tanto su objetivo es clasificar los datos en algún número de grupos previamente especificado, permitiendo a cada dato pertenecer a todos los grupos con un cierto grado de pertenencia. El grado de pertenencia de cada dato $\boldsymbol{X}_j$ ($j \in \{1,2,\ldots,n\}$) al grupo $G_i (i \in \{1,2,\ldots,c\})$ se representa mediante u_{ij} y satisface las siguientes restricciones:

$$0 \leq u_{ij} \leq 1 \tag{2-11}$$

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j \tag{2-12}$$

$$0 < \sum_{j=1}^{n} u_{ij} < n, \forall i \tag{2-13}$$

El valor de u_{ij} será mayor cuanto más próximo se encuentre $\boldsymbol{X}_j$ de $\boldsymbol{m}_i$, centro del grupo G_i. El objetivo de *FCM* es minimizar la siguiente función de coste (Xu & Wunsch, 2009):

$$J(\mathbf{U}, \mathbf{M}) = \sum_{i=1}^{c} \sum_{j=1}^{n} \left(u_{ij}\right)^m {D_{ij}}^2 \tag{2-14}$$

donde:

- $\mathbf{U} = [u_{ij}]_{c \times n}$ es la matriz de partición difusa.
- $\mathbf{M} = [\boldsymbol{m}_1, \ldots, \boldsymbol{m}_c]$ es la matriz de centros de los grupos.

- $m \in [1, +\infty)$ es el parámetro que controla cómo de difusos son los grupos. Valores más grandes este parámetro favorece agrupamientos más difusos. En la aplicación este parámetro es solicitado al usuario aunque se recomienda un valor de $m = 2$.
- D_{ij} es la distancia entre $\boldsymbol{X}_j$ y $\boldsymbol{m}_i$.

La función de coste (2-14) se minimiza mediante el algoritmo *FCM* cuyos pasos son (Xu & Wunsch, 2009) :

1. Fijar el valor de c ($2 \leq c < n$), seleccionar un valor para el parámetro m, inicializar la matriz de centros **M** aleatoriamente y tomar un valor pequeño de $\varepsilon > 0$. Comenzar el proceso iterativo tomando un valor inicial $t = 0$.
2. Actualizar la matriz **U**:

$$u_{ij}^{(t+1)} = \begin{cases} 1/\left(\sum_{l=1}^{c} \left(D_{lj}/D_{ij}\right)^{2/(1-m)}\right), & \text{si } \boldsymbol{I}_j = \emptyset \\ 1/\left|\boldsymbol{I}_j\right|, & \text{si } \boldsymbol{I}_j \neq \emptyset, i \in \boldsymbol{I}_j, i \in \{1, \ldots, c\}, j \in \{1, \ldots, n\} \\ 0, & \text{si } \boldsymbol{I}_j \neq \emptyset, i \notin \boldsymbol{I}_j \end{cases} \quad (2\text{-}15)$$

   donde $\boldsymbol{I}_j = \{i / i \in [1, c], \boldsymbol{X}_j = \boldsymbol{m}_i\}$.

3. Actualizar la matriz **M**:

$$\boldsymbol{m}_i^{(t+1)} = \left(\sum_{j=1}^{n} \left(u_{ij}^{(t+1)}\right)^m \boldsymbol{X}_j\right) / \left(\sum_{j=1}^{n} \left(u_{ij}^{(t+1)}\right)^m\right), i \in \{1, \ldots., c\} \quad (2\text{-}16)$$

4. Repetir los pasos 2 y 3 hasta que $\left\|\mathbf{M}^{(t+1)} - \mathbf{M}^{(t)}\right\| < \varepsilon$.

Los principales problemas de los que *FCM* adolece son el desconocimiento previo del número de grupos, la falta de información para identificar una partición inicial apropiada y la presencia de ruido y valores atípicos.

No obstante, *FCM* es un algoritmo de agrupamiento que se aplica a una amplia gama de problemas relacionados con Ingeniería (Sudha, Raju, & Sekhar, 2012), medio ambiente (Huang, et al., 2013), Geología (Hussain, 2012), análisis de imágenes y diagnóstico médico (Oke, Adedeji, Alade, & Adewus, 2012; Rakesh & Ravi, 2012), etc …

2.4. Validación

Como se ha comentado en el apartado anterior, las técnicas *K-medias* y *FCM* dependen del número de grupos de elección, siendo la elección de un valor apropiado una tarea difícil en situaciones sin supervisión. Por tanto, son necesarios unos criterios de evaluación eficaces que proporcionen al usuario información acerca de la fiabilidad de los resultados obtenidos en un determinado agrupamiento. Para abordar este problema, las medidas de validación tratan de determinar con qué precisión las agrupaciones obtenidas pueden reflejar una estructura particular de los datos.

Debido a que los algoritmos implementados en la aplicación desarrollada en esta tesis se aplican con el fin de interpretar y comprender datos sobre los que se sabe poco (aprendizaje no supervisado), la validación se convierte en un último paso de facto en el proceso de agrupación. Por lo tanto, la validación es crucial para una división correcta de los datos cuando se desconoce el número de conglomerados.

Los dos criterios generales casi universales utilizados por todas las medidas de validación para evaluar el agrupamiento obtenido son la compacidad y la separación. Un buen agrupamiento deberá ser capaz de crear grupos con datos que son similares entre sí (compacidad), pero disimilares respecto de datos pertenecientes a otros grupos (separación). Casi la totalidad de las medidas de validación tratan de medir la compacidad, la separación y/o cómo ambas se relacionan entre sí. Si bien los métodos para determinar la separación o compacidad a veces pueden variar, hay que señalar que la mayoría utiliza los centros de los grupos obtenidos como la base para medir la compacidad, y la separación, con lo que grupos con estructura hiperesférica obtienen mejores resultados en la validación.

Existen numerosas medidas de validación y, a menudo, diferentes medidas suelen producir resultados dispares. En la aplicación desarrollada en esta tesis se incluyen tres medidas para la validación de los agrupamientos, éstas son *Dunn*, *Silhouette* (*Silh*) y *Suma de Cuadrados* o *SS* (del inglés *Sum of Squares*). Las razones

de tal elección se fundamentan en base a criterios de popularidad, eficiencia y simplicidad en la implementación. La descripción de estas medidas es la siguiente (Baarsch & Celebi, 2012):

- ***Dunn***: Es una de las medidas más frecuentemente citada. Para su cálculo se obtiene la raíz cuadrada de la distancia mínima entre dos grupos (obtenida como la mínima distancia entre todos los pares de datos de diferente grupo) como medida de separación y se divide por la raíz cuadrada de la distancia máxima entre dos datos del mismo grupo como medida de compacidad. Su expresión analítica es:

$$Dunn = \frac{\sqrt{Mínima\ distancia\ entre\ grupos}}{\sqrt{Máxima\ distancia\ en\ un\ mismo\ grupo}} \qquad (2\text{-}17)$$

  Dado que un agrupamiento se considera mejor cuanto mayor sea la distancia entre grupos (mayor separación entre grupos) y menor la que existe entre datos del mismo grupo (mayor compacidad), cuanto mayor sea el valor de la medida de *Dunn*, mejor agrupamiento será el obtenido. Debido a que la medida sólo utiliza los valores máximos y mínimos en su cálculo, el método de *Dunn* es altamente susceptible a la influencia de ruido, valores atípicos, o situaciones en las que dos grupos en concreto puedan estar próximos.

- ***Silhouette* (*Silh*)**: Al igual que en el anterior, este método también contempla la compacidad frente a la separación. En primer lugar, para cada punto se calcula la diferencia entre la distancia media desde ese punto al resto de puntos de su grupo y el mínimo de la distancia media entre ese punto y los puntos de cada uno del resto de grupos. Esta diferencia se divide por un término de normalización que es el mayor de los dos promedios:

$$S_{\boldsymbol{x}_i} = \frac{b_{q,i} - a_{p,i}}{\max\{a_{p,i}, b_{p,i}\}} \qquad (2\text{-}18)$$

  Es decir, si $\boldsymbol{x}_i$ es un dato perteneciente al grupo C_p entonces se define $b_{q,i} = \min d_{q,i}$, donde $d_{q,i}$ es la distancia media entre el punto $\boldsymbol{X}_i$ y los

puntos pertenecientes del grupo C_q. Por otro lado, $a_{p,i}$ es la distancia media entre el punto $\boldsymbol{X}_i$ y cada punto del grupo C_p que no sea el propio $\boldsymbol{X}_i$. En segundo lugar, *Silhouette* se calcula hallando el promedio $S_{\boldsymbol{x}_i}$ para todos los datos del conjunto:

$$Silh = \frac{1}{n}\sum_{i=1}^{n} S_{\boldsymbol{x}_i} \tag{2-19}$$

A diferencia de *Dunn* la medida *Silhouette* relaciona la separación frente a la compacidad mediante sustracción en lugar de división. Conforme la puntuación es más cercana a 1, mejor es el agrupamiento obtenido.

- ***Suma de Cuadrados* (*SS*)**: El método de la *Suma de Cuadrados* (*SS*, del inglés *Sum of Squares*) se basa en una relación entre la traza de la matriz de dispersión entre grupos o *BCSM* (del inglés, *Between Cluster Scatter Matrix*) y la traza de la matriz de dispersión dentro de cada grupo o *WCSM* (del inglés, *Within Cluster Scatter Matrix*). Su expresión es:

$$SS = \frac{traza(WCSM)}{traza(BCSM)} * k \tag{2-20}$$

La $traza(BCSM)$ es simplemente la suma de los cuadrados de las distancias entre el centro de cada grupo y el centro global del conjunto de datos, ponderado por el tamaño del clúster y la $traza$ $(WCSM)$ es la suma de los cuadrados de las distancias entre el centro de cada grupo y cada punto en el clúster. El parámetro *k* se corresponde con el número de grupos y se utiliza como normalizador del cociente. Puesto que *SS* divide compacidad entre separación, una puntuación inferior indica mejores agrupamientos.

Según el análisis entre diversas medidas de validación realizado por (Baarsch & Celebi, 2012), el método *Suma de Cuadrados* y *Silhouette* son los que obtuvieron mejores resultados.

Capítulo 3

Técnicas de reducción de la dimensionalidad

Resumen

Este capítulo se centra en otro de los pilares fundamentales de esta tesis: la reducción de la dimensionalidad. En la introducción se exponen los conceptos básicos de este campo dejando un apartado para la descripción de las variedades o manifolds, término básico acerca de la estructura intrínseca de los datos imprescindible para la comprensión de las estrategias que se hallan tras las técnicas de reducción de la dimensionalidad.

Más adelante se incluye la notación matemática que se va a utilizar y la descripción de las técnicas de reducción de la dimensionalidad que serán utilizadas en el desarrollo de la aplicación en Processing. Finalmente, se detalla el modo en que se realizará la evaluación de la calidad en la reducción de la dimensionalidad.

3.1. Introducción

Como ya fue expuesto en el primer capítulo, el mundo es en esencia multidimensional y el número de características (también llamadas variables,

atributos o parámetros) que describen un determinado fenómeno puede llegar a ser enorme. Más características implican más información y, potencialmente, una mayor precisión. Paradójicamente cuantas más características se tienen, más costosa y difícil es la extracción de información.

Sin embargo, las variables que pueden caracterizar un conjunto de datos no todas suelen ser independientes unas de otras y no todas suelen ser importantes o significativas. En consecuencia, la gestión eficiente y la extracción del conocimiento subyacente requiere tomar la redundancia en cuenta. Esa gran cantidad de variables debe resumirse en un conjunto más pequeño sin, o con menos, redundancia. Esta es la meta de la reducción de la dimensionalidad, RD o *DR* (del inglés, *Dimensionality Reduction*), que es una de las herramientas clave para el análisis de datos de alta dimensión (Lee & Verleysen, 2007).

Desde un punto de vista práctico, la detección de las características que son relevantes y cómo interactúan unas con otras es esencial a la hora de descubrir y extraer información del conjunto de datos.

Desde un punto de vista teórico, cuando la dimensionalidad de datos crece, las propiedades buenas y conocidas de los espacios de dos o tres dimensiones euclídeas usuales dan paso a fenómenos extraños y molestos. Todas las dificultades que se presentan cuando se trata con datos de alta dimensión a menudo se denominan la "maldición de la dimensionalidad" (Köppen, 2000).

Existen dos enfoques principales de trabajo utilizados en la reducción de la dimensionalidad (Masaeli, Fung, & Dy, 2010):

- **Selección de características:** También conocida como selección de variables, consiste en la selección de un subconjunto de las características originales no permitiéndose que éstas sean transformadas o modificadas. Al analizar los datos, no necesariamente todas las características están relacionadas con la información subyacente que el usuario desea capturar. En este caso, las variables irrelevantes pueden ser eliminadas del conjunto de datos. La selección de características típicamente suele utilizarse cuando se desea mantener su significado original y se persigue determinar cuáles son importantes. Una vez seleccionadas las características, sólo éstas deben ser calculadas o recopiladas. Más información sobre algoritmos de selección de características puede encontrarse en (Guyon & Elisseeff, 2003; Yu & Liu, 2004).
- **Transformación de características:** Incluso cuando se supone que

> todas las variables son relevantes, la dimensionalidad de los datos observados todavía puede ser más grande de lo necesario. En ese caso, en lugar de retirar de manera arbitraria una variable, se crea un nuevo conjunto de variables más reducido que el original mediante la aplicación de una transformación, o proyección, desde el espacio multidimensional original a un espacio de menos dimensiones. Al contrario que en la selección, en la transformación el espacio de características originales pasa a ser un reducido, aunque nuevo espacio informativo donde queda modificada la representación original de las variables. En este caso todas las características de entrada son necesarias para obtener la dimensión reducida. El nuevo conjunto, obviamente, debe contener un número menor de variables, pero también debe preservar las características interesantes del conjunto inicial. En otras palabras, se busca una transformación de las variables con algunas propiedades bien definidas. Estas propiedades deben asegurarse de que la transformación no altera el contenido de la información transmitida por el conjunto de datos inicial, sino sólo representarlo en una forma diferente.

Con el fin de generar visualizaciones intuitivas la Representación Visual de Datos necesita que el número de dimensiones se reduzca generalmente a dos o tres. Por tanto, las estrategias interesantes para el desarrollo de esta tesis son aquellas que se centran en la transformación de características.

3.2. *Manifolds*

Desde un punto de vista geométrico, cuando dos o más variables dependen unas de otras, su distribución conjunta no abarca todo el espacio. En realidad, la dependencia induce alguna estructura en la distribución, en forma de un lugar geométrico que puede ser visto como un tipo de objeto en el espacio. Como se mencionó anteriormente la reducción de dimensionalidad tiene por objetivo dar una nueva representación de estos objetos preservando su estructura.

En Matemáticas, la Topología estudia las propiedades de los objetos que se preservan a través de deformaciones, torsiones y estiramientos, garantizando de este modo que la estructura intrínseca de los objetos no se altera. Por ejemplo, un círculo es topológicamente equivalente a una elipse, y una esfera es equivalente a un elipsoide.

Una de las ideas centrales de la Topología es que los objetos espaciales

puedan ser tratados como objetos en sí mismos independientemente de cómo se representen o sean dispuestos en el espacio. En otras palabras, la Topología se utiliza para abstraer la conectividad intrínseca de los objetos sin tener en cuenta su forma detallada. Si dos objetos tienen las mismas propiedades topológicas, se dice que son homeomorfos (Edelsbrunner, 2014).

Los objetos con los que se trabaja en Topología se definen formalmente como espacios topológicos. Un espacio topológico es un conjunto para el cual se especifica una topología. Para un conjunto Y, una topología T se define como una colección de subconjuntos de Y con las siguientes propiedades (Lee J. , 2010):

- Trivialmente, $\emptyset \in T$ e $Y \in T$.
- Siempre que dos conjuntos están en T, entonces también lo está su intersección.
- Cuando dos o más conjuntos están en T, entonces también lo está su unión.

Esta definición de topología se mantiene para espacios cartesianos ($\mathbb{R}^D$). Por ejemplo, la topología natural asociada a $\mathbb{R}$, el conjunto de los números reales, es la unión de todos sus intervalos abiertos.

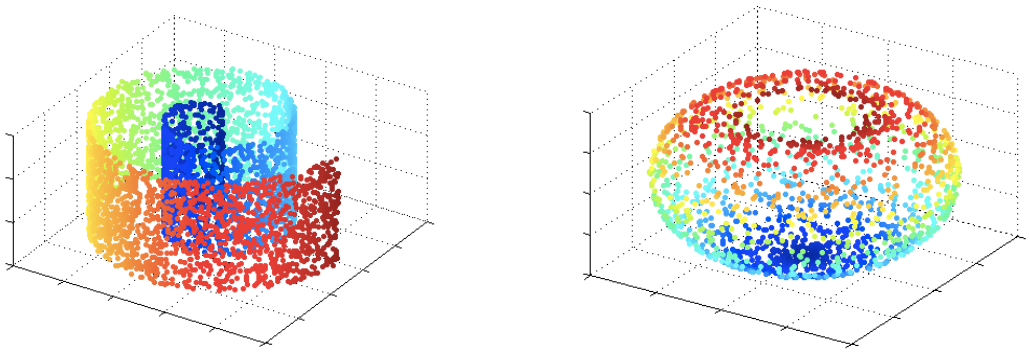
Desde un punto de vista más geométrico un espacio topológico también se puede definir mediante entornos. El entorno de un punto $\mathbf{x} \in \mathbb{R}^D$ es, a menudo, definido como una bola abierta de radio $\varepsilon > 0$ ó $B_\varepsilon(\mathbf{x})$ y es el conjunto de puntos dentro de una esfera de dimensión D de radio $\varepsilon > 0$ centrada en $\mathbf{x}$. Un conjunto que contiene un entorno abierto también se llama entorno. Entonces, un espacio topológico es tal que:

- A cada punto $\boldsymbol{x}$ le corresponde, por lo menos, un entorno $U(\boldsymbol{x})$ de tal modo que $\boldsymbol{x} \in U(\boldsymbol{x})$.
- Si $U(\boldsymbol{x})$ y $V(\boldsymbol{x})$ son entornos del mismo punto, entonces existe un entorno $W(\boldsymbol{x})$ tal que $W(\boldsymbol{x}) \subset U(\boldsymbol{x}) \cup V(\boldsymbol{x})$.
- Si $\mathbf{z} \in U(\boldsymbol{x})$, entonces existe un entorno $V(\boldsymbol{z})$ de $\mathbf{z}$ tal que $V(\boldsymbol{z}) \subset U(\boldsymbol{x})$.
- Para dos puntos distintos, existen dos entornos disjuntos de estos puntos.

En este marco, una variedad topológica o *manifold* $\mathcal{M}$ es un espacio topológico que es localmente euclídeo, lo que significa que alrededor de cada punto de $\mathcal{M}$ existe un entorno que es topológicamente equivalente a una bola unidad abierta en $\mathbb{R}^D$ (existe un homeomorfismo entre ambos) (Loring, 2010). En general, cualquier objeto que es casi "plano" a pequeñas escala es un *manifold*. Por ejemplo, la Tierra es esférica, pero se ve plana en la escala humana.

Una inmersión es una representación de un objeto topológico (un *manifold*, un gráfico, etc.) en un determinado espacio, por lo general $\mathbb{R}^D$ para alguna D, de tal manera que sus propiedades topológicas se conservan. Un espacio X está incrustado en otro espacio Y cuando las propiedades de Y restringidas a X son las mismas que las propiedades de X. Así, un *manifold* $\mathcal{M}$ que es un subconjunto de $\mathbb{R}^D$, puede tener una dimensión desde un punto de vista geométrico y su dimensión puede ser menor que D. A esta dimensión se le denomina dimensión intrínseca del *manifold* (Lee & Verleysen, 2007).

Intuitivamente, un *manifold* es una generalización de las curvas y las superficies de $\mathbb{R}^3$ en espacios de mayor dimensión como $\mathbb{R}^D$. En cada punto es localmente euclídeo y tiene un entorno homeomorfo a una bola abierta de $\mathbb{R}^D$. De este modo, las coordenadas en estos entornos permiten llevar a cabo cálculos como en un espacio euclídeo, por lo que muchos conceptos de $\mathbb{R}^D$, como diferenciabilidad, derivadas en un punto, espacios tangentes y formas diferenciales se trasladan al *manifold* (Loring, 2010).



(a) Swiss roll *(b) fish bowl*

Figura 3-1: *Ejemplos de manifolds artificiales.*

Como ya se ha comentado al principio de este apartado, el objetivo de la reducción de la dimensionalidad es dar una nueva representación a estas estructuras denominadas *manifolds* en espacios de menor dimensión preservando

su estructura. En la práctica, sin embargo, un *manifold* no es más que el soporte subyacente a una distribución de datos que se conoce sólo a través de una muestra de datos finita. Este es un problema importante ya que las técnicas de reducción de la dimensionalidad deben trabajar con esos datos parciales y limitados. También se ha de tener en cuenta que el modelo de *manifold* no contempla el ruido generado por datos anómalos. En este caso, los puntos correspondientes a esos datos ya no se encuentran en el mismo *manifold* sino solamente cerca.

Con el objeto de ilustrar las ventajas e inconvenientes de los distintos métodos de reducción de la dimensionalidad, existen en la literatura diferentes *manifolds* artificiales que son repetidamente utilizados. En la figura 3-1 se muestran dos de los más comunes: *Swiss roll* y *fish bowl*.

3.2.1. Distancias geodésica y de grafo

La figura 3-2 muestra un *manifold* (línea gris) cuya dimensión intrínseca es igual a uno pero que se encuentra definido en un espacio de dos dimensiones. Intuitivamente, para preservar su estructura debe esperarse que su reducción a una dimensión desenrolle la estructura disponiéndola recta. Con este planteamiento la distancia euclídea en el espacio de dos dimensiones no describe fielmente las distancias entre puntos del *manifold*, excepto para distancias pequeñas en las que puede considerársele casi lineal. Si se presta atención a la distancia euclídea entre los puntos A y B (línea roja), ésta es corta en el espacio de dos dimensiones ya que el *manifold* se pliega sobre sí mismo. Sin embargo, si el *manifold* se desenrolla para su reducción a una dimensión, la nueva distancia medida entre A y B es mucho mayor (línea azul). A esta distancia medida sobre el *manifold* se le denomina distancia geodésica por su analogía con las curvas trazadas en la superficie de la Tierra.

En contraste con la distancia euclídea, la distancia geodésica no depende tanto como aquélla de cómo el *manifold* se encuentre ubicado en el espacio. En el ejemplo de la figura 3-2, la distancia geodésica (línea azul) es la misma en el espacio de una y dos dimensiones.

Formalmente, la distancia geodésica es bastante complicada de calcular a partir de la expresión analítica de un *manifold* y más todavía cuando solamente se dispone de algunos puntos que incluso pueden no estar exactamente en él. Es por ello que la distancia geodésica ha de estimarse mediante la longitud de un camino que pase a través de cierto número de puntos del *manifold*. En este caso, la solución más sencilla para el cálculo de cada tramo del camino es la distancia

euclidiana. Este camino ha de seguir el *manifold* subyacente, al menos, aproximadamente. Por tanto, sus tramos no pueden unir directamente un punto del conjunto con cualquier otro.

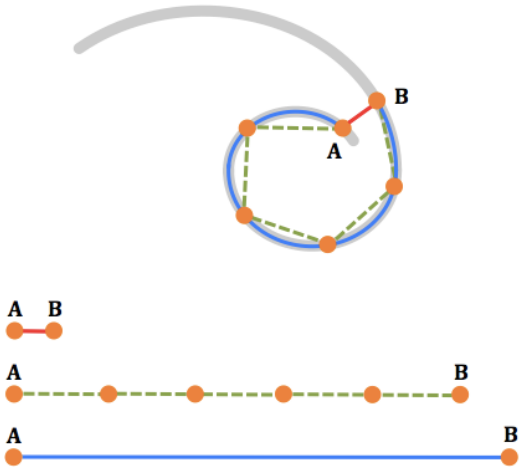


Figura 3-2: *Distintas distancias entre los puntos A y B pertenecientes a un manifold (línea gris). La línea roja representa la distancia euclídea en el espacio de dos dimensiones, la línea azul es la distancia geodésica medida a lo largo del manifold y la línea verde simboliza la distancia de grafo.*

Con el fin de obtener una buena aproximación a la distancia geodésica, es necesaria una fina discretización del *manifold*, permitiéndose entonces sólo los tramos más cortos para construir los caminos. En la práctica, existen varias reglas simples que pueden lograr este objetivo. Dos ejemplos son la regla K, que solamente permite tramos desde un punto a los K puntos más cercanos y la regla ε, que sólo permite tramos desde un punto a todos los otros que estén situados dentro de una bola con un radio ε predeterminado (Bernstein, de Silva, Langford, & Tenenbaum, 2000).

Matemáticamente, el conjunto de puntos de datos junto con el conjunto de tramos permitidos constituye un grafo $G = (V, E)$ donde el conjunto de vértices V son los puntos del *manifold* y el conjunto de aristas E son los tramos permitidos. De este modo, el camino que une un punto A y uno B es un subconjunto ordenado de V donde el primer vértice es A y el último B de tal manera que los tramos que unen a cada dos vértices consecutivos del camino son elementos de E. De esta manera, la longitud de un camino se define como la suma de las longitudes de los tramos que lo forman.

En esta situación, faltaría calcular los caminos más cortos entre los distintos puntos con un algoritmo como los de *Dijkstra* (Cormen, Leiserson, Rivest, & Stein, 2009), *Bellman Ford* (Mawale & Gandole, 2011) o *Floyd-Warshall* (Hougardy,

2010). Una vez obtenidos, la longitud del camino que une dos puntos cualesquiera del *manifold* se define como su distancia de grafo. En la figura 3-2 esta distancia se representa mediante la línea verde discontinua. Una demostración de que la distancia de grafo se aproxima a la distancia geodésica se proporciona en (Bernstein, de Silva, Langford, & Tenenbaum, 2000).

Los conceptos de distancia geodésica y de grafo son de gran importancia en el ámbito de la reducción de la dimensionalidad donde son utilizados principalmente en técnicas cuyo objetivo es preservar la estructura global del *manifold* como *Isomap* (ver sección 3.4.7.).

3.3. Notación

El problema de la reducción de la dimensionalidad comienza admitiendo que se dispone de un conjunto de n datos. Cada dato se representa por un vector $\boldsymbol{X}_i$ ($i \in \{1,2,\ldots,n\}$) en un espacio de dimensión D, donde cada componente es el valor que toma el dato respecto de una determinada variable (también denominada característica o atributo). El conjunto de datos se dispone por filas para formar la matriz $\boldsymbol{X}$ de dimensiones $n \times D$ ($\boldsymbol{X}_{n\times D}$). De igual modo, se supone que los datos están situados en o cerca de un *manifold* con una dimensionalidad intrínseca menor que la dimensión del espacio donde se encuentra sumergido, es decir, D. En este contexto, la reducción de la dimensionalidad consiste en transformar ese conjunto de datos en un nuevo conjunto de datos $\boldsymbol{Y}$ con dimensionalidad d, donde $d < D$, de tal modo que se conserve la estructura de los datos originales lo más posible. En general, lo que se persigue es que d sea la dimensión intrínseca del *manifold* aunque en aplicaciones orientadas a la visualización de datos d no puede ser superior a 3. Al conjunto $\boldsymbol{Y}$ se le denomina también inmersión o proyección del conjunto $\boldsymbol{X}$.

Una vez definido el problema, la notación que se utilizará en adelante es la siguiente:

- n : Número de datos.
- D : Dimensión del espacio original.
- d : Dimensión del espacio de inmersión.
- $\boldsymbol{X}$: Matriz de dimensiones $n \times D$ ($\boldsymbol{X}_{n\times D}$) donde se disponen por filas los diferentes datos del espacio de alta dimensión.

- $\boldsymbol{X}_i$: Punto que representa al dato i $(i \in \{1,2,\ldots.,n\})$ en el espacio de alta dimensión también denominado espacio original. Es la *i*-ésima fila de la matriz $\boldsymbol{X}$ y su dimensión es *D*.
- $\boldsymbol{X}^j$: Valores que toma la variable j $(j \in \{1,2,\ldots.,D\})$ del espacio de alta dimensión. Es la *j*-ésima columna de la matriz $\boldsymbol{X}$ y su dimensión es *n*.
- $\boldsymbol{Y}$: Matriz de dimensiones $n \times d$ $(\boldsymbol{Y}_{n\times d})$ donde se disponen por filas los diferentes datos del espacio de baja dimensión.
- $\boldsymbol{Y}_i$: Punto que simboliza el dato i $(i \in \{1,2,\ldots.,n\})$ en el espacio de baja dimensión (homólogo de $\boldsymbol{X}_i$). Es la *i*-ésima fila de la matriz $\boldsymbol{Y}$ y su dimensión es *d*.
- $\boldsymbol{Y}^j$: Valores que adquiere la variable j $(j \in \{1,2,\ldots.,D\})$ del espacio de baja dimensión. Es la *j*-ésima columna de la matriz $\boldsymbol{Y}$ y su dimensión es *n*.

3.4. Técnicas de reducción de la dimensionalidad utilizadas

Existen numerosas técnicas para llevar a cabo la reducción de la dimensionalidad que suelen clasificarse atendiendo a diversos criterios. Una primera clasificación es separarlas entre aquellas que son lineales y no lineales. Las técnicas lineales se caracterizan por asumir que los datos se encuentran en o cerca de un *manifold* lineal inmerso en un espacio de dimensión mayor que la suya intrínseca. Del otro lado, las técnicas no lineales no adoptan la hipótesis de linealidad, por lo cual pueden identificar *manifolds* más complejos.

A su vez, las técnicas no lineales se clasifican en dos grandes categorías: globales y locales (de Silva & Tenenbaum, 2002). Las técnicas globales intentan preservar la geometría en todas las escalas, representando puntos cercanos en el *manifold* como puntos cercanos en el espacio de baja dimensión y puntos lejanos como puntos lejanos. Las técnicas locales tratan de preservar la geometría local de los datos y su principal objetivo es representar puntos cercanos en el *manifold* como puntos cercanos en la representación de baja dimensión.

La principal ventaja del enfoque global es que tiende a dar una representación más fiel de la estructura general de los datos. Los enfoques locales tienen dos ventajas principales: por un lado, la eficiencia computacional, ya que implican cálculos matriciales más sencillos al preocuparse solamente de las relaciones entre puntos cercanos y, por otro, su capacidad de representación, ya

que pueden obtener resultados útiles en una gama más amplia de *manifolds* cuya geometría local puede ser similar a la euclídea pero cuya geometría global puede no serlo.

De entre todas las técnicas de reducción de la dimensionalidad, se han seleccionado once para incluirlas en la aplicación desarrollada en esta tesis atendiendo a razones de popularidad, diversidad de criterio o su probada eficacia en diferentes contextos de aplicación tal y como se argumenta para cada una en los siguientes subapartados. La figura 3-3 muestra estas técnicas clasificadas.

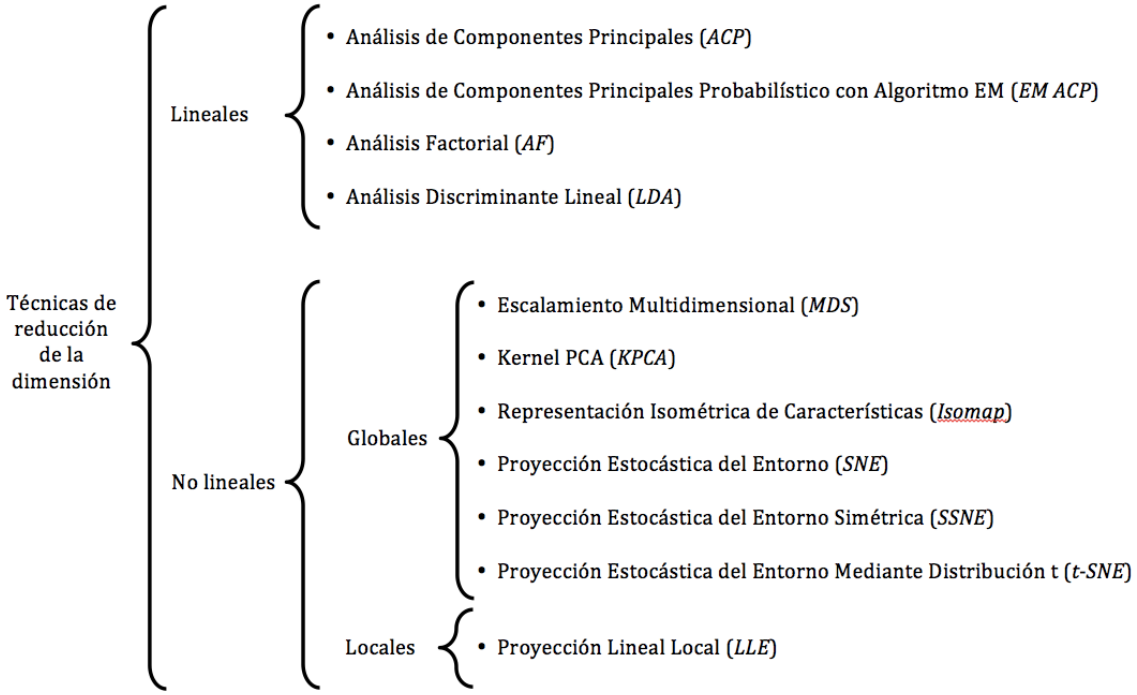


Figura 3-3: *Clasificación de las distintas técnicas de reducción de la dimensionalidad utilizadas.*

3.4.1. Análisis de Componentes Principales (ACP)

El Análisis de Componentes Principales, ACP o *PCA* (del inglés, *Principal Component Analysis*) es la técnica lineal más popular. La idea central es reducir la dimensionalidad del conjunto de datos, que consta de un gran número de variables interrelacionadas, conservando tanto como sea posible la variación presente en dicho conjunto (Jolliffe, 2002). Esto se logra mediante la transformación respecto de un nuevo conjunto de variables, las componentes principales, que son no correlacionadas, y que están ordenadas de tal forma que unas pocas de las primeras conservan la mayor parte de la variación presente en todas las variables

originales. Con este fin se busca una base lineal de dimensionalidad reducida para representar los datos en la que la cantidad de varianza en éstos sea máxima (ver figura 3-4).

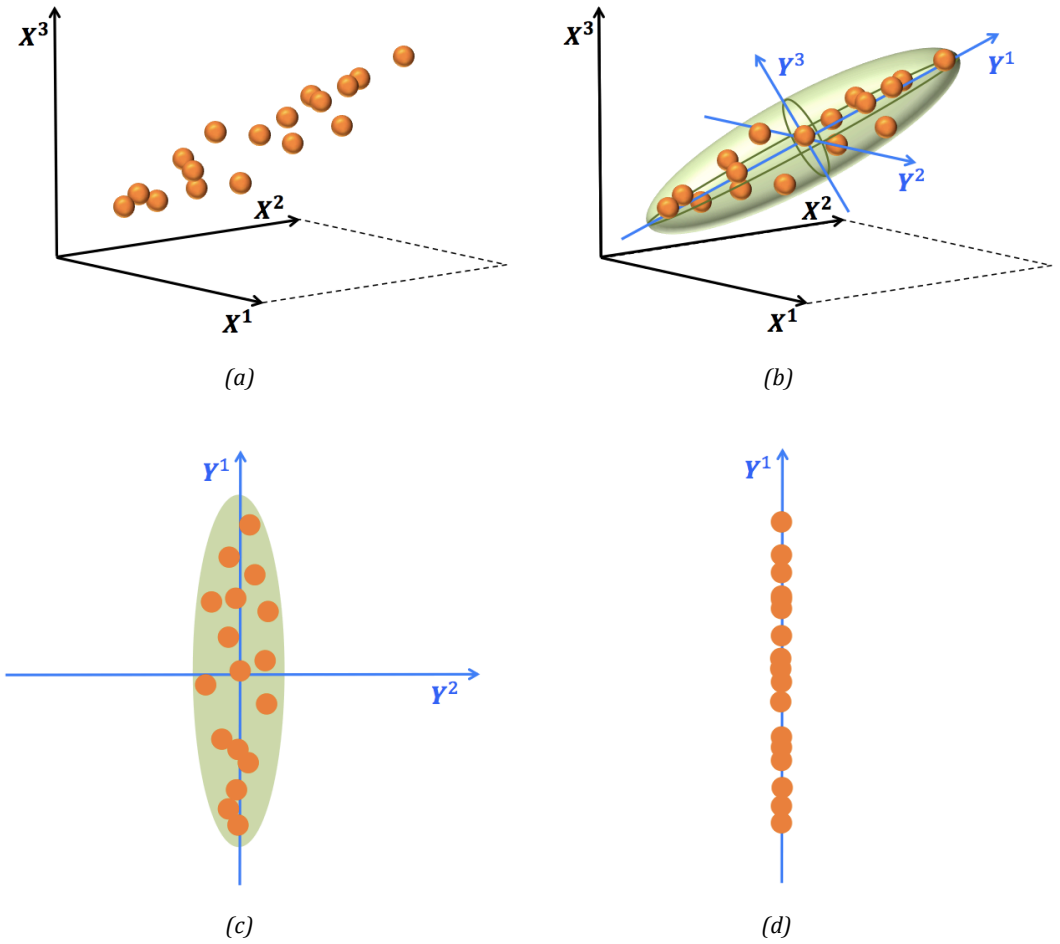


Figura 3-4: *Análisis de Componentes Principales. En (a) se halla un conjunto de datos tridimensional representado respecto de sus tres variables* $\boldsymbol{X^j}$*. En (b) ese mismo conjunto aparece representado respecto de las variables incorreladas* $\boldsymbol{Y^j}$ *obtenidas mediante transformación lineal a partir de las anteriores en las direcciones de máxima varianza, son las componentes principales. En (c) el conjunto de datos es expresado en dos dimensiones respecto de las dos primeras componentes principales que son las que más varianza representan y en (d) respecto de una única dimensión, la de máxima varianza.*

En términos matemáticos, el modelo de ACP supone que existen D variables observadas en un conjunto de n elementos dispuestas en un vector aleatorio $\boldsymbol{x}$ que resultan de una transformación lineal según la matriz $\mathbf{W}$ respecto de d variables latentes y desconocidas reunidas en el vector $\boldsymbol{y}$. Los valores que toma la variable i-ésima de $\boldsymbol{x}$ en el conjunto de datos se recogen en la columna $\boldsymbol{X}^j$ de la matriz $\boldsymbol{X}$. De este modo:

$$\boldsymbol{x} = \mathbf{W}\,\boldsymbol{y} \tag{3-1}$$

Se asume que todas las variables latentes siguen una distribución normal o de *Gauss* y que las columnas de $\mathbf{W}$ forman un sistema ortonormal. De este modo, $\mathbf{W}$ es una matriz de dimensiones $D \times d$ tal que $\mathbf{W}^T\mathbf{W} = \mathbf{I}_d$. De igual modo, las variables latentes son centradas y no están correladas, es decir, $\mathrm{E}(\boldsymbol{y}) = \mathbf{0}_d$ y su matriz de covarianzas $\mathrm{Cov}(\boldsymbol{y}) = \mathrm{E}(\boldsymbol{y}\boldsymbol{y}^T)$ es diagonal. Las variables observadas, se suponen centradas, es decir, $\mathrm{E}(\boldsymbol{x}) = \mathbf{0}_D$. El objetivo de ACP es obtener la matriz $\mathbf{W}$ que permita obtener los valores de cada dato en las variables latentes a partir de sus valores en las variables originales.

Suponiendo que la matriz de covarianzas de $Cov(\boldsymbol{x})$ es conocida, tenemos que:

$$Cov(\boldsymbol{x}) = \mathrm{E}(\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}) = \mathrm{E}(\mathbf{W}\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}\mathbf{W}^T) = \mathbf{W}\mathrm{E}(\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}})\mathbf{W}^T = \mathbf{W}Cov(\boldsymbol{y})\mathbf{W}^T \tag{3-2}$$

Y puesto que $\mathbf{W}^T\mathbf{W} = \mathbf{I}_d$, multiplicando a izquierda y derecha por $\mathbf{W}^T$ y $\mathbf{W}$ respetivamente, se obtiene:

$$Cov(\boldsymbol{y}) = \mathbf{W}^T Cov(\boldsymbol{x})\mathbf{W} \tag{3-3}$$

Seguidamente, la matriz $\mathrm{Cov}(\boldsymbol{x})$ puede ser descompuesta en sus valores y vectores propios (Saad, 2011) :

$$Cov(\boldsymbol{x}) = \mathbf{V}\,\mathbf{\Lambda}\,\mathbf{V}^{\mathrm{T}} \tag{3-4}$$

donde $\mathbf{V}$ es una matriz de vectores propios normalizados y $\mathbf{\Lambda}$ una matriz diagonal que contiene los valores propios asociados en orden descendente. Puesto que la matriz $Cov(\boldsymbol{x})$ es simétrica y definida como semipositiva, los vectores propios son ortogonales y los valores propios no negativos. Finalmente, se obtiene:

$$Cov(\boldsymbol{y}) = \mathbf{W}^T\mathbf{V}\,\mathbf{\Lambda}\,\mathbf{V}^{\mathrm{T}}\mathbf{W} \tag{3-5}$$

De donde se obtiene la expresión de $\mathbf{W}$:

$$\mathbf{W} = \mathbf{V}\mathbf{I}_{D\times d} \tag{3-6}$$

siendo $\mathbf{I}_{D\times d}$ la matriz de dimensiones $D\times d$ con $\mathbf{I}_i^j = 1$ si $i = j$ y $\mathbf{I}_i^j = 0$ si $i \neq j$. Ello es debido a que, al sustituir (3-6) en la ecuación (3-5) y dado que $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{V}\,\mathbf{V}^{\mathrm{T}} = \mathbf{I}$ por ser $\mathbf{V}$ ortogonal, tenemos $Cov(\boldsymbol{y}) = \mathbf{I}_{d\times D}\mathbf{V}\,\boldsymbol{\Lambda}\mathbf{I}_{D\times d}$. De este modo, las variables originales quedan expresadas como combinación de unas nuevas variables latentes incorreladas entre sí y de máxima varianza.

En cuanto a las nuevas coordenadas de los datos en el espacio de baja dimensión, se necesita centrar los datos originales de la matriz $\boldsymbol{X}$ y, a partir de éstos, calcular la matriz de covarianzas. Si la matriz $\boldsymbol{Z}$ representa a los datos originales centrados, $\boldsymbol{Z}^i = \boldsymbol{X}^i - \mathrm{E}(\boldsymbol{X}^i)$, para cada variable i, la matriz de covarianzas de las variables originales puede estimarse mediante la expresión:

$$\widehat{Cov}(\boldsymbol{x}) = \frac{1}{n}\boldsymbol{Z}^T\boldsymbol{Z} \tag{3-7}$$

ACP se ha aplicado con éxito en un gran número de dominios tales como clasificación de moneda (Huber, Ramoser, Mayer, Penz, & Rubik, 2005), diagnóstico de fallos en engranajes (Li, Yan, Yuan, Peng, & Li, 2011) o variación en los perfiles antioxidantes de frutas y verduras (Patras, Brunton, Downey, Rawson, Warriner, & Gernigon, 2011) entre muchos otros.

Los principales inconvenientes de ACP son:

- Como el tamaño de la matriz de covarianza es proporcional a la dimensionalidad de los datos, el cálculo de los vectores propios podría no ser factible para datos de muy alta dimensión.
- Cálculo de la matriz de covarianzas en conjuntos de datos en los que $n < D$. En este caso, el problema puede superarse mediante el cálculo de los vectores propios del cuadrado de la matriz de distancia euclidiana $\boldsymbol{X}\boldsymbol{X}^T$ en lugar de los correspondientes a la matriz de covarianzas que es proporcional a $\boldsymbol{X}^T\boldsymbol{X}$.
- Adaptación a conjuntos de datos con estructura no lineal.

Otros métodos como *EM PCA* (ver sección 3.4.2.) o *MDS* (ver sección 3.4.5.) han sido desarrollados a partir de ACP con la intención de mejorarlo.

3.4.2. Análisis de Componentes Principales Probabilístico con algoritmo *Expectation-Maximization* (*EM ACP*)

Como se comentado en la anterior sección, uno de los defectos de la técnica ACP es el problema que tiene con conjuntos que constan de un gran número de datos o cuando estos datos son de alta dimensionalidad. Intentar diagonalizar la matriz de covarianza muestral, o incluso el mismo cálculo de esta matriz, en estas situaciones puede hacer que surjan dificultades en cuanto a complejidad computacional. En general, lo mejor es evitar por completo el cálculo de la matriz de covarianza muestral de forma explícita. El algoritmo *Expectation-Maximization* o algoritmo *EM* (McLachlan & Krishnan, 2007), que en Estadística se utiliza para encontrar estimadores de máxima verosimilitud de parámetros en modelos probabilísticos que dependen de variables no observables, puede utilizarse para el cálculo de los principales vectores propios de un conjunto de datos.

Otro inconveniente de los enfoques estándar de ACP es que no es obvio cómo tratar adecuadamente el problema de la falta de datos. Este problema también puede abordarse mediante el empleo del algoritmo *EM* para una estimación de máxima verosimilitud de los valores que faltan en cada iteración (Vitelleschi & Quaglino, 2010).

