Aprendizaje de similitudes entre pares de objetos mediante clasificación supervisada

VNIVERSITAT DE VALÈNCIA

TESIS DOCTORAL

Autora: Emilia López Iñesta

Memoria que presenta para optar al título de Doctora

Dirigida por:

Dr. Miguel Arevalillo Herráez
Dr. Francisco Grimaldo Moreno

Programa de Doctorado en Tecnologías de la Información, Comunicaciones y Computación

Departament d'Informàtica

Escola Tècnica Superior d'Enginyeria

Universitat de València

Mayo 2017

Documento maquetado con TEXIS v.1.0+.

Agradecimientos

A mi familia. Aquí y Allá.

Quisiera comenzar esta sección haciendo mención a mis directores de Tesis, los Doctores Miguel Arevalillo Herráez y Francisco Grimaldo Moreno. Desde estas líneas quiero expresaros mi gratitud por toda vuestra ayuda (que ha sido *mucha*) a lo largo de estos años, por vuestra generosidad, disponibilidad y todas vuestras enseñanzas en el plano laboral y personal.

Un reconocimiento a José Manuel Claver, coordinador del Programa de Doctorado, así como a todos los integrantes de la comisión de Doctorado del Departamento de Informática y Pilar Gómez Arce por todas las gestiones.

A los profesores del Departamento de Informática y a los miembros del grupo de innovación educativa, gracias por ser siempre tan atentos, interesarse por el progreso de esta Tesis y enseñarme tantas cosas. Gracias Ariadna, Paco, Miguel L., Fernando, Max, Miguel G., Juan G., David Arnau, etc.

No puedo dejar de mencionar y agradecer la atención y ayuda de todos los investigadores y profesores del *Centro de Investigación en Inteligencia Artificial* (CIIA) de la Universidad Veracruzana (México): Alejandro, Nicandro, Erandi, Héctor Gabriel, Rubén y sus respectivas familias (Adriana, Sofi, Christina, María, etc.) que año a año han conocido la evolución de este trabajo, me han aportado ideas y me han hecho sentir siempre como en casa.

Por otro lado, un agradecimiento a los investigadores del *Labbs* del Consiglio Nazionale delle Ricerche (CNR) de Roma: Mario, Daniele, Giulia A., Federica, Giulia B., etc. Un recuerdo muy especial para Rosaria Conte.

Als membres de *l'Associació Catalana d'Intel·ligència Artificial* (ACIA), gràcies per les vostres aportacions i comentaris que sense dubte han fet millorar aquest treball. Gràcies Àngel, Eva, Dionís, Zoe i tants altres.

El desarrollo de esta Tesis hubiera sido mucho más complicado sin el apoyo del Decanato de la Facultad de Ciencias del Deporte y Actividad Física de la Universidad Católica de Valencia: muchas gracias por todos los cambios de horarios así como permisos para poder irme de estancia, entre otras cosas. Gracias Amparo, Luis, Chelo, Javier y Pilar.

Mi gratitud a todos mis compañeros profesores y especialmente a Mario

Zacarés: gracias por tu apoyo y generosidad estos años. No olvido al resto: Laura R., María F., Julio, Cristina, Inma, Carmen, Laura J., Víctor, Josep, Nacho, Héctor, Isaac, Verónica y un largo etcétera. Una mención a "nuestros" informáticos, Andrés, Daniel y Javier, por instalar *R* una y otra vez.

Al Grup de Recerca *Alcàntera*: gràcies per obrir-me aquest món de la Sociolingüística i permetre'm fer experiments. Començarem fa poc, però estic segura que farem coses molt interessants.

Al equipo de la "Falla Immaterial", ejemplo de entrega y trabajo. Nadie pensó que una falla así se pudiera proyectar en el Palau de la Generalitat... ¡y lo logramos! Gracias Fran, Emilio, Paula, Dani, Juanjo, Eusebio, Carlos, Alejandro y Santi.

A mis amigas de Alicante: Lorena, Ana, las Marías y especialmente a Rosa. Aunque repartidas por el mundo, desde Guinea hasta Washington, buscamos siempre un hueco para encontrarnos en "la millor terreta del món".

A mis amigos de la Facultad de Matemáticas: Lola, Paloma, Ángela, Mª José, Mario, Pablo, Juan, Alfred, Mayte, Mª Carmen, María R., Ruth, David... Gracias por ser mi mayor club de fans y seguirme en mis múltiples proyectos, ya sea un DatabeersVlc, una Falla Immaterial o Milmots.

A mis compañeros y amigos del Colegio Mayor Rector Peset: Belén B., Ana, Inma, Belén A., Vicent, Pep, Lucía, Adriana y resto de la tropa. Gracias por esos maravillosos años y por esos fines de semana de reencuentro con vuestras familias en Ayna, las excursiones a la Fresneda y a Castellnovo.

A todos los amigos de *Alababarada*, por llenar de música y modernez este mundo, que sin duda sería mucho menos interesante y divertido. Gracias Yolanda, Chema, Cristina, David, Mª José, Emili, Rosa, Raquel, Salva, Rafa, Armando, Isabel, Elvira, Pili... ¡Sois muchos!

Gracias a Julia, Esther, Vicente y François, por acordaros siempre de mí.

A mis amigos y allegados de Benimàmet: Alonso, Belén, Óscar (traductor oficial), José Manuel, Manolo, PD, María, Isa, Andrea, Mónica, Bea, ect.

A mi familia en Burjassot: Paco, Cande, Amparo, Carlos, Toni, las *hermanas* y compañía. Gracias por cuidarme tanto y tratarme con tanto cariño.

Y por supuesto a mi familia, tíos y primos, en Alicante, Gandía y Logroño, sin su constante ánimo, llegar al final no hubiera sido posible. Un reconocimiento especial a mis padres, Lázaro y Emilia, y hermanos, Andrés y Vicky, por entender todas mis ausencias, apoyarme y ayudarme siempre.

A tu estimat meu, et deixe per al Final, però saps que ets el primeR. Gràcies per lA teua pacièNcia i esTima al llarg d′aquEsts anys que no han sigut fàcils. Sembla que per fi, podrem començar una nova etapa, anar al Palmar i descobrir el món sense parlar *tant* de problemes NP-H, fer un "viatge als somnis polars", riure′ns de si "Eres PC o Eres Mac" i recordar que "Nada debería fallar".

València, maig de 2017

Resumen

El uso de medidas de similitud, distancias o métricas se encuentra en la base del funcionamiento de numerosas técnicas estándar de clasificación, resultando además, una tarea fundamental e importante en las áreas de estudio del Aprendizaje Automático (*Machine Learning*) y el Reconocimiento de Patrones (*Pattern Recognition*). Dado que el cálculo de la similitud entre dos objetos puede ser muy diferente en función del contexto, la construcción inteligente de estas medidas a partir de los datos disponibles, puede ayudar en la obtención de clasificadores más robustos y mejorar los resultados en la tarea específica que se propone resolver.