El objetivo de *EM ACP* es capturar la estructura de covarianza de D variables observadas y recogidas en el vector aleatorio $\boldsymbol{x}$ usando menos parámetros de los requeridos en una matriz de covarianza completa. El planteamiento consiste en definir un modelo de probabilidad, reformulando ACP como una solución de máxima verosimilitud para un modelo de variables latentes (Guan & Dy, 2009). Ello se consigue suponiendo que $\boldsymbol{x}$ se produjo como una transformación lineal de un conjunto de d variables latentes que conforman el vector aleatorio $\boldsymbol{y}$ más la adición de ruido gaussiano. Denotando la transformación mediante la matriz $\boldsymbol{C}$ (de dimensiones $D \times d$), y el ruido gaussiano por $\boldsymbol{v}$ (de dimensión D y con matriz de covarianza $Cov(\boldsymbol{v}) = \boldsymbol{R}$) el modelo generativo puede escribirse como:

$$\boldsymbol{x} = \boldsymbol{C}\boldsymbol{y} + \boldsymbol{v} \tag{3-8}$$

donde $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ y $\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R})$.

Las variables latentes $\boldsymbol{y}$ están centradas y son independientes e idénticamente distribuidas según una curva normal. Por tanto, el modelo puede escribirse de la forma:

$$\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{C}\boldsymbol{C}^T + \boldsymbol{R}) \tag{3-9}$$

donde se ha sustraído de $\boldsymbol{x}$ el vector de medias.

Para poder obtener los parámetros de la distribución, es necesario exigir que $d < D$ y que la estructura de la covarianza del ruido gaussiano $\boldsymbol{v}$ quede restringida mediante la imposición de condiciones sobre $\boldsymbol{R}$.

Hay dos principales problemas cuando se trabaja con los modelos lineales descritos anteriormente. El primer problema es, suponiendo conocidas las matrices $\boldsymbol{C}$ y $\boldsymbol{R}$, deducir los estados ocultos $\boldsymbol{y}$ dadas las n observaciones (puntos) del conjunto de datos. Puesto que los puntos son independientes, el interés se centra en la probabilidad a *posteriori* de un estado oculto conociendo únicamente la observación correspondiente. Esto puede ser fácilmente calculado mediante proyección lineal resultando la distribución de densidad asimismo gaussiana (Roweis, 1997):

$$P(\boldsymbol{y}/\boldsymbol{x}) = \frac{P(\boldsymbol{x}/\boldsymbol{y})P(\boldsymbol{y})}{P(\boldsymbol{x})} = \frac{\mathcal{N}(\boldsymbol{C}\boldsymbol{y}, \boldsymbol{R})|_x \ \mathcal{N}(\mathbf{0}, \boldsymbol{I})|_y}{\mathcal{N}(\mathbf{0}, \boldsymbol{C}\boldsymbol{C}^T + \boldsymbol{R})|_x} \tag{3-10}$$

$$P(\boldsymbol{y}/\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\beta}\boldsymbol{x}, \boldsymbol{I} - \boldsymbol{\beta}\boldsymbol{C})|_y \quad \text{donde} \quad \boldsymbol{\beta} = \boldsymbol{C}^T(\boldsymbol{C}\boldsymbol{C}^T + \boldsymbol{R})^{-1} \tag{3-11}$$

de donde se puede obtener el valor esperado $\boldsymbol{\beta}\boldsymbol{x}$ del estado oculto y una estimación de la incertidumbre mediante la varianza $\boldsymbol{I} - \boldsymbol{\beta}\boldsymbol{C}$. La reconstrucción de $\boldsymbol{x}$ a partir de $\boldsymbol{y}$ resulta de $P(\boldsymbol{x}/\boldsymbol{y}) = \mathcal{N}(\boldsymbol{C}\boldsymbol{y}, \boldsymbol{R})|_x$ y la probabilidad para cualquier punto puede evaluarse mediante la expresión (3-9).

El segundo problema es el de ajustar las matrices $\boldsymbol{C}$ y $\boldsymbol{R}$ de tal forma que haga que el modelo asigne la probabilidad más alta a los datos observados, este proceso se realiza empleando el algoritmo *EM*. Este algoritmo alterna dos pasos de forma iterativa:

- **Paso E**, donde se estima el estado latente utilizando la expresión de inferencia (3-11).
- **Paso M**, donde se calculan estimadores de máxima verosimilitud para las matrices $\boldsymbol{C}$ y $\boldsymbol{R}$ de tal forma que se maximice la probabilidad del estado observado y del estado latente calculado en el paso E. Denominando $\boldsymbol{W} = \boldsymbol{C}\boldsymbol{C}^T + \boldsymbol{R}$ y siendo $\boldsymbol{S}$ la matriz de covarianzas de la

muestra de las observaciones, el logaritmo de la correspondiente función de probabilidad es (Li, Yeung, & Zhang, 2009):

$$\mathcal{L} = \ln\big(P(\boldsymbol{x})\big) = \ln\big(\mathcal{N}(\mathbf{0}, \boldsymbol{W})\big) = \\ = -\frac{n}{2}[Dln(2\pi) + \ln|\boldsymbol{W}| + traza(\boldsymbol{W}^{-1}\boldsymbol{S})] \qquad (3\text{-}12)$$

EM ACP se ha aplicado en contextos como el análisis espectrofotométrico de Supernovas (Saunders, et al., 2014), reconocimiento facial (Rujirakul, So-In, & Arnonkijpanich, 2014) o identificación de voz (Chien & Ting , 2004) entre otros.

3.4.3. Análisis Factorial (AF)

El Análisis Factorial, AF o *FA* (del inglés, *Factor Analysis*) (Cudeck & MacCallum, 2007) se originó en Psicometría, donde se utiliza para identificar rasgos latentes en datos obtenidos a nivel de encuesta. En ocasiones, un fenómeno de interés es complejo y no directamente medible. En ese caso, es necesario hacer una serie de preguntas sobre el fenómeno y luego combinar adecuadamente los resultados con el fin de obtener una sola medida o "factor". Este factor, a continuación, se convierte en la medida del fenómeno inobservable o latente.

AF es un modelo generativo que supone que los datos observados se han producido a partir de un conjunto de variables no observadas latentes (llamadas factores) a través de la ecuación:

$$\boldsymbol{x} = \boldsymbol{\Lambda y} + \boldsymbol{u} \qquad (3\text{-}13)$$

donde $\boldsymbol{x}$ es el vector aleatorio de dimensión D correspondiente a las variables observadas, $\boldsymbol{y}$ es el vector aleatorio de factores comunes con dimensión d y $\boldsymbol{u}$ es el vector de los factores específicos de dimensión D. Se tienen en cuenta, además, las siguientes suposiciones:

- $E(\boldsymbol{x}) = \mathbf{0}$ (si no fuera el caso, se sustrae el vector de medias)
- $E(\boldsymbol{y}) = \mathbf{0}\ \ y\ \ Cov(\boldsymbol{y}) = \mathbf{I}$
- $E(\boldsymbol{u}) = \mathbf{0}\ \ \ y\ \ Cov\big(\boldsymbol{u}_i, \boldsymbol{u}_j\big) = 0\ \ para\ \ i \neq j$

- $Cov(\boldsymbol{y}, \boldsymbol{u}) = 0$
- Los factores siguen una distribución normal multivariante.

En estas condiciones, AF puede considerarse como un caso particular de *EM ACP* en el que la matriz de covarianzas es diagonal.

La covarianza de las variables observadas es

$$Cov(\boldsymbol{x}) = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + Cov(\boldsymbol{u}) \tag{3-14}$$

donde la matriz de covarianza de los factores específicos tiene que estimarse a partir de los datos y la covarianza de los datos a partir de la covarianza muestral. La matriz $\boldsymbol{\Lambda}$ se resuelve mediante factorización de la matriz $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T = Cov(\boldsymbol{x}) - Cov(\boldsymbol{u})$. Esta factorización no es única, ya que para cualquier rotación ortogonal de $\boldsymbol{\Lambda}$ resulta la misma descomposición de $Cov(\boldsymbol{x}) - Cov(\boldsymbol{u})$ (Williams , Brown, & Onsman, 2012). Este aspecto es aprovechado con el fin de obtener soluciones más simples.

Dado que normalmente el número de parámetros para la estimación de $\boldsymbol{\Lambda}$ y $Cov(\boldsymbol{u})$ es menor que el número de parámetros de la covarianza muestral de los datos, en general no hay una solución exacta del problema. En este caso, también es utilizado, como en *EM ACP*, el algoritmo *Expectation-Maximization* para estimar $\boldsymbol{\Lambda}$ y $Cov(\boldsymbol{u})$.

Se distinguen dos tipos de análisis factorial (Schmitt, 2011; Méndez & Rondón, 2012):

- **Análisis factorial exploratorio:** Se centra en encontrar o establecer una estructura interna, generando nuevos factores a partir de un conjunto de variables, o reducir el número de éstas. En el primer escenario, se establece cuál es la contribución de las variables originales a cada uno de estos nuevos factores, mientras que en el segundo se eliminan del análisis aquellas variables que sean poco relevantes o que estén muy correlacionadas con otras.
- **Análisis factorial confirmatorio:** Tiene por meta evaluar hasta qué punto un conjunto de factores organizados teóricamente se ajusta a los datos.

AF es un método utilizado en las ciencias del comportamiento tales como ciencias sociales (White, Bray, & Ollendick, 2012) o marketing (Persaud & Azhar, 2012), en medicina (de Stefani, et al., 2012) y en otras ciencias aplicadas (Zhang, et al., 2011; Dehak, Kenny, Dehak, Dumouchel, & Ouellet, 2011).

3.4.4. Análisis Discriminante Lineal (*LDA*)

El Análisis Discriminante lineal más conocido como *LDA* (del inglés, *Linear Discriminant analysis*) es una técnica de aprendizaje supervisado para clasificar datos. La idea central de *LDA* es obtener una proyección de los datos en un espacio de menor (o incluso igual) dimensión que los datos entrantes, con el fin de que la separabilidad de las clases sea la mayor posible. Así como ACP busca minimizar el error de representación cometido, este no es el objetivo de *LDA* que, como se ha comentado anteriormente, se centra en encontrar la combinación o transformación de variables que garantice una máxima separabilidad entre clases.

Existen varias implementaciones de *LDA*, entre ellas se encuentra *Fisher-LDA* (Welling, 2005). Considérese la versión más simple del problema que consiste en encontrar el vector $\boldsymbol{\omega}$ que proyecte los datos a un espacio de una sola dimensión con el fin de obtener la mayor separabilidad entre clases.

Así, tenemos *n* patrones $\boldsymbol{X}_1, .., \boldsymbol{X}_n$ *D*-dimensionales etiquetados en *c* clases donde cada clase cuenta con n_i patrones ($i \in \{1, .., c\}$). Se busca un vector $\boldsymbol{\omega}$ en el cual obtener las *n* proyecciones unidimensionales de los patrones $\boldsymbol{Y}_1, .., \boldsymbol{Y}_n$.

Lo que busca *Fisher-LDA* es maximizar la siguiente función objetivo (Welling, 2005):

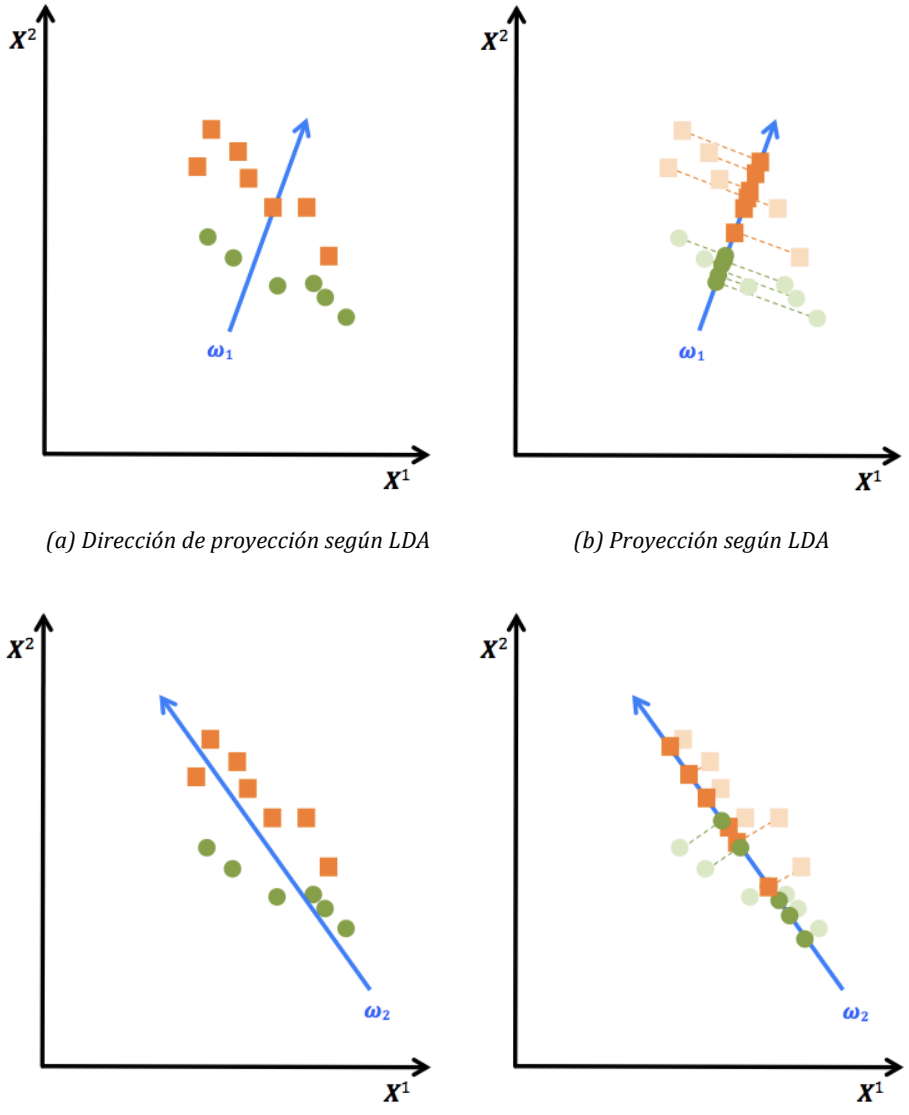
$$J(\boldsymbol{\omega}) = \frac{\boldsymbol{\omega}^T \boldsymbol{S}_B \boldsymbol{\omega}}{\boldsymbol{\omega}^T \boldsymbol{S}_W \boldsymbol{\omega}} \tag{3-15}$$

donde $\boldsymbol{S}_B$ es la matriz de dispersión interclase y $\boldsymbol{S}_W$ es la matriz de dispersión intraclase, es decir:

$$\boldsymbol{S}_B = \sum_{i=1}^{c} n_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T \tag{3-16}$$

$$\boldsymbol{S}_W = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (\boldsymbol{X}_j - \boldsymbol{\mu}_i)(\boldsymbol{X}_j - \boldsymbol{\mu}_i)^T \tag{3-17}$$

donde $\boldsymbol{\mu}_i$ es el vector media de cada clase, $\boldsymbol{\mu}$ el vector media de todos los datos y n_i la cantidad de datos de la clase *i*.



(a) Dirección de proyección según LDA *(b) Proyección según LDA*

(c) Dirección de proyección según ACP *(d) Proyección según* ACP

Figura 3-5: *Proyección de un conjunto de datos de dos dimensiones diferenciados en dos clases. En (a)* $\boldsymbol{\omega_1}$ *indica la dirección de proyección según LDA y en (b) se muestra la proyección de los datos según esa dirección observándose que ambas clases quedan separadas. En (c)* $\boldsymbol{\omega_2}$ *está dispuesto en la dirección de máxima varianza de los datos (según* ACP*) y en (d) se observa que en la proyección sobre este vector no logra separar correctamente dos clases de datos.*

Lo que se pretende es encontrar el vector $\boldsymbol{\omega}$ de proyección que maximice $J(\boldsymbol{\omega})$. Este vector debe cumplir:

$$\boldsymbol{S}_B\boldsymbol{\omega} = \lambda \boldsymbol{S}_W\boldsymbol{\omega} \tag{3-18}$$

Si $\boldsymbol{S}_W$ es no singular puede resolverse el problema de valores propios para la matriz $\boldsymbol{S}_W^{-1}\boldsymbol{S}_B$:

$$\boldsymbol{S}_W^{-1}\boldsymbol{S}_B\boldsymbol{\omega} = \lambda\boldsymbol{\omega} \tag{3-19}$$

Consecuentemente, el vector $\boldsymbol{\omega}$ que maximiza $J(\boldsymbol{\omega})$ es aquel con mayor valor propio asociado. En la figura 3-5 se ilustra una situación en la cual existen dos clases en el conjunto de datos y dos características de los mismos ($\boldsymbol{X}^1$ y $\boldsymbol{X}^2$). Ambos conjuntos bidimensionales son proyectados en una sola dimensión según dos vectores distintos. En la figura 3-5(b) la proyección sobre $\boldsymbol{\omega}_1$ es la adecuada según *LDA* y las clases se separan. En la figura 3-5(d) la proyección sobre $\boldsymbol{\omega}_2$ es en la dirección de máxima varianza (según ACP) y puede observarse que, en este caso, no se logra el objetivo de separar ambas clases.

Para el caso de que la proyección sea en un espacio de *d* dimensiones, el problema a resolver es el mismo eligiendo los *d* vectores propios con valores propios asociados más grandes.

Pueden encontrarse aplicaciones de *LDA* en caracterización bioquímica (Park, Pande, Shrestha, Clubb, Applegate, & Jo, 2012), reconocimiento facial (Shu, Gao, & Lu, 2012) o medicina (Luo, Kim, Dighe, & Kim, 2011) aparte de otros.

3.4.5. Escalado Multidimensional (*MDS*)

El Escalado Multidimensional o *MDS* (del inglés, *Multidimensional Scaling*) (Borg & Groenen, 2005) representa una colección de técnicas no lineales que asignan a la representación de datos de alta dimensión a una representación de menor dimensión intentando conservar las disimilitudes entre pares de datos tanto como sea posible.

Al implementar esta técnica en la aplicación desarrollada, la disimilitud entre pares de datos se representa mediante su distancia euclídea. La calidad de la

representación en una dimensión menor se expresa mediante la función de tensión, que constituye una medida del error entre las distancias correspondientes a los pares de datos en el espacio de alta dimensión y las correspondientes a sus representaciones en baja dimensión. La función de tensión en su forma más simple se define por (van der Maaten , Postma, & van den Herik , 2009):

$$\phi(\boldsymbol{Y}) = \sum_{ij} \left(\left\| \boldsymbol{X}_i - \boldsymbol{X}_j \right\| - \left\| \boldsymbol{Y}_i - \boldsymbol{Y}_j \right\| \right)^2 \tag{3-20}$$

en la que $\left\| \boldsymbol{X}_i - \boldsymbol{X}_j \right\|$ representa la distancia euclídea entre los puntos correspondientes a los datos de alta dimensión $\boldsymbol{X}_i$ y $\boldsymbol{X}_j$ mientras que $\left\| \boldsymbol{Y}_i - \boldsymbol{Y}_j \right\|$ es la distancia euclídea entre los puntos $\boldsymbol{Y}_i$ e $\boldsymbol{Y}_j$ que representan a los datos en baja dimensión. Otro tipo de función de tensión es la función de coste de *Sammon* cuya expresión es:

$$\phi(\boldsymbol{y}) = \frac{1}{\sum_{ij} \left\| \boldsymbol{X}_i - \boldsymbol{X}_j \right\|} \sum_{i \neq j} \frac{\left(\left\| \boldsymbol{X}_i - \boldsymbol{X}_j \right\| - \left\| \boldsymbol{Y}_i - \boldsymbol{Y}_j \right\| \right)^2}{\left\| \boldsymbol{X}_i - \boldsymbol{X}_j \right\|} \tag{3-21}$$

donde puede observarse que pone mayor énfasis en retener aquellas distancias que son pequeñas en el espacio de alta dimensión.

La minimización de la función de tensión puede realizarse, entre otros métodos, mediante la descomposición en vectores propios de la matriz de distancias por parejas (van der Maaten , Postma, & van den Herik , 2009).

MDS es ampliamente utilizado en aplicaciones biológicas (Rambaut, Pybus, Nelson, Viboud, Taubenberger, & Holmes, 2008), de economía (Machado, Duarte, & Duarte, 2011) o de localización (Wu, Sheng, & Zhang, 2006) entre otras. La popularidad de *MDS* ha dado lugar a la propuesta de variantes entre las que cabe destacar *SNE* (ver sección 3.4.8.).

3.4.6. *Kernel PCA* (*KPCA*)

Kernel PCA o *KPCA* (Shawe-Taylor & Cristianini , 2004) es una técnica no lineal cuyo enfoque está muy relacionado con *MDS*. *KPCA* extiende las propiedades

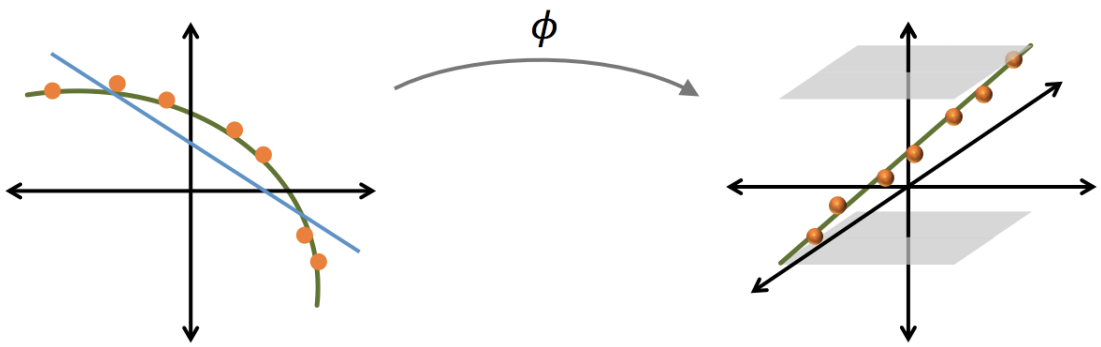
algebraicas de *MDS* a *manifolds* no lineales, sin tener en cuenta su significado geométrico.

La primera idea de *KPCA* consiste en trabajar con la matriz de productos escalares, proporcional a $\boldsymbol{X}\boldsymbol{X}^T$ como en *MDS*, en vez de con la matriz de covarianza de la muestra, proporcional a $\boldsymbol{X}^T\boldsymbol{X}$ como en ACP.

La segunda idea de *KPCA* es dotar de una estructura lineal al *manifold* subyacente $\mathcal{M}$. Para ello, *KPCA* utiliza una función:

$$\begin{aligned} \phi: \mathcal{M} \subset \mathbb{R}^D &\rightarrow \mathbb{R}^Q \\ \boldsymbol{X}_i &\rightarrow \boldsymbol{Z}_i = \phi(\boldsymbol{X}_i) \end{aligned} \tag{3-22}$$

donde Q puede ser cualquier dimensión, posiblemente mayor que D o incluso infinita. La única suposición sobre esta función es que los datos asignados abarcan un subespacio lineal de $\mathbb{R}^Q$. En la figura 3-6 se ilustra esta idea. En 3-6(a) se muestra un conjunto de datos (círculos naranjas) siguiendo la estructura de una línea curva (línea verde) que no puede describirse convenientemente usando la técnica ACP (línea azul). Para describir correctamente este conjunto de datos por el método estándar ACP serían necesarios al menos dos componentes principales. En 3-6(b) se muestra el conjunto de datos después de su transformación a través de la función ϕ. Como puede apreciarse, en este espacio transformado y ampliado puede describirse con precisión el conjunto de datos usando una sola componente principal dado que el conjunto de datos sigue una estructura lineal.



(a) *(b)*

Figura 3-6: *Transformación mediante la función $\boldsymbol{\phi}$ de un conjunto de datos (puntos naranjas) a un espacio de mayor dimensión adquiriendo una estructura lineal.*

Una vez elegida la función ϕ, se calcula la matriz $\mathbf{\Phi}$ cuyos elementos son los productos escalares de las representaciones de los datos en $\mathbb{R}^Q$:

$$\mathbf{\Phi} = \left[\langle \phi(\boldsymbol{X}_i) \cdot \phi(\boldsymbol{X}_j)\rangle\right]_{1\leq i,j\leq n} = \left[\langle \boldsymbol{Z}_i \cdot \boldsymbol{Z}_j\rangle\right]_{1\leq i,j\leq n} \tag{3-23}$$

A continuación, al igual que en *MDS*, la matriz simétrica $\mathbf{\Phi}$ se descompone en sus valores y vectores propios. Para que esto pueda realizarse, $\mathbf{\Phi}$ debe ser semidefinida positiva, es decir, los valores $\boldsymbol{Z}_i = \phi(\boldsymbol{X}_i)$ han de centrarse. El centrado de $\boldsymbol{Z}_i$ no puede realizarse directamente ya que la representación es desconocida. Sin embargo, puede lograrse de manera implícita mediante la realización de un doble centrado en la matriz $\mathbf{\Phi}$ (Lee & Verleysen, 2007). Es entonces cuando $\mathbf{\Phi}$ puede ya descomponerse en sus valores y vectores propios.

En cuanto a la función ϕ, dado que únicamente se utiliza en los productos escalares, puede permanecer desconocida si existe una función *kernel* (núcleo) κ, que es una función que proporciona directamente el valor del producto escalar a partir de $\boldsymbol{X}_i$ y $\boldsymbol{X}_j$, es decir:

$$\begin{aligned} &\kappa : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R} \\ &(\boldsymbol{X}_i, \boldsymbol{X}_j) \to \kappa(\boldsymbol{X}_i, \boldsymbol{X}_j) = \langle \phi(\boldsymbol{X}_i) \cdot \phi(\boldsymbol{X}_j)\rangle \end{aligned} \tag{3-24}$$

Necesariamente, κ no puede ser cualquier función: su reformulación como un producto escalar implica satisfacer una serie de condiciones. Más precisamente, el teorema de *Mercer* de análisis funcional (Steinwart & Scovel, 2012) establece que:

- Si κ es un *kernel* continuo de un operador integral positivo $\mathcal{K}$ definido como:

$$\begin{aligned} &\mathcal{K} : L_2 \to L_2 \\ &f \to \mathcal{K}f \end{aligned} \tag{3-25}$$

con $(\mathcal{K}f)(\boldsymbol{v}) = \int \kappa(\boldsymbol{u}, \boldsymbol{v}) f(\boldsymbol{v}) d\boldsymbol{v}$

- Si $\mathcal{K}$ es definido positivo, es decir,

$$\int f(\boldsymbol{u})\kappa(\boldsymbol{u},\boldsymbol{v})f(\boldsymbol{v})\boldsymbol{du}\boldsymbol{dv} > 0, \qquad \text{si } f \neq 0 \tag{3-26}$$

Entonces κ puede ser expresado como:

$$\kappa(\boldsymbol{u},\boldsymbol{v}) = \sum_{q=1}^{\infty} \lambda_q \, \phi_q(\boldsymbol{u}) \, \phi_q(\boldsymbol{v}) \tag{3-27}$$

siendo λ_q coeficientes positivos (los valores propios) y ϕ_q funciones ortogonales, es decir,

$$\langle \phi_{q_1} \cdot \phi_{q_2} \rangle = \begin{cases} 0 \text{ si } q_1 \neq q_2 \\ 1 \text{ si } q_1 = q_2 \end{cases} \tag{3-28}$$

A partir de la ecuación (3-27) se obtiene fácilmente que

$$\phi(\boldsymbol{X}_i) = \sum_{q=1}^{\infty} \sqrt{\lambda_q} \, \phi_q(\boldsymbol{X}_i) \tag{3-29}$$

es una función para la cual κ actúa como producto escalar, es decir, $\kappa\,(\boldsymbol{X}_i\,, \boldsymbol{X}_j) = \langle \phi(\boldsymbol{X}_i) \cdot \phi(\boldsymbol{X}_j) \rangle$.

En la práctica hay tres tipos de funciones *kernel* básicas:

- Función *kernel* polinómica: $\kappa(\boldsymbol{u},\boldsymbol{v}) = (\langle \boldsymbol{u} \cdot \boldsymbol{v} \rangle)^p$, con $p \in \mathbb{Z}$.
- Función *kernel* sigmoidal: $\kappa(\boldsymbol{u},\boldsymbol{v}) = \tanh(\langle \boldsymbol{u} \cdot \boldsymbol{v} \rangle + b)$.
- Función *kernel* de base radial gaussiana: $\kappa(\boldsymbol{u},\boldsymbol{v}) = exp\left(-\frac{\|\boldsymbol{u}-\boldsymbol{v}\|^2}{2\sigma^2}\right)$. Este tipo de función *kernel* es el utilizado en la implementación de *KPCA* en la aplicación desarrollada en esta tesis.

Este enfoque sobre las funciones *kernel* juega un papel clave en una gran familia de los métodos llamados Máquinas de Vectores Soporte o *SVM* (del inglés, *Support Vector Machines*) (Steinwart & Christmann, 2008).

KPCA es un método que es utilizado en numerosas aplicaciones entre las que pueden citarse algunas relacionadas con reconocimiento de patrones (Peng, Cao, Liu, & Pan, 2013), tratamiento de señales (Widjaja, Varon, Dorado, Suykens, & Van Huffel, 2012; Cheng, Yang, Zheng, Li, & Ren, 2015), autentificación (Damavandinejadmonfared & Varadharajan, 2015), etc.

En comparación con métodos mencionados anteriormente como ACP, *KPCA* tiene la ventaja de poder tratar con *manifolds* no lineales. Sin embargo, la principal dificultad de *KPCA*, estriba en la elección de una función *kernel* adecuada junto con unos valores correctos de sus parámetros.

3.4.7. Representación Isométrica de Características (*Isomap*)

La Representación Isométrica de Características o *Isomap* (del inglés, *Isometric feature Mapping*) (de Silva & Tenenbaum, 2003) es un método de reducción de la dimensión no lineal cuyo objetivo es preservar las propiedades globales de los datos originales en su representación en el espacio de baja dimensión.

A pesar de la eficacia demostrada por técnicas como *MDS* basadas en las distancias euclídeas, éstas no tienen en cuenta la distribución de los datos que se encuentran en un entorno. Como ya se indicó en la sección 3.2.1. si los datos de alta dimensión se encuentran en, o cerca de, un *manifold* curvado, las técnicas basadas en distancias euclídeas pueden considerar dos puntos como cercanos, mientras que su distancia sobre el *manifold*, distancia geodésica, es mucho mayor que la distancia típica de punto a punto. *Isomap* resuelve este problema al intentar preservar esta distancia geodésica entre pares de datos.

Lo importante en *Isomap* es encontrar una forma eficiente para calcular la verdadera distancia geodésica entre observaciones, dado que sólo se conocen las distancias euclídeas en el espacio de alta dimensión. *Isomap* asume que la distancia entre puntos en el espacio observación es una medida exacta de la distancia del *manifold* solamente a nivel local (respecto de los *k* vecinos más próximos, parámetro éste necesario y que es solicitado por la aplicación desarrollada en esta tesis). En otro caso, la distancia geodésica debe calcularse sobre caminos en el *manifold* mediante la construcción de un grafo en el que cada punto se conecta con sus *k* vecinos más próximos.

El procedimiento llevado a cabo por *Isomap* consta de tres pasos principales (de Silva & Tenenbaum, 2003):

1. La cuestión principal en *Isomap* es encontrar una forma eficiente de calcular la verdadera distancia geodésica entre los datos, dado que solamente pueden conocerse las distancias euclídeas en el espacio de alta dimensión. Esta técnica propone calcular las distancias geodésicas entre datos mediante la construcción de un grafo de vecindad G en el que cada dato $\boldsymbol{X}_i$ se conecta con sus k vecinos más próximos del conjunto de datos.
2. Una vez construido el grafo, la distancia geodésica entre dos puntos se estima recorriendo el camino más corto que los une a través de aquél. La opción utilizada en la aplicación desarrollada en esta tesis es el algoritmo de *Dijkstra* (Cormen, Leiserson, Rivest, & Stein, 2009). De este modo, se construye una matriz de distancias geodésicas entre todos los pares de puntos.
3. La representación de cada punto en el espacio de baja dimensión se calcula aplicando *MDS* sobre la matriz de distancias geodésicas.

Este algoritmo sufre de ciertas debilidades (van der Maaten , Postma, & van den Herik , 2009) como su inestabilidad topológica al poder construir conexiones erróneas o cortocircuitadas en el grafo y problemas con *manifolds* no convexos o en los que existen zonas con poca densidad de puntos. No obstante, *Isomap* se ha aplicado con éxito en trabajos tales como análisis de fenómenos atmosféricos (Hannachi & Turner, 2013), reconocimiento de perfiles (Jyotsna, Akhil, & Arun, 2013) o localización de sensores en redes inalámbricas (Wang, Chen, Sun, & Shen, 2009) entre muchos otros.

3.4.8. Proyección Estocástica del Entorno (*SNE*)

La Proyección Estocástica del Entorno o *SNE* (del inglés, *Stochastic Neighbor Embedding*) es una técnica de reducción de la dimensión cuyo objetivo es preservar de manera óptima el entorno de los datos basándose en métodos probabilísticos (Hinton & Roweis, 2002).

A cada dato, que es considerado un punto en el espacio de alta dimensión, se le asigna una distribución de probabilidad normal respecto de la cual se estima el potencial del resto de datos como vecino suyo (de su entorno).

Así, para cada punto $\boldsymbol{X}_i$ y cada vecino potencial $\boldsymbol{X}_j$, se empieza calculando la probabilidad condicional $p_{j|i}$ de que $\boldsymbol{X}_i$ pueda tener a $\boldsymbol{X}_j$ como vecino suyo:

$$p_{j|i} = \frac{exp\left(-d_{j|i}^2\right)}{\sum_{k \neq i} exp\left(-d_{k|i}^2\right)} \tag{3-30}$$

Las disimilitudes $d_{j|i}$ pueden obtenerse como parte de la definición del problema (no necesitando ser simétricas) o bien pueden ser calculadas utilizando el cuadrado de la distancia euclídea entre los puntos en el espacio de alta dimensión (opción escogida al implantar este método en la aplicación de esta tesis), en este caso:

$$d_{j|i}^2 = \frac{\left\|\boldsymbol{X}_i - \boldsymbol{X}_j\right\|^2}{2\sigma_i^2} \tag{3-31}$$

donde σ_i^2, varianza de la distribución normal centrada en $\boldsymbol{X}_i$, puede establecerse directamente de acuerdo con la experiencia sobre el problema a tratar o mediante una búsqueda binaria de tal modo que haga que la entropía de la distribución sobre los puntos vecinos sea igual a $\log(k)$, donde k es el número efectivo de vecinos locales, el cual es escogido a mano (Hinton & Roweis, 2002). Este parámetro se solicita en la aplicación desarrollada.

En el espacio de baja dimensión se procede de igual modo utilizando entornos con distribución normal pero con una varianza fija (que, sin pérdida de generalidad, se establece en $\frac{1}{2}$). Así, la probabilidad condicional $q_{j|i}$ de que el punto $\boldsymbol{y}_i$ tenga al punto $\boldsymbol{y}_j$ en su entorno viene dada por la expresión:

$$q_{j|i} = \frac{exp\left(-\left\|\boldsymbol{Y}_i - \boldsymbol{Y}_j\right\|^2\right)}{\sum_{k \neq i} exp(-\|\boldsymbol{Y}_i - \boldsymbol{Y}_k\|^2)} \tag{3-32}$$

El objetivo en la reducción de la dimensión es hacer coincidir estas dos distribuciones tanto como sea posible. Esto se consigue minimizando una función de coste que es una suma de las divergencias de Kullback-Leibler (Pérez-Cruz, 2008) entre las distribuciones originales ($p_{j|i}$) y las inducidas ($q_{j|i}$) para cada dato:

$$\mathcal{J} = \sum_i \sum_j p_{j|i} \, log \frac{p_{j|i}}{q_{j|i}} \tag{3-33}$$

cuya expresión diferencial con respecto a cada dato $\boldsymbol{Y}_i$ es:

$$\frac{\partial \mathcal{J}}{\partial \boldsymbol{Y}_i} = 2 \sum_j (\boldsymbol{Y}_i - \boldsymbol{Y}_j)(p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j}) \tag{3-34}$$

que puede interpretarse fácilmente como una suma de fuerzas de tracción o repulsión de $\boldsymbol{Y}_i$ hacia $\boldsymbol{Y}_j$ según si se observa que es un dato vecino con más o menos frecuencia de la esperada. De este modo, los valores de cada dato en el espacio de baja dimensión se van actualizando de forma iterativa añadiendo al término diferencial un término de momento con el fin de acelerar la optimización. La expresión de actualización es (van der Maaten & Hinton, 2008):

$$\boldsymbol{Y}_i^{(k)} = \boldsymbol{Y}_i^{(k-1)} + \eta \frac{\partial \mathcal{J}}{\partial \boldsymbol{Y}_i}\Big|_{\boldsymbol{y}_i = \boldsymbol{y}_i^{(k-1)}} + \beta(k)(\boldsymbol{Y}_i^{(k-1)} - \boldsymbol{Y}_i^{(k-2)}) \tag{3-35}$$

donde η es la tasa de aprendizaje, $\beta(k)$ es el momento en la iteración k e $\boldsymbol{Y}_i^{(k)}$ es la solución para $\boldsymbol{Y}_i$ en la iteración k. La implementación de esta técnica en la aplicación desarrollada en esta tesis añade una fluctuación aleatoria que va disminuyendo conforme avanzan las iteraciones. Esto permite orientarse hacia mejores óptimos locales aunque con ello el proceso se ralentiza. En un primer momento, los valores de $\boldsymbol{Y}_i$ se inicializan situándolos en lugares al azar cerca del origen.

Aunque *SNE* construye razonablemente buenas visualizaciones, se ve obstaculizada por una función de coste que es difícil de optimizar y por un problema que suele denominarse "problema de hacinamiento" (van der Maaten & Hinton, 2008). El “problema de hacinamiento” suele ocurrir cuando en un espacio de alta dimensión tenemos un conjunto de datos cuyos puntos se encuentran en un *manifold* cuya dimensión intrínseca es mayor que la del espacio de baja dimensión en el que queremos representarlo. En este supuesto, las distancias entre los puntos en una representación en baja dimensión no pueden representar fielmente las distancias entre los puntos en un *manifold* de dimensión intrínseca mayor. Es por ello que, si se pretende modelar pequeñas distancias con precisión, la mayor parte

de los puntos que se encuentran a una distancia moderada de un determinado punto $\boldsymbol{X}_i$ en alta dimensión tendrán que colocarse demasiado lejos en baja dimensión. En *SNE*, los valores de las probabilidades correspondientes al punto $\boldsymbol{X}_i$ con respecto a cada uno de esos otros puntos tan lejanos son pequeños; no obstante, el gran número de tales probabilidades termina aglomerándolos e impidiendo la formación de espacios entre agrupaciones naturales.

Esos problemas son abordados en las versiones de esta técnica que se tratan en las dos siguientes secciones: *SSNE* y *t-SNE.*

3.4.9. Proyección Estocástica del Entorno Simétrica (*SSNE*)

La Proyección Estocástica del Entorno Simétrica o *SSNE* (del inglés, *Symmetric Stochastic Neighbor Embedding*) se basa en *SNE* definiendo nuevas probabilidades de forma conjunta tanto en el espacio de alta como de baja dimensión. De este modo se tienen las probabilidades p_{ij} y q_{ij} sustituyendo a las $p_{j|i}$ y $q_{j|i}$ condicionales de *SNE* (expresiones (3-30) y (3-32) respectivamente). Las nuevas probabilidades son (Cook, Sutskever, Mnih, & Hinton, 2007):

$$p_{ij} = \frac{exp\left(-\left\|\boldsymbol{X}_i - \boldsymbol{X}_j\right\|^2 / 2\sigma^2\right)}{\sum_{k \neq l} exp(-\|\boldsymbol{X}_k - \boldsymbol{X}_l\|^2 / 2\sigma^2)} \tag{3-36}$$

$$q_{ij} = \frac{exp\left(-\left\|\boldsymbol{Y}_i - \boldsymbol{Y}_j\right\|^2\right)}{\sum_{k \neq l} exp(-\|\boldsymbol{Y}_k - \boldsymbol{Y}_l\|^2)} \tag{3-37}$$

No obstante, esto puede causar problemas con datos atípicos, es decir, puntos $\boldsymbol{X}_i$ tales que su distancia $\left\|\boldsymbol{X}_i - \boldsymbol{X}_j\right\|^2$ es grande para cualquier otro punto. Esto provoca que su representación $\boldsymbol{Y}_i$ en el espacio de baja dimensión tenga poca influencia en la función de coste y, por tanto, no quede bien determinada (van der Maaten & Hinton, 2008). Este problema se soslaya definiendo las probabilidades p_{ij} como probabilidades condicionales simétricas, es decir:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n} \tag{3-38}$$

Esto asegura que $\sum_j p_{ij} > \frac{1}{2n}$ para todos los puntos $\boldsymbol{X}_i$ con lo cual todos tienen una contribución significativa en la función de coste. En el espacio de baja dimensión, la expresión utilizada es la que muestra la ecuación (3-37).

Para *SSNE*, la expresión diferencial de la función de coste queda de una forma más simplificada y rápida de computar que en *SNE*:

$$\frac{\partial \mathcal{J}}{\partial \boldsymbol{y}_i} = 4 \sum_j (\boldsymbol{Y}_i - \boldsymbol{Y}_j)(p_{ij} - q_{ij}) \tag{3-39}$$

Según se afirma en (van der Maaten & Hinton, 2008) basándose en ciertos experimentos, *SSNE* parece producir resultados tan buenos como *SNE* y, en algunos casos, incluso mejores.

3.4.10. Proyección Estocástica del Entorno mediante distribución t (*t-SNE*)

La Proyección Estocástica del Entorno mediante distribución t o *t-SNE* (del inglés, *t-Distributed Stochastic Neighbor Embedding*) es una técnica desarrollada a partir de *SNE* cuyo fin es aliviar los problemas que esta otra presenta. La función de coste utilizada por *t-SNE* difiere de la utilizada por *SNE* de dos maneras (van der Maaten & Hinton, 2008):

1. Utiliza una versión simétrica de la función de coste *SNE* con expresiones diferenciales más simples.
2. Utiliza una distribución *t de Student* en lugar de una gaussiana para calcular la similitud entre dos puntos en el espacio de pocas dimensiones. *t-SNE* emplea una distribución más abrupta en sus extremos en el espacio de pocas dimensiones para aliviar tanto el problema de hacinamiento como los problemas de optimización de *SNE*.

En el espacio de alta dimensión *t-SNE* convierte las distancias en probabilidades utilizando una distribución gaussiana mientras que, en el espacio de baja dimensión, emplea una distribución de probabilidad *t de Student* que es

menos suave en los extremos. Esto permite que una distancia moderada en el espacio de alta dimensión sea fielmente modelada por una distancia mucho mayor en el espacio de baja dimensión y, como consecuencia, eliminar las fuerzas de atracción no deseadas entre puntos que representan a datos moderadamente diferentes (van der Maaten & Hinton, 2008).

Así pues, *t-SNE* utiliza la misma expresión que *SSNE* para las probabilidades conjuntas en alta dimensión, mientras que en el espacio de baja dimensión la expresión de la probabilidad de elección en el entorno de vecindad para el punto $\boldsymbol{y}_i$ respecto del resto tiene la expresión:

$$q_{ij} = \frac{\left(1 + \left\|\boldsymbol{Y}_i - \boldsymbol{Y}_j\right\|^2\right)^{-1}}{\sum_{k \neq l}(1 + \|\boldsymbol{Y}_k - \boldsymbol{Y}_l\|^2)^{-1}} \tag{3-40}$$

Computacionalmente, el cálculo de la densidad de probabilidad de un punto bajo una distribución *t de Student* es más rápido que bajo una distribución normal porque no implica una exponencial, a pesar de que la distribución *t de Student* es equivalente a una mezcla infinita de curvas normales con diferentes varianzas.

La expresión diferencial de la función de coste queda del siguiente modo (van der Maaten & Hinton, 2008):

$$\frac{\partial \mathcal{J}}{\partial \boldsymbol{Y}_i} = 4 \sum_j \left(\boldsymbol{Y}_i - \boldsymbol{Y}_j\right)(p_{ij} - q_{ij}) \left(1 + \left\|\boldsymbol{Y}_i - \boldsymbol{Y}_j\right\|^2\right)^{-1} \tag{3-41}$$

t-SNE ha probado su eficacia principalmente en aplicaciones donde el objetivo es la visualización de datos (Brockmeier, Kriminger, Sanchez, & Principe, 2011; Ridgway & Ashburner, 2012), aunque también en otros contextos como, por ejemplo, la clasificación (Dupont & Ravet, 2013).

3.4.11. Proyección Lineal Local (*LLE*)

La Proyección Lineal Local o *LLE* (del inglés, *Local Linear Embedding*) es un método de reducción de la dimensionalidad no lineal que coincide con *Isomap* en la construcción de un grafo en el conjunto de los puntos representantes de los datos en alta dimensión. En contraste con *Isomap*, *LLE* intenta preservar exclusivamente

las propiedades locales en el entorno de cada punto por lo que se engloba dentro de las técnicas locales. Ello permite a *LLE* ser menos sensible a la existencia de conexiones erróneas en el grafo que *Isomap* y obtener mejores resultados al aplicarse a *manifolds* no convexos.

En *LLE*, para garantizar la preservación las propiedades locales del conjunto de datos en alta dimensión, éstos se expresan como una combinación lineal de sus vecinos más cercanos. Posteriormente, en la representación de pocas dimensiones de los datos, *LLE* intenta retener los pesos de esas combinaciones lineales lo más posible para cada uno de los datos. Por lo tanto, *LLE* encaja un hiperplano a través de los puntos $\boldsymbol{X}_i$ y sus k vecinos más próximos (parámetro solicitado en la aplicación), asumiendo de este modo que el *manifold* en el que están situados es localmente lineal. La suposición de linealidad local implica que la reconstrucción de los pesos de datos $\boldsymbol{X}_i$ es invariante respecto de traslaciones, rotaciones y cambios de escala. Debido a la invariancia de estas transformaciones, cualquier proyección lineal del hiperplano a un espacio de baja dimensionalidad conserva los pesos de reconstrucción en el espacio de baja dimensionalidad (Chen & Liu, 2011).

El método *LLE* consta de tres pasos (ver figura 3-7):

1. Para cada $\boldsymbol{X}_i$ encontrar los k vecinos más próximos (figura 3-7(a)).
2. Calcular la matriz de pesos $\mathbf{W}$ que minimice la suma de cuadrados residual en la reconstrucción de cada $\boldsymbol{X}_i$ a partir de sus vecinos (figura 3-7(b)):

$$SCR(\mathbf{W}) = \sum_{\boldsymbol{i}} \left\| \boldsymbol{X}_i - \sum_{\boldsymbol{j} \neq \boldsymbol{i}} \mathrm{W}_{ij} \boldsymbol{X}_j \right\|^2 \tag{3-42}$$

   donde $\mathrm{W}_{ij} = 0$ si $\boldsymbol{X}_j$ no es uno de los k vecinos más próximos a $\boldsymbol{X}_i$ y $\sum_j \mathrm{W}_{ij} = 1$ para cada i.

3. Encontrar las coordenadas $\boldsymbol{y}$ que minimicen el error de reconstrucción utilizando los pesos calculados en el paso 2 (figura 3-7(c)). Para ello hay que minimizar la función de coste:

$$\phi(\boldsymbol{Y}) = \sum_{\boldsymbol{i}} \left\| \boldsymbol{Y}_i - \sum_{\boldsymbol{j} \neq \boldsymbol{i}} \mathrm{W}_{ij} \boldsymbol{Y}_j \right\|^2 \tag{3-43}$$

que puede escribirse como:

$$\phi(\boldsymbol{Y}) = (\mathbf{I} - \mathbf{W}\boldsymbol{Y})^2 = \boldsymbol{Y}^T(\mathbf{I} - \mathbf{W})^T(\mathbf{I} - \mathbf{W})\boldsymbol{Y} \tag{3-44}$$

por lo que las coordenadas $\boldsymbol{Y}_i$ en el espacio de inmersión, de dimensión d, que minimizan la función de coste pueden hallarse calculando los d vectores propios correspondientes a los d valores propios más pequeños distintos de 0 de $(\mathbf{I} - \mathbf{W})^T(\mathbf{I} - \mathbf{W})$, donde $\mathbf{I}$ representa la matriz identidad (van der Maaten , Postma, & van den Herik , 2009).

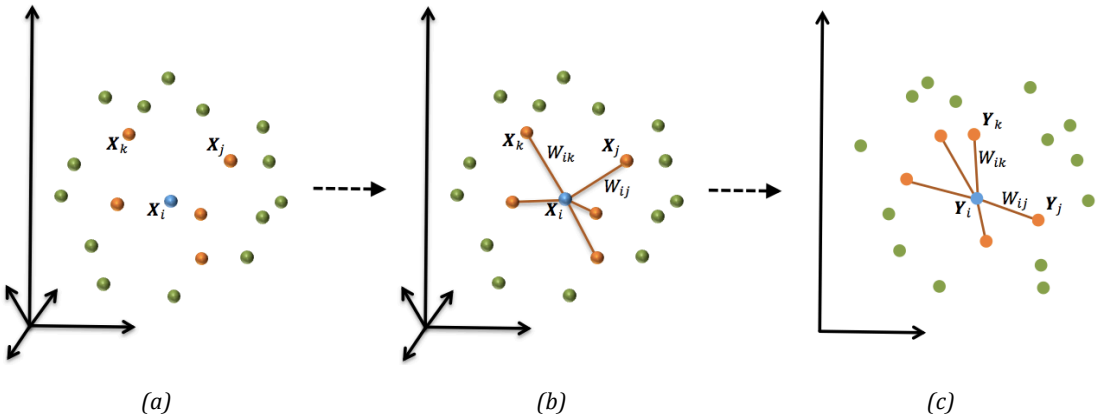


Figura 3-7: *Proceso LLE.*

LLE ha sido utilizado con éxito principalmente en aplicaciones de tratamiento de imágenes (Li, Hao, & Zhang, 2008; Xie & Mu, 2008; Morariu & Camps, 2006). Por otro lado, *LLE* puede presentar problemas en *manifolds* en los que existen zonas con poca densidad de puntos (agujeros) y tiende a colapsar los datos en un solo punto en los casos en que la dimensión de destino es demasiado baja (van der Maaten , Postma, & van den Herik , 2009).

3.5. Evaluación de la calidad en la reducción de la dimensionalidad

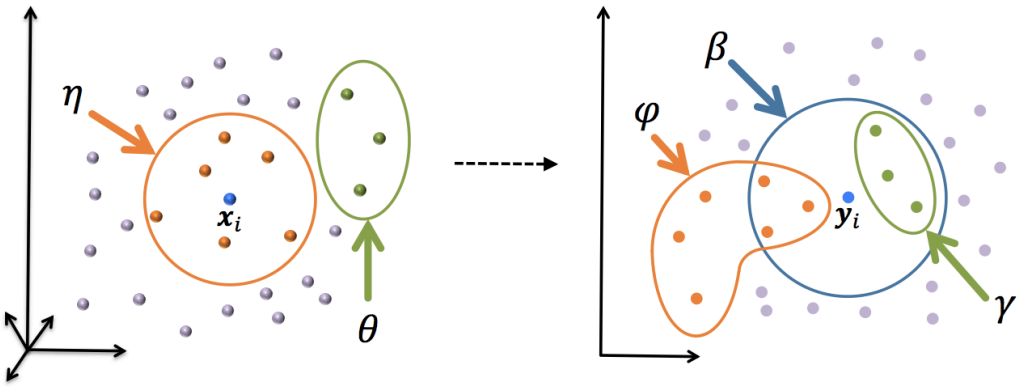
Al plantearse una reducción de la dimensionalidad con el objetivo de visualizar un conjunto de datos en dos o tres dimensiones es que se preserve lo mejor posible la estructura intrínseca de éste. Por este motivo es importante

establecer criterios que permitan objetivamente evaluar la calidad de la transformación y determinar si los resultados obtenidos son satisfactorios.

En la aplicación desarrollada en esta tesis se utilizan dos criterios de evaluación que permiten calificar la bondad de una inmersión. Los valores obtenidos al aplicar estos criterios son facilitados al usuario con lo que éste dispondrá de información objetiva acerca de lo adecuada que pueda resultar una determinada técnica de reducción de la dimensión para el problema particular que se aborde aparte de su experiencia y criterio personal. Ambos criterios fijan su atención en determinar cómo se altera en el proceso de reducción de la dimensión el entorno de vecindad. Para ello se tienen en cuenta el conjunto de los k vecinos más próximos de cada dato $\boldsymbol{X}_i$ en alta dimensión y se analizan las coincidencias con el conjunto de los k datos cuyas representaciones en baja dimensión son las más cercanas a $\boldsymbol{Y}_i$.

La notación utilizada es la siguiente (ver figura 3-8):

- $\boldsymbol{\eta}$: conjunto de los k vecinos más cercanos a $\boldsymbol{X}_i$ en el espacio original.
- $\boldsymbol{\beta}$: conjunto de los k vecinos más cercanos a $\boldsymbol{Y}_i$ en el espacio de inmersión.
- $\boldsymbol{\varphi}$: conjunto de las proyecciones de los elementos de $\boldsymbol{\eta}$.
- $\boldsymbol{\gamma}$: conjunto de los k_n elementos de $\boldsymbol{\beta}$ en el espacio de inmersión que son proyecciones de aquellos elementos que no pertenecen a $\boldsymbol{\eta}$, es decir: $\boldsymbol{\gamma} = \boldsymbol{\beta} - (\boldsymbol{\beta} \cap \boldsymbol{\varphi})$.
- $\boldsymbol{\theta}$: conjunto de los k_n elementos en el espacio original cuyas proyecciones forman $\boldsymbol{\gamma}$.



(a) Conjuntos del espacio original *(b) Conjuntos del espacio de inmersión*

Figura 3-8: *Representación gráfica de los distintos conjuntos de datos en el espacio original y de inmersión.*

Los criterios de evaluación de la calidad propuestos son el Error de Conservación de Vecindarios (ECV) y el Promedio de Vecinos Conservados (PVC) (Valencia, Daza, Acosta, & Castellanos, 2010).

3.5.1. Error de Conservación de Vecindarios (ECV)

El ECV está basado en la conservación de la geometría local; su expresión matemática es:

$$ECV = \frac{1}{2n}\sum_{i=1}^{n}\left\{\frac{1}{k}\sum_{j=1}^{k}\left(M_{(\boldsymbol{x}_i,\boldsymbol{\eta}_j)} - M_{(\boldsymbol{y}_i,\boldsymbol{\varphi}_j)}\right)^2 + \frac{1}{k_n}\sum_{j=1}^{k_n}\left(M_{(\boldsymbol{x}_i,\boldsymbol{\theta}_j)} - M_{(\boldsymbol{y}_i,\boldsymbol{\gamma}_j)}\right)^2\right\} \quad (3\text{-}45)$$

Donde M simboliza la distancia euclídea estandarizada para obtener un valor máximo igual a 1. En una inmersión ideal $ECV = 0$.

3.5.2. Promedio de Vecinos Conservados (PVC)

Se basa en calcular el número de observaciones que se conservan como parte de los vecindarios tras la inmersión. Este promedio se define como:

$$PVC = \frac{1}{n}\sum_{i=1}^{n}\frac{|\boldsymbol{\beta}_i \cap \boldsymbol{\varphi}_i|}{\boldsymbol{k}_i} \quad (3\text{-}46)$$

donde el operador $|\cdot|$ expresa el cardinal del conjunto. En una situación ideal el valor de PVC es 1 lo cual indicaría que, para todos los elementos del conjunto de entrada, sus *k* vecinos son conservados como tales tras la inmersión.

Capítulo 4

Desarrollo de una aplicación en *Processing* para la visualización eficiente de datos

Resumen

En respuesta a las necesidades expuestas en el capítulo 1 surge la idea de desarrollar una aplicación para la visualización de datos que combine sencillez y eficiencia dentro de un contexto intuitivo para el usuario.

En este capítulo se describen las principales características de esta aplicación así como su manejo y posibilidades.

4.1. Características de la aplicación *VDE*

Con el fin de abordar el problema de la visualización gráfica de datos multidimensionales se diseña e implementa la aplicación *Entorno Visual de Datos* o *VDE* (del inglés, *Visual Data Enviroment*). Como se comentó en el apartado 1.5, esta aplicación se desarrolla mediante el lenguaje de programación *Processing*© y sus

principales características son las siguientes:

- Ofrece un entorno para la representación de datos multidimensionales en un espacio de dos dimensiones utilizando técnicas de reducción de la dimensionalidad. Esto permite al usuario tener una visualización de la disposición espacial de los datos en íntima relación con la estructura original de éstos.
- Permite el acceso a la visualización de grandes cantidades de datos utilizando métodos de agrupamiento a partir de los cuales referir subconjuntos de datos similares mediante un único representante facilitando la visión general del conjunto sin pérdida de información detallada. De igual modo, admite la realización de agrupamientos en el conjunto de datos, bien a través de una clasificación previa de los mismos o a través de la selección de uno de los métodos de agrupamiento propuestos por la aplicación.
- Facilita el acceso a la información contenida en distintas formas y niveles de detalle. En referencia a los datos, el usuario puede acceder desde la información más detallada de cada uno a la más general sobre los agrupamientos que puedan realizarse de los mismos. En referencia a las variables o características que los componen, puede visualizarse el comportamiento individual de cada una de ellas y su distribución a lo largo de todo el conjunto.
- Posibilita la interacción del usuario con la visualización del conjunto de datos en distintas formas:
    - Controlando la navegación por el entorno (desplazando, ampliando y reduciendo el espacio visualizado).
    - Permitiendo el control de los distintos agrupamientos en el conjunto de datos.
    - Seleccionando y filtrando los datos de forma individual o grupal.
    - Escogiendo la forma de visualizar los datos según la característica, o variable, que se quiera destacar.
    - Interviniendo en el modo de representación gráfica del conjunto de datos mediante la elección de la técnica de reducción de la dimensionalidad.

- Proporciona información numérica sobre el conjunto de datos y sobre la representación visual con la que se está trabajando (indicadores de evaluación de la calidad en la reducción de la dimensionalidad y de validación en caso de que los datos se hallen agrupados, entre otros).

Estas características convierten a *VDE* en una herramienta de gran utilidad para la visualización eficiente de datos que ofrece unas posibilidades óptimas al usuario en su objetivo de obtener información de interés a partir de grandes conjuntos de datos multidimensionales.

A diferencia otras herramientas y técnicas utilizadas en la Representación Visual de Datos, *VDE* ofrece un entorno visual de fácil comprensión y manejo, sencillo e intuitivo, apto incluso para usuarios no experimentados. Por otro lado, las técnicas de Minería de Datos que *VDE* tiene implementadas, y que ya han sido analizadas en los capítulos 2 y 3, hacen de esta aplicación una herramienta potente y eficaz.

El resto de este capítulo se dedica a la descripción y uso de *VDE* así como a sus requisitos de funcionamiento.

4.2. Inicio y ejecución de *VDE*

Los programas implementados en lenguaje *Processing* pueden ser exportados como aplicaciones ejecutables independientes a diferentes plataformas como *Linux*© *y Windows*®. Para las últimas versiones del sistema operativo *Mac OS X*®*,* las aplicaciones exportadas de *Processing* pueden generar problemas que en el momento de la elaboración de esta tesis[1] no han sido resueltos. De hecho, en las últimas versiones de *Processing 2.2.1* (estable), *3.0a5* y *3.0a7* la opción de exportar al sistema operativo *Mac OS X* aparece inhabilitada. Ello implica que para que la aplicación pueda utilizarse correctamente en este sistema operativo es necesario hacerlo a través del propio entorno *Processing*.

Para que la aplicación pueda funcionar es necesario tener instalada la versión de *Java jdk-7* correspondiente al sistema operativo en el que se quiera ejecutar teniendo en cuenta si éste es de 32 ó 64 bits. Esta versión de *Java*® puede descargarse gratuitamente desde la página web de *Oracle*® (*www.oracle.com*).

VDE puede ser descargada desde la página web *www.uv.es/misai/VDE.htm*

[1] Junio de 2015

de forma gratuita. Una vez descargada y con la versión de *Java jdk-7* instalada, se accede a la aplicación desde la carpeta *aplicacion_VDE* que puede ser ubicada en cualquier directorio del ordenador. Dentro de esta carpeta se encuentran las subcarpetas correspondientes a los sistemas operativos *Linux* y *Windows* distinguiéndose las versiones de 32 y 64 bits. También se encuentra la subcarpeta *VDE* dentro de la cual se halla la aplicación lista para ser utilizada con *Processing* (figura 4-1).

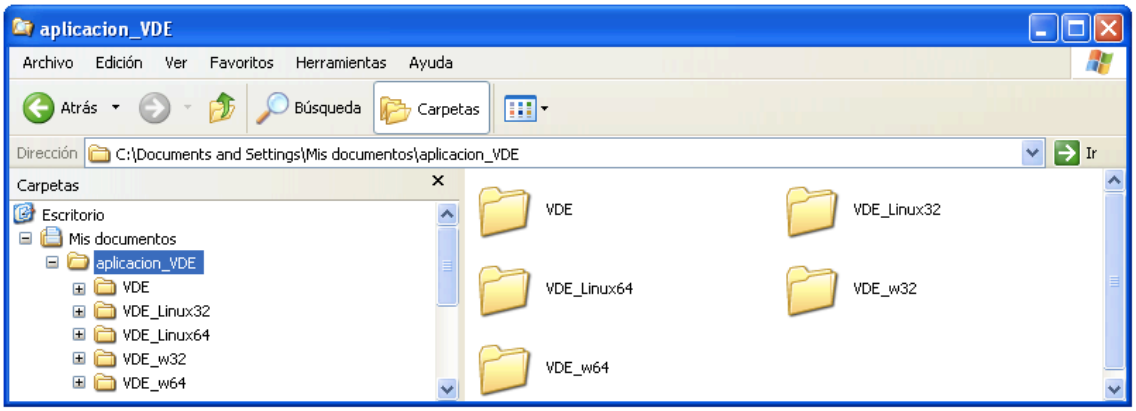


Figura 4-1: *Contenido de la carpeta aplicación_VDE.*

A continuación, se distingue el modo de ejecutar la aplicación según el sistema operativo:

1. **Sistemas operativos *Linux* y *Windows*:** Acceder a la carpeta correspondiente y pulsar dos veces con el ratón en el archivo ejecutable *VDE* (ver figura 4-2).

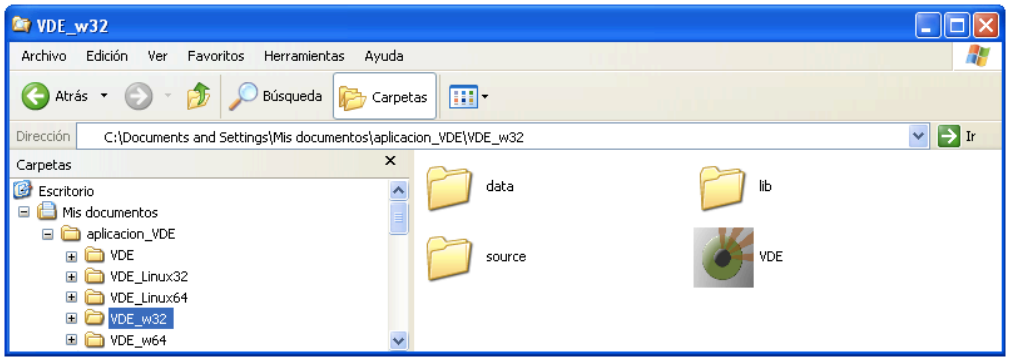


Figura 4-2: *Acceso a la aplicación para un sistema operativo Windows de 32 bits.*

2. **Sistema operativo *Mac OS X*:** Como se ha comentado anteriormente,

para utilizar la aplicación *VDE* en este sistema operativo es necesario hacerlo desde el programa *Processing* que puede obtenerse gratuitamente desde su página web (*www.processing.org*) accediendo al apartado de descargas (se recomienda la versión *2.2.1*).

Una vez instalado *Processing* éste crea un directorio con el mismo nombre en la carpeta *Documentos* (ver figura 4-3).

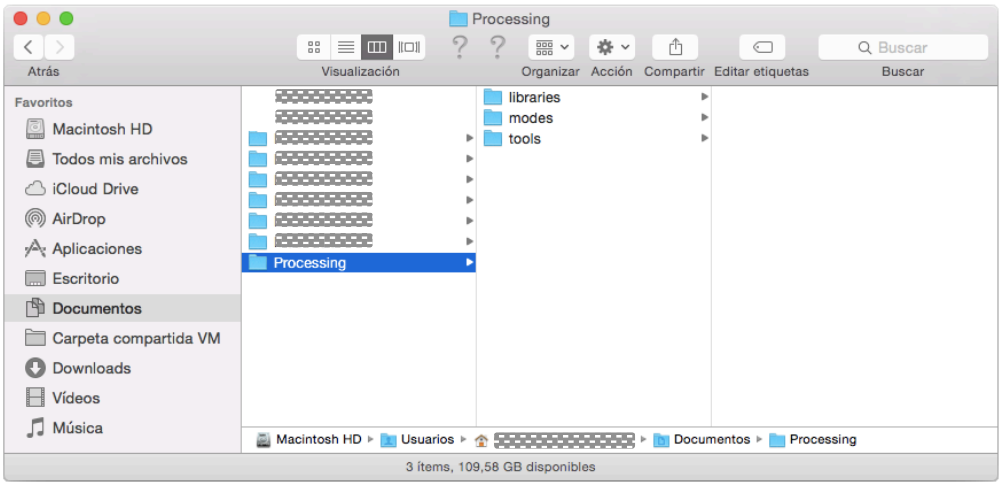


Figura 4-3: *Contenido de la carpeta creada por Processing en el directorio Documentos.*

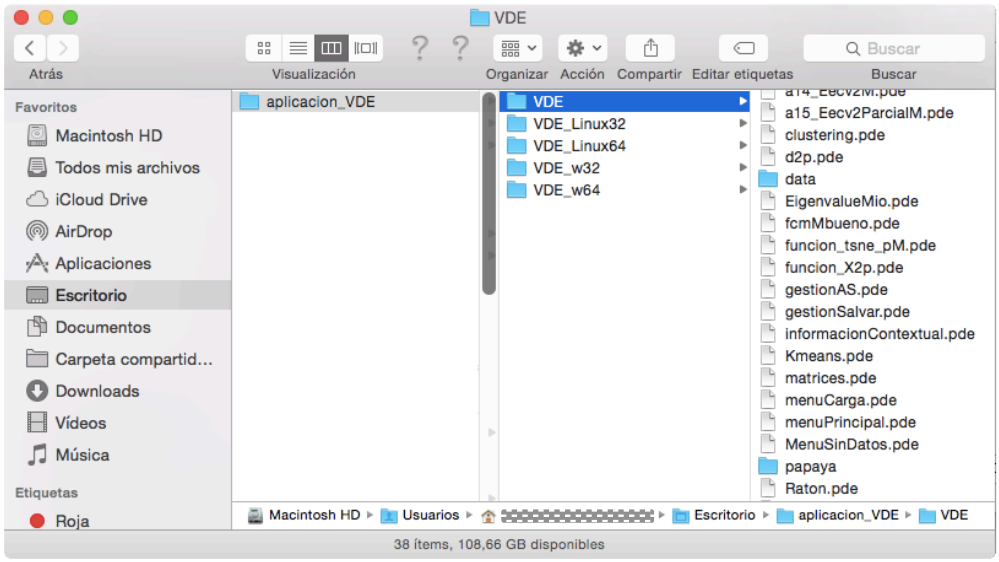


Figura 4-4: *Situación de los directorios papaya y data .*

Para que la aplicación pueda ser utilizada directamente por *Processing* es necesario copiar en el directorio *libraries* (ver figura 4-3) la

carpeta *papaya* que se encuentra en el subdirectorio *VDE* de la aplicación (ver figura 4-4). Una vez copiado este subdirectorio dentro de la carpeta creada por *Processing* queda como se muestra en la figura 4-5 y la aplicación está lista para ser utilizada.

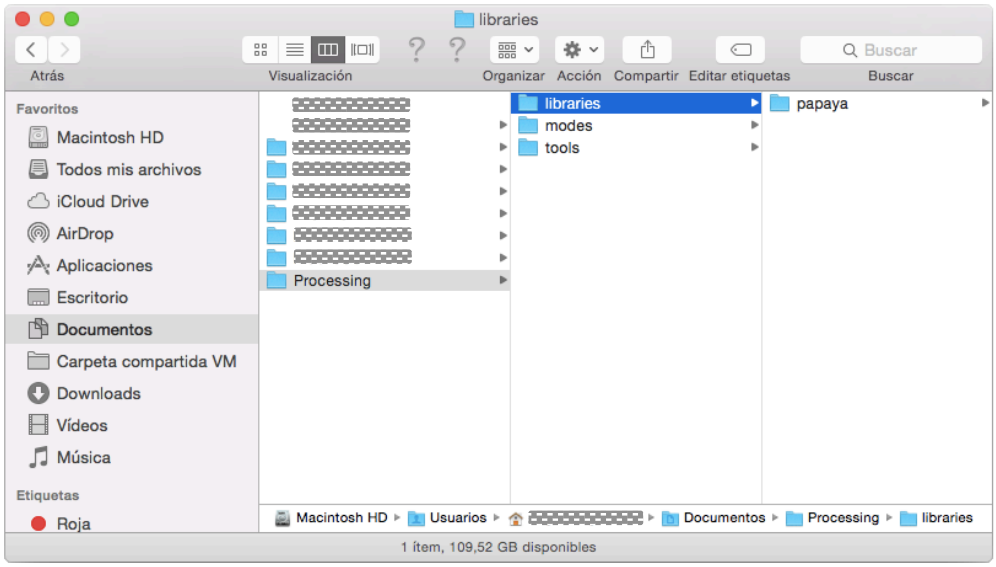


Figura 4-5: *Situación del directorio papaya para su utilización por Processing.*

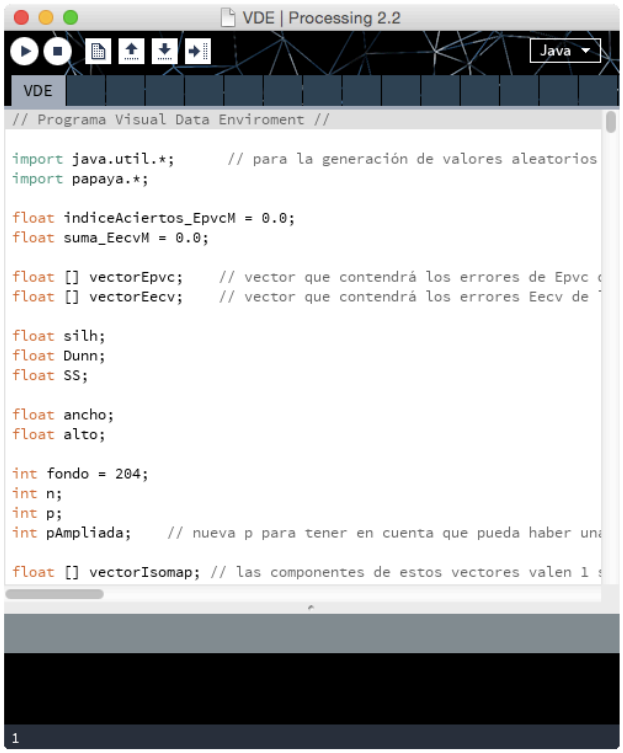


Figura 4-6: *Aplicación VDE en el entorno Processing lista para comenzar presionando el botón*  .

El paso final es comenzar la aplicación desde el entorno *Processing*. Para ello bastará presionar dos veces con el ratón sobre cualquier archivo con extensión *.pde* de la carpeta *VDE* o bien seleccionarlo desde el propio entorno *Processing*. El resultado es el que se muestra en la figura 4-6. Para poder empezar con la aplicación basta con pulsar sobre el botón  situado en la parte superior izquierda.

Una vez en funcionamiento la aplicación, ésta es idéntica independientemente del sistema operativo utilizado. Su aspecto inicial se muestra en la figura 4-7.

Finalmente, queda añadir que el modo de inicio de *VDE* descrito para el sistema operativo *Mac OS X* puede llevarse a cabo de forma análoga para el resto de sistemas operativos.

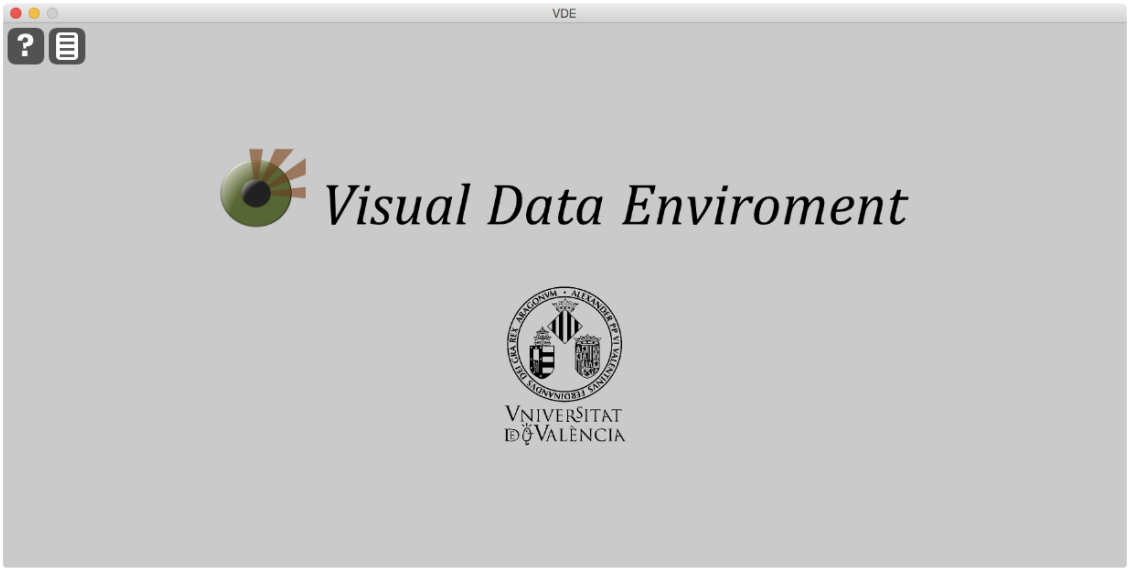


Figura 4-7: *Pantalla inicial de VDE.*

4.3. Introducción de datos

La introducción de los datos con los que la aplicación *VDE* ha de trabajar se realiza a través de archivos de texto de extensión *.txt* o *.csv*. *VDE* funciona con dos tipos de archivos de texto: aquellos elaborados por el usuario y los que genera la propia aplicación.

Previamente a la elaboración del archivo por parte del usuario, es importante señalar que si los datos están ya clasificados puede facilitarse esta información a *VDE*. Para ello, se ha de enumerar cada uno de los grupos de clasificación mediante números enteros a partir de uno. A continuación, se ha de crear una nueva variable que se denominará *clase* cuyos valores para cada uno de los datos será el número entero correspondiente al grupo al que pertenecen.

Los archivos elaborados por el usuario han de tener una estructura concreta para ser utilizados por *VDE*:

1. La primera fila debe contener el nombre de cada una de las variables de que constan los datos separados mediante tabulación para el caso de archivos del tipo *.txt* o por punto y coma en el caso de archivos con formato *.csv*. Como la primera columna del archivo se corresponde con el nombre o etiqueta de los datos, el primer elemento de que consta esta fila es ignorado por la aplicación. En caso de que los datos estén clasificados, se podrá añadir la variable *clase* al final de todas las variables.
2. Cada una de las siguientes filas se corresponde con cada uno de los datos que queremos introducir siguiendo la misma estructura que la descrita para las variables. El primer elemento de la fila se corresponde con el nombre o etiqueta del dato y los siguientes con el valor que éste adquiere en cada una de las variables enumeradas en la primera fila, todos ellos separados por tabulación en el caso de archivos con extensión *.txt* o mediante punto y coma para archivos de extensión *.csv*. Para la introducción de valores decimales se debe utilizar el símbolo del punto como separación entre la parte entera y la decimal. Si los datos están previamente clasificados, se utilizará la última columna para expresar la clase a la que el dato pertenece utilizando el número entero correspondiente como se comentó anteriormente.
3. El nombre del archivo no puede tener el formato *dataVDE#.txt* donde *#* es un número natural ya que este tipo de archivos son los que *VDE* genera en el proceso de salvar datos y el programa daría error en el momento de cargarse (ver apartado 4.5.3.6.).

Todos los archivos de datos con los que trabaja *VDE* han de situarse para poder ser adquiridos por la aplicación en el directorio *data* situado en la carpeta *aplicación_VDE* (ver figuras 4-2 y 4-4).

4.4. Modos de funcionamiento

Como ya se comentó en los capítulos 1 y 2, uno de los principales problemas con respecto a la Representación Visual de Datos es trabajar con grandes cantidades de datos. Conjuntos muy numerosos provocan que su representación gráfica no sea todo lo clara que sería deseable. En estos casos, la obtención de información se convierte en una tarea complicada para el análisis visual.

VDE está diseñado para representar conjuntos con un máximo de 700 datos. En caso de que la cantidad de datos sea superior a esa cifra, *VDE* realiza un análisis de agrupamiento mediante el cual clasifica los datos en 700 grupos. Esta labor se realiza a partir de los métodos de agrupamiento implementados escogiendo entre ellos aquel que obtenga mejores resultados en la validación (ver capítulo 2). A partir de ese momento la representación visual se realiza no de los datos originales sino de los centroides de cada uno de los grupos obtenidos (denominados representantes).

Por tanto, *VDE* contempla dos modos de funcionamiento distinguiendo entre que la cantidad de datos en el conjunto analizado sea inferior a 700 (modo de funcionamiento normal) o que sea superior (modo de funcionamiento agrupado). Aunque el procedimiento de obtención de información en *VDE* y su manejo es el mismo para ambos modos, la información mostrada en cada caso puede ser diferente. En el siguiente apartado se explica el manejo de la aplicación en modo de funcionamiento normal y en el apartado 4.6. se explica el modo de funcionamiento agrupado.

4.5. Manejo de la aplicación. Modo de funcionamiento normal

Para ilustrar el manejo de la aplicación en modo de funcionamiento normal se utilizan los archivos *paises_clase.txt* y *paises.txt* cuyos datos han sido extraídos de (Baillo & Grané, 2008). Estos archivos están incluidos en el directorio *data* de la aplicación.

Los datos que aparecen en *paises_clase.txt* se muestran en la tabla 4-1 y se corresponden con 11 indicadores económicos y sociales de 96 países más un indicador del continente al que pertenece cada país (datos clasificados). El archivo *paises.txt* contiene los mismos datos pero sin la información sobre el continente al que pertenece cada país (datos sin clasificar).

Tabla 4-1: *Indicadores económicos y sociales sobre países del mundo (Baillo & Grané, 2008). Datos correspondientes al archivo paises_clase.txt.*

Países	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
Albania	1	30	41	2199	3903	12	94	53	0	341	1.2	4
Angola	3	124	46	4422	955	6	57	19	0.7	89	0.5	1
Arabia Saudí	4.3	21	13	133540	91019	96	497	1	0	4566	13.1	3
Argelia	2.5	34	24	44609	19883	42	180	2	0.8	906	3	1
Argentina	1.3	22	31	278431	65962	160	1043	22	0.1	1504	3.5	2
Australia	1.4	6	43	337909	167155	510	933	19	0	5341	15.3	5
Austria	0.6	6	41	216547	53259	465	304	47	-0.4	3301	7.2	4
Bangladesh	2	79	42	28599	9891	2	220	6	4.1	64	0.2	3
Bélgica	0.3	8	40	250710	72236	457	917	20	-0.3	5120	10.1	4
Benín	3	95	48	2034	6	5	26	45	1.3	20	0.1	1
Bielorrusia	0.4	13	49	21356	31397	190	295	31	-0.4	2392	9.9	4
Bolivia	2.3	69	37	5905	2824	35	201	45	1.2	373	1	2
Brasil	1.6	44	35	579787	260682	75	246	66	0.6	718	1.4	2
Bulgaria	-0.6	15	48	11225	381333	335	1544	33	-0.2	2438	6.4	4
Camerún	2.9	56	38	8615	2740	4	38	44	0.6	103	0.2	1
Canadá	1.3	6	45	573695	554227	590	1602	49	-1.1	7854	14.4	2
Colombia	1.8	26	37	70263	43354	100	174	52	0.7	622	1.8	2
Congo	3.1	90	43	1784	435	8	20	58	0.2	331	1.6	1
Corea del Norte	1.8	26	45	12870	38000	47	687	74	0	1129	11.2	3
Corea del Sur	0.9	10	40	435137	164993	415	632	66	0.1	2982	6.6	3
Costa de Marfil	3.4	86	33	9248	2305	8	66	34	1	103	0.5	1
Costa Rica	2.5	13	30	8884	4772	164	780	28	3	558	1.2	2
Cuba	0.9	9	38	7150	10982	32	870	16	1	923	2.6	2
Chile	1.6	12	32	59151	25276	132	1626	12	-0.1	1012	2.6	2
China	1.3	34	45	744890	928083	34	461	13	0.7	664	2.3	3
Dinamarca	0.2	6	46	156027	40097	613	233	12	0	3977	10.4	4
R. Dominicana	2	37	29	11390	6182	79	446	22	2.9	337	1.4	2
Ecuador	2.3	36	26	15997	8256	61	581	43	1.8	565	1.8	2
Egipto	2.2	56	29	45507	51947	46	956	0	0	600	1.5	1
El Salvador	1.8	36	34	9057	3211	53	245	6	2.3	370	0.7	2
E. Árabes Unidos	5.8	16	13	42806	18870	283	884	0	0	10531	33.9	3
Eslovaquia	0.3	11	48	15848	24740	208	337	38	0.1	3243	7	4
España	0.2	7	36	532347	161654	385	781	51	0	2458	5.7	4
Etiopía	2.6	112	41	5722	1293	2	51	13	0.3	22	0.1	1
Filipinas	2.3	39	37	71865	27062	21	686	26	3.4	316	0.8	3
Finlandia	0.4	5	48	105174	65546	550	440	77	0	5997	8.2	4
Francia	0.5	6	44	1451951	476200	558	665	25	-0.1	4042	6.3	4
Gabón	2.9	89	44	3759	993	30	57	71	0.6	652	5.5	1
Ghana	3	73	51	6719	6115	4	35	42	1.4	93	0.2	1
Grecia	0.5	8	36	85885	40623	493	523	47	0	2260	7.2	4
Guatemala	2.9	44	26	14255	3161	27	139	39	1.8	210	0.6	2
Haití	2	72	43	1777	362	8	7	1	5.1	29	0.1	2
Países Bajos	0.6	6	40	371039	79647	525	518	10	-0.3	4580	9.2	4
Honduras	3	45	30	3566	2672	29	294	41	2.2	204	0.6	2
Hungría	-0.3	11	44	42129	33486	185	661	18	-0.5	2383	5.8	4
India	1.9	68	32	319660	386500	13	612	17	0.6	248	0.9	3
Irak	2.7	108	18	24600	27060	33	4575	4	0.1	1213	3.4	3
Irán	3.2	45	24	113400	79128	79	1362	11	0	1505	4	3
Irlanda	0.1	6	33	52765	17105	365	233	6	-1.2	3137	8.7	4
Islandia	1.1	4	44	6686	4780	555	636	1	0	7932	6.8	4
Indonesia	1.7	51	40	190105	53414	17	96	60	1.1	366	1	3
Israel	2.7	8	40	87875	32781	418	408	6	-0.3	2717	8.1	3
Jordania	4.7	31	21	6354	5076	73	173	1	-1	1067	3	3
Kenia	2.9	58	46	7583	3539	9	87	2	0.6	110	0.2	1
Kuwait	-0.3	11	28	28941	22798	230	525	0	0	8622	11.2	3
Líbano	2.3	32	28	10673	5184	82	271	8	0.6	964	2.9	3
Libia	3.6	61	21	23400	17800	59	880	0	-1.4	2499	8.1	1
Malasia	2.5	12	37	78321	39093	166	768	54	2.1	1699	3.8	3
Marruecos	2	55	35	29545	11100	43	427	20	-1.4	327	1.1	1
México	2.1	33	31	304596	147926	96	899	25	1.3	1561	3.8	2
Mozambique	1.8	113	48	1353	490	3	55	22	0.8	40	0.1	1
Birmania	1.8	83	43	35840	3500	3	101	44	1.3	49	0.1	3
Nepal	2.5	91	40	4391	927	4	150	37	1	28	0.1	3
Nicaragua	3.1	46	36	1659	1688	23	367	50	1.9	300	0.6	2
Nigeria	2.9	80	36	28411	15530	4	41	17	0.7	162	0.9	1
Noruega	0.5	5	46	136077	113488	556	488	31	-1.4	5318	14.1	4
Nueva Zelanda	1	7	44	51655	35135	479	589	28	0	4245	7.6	5
Omán	4.5	18	15	10578	6187	77	564	19	0	2392	5.3	3
Pakistán	3	90	26	59991	58529	16	2053	2	3.5	254	0.6	3
Panamá	1.9	23	34	7253	3380	114	754	42	1.9	618	1.7	2
Paraguay	2.7	41	29	8158	36415	31	109	32	2.8	299	0.6	2
Perú	2.1	47	29	55019	15563	47	300	53	0.4	367	1	2
Polonia	0.4	14	46	107829	135347	148	321	28	-0.1	2401	8.9	4
Portugal	-0.1	7	43	96829	31380	361	739	34	-0.5	1827	4.8	4
Reino Unido	0.3	6	43	1094734	325383	502	205	10	-1.1	3732	9.8	4
Rep. Checa	0	8	47	39990	58705	236	266	34	0	3868	13.1	4
Rumanía	0	23	44	33488	55136	131	1134	27	0	1733	5.4	4
Senegal	2.8	62	42	5070	1002	10	202	39	0.7	97	0.4	1
Singapur	1.8	4	38	79831	20046	478	84	7	2.3	8103	17.7	3
Siria	3.1	32	26	15780	15186	63	435	4	-4.3	997	3.3	3
Sri Lanka	1.3	16	35	12616	4387	11	503	27	1.4	97	0.3	3
Sudán	2.2	77	28	7510	1333	3	633	18	1.1	66	0.1	1
Suecia	0.6	4	48	209720	142895	681	341	68	0	5723	6.6	4
Suiza	0.8	6	40	286014	65724	613	173	30	-0.6	3629	6.4	4
Suráfrica	2.3	50	37	130918	189316	95	359	4	-0.8	2146	7.5	1
Tailandia	1.3	35	46	159630	71177	59	602	25	3.5	769	2	3
Tanzania	3.1	82	49	3703	1913	3	40	38	1.2	34	0.1	1
Túnez	2.1	39	30	16369	6714	58	381	4	-1.9	595	1.6	1
Turquía	1.9	48	35	169452	78322	212	585	26	0	957	2.5	4
Ucrania	0.1	15	49	84084	202995	157	673	16	-0.3	3180	11.7	4
Uruguay	0.6	18	40	16458	7617	196	241	4	-0.6	629	1.6	2
Venezuela	2.4	23	33	65382	73116	111	382	52	1.2	2186	5.7	2
Vietnam	2.2	41	49	17634	12270	11	414	26	1.5	101	0.3	3
Yemen	4.2	100	29	4044	2159	12	335	8	0	206	0.7	3
Zambia	2.6	109	45	3605	7785	8	186	43	1.1	149	0.3	1
Zimbabue	2.8	55	44	5933	7334	14	136	23	0.7	438	1.8	1

Las variables observadas son: X_1 = Tasa anual de crecimiento de la población, X_2 = Tasa de mortalidad infantil por cada 1000 nacidos vivos, X_3 = Porcentaje de mujeres en la población activa, X_4 = PNB en 1995 (en millones de dólares), X_5 = Producción de electricidad (en millones de kW/h), X_6 = Líneas telefónicas por cada 1000 habitantes, X_7 = Consumo de agua *per cápita*, X_8 = Proporción de la superficie del país cubierta por bosques, X_9 = Proporción de deforestación anual, X_{10} = Consumo de energía *per cápita* y X_{11} = Emisión de CO_2 *per cápita*. La última columna, X_{12}, es la variable *clase* que indica la clasificación de los países según el continente al que pertenecen con el código: 1 = África, 2 = América, 3 = Asia, 4 = Europa, 5 = Oceanía.

En los siguientes subapartados se explica cómo opera la aplicación *VDE* en modo de funcionamiento normal.

4.5.1. Aplicación sin datos disponibles

Al comenzar la aplicación no se dispone de datos que representar por lo que ésta es la primera acción a realizar. La pantalla que aparece es la que se muestra en la figura 4-7. En ella únicamente hay dos botones en la esquina superior izquierda que pueden seleccionarse presionando con el ratón:

- El botón  despliega una sencilla ayuda acerca de cómo empezar con la aplicación.
- El botón  abre un menú en el cual puede seleccionarse salir de la aplicación o cargar un archivo. Seleccionando esta última opción, se despliega un listado con todos los archivos de extensión *.txt* y *.csv* en el directorio *data* de *VDE* (ver figura 4-8).

Figura 4-8: *Detalle de los archivos con extensión .txt contenidos en el directorio data de VDE.*

Desde este listado puede seleccionarse el archivo de datos a cargar en *VDE*. Si el archivo que se desea emplear no tiene el formato correcto la aplicación avisará de esta circunstancia mediante un cuadro informativo y el fichero no podrá ser adquirido por la aplicación para su análisis (ver apartado 4.3.).

Una vez escogido el archivo, si éste ha sido generado por *VDE* los datos son cargados instantáneamente. De no ser así *VDE* solicitará se le facilite el valor de dos parámetros:

- ***Vecindario*:** Este apunte hace referencia al número de vecinos de cada dato, es decir, la cantidad de elementos a tener en cuenta en su entorno para ser utilizado por las técnicas de reducción de la dimensionalidad que lo requieren (ver capítulo 3). En caso de no facilitarse, la aplicación tomará por defecto el valor 15.
- ***Difusión*:** Necesario para el método de agrupamiento *FCM* como ya se explicó en el capítulo 2. En caso de no proporcionarse, *VDE* dará un valor de 2 a este parámetro.