En los últimos años, el aprendizaje de métricas (*Metric Learning*) y medidas de similitud (*Similarity Learning*) ha recibido un creciente interés de la comunidad científica. Dada la información disponible en forma de ejemplos etiquetados con una categoría o clase, el objetivo del aprendizaje de métricas es aprender una distancia métrica de acuerdo al siguiente principio: las distancias entre pares similares (es decir, de la misma clase) han de ser pequeñas, mientras que las distancias entre pares diferentes (es decir, de diferentes clases) han de ser mayores. De la misma manera, el aprendizaje de similitud intenta aprender una función de similitud que asocie grandes puntuaciones (*scores*) a pares similares y pequeñas puntuaciones a pares diferentes. Un caso particular del aprendizaje de similitudes consiste en el empleo de métodos de clasificación para el aprendizaje de medidas de similitud (*Classification-based Similarity Learning*). En todos estos métodos, el rendimiento depende en gran medida de la representación de las características de los datos disponibles.

Así, en esta Tesis se presenta un método de clasificación enriquecido que sigue un enfoque híbrido que combina la extracción de características (*Feature Extraction*) y la ampliación de las mismas (*Feature Expansion*). En particular, se propone una transformación de datos y el uso de un conjunto de distancias métricas y no métricas para complementar y enriquecer la información proporcionada por los vectores de características de los ejemplos de entrenamiento. Si bien es cierto que esto aumenta la dimensión del problema en cuestión, también supone una inyección de conocimiento adicional debido a que el uso de las medidas de distancias supone un emparejamien-

to implícito entre los vectores de características de dos objetos. Además, se analiza si la nueva información añadida compensa el aumento de dimensión que ello implica, así como la influencia de los diferentes formatos de datos de entrada y el tamaño de entrenamiento sobre el rendimiento del clasificador.

La propuesta se compara con métodos de aprendizaje de métricas y los resultados obtenidos muestran rendimientos comparables en favor del método propuesto en distintos contextos y empleando diferentes bases de datos.

Resum

L'ús de mesures de similitud, distàncies o mètriques es troba en la base del funcionament de nombroses tècniques estàndard de classificació, resultant a més, una tasca fonamental i important en les àrees d'estudi de l'Aprenentatge Automàtic (*Machine Learning*) i el Reconeixement de Patrons (*Pattern Recognition*). Atés que el càlcul de la similitud entre dos objectes pot ser molt diferent en funció del context, la construcció intel·ligent d'aquestes mesures a partir de les dades disponibles, pot ajudar en l'obtenció de classificadors més robusts i millorar els resultats en la tasca específica que es proposa resoldre.

En els últims anys l'aprenentatge de mètriques (*Metric Learning*) i mesures de similitud (*Similarity Learning*) ha rebut un creixent interés de la comunitat científica. Donada la informació disponible en forma d'exemples etiquetats amb una categoria o classe, l'objectiu d'aprenentatge de mètriques és aprendre una distància mètrica d'acord al següent principi: les distàncies entre parells similars (és a dir, de la mateixa classe) han de ser xicotetes, mentre que les distàncies entre parells diferents (és a dir, de diferents classes) han de ser majors. De la mateixa manera, l'aprenentatge de similitud intenta aprendre una funció de similitud que associe grans puntuacions (*scores*) a parells similars i xicotetes puntuacions a parells diferents. Un cas particular de l'aprenentatge de similituds consisteix en l'ús de mètodes de classificació per a l'aprenentatge de mesures de similitud (*Classification-based Similarity Learning*). En tots aquests mètodes, el rendiment depén en gran manera de la representació de les característiques de les dades disponibles.

Així, en aquesta Tesi es presenta un mètode de classificació enriquit que segueix un enfocament híbrid que combina l'extracció de les característiques (*Feature Extraction*) i l'ampliació de les mateixes (*Feature Expansion*). En particular, es proposa una transformació de dades i l'ús d'un conjunt de distàncies mètriques i no mètriques per a complementar la informació proporcionada pels vectors de característiques dels exemples d'entrenament. Si bé és cert que açò augmenta la dimensió del problema en qüestió, també suposa una injecció de coneixement addicional a causa que l'ús de les mesures de distàncies suposa un emparellament implícit entre els vectors de característiques de dos objectes. A més, s'analitza si la nova informació afegida compensa l'augment de dimensió que açò implica, així com la influència dels

diferents formats de dades d'entrada i la grandària d'entrenament sobre el rendiment del classificador.

La proposta es compara amb mètodes d'aprenentatge de mètriques i els resultats obtinguts mostren rendiments comparables en favor del mètode proposat en diferents contextos i emprant diferents bases de dades.

Abstract

The use of measures of similarity, distances or metrics is a core central issue for many standard classification techniques, becoming a fundamental and important task in the areas of study of Machine Learning and Pattern Recognition. Since computing the similarity between two objects may be very different depending on the context, the intelligent construction of these measures from the available data can help in obtaining more robust classifiers and improve the results in the specific task that It is proposed to resolve.

In recent years, Metric Learning and Similarity Learning techniques have received a growing interest from the scientific community. Given the available information in the form of labeled examples with a category or class, the main goal of Metric Learning is to learn a metric distance according to the following principle: the distances between similar pairs (i.e., pairs of objects with the same class) must be small, while the distances between different pairs (i.e., different classes) must be greater. Likewise, Similarity Learning attempts to learn a similarity function that associates large scores with similar pairs and small scores to different pairs. A particular case of Similarity Learning is the use of classification methods for learning similarity measures known as Classification-based Similarity Learning. In all these methods, the performance depends to a great extent on the features representation of the available data.

Thus, this Thesis presents an enriched classification method that follows a hybrid approach combining Feature Extraction and Feature Expansion techniques. In particular, we propose a data transformation and the use of a set of metric and non-metric distances to complement the information provided by the feature vectors of the training examples. While this increases the dimensionality of the problem in question, it also implies an additional injection of knowledge because the use of distance measures implies an implicit match between the characteristics of two objects. In addition, we analyze whether the new information added compensates for the dimensionality increasement involved, as well as the influence of different data input formats and training size on classifier performance.

The proposal is compared with metric learning methods and the results obtained show comparable yields in favor of the proposed method in different

contexts and using different databases.