El modo en que estos parámetros se solicitan se realiza a partir de dos cuadros en los cuales el usuario introduce los valores. En ambos casos, los valores introducidos se aceptan presionando la tecla *Intro*.

Una vez se tienen los valores de los parámetros, la aplicación comienza un proceso de análisis en el cual se aplican las técnicas de reducción de la dimensionalidad y los métodos de agrupamiento (en caso de ser necesario). Este proceso puede prolongarse durante algunos minutos en función del número de datos. Durante este tiempo *VDE* muestra en pantalla el progreso de los cálculos (ver figura 4-9).

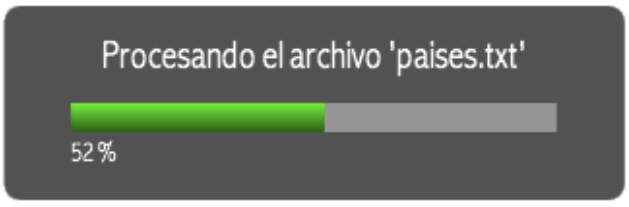


Figura 4-9: *Indicación del proceso de carga del archivo de datos paises.txt.*

4.5.2. Aplicación con datos disponibles. Representación

Cuando *VDE* dispone de datos, éstos son representados gráficamente en dos dimensiones. La figura 4-10 muestra la pantalla de *VDE* con los datos del archivo *paises.txt* para un valor de 12 en el parámetro *vecindario* y de 2 en el de *difusión*.

Cada dato está representado en el plano mediante un círculo rojo en una posición acorde a los valores obtenidos al aplicar la técnica de reducción de la dimensionalidad ACP. En la esquina inferior izquierda se muestra información relativa al conjunto de datos tales como el archivo desde el que han sido adquiridos, el número de datos (denominados elementos en la aplicación), número de variables y los valores de PVC y ECV calculados para ACP (definidos en el capítulo 3).

En caso de haber cargado el archivo *paises_clase.txt*, la pantalla mostraría exactamente la misma información ya que la columna *clase* de este archivo no es tenida en cuenta por *VDE* como una variable más sino solamente para informar acerca de un agrupamiento preestablecido.

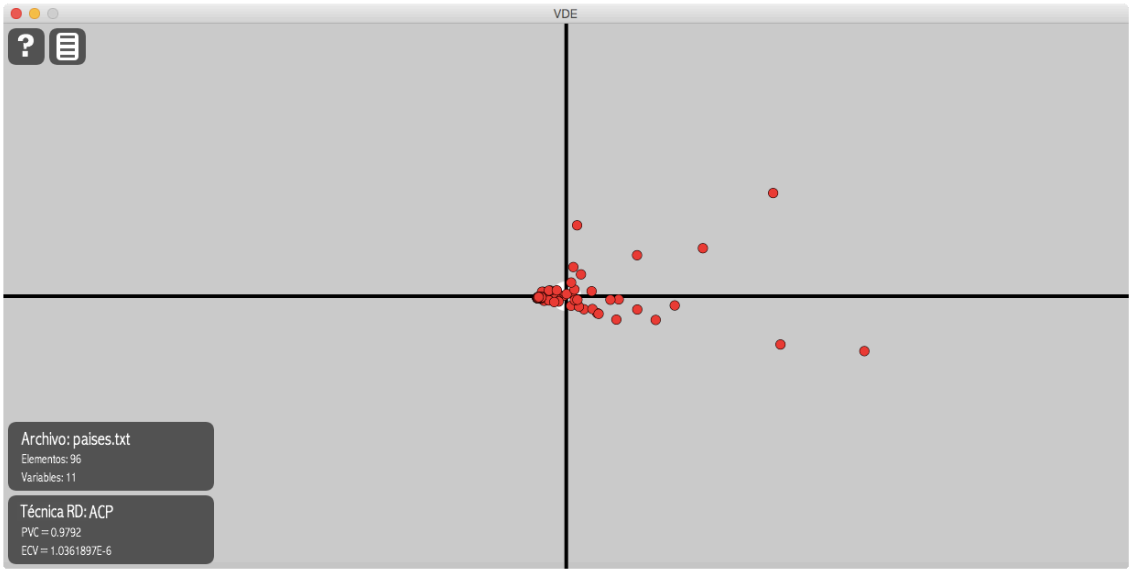
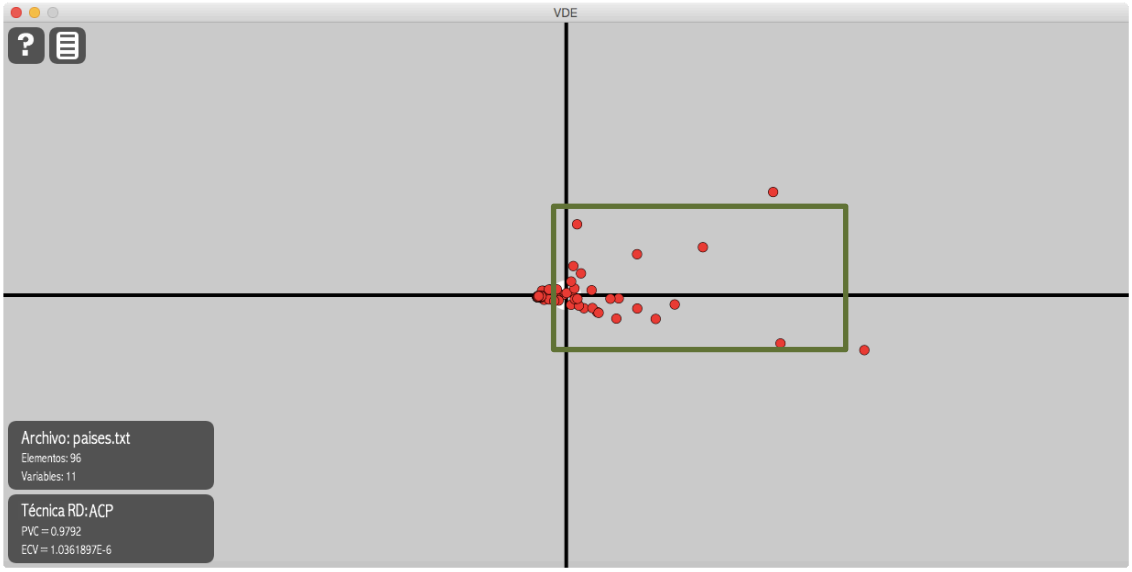


Figura 4-10: *Pantalla de VDE con los datos del archivo de datos paises.txt.*

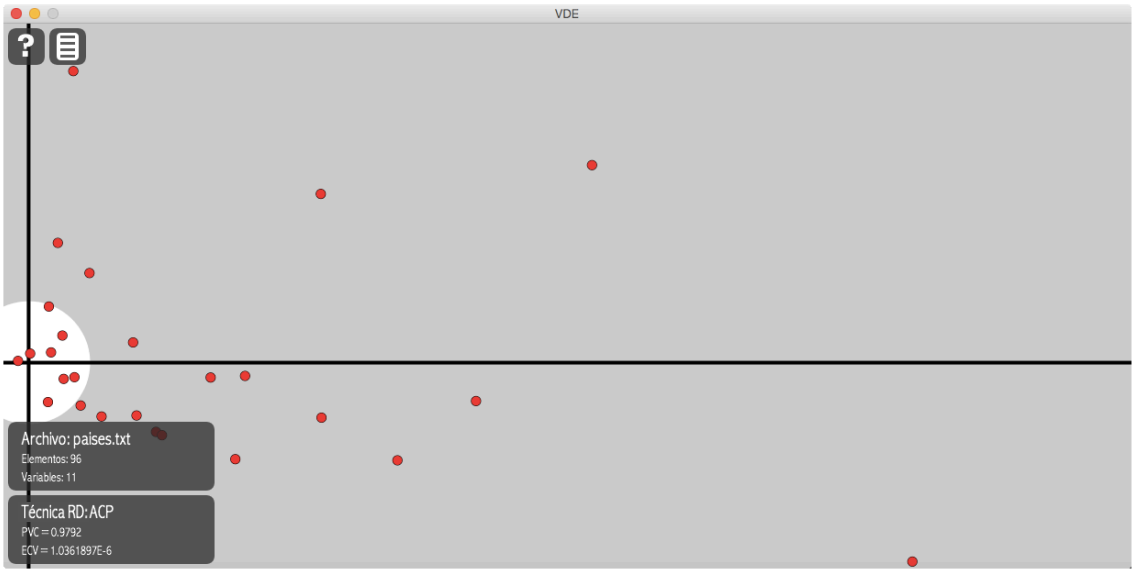
La visualización de los datos puede ser adaptada a las necesidades del usuario mediante desplazamiento, ampliación y reducción de la representación gráfica mostrada. Para realizar desplazamientos basta con mantener presionado el botón izquierdo del ratón en cualquier parte de la pantalla y realizar con él el movimiento deseado. Para ampliar y reducir la zona representada en pantalla se

utilizan las teclas de dirección arriba y abajo respectivamente.

Cuando se realiza una ampliación, o reducción, las distancias entre los elementos en pantalla aumentan o disminuyen manteniéndose constante el tamaño del círculo que representa a cada dato. La figura 4-11(b) muestra la ampliación de la zona recuadrada en verde de la figura 4-11(a).



(a)



(b)

Figura 4-11: *La zona delimitada por el recuadro verde que aparece en (a) se muestra ampliada en (b). En esta ampliación se puede observar cómo el tamaño del círculo que representa cada dato no varía con respecto a la representación sin ampliar.*

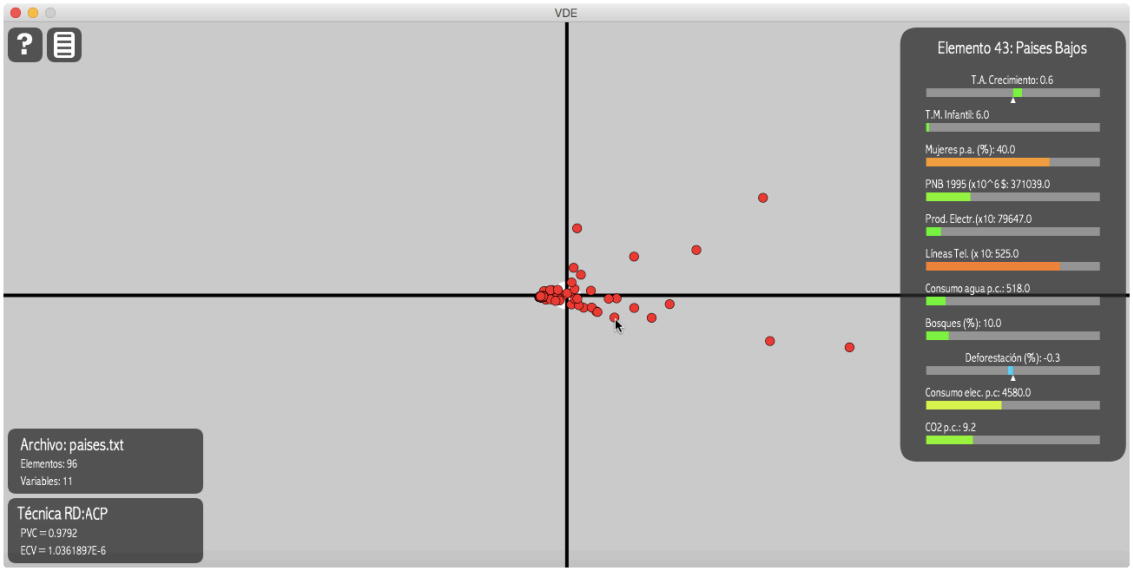


Figura 4-12: *Cuadro de información del elemento señalado por el puntero del ratón.*

Al situar el puntero del ratón sobre alguno de los puntos, *VDE* muestra en pantalla un cuadro con la información correspondiente al dato al que representa (ver figura 4-12). En este recuadro se indica en primer lugar el orden que el dato (denominado elemento) ocupa dentro de la tabla en el archivo cargado así como su nombre o etiqueta identificativa. A continuación se dispone un listado con los valores que toma el elemento en cada una de las variables de forma numérica y gráfica mediante barras. Es especialmente importante esta información gráfica ya que sitúa el valor de cada una de las variables en este dato dentro del rango total de variación de cada una en todo el conjunto de datos. Cada barra representa una amplitud que depende de si la variable a la que representa toma valores negativos en algún dato del conjunto o no:

- Si la variable solamente toma valores positivos, su barra representa un rango de variación que va desde su valor mínimo hasta su valor máximo en el conjunto de datos (ver figura 4-13(a)).
- Si, por el contrario, la variable toma valores negativos en algunos datos, *VDE* calcula el valor absoluto de todas sus anotaciones y halla el máximo valor en el conjunto de datos. El rango de variación se sitúa entonces entre este valor en negativo y el mismo valor en positivo. En este caso, la aplicación facilita la referencia del cero señalándose su posición mediante un triángulo que sitúa en la parte inferior de la barra (ver figura 4-13(b)).

(a) *(b)*

Figura 4-13: *Detalles del cuadro de información mostrado en la figura 4-12. En (a) se muestra la representación de una variable que solo toma valores positivos en el conjunto de datos y en (b) la de una variable que puede tomar tanto valores positivos como negativos.*

Si el cursor está situado sobre una zona en la que hay de dos a cuatro puntos, *VDE* advierte de esta situación mostrando un cuadro con los números de los elementos que hay en dicha zona. Al mismo tiempo solicita introducir mediante el teclado el número de uno de ellos por si se desea mostrar su estadística (ver figura 4-14). Si la cantidad de puntos en la zona en la que está situado el puntero del ratón es superior a cuatro, *VDE* mostrará un cuadro informando sobre la cantidad de elementos que se hallan en dicha zona y recomendará ampliar la misma para poderlos distinguir.

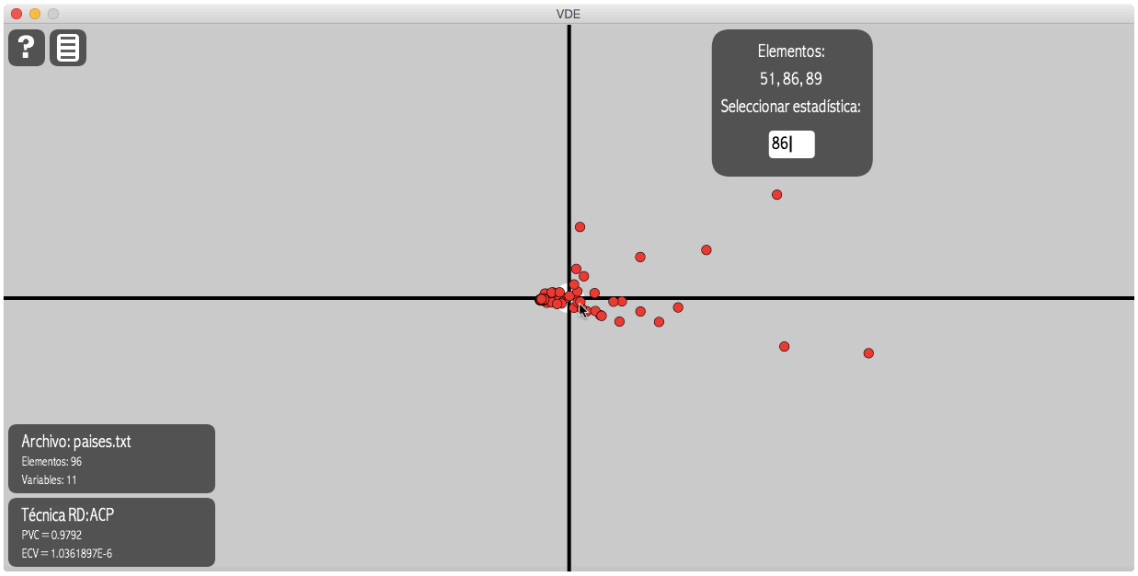


Figura 4-14: *En la zona donde se halla situado el puntero del ratón están situados los tres elementos que se señalan en el cuadro superior: el número 51, el 86 y el 89. Introduciendo mediante el teclado uno de ellos y aceptando con la tecla Intro, VDE mostrará a la derecha el cuadro correspondiente a su información.*

4.5.3. Aplicación con datos disponibles. Menú

Al igual que al comienzo de la aplicación cuando *VDE* no dispone de datos, en la esquina superior izquierda se encuentran los botones de ayuda y menú. El

botón de ayuda despliega una breve descripción de las opciones que se presentan al seleccionar el botón de menú mientras que en éste las posibilidades son: *Cargar archivo*, *Vista por variables*, *Buscar un elemento*, *Técnica RD* (reducción de la dimensionalidad), *Agrupamiento*, *Autozoom*, *Salvar Archivo* y *Salir*. La figura 4-15 muestra este menú desplegado.

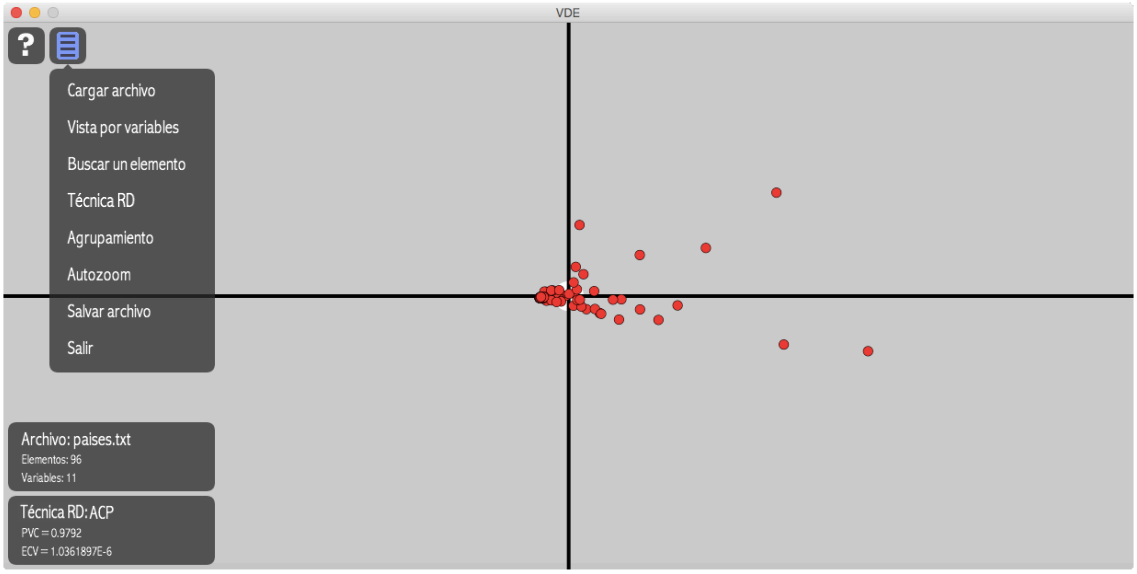


Figura 4-15: *Menú de la aplicación VDE.*

Las opciones *Cargar archivo* y *Salir* son las mismas que las ya comentadas en el menú de *VDE* cuando no dispone de datos. El resto se describe en los siguientes subapartados.

4.5.3.1. Vista por variables

Seleccionando *Vista por variables, VDE* muestra un submenú donde aparece un listado con el nombre de todas las variables observadas en el conjunto de datos más la opción *Vista general* (ver figura 4-16). En esta imagen, el archivo de datos es *paises.txt* cuyos datos no están previamente clasificados; en el caso de que el archivo cargado fuera *paises_clase.txt* el listado sería el mismo ya que la columna *clase* de este archivo no se corresponde con una variable medida en el conjunto de datos sino con un agrupamiento previo.

Seleccionar la opción *Vista general* dentro de este submenú permite visualizar los datos en pantalla tal y como se tienen en un principio, es decir, cada

elemento gráfico correspondiente a un dato se representa mediante un círculo de color rojo y el tamaño es el mismo para todos ellos.

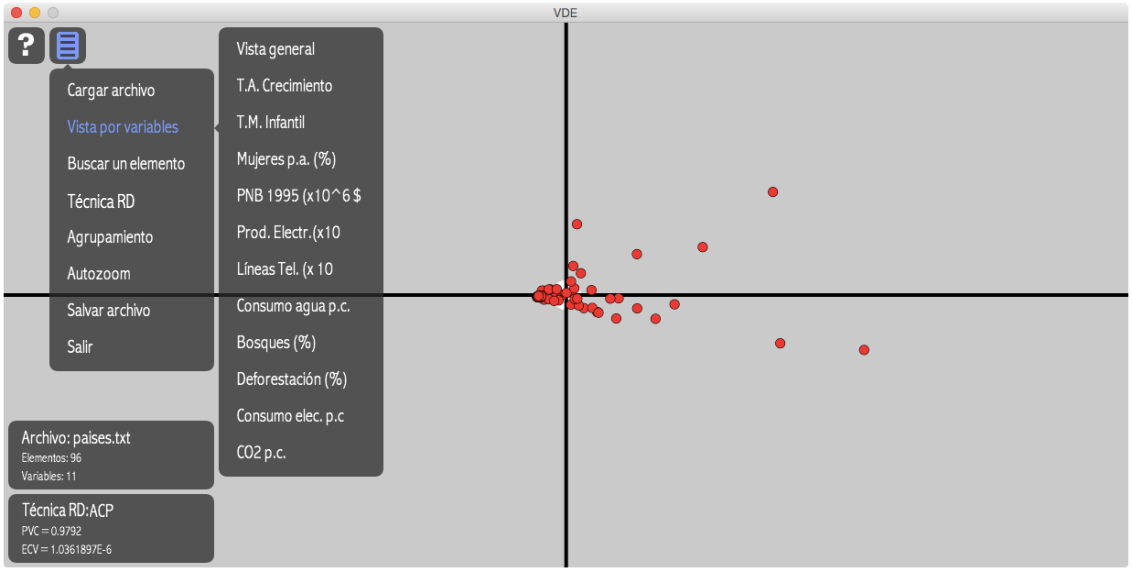


Figura 4-16: *Submenú Vista por variables.*

Por otro lado, seleccionar cualquiera de las variables que aparecen en este submenú permite visualizar la representación gráfica de los datos de forma distinta: en este caso cada elemento tendrá un color y un tamaño acorde al valor del dato al que hace referencia según la variable seleccionada. En la figura 4-17 se ha elegido la variable *Líneas Telefónicas* y se ha realizado una ampliación de la pantalla para una mejor visualización. En este caso, *VDE* presenta un cuadro en la parte central superior donde se informa de la variable seleccionada y del rango de variación de la misma en el conjunto de datos. Dado que la variable toma valores en el conjunto de datos entre 2 y 681, en el cuadro aparece una barra de colores que va desde el verde para el valor 2 al rojo para el valor 681. Cada elemento en pantalla se visualiza ahora según un color y un tamaño proporcional a su valor en esta variable. Así, elementos que presenten un valor pequeño en esta variable, se visualizarán mediante un círculo de menor tamaño y un color verde o cercano al verde mientras que aquellos elementos que tengan un valor grande en esta variable aparecerán como círculos de mayor tamaño y un color anaranjado o cercano al rojo.

Si la variable seleccionada toma valores negativos y positivos, el rango de variación va, como ya se comentó en el apartado 4.5.2, desde el máximo del valor absoluto tomado en negativo y este mismo valor en positivo. En la figura 4-18 se muestra la representación del conjunto de datos seleccionando la variable

Deforestación. En esta figura se puede observar que el rango de variación para esta variable va desde -5,1 a 5,1. Los valores positivos se representan en una gama de colores desde el verde para valores cercanos a cero hasta el rojo para valores grandes, mientras que, para valores negativos, la escala de colores va desde el azul claro para valores cercanos a cero hasta el púrpura para valores más alejados. En este caso, el tamaño de representación de los datos es grande para valores alejados de cero y pequeño para valores cercanos independientemente de su signo.

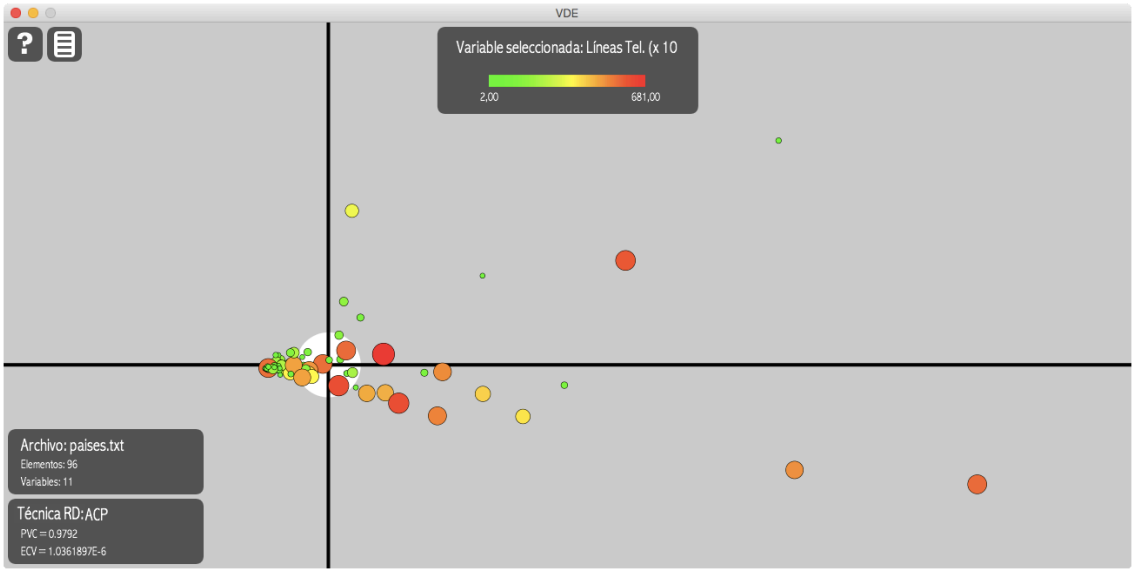


Figura 4-17: *Representación gráfica del conjunto de datos del archivo paises.txt al seleccionar la visualización según la variable Líneas Telefónicas.*

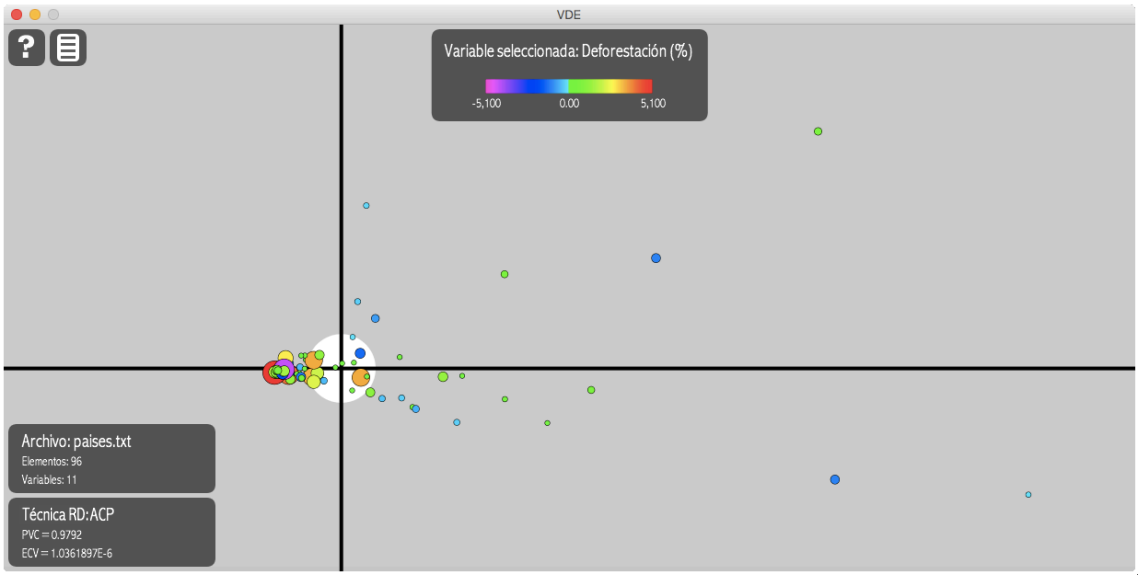


Figura 4-18: *Representación gráfica del conjunto de datos del archivo paises.txt al seleccionar su visualización según la variable Deforestación.*

Cuando la visualización de datos se realiza respecto de una variable, al señalar un elemento mediante el puntero del ratón además de visualizar su estadística como en *Vista general*, su valor en esta variable se señala gráficamente mediante un cursor intermitente de color blanco en la barra de color del cuadro superior. En la figura 4-19 se ilustra este aspecto.

Figura 4-19: *Detalle de la información de un elemento cuando hay una variable seleccionada. El valor del elemento para esa variable se representa mediante un cursor intermitente de color blanco en el cuadro superior a lo largo de la barra que representa el rango de variación de la variable en el conjunto de datos.*

Para volver a la visualización de los datos sin que ninguna variable en especial aparezca destacada, bastará con seleccionar *Vista general* dentro de este mismo apartado de *Vista por variables.*

La posibilidad de poder visualizar un conjunto de datos pudiendo destacar el comportamiento de una variable concreta ofrece al usuario una herramienta de gran utilidad en la labor de descubrir comportamientos, redundancias, tendencias y patrones de una forma sencilla y muy intuitiva.

4.5.3.2. Buscar un elemento

Siempre es interesante para el usuario tener la posibilidad de acceder a la situación gráfica y la información de elementos concretos para analizar su situación respecto al resto e indagar acerca de su entorno próximo. *VDE* facilita la labor de búsqueda de elementos concretos seleccionando el submenú *Buscar un elemento*.

Cuando se escoge esta opción, aparece un cuadro en el cual *VDE* solicita introducir el número de elemento que se desea buscar (ver figura 4-20).

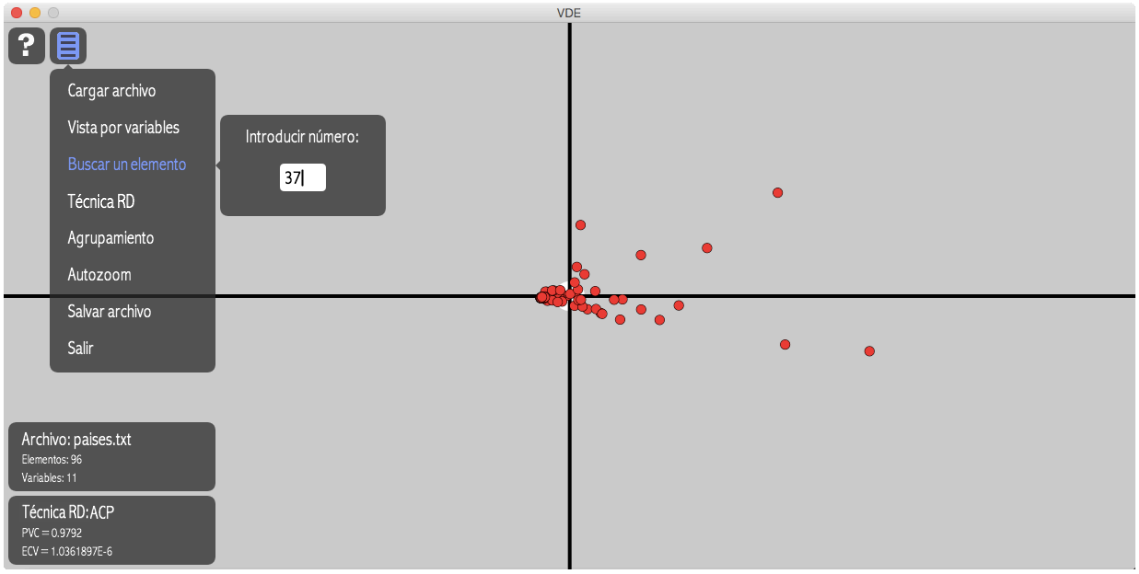


Figura 4-20: *Submenú Buscar un elemento.*

VDE admite solamente la introducción de caracteres numéricos. Una vez introducido el número de elemento deseado, la aceptación se realiza mediante la tecla *Intro*. Si el número introducido es superior a la cantidad de elementos la aplicación informa de esta circunstancia; en caso contrario, se mostrará el conjunto de datos centrando la pantalla en el elemento que se ha seleccionado. Éste aparecerá resaltado de forma intermitente y se facilitará su estadística (ver figura 4-21).

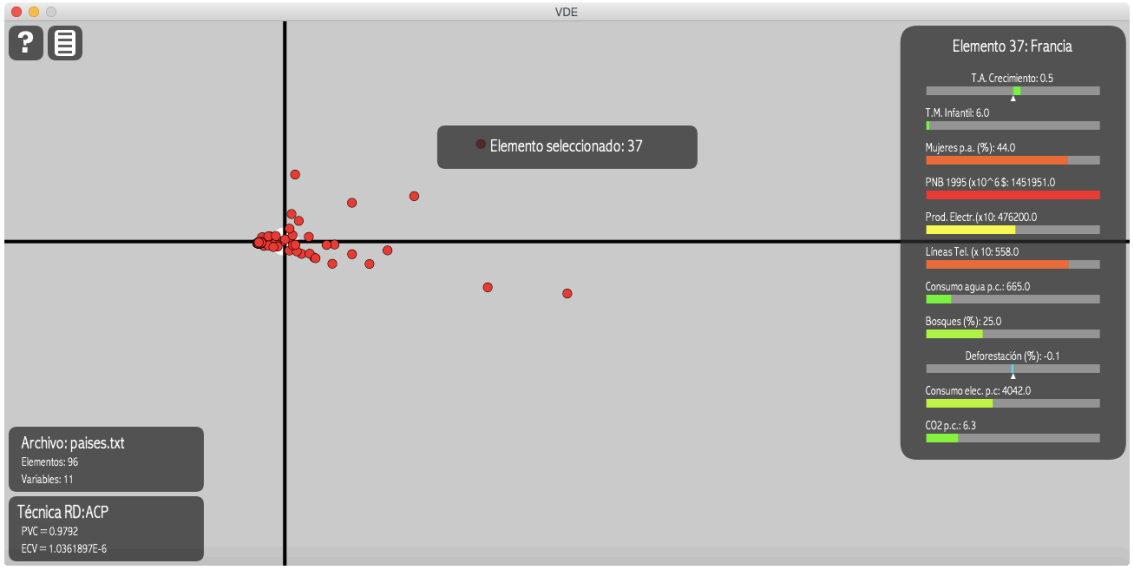


Figura 4-21: *Localización de un elemento concreto.*

Al seleccionar un elemento y quedar la pantalla centrada en él, si se realizan ampliaciones (o reducciones), este elemento siempre quedará centrado. Ello permite analizar su entorno a diferentes escalas.

Al presionar el botón izquierdo del ratón sobre cualquier parte de la pantalla o las teclas de dirección arriba o abajo, la información sobre el elemento desaparece y éste deja de parpadear.

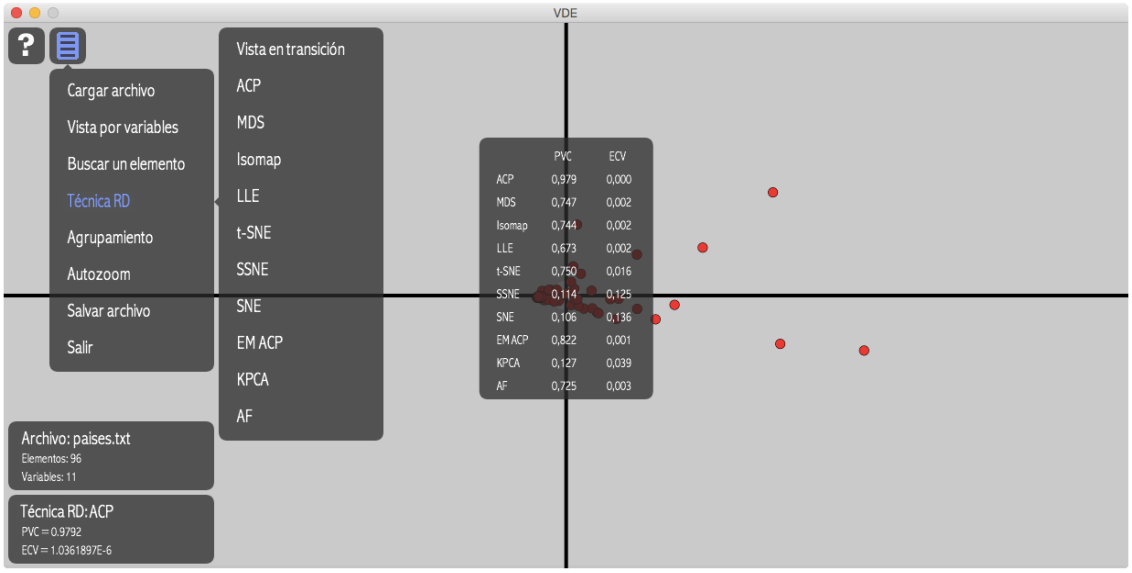
4.5.3.3. Técnica RD (reducción de la dimensionalidad)

Este submenú permite al usuario visualizar en pantalla el conjunto de datos según las diferentes técnicas de reducción de la dimensionalidad analizadas en el capítulo 3. Dado que, como se comentó en ese capítulo, la técnica de reducción de la dimensionalidad *LDA* es supervisada, *VDE* sólo permite su utilización cuando los datos están previamente clasificados. Es decir, *LDA* solamente estará disponible cuando el archivo cargado en *VDE* disponga de la columna denominada *clase* dispuesta como se comentó en el apartado 4.3. De igual modo, al seleccionar esta opción se facilita un cuadro informativo en la parte central de la pantalla donde aparecen las distintas técnicas de reducción de la dimensionalidad con sus valores de PVC y ECV para que el usuario pueda compararlas fácilmente. En la figura 4-22 se muestra el menú *Técnica RD* para los archivos *paises.txt* y *paises_clase.txt* cuyos datos están previamente sin clasificar y clasificados, respectivamente.

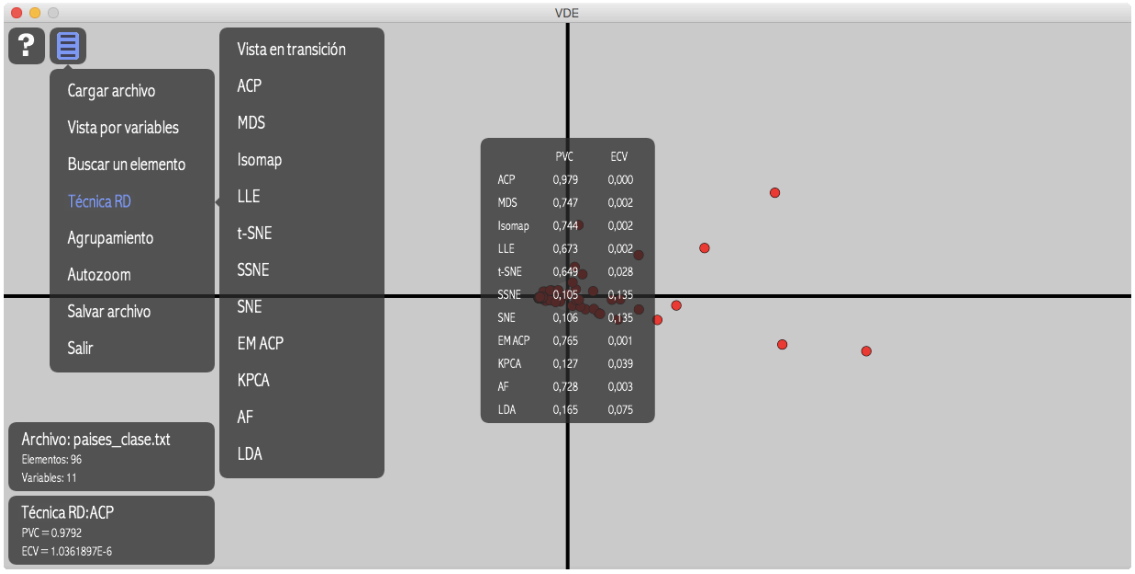
Al seleccionar la técnica de reducción de la dimensionalidad deseada, *VDE* muestra en pantalla la disposición de los elementos según el resultado obtenido en aquélla y en el cuadro de la esquina inferior izquierda actualiza la información comunicando la nueva técnica junto con sus valores de PVC y ECV. En la figura 4-23 se muestran los datos del archivo *paises.txt* para la técnica *EM ACP*. En aquellas técnicas que hacen uso del parámetro *vecinos*, se muestra el valor de éste junto con el nombre de la técnica (ver figura 4-24).

Las técnicas *LLE* e *Isomap*, como se explicó en el capítulo 3, realizan la reducción de la dimensionalidad a partir de un grafo en el cual cada elemento queda conectado directamente con solo un cierto número de elementos. Ello implica que estas técnicas pueden obtener, en ocasiones, subconjuntos de datos conectados entre sí pero sin conexión entre ellos. En estas situaciones, *VDE* solamente muestra en pantalla aquellos elementos que pertenecen al subconjunto con mayor número de datos. Si se desea aumentar el número de datos mostrado por estas técnicas, bastará con aumentar el parámetro de cantidad de vecinos al cargar el conjunto de datos como se comentó en el apartado 4.5.1. Puede

observarse en la figura 4-24 el conjunto de datos del archivo *paises.txt* según la técnica *LLE* para un entorno de 12 elementos vecinos. En el cuadro de la esquina inferior izquierda se muestra junto el nombre de la técnica el valor del parámetro *vecinos* y más arriba, junto con la cantidad de elementos del conjunto, cuántos de ellos aparecen representados gráficamente (en este caso 95 elementos de un total de 96).



(a)



(b)

Figura 4-22: *En (a) se muestra el submenú Técnica RD de VDE para datos sin clasificar y en (b) para datos previamente clasificados*

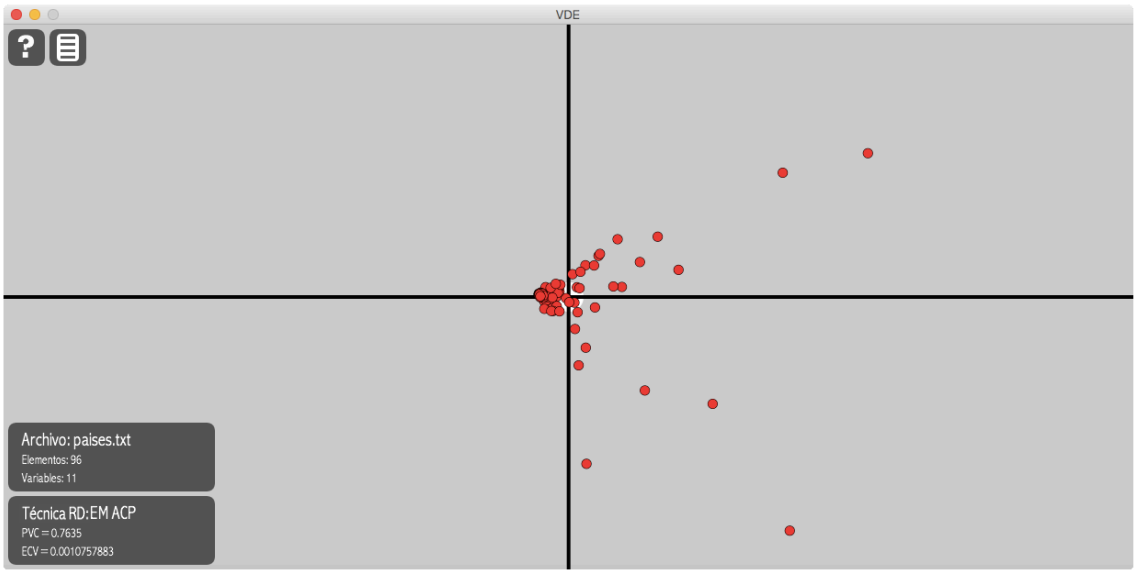


Figura 4-23: *Representación gráfica del conjunto de datos del archivo paises.txt según la técnica de reducción de la dimensionalidad EM ACP.*

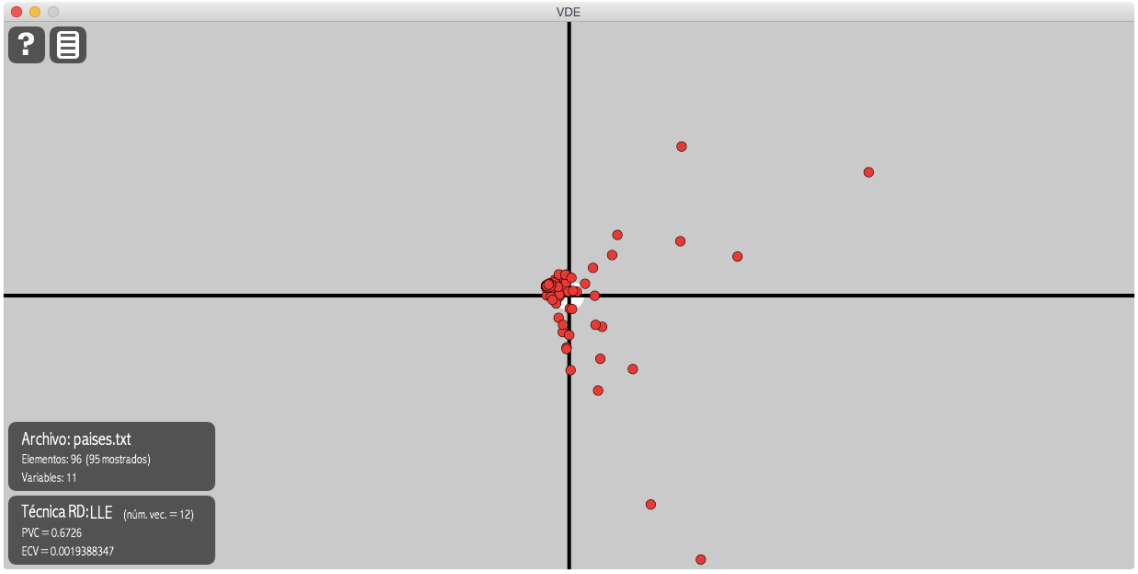


Figura 4-24: *Representación gráfica del conjunto de datos del archivo paises.txt según el método de reducción de la dimensionalidad LLE.*

Dentro del submenú *Técnica RD*, *VDE* ofrece la opción *Vista en transición*. Seleccionándola, *VDE* representa en pantalla cada dato ponderando los valores de las coordenadas que lo representan en dos técnicas de reducción de la dimensionalidad. Esta ponderación puede ser controlada gráficamente por el usuario.