Abreviaturas

Abreviatura	Significado y página primera aparición
AP	Average Precision 81
CBIR	Content Based Image Retrieval 13
CNN	Convolutional Neural Networks 13
ComProd	Product Combinations 28
CombSum	Sum Combinations 28
CSL	Classification-based Similarity Learning 11
DIST-LR	pool of Distances using Logistic Regression 109
DIST-SVM	pool of Distances using a linear SVM 109
ECSL	Enriched Classification Similarity Learning 103
ECSL-LR	ECSL using Logistic Regression 106
ECSL-SVM	ECSL using a linear SVM 106
EER	Equal Error Rate 106
EXPAN	Expansion 101
EXTRAC	Extraction 101
ICA	Independent Component Analysis 32
IoE	Internet of Everything 4
IoT	Internet of Things 3
IR	Information Retrieval 4
ITML	Information Theoretic Metric Learning 74
JCR	Journal Citation Report 37
KISSME	Keep It Simple and Straightforward MEtric 98
k-NN	k-Nearest Neighbor 11
LDA	Linear Discriminant Analysis 32
LDML	Logistic Discriminant Metric Learning 98
LE	Laplacian Eigenmaps 32

Abreviatura	Significado y página primera aparición
LFW	Labeled Faces in the Wild dataset 104
LFW-Attr	LFW dataset Attributes version 104
LFW-SIFT	LFW dataset SIFT version 104
LMNN	Large Margin Nearest Neighbour 98
LR	Logistic Regression 103
MAHAL	Mahalanobis 106
MAP	Mean Average Precision 62
MFCC	Mel-Frequency Cepstral Coefficient 40
ML	Metric Learning 10
OWA	Ordered Weighted Averaging 27
PCA	Principal Component Analysis 14
POLY	Polynomial 112
PSD	Positive Semidefinite 73
RBF	Radial basis function 63
RCA	Relevant Component Analysis 74
RGB	Red Green Blue 23
SBE	Sequential backward elimination 30
SFS	Sequential Forward Selection 30
SL	Similarity Learning 10
SVM	Support Vector Machine 13
SVM-RFE	SVM-Recursive Feature Elimination 30

Tabla de símbolos

Notación general

Notación	Descripción
X	Colección de objetos
x_i	Vector de característica de un objeto genérico
$\mathbf{x}_i$	Vector de coordenadas objeto x_i
S	Conjunto pares objetos similares entrenamiento
D	Conjunto pares objetos disimilares entrenamiento
R	Conjunto ternas objetos entrenamiento
M	Matriz semidefinida positiva
$\mathbb{S}_+^d$	Cono matrices simétricas Semidefinidas Positivas
d	Dimensión espacio entrada
d_M	Distancia de Mahalanobis
δ	Distancia genérica
x_i'	Vector x_i transpuesto
M'	Matriz M transpuesta
S_M	Función bilinear de similitud
$N(x_i, x_j)$	Término de normalización
$\|\|$	Concatenación
k	Índice pares de objetos $k = 1, ..., K$
p_k	Pares de objetos
l_k	Etiqueta similar/disimilar
$\mathbf{p_k}$	Vectores de características de pares de objetos concatenados
$\mathbf{p_k'}$	Nuevo conjunto de vectores de características después de realizar transformación
T	Descriptores
p	Parámetro distancia de Minkowski
$\mathbf{x}_i^{(t)}$	Conjunto de características correspondientes al descriptor t en $\mathbf{x}_i$
$\delta_h^{(t)}$	Distancia entre los descriptores de dos objetos $\mathbf{x}_i^{(t)}$ y $\mathbf{x}_j^{(t)}$

Notación	Descripción
k_e	Kernel
Σ	Matriz de covarianzas
μ	Media
σ	Desviación típica
r	Rango de M
W	Matriz de transformación lineal

Índice

Índice de figuras

Índice de Tablas

Parte I

Presentación de la Tesis Doctoral

Capítulo 1

Introducción

Resumen:

Esta Tesis Doctoral investiga el problema del aprendizaje de similitudes entre pares de objetos. En este primer capítulo introductorio, se presentan la motivación y justificación del trabajo realizado y los objetivos de la Tesis en las secciones 1.1 y 1.2, respectivamente. Asimismo, se detalla la estructura de esta memoria en el apartado 1.3.

1.1. Motivación y justificación

Los avances tecnológicos han permitido en las últimas décadas la transmisión y recopilación de datos de una manera cada vez más sencilla y eficaz. De un lado se tiene que las cámaras de tráfico o videovigilancia, vehículos, sensores medioambientales y en general, todo tipo de dispositivos relacionados con el "Internet de las Cosas" (*IoT, Internet of the Things*)[1] contribuyen cada día a la generación de grandes bases de datos que conviene analizar de forma inteligente y automática, con el objetivo de transformar esos datos en información útil en la que basar la toma de decisiones.

Por otro lado, se debe tener en cuenta que se ha producido un cambio de paradigma en el que el comportamiento y hábitos de las personas ha hecho que pasemos de ser únicamente consumidores de datos a ser a la vez productores de datos a través del uso que se realiza de teléfonos móviles, relojes inteligentes, redes sociales o en términos más globales, de las relaciones que se establecen con el llamado "Internet del Todo" (*IoE, Internet of*

[1] El término "Internet de las Cosas" (IoT) fue acuñado por K. Ashton (2009) y se refiere a la red de sensores y objetos físicos interconectados digitalmente y que disponen de tecnología integrada (Bluetooth, Wi-Fi, etc.) para recolectar información e interactuar con el mundo físico además de utilizar los estándares existentes de Internet para proporcionar servicios de transferencia de la información disponible, análisis, aplicaciones y comunicaciones (Chui et al., 2010; Gubbi et al., 2013)

Figura 1.1: Internet del Todo

Everything)[2] como se puede observar en la figura 1.1[3] .

En consecuencia, en la actualidad nos enfrentamos a una gran explosión de información donde el problema del almacenamiento de datos de cualquier tipo (imágenes, vídeos, textos, etc.) ha quedado prácticamente superado y se necesitan mecanismos adecuados para la organización y análisis de los datos, así como para la fusión y/o enriquecimiento de estos datos con información auxiliar proveniente bien de fuentes externas o bien de transformaciones que se pueden realizar a partir de los propios datos de los que se dispone. Ante este escenario surge la denominada "Ciencia de datos" o *Data Science* que incluye áreas como las matemáticas, la estadística y la computación avanzada, así como técnicas de reconocimiento de patrones y visualización de datos entre otras. El gran reto de la Ciencia de datos reside en la capacidad de gestionar y analizar los datos disponibles para comprender su estructura, identificar relaciones interesantes y extraer conocimiento.

En el proceso de esta transformación de datos en conocimiento resulta imprescindible una búsqueda y recuperación efectiva de la información (*IR, Information Retrieval*) que se puede definir como la disciplina que se encar-

[2] La empresa Cisco define el "Internet del Todo" (IoE) como la interacción de cuatro pilares fundamentales: personas, datos, objetos y procesos que consiguen que sus conexiones en red sean más relevantes y valiosas que nunca, convirtiendo la información en acciones que crean nuevas capacidades, experiencias más ricas y oportunidades económicas sin precedentes para las empresas, individuos y países (Evans, 2012).

[3] Imagen extraída del repositorio `http://www.thinkstockphotos.ca/`

ga de la búsqueda y acceso eficiente a recursos digitales que se encuentran en grandes colecciones. Estos recursos pueden ser de muy distinta índole: páginas web, libros, vídeos, fotografías, etc.

Existen múltiples aplicaciones que hacen uso de sistemas relacionados con la recuperación de la información. Entre éstas se pueden señalar como ejemplos el estudio de imágenes para el diagnóstico de enfermedades en medicina, la identificación de huellas dactilares en biometría o los sistemas de reconocimiento automático de caras para autenticación o investigación criminal. Otras aplicaciones muy utilizadas en nuestro día a día son los motores de búsqueda en Internet como Google[4] o Mozilla Firefox[5], las búsquedas realizadas por voz a través de los asistentes personales inteligentes como Siri[6] o Cortana[7], los sistemas de identificación de música (p.e., Shazam[8] o Soundhound[9]) o plataformas de series de televisión, vídeos, películas, música o comercio electrónico que incluyen sistemas de recomendación y sugerencias como Netflix[10], YouTube[11], Itunes[12], Spotify[13], eBay[14] o Amazon[15].

En las aplicaciones citadas anteriormente, el contenido de un objeto cualquiera (canción, imagen, fichero de audio, etc.) está representado habitualmente mediante un vector numérico de características (*feature vector*). Al realizar una búsqueda de información concreta, se compara la descripción de un objeto a través de su vector de características con los vectores del resto de objetos en la base de datos en la que se realiza la consulta, obteniéndose un conjunto de resultados ordenados en base a la relevancia con la búsqueda realizada. Por consiguiente, las comparaciones son una tarea esencial tanto en la búsqueda y recuperación de información como en muchos de los métodos de Reconocimiento de Patrones y Aprendizaje Automático.

La forma más sencilla de comparar pares de objetos es mediante el uso de funciones de similitud o disimilitud (distancias, métricas u otra función). Un ejemplo muy utilizado es la distancia Euclídea, aunque presenta una serie de desventajas ya que depende de la tarea específica a realizar y del contexto, es decir, de las particularidades del problema en cuestión (McFee y Lanckriet, 2010). Además, trata a todos los atributos que describen los objetos de igual manera, no teniendo en cuenta aspectos como la importancia relativa de cada atributo y la correlación que pueda existir entre ellos, así como la naturaleza

[4] https://www.google.com
[5] https://www.mozilla.org
[6] http://www.apple.com/es/ios/siri/
[7] https://www.microsoft.com/es-es/windows/cortana
[8] https://www.shazam.com
[9] https://soundhound.com/
[10] https://www.netflix.com
[11] https://www.youtube.com/
[12] https://www.apple.com/es/itunes/music/
[13] https://www.spotify.com/es/
[14] http://www.ebay.es/
[15] https://www.amazon.es

de los datos en consideración y la estructura interna de la información que se está analizando.

Para paliar estos inconvenientes, surgen enfoques como el aprendizaje de distancias métricas que tratan de aprender una medida de similitud entre objetos a partir de los vectores de características que los describen de manera que se cumplan las propiedades de distancia. Sin embargo, existen aplicaciones (p.e., sistemas de recomendación o de recuperación de imágenes) en las que no son necesarias estas propiedades y por tanto, se puede flexibilizar su cumplimiento. En particular en esta Tesis, se presenta una propuesta con la finalidad de obtener un ranking de pares de objetos en base a su similitud empleando métodos de clasificación supervisada. Estos métodos producen una puntuación en lugar de una métrica que es especialmente útil cuando el objetivo es clasificar u ordenar pares de elementos de acuerdo a su similitud.

A continuación, se enuncian los objetivos planteados en este trabajo.

1.2. Objetivos

Esta Tesis Doctoral tiene como **objetivo principal** presentar un método de aprendizaje de puntuaciones (*scores*) que permita establecer un ranking de pares de objetos en base a su similitud empleando técnicas de clasificación supervisada. Para obtener estos *scores*, se proporciona a los clasificadores distintas configuraciones de datos de entrada con la finalidad de analizar las ventajas e inconvenientes de cada una de ellas.

Como **objetivos específicos** del trabajo se señalan los siguientes:

- Estudiar el rendimiento de métodos de clasificación en el contexto del aprendizaje de puntuaciones de similitud mediante técnicas de clasificación supervisada.
- Analizar la influencia de la representación del formato de los datos de entrada a los clasificadores para obtener puntuaciones de similitud.
- Identificar formatos de representación de entrada de los datos mediante técnicas de fusión, extracción y/o expansión de características que mejoren el resultado de las técnicas de clasificación.
- Aplicar métodos de evaluación y comparación del rendimiento de técnicas de aprendizaje de distancias y clasificadores para obtención de un ranking de similitud.

Una vez identificados los objetivos, se presenta en la siguiente sección la estructura de esta memoria de Tesis Doctoral.

1.3. Estructura de la Memoria de Tesis Doctoral

Esta Tesis se presenta como compendio de publicaciones, según la normativa de la Escuela de Doctorado de la Universitat de València y su Programa de Doctorado en Tecnologías de la Información, Comunicaciones y Computación del Departamento de Informática.

Atendiendo a los requisitos establecidos, el presente documento se ha estructurado en tres partes principales.

La primera parte contiene los dos primeros capítulos de la memoria de la Tesis Doctoral. En el capítulo 1 se expone la motivación y justificación de este trabajo y los objetivos, así como la estructura del presente documento. Seguidamente, en el capítulo 2 se realiza un resumen global de la temática a estudiar que incluye el planteamiento del problema a investigar y el recordatorio de una serie de conceptos preliminares necesarios para abordar la lectura del resto de la Tesis Doctoral, además de aspectos relacionados con el estado del arte y la metodología empleada en este trabajo.

En la segunda parte de la Tesis, el capítulo 3 presenta un resumen de las contribuciones realizadas a través de las publicaciones presentadas en las distintas revistas impactadas en el *Journal Citation Report (JCR)*. A continuación, en el capítulo 4 se exponen las conclusiones más importantes que se derivan de esta Tesis Doctoral y las futuras líneas de trabajo.

Por último, como indica la normativa, en los anexos A, B y C se puede consultar la versión íntegra de los artículos publicados.

Capítulo 2

Contextualización del trabajo

Resumen:

En este capítulo se expone el planteamiento del problema a estudiar en el apartado 2.1 junto con una serie de definiciones y conceptos preliminares, así como aspectos metodológicos necesarios para el posterior desarrollo de la memoria de tesis.

En particular, se recuerda la importancia del cómputo del grado de similitud entre objetos en la sección 2.2 definiendo los conceptos de distancia, métrica y similaridad[1] desde un punto de vista matemático y exponiendo la relación existente entre ellos. A continuación, se revisan las funciones de distancias métricas y no métricas más utilizadas en $\mathbb{R}^n$ y que se emplearán en la experimentación de los trabajos incluidos en la Tesis.

El problema del aprendizaje de funciones de distancia está estrechamente relacionado con el problema de la representación de datos para describir un objeto. Así, para ilustrar esta relación en la sección 2.3 se introduce la importancia de una buena representación de los objetos en el espacio de características con el fin de poder calcular la similitud entre ellos. En este sentido, es necesario recordar técnicas relacionadas con la combinación de características, así como algunos procesos asociados a la extracción de características para convertir la información "en bruto" o sin procesar que caracteriza a los objetos en información útil y significativa.