La figura 4-25 muestra la representación gráfica de *VDE* respecto del conjunto de datos del archivo *paises.txt* en una *Vista en transición* para las técnicas de reducción de la dimensionalidad AF y *SNE*. En la esquina inferior izquierda *VDE* sitúa una barra en cuyos extremos se hallan dos cuadros indicando las técnicas seleccionadas y un cursor de color azul sobre esa barra. Este cursor puede ser desplazado a lo largo de la barra al mantener el ratón presionado sobre él. La representación gráfica de los datos en pantalla se corresponde con una ponderación entre los valores de las coordenadas obtenidas en las técnicas de reducción de la dimensión situadas en los extremos de la barra, en este caso AF y *SNE*. Cuanto más cerca se sitúa el cursor de uno de los extremos de la barra mayor es el peso que se asigna a los valores de las coordenadas obtenidas por la técnica situada en ese extremo. En caso de situar el cursor totalmente en uno de los extremos, la representación gráfica de los datos coincide con la de la técnica que se halla situada en ese extremo no teniéndose en cuenta la técnica situada en el otro extremo.

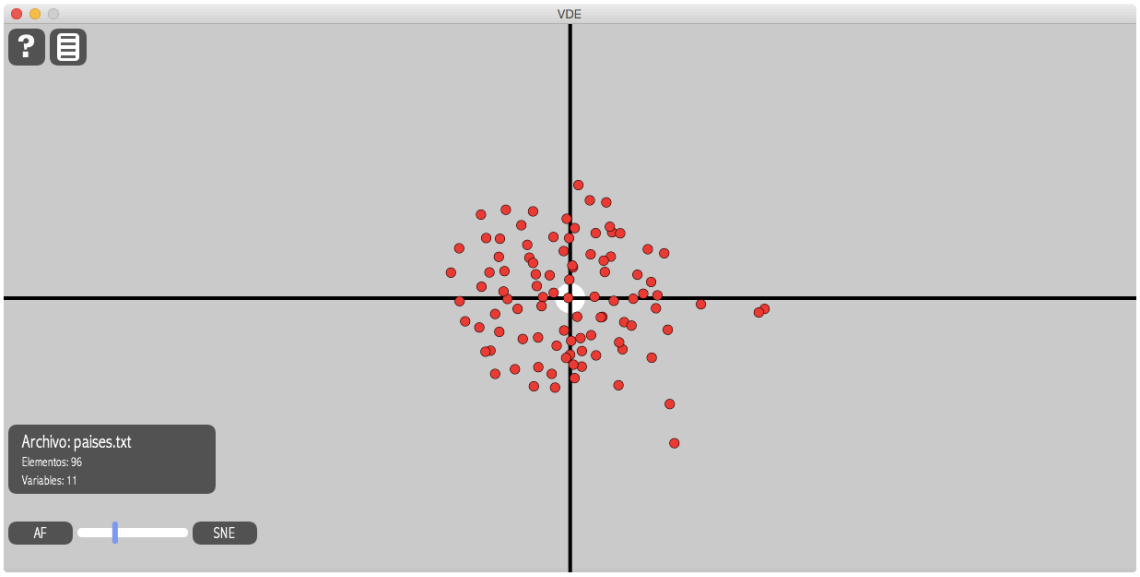


Figura 4-25: *Vista en transición del conjunto de datos del archivo paises.txt entre las técnicas de reducción de la dimensionalidad* AF *y SNE.*

Las técnicas de reducción de la dimensionalidad que se hallan situadas en los extremos de la barra pueden cambiarse presionando sobre éstos con el botón izquierdo del ratón. Al seleccionar uno de los extremos aparece un desplegable con todas las técnicas disponibles desde el cual puede escogerse aquella que se desee situar en ese extremo. El procedimiento es análogo para el otro extremo (ver figura 4-26).

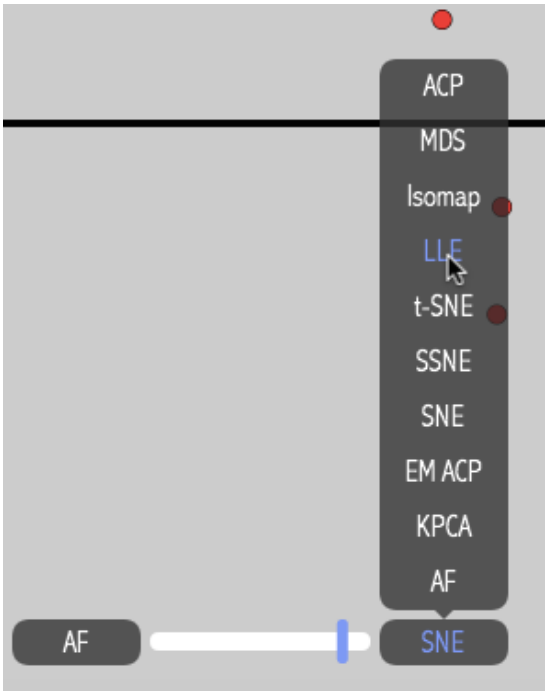


Figura 4-26: *Detalle de la selección de la técnica de reducción de la dimensión para la opción de Vista en transición en uno de los extremos.*

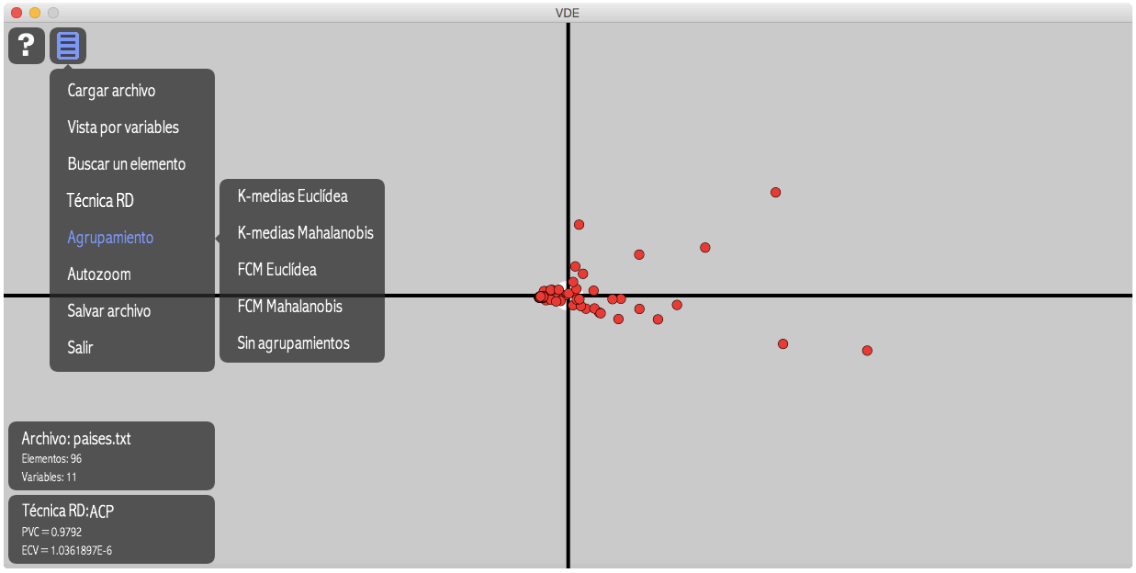
4.5.3.4. Agrupamiento

VDE permite al usuario realizar agrupamientos en el conjunto de datos salvo que el modo de funcionamiento sea agrupado, es decir, cuando el número de elementos es superior a 700 ya que, en este caso, los datos ya han sido previamente clasificados.

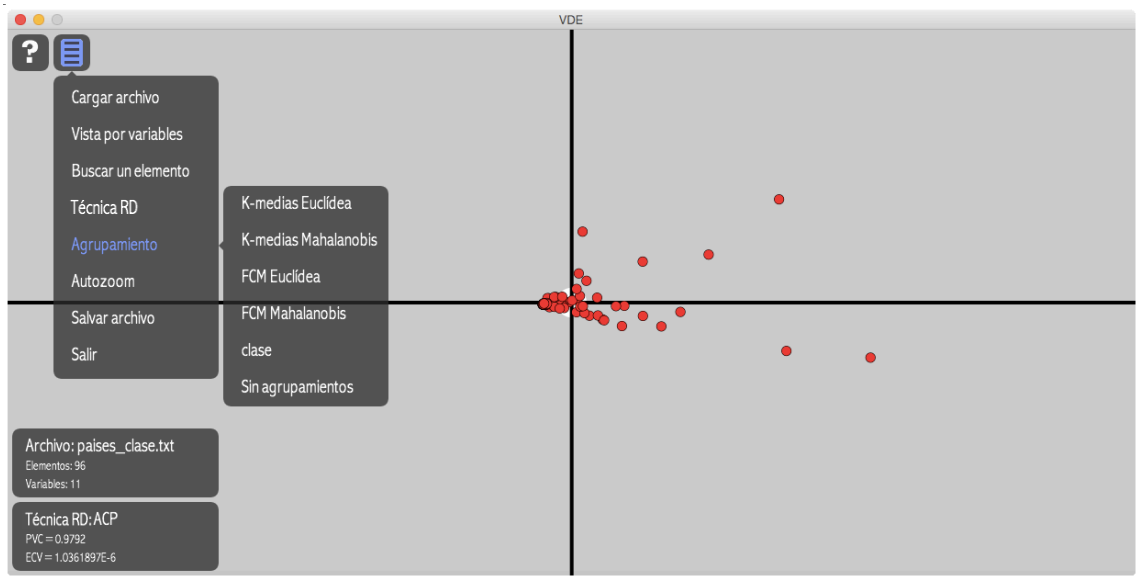
En el modo de funcionamiento normal, la opción *Agrupamiento* permite la realización de grupos según los métodos *K-Medias* y *FCM* expuestos en el capítulo 2. Para cada uno de estos dos métodos, *VDE* ofrece la posibilidad de medir la disimilitud entre datos utilizando las distancias *Euclídea* o *Mahalanobis* también vistas en el mismo capítulo (ver figura 4-27(a)).

Si existe una clasificación previa de los datos, *VDE* muestra además dentro del submenú *Agrupamiento* la opción de visualizar dicho agrupamiento escogiendo la opción *clase* (ver figura 4-27(b)).

Al seleccionar una de las cuatro posibilidades para el agrupamiento que no son la opción *clase*, la aplicación *VDE* muestra un cuadro en el cual se pide introducir el número de grupos en el que se desea clasificar el conjunto de datos. La aceptación de la clasificación requerida se efectúa mediante la tecla *Intro* (ver figura 4-28).



(a)



(b)

Figura 4-27: *En (a) se muestra el submenú Agrupamiento para datos sin clasificación previa y en (b) para datos previamente clasificados.*

Una vez *VDE* realiza el análisis, muestra el resultado en pantalla utilizando distintos colores para representar los diferentes grupos. Así, todos los elementos de un mismo grupo se representan con el mismo color siendo éste distinto al de cualquier otro grupo. En la parte inferior de la pantalla se detalla un cuadro con la información sobre el agrupamiento realizado: el número de grupos, el método de agrupamiento y las técnicas de validación expuestas en el capítulo 2 (*Dunn*, *Silhouette (Sil) y Suma de Cuadrados (SS))* (ver figura 4-29). Si el método de

agrupamiento seleccionado es *FCM*, la aplicación informa, además, del valor del parámetro *difusión*.

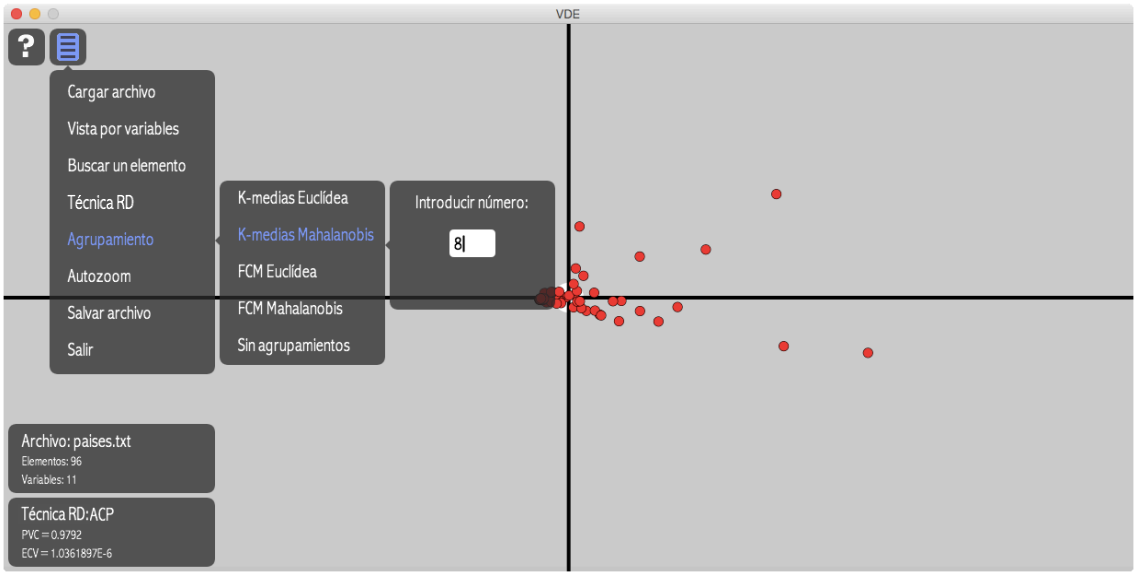


Figura 4-28: *Solicitud para clasificar los datos del archivo paises.txt en 8 grupos según el método K-Medias utilizando la distancia de Mahalanobis.*

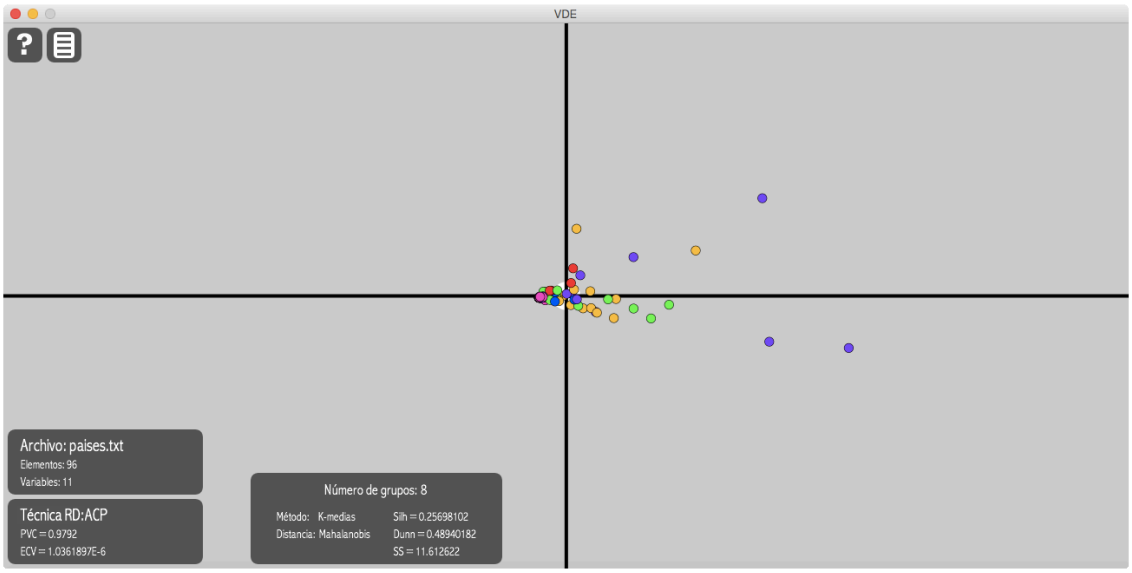


Figura 4-29: *Agrupamiento realizado por VDE.*

Si la opción seleccionada en este submenú es *clase*, *VDE* muestra en pantalla el agrupamiento preestablecido. En este caso, no se realiza ningún cálculo sobre validación (ver figura 4-30).

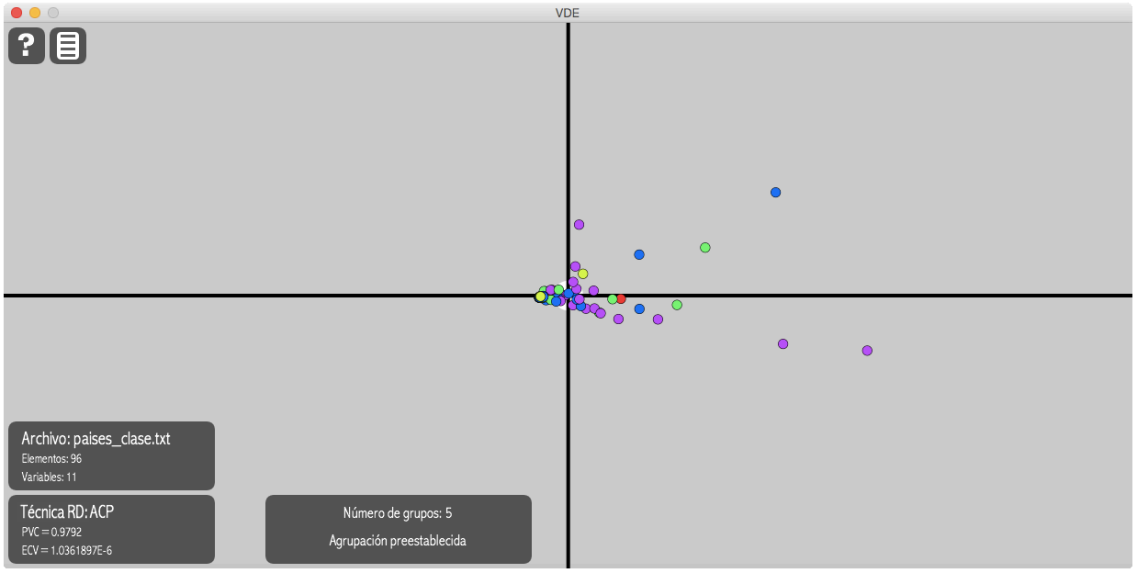


Figura 4-30: *Visualización en VDE del agrupamiento preestablecido en el archivo paises_clase.txt.*

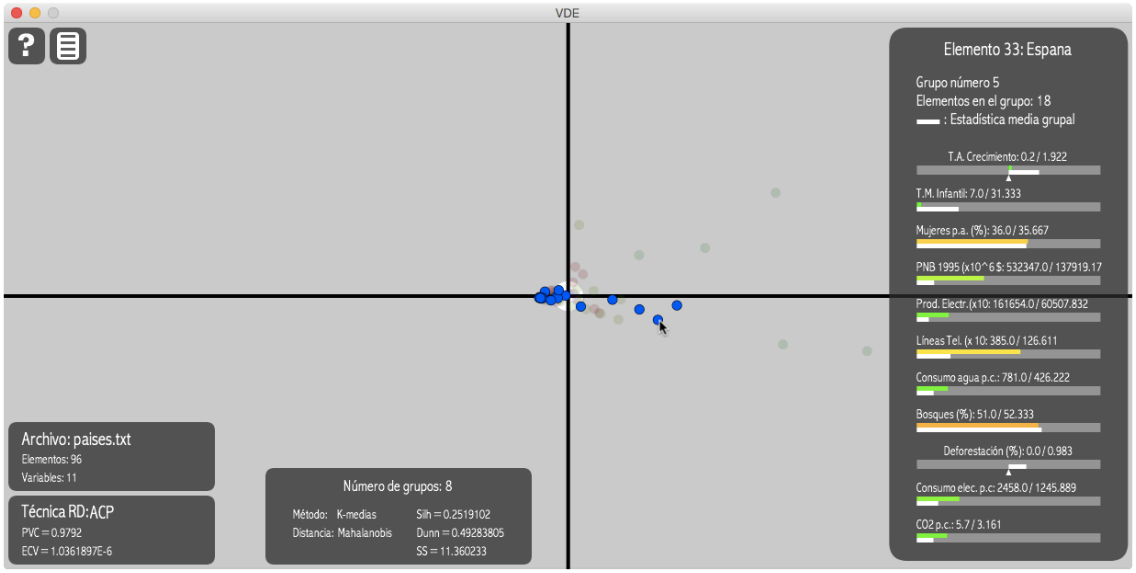


Figura 4-31: *Al realizar un análisis de agrupamiento, VDE muestra en pantalla destacado el grupo de elemento señalado con el puntero del ratón. En el cuadro de información del elemento aparece su estadística junto con la media de su grupo.*

Si sobre el conjunto de datos se ha realizado un agrupamiento, al situar el ratón sobre un elemento, *VDE* resalta en pantalla todos los elementos de su grupo atenuando la representación del resto. En la parte derecha de la pantalla aparece la estadística del elemento junto con la del grupo al que pertenece (ver figura 4-31). En este cuadro aparece el número del elemento, su etiqueta, el número del grupo al que el elemento pertenece y la cantidad de elementos de que consta el grupo.

Más abajo, aparecen los valores del elemento para cada variable junto con los valores medios de todos los elementos de su grupo de forma numérica y gráfica. En este último caso, la barra que representa el valor de la media grupal aparece en color blanco.

Si al mostrar *VDE* un agrupamiento, se selecciona la visualización del conjunto de datos según una variable, el color de representación de cada elemento se mantiene de acuerdo al grupo al que pertenece y no al valor que éste toma en dicha variable. No obstante, el tamaño de cada elemento sigue manteniendo su relación con el valor que toma en esa variable (ver figura 4-32).

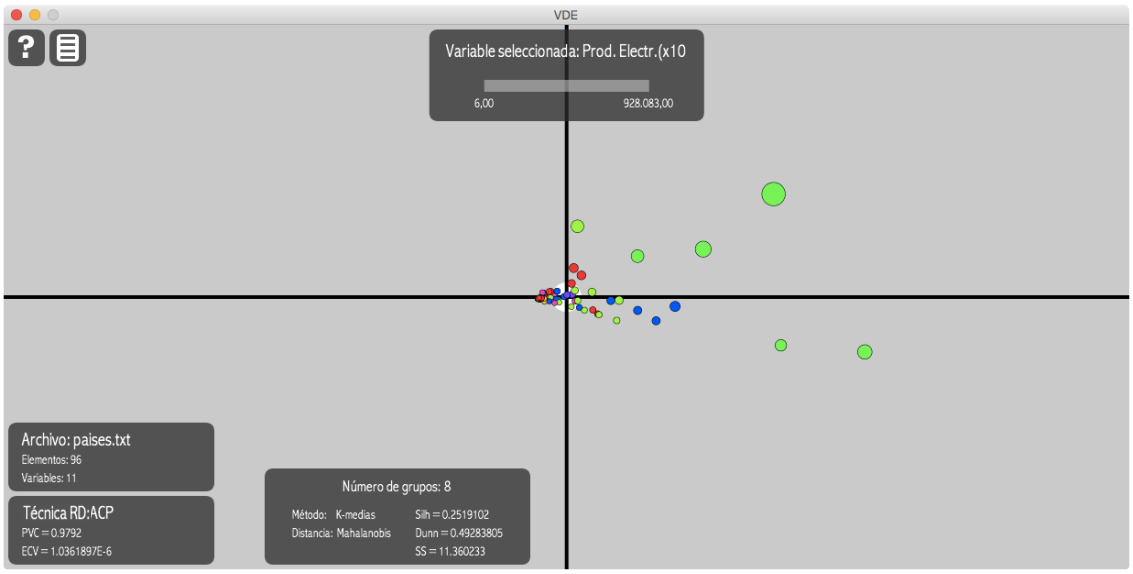


Figura 4-32: *Representación del conjunto de datos del archivo paises.txt en modo agrupado y con variable seleccionada.*

En el caso de que la variable seleccionada para la visualización de datos tome valores positivos y negativos, el tamaño del círculo que representa un dato no discrimina entre si éste toma un valor positivo o negativo. En este caso, *VDE* representa aquellos datos que toman valores negativos mediante una línea horizontal a lo largo del diámetro del círculo que lo representa (ver figura 4-33).

Cuando *VDE* aplica el método *FCM* para realizar el agrupamiento asigna cada elemento al grupo respecto del cual ha obtenido un mayor grado de pertenencia (ver capítulo 2). Ello puede llevar a situaciones en las que si la cantidad de grupos solicitado es alto alguno de ellos quede sin elementos y se presente el conjunto de datos distribuido en un número de grupos menor que el requerido. Si se da esta circunstancia, *VDE* muestra un cartel advirtiendo de la situación y expresa en el cuadro inferior esta información (ver figura 4-34).

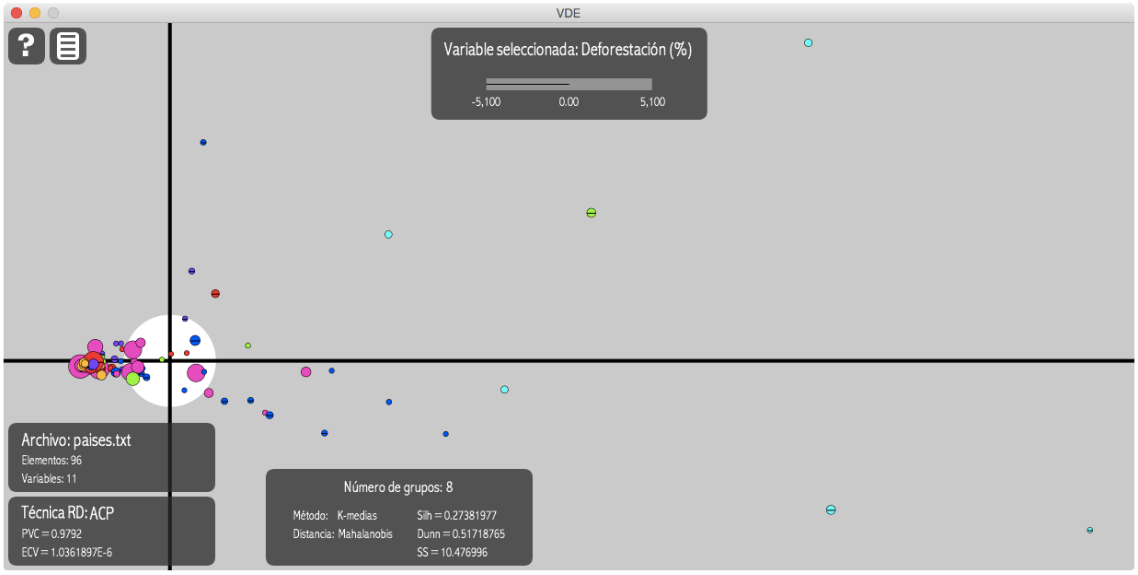


Figura 4-33: *Representación del conjunto de datos del archivo paises.txt con los datos agrupados y seleccionando la variable Deforestación que puede tomar valores positivos y negativos.*

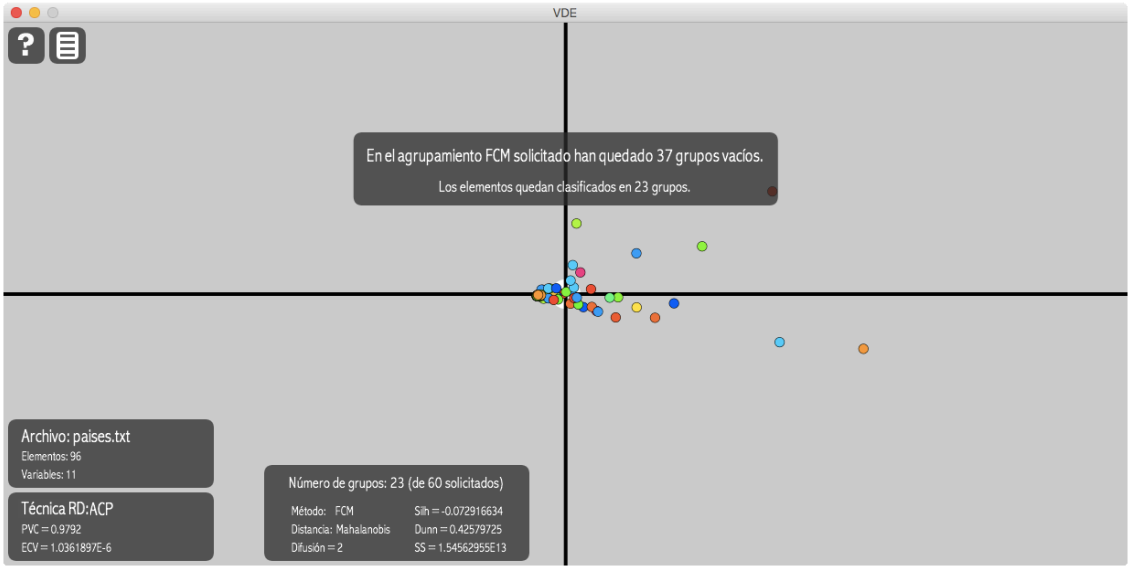


Figura 4-34: *La figura muestra un agrupamiento según FCM utilizando la distancia de Mahalanobis. A pesar de haberse solicitado distribuir el conjunto en 37 grupos éste queda clasificado solamente en 23. El cuadro central de advertencia desaparece al presionar el botón del ratón en cualquier parte de la pantalla mientras que en el cuadro inferior la información queda permanente.*

4.5.3.5. Autozoom

Seleccionando la opción *Autozoom*, *VDE* representa los datos situando los ejes de referencia centrados en pantalla y con un grado de ampliación suficiente para poder visualizar todo el conjunto.

4.5.3.6. Salvar archivo

La opción *Salvar archivo* permite guardar todos los resultados obtenidos por *VDE* para el conjunto de datos que en ese momento está cargado en la aplicación. Esto permite poder recuperar toda esta información de manera instantánea sin que *VDE* tenga que volver a realizar los cálculos.

Los archivos generados por *VDE* son archivos con el nombre *dataVDE#.txt* donde # es un número que la aplicación asigna con el objetivo de no coincidir con otros archivos que puedan tener el mismo nombre y sobrescribirlos. Estos archivos se guardan en la carpeta *data* del directorio de la aplicación.

Cada archivo generado por *VDE* declara en su primera línea el nombre del archivo de datos y a continuación expresa los valores obtenidos en los diferentes cálculos realizados.

4.6. Manejo de la aplicación. Modo de funcionamiento agrupado

Cuando el número de datos es superior a 700, *VDE* opera en modo de funcionamiento agrupado. En este caso, como ya se comentó en el apartado 4.4., la aplicación clasifica los elementos del conjunto de datos en 700 grupos y procede a la representación gráfica de los centroides (denominados representantes) de esos grupos, no de los datos originales. El objetivo principal que se persigue con esta estrategia es el de facilitar la labor del usuario en la exploración.

Este modo de funcionamiento es prácticamente igual que el modo normal salvo por lo comentado en el párrafo anterior y que, en este caso, la posibilidad de realizar agrupamientos no está disponible, es decir, el submenú *Agrupamiento* está deshabilitado. Lo importante ahora es saber cómo se puede manejar la aplicación para disponer de toda la información en este nuevo escenario en el cual se trabaja con representantes de grupos de datos en lugar de directamente con los propios datos.

Para ilustrar este modo de funcionamiento, se utilizará la base de datos *Yeast* obtenida del *UCI Machine Learning Repository* (Lichman, 2013). Este conjunto consta de 1484 datos. Cada dato se refiere a una proteína en la cual se han medido 8 variables o características (*mcg*, *gvh*, *alm*, *mit*, *erl*, *pox*, *vac* y *nuc*).

En la figura 4-35 se muestra el resultado obtenido por *VDE*. Como el

número de elementos del conjunto es superior a 700 el modo de funcionamiento de la aplicación pasa a ser agrupado. La primera diferencia que puede apreciarse con respecto al modo de funcionamiento normal es que en el cuadro inferior se señala que el conjunto consta de 1484 elementos y que hay 700 representantes. Estos representantes son los que aparecen gráficamente.

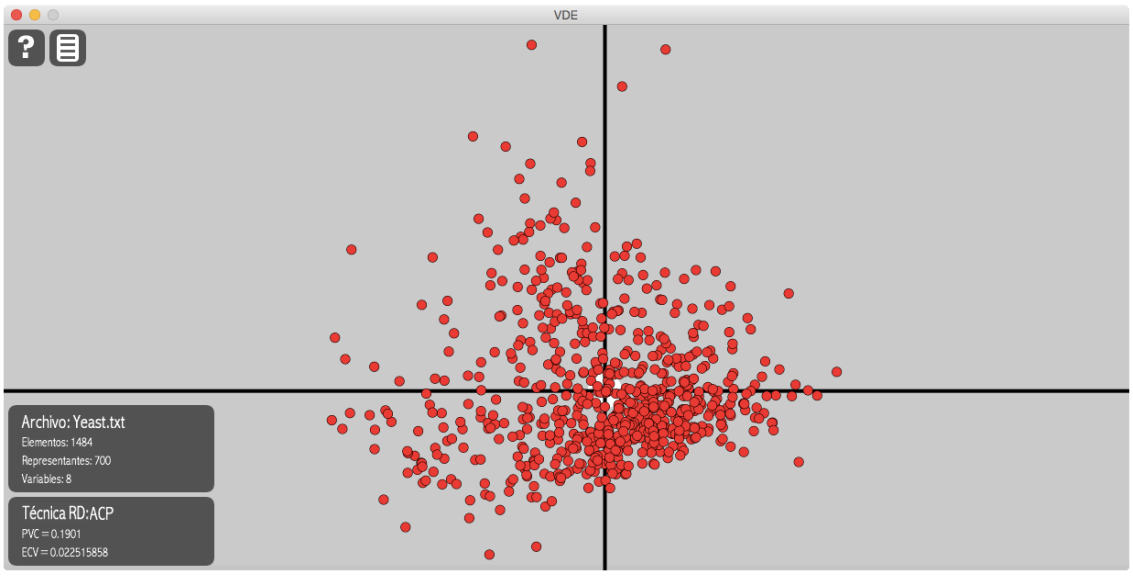


Figura 4-35: *Aplicación VDE en modo de funcionamiento agrupado.*

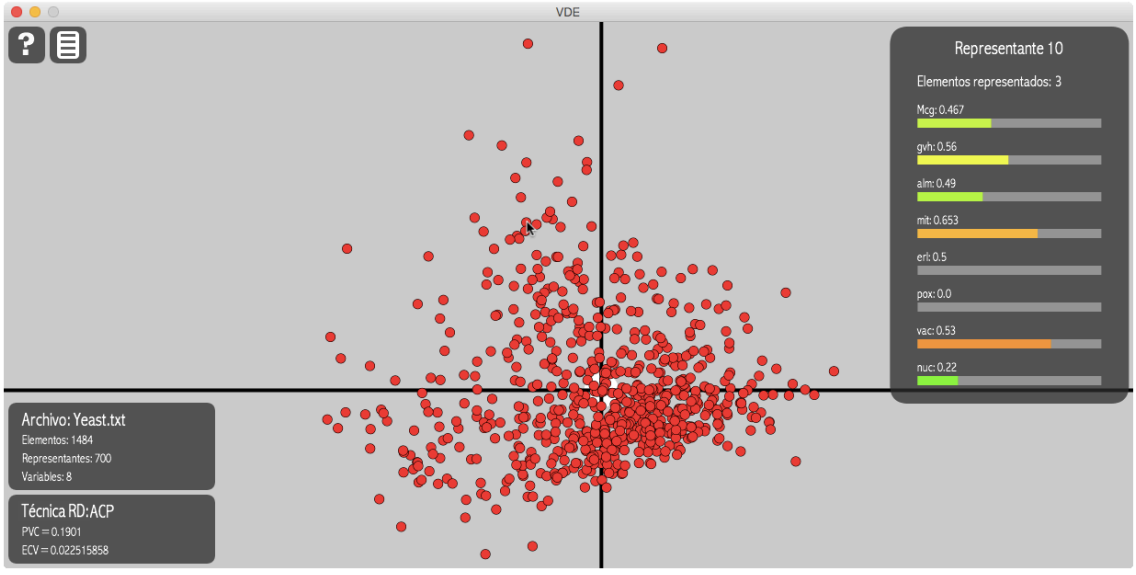


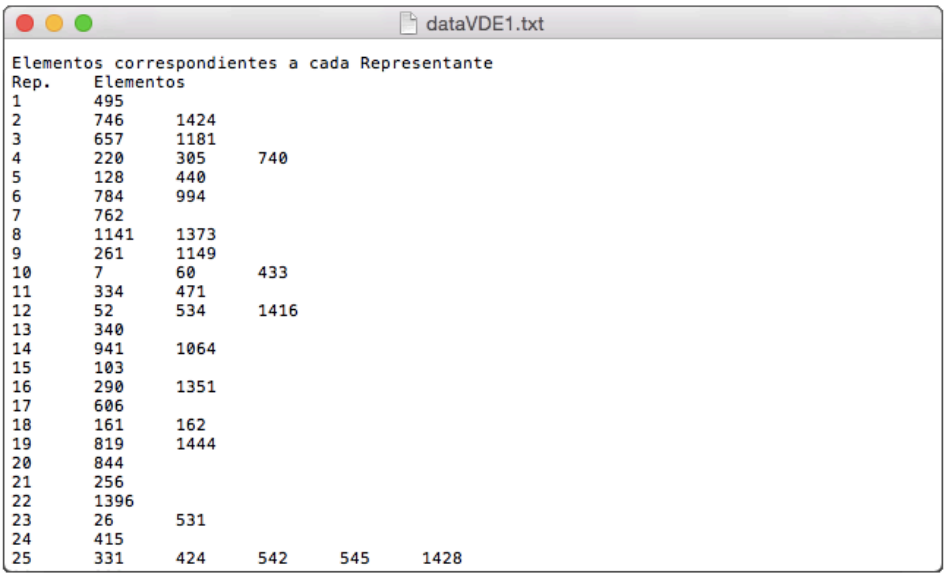
Figura 4-36: *Detalle de la información sobre un representante en modo de funcionamiento agrupado.*

En este modo de funcionamiento, la información que se muestra del elemento señalado por el puntero del ratón es la correspondiente al centroide del

grupo al que representa. En este caso se muestra al principio de la estadística el número de representante (número de grupo que asigna *VDE*) y la cantidad de datos del conjunto original que representa (elementos). En la figura 4-36 se muestra la estadística del representante 10 donde puede observarse que incorpora 3 elementos, es decir, es un grupo formado por tres datos originales. La información que ahí aparece de cada variable se corresponde con la media de los valores de cada uno de los elementos a los que se representa.

Con el objetivo de no saturar la información presentada en pantalla, *VDE* facilita cuáles son los datos originales que componen cada grupo mediante el fichero generado con la opción *Salvar archivo*. Cuando *VDE* genera un archivo estando en modo de funcionamiento agrupado, en la parte final de este archivo aparece un listado en el cual se detalla para cada uno de los representantes qué elementos del conjunto original incorpora. En la figura 4-37 aparece la forma en que esta información se facilita en el archivo *dataVDE1.txt* respecto de los primeros 25 representantes.

En esta figura puede observarse, por ejemplo, que el representante número 1 solamente consta del elemento número 495 o que el representante número 25 hace referencia a los elementos número 331, 424, 542, 545 y 1428. En el caso del representante número 10 anteriormente mencionado, puede observarse que los tres elementos que representa son el 7, el 60 y el 433.


dataVDE1.txt

```
Elementos correspondientes a cada Representante
Rep.    Elementos
1       495
2       746     1424
3       657     1181
4       220     305     740
5       128     440
6       784     994
7       762
8       1141    1373
9       261     1149
10      7       60      433
11      334     471
12      52      534     1416
13      340
14      941     1064
15      103
16      290     1351
17      606
18      161     162
19      819     1444
20      844
21      256
22      1396
23      26      531
24      415
25      331     424     542     545     1428
```

Figura 4-37: *Detalle del archivo dataVDE1.txt generado por VDE donde se muestran los elementos que corresponden a los primeros 25 representantes del conjunto de datos.*

El acceso a la información de los datos originales se realiza a través del

submenú *Buscar un elemento*. El procedimiento es exactamente el mismo que en el caso de funcionamiento normal. Al introducir el número del elemento deseado en este submenú, *VDE* se centra en el representante del grupo en el que se halla el elemento y la estadística que se muestra es la de ambos. En la figura 4-38 se muestra la información correspondiente al elemento número 60. En este caso, *VDE* informa del número del elemento, su etiqueta, el número del representante del grupo al que pertenece y de la cantidad de elementos en dicho grupo. Más abajo aparecen los valores del elemento en cada variable conjuntamente con los valores correspondientes a los valores medios de todos los elementos del grupo, es decir, del representante, que son mostrados en color blanco. De este modo, pueden compararse fácilmente ambas estadísticas.

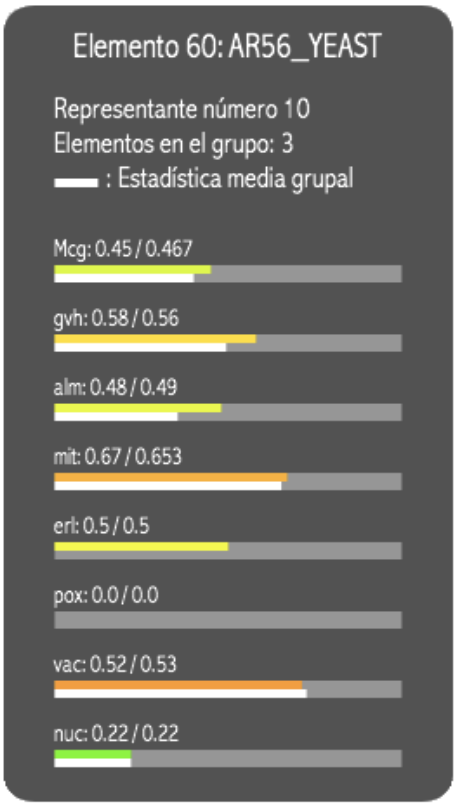


Figura 4-38: *Información mostrada por VDE en modo de funcionamiento agrupado al localizar un elemento mediante la opción Buscar un elemento. En la figura se muestra la estadística del elemento buscado, cuyo número es 60, conjuntamente con la de su representante, cuyo número es 10.*

Para el resto de funcionalidades como navegación por el entorno, *Técnica RD* (reducción de la dimensionalidad), *Vista por variables*, *Autozoom*, etc … el manejo de la aplicación *VDE* es igual en ambos modos de funcionamiento.

Capítulo 5

Casos de estudio

Resumen

Este capítulo ilustra la utilidad de la herramienta de visualización VDE en su aplicación a distintos problemas reales, todos ellos obtenidos del *UCI Machine Learning Repository* (Lichman, 2013)*. El primer problema analizado es sobre la base de datos Semillas (Seeds) cuyos datos se corresponden con medidas obtenidas a partir de semillas de trigo pertenecientes a tres clases diferentes. El segundo caso estudiado versa sobre el conjunto Iris que se compone de datos obtenidos en medidas realizadas en tres variedades de plantas del género Iris. El tercer problema abordado se refiere a la base de datos Wine cuyos elementos se elaboran en consideración a los resultados obtenidos en un análisis químico sobre vinos correspondientes a tres cultivos diferentes en una misma región de Italia. El cuarto conjunto estudiado es E.coli que presenta un problema de clasificación de proteínas en distintas localizaciones celulares a partir de sus secuencias de aminoácidos. Por último, se trata la base de datos Modelado del usuario donde cada elemento se corresponde con los valores obtenidos teniendo en cuenta la medición de diferentes atributos en un determinado usuario con el objeto de poderlo clasificar correctamente en un determinado nivel de conocimiento.*

Con estos casos de estudio se demuestra la validez de la herramienta de análisis visual de datos VDE para la extracción de información y conocimiento proponiendo un entorno sencillo y fácil de interpretar.

5.1. Introducción

En el apartado 4.1 se han puesto de manifiesto las principales características de *VDE* como respuesta al problema de la extracción de conocimiento en conjuntos con gran cantidad de datos multidimensionales. También se destacó en ese apartado como aspectos principales de esta aplicación el ofrecer un entorno intuitivo y de fácil manejo mediante el cual el usuario pueda visualizar los datos de una forma estrechamente vinculada a la estructura original de los datos gracias a las técnicas de reducción de la dimensionalidad.