Para finalizar, teniendo en cuenta que la descripción de los objetos a menudo tiene una representación en un espacio de alta dimensión, en la sección 2.4 se dan más detalles de la conexión existente entre el problema de representación de características, el aprendizaje de métricas y la reducción de la dimensión.

[1] Los términos *similaridad* y *disimilaridad* son traducciones de las palabras *similarity* y *dissimilarity* en inglés y no aparecen en el Diccionario de la lengua española de la Real Academia Española. Sin embargo, dado que son palabras muy empleadas en el ámbito de estudio y trabajo en el que se enmarca esta Tesis y con el objetivo de evitar el uso continuo de las palabras similitud y disimilitud, se relajará esta norma.

2.1. Planteamiento del problema

Dado que la precisión de muchos algoritmos de Aprendizaje Automático depende críticamente de una distancia definida sobre el espacio de entrada, la posibilidad de automatizar el proceso de la medición de la similitud entre objetos o el diseño de sistemas que adquieran dicha capacidad, ha recibido últimamente un creciente interés de la comunidad científica.

Las revisiones bibliográficas de Kulis (2013) y Bellet et al. (2015) recopilan muchas de las numerosas publicaciones de las últimas décadas que han presentando soluciones alternativas y más sofisticadas al uso de medidas convencionales (distancia Euclídea, Manhattan, Coseno, etc.). Éstas consisten en el aprendizaje de una función de (di)similitud a partir de un conjunto más reducido de la colección disponible de objetos X, llamado conjunto de entrenamiento (*training set*). Así, el Aprendizaje de Funciones de Distancia o Métricas (*Metric Learning*, ML) y el Aprendizaje de medidas de Similitud (*Similarity Learning*, SL) son dos enfoques comunes para hacer frente a esta cuestión dentro del campo del Aprendizaje Automático. Además, si los objetos están conceptualmente definidos a través de una etiqueta que indica a qué categoría o clase pertenece, nos encontramos en el caso particular de Aprendizaje Automático Supervisado.

De manera genérica, se considerará que dada una colección de objetos $X = \{x_1, x_2, ..., x_n\}$, representados a través de sus vectores de características que describen convenientemente todos sus atributos relevantes en un espacio vectorial d-dimensional, $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\} \subseteq \mathbb{R}^d$, se puede interpretar cada objeto como un punto en el espacio de sus características (*feature space*).

A partir de ejemplos etiquetados, se pueden definir restricciones que relacionen los ejemplos disponibles basándose en la asunción de que la distancia entre ejemplos etiquetados con la misma clase debe ser menor que la distancia entre ejemplos de clases distintas. Estas restricciones pueden ser construidas por pares o pueden ser definidas de manera relativa (por ejemplo, empleando ternas) de la siguiente manera:

$$S = \{(x_i, x_j) : x_i \; y \; x_j \; deben \; ser \; similares\}$$
$$D = \{(x_i, x_j) : x_i \; y \; x_j \; deben \; ser \; disimilares\}$$
$$R = \{(x_i, x_j, x_k) : x_i \; debe \; ser \; más \; similar \; a \; x_j \; que \; a \; x_k\}$$

En el caso del **aprendizaje supervisado de distancias (métricas)**, el objetivo es definir una distancia parametrizada por una matriz M, $d_M = d_M(x_i, x_j) = (\mathbf{x}_i - \mathbf{x}_j)' M (\mathbf{x}_i - \mathbf{x}_j) \; \forall x_i, x_j \in X$, donde M es una matriz semidefinida positiva aprendida de los datos de entrenamiento al minimizar (o maximizar) algunos criterios relacionados con el rendimiento (*performance*) de la función d_M utilizando las restricciones definidas anteriormente.

La distancia aprendida actúa como una transformación del espacio original en el que están representados los ejemplos para satisfacer las restricciones impuestas, de manera que aquellos ejemplos que son semánticamente similares estarán ahora más cerca en el nuevo espacio, mientras que aquellos que son distintos se alejarán, tal y como muestra la figura 2.1[2].

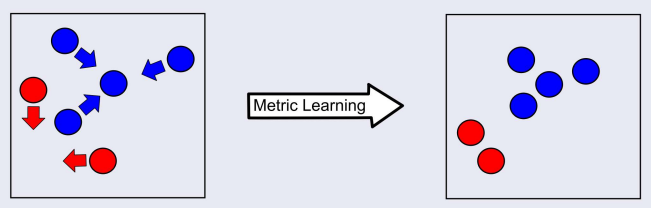


Figura 2.1: Cambio de representación en el espacio original de características inducido por la distancia aprendida

La ventaja principal de los métodos de aprendizaje de métricas es que la distancia d_M es una pseudométrica, por lo tanto d_M se puede integrar en los enfoques de clasificación existentes que asumen espacios pseudómetricos. Un ejemplo se tiene en uno de los primeros trabajos presentados en este área por Xing et al. (2002) en el que la métrica aprendida se utilizó para mejorar el rendimiento del algoritmo del k-vecino más cercano (*k-nearest neighbor, k-NN*) . Con el mismo objetivo surgen los estudios de Domeniconi et al. (2005) y Davis et al. (2007a). Otros trabajos interesantes aplicados a contextos como el de reconocimiento facial o identificación de personas son los de Köstinger et al. (2012) o Liao et al. (2015).

Asimismo, el **Aprendizaje de medidas de Similitud** intenta aprender una función de similitud de tal manera que se obtengan valores altos de puntuaciones o *scores* para pares similares y valores pequeños de puntuaciones o *scores* para pares disimilares. La mayoría de los trabajos de aprendizaje de similitud se centra en aprender una función de similitud definida por $S_M = S_M(x_i, x_j) = \mathbf{x}_i' M \mathbf{x}_j$ donde M, como en el caso anterior, es una matriz que se ha de aprender. Su aplicación ha sido probada en distintos escenarios como la recuperación de imágenes (Chechik et al., 2010), los sistemas de recomendación de música (McFee et al., 2012) o la construcción de *rankings* con los que realizar sugerencias de productos para los clientes de un negocio (Burhanuddin et al., 2015).

Un caso particular del aprendizaje de medidas de similitud, consiste en el uso de métodos de Clasificación para la obtención de los *scores* que estén relacionados con la similitud entre pares de objetos (Chen et al., 2009; Brunner et al., 2012). En este planteamiento, conocido como ***Classification-based Similarity Learning* (CSL)** o *Scoring-based (Dis)Similarity Learning*, el

[2]Imagen tomada de Bellet et al. (2015)

conjunto de restricciones anteriormente definido en el contexto del aprendizaje supervisado de distancias, se utiliza para entrenar a un clasificador.

En su forma más simple, cada restricción por pares se puede representar mediante la concatenación de los vectores de características de los dos objetos. Durante el entrenamiento, esta representación de los objetos se da junto con una etiqueta de clase con dos posibles valores que representan si los pares de objetos son similares o no, por tanto se plantea un problema de clasificación binaria.