En este capítulo se procede a la utilización de *VDE* en el análisis de cinco conjuntos de datos reales. En cada caso se realiza una descripción de la base de datos y una exposición y análisis de los resultados obtenidos en el uso de la herramienta. La metodología a seguir para la obtención de estos resultados es básicamente la misma en todos los casos con la excepción de la aplicación en el análisis del conjunto de datos *Modelado del usuario* donde se procede únicamente al análisis supervisado debido a sus características particulares (ver apartado 5.6). En resumen, esta metodología es la siguiente:

- Realización de un análisis no supervisado.
    - Discusión sobre el método de reducción de la dimensionalidad que mejor realice la representación gráfica del conjunto de datos en dos dimensiones.
    - Inspección visual de la distribución de las diferentes variables o atributos a partir de la representación gráfica adoptada.
    - Análisis de agrupamiento.
- Realización de un análisis supervisado.
    - Visualización del conjunto de datos según su agrupamiento original y comparación de resultados con los obtenidos en el análisis no supervisado.
    - Estudio de la técnica supervisada de reducción de la dimensionalidad *LDA* para la representación gráfica del conjunto de datos en comparación con la utilizada en el análisis no supervisado.
- Observación de datos concretos por ser especialmente interesantes en el examen del conjunto como, por ejemplo, datos anómalos.

En cada apartado se reflexiona acerca de los resultados que se obtienen y se analiza la importancia de los mismos en la extracción de conocimiento sobre el conjunto de datos.

Finalmente, cabe destacar que *VDE* no está concebida como una herramienta que obtenga resultados de forma automática sino que posibilita un entorno óptimo en el que el usuario puede aprovechar su experiencia a la hora de extraer conocimiento. De este modo, la metodología aquí propuesta para el análisis de los distintos conjuntos de datos es simplemente una referencia de algunas de las posibilidades que se pueden manejar. El conocimiento que el usuario tenga sobre el contexto en el que los datos han sido elaborados y los intereses y motivaciones que éste tenga son elementos esenciales a la hora de aprovechar al máximo las posibilidades de esta aplicación.

5.2. Conjunto de datos *Semillas* (*Seeds*)

5.2.1. Descripción de la base de datos

Esta base de datos (Lichman, 2013) se elabora tomando en consideración diferentes propiedades geométricas medidas a partir de una técnica de rayos X en semillas de trigo. El grupo de semillas examinado está compuesto por granos pertenecientes a tres variedades de trigo denominadas *"Kama"*, *"Rosa"* y *"Canadian"*, existiendo un total de 70 elementos en cada una, seleccionados al azar para el experimento. Estos granos fueron extraídos a partir de las cosechas obtenidas en campos experimentales por parte del *Institute of Agrophysics of the Polish Academy of Sciences* en Lublin. Cada dato consta de 7 atributos medidos en cada grano de trigo: *área* (A), *perímetro* (P), *compacidad* $\left(C = \frac{4*\pi*A}{P^2}\right)$, *longitud*, *anchura*, *coeficiente de asimetría* y *longitud de la ranura del grano*; todos ellos de valor continuo (real).

5.2.2. Análisis de resultados

El conjunto de datos se analiza en *VDE* con unos valores de 15 para el parámetro *vecindario* y de 2 para *difusión*. Las técnicas de reducción de la dimensión que obtienen los mejores resultados en la evaluación de la calidad son

ACP, *Isomap*, *SNE* y *t-SNE* (ver tabla 5-1). En el caso de ACP, el valor obtenido para PVC es de 0,93 que es significativamente alto, lo cual indica que la inmersión mostrada por esta técnica es muy fiel a la estructura original de los datos en alta dimensión. La representación del conjunto de datos según la técnica de reducción de la dimensionalidad ACP se muestra en la figura 5-1.

Tabla 5-1: *Valores de* PVC *y* ECV *obtenidos por VDE en la evaluación de la calidad de la inmersión realizada en el conjunto de datos Semillas por los métodos* ACP, *Isomap, SNE y t-SNE.*

	PVC	ECV
ACP	0,930	0,000
Isomap	0,874	0,001
SNE	0,869	0,001
t-SNE	0,812	0,003

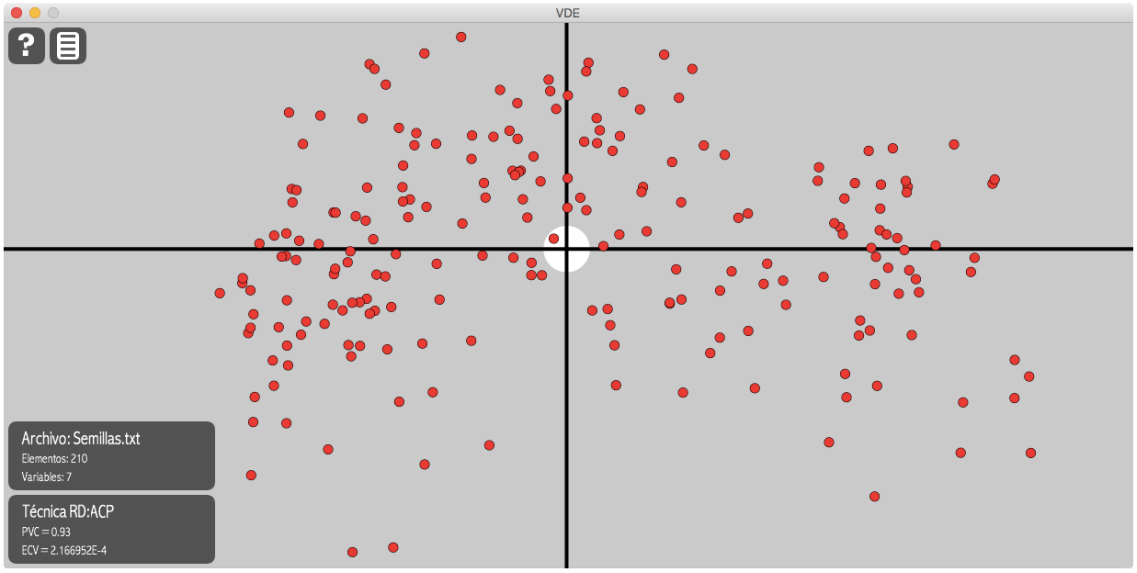


Figura 5-1: *Conjunto de datos Semillas según la técnica de reducción de la dimensionalidad* ACP.

Procediendo al análisis visual de la distribución de los distintos atributos en el conjunto de datos, puede observarse que aquellos que explican mejor la disposición espacial de éstos en cada eje coordenado son el *área*, la *anchura*, la *longitud de la ranura del grano* y el *coeficiente de asimetría*. Como puede observarse en la figura 5-2 los tres primeros atributos se distribuyen a lo largo del

eje horizontal desde valores menores en los datos situados a izquierda a mayores para los situados a la derecha (figuras 5-2(a), 5-2(b) y 5-2(c)) mientras que el *coeficiente de asimetría* varía a lo largo del eje vertical desde valores pequeños en la parte superior a valores mayores en la parte inferior (figura 5-2(d)). Esta información es especialmente importante ya que ayuda a comprender las dos componentes principales en la representación de los datos mediante ACP (ver apartado 3.4.1.). Para el resto de atributos se observa un comportamiento similar al de los tres primeros a lo largo del eje horizontal a excepción del de *compacidad*.

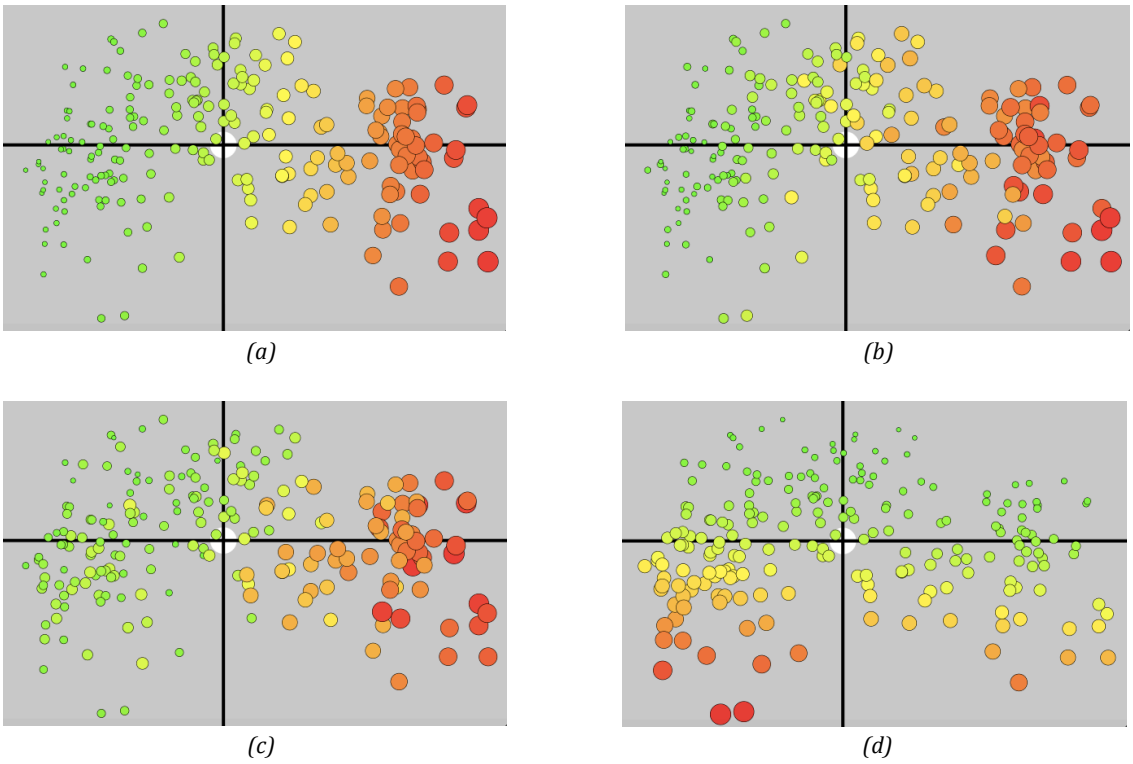
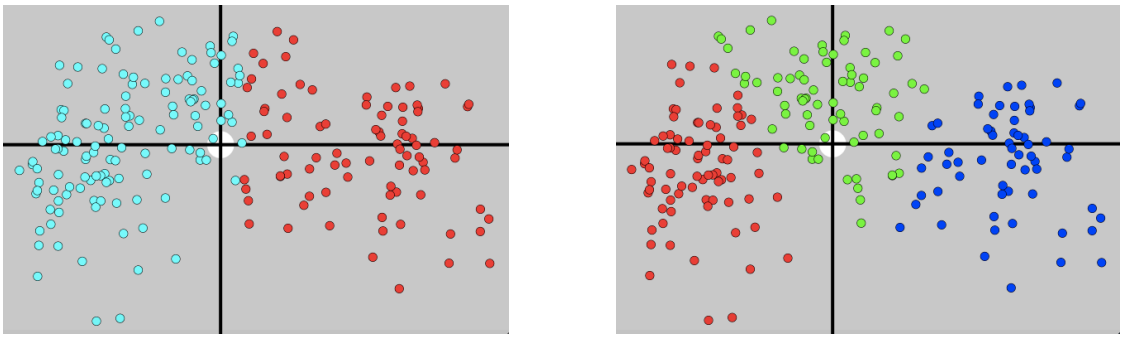


Figura 5-2: *Visualización del conjunto de datos Semillas en VDE según las variables área (a), anchura (b), longitud de la ranura del grano (c) y coeficiente de asimetría (d).*

Tabla 5-2: *Valores de los índices de validación obtenidos por VDE en el agrupamiento del conjunto Semillas para el método FCM utilizando la distancia Euclídea.*

	Silh	*Dunn*	*SS*
2 grupos	0,631	0,831	1,086
3 grupos	0,598	0,704	0,815

En cuanto al análisis de agrupamiento, el método que obtiene mejores resultados, según los índices de validación presentados en el capítulo 2, es *FCM* utilizando la distancia *Euclídea* para 2 ó 3 grupos (ver tabla 5-2). El resultado visual de estos agrupamientos presentado por *VDE* se muestra en la figura 5-3.



(a) *(b)*

Figura 5-3: *Distribución del conjunto de datos Semillas según el método de agrupamiento FCM utilizando la distancia Euclídea. Resultado para 2 grupos (a) y 3 grupos (b).*

Al estar el conjunto de datos previamente clasificado, el correspondiente agrupamiento puede también visualizarse en *VDE*. El resultado se muestra en la figura 5-4 donde los elementos de la clase *"Canadian"* aparecen en color rojo, los de la clase "*Kama"* en verde y los de la clase "*Rosa"* en azul.

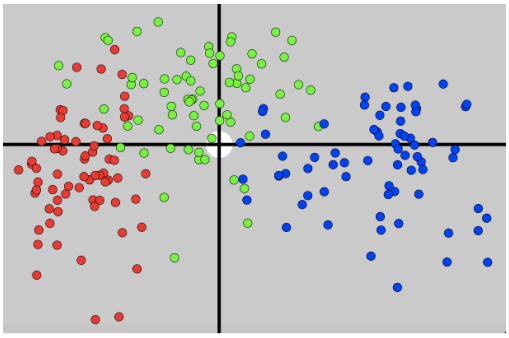


Figura 5-4: *Agrupamiento original del conjunto de datos Semillas. Los elementos de la clase "Canadian" aparecen en color rojo, los de la clase "Kama" en verde y los de la clase "Rosa" en azul.*

Haciendo uso del agrupamiento original del conjunto de datos, análisis supervisado, puede obtenerse la siguiente información:

- Comparando las figuras 5-3(b) y 5-4 puede observarse que el agrupamiento realizado por *FCM* utilizando la distancia *Euclídea* para 3

grupos distribuye los datos de forma muy similar a la original (concretamente clasifica de forma correcta un 89,5 % de los datos).

- Mediante una simple inspección visual, se constata que el atributo que mejor diferencia los diferentes grupos entre sí es el *área* (ver figura 5-5).

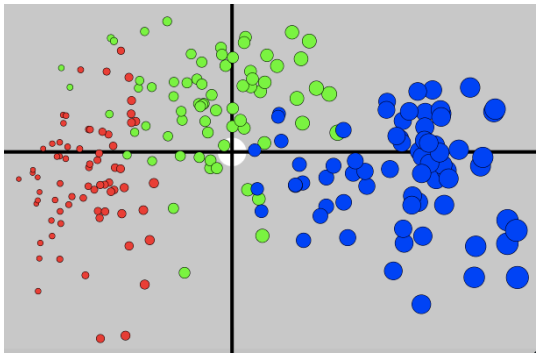


Figura 5-5: *Visualización del agrupamiento original del conjunto de datos Semillas de acuerdo con el atributo área. Los elementos de la clase "Canadian" aparecen en color rojo, los de la clase "Kama" en verde y los de la clase "Rosa" en azul.*

- Analizando las estadísticas de los elementos que no han sido correctamente clasificados, puede obtenerse información acerca del error en la clasificación. Así, por ejemplo, se observa que las clases "*Canadian*" y "*Rosa*" (cuyos elementos están representados en rojo y azul respectivamente) quedan bien separadas (no hay elementos de una clase que se clasifiquen en la otra). Sin embargo, existen problemas entre esas dos clases y la clase "*Kama*" (cuyos elementos aparecen representados en color verde). En los elementos de la clase "*Kama*" que son clasificados como "*Canadian*" (por ejemplo el número 62) se observa que una variable importante en la clasificación errónea es el *coeficiente de asimetría* y lo mismo ocurre al contrario, elementos de la clase "*Canadian*" son mal clasificados en la clase "*Kama*" debido a este mismo atributo (por ejemplo los elementos 180 y 202). Ello indica que este atributo no ha de ser tenido en cuenta para distinguir ambas clases. Sin embargo, en la distinción de las clases "*Kama*" y "*Rosa*" es precisamente el atributo *coeficiente de asimetría* uno de los que ha de ser tenido en cuenta a la hora de distinguir ambas clases (este atributo aparece más acorde a la media grupal para la clase correcta como puede observarse en los elementos 38 y 136). En este punto hay que recordar que el *coeficiente de asimetría* es el atributo que más influye con diferencia en la disposición vertical del conjunto de datos representado (segunda componente principal).

- En el análisis supervisado, siempre es interesante observar la representación del conjunto de datos según la técnica *LDA* a pesar de que no obtenga un buen resultado en la evaluación de la calidad ya que, como se comentó en el capítulo 3, es una técnica que utiliza la información acerca de la clasificación previa de los datos para realizar la inmersión del conjunto. En este caso, puede observarse que *LDA* presenta las clases del conjunto de datos de forma más compacta y diferenciada que ACP (ver figura 5-6).

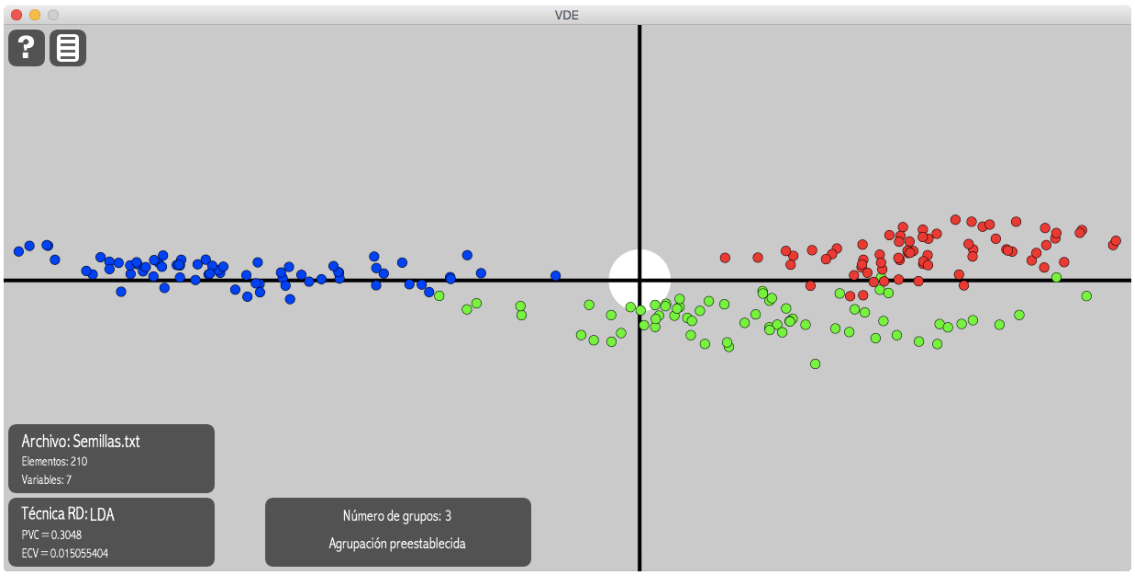


Figura 5-6: *Conjunto de datos Semillas según la técnica de reducción de la dimensionalidad LDA. Los elementos de la clase "Canadian" aparecen en color rojo, los de la clase "Kama" en verde y los de la clase "Rosa" en azul.*

5.3. Conjunto de datos *Iris*

5.3.1. Descripción de la base de datos

La base de datos *Iris* (Lichman, 2013) toma en consideración medidas efectuadas en pétalos y sépalos de plantas de la especie Iris. El grupo examinado está compuesto por plantas pertenecientes a tres variedades denominadas *"Setosa"*, *"Versicolour"* y *"Virginica"*, contando con un total de 50 elementos cada una. Cada dato consta de 4 atributos medidos en cada flor: *longitud del sépalo*, *anchura del sépalo*, *longitud del pétalo* y *anchura del pétalo.* Todos ellos de valor continuo (real) y expresados en centímetros.

5.3.2. Análisis de resultados

El análisis con *VDE* se efectúa con unos valores de 15 en *vecindario* y de 2 en *difusión*. Para este valor del parámetro *vecindario*, el grafo construido por los métodos *Isomap* y *LLE* no logra conectar todos los elementos del conjunto (solamente 99 de los 150). Sin embargo, aumentando el valor de *vecindario* para llegar a conectar todos los elementos del conjunto, los valores obtenidos en la evaluación de la calidad empeoran respecto a los conseguidos con el valor propuesto al principio de este párrafo. Es por ello que el valor de *vecindario* se mantiene en 15 para el análisis con *VDE*.

Tabla 5-3: *Valores de* PVC *y* ECV *obtenidos por VDE en la evaluación de la calidad de la inmersión realizada en el cojunto de datos Iris por los métodos SNE,* ACP *y t-SNE.*

	PVC	ECV
SNE	0,791	0,001
ACP	0,747	0,002
t-SNE	0,740	0,002

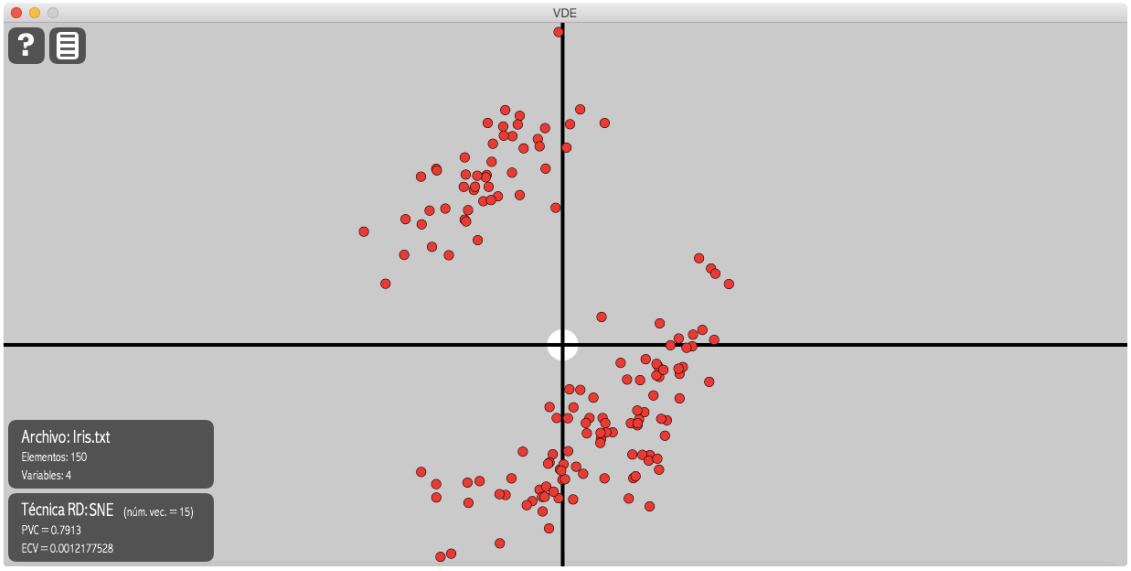
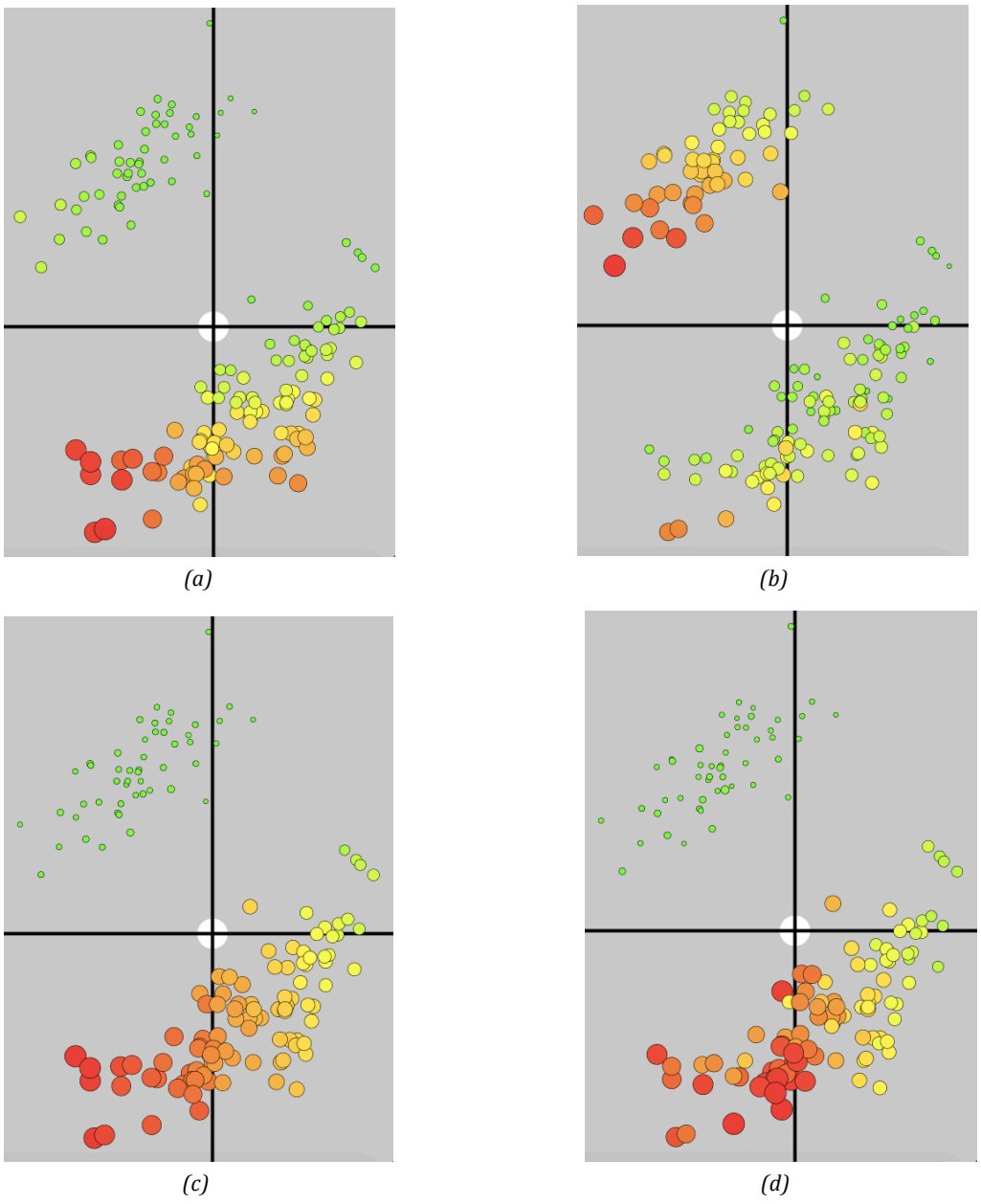


Figura 5-7: *Conjunto de datos Iris según la técnica de reducción de la dimensionalidad SNE.*

Las técnicas de reducción de la dimensión que obtienen los mejores resultados en la evaluación de la calidad son *SNE*, ACP y *t-SNE* (ver tabla 5-3). La

representación del conjunto de datos según *SNE* se muestra en la figura 5-7. En esta representación pueden distinguirse dos subconjuntos de elementos claramente distanciados, razón que explicaría la dificultad de formar un grafo que abarque a la totalidad del conjunto para los métodos *LLE* e *Isomap* con un valor pequeño en el parámetro *vecindario*.



(a) *(b)* *(c)* *(d)*

Figura 5-8: *Visualización del conjunto de datos Semillas en VDE según las variables longitud del sépalo (a), anchura del sépalo (b), longitud del pétalo (c) y anchura del pétalo (d).*

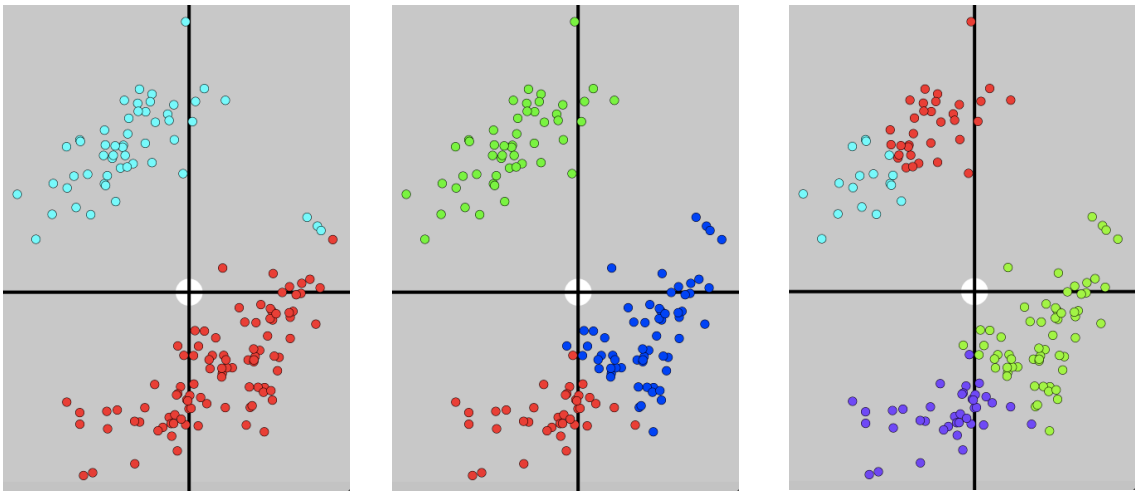
Los resultados sobre la distribución de los distintos atributos en el conjunto de datos se presentan en la figura 5-8. En la inspección visual de los resultados

puede apreciarse que en todos los atributos a excepción de la *anchura del sépalo* (figura 5-8(b)) hay una distinción entre los valores del subconjunto situado en la parte superior (que toma valores pequeños en todos ellos) y el subconjunto situado en la parte inferior donde los atributos aparecen en sentido creciente desde la parte superior derecha hasta la parte inferior izquierda (figuras 5-8(a), 5-8(c) y 5-8(d)). Dado que la diferencia entre los valores en el subgrupo situado en la parte superior e inferior es más acusada en las variables *longitud del pétalo* y *anchura del pétalo* (figuras 5-8(c) y 5-8(d)), éstas serán las más determinantes a la hora de formar agrupamientos y elaborar patrones de comportamiento.

Procediendo a realizar el análisis de agrupamiento, el método que obtiene mejores resultados, según los índices de validación, es *K-medias* utilizando la distancia *Euclídea* para 2, 3 ó 4 grupos (ver tabla 5-4). El resultado visual presentado por *VDE* es el que se muestra en la figura 5-9.

Tabla 5-4: *Valores de los índices de validación obtenidos por VDE en el agrupamiento del conjunto Iris para el método K-medias utilizando la distancia Euclídea.*

	Silh	*Dunn*	*SS*
2 grupos	0,767	0,961	0,576
3 grupos	0,660	0,798	0,393
4 grupos	0,542	0,503	0,471



(a) *(b)* *(c)*

Figura 5-9: *Distribución del conjunto de datos Semillas según el método de agrupamiento K-medias utilizando la distancia Euclídea. Resultado para 2 grupos (a), 3 grupos (b) y 4 grupos (c).*

La clasificación en cuatro grupos además de obtener resultados inferiores en la validación (ver tabla 5-4) también se observa en la figura 5-9 que el agrupamiento propuesto es básicamente igual que en la distribución en tres grupos con la diferencia de que el grupo superior queda dividido en dos subgrupos, hecho poco probable de acuerdo a la homogeneidad de este grupo ya comentada en el análisis por variables. También es significativo que la clasificación en dos grupos no distinga perfectamente los dos subgrupos que aparecen separados gráficamente (resultado que también se obtiene al realizar el agrupamiento según *FCM*). Una explicación a este resultado puede ser debida al comportamiento de la variable *longitud del sépalo* (ver figura 5-8(a)).

El agrupamiento original de los datos se muestra en la figura 5-10. En este agrupamiento se observa que la clase "*Setosa*" (elementos de color verde) se corresponde con el subgrupo situado en la parte superior mientras que el subgrupo de la parte inferior se compone de los elementos de las clases "*Virginica*" (en rojo) y "*Versicolour*" (en azul).

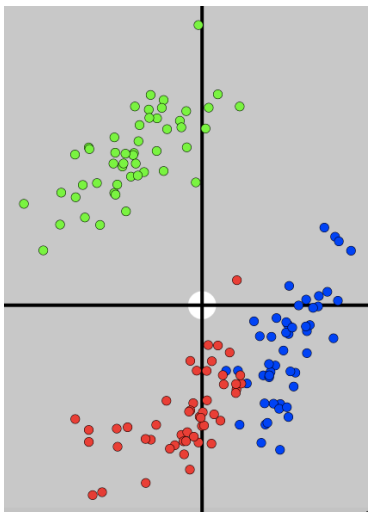
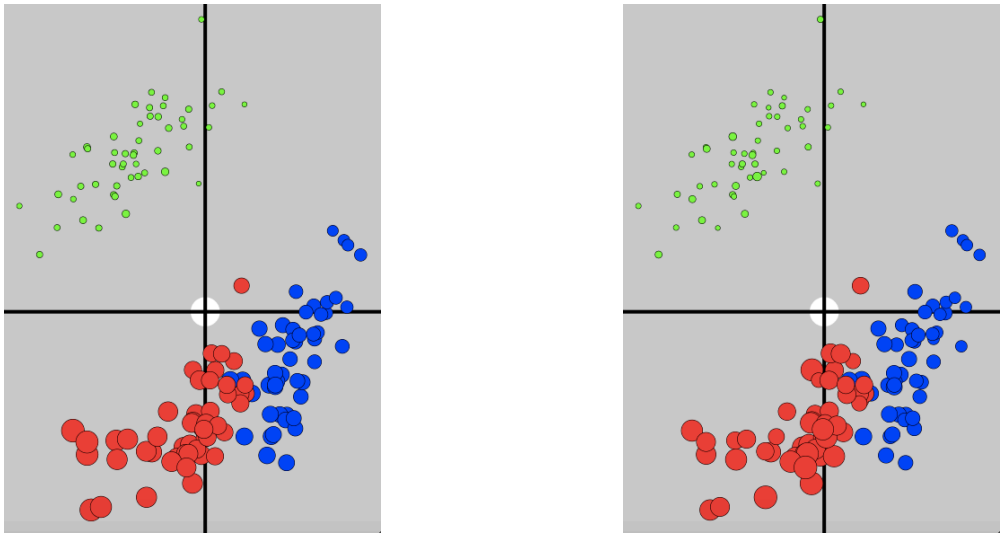


Figura 5-10: *Agrupamiento original del conjunto de datos Iris. Los elementos de la clase "Virginica" aparecen en color rojo, los de la clase "Setosa" en verde y los de la clase "Versicolour" en azul.*

Comparando las figuras 5-9(b) y 5-10 puede observarse que el agrupamiento realizado por *K-medias* utilizando la distancia *Euclídea* para 3 grupos distribuye los datos de forma muy similar a la original. La clase "*Setosa*" queda perfectamente clasificada y diferenciada de las otras dos mientras que entre las clases "*Virginica*" y "*Versicolour*" hay ciertos errores de clasificación. En este caso, el agrupamiento propuesto por *K-medias* para 3 grupos clasifica correctamente un 89,5 % de los datos.

Analizando la distribución de los atributos en los grupos originales, se puede observar que los dos que más influencia tienen a la hora de distinguirlos son la *longitud del pétalo* y la *anchura del pétalo*, conclusión a la que ya se llegó anteriormente en el análisis de variables (ver figura 5-11). Ambas variables son apropiadas para distinguir las tres clases aunque visualmente se observa más adecuada la variable *anchura del pétalo* para distinguir las clases "*Virginica*" y "*Versicolour*".



(a) *(b)*

Figura 5-11: *Visualización del agrupamiento original del conjunto de datos Iris de acuerdo con los atributos longitud del pétalo (a) y anchura del pétalo (b). Los elementos de la clase "Virginica" aparecen en color rojo, los de la clase "Setosa" en verde y los de la clase "Versicolour" en azul.*

En cuanto a los elementos mal clasificados, se observa que hay catorce elementos de la clase "*Virginica*" que son clasificados como "*Versicolour*". En las estadísticas de estos elementos puede observarse que la única variable que los hubiera clasificado en la clase correcta ("*Virginica*") es la *anchura del pétalo*, por tanto se confirma una vez más la importancia de ésta en la correcta clasificación de los elementos. Por otro lado, los elementos de la clase "*Versicolour*" que son clasificados como "*Virginica*" son solamente dos: el número 53 y el 78. En cuanto al elemento 53, si se atiende a la información de la variable *anchura del pétalo* se clasificaría correctamente mientras que el elemento 78 posee unas estadísticas propias de la clase "*Virginica*" en todos sus atributos a pesar de ser de la clase "*Versicolour*".

Finalmente, se procede al análisis visual del conjunto de datos según la técnica supervisada de reducción de la dimensionalidad *LDA*; la representación

gráfica se muestra en la figura 5-12. El resultado es muy similar al obtenido en la representación anterior: la clase "*Setosa*" aparece claramente separada de las otras dos y las clases "*Virginica*" y "*Versicolour*" vuelven a mostrarse muy próximas aunque quedan sensiblemente mejor diferenciadas. No obstante, es poco significativa la mejora que aporta *LDA* con respecto a la propuesta obtenida utilizando *SNE*.

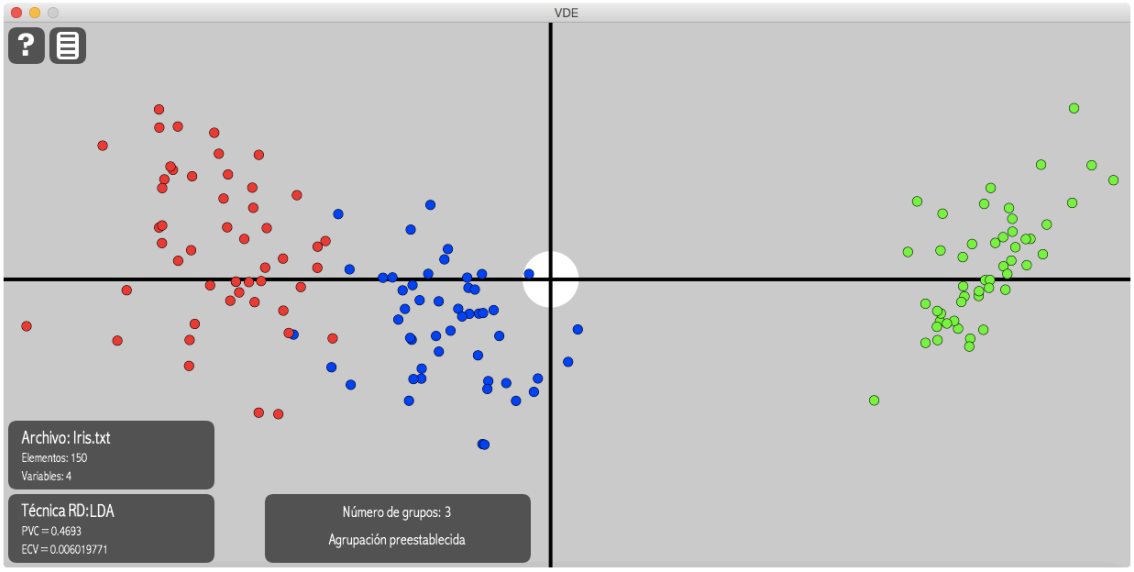


Figura 5-12: *Conjunto de datos Iris según la técnica de reducción de la dimensionalidad LDA. Los elementos de la clase "Virginica" aparecen en color rojo, los de la clase "Setosa" en verde y los de la clase "Versicolour" en azul.*

5.4. Conjunto de datos *Wine*

5.4.1. Descripción de la base de datos

La base de datos *Wine* (Lichman, 2013) refleja un análisis químico de vinos correspondientes a tres cultivos diferentes en una misma región de Italia. Este conjunto consta de 178 elementos de los cuales 59 son de la clase *"1"*, 71 de la clase *"2"* y 48 de la clase *"3"*. Cada dato consta de 13 atributos que se corresponden con las cantidades de medias de cada uno de los siguientes componentes: *alcohol*, *ácido málico*, *cenizas*, *alcalinidad de las cenizas*, *magnesio*, *fenoles en total*, *flavonoides*, *fenoles no flavonoides*, *pro-antocianidinas*, *intensidad de color*, *matiz de color*, *OD280/OD315* y *prolina*.

5.4.2. Análisis de resultados

Los valores de los parámetros *vecindario* y *difusión* escogidos son 15 y 2 respectivamente. Las técnicas de reducción de la dimensión que obtienen los mejores resultados en la evaluación de la calidad son ACP, *Isomap* y *t-SNE* (ver tabla 5-5). La representación del conjunto de datos según ACP se muestra en la figura 5-13.

Tabla 5-5: *Valores de* PVC *y* ECV *obtenidos por VDE en la evaluación de la calidad de la inmersión realizada en el cojunto de datos Wine por los métodos ACP, Isomap y t-SNE.*

	PVC	ECV
ACP	0,993	0,000
Isomap	0,928	0,000
t-SNE	0,913	0,001

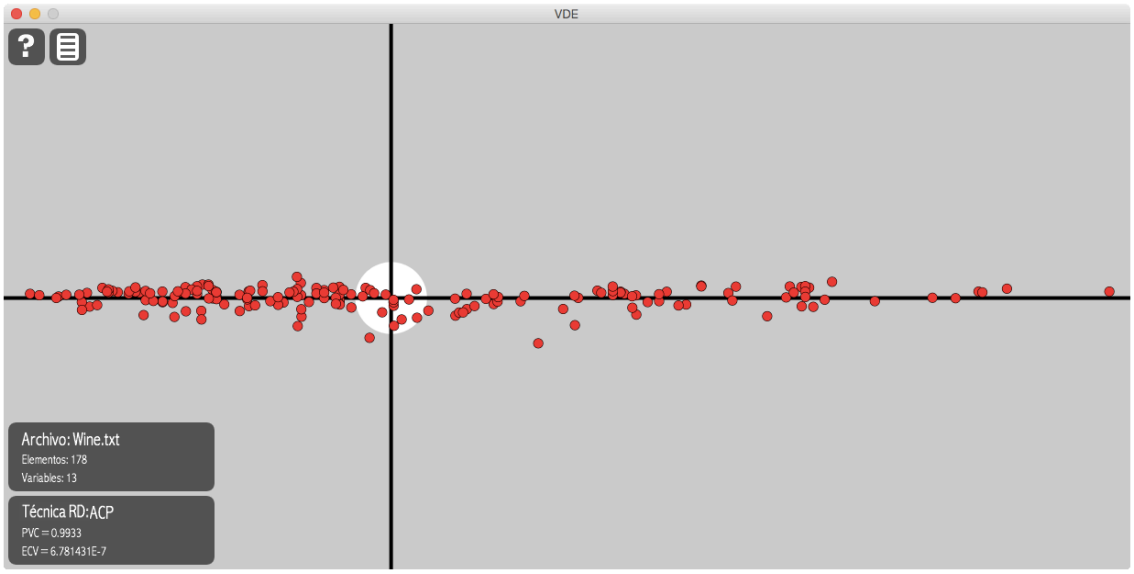
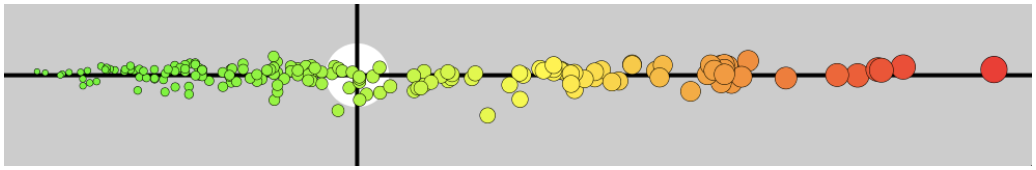


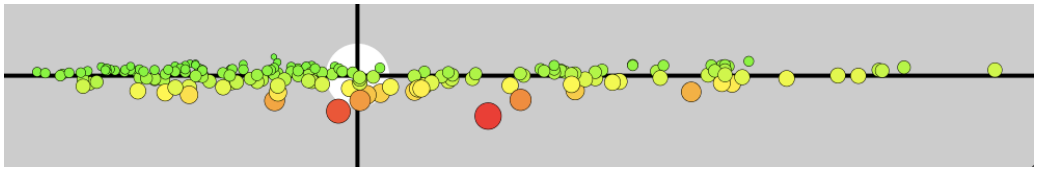
Figura 5-13: *Conjunto de datos Wine según la técnica de reducción de la dimensionalidad* ACP.

Analizando visualmente la distribución de los distintos atributos en el conjunto de datos, puede observarse que aquellos que explican mejor las componentes principales son *prolina* en el eje horizontal y *magnesio* en el vertical

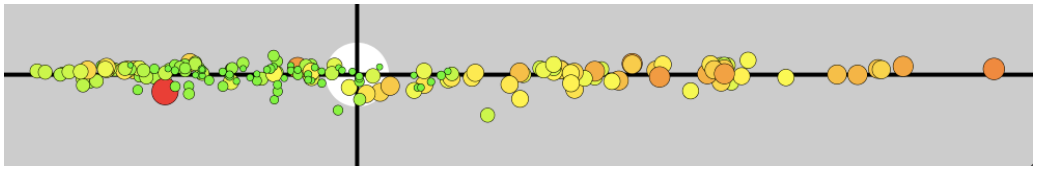
(figuras 5-14(a) y 5-14(b) respectivamente). Dado que se observa una disposición predominante en el eje horizontal puede deducirse que aquellos atributos cuya distribución sea acorde a este eje pueden ser importantes para explicar el comportamiento del conjunto. En este sentido, los atributos *flavonoides* (ver figura 5-14(c)) e *intensidad del color* (ver figura 5-14(d)) deben ser tenidos en cuenta aunque su influencia sea menor que en el caso del atributo *prolina*.