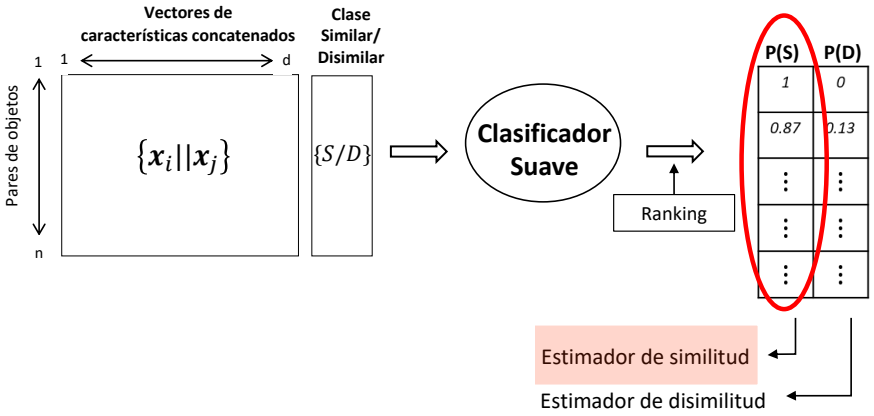


Figura 2.2: Método de Clasificación (CSL) para la obtención de *scores* relacionados con la similitud entre pares de objetos

Una vez que el clasificador ha sido entrenado, se utiliza para obtener resultados que están relacionados con la similitud entre los pares de objetos, es decir, el clasificador en lugar de predecir las dos clases (similar, disimilar), da una estimación de la probabilidad de pertenencia a cada clase. A estos clasificadores se les denomina *soft classifiers* o clasificadores suaves (Liu et al., 2011) y proporcionan un valor entre 0 y 1, al que también se le denomina *score* y determina si está más cerca de la clase similar o disimilar, como muestra la figura 2.2.

Aunque estos valores no satisfacen las propiedades de una métrica, no es necesariamente una desventaja ya que el uso de distancias no métricas, puede resultar muy útil en la comparación de objetos complejos, como indican Tversky (1977) y Pekalska et al. (2006). De hecho, se ha demostrado su buen rendimiento en determinados contextos como la clasificación de datos empleando el algoritmo *k-NN* superando a diferentes técnicas de aprendizaje de distancias métricas del estado del arte como muestran en su trabajo Woznica y Kalousis (2010). Por otro lado, cuando el propósito es, por ejemplo, establecer un *ranking* u ordenación de los objetos en función de su similitud, las propiedades de las métricas no son necesarias y por tanto, los métodos

de clasificación CSL son un enfoque alternativo competitivo al aprendizaje supervisado de distancias tal y como se demuestra en esta Tesis Doctoral.

Los diferentes métodos de CSL generalmente difieren en el mecanismo de clasificación utilizado y/o en la representación de los objetos. De hecho, el formato de los datos de entrada es generalmente un factor importante que afecta al rendimiento de la clasificación, ya que esto determina hasta qué punto las clases son separables. Por lo tanto, a través de una selección de la representación adecuada del formato de entrada de los datos al clasificador se puede obtener una puntuación de similitud más precisa.

Dado que la representación de los datos desempeña un papel clave en el rendimiento de los métodos de aprendizaje automático, la transformación de los datos originales en una entrada más significativa e informativa se ha convertido en un tema de investigación muy activo (Arevalillo-Herráez et al., 2008a; Bengio et al., 2013) comúnmente conocido como *Extracción de Características* (Guyon et al., 2006), y que se refiere a los métodos que se ocupan de cualquier transformación o combinación de los datos en bruto (*raw data*) en vectores de características.

Tradicionalmente, en CSL cada objeto del conjunto de entrenamiento se ha representado mediante un vector de características que corresponde a varios atributos numéricos definidos en un espacio multidimensional. El contenido de los objetos puede ser definido mediante un único vector de características que define completamente al objeto. Sin embargo, en numerosas ocasiones los objetos no tienen una única representación o el empleo de una única caracterización conduce a la obtención de resultados poco eficientes (Faria et al., 2014), lo que conlleva el empleo de varios descriptores, es decir, un conjunto de subvectores agrupados que caracterizan al objeto.

Ésta es una situación frecuente en aplicaciones como el reconocimiento de rostros, la clasificación de audio o la Recuperación de Imágenes Basadas en Contenido (*Content Based Image Retrieval, CBIR*), donde los objetos están representados usualmente por varios descriptores que se relacionan con características diferentes, pero complementarias. Esta fusión de información puede ser tan sencilla como la unión o concatenación (*binding*) de distintos descriptores o más compleja en el caso del empleo de técnicas propias de Inteligencia Artificial como Máquinas de Soporte Vectorial (*SVM, Support Vector Machines*) o Algoritmos evolutivos (Faria et al., 2014).

Como ejemplo, se puede considerar una imagen, donde es habitual caracterizar su contenido a través del uso del histograma de color o de descriptores relacionados con su textura o forma. Estos descriptores se denominan también descriptores de bajo nivel (*low-level descriptors*) y se extraen de manera manual (*hand-crafted features*) según un determinado algoritmo predefinido basado en el conocimiento de un experto tal y como señalan Bengio y Courville (2013). Como alternativa, recientemente, se están empleando redes neuronales convolucionales (*CNN, Convolutional Neural Networks*) para tratar

de obtener representaciones más significativas de los objetos conocidas como características aprendidas o *learned features* (Antipov et al., 2015).

Aunque las CNN han mostrado resultados notables para el aprendizaje de características en tareas como el reconocimiento de objetos o la clasificación de imágenes, estudios como el de Wan et al. (2014) apuntan que las características aprendidas pueden ser o no mejores que las características tradicionales extraídas manualmente, pero con esquemas de refinación adecuados se pueden obtener representaciones que superan el rendimiento de las características manuales convencionales. Sin embargo, otros trabajos como el de Wu et al. (2016) o Park et al. (2016), indican que en determinados contextos como el de identificación de personas funciona mejor la combinación o fusión de las características aprendidas con otras extraídas manualmente.

Además del formato de entrada de los datos, un aspecto importante a tener en cuenta es que el rendimiento de muchos algoritmos de clasificación depende en gran medida del tamaño y la dimensión de los datos de entrenamiento. El uso de distintos tipos de descriptores así como su combinación es una cuestión muy relacionada con el problema de representación de los datos. A pesar de que la descripción del objeto es mucho más completa cuando se utilizan varios descriptores, esto puede conducir a la obtención de un espacio de dimensión muy alta. De ahí que se plantee cuál debería ser la dimensión ideal de los datos de entrada al clasificador para obtener un buen desempeño, considerando que a mayor dimensión generalmente se obtiene una mejor clasificación a costa de aumentar la carga computacional y el riesgo de sobreentrenamiento (Liu y Zheng, 2006).

La alta dimensión también afecta a los métodos de aprendizaje de métricas. Suponiendo una dimensión d de los datos, estos métodos calculan una matriz M de transformación semidefinida positiva $d \times d$ que rota y escala el espacio de características original. Esta matriz M tiene en cuenta tanto la correlación entre las características como la importancia relativa de cada característica y se utiliza para proyectar linealmente los datos en un nuevo espacio donde las restricciones de similitud impuestas por los datos de entrenamiento se satisfacen mejor.