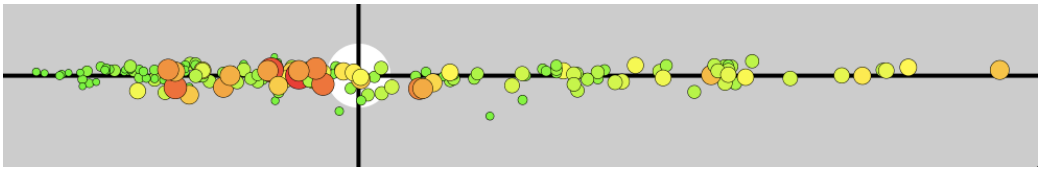
(a)



(b)



(c)



(d)

Figura 5-14: *Visualización del conjunto de datos Wine en VDE según las variables prolina (a), magnesio (b), flavonoides (c) e intensidad del color (d).*

El análisis de agrupamiento no resulta muy revelador para este conjunto de datos. Los índices de validación no precisan de forma clara la cantidad de grupos en los que clasificar el conjunto de datos. Entre los métodos, tampoco se aprecia diferencia significativa entre *FCM* y *K-medias.* En la tabla 5-3 se detallan los valores de estos índices para 3, 4 y 5 grupos que es donde se han apreciado mejores resultados de validación. En la figura 5-15 se muestran los distintos agrupamientos obtenidos en estos casos. Puede observarse que los resultados son

muy similares para los métodos *FCM* y *K-medias* tendiéndose, en ambos casos, a organizar distintos grupos a lo largo del eje horizontal como era de esperar.

Tabla 5-6: *Valores de los índices de validación obtenidos por VDE en distintos agrupamientos del conjunto Wine para los métodos FCM y K-medias.*

	FCM			*K-medias*		
	Silh	*Dunn*	*SS*	*Silh*	*Dunn*	*SS*
3 grupos	0,662	0,558	0,445	0,671	0,534	0,467
4 grupos	0,660	0,589	0,327	0,658	0,579	0,328
5 grupos	0,630	0,542	0,282	0,646	0,561	0,275

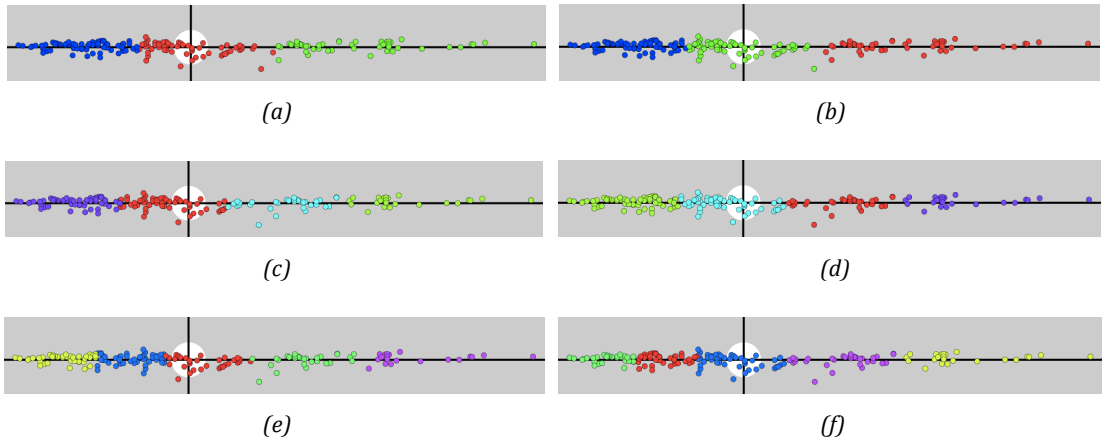


Figura 5-15: *Distribución del conjunto de datos Wine. En las figuras (a), (c) y (e) se presentan los resultados según el método FCM para 3, 4 y 5 grupos respectivamente mientras que en (b), (d) y (f) se muestran los obtenidos según K-medias también para 3, 4 y 5 grupos de forma respectiva.*

El agrupamiento original de los datos se muestra en la figura 5-16. En este agrupamiento se observa que la clase "*1*" (elementos de color verde) queda algo más diferenciada de las otras dos situándose en la parte derecha de la representación según ACP mientras que los subgrupos correspondientes a las clases "*3*" (en rojo) y "*2*" (en azul) ocupan la parte izquierda no siendo nada clara la diferenciación entre ambas.

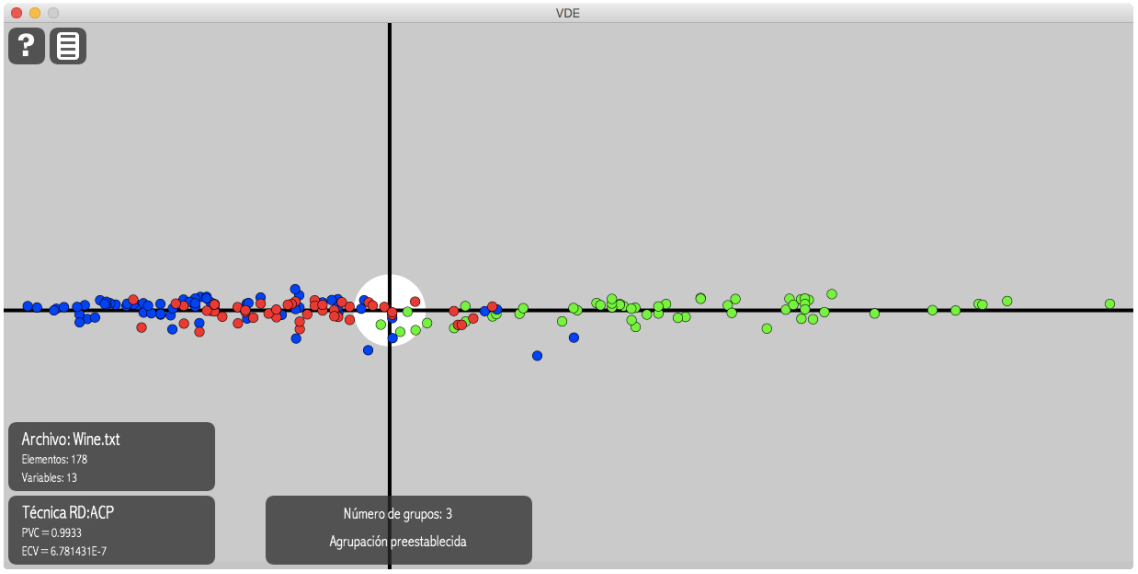


Figura 5-16: *Agrupamiento original del conjunto de datos Wine. Los elementos de la clase "1" aparecen en color verde, los de la clase "2" en azul y los de la clase "3" en rojo.*

Para realizar el análisis que pueda facilitar información acerca del comportamiento de cada variable en cada una de las clases en las que el grupo de datos queda clasificado, interesa representar el conjunto de datos mediante aquella técnica que sea capaz de separarlas gráficamente de la forma más clara. En este caso, el método supervisado *LDA* obtiene un resultado adecuado para este fin (ver figura 5-17).

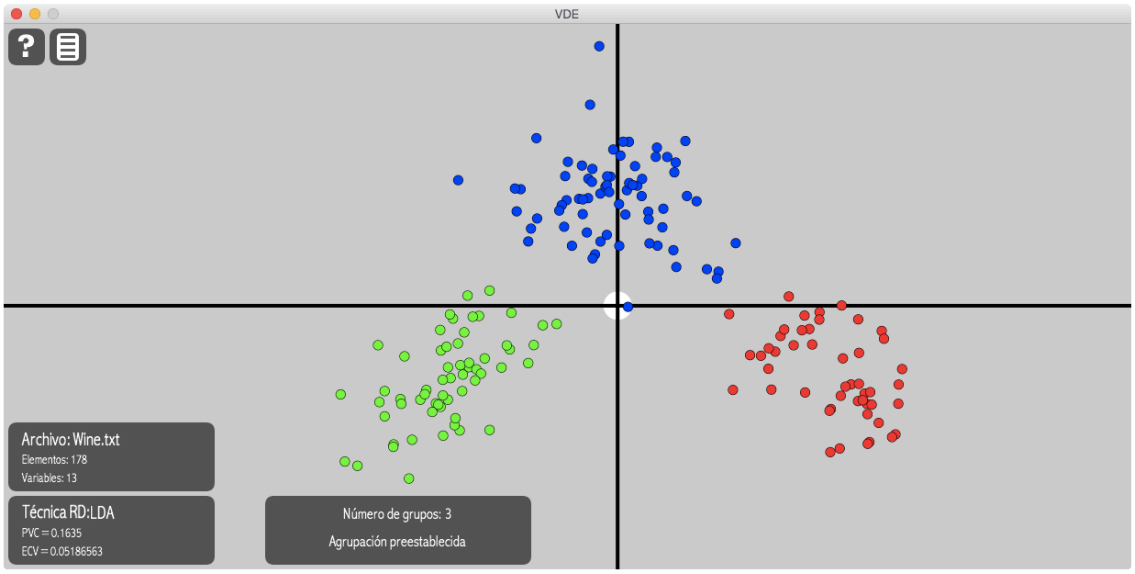
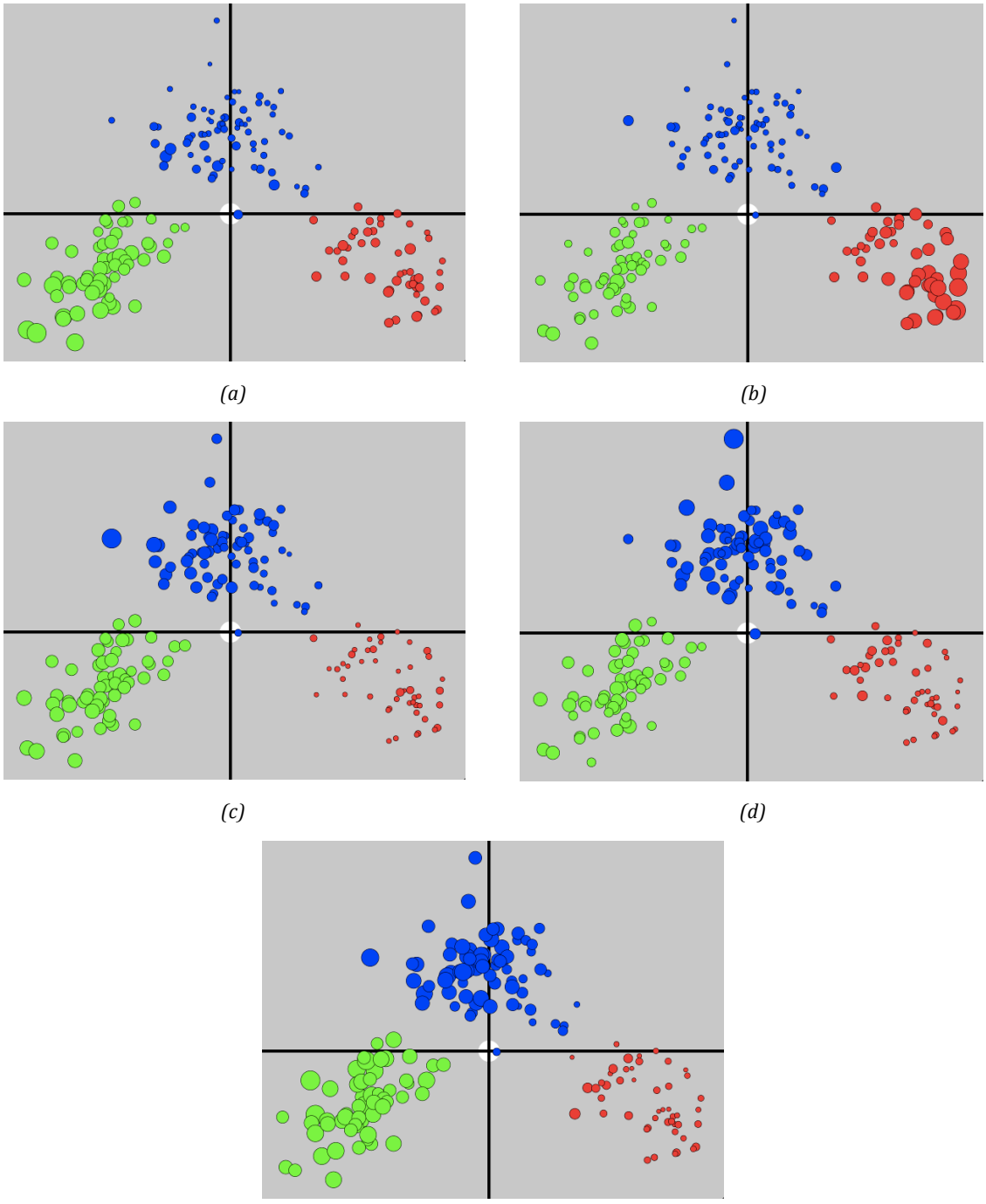


Figura 5-17: *Conjunto de datos Wine según la técnica de reducción de la dimensionalidad LDA. Los elementos de la clase "1" aparecen en color verde, los de la clase "2" en azul y los de la clase "3" en rojo.*



(a) *(b)* *(c)* *(d)* *(e)*

Figura 5-18: *Visualización del agrupamiento original del conjunto de datos Wine de acuerdo con los atributos prolina (a), intensidad del color (b), flavonoides (c), matiz de color (d) y OD280/OD315 (e) según la técnica LDA. Los elementos de la clase "1" aparecen en color verde, los de la clase "2" en azul y los de la clase "3" en rojo.*

Procediendo al análisis supervisado de las variables, aquellas que aportan más información acerca de las clases en las que está distribuido el conjunto de datos son *prolina*, *intensidad del color*, *flavonoides*, *matiz de color* y *OD280/OD315*

(ver figura 5-18). En la figura 5-18(a), correspondiente a la variable *prolina*, puede observarse cómo la clase *"1"* (elementos de color verde) se diferencia de las otras dos tomando valores notablemente superiores en esta clase. Esta variable ya fue apuntada en el análisis correspondiente como importante en el estudio del conjunto de datos por su influencia en la primera componente principal en el análisis ACP y el hecho de que con esta técnica la clase *"1"* quede representada de forma diferenciada respecto de las otras dos (ver figura 5-16) hace que sea un atributo adecuado para la diferenciación de esta clase. La figura 5-18(b) se corresponde con la distribución de la variable *intensidad de color* que es la que mejor diferencia la clase *"2"* (color azul) de las otras dado el bajo valor que en general toman sus elementos en esta variable. La distribución de las variables *flavonoides*, *matiz de color* y *OD280/OD315* se muestra respectivamente en las figuras 5-18(c), 5-18(d) y 5-18(e). En estas figuras se observa que pueden diferenciar la clase *"3"* de las otras dos dado los bajos valores de estas variables en los elementos de esta clase respecto de las otras dos.

Otra conclusión interesante que puede deducirse del análisis de variables en las distintas clases es que así como las clases *"1"* y *"3"* tienen un comportamiento más homogéneo (tamaños semejantes en la representación gráfica) en los cinco atributos comentados (ver figura 5-18), la clase *"2"* presenta un comportamiento más disperso (tamaños más dispares en la representación) lo que hace que sea más difícil su diferenciación respecto de las otras dos clases.

5.5. Conjunto de datos *E.coli*

5.5.1. Descripción de la base de datos

La base de datos *E.coli* presenta un problema de clasificación de proteínas E.coli en distintas localizaciones celulares a partir de sus secuencias de aminoácidos. El conjunto consta de 336 elementos donde cada uno de ellos cuenta con 7 atributos obtenidos a partir de distintos métodos de reconocimiento, detección y medición realizados sobre las proteínas. Estos atributos son *mcg*, *gvh*, *aac*, *alm1* y *alm2* que toman valores continuos reales, y *lip* y *chg* que toman únicamente dos valores (Lichman, 2013). Cada dato, además, posee un nombre que es utilizado por *VDE* como etiqueta del mismo.

Los datos están clasificados en 8 clases que hacen referencia al lugar de localización de la proteína. Estas clases son *"cp"* (citoplasma) con 143 elementos,

"im" (membrana interna sin secuencia de señal) con 77 elementos, *"pp"* (periplasma) con 52 elementos, *"imU"* (membrana interna con secuencia de señal no escindible) con 35 elementos, *"om"* (membrana externa) con 20 elementos, *"omL"* (lipoproteína en membrana externa) con 5 elementos, *"imL"* (lipoproteína en membrana interna) con 2 elementos e *"imS"* (membrana interna con secuencia de señal escindible) con 2 elementos.

5.5.2. Análisis de resultados

Los valores escogidos para los parámetros *vecindario* y *difusión* son 15 y 2 respectivamente. Para este valor de *vecindario*, el grafo construido por las técnicas *Isomap* y *LLE* no logra conectar todos los elementos del conjunto (solamente 325 de los 336). Aumentando el valor de este parámetro se consigue conectar todos los elementos al utilizar esas técnicas de reducción de la dimensionalidad pero no se mejora la evaluación de la calidad por lo que se mantiene el valor inicial de 15 en el análisis con *VDE*.

Tabla 5-7: *Valores de* PVC *y* ECV *obtenidos por VDE en la evaluación de la calidad de la inmersión realizada en el cojunto de datos E.coli por los métodos t-SNE, SNE e Isomap.*

	PVC	ECV
t-SNE	0,640	0,016
SNE	0,495	0,016
Isomap	0,443	0,016

La técnica de reducción de la dimensión que obtiene el mejor resultado en la evaluación de la calidad es *t-SNE* seguida de *SNE* e *Isomap* (ver tabla 5-7). La representación del conjunto de datos según *t-SNE* se muestra en la figura 5-19. En un primer análisis visual se pueden observar tres subgrupos diferenciados. Si se selecciona la opción *Vista en transición* del submenú *Técnica RD* entre los métodos *t-SNE* y *LLE* o *Isomap* se advierte que el subgrupo central, que está formado por sólo 10 elementos, no aparece representado por estos dos últimos métodos (de los 11 que quedan fuera del grafo construido para estas técnicas). De ahí la dificultad por parte de *Isomap* y *LLE* en conectar a todos los elementos del conjunto.

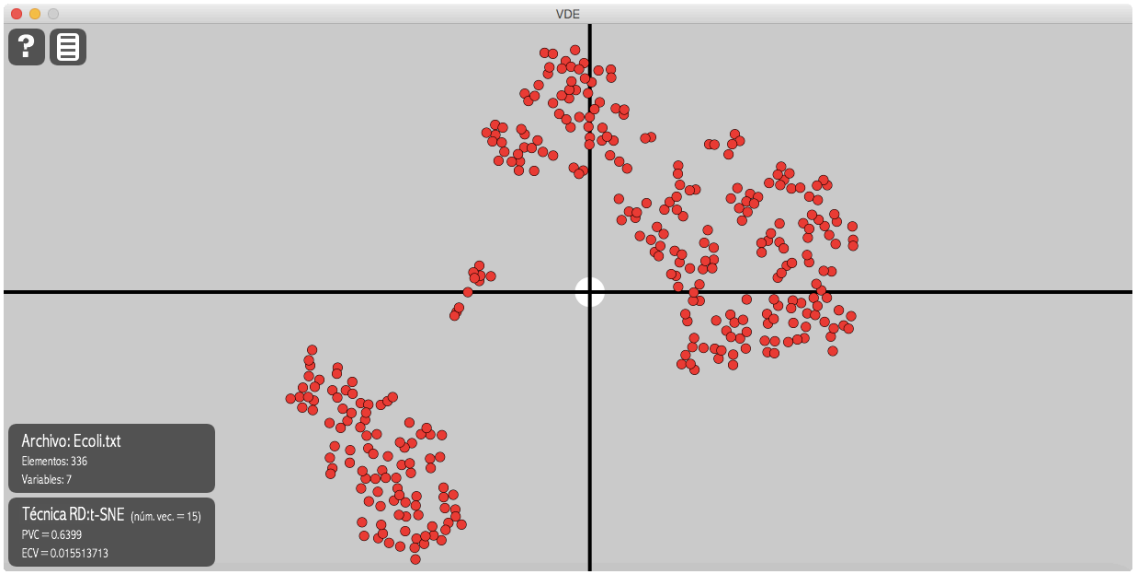
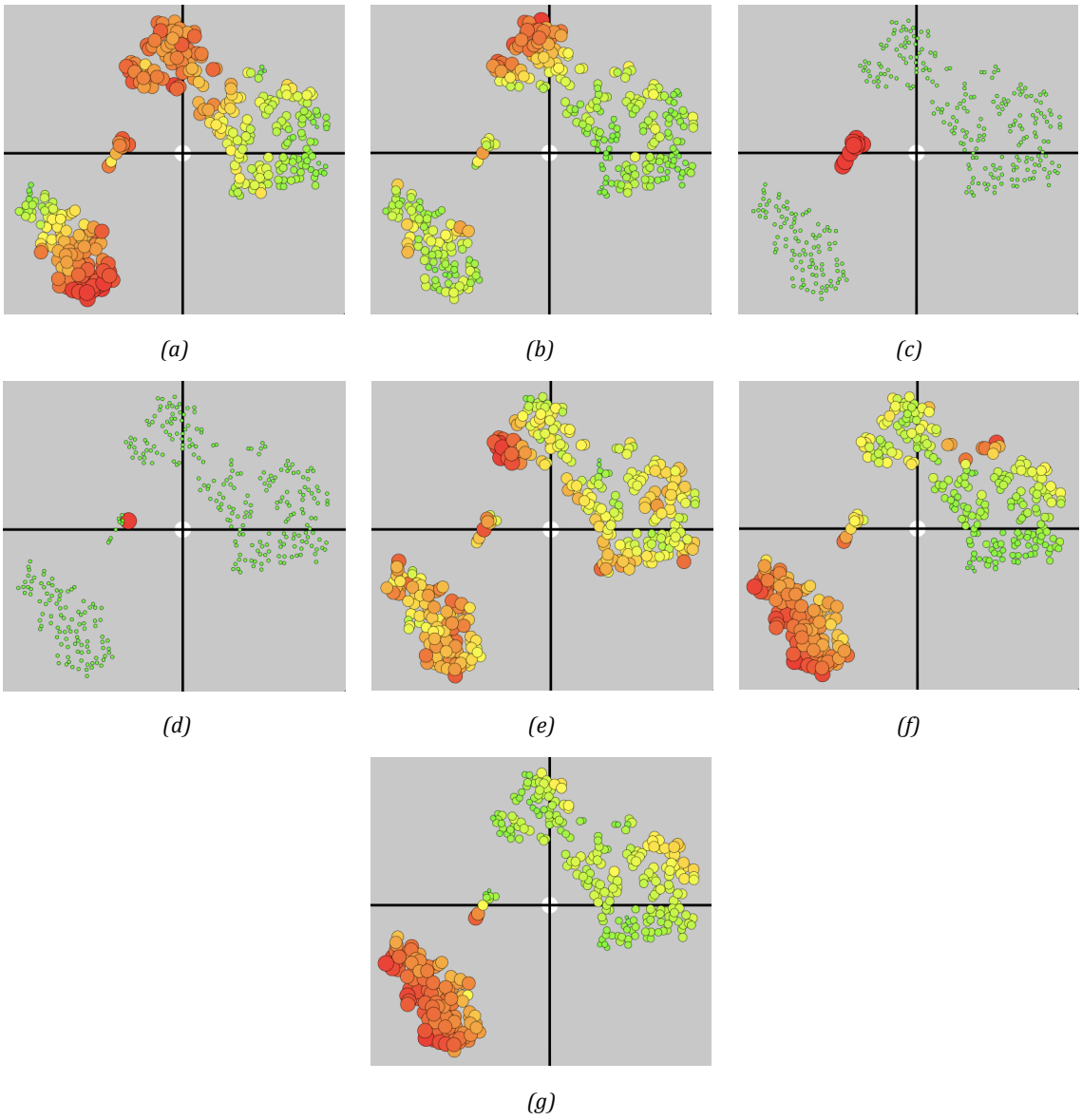


Figura 5-19: *Conjunto de datos E.coli según la técnica de reducción de la dimensionalidad t-SNE.*

En el análisis de variables se consiguen los resultados mostrados en la figura 5-20 a partir de los cuales puede obtenerse la siguiente información:

- El atributo *lip* (figura 5-20(c)) diferencia perfectamente el subgrupo de elementos central de los otros dos.
- Los atributos *alm1* y *alm2* (figuras 5-20(f) y 5-20(g) respectivamente) establecen una clara distinción entre los dos grandes subgrupos, principalmente *alm2* que obtiene unos valores notoriamente más elevados en el subgrupo representado en la parte inferior de la pantalla. En cuanto al subgrupo central, éste obtiene valores similares en *alm1* y dispares en *alm2*.
- El atributo *chg* (figura 5-20(d)) toma un valor de 0,5 para todos los elementos del conjunto de datos a excepción del elemento 223 (*NLPA_ECOLI*) donde toma un valor de 1.
- El atributo *mcg* (figura 5-20(a)) establece distinción dentro de cada uno de los dos grandes subgrupos y toma valores similares en el subgrupo central. En el subgrupo situado de la parte alta se observa una tendencia decreciente desde su parte superior hacia la inferior y al contrario en el subgrupo ubicado en la parte baja de la representación gráfica.
- El atributo *gvh* (figura 5-20(b)) mantiene valores aceptablemente homogéneos en todo el conjunto a excepción de una colección de elementos situados en la parte alta del subgrupo representado en la parte superior.

- El atributo *aac* (figura 5-20(e)) tiene un comportamiento parecido a *gvh* aunque la colección de elementos donde toma un valor marcadamente superior es más reducido.



(a) *(b)* *(c)*

(d) *(e)* *(f)*

(g)

Figura 5-20: *Visualización del conjunto de datos E.coli según las variables mcg (a), gvh (b), lip (c), chg (d), aac(e), alm1(f) y alm2 (g).*

En referencia al análisis de agrupamiento, la información obtenida en el estudio de variables ya hace ver la posible existencia de una cantidad de grupos superior a los tres que aparecen definidos en la representación gráfica según *t-SNE* debido principalmente a la información obtenida a partir de los atributos *mcg*, *gvh* y *aac* que rompen con la homogeneidad en estos subgrupos representados

gráficamente. El método que obtiene mejores resultados según los índices de validación en el análisis de agrupamiento es *K-medias* con la distancia *Euclídea* aunque no queda patente el número óptimo de grupos (ver tabla 5-8).

Tabla 5-8: *Valores de los índices de validación obtenidos por VDE en distintos agrupamientos del conjunto E.coli para el método K-medias utilizando la distancia Eucídea.*

	Silh	*Dunn*	*SS*
2 grupos	0,542	0,662	2,838
3 grupos	0,542	0,616	1,996
4 grupos	0,500	0,495	2,170
5 grupos	0,455	0,445	2,570
6 grupos	0,413	0,474	2,380
7 grupos	0,34	0,45	2,700
8 grupos	0,336	0,437	2,800
9 grupos	0,380	0,416	2,880
10 grupos	0,344	0,446	3,030

El agrupamiento original de los datos se muestra en la figura 5-21. Comparando esta figura con las de la figura 5-20, puede establecerse una caracterización de los distintos grupos de acuerdo con los valores que sus elementos toman en cada atributo y de este modo obtener conocimiento acerca del conjunto de datos. De esta comparación, y atendiendo a la información obtenida en el análisis por variables, pueden extraerse la siguientes conclusiones:

- El atributo *lip* (figura 5-20(c)) es idóneo para distinguir las clases *"omL"* e *"imL"* del resto ya que todos sus elementos se sitúan en el subgrupo central que es donde este atributo toma valores altos. No obstante, existen en este subgrupo elementos pertenecientes a otras clases que deberán ser diferenciados de las clases *"omL"* e *"imL"* atendiendo a otros

atributos. Si un elemento adquiere valores bajos en el atributo *lip* puede descartarse su pertenencia a las clases "*omL*" o *"imL"*.

- Los atributos *alm1* y *alm2* (figuras 5-20(f) y 5-20(g)) distinguen los otros dos subgrupos entre sí. Estos atributos adquieren valores altos en el subgrupo situado en la parte inferior de la representación gráfica, que es donde se hallan principalmente los elementos de las clases *"im"* e *"imU"*, y valores bajos en los elementos del subgrupo situado en la parte superior donde se encuentran básicamente las clases *"cp"*, *"om"* y *"pp"*.
- El atributo *mcg* (ver figura 5-20(a)) establece distinción entre los elementos del subgrupo situado en la parte inferior de la representación visual. Ello permite diferenciar las clases *"im"* e *"imU"* entre sí ya que este atributo adquiere valores altos en la parte inferior de este subgrupo (donde están situados los elementos de la clase *"imU"*) y bajos en la parte superior del mismo (donde están situados los elementos de la clase *"im"*).
- De igual modo, el atributo *mcg* es adecuado para distinguir las clases en el subgrupo representado en la parte alta. Mientras que este atributo adquiere valores altos en la parte superior de este subgrupo (donde se sitúan casi exclusivamente los elementos de las clases *"imS"*, *"om"* y *"pp"*), toma valores bajos en la parte inferior (donde se sitúa la representación gráfica de la clase *"cp"*). Esto permitiría distinguir la clase *"cp"* del resto de clases dentro de este subgrupo.

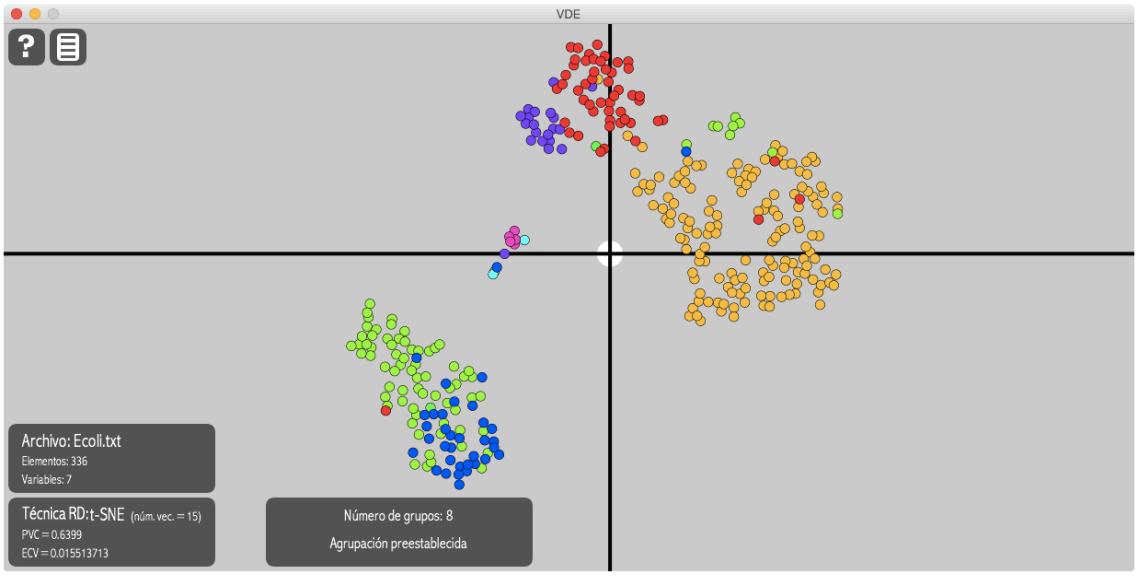


Figura 5-21: *Agrupamiento original del conjunto de datos E.coli. Los elementos de la clase "cp" aparecen en color amarillo, los de la clase "im" en verde claro, los de la clase "imS" color verde oscuro, los de la clase "imL" en azul claro, los de la clase "imU" en azul oscuro, los de la clase "om" en color morado, los de la clase "omL" en un color púrpura y los de la clase "pp" en rojo.*

- El atributo *gvh* puede distinguir la clase *"imS"* (donde toma valores bajos en sus dos elementos) de las clases "om" y *"pp"*
- El atributo *aac* (ver figura 5-20(e)) distingue los elementos de la clase *"om"* del resto ya que toma en ella valores más altos que en el resto.

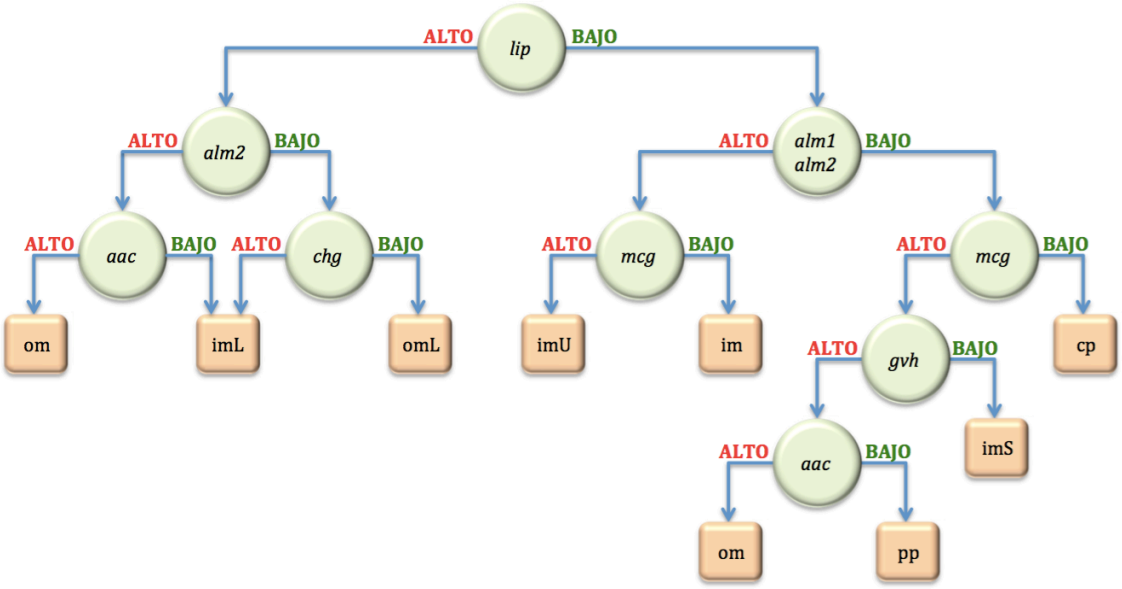


Figura 5-22: *Diagrama en forma de árbol sobre la caracterización de las distintos clases en relación con los valores de los atributos. Cada nodo circular representa a un atributo y cada nodo cuadrangular una clase. Para cada elemento se estima en cada nodo circular, si el valor en el atributo correspondiente es alto, debiéndose seguir en este caso el itinerario de la izquierda, o si es bajo, tomando en este caso el trayecto de la derecha. Al final del recorrido se estima la clase a la que el elemento pertenece.*

Toda esta información puede sintetizarse a modo de diagrama en forma de árbol como se muestra en la figura 5-22. Para cada elemento que se desee clasificar se comienza por la parte superior del árbol estimando si su valor en el atributo *lip* es alto o bajo. Si este valor es alto, el elemento pertenece como se ha comentado al subgrupo situado en la parte central de la representación gráfica según *t-SNE*. En este subgrupo se hallan todos los elementos de la clase "*omL"* e "*imL"* además de otros tres elementos alejados de las zonas donde sus clases son mayoritarias.

Afinando en la distinción de los elementos del subgrupo central, puede observarse que el atributo *alm2* diferencia a la clase "*omL*" del resto al tomar valores inferiores en este grupo respecto del resto a excepción del elemento 223 (*NLPA_ECOLI*) de la clase *"imL"* que puede diferenciarse de la clase *"omL"* mediante el atributo *chg* (único elemento que toma en este parámetro el valor 1)(ver figuras 5-20(g) y 5-20(d) en comparación con la figura 5-21). Si los atributos *lip* y *alm2* toman valores altos, la variable *aac* puede diferenciar el elemento que falta de la clase *"imL"* del resto por tomar en éste un valor inferior. De los tres elementos que

quedan, el perteneciente a la clase *"om"* es el que toma el valor más alto en el atributo *aac* siendo los otros dos elementos (el 183 (*MCP3_ECOLI*) de la clase *"im"* y 252 (*RHAT_ECOLI*) de la clase *"imU"*) más complicados de clasificar. Todo el razonamiento de este párrafo queda plasmado en la parte superior izquierda del diagrama una vez se ha observado un valor alto en el atributo *lip*.

Volviendo a la parte superior del diagrama, para un valor bajo del atributo *lip* se encuentran los otros dos grandes subgrupos que aparecen en la representación según *t-SNE*. En este caso, los atributos *alm1* y *alm2* pueden diferenciarlos tal y como se comentó en el análisis por variables. Para valores altos de estos atributos se encuentran las clases *"imU"* e *"im"* correspondientes al subgrupo representado en la parte inferior. Estas dos clases pueden distinguirse entre sí a partir de la variable *mcg* que suele tomar valores altos en la clase *"imU"* y bajos en la clase *"im"* como ya se argumentó. Este planteamiento queda plasmado en el diagrama siguiendo el itinerario correspondiente a un valor bajo en el atributo *lip* y un valor alto en los atributos *alm1* y *alm2*.

Regresando de nuevo a la parte superior del diagrama y atendiendo a la información obtenida en el análisis por variables, los elementos situados en el subgrupo representado en la parte superior de la pantalla mediante *t-SNE* tienen un valor en el atributo *lip* bajo junto con valores bajos de *alm1* y *alm2*. En esta parte del diagrama, de nuevo la variable *mcg* ayuda a diferenciar la clase *"cp"* del resto de clases ya que sus elementos toman valores bajos en esta variable. Si por el contrario *mcg* toma un valor alto entonces el atributo *gvh* es el que distingue para valores bajos la clase "*imS"* (de tan sólo dos elementos) de las clases *"om"* y *"pp"* donde toma valores altos. Finalmente, estas dos clases pueden diferenciarse entre sí mediante el atributo *aac* que es, de hecho, el que mejor distingue la clase *"om"* del resto al adquirir en ella valores altos. Este argumento es el que se refleja en la parte inferior derecha del diagrama de la figura 5-22.

Para finalizar, indicar que el análisis del conjunto de datos mediante la técnica de reducción de la dimensión *LDA* no mejora la visualización de las distintas clases respecto a *t-SNE*.

5.6. Conjunto de datos *Modelado del usuario*

5.6.1. Descripción de la base de datos

El modelado del usuario consiste en la realización de tareas adaptadas a

cada individuo y a cada situación. Este modelado se realiza en función de la construcción de un modelo que guarda las características, o atributos, de un determinado usuario y es utilizado principalmente en entornos de aprendizaje en línea (Kahraman, Sagiroglu, & Colak, 2013). La base de datos *Modelado del usuario* (Lichman, 2013) consta de un conjunto de entrenamiento y un conjunto de validación de 258 y 145 elementos respectivamente donde cada uno de ellos cuenta con 5 atributos: *STG* (medida del tiempo de estudio), *SCG* (medida del número de repeticiones), *PEG* (medida del rendimiento del usuario en los exámenes), *STR* (medida del tiempo de estudio) y *LPR* (grado de aprendizaje). Todos ellos toman valores reales normalizados.

Los elementos están clasificados en 4 clases o niveles de conocimiento cada uno de los cuales cuenta con la cantidad de elementos en el conjunto de entrenamiento y validación que se detalla en la tabla 5-9.

Tabla 5-9: *Cantidad de elementos de la base de datos Modelado del usuario en cada una de sus clases para los conjuntos de entrenamiento y validación.*

Nivel de conocimiento (clase)	*Entrenamiento*	*Validación*
"muy bajo"	24	26
"bajo"	83	46
"medio"	88	34
"alto"	63	39

5.6.2. Análisis de resultados

Como se ha comentado en el apartado 5.1 el enfoque metodológico para este conjunto de datos se va a centrar en el análisis supervisado. El objetivo será analizar cómo *VDE* puede ayudar a caracterizar las clases de una forma visual e intuitiva a partir del conjunto de entrenamiento y verificar el resultado en conjunto de validación.

El conjunto de entrenamiento se analiza en *VDE* con valores de 15 y 2 para los parámetros *vecindario* y *difusión* respectivamente. La técnica de reducción de la dimensionalidad que obtiene mejor resultado en cuanto a la separación de las distintas clases es *LDA* a pesar de que *t-SNE* es la que obtiene un mejor resultado

en la evaluación de la calidad con un valor de PVC de 0,62.

La representación del conjunto de datos según *LDA* se muestra en la figura 5-23 donde los elementos aparecen ya representados por color según su clase original. En esta figura puede observarse cómo los elementos de la clase *"muy bajo"* (en color verde) se sitúan en la parte izquierda del conjunto seguidos según el eje horizontal hacia la derecha por los elementos de las clases *"bajo"*, *"medio"* y *"alto"* (en color azul, morado y rojo respectivamente). Esta disposición concuerda con la lógica de que las clases con un nivel de conocimiento parecido se encuentren más próximas que aquellas con un nivel de conocimiento más disparejo. También puede observarse cómo la clase *"alto"* aparece más aislada del resto mientras que la clase *"bajo"* aparece menos separada de las clases *"muy bajo"* y *"medio"*.

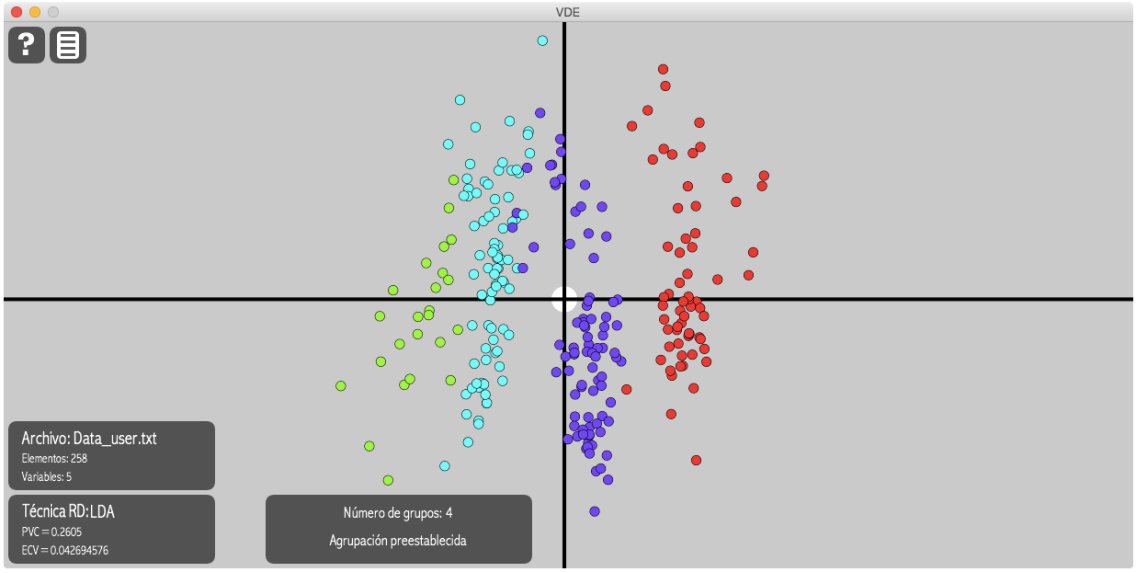


Figura 5-23: *Agrupamiento original del conjunto de datos de entrenamiento Modelado del usuario. Los elementos de la clase "muy bajo" aparecen en color verde, los de la clase "bajo" en azul claro, los de la clase "medio" color morado y los de la clase "alto" en rojo.*

Los resultados sobre la distribución de los distintos atributos en el conjunto de datos se muestra en la figura 5-24.

En una primera inspección visual, puede observarse que el atributo que mejor comportamiento tiene a la hora de distinguir las diferentes clases es *PEG* (ver figura 5-24(e)) debido a que toma valores poco dispersos dentro de cada una de las clases pero sí notablemente diferentes en las distintas clases, siendo crecientes conforme aumenta el nivel de conocimiento (inferiores en la clase *"muy bajo"* y superiores en la clase *"alto"*). Otro atributo interesante a observar es *LPR* (ver figura 5-24(d)) ya su distribución sigue una progresión creciente según el eje

vertical en sentido ascendente. Es especialmente significativo el observar que aquellos elementos de la clase *"medio"* que tienen valores bajos en *PEG* son los que obtienen mayores valores en *LPR* dentro del grupo y viceversa. Algo similar ocurre con la clase *"alto"* aunque de forma menos acusada. En cuanto al resto de atributos, no se observa un comportamiento tan determinante a la hora de distinguir las diferentes clases además de tener un rango de variación amplio dentro de cada una.