En general, los métodos de aprendizaje de métricas funcionan particularmente bien y muestran un excelente rendimiento con datos de baja dimensión, pero no se adaptan bien a problemas con grandes dimensiones. De hecho, el aprendizaje en entornos de alta dimensión es propenso a un ajuste excesivo (*overfitting*) y además, requiere una gran cantidad de datos de entrenamiento (Liu et al., 2015).

Una solución habitual para tratar los problemas de alta dimensión consiste en la aplicación de técnicas clásicas de reducción de la dimensión como el análisis de componentes principales (PCA). Otra alternativa frecuente en aprendizaje automático es la sustitución de los vectores de características de los objetos por una representación basada en distancias o disimilarida-

des (Lee et al., 2010). Esta idea fue introducida inicialmente por Pekalska y Duin (2005) y consiste en representar cada objeto mediante un vector de (dis)similaridades a un conjunto de prototipos (i.e., un conjunto de objetos representativos o puntos en el espacio multidimensional en el que se trabaja). De esta manera, el espacio de representación original es reemplazado por nuevas características que son las distancias a algunos objetos concretos y fijos.

El uso de las distancias para representar objetos permite plantear una serie de cuestiones que se han explorado en esta Tesis y que constituyen las principales aportaciones de la misma:

- Se propone la definición de una función de transformación basada en la aplicación de múltiples distancias sobre las características originales de pares de objetos libremente seleccionados, en lugar de usar un conjunto de representación fijo, para generar un nuevo formato de datos de entrada a un clasificador con el objetivo de mejorar el cómputo de una puntuación de similitud entre pares de objetos y plantear un método alternativo al aprendizaje supervisado de distancias como muestran los trabajos López-Iñesta et al. (2015a) y López-Iñesta et al. (2017b). El contenido de este último trabajo se puede consultar en anexo B.

- Por un lado se abordan contextos donde los objetos están representados por varios descriptores que se relacionan con diferentes características. En este enfoque, el cálculo de las distancias entre las características de pares de objetos agrupadas por descriptor se utiliza como una capa de preprocesado de los datos originales para reducir la dimensión de los datos de entrada al clasificador lo que supone una ventaja frente a los métodos clásicos que emplean la concatenación de los vectores de pares de objetos. Estos resultados se pueden comprobar en los estudios López-Iñesta et al. (2014a), López-Iñesta et al. (2014b) y López-Iñesta et al. (2015b), este último en el anexo A.

- Por otro lado, para que el método sea más general, se consideran contextos de uso en los que no sea necesaria la existencia de descriptores en la caracterización de los objetos. En este caso, se realiza un estudio en el que se presenta un enfoque híbrido que combina tanto la extracción de características como la expansión de características. En particular, se propone un método de extracción de características a través de una transformación de los datos originales que contribuye a la separación de las clases. A continuación, el resultado de esta transformación, se fusiona con nuevas características construidas a partir de un conjunto de distancias estándar que se utilizan para ampliar y complementar la información proporcionada por la nueva representación de los vectores de características dando al clasificador una entrada enriquecida y más significativa. Este nuevo formato convierte al método CSL en un

procedimiento que obtiene mejores resultados que otras técnicas de extracción de características y de aprendizaje de métricas del estado del arte tal y como se demuestra en el artículo López-Iñesta et al. (2017a) en el anexo C.

Se presenta por tanto, una doble aplicación de las distancias sobre el formato de representación de los objetos. Asimismo, se establece un debate interesante entre en qué situaciones resulta más conveniente reducir la dimensión de los datos de entrada empleando distancias como un método de preprocesado o por el contrario, fusionar y expandir una representación vectorial de los objetos mediante su empleo. Esta segunda opción, supone un enriquecimiento a los vectores de características al emplear la información auxiliar que aportan las distancias. A pesar de que esto implica un aumento de la dimensión proporcional al número de distancias calculadas, en general ésta es mucho menor en comparación a la dimensión de los datos originales y además se puede aprovechar la sinergia de la combinación de las características (o una transformación de las mismas) con otras nuevas características construidas a partir de distancias tal y como indica la literatura revisada (Tsai et al., 2011; Guo et al., 2014).

Además, otros aspectos interesantes a señalar y cubiertos en esta Tesis son:

- El aumento del rendimiento del método está basado en la consideración de varias funciones de distancia (métricas y no métricas) simultáneamente en el formato de representación de los objetos. Esto se debe a que la naturaleza distinta de cada función de distancia puede contribuir al aprendizaje mediante la captura de una relación diferente entre las características de cada par de objetos.

- El formato de entrada de los pares de datos tiene efectos sobre el rendimiento de un clasificador en función de si éste es lineal o no lineal. De esta manera surgen preguntas como: ¿sería posible que un clasificador no lineal que utiliza una representación "en bruto" (sin la adición de características basadas en la distancia) pudiera producir mejores resultados que el clasificador con las características concatenadas?. En otras palabras, ¿es la adición de nuevas funciones sólo necesaria cuando se emplean clasificadores "más sencillos"? ¿Podrían los clasificadores con modelos no lineales superar fácilmente a los clasificadores lineales en esta situación?

- El método empleado funciona independientemente de si los objetos tienen una representación basada en características que emplea uno o varios descriptores o si las características han sido extraídas manualmente o aprendidas a través de otras técnicas.

- La propuesta no se limita en modo alguno al uso de un clasificador particular. Por el contrario, el enfoque presentado está abierto a la utilización de métodos de clasificación alternativos. A título ilustrativo, en esta Tesis se han empleado la Regresión Logística y la SVM.

La elección de la función de distancia más adecuada depende de la tarea a realizar (Papa et al., 2009), así diferentes autores han analizado el desempeño de varias medidas de distancia para tareas específicas (Aggarwal et al., 2001; Howarth y Rüger, 2005).

En este trabajo no se trata de elegir la distancia más apropiada para una tarea concreta, sino que se busca un método que tenga una aplicación general en múltiples contextos, casos de estudio o bases de datos. Por ello se propone el uso de una representación de los objetos más allá de una representación vectorial en la que se incluye una combinación de varias funciones de distancia simultáneamente con el objetivo de aumentar el rendimiento de un clasificador que se empleará para proporcionar una puntuación mediante la que establecer un ranking de similitud de pares de objetos.

Para conseguir tal efecto, parece lógico que la representación empleada deba reflejar las relaciones de similitud o disimilitud entre los objetos de una manera significativa. Es por ello que se emplean distintas distancias métricas y no métricas con el propósito de proporcionar información auxiliar útil ya que la combinación de diferentes medidas de disimilitud puede enfatizar distintos tipos de relación y de información sobre los objetos y clases a distinguir, tal y como apuntan autores como Lee et al. (2007) o Ibba et al. (2010).