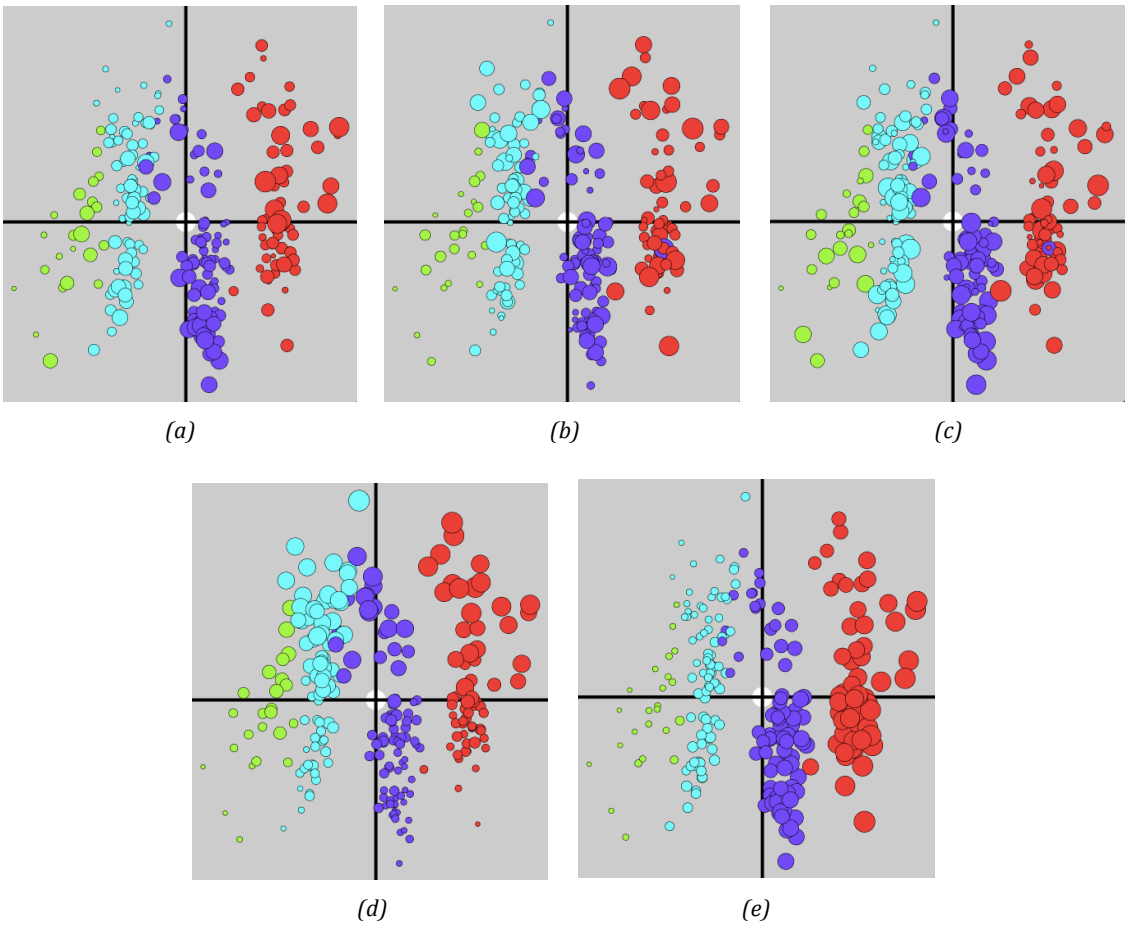


Figura 5-24: *Visualización del conjunto de datos Modelado del usuario en VDE según la distribución de las variables STG (a), SCG (b), STR (c), LPR (d) y PEG (e).*

Con este planteamiento, es sencillo determinar visualmente para cada uno de los atributos *LPR* y *PEG* cuáles son sus valores máximos y mínimos en cada una de las clases por la forma ya comentada en que éstos se distribuyen en el conjunto de datos y establecer un diagrama de flujo que permita clasificar los elementos. Una propuesta para este diagrama se muestra en la figura 5-25.

Como puede observarse en el diagrama de flujo, en casi todos los nodos de

decisión (con forma de rombo) aparecen condiciones relacionadas con los atributos *PEG* y *LPR* ya que son los que mejor distinguen las clases. Solamente en dos de estos nodos se ha tenido en cuenta los atributos *SCG* y *STR*. En ambos casos es debido a la dificultad de diferenciar la clase *"bajo"* de las clases *"medio"* y *"muy bajo"* como ya se comentó en la inspección visual del conjunto de datos. El modo en el que se logra construir el diagrama se describe a continuación.

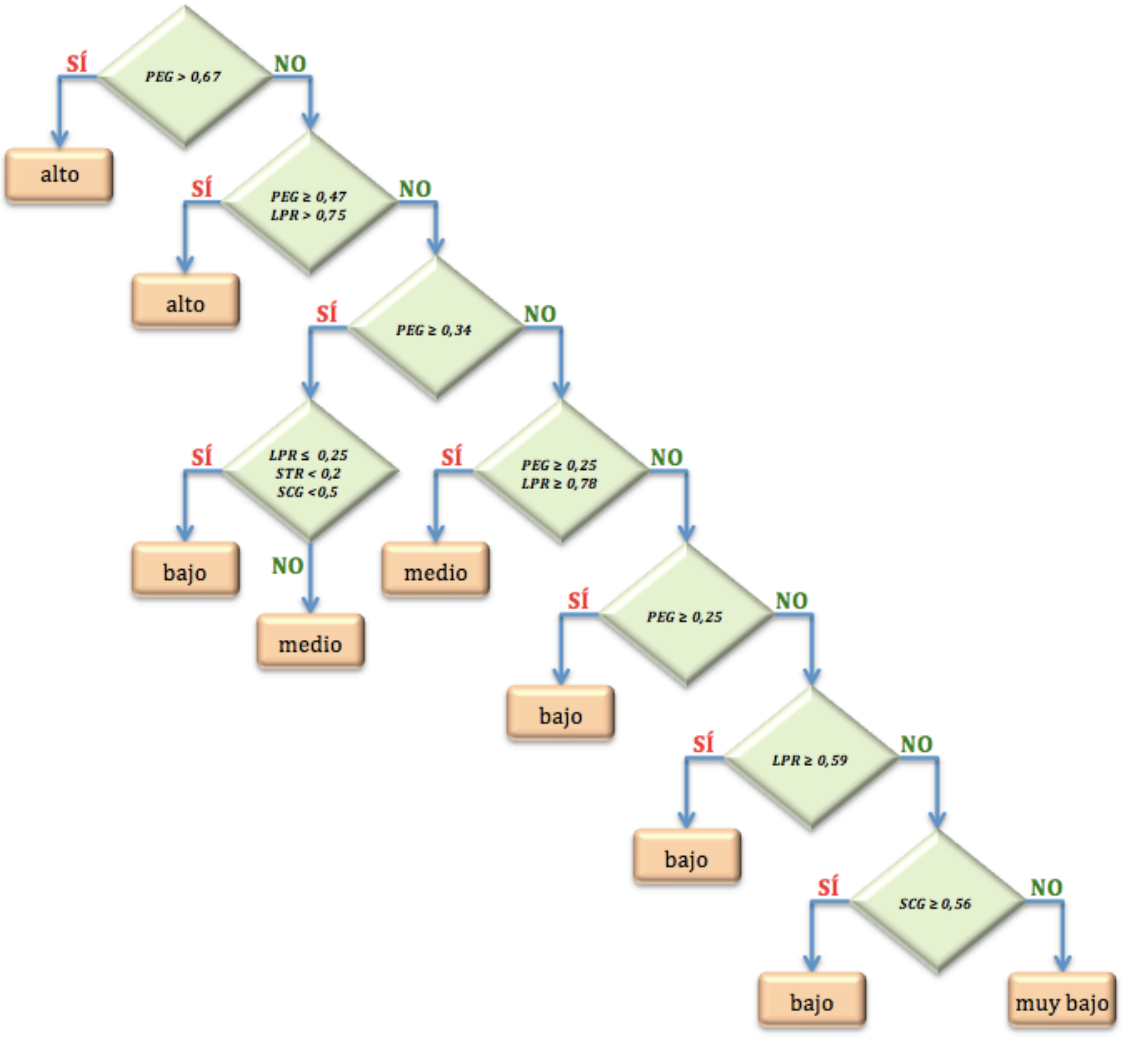


Figura 5-25: *Diagrama de flujo para la clasificación de un elemento en relación con los valores de sus atributos. En cada nodo en forma de rombo se estima si el atributo o atributos del elemento cumplen con las condiciones en él señaladas. En caso afirmativo, se sigue el itinerario de la izquierda y en caso contrario el de la derecha. Al final del recorrido se obtiene la clase a la que el elemento pertenece (nodo final en forma rectangular).*

Para construir el primer nodo ($PEG > 0{,}67$), se establece mediante inspección visual el máximo valor de *PEG* para la clase *"medio"* (círculos de mayor diámetro en esta clase en el modo de visualización según la viariable *PEG* (ver

figura 5-24(e)). Como en esta clase no se supera el valor de 0,67, si un elemento tiene un valor superior a éste se considera de la clase *"alto"*.

En caso contrario, para elaborar el siguiente nodo vuelve a inspeccionarse el atributo *PEG* ahora en la clase *"alto"* para hallar su valor mínimo resultando 0,47. Observando los elementos de la clase *"medio"*, aquellos que superan este valor en *PEG* son los situados en la parte baja de la representación visual según esta variable, mientras que los elementos de la clase "alto" que tienen menor puntuación en *PEG* se sitúan en la parte superior (ver figura 5-24(e)). Este comportamiento es el contrario que el que ocurre con el atributo *LPR* (ver figura 5-24(d)) observándose que para un elemento con valor de *PEG* mayor o igual a 0,47 si su atributo *LPR* es mayor de 0,75 pertenece a la clase *"alto"*.

En caso de que las condiciones anteriores no se cumplan, estamos ante un elemento que no debe clasificarse en la clase *"alto"* y nos situamos en el tercer nodo decisorio del diagrama. El problema ahora es distinguir las clases *"medio"* y *"bajo"*. Para ello, se estima la condición $PEG \geq 0{,}34$ adecuada para poderlas diferenciar. No obstante, como todavía existen algunos elementos de la clase *"bajo"* que cumplen esta condición, se hace necesario distinguirlos. Es por ello que se añade al diagrama un nuevo nodo en caso de que el atributo *PEG* sea mayor o igual que 0,34. En este caso, al observar los elementos de la clase *"bajo*" que cumplen la condición $PEG \geq 0{,}34$ se construye el nuevo nodo decisorio atendiendo a los atributos *LPR*, *STR* y *SCG* para lograr distinguirlos de la clase *"medio"*: aquellos elementos con valores de $LPR \leq 0{,}25$, $STR < 0{,}2$ y $SCG < 0{,}5$ se ubican en la clase *"bajo"*. En caso de que no se cumplan estas condiciones, se clasifican en la clase *"medio"*.

Continuando con el diagrama, nos encontramos ya con los elementos cuyo atributo *PEG* es menor que 0,34. En este caso, tampoco puede afirmarse que necesariamente pertenezcan a la clase *"bajo"*. De forma análoga al caso anterior, se observa la existencia de elementos pertenecientes la clase *"medio"* que cumplen esta condición ($PEG < 0{,}34$). Mediante inspección visual de estos elementos, se observa que pueden llegar a distinguirse de la clase *"bajo"* para unos valores de *PEG* mayores o iguales que 0,25 y de *LPR* mayores o iguales que 0,78. En caso de no cumplirse estas condiciones, el elemento ya no se clasifica en las clases *"alto"* ni *"medio"*.

Queda ahora por distinguir en los siguientes nodos las clases *"bajo"* y *"muy bajo"*. El siguiente paso para distinguir ambas clases vuelve a basarse en el atributo *PEG*. Analizando el valor máximo de este atributo en la clase *"muy bajo"* se obtiene que si $PEG \geq 0{,}25$ el elemento en cuestión pertenece a la clase *"bajo"*. Esto se refleja en el siguiente nodo.

Si $PEG < 0{,}25$ todavía quedan elementos de la clase *"bajo"* que hay que caracterizar y distinguir, ahora, de la clase "muy bajo". Analizando estos elementos en la representación gráfica se elaboran los dos últimos nodos decisorios del diagrama teniendo en cuenta los atributos *LPR* y *SCG*. Los elementos que tienen un valor en el atributo *PEG* menor de 0,25 y, además, un valor de *LPR* mayor o igual que 0,59 se clasifican en la clase *"bajo"*. En caso contrario, se vuelven a inspeccionar los elementos de la clase *"bajo"* que quedan observándose que, a diferencia de los elementos de la clase *"muy bajo"*, presentan normalmente un valor de *SCG* mayor o igual que 0,56. Si estas dos últimas condiciones que corresponden a los dos últimos nodos no se cumplen, el elemento se considera dentro de la clase *"muy bajo"*.

Tabla 5-10: *Porcentaje de elementos clasificados correctamente con el diagrama de la figura 5-25 en cada una de las clases de la base de datos Modelado del usuario para los conjuntos de entrenamiento y validación.*

Nivel de conocimiento (clase)	*Entrenamiento (%)*	*Validación (%)*
"muy bajo"	87,5	73,1
"bajo"	98,8	93,5
"medio"	93,2	79,4
"alto"	98,4	100

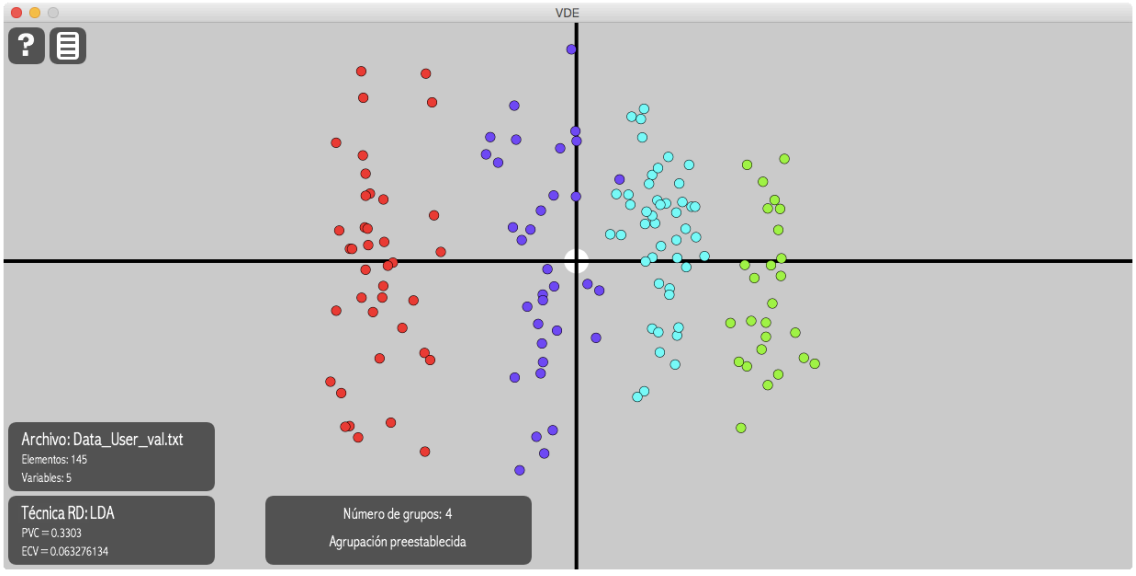


Figura 5-26: *Agrupamiento original del conjunto de datos de validación Modelado del usuario. Los elementos de la clase "muy bajo" aparecen en color verde, los de la clase "bajo" en azul claro, los de la clase "medio" color morado y los de la clase "alto" en rojo.*

Con este esquema, obtenido a partir de los datos correspondientes al conjunto de entrenamiento, se logra clasificar correctamente en este conjunto 247 de los 258 elementos de que consta (un 95,7 %). En cuanto al conjunto de validación, el diagrama logra clasificar correctamente 128 de los 145 elementos (un 88,3 %). En la tabla 5-10 se detalla el porcentaje de elementos correctamente clasificados en cada clase para los conjuntos de entrenamiento y validación. En la figura 5-26 se muestra la representación gráfica del conjunto de validación obtenida por *VDE* mediante la técnica de reducción de la dimensionalidad *LDA*.

Capítulo 6

Conclusiones y propuestas de futuro

En este capítulo se recogen tanto las conclusiones derivadas de la presente tesis como las líneas de investigación futuras que se pretenden desarrollar.

6.1. Conclusiones

La presentación gráfica de datos goza de numerosas ventajas sobre cualquier otra forma de representación por su atractivo, por el interés que provoca y por la facilidad con la que es capaz de transmitir información de una forma rápida y directa. No obstante la visualización de datos es una de las dificultades más habituales e importantes cuando el volumen de datos y/o la dimensionalidad de éstos son altos.

Esta tesis logra el objetivo de diseñar e implementar una aplicación que es capaz de representar gráficamente gran cantidad de datos multidimensionales de tal modo que la extracción de información y conocimiento de forma visual se realiza de forma eficaz incluso para aquellos usuarios no familiarizados con las técnicas de Minería de Datos. La aplicación *VDE* ofrece un entorno de visualización sencillo y previsible a la vez que competente que permite al usuario aplicar su

instinto y experiencia convirtiéndolo en el verdadero protagonista en la labor de extraer la información implícita en grandes colecciones de datos.

Con respecto a otras herramientas de Representación Visual de Datos, *VDE* reúne un conjunto de cualidades que, en su conjunto, la diferencian claramente del resto. Una de esas cualidades es la representación de datos multidimensionales de forma intuitiva gracias a las técnicas de reducción de la dimensionalidad. Ello permite al analista poder interpretar fácilmente los resultados obtenidos por la aplicación sin la necesidad de tener que descifrar complejas representaciones con las que puede no estar familiarizado y que pueden requerirle un adiestramiento previo. Esto le concede la ventaja de poder centrarse totalmente en los aspectos concernientes al conjunto de datos a analizar. Otra cualidad importante de la aplicación es su rasgo genérico, siendo *VDE* un instrumento de visualización de datos de propósito general. Efectivamente, esta aplicación puede emplearse a una gran variedad de conjuntos de datos sin importar el contexto en el cual han sido elaborados ni su estructura intrínseca.

Las bondades de *VDE* se extienden más allá de su carácter versátil e intuitivo. Esta aplicación presenta un medio interactivo donde el usuario puede extraer la información y adaptar la visualización del conjunto de datos según marquen sus designios y necesidades. En este sentido, un amplio espectro de prestaciones se ponen al servicio de quien utilice la aplicación para facilitar el análisis visual de los datos desde diferentes perspectivas y a distintos niveles de concreción. Entre otras, pueden citarse las siguientes:

- Facilita la visualización gráfica de los datos según diferentes técnicas de reducción de la dimensionalidad o de forma combinada entre ellas. Esto permite al usuario escoger la representación más adecuada en cada caso.
- Permite el acceso a la visualización de grandes cantidades de datos utilizando métodos de agrupamiento que facilitan su visualización de forma simplificada sin pérdida de información.
- Ofrece la posibilidad de realizar agrupamientos en el conjunto de datos, bien a través de una clasificación previa de los mismos o a través de la selección de uno de los métodos con que cuenta la aplicación.
- Proporciona el acceso gráfico y numérico a la información contenida en distintas formas y niveles de detalle. En este sentido, *VDE* permite modos de visualización centrados en los propios datos de forma individual o agrupada, en los atributos que los componen o incluso de forma combinada.

- Posibilita una interacción natural del usuario con la visualización del conjunto de datos.

El alcance de las posibilidades de *VDE* se ha puesto de manifiesto en su aplicación a cinco conjuntos de datos reales. En todos los casos, *VDE* ha demostrado su capacidad para la obtención de conocimiento esencial que ayude a la comprensión de cómo ese conjunto de datos se encuentra organizado y revelar qué información es importante para explicar esa organización.

En el estudio de los conjuntos *Semillas* e *Iris*, *VDE* demuestra ser una herramienta de gran eficacia. En ambos casos, a partir de la representaciones visuales de datos que se obtienen, el usuario es capaz de comprender la estructura en la que se disponen. Teniendo en cuenta el análisis por variables, tanto supervisado a partir del agrupamiento original como no supervisado, se puede comprobar con suma sencillez cuáles son las que mejor explican la organización de los datos y cómo caracterizan los distintos grupos. Por otro lado, el análisis comparativo entre los agrupamientos propuestos por la aplicación con respecto a los agrupamientos originales descubre datos clasificados incorrectamente. El análisis visual de estos datos pone de manifiesto las razones por las cuales su clasificación no es acertada. Esta información al servicio del analista le dota de un importante conocimiento acerca de los conjuntos en cuestión.

En referencia a los resultados obtenidos con el conjunto *Wine*, vuelve a ponerse de manifiesto la utilidad de *VDE*. En este caso, es el análisis supervisado que ofrece una representación mediante la técnica de reducción de la dimensionalidad *LDA* que logra diferenciar las clases en las que los datos se hayan clasificados respetando su estructura original. Esto hace que sea posible mediante el análisis de variables descubrir cómo estas se distribuyen en cada clase de forma sencilla y así conseguir caracterizarlas. Una vez más, el análisis visual de los atributos en datos que aparecen próximos en la representación gráfica y que pertenecen a distintas clases aporta conocimiento valioso sobre el conjunto de datos.

Procediendo de modo similar a los anteriores, también se ha comprobado la capacidad de *VDE* para la transmisión de información visual en los conjuntos *E.coli* y *Modelado del usuario*. En ambos casos se ha logrado a partir del análisis visual de los datos establecer un esquema de diferenciación de las diferentes clases a partir del análisis visual. En el caso de *E.coli* se ha propuesto un esquema a modo de diagrama en árbol (ver figura 5-22) que permite comprender de forma cualitativa el comportamiento del conjunto de datos. En referencia a la base de datos *Modelado del usuario* el análisis visual junto con la observación de algunos valores

concretos permite establecer un diagrama de flujo (ver figura 5-25) que ayuda a caracterizar las diferentes clases en las que el conjunto está organizado.

Aludiendo a estos dos últimos conjuntos, cabe admitir que otras técnicas de Minería de Datos pueden obtener mejores resultados que los logrados mediante el análisis en *VDE* pero el conocimiento que aportan al analista no es tan significativo como el obtenido por esta aplicación. Sin lugar a dudas, el conocimiento que aporta el análisis de resultados mediante *VDE* goza de total significado para el analista dado que es éste quien, a partir del estudio visual de los resultados facilitados por la aplicación, su experiencia e intuición, construye ese conocimiento con total comprensión de su porqué.

En relación a la utilización de *VDE* en todos estos conjuntos de datos es importante mencionar que el analista experimentado hubiese aprovechado esta herramienta de visualización de un modo distinto y más apropiado al aquí mostrado. El propósito al aplicar *VDE* en todos estos casos es el de ilustrar las excelentes posibilidades que ofrece este instrumento de uso abierto al servicio un usuario no experto en análisis y visualización de datos.

Como resumen de este apartado de conclusiones, se puede afirmar que *VDE* es una aplicación capaz de representar de forma coherente gran cantidad de datos multidimensionales en dos dimensiones. Es intuitiva, potente, sumamente sencilla de utilizar e interpretar, versátil y con una variedad de servicios eficaces de apoyo con el fin de que el usuario logre un conocimiento significativo a partir de cualquier conjunto de datos.

6.2. Propuestas de futuro

VDE es una aplicación en constante evolución cuyo principal objetivo de desarrollo es incrementar los servicios que proporciona para el análisis de datos sin menoscabo de su sencillez en el manejo y su facilidad para presentar la información de forma simple y comprensible.

En este sentido, varios son los aspectos a trabajar en el avance de esta herramienta. Los principales son:

- Ampliar el número de técnicas de reducción de la dimensionalidad con el fin de poder adaptar la representación gráfica de forma adecuada a la mayor cantidad posible de conjuntos de datos.

- Aumentar la cantidad de medidas de proximidad entre datos de tal forma que se mejore su versatilidad. En este sentido, se hace necesario contar con medidas válidas para atributos categóricos.
- Incrementar la propuesta de métodos de agrupamiento a utilizar con la intención de obtener clasificaciones de datos adecuadas que ayuden a la comprensión del conjunto sin necesidad de supervisión.

No obstante todo lo anterior, la línea de evolución de *VDE* ha de mantener un firme compromiso con la facilidad de uso y comprensión que ofrece al usuario ya que es una de sus piedras angulares. De este modo, siempre ha de ser importante valorar cómo la dotación de nuevas posibilidades para el análisis de datos o la cantidad de información mostrada en cada caso (gráfica o numérica) puede influir en este aspecto.

Por todo ello, como ya se comentó en el capítulo 4, la aplicación está disponible para su descarga desde la página web *www.uv.es/misai/VDE.htm*. Desde esta misma página se anima a otros analistas que la utilicen a efectuar sugerencias y propuestas de mejora para su desarrollo.

Bibliografía

Abdesselam, R., & Zighed, D. (2012). Statistical comparisons for the topological equivalence of proximity measures. *Proceedings, 2nd Stochastic Modeling Techniques and Data Analysis International Conference, SMTDA 2012*.

Alpaydin, E. (2010). *Introduction to Machine Learning - 2nd Edition.* Cambridge, Massachusetts, USA: The MIT Press.

Baarsch, J., & Celebi, M. (2012). Investigation of Internal Validity Measures for K-Means Clustering. *Proceedings of the International MultiConference of Engineers and Computer Scientists, IMECS*, *1*, págs. 14-16.

Baillo, A., & Grané, A. (2008). *100 problemas resueltos de Estadística Multivariante (implementados en MATLAB).* Las Rozas, Madrid, España: Delta, Publicaciones Universitarias.

Bansal, K., & Sood, S. (2011). Data Visualization A Tool of Data Mining. *International Journal of Computer Science and Technology , 2* (3), 197–198.

Bernstein, M., de Silva, V., Langford, J. C., & Tenenbaum, J. B. (2000). *Graph approximations to geodesics on embedded manifolds.* Technical Report, Stanford University, Department of Psychology.

Bohnacker, H., Gross, B., Laub, J., & Lazzeroni, C. (2012). *Generative Design: Visualize, Program, and Create with Processing.* New York, NY, USA: Princeton Architectural Press.

Borg, I., & Groenen, P. (2005). *Modern Multidimensional Scaling: Theory and Applications.* New York, NY, USA: Springer Science & Business Media.

Boriah, S., Chandola, V., & Kumar, V. (2008). Similarity Measures for Categorical Data: A Comparative Evaluation. *Proceedings of the eighth SIAM International Conference on Data Mining*, págs. 243–254.

Brockmeier, A., Kriminger, E., Sanchez, J., & Principe, J. (2011). Latent State Visualization of Neural Firing Rates. *Neural Engineering (NER), 2011 5th International IEEE/EMBS Conference*, págs. 144-147.

Chen , J., & Liu, Y. (2011). Locally Linear Embedding: a survey. *Artificial Intelligence Review , 36* (1), 29-48.

Chen, C.-h., Härdle, W., & Unwin, A. (2008). *Handbook of Data Visualization.* Santa Clara, CA, USA: Springer-Verlag.

Cheng, L., Yang, J., Zheng, D., Li, B., & Ren, J. (2015). The Health Monitoring Method of Concrete Dams Based on Ambient Vibration Testing and Kernel Principle Analysis. *Shock and Vibration, 2015*.

Chien, J.-T., & Ting , C.-W. (2004). Speaker Identification Using Probabilistic PCA Model Selection. *Proceedings of the 8th International Conference on Spoken Language Processing, INTERSPEECH 2004 – ICSLP*, págs. 1785-1788.

Coenen, F. (2011). Data mining: past, present and future. *The Knowledge Engineering Review, 26* (1), 25-29.

Cook, J., Sutskever, I., Mnih, A., & Hinton, G. (2007). Visualizing Similarity Data with a Mixture of Maps. *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics, 1*, págs. 67-74. Citeseer.

Cormen, T., Leiserson, C., Rivest, R., & Stein, C. (2009). *Introduction to Algorithms. Third Edition.* Cambridge, Massachusetts, USA: The MIT Press.

Cudeck, R., & MacCallum, R. (2007). *Factor Analysis at 100: Historical Developments and Future Directions.* Mahwah, New Jersey, USA: Lawrence Erlbaum Associates.

Damavandinejadmonfared, S., & Varadharajan, V. (2015). A New Extension of Kernel Principal Component Analysis for Finger Vein Authentication. *Proceedings of the 38th Australasian Computer Science Conference (ACSC 2015)*, págs. 59-63.

de Oliveira Martins, L., Junior, G., Silva, A., de Paiva, A., & Gattass, M. (2009). Detection of Masses in Digital Mammograms using K-means and Support Vector Machine. *ELCVIA: Electronic Letters on Computer Vision and Image Analysis , 8* (2), 39-50.

de Silva, V., & Tenenbaum, J. B. (2002). Global versus local methods in nonlinear dimensionality reduction. *Advances in Neural Information Processing Systems*, 705-712.

de Silva, V., & Tenenbaum, J. B. (2003). Unsupervised learning of curved manifolds. En M. Hansen, C. C. Holmes, B. K. Mallick, & B. Yu (Edits.), *Nonlinear Estimation and Classification*, 453-465. New York, NY, USA: Springer.

de Stefani, E., Ronco, A. L., Boffetta, P., Deneo-Pellegrini, H., Correa, P., Acosta, G., y otros. (2012). Nutrient-derived Dietary Patterns and Risk of Colorectal Cancer: a Factor Analysis in Uruguay. *Asian Pac J Cancer Prev , 13* (1), 231-235.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-End Factor Analysis For Speaker Verification. *Audio, Speech, and Language Processing , 19* (4), 788-798.

Dubey, S., Dixit, P., Singh, N., & Gupta, J. (2013). Infected Fruit Part Detection using K-Means Clustering Segmentation Technique. *International Journal of Artificial Intelligence and Interactive Multimedia , 2* (2), 65-72.

Dupont, S., & Ravet, T. (2013). Improved Audio Classification Using a Novel Non-Linear Dimensionality Reduction Ensemble Approach. *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR '13)*, págs. 287-292.

Edelsbrunner, H. (2014). Topological Spaces. En H. Edelsbrunner, *A Short Course in Computational Geometry and Topology* (págs. 57-63). Berlin, Germany: Springer International Publishing.

Everitt, B., & Hothorn, T. (2011). *An Introduction to Applied Multivariate Analysis with R .* New York, NY, USA: Springer .

Everitt, B., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis.* London, UK: John Wiley & Sons, Ltd.

Fürnkranz, J., Gamberger, D., & Lavrač, N. (2012). *Foundations of Rule Learning.* Springer-Verlag Berlin Heidelberg.

Fry, B. (2008). *Visualizing Data: Exploring and Explaining Data with the Processing Environment.* Sebastopol, CA, USA: O'Reilly Media, Inc.

Gantz, J., & Reinsel, D. (2012). *The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the Far East.* IDC Tech Report.

Guan, Y., & Dy, J. (2009). Sparse Probabilistic Principal Component Analysis. *Proceedings of the International Conference on Artificial Intelligence and Statistics*, págs. 185-192.

Gupta, I., Kumar, A., Singh, C., & Kumar, R. (2015). Detection and Mapping of Water Quality Variation in the Godavari River Using Water Quality Index, Clustering and GIS Techniques. *Journal of Geographic Information System, 7* (2), 71-84.

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research, 3*, 1157-1182.

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques, Second Edition.* San Francisco, CA, USA: Morgan Kaufmann Publishers.

Hannachi, A., & Turner, A. (2013). Isomap nonlinear dimensionality reduction and bimodality of Asian monsoon convection. *Geophysical Research Letters, 40* (8), 1653-1658.

Harris, R. (1999). *Information Graphics: A Comprehensive Illustrated Reference.* New York, NY, USA: Oxford University Press.

Heer, J., Bostock, M., & Ogievetsky , V. (2010). A Tour Through the Visualization Zoo. *Communications ACM , 53* (6), 59-67.

Hinton, G., & Roweis, S. (2002). Stochastic Neighbor Embedding. *Advances in Neural Information Processing Systems*, *15*, 833-840.

Hougardy, S. (2010). The Floyd-Warshall Algorithm on Graphs with Negative Cycles. *Information Processing Letters, 110* (8), 279-281.

Huang, M., Hao, L., Guo, X., Hu, C., Gu, X., Zhao, W., y otros. (2013). Characterization of secondary organic aerosol particles using aerosol laser time-of-flight mass spectrometer coupled with FCM clustering algorithm. *Atmospheric Environment, 64*, 85-94.

Huber, R., Ramoser, H., Mayer, K., Penz, H., & Rubik, M. (2005). Classification of coins using an eigenspace approach. *Pattern Recognition Letters, 26* (1), 61-75.

Hussain, R. (2012). Synthetic Aperture Radar(SAR) images features clustering using Fuzzy c-means(FCM) clustering algorithm. *Computational Ecology and Software, 2* (4), 220-225.

Ilango, M., & Mohan, V. (2010). A Survey of Grid Based Clustering Algorithms. *International Journal of Engineering Science and Technology, 2* (8), 3441-3446.

Jolliffe, I. (2002). *Principal Component Analysis.* New York, NY, USA: Springer.

Jumb, V., Sohani, M., & Shrivas, A. (2014). Color Image Segmentation Using K-Means Clustering and Otsu's Adaptive Thresholding. *International Journal of Innovative Technology and Exploring Engineering (IJITEE), 3* (9), 72-76.

Jyotsna, E., Akhil, P., & Arun, K. (2013). Silhouette based human action recognition using PCA and ISOMAP . *International Journal of Advanced Research in Computer and Communication Engineering, 2* (11), 4192-4198.

Köppen, M. (2000). The curse of dimensionality. *5th Online World Conference on Soft Computing in Industrial Applications (WSC5).*

Kahraman, H., Sagiroglu, S., & Colak, I. (2013). Developing intuitive knowledge classifier and modeling of users' domain dependent data in web . *Knowledge Based Systems, 37*, 283-295.

Keim, D. A. (2002). Information Visualization and Visual Data Mining. *IEEE Tansactios on Visualization and Computer Graphics, 8* (1), 1-8.

Kirk, A. (2012). *Data Visualization: A Successful Design Process .* Birmingham, UK: Packt Publishing Ltd .

Kotsiantis, S. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica 31*, 249–268.

Koupaie, H., Ibrahim, S., & Hosseinkhani, J. (2013). Outlier Detection in Stream Data by Clustering Method. *International Journal of Advanced Computer Science and Information Technology, 2* (3), 25-34.

Kurgan, L., & Musilek, P. (2006). A survey of Knowledge Discovery and Data Mining process models. *The Knowledge Engineering Review, 21* (1), 1-24.

Lee, J. (2010). *Introduction to Topological Manifolds.* New York, NY, USA: Springer Science & Business Media.

Lee, J., & Verleysen, M. (2007). *Nonlinear dimensionality reduction.* New York, NY, USA: Springer Science & Business Media.

Lesot , M., Rifqi, M., & Benhadda, H. (2009). Similarity measures for binary and numerical data: a survey. *International Journal of Knowledge Engineering and Soft Data Paradigms, 1* (1), 63-84.

Li, J., Hao, P., & Zhang, C. (2008). Transferring Colours to Grayscale Images by Locally Linear Embedding. *Proceedings of the British Machine Vision Conference*, págs. 835-844.

Li, W., Yeung, D., & Zhang, Z. (2009). Probabilistic Relational PCA. *Advances in Neural Information Processing Systems*, 1123-1131.

Li, Z., Yan, X., Yuan, C., Peng, Z., & Li, L. (2011). Virtual prototype and experimental research on gear multi-fault diagnosis using wavelet-autoregressive model and principal component analysis method. *Mechanical Systems and Signal Processing, 25* (7), 2589–2607.

Lichman, M. (2013). *UCI Machine Learning Repository*. (University of California, School of Information and Computer Sciences. Irvine, CA) Obtenido el 31 de enero de 2015, de http://archive.ics.uci.edu/ml

Loring, W. (2010). *An Introduction to Manifolds.* New York, NY, USA: Springer Science+Business Media .

Luo, S., Kim, E., Dighe, M., & Kim, Y. (2011). Thyroid nodule classification using ultrasound elastography via linear discriminant analysis. *Ultrasonics, 51* (4), 425-431.

Müller, H., & Hamm, U. (2014). Stability of market segmentation with cluster analysis – A methodological approach . *Food Quality and Preference, 34*, 70-78.

Machado, J., Duarte, G., & Duarte, F. (2011). Identifying economic periods and crisis with the multidimensional scaling. *Nonlinear Dynamics, 63* (4), 611-622.

Mann, A., & Kaur, N. (2013). Survey Paper on Clustering Techniques. *International Journal of Science, Engineering and Technology Research (IJSETR), 2* (4), 803-806.

Masaeli, M., Fung, G., & Dy, J. (2010). From Transformation-Based Dimensionality Reduction to Feature Selection. *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, págs. 751-758.

Mawale, M., & Gandole, Y. (2011). Analysis Of Optimal Route Algorithms Under Constraint Conditions. *International Journal of Computer Science and Information Technologies, 2* (6), 2614-2619.

McLachlan, G., & Krishnan, T. (2007). *The EM algorithm and extensions*. Hoboken, New Jersey, USA: John Wiley & Sons.

Méndez, C., & Rondón, M. (2012). Introducción al análisis factorial exploratorio. *Revista Colombiana de Psiquiatría, 41* (1), 197-207.

Morán Álvarez , A. (2012). Análisis y predicción de perfiles de consumo energético en edificios públicos mediante técnicas de Minería de Datos. *Tesis Doctoral*. Universidad de Oviedo.

Morariu, V., & Camps, O. (2006). Modeling Correspondences for Multi-Camera Tracking Using Nonlinear Manifold Learning and Target Dynamics. *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference, 1*, págs. 545-552. IEEE.

Noble, J. (2009). *Programming Interactivity: A Designer's Guide to Processing, Arduino, and Openframeworks.* Sebastopol, CA, USA: O'Reilly Media, Inc.

Oded, M., & Rokach, L. (2010). *The Data Mining and Knowledge Discovery Handbook.* New York, NY, USA: Springer.

Oke, O., Adedeji, T., Alade, O., & Adewus, E. (2012). Fuzzy kc-means Clustering Algorithm for Medical Image Segmentation. *Journal of Information Engineering and Applications, 2* (6), 21-32.

Parimala, M., Lopez, D., & Senthilkumar, N. (2011). A Survey on Density Based Clustering Algorithms for Mining Large Spatial Databases. *International Journal of Advanced Science and Technology, 31* (1), 59-66.

Park, J., Pande, P., Shrestha, S., Clubb, F., Applegate, B., & Jo, J. (2012). Biochemical characterization of atherosclerotic plaques by endogenous multispectral fluorescence lifetime imaging microscopy. *Atherosclerosis , 220* (2), 394-401.

Patras, A., Brunton, N., Downey, G., Rawson, A., Warriner, K., & Gernigon, G. (2011). Application of principal component and hierarchical cluster analysis to classify fruits and vegetables commonly consumed in Ireland based on in vitro antioxidant activity. *Journal of Food Composition and Analysis, 24* (2), 250-256.

Peng, Z., Cao, C., Liu, Q., & Pan, W. (2013). Human Walking Pattern Recognition Based on KPCA and SVM with Ground Reflex Pressure Signal. *Mathematical Problems in Engineering, 2013*.

Pérez-Cruz, F. (2008). Kullback-Leibler Divergence Estimation of Continuous Distributions. *IEEE International Symposium on Information Theory*, págs. 1666-1670.

Persaud, A., & Azhar, I. (2012). Innovative mobile marketing via smartphones: Are consumers ready? *Marketing Intelligence & Planning, 30* (4), 418-443.

Rakesh, M., & Ravi, T. (2012). Image Segmentation and Detection of Tumor Objects in MR Brain Images Using FUZZY C-MEANS (FCM) Algorithm. *International Journal of Engineering Research and Application, 2* (3), 2088-2094.

Rambaut, A., Pybus, O., Nelson, M., Viboud, C., Taubenberger, J., & Holmes, E. (2008). The genomic and epidemiological dynamics of human influenza A virus. *Nature, 453* (7195), 615-619.

Ridgway, G., & Ashburner, J. (2012). Visualising the distribution of subjects using t-distributed stochastic neighbour embedding (t-SNE). *Organization of Human Brain Mapping Conference.*

Rokach, L., & Maimon, O. (2005). Clustering methods. En *Data Mining and Knowledge Discovery Handbook* (págs. 321-352). Ney York, NY, USA: Springer.

Roweis, S. (1997). EM Algorithms for PCA and SPCA. *Advances in Neural Information Processing Systems*, 626-632.

Rujirakul, K., So-In, C., & Arnonkijpanich, B. (2014). PEM-PCA: A Parallel Expectation-Maximization PCA Face Recognition Architecture. *The Scientific World Journal,* 1-16.

Saad, Y. (2011). *Numerical Methods for Large Eigenvalue Problems: Revised Edition*. Philadelphia, PA, USA: Siam.

Saunders, C., Aldering, G., Bailey, S., Birchall, D., Childress, M., Fakhouri, H., y otros. (2014). Principal Component Analysis of Type Ia Supernova Spectrophotometric Time Series. *American Astronomical Society Meeting Abstracts, 223.*

Schmitt, T. (2011). Current Methodological Considerations in Exploratory and Confirmatory Factor Analysis. *Journal of Psychoeducational Assessment, 29* (4), 304-321.

Shawe-Taylor, J., & Cristianini , N. (2004). *Kernel Methods for Pattern Analysis.* New York, NY, USA: Cambridge University Press.

Shmueli, G., Patel, N., & Bruce, P. (2005). *Data Mining In Excel: Lecture Notes and Cases.* Arlington, Virginia, USA: Resampling Stats, Inc.

Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., y otros. (2013). Real-Time Human Pose Recognition in Parts from Single Depth Images. *Communications of the ACM, 56* (1), 116-124.

Shu, X., Gao, Y., & Lu, H. (2012). Efficient linear discriminant analysis with locality preserving for face recognition. *Pattern Recognition, 45* (5), 1892-1898.

Simoff, S., Böhlen , M., & Mazeika, A. (2008). *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics.* Berlin, Germany: Springer-Verlag.

Smiciklas, M. (2012). *The Power of Infographics: Using pictures to communicate and connect with your audiences.* Indianapolis, Indiana, USA: QUE Publising.

Stahl, F., Gabrys, B., Medhat Gaber, M., & Berendsen, M. (2013). An Overview on Interactive Visual Data Mining Techniques for Knowledge Discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3* (4), 239–256.

Stefaner, M., & GmbH, R. (s.f.). *OECD Better Life Index*. Obtenido el 29 de enero de 2015, de http://oecdbetterlifeindex.org

Steinwart, I., & Christmann, A. (2008). *Support Vector Machines.* New York, NY, USA: Springer Science & Business Media.

Steinwart, I., & Scovel, C. (2012). Mercer's Theorem on General Domains: On the Interaction between Measures, Kernels, and RKHSs. *Constructive Approximation, 35* (3), 363-417.

Sudha, K., Raju, Y., & Sekhar, A. (2012). Fuzzy C-Means clustering for robust decentralized load frequency control of interconnected power system with Generation Rate Constraint. *International Journal of Electrical Power & Energy Systems, 37* (1), 58-66.

Sudhamathy, G., & Venkateswaran, C. (2011). Web Log Clustering Approaches: A Survey. *International Journal on Computer Science and Engineering, 3* (7), 2896-2903.

Sumathi, S., & Sivanandam, S. (2006). *Introduction to Data Mining and its Applications.* New York, NY, USA: Springer.

Usama Fayyad, G. P.-S. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine, 17* (3), 37-53.

Valencia, J., Daza, G., Acosta, C. D., & Castellanos, G. (2010). Comparación de Métodos de Reducción de Dimensión Basados en Análisis por Localidades. *Tecno Lógicas, 25*, 131-150.

van der Maaten, L., Postma, E., & van den Herik, J. (2009). *Dimensionality Reduction: A Comparative Review.* Tilburg University Technical Report. TiCC-TR 2009-005.

van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research, 9*, 2579-2605.

Vijayan, A., & Balasundaram, S. (2013). Effective Web-Service Discovery Using K-Means Clustering. *Distributed Computing and Internet Technology, proceedings of the 9th International Conference*, págs. 455-464.

Vitelleschi, M., & Quaglino, M. (2010). Modelos PCA a partir de conjuntos de datos con información faltante.¿ Se afectan sus propiedades? *SaberEs, 2*, 105-109.

Wang, C., Chen, J., Sun, Y., & Shen, X. (2009). Wireless Sensor Networks Localization with Isomap. *Communications, 2009. ICC'09. IEEE International Conference*, págs. 1-5. IEEE.

Welling, M. (2005). *Fisher Linear Discriminant Analysis.* University of Toronto, Department of Computer Science.

White, S., Bray, B., & Ollendick, T. (2012). Examining Shared and Unique Aspects of Social Anxiety Disorder and Autism Spectrum Disorder Using Factor Analysis. *Journal of Autism and Developmental Disorders, 42* (5), 874-884.

Widjaja, D., Varon, C., Dorado, A., Suykens, J., & Van Huffel, S. (2012). Application of Kernel Principal Component Analysis for Single-Lead-ECG-Derived Respiration. *Biomedical Engineering, IEEE Transactions on, 59* (4), 1169-1176.

Williams , B., Brown, T., & Onsman, A. (2012). Exploratory factor analysis: A five-step guide for novices. *Australasian Journal of Paramedicine, 8* (3).

Wu, C., Sheng, W., & Zhang, Y. (2006). Mobile Self-Localization using Multi-Dimensional Scaling in Robotic Sensor Networks. *The International Journal of Intelligent Control and Systems, 11* (3), 163-175.

Xiang, S., Nie, F., & Zhang, C. (2008). Learning a Mahalanobis distance metric for data clustering and classification. *Pattern Recognition, 41* (12), 3600-3612.

Xie, Z., & Mu, Z. (2008). Ear Recognition Using LLE and IDLLE Algorithm. *Pattern Recognition, 2008. ICPR 2008. 19th International Conference*, págs. 1-4.

Xu, R., & Wunsch, D. (2009). *Clustering.* Hoboken, NJ, USA: John Wiley & Sons.

Yu, L., & Liu, H. (2004). Efficient Feature Selection via Analysis of Relevance and Redundancy. *Journal of Machine Learning Research, 5*, 1205–1224 .

Zaixian, X. (2011). Exploratory Visualization of Data Pattern Changes in Multivariate Data Streams. *Tesis Doctoral*. Worcester Polytechnic Institute.

Zhang, Q., Jimenez, J. L., Canagaratna, M. R., Ulbrich, I., Ng, N., Worsnop, D., y otros. (2011). Understanding atmospheric organic aerosols via factor analysis of aerosol mass spectrometry: a review. *Analytical and Bioanalytical Chemistry, 401* (10), 3045-3067.