La consideración de emplear las distancias no métricas reside en que en los últimos años han surgido distintos trabajos que consideran útiles e informativas aquellas medidas que no cumplen algunas propiedades propias de las métricas (Pekalska et al., 2006; Skopal y Bustos, 2011; Scheirer et al., 2014). De hecho, investigadores como Amos Tversky demostraron años antes empíricamente en sus trabajos (Tversky, 1977; Tversky y Gati, 1982) que en determinados contextos, la similitud no cumple necesariamente propiedades como la simetría o la propiedad triangular que son asumidas por las distancias. Su razonamiento se basa en que desde un punto de vista psicológico, el concepto de similitud para los humanos es inherentemente subjetivo y juzgamos la similitud mediante medidas no métricas. Al comparar pares de objetos buscamos la parte concordante entre estos y aquello que los diferencia (Tversky, 1977; Bustos y Skopal, 2006). Además, autores como Jacobs et al. (2000) y Howarth y Rüger (2005), han demostrado que las funciones de distancia que son robustas a valores extremos o *outliers* y aspectos no relevantes en la coincidencia de pares de objetos son distancias no métricas.

En las siguientes secciones de este capítulo, se definen formalmente los conceptos de distancia, métrica y similaridad. Además, se recuerdan las medidas de distancia más empleadas entre vectores de características. Asimismo

también se relacionará el problema de aprendizaje de distancias métricas con el de representación de características.

2.2. Concepto y cómputo de la similitud

La noción de similitud juega un rol importante en las técnicas de Inteligencia Artificial y Aprendizaje automático y ha sido tema de estudio en las últimas décadas debido a sus múltiples aplicaciones. Entre éstas figuran desde encontrar y agrupar objetos que se parezcan basándose en sus características hasta reconocer patrones o etiquetar y clasificar objetos.

A pesar de que detectar que un grupo de objetos es similar o disimilar pueda ser una tarea sencilla e intuitiva de realizar para un humano, enseñarle a un ordenador o a un aprendiz artificial este tipo de tareas resulta complicado y es un tema de investigación que involucra a distintas áreas como Visión Computacional, Aprendizaje automático o Recuperación de la Información, tal y como indican Bustos y Skopal (2006).

Otro aspecto a tener en cuenta, es que en muchas ocasiones capturar el parecido o similitud entre dos objetos depende del contexto considerado, por ello muchas investigaciones plantean que debe ser aprendida a partir de los datos disponibles.

Atendiendo a estas consideraciones, se puede afirmar que la similitud en sí es un concepto complejo. Matemáticamente, la similitud entre objetos puede ser modelada a través de una medida de similaridad o disimilaridad, que hace referencia a un valor numérico para indicar el grado de cercanía o lejanía entre dos objetos o cuán parecidos o distintos son.

Existe un gran número de coeficientes de similaridad, disimilaridad y de distancias en la literatura. En esta sección se resumen las medidas más importantes para el desarrollo de esta Tesis Doctoral, a la vez que se definirán conceptos relacionados como el de kernel o divergencia.

2.2.1. Conceptos de distancia, métrica y similaridad

Dados dos objetos representados vectorialmente, es posible calcular la similitud entre ellos mediante distintas funciones que calculan un valor numérico que da una idea de lo lejos o cerca que están los dos vectores en el espacio en el que se encuentran definidos. La abstracción del concepto de "cercanía o lejanía" se puede formalizar matemáticamente a través de un *espacio métrico* (Kelley, 1955; Zezula et al., 2005; Bustos y Skopal, 2006). De hecho algunos autores como Santini y Jain (1999) parten de la hipótesis de que la similitud (o disimilitud) entre objetos es una distancia en algún espacio de características, que se asume como un espacio métrico.

Se definen a continuación formalmente los conceptos de espacio métrico, distancia, métrica y similaridad.

Definición 2.2.1. Dado un conjunto de objetos X, se define como un espacio métrico al par (X, δ), donde δ es una aplicación $\delta : X \times X \to \mathbb{R}$, llamada distancia, tal que a cada par de objetos representados por dos puntos en el espacio x_i, x_j le hace corresponder un número real $\delta(x_i, x_j)$ cumpliendo las siguientes propiedades generales:

- P1. $\forall x_i, x_j \in X\,, \delta(x_i, x_j) \geqslant 0$ (No negatividad)
- P2. $\forall x_i \in X\,, \delta(x_i, x_i) = 0$ (Reflexiva)
- P3. $\forall x_i, x_j \in X\,, x_i \neq x_j \Rightarrow \delta(x_i, x_j) > 0$ (Positividad)
- P4. $\forall x_i, x_j \in X\,, \delta(x_i, x_j) = \delta(x_j, x_i)$ (Simetría)
- P5. $\delta(x_i, x_k) \leqslant \delta(x_i, x_j) + \delta(x_j, x_k)\, \forall x_i, x_j, x_k \in X$ (Desigualdad triangular)

Observación 2.2.1. Las propiedades P2 y P3 se pueden expresar como una única propiedad: $\delta(x_i, x_j) = 0 \Leftrightarrow x_i = x_j\ \forall x_i, x_j \in X$ que se conoce como la *Identidad de los indiscernibles*.

Definición 2.2.2. Si los elementos de un espacio métrico (X, δ) son tuplas de valores reales, entonces el par (X, δ) se llama espacio vectorial. Un espacio vectorial X n-dimensional es un caso particular de un espacio métrico donde los objetos son representados a través de n coordenadas de números reales $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ (Chávez et al., 2001).

En un espacio métrico, la única operación posible sobre objetos es el cálculo de una función de distancia entre pares de objetos que satisface la desigualdad triangular. Por el contrario, los objetos en un espacio vectorial satisfacen algunas propiedades adicionales. Además del cálculo de la distancia entre vectores, la representación vectorial nos permite realizar operaciones como adición y sustracción de vectores. De este modo, se pueden construir nuevos vectores a partir de vectores anteriores (Zezula et al., 2005).

Todas las funciones δ que satisfagan las propiedades P1 a P5 previas se denominan **métricas**, pero estas condiciones se pueden relajar tal y como indican en sus trabajos Cuadras (1989) y Deza y Deza (2009), dando lugar a otras definiciones, como las de **distancia o disimilaridad**, de las que se resumen algunas en la tabla 2.1. Para mayor detalle, se puede consultar la *Enciclopedia de distancias* de Deza y Deza (2009).

Definición 2.2.3. En general, se conoce como **no métrica** a aquella función que no cumple alguna de las propiedades de distancia anteriores, siendo la desigualdad triangular (P5) la propiedad que suele incumplirse más habitualmente (Skopal y Bustos, 2011).

Otra manera de calcular la similitud o semejanza entre objetos, es a través del concepto de *similaridad*.

Tabla 2.1: Clasificación de las distancias según sus propiedades.

Clasificación	Propiedades
Distancia	P1,P2,P3
Distancia semimétrica	P1,P2,P3,P4
Distancia quasimétrica	P1,P2,P3,P5
Distancia pseudométrica	P1,P2,P4,P5
Distancia Métrica	P1,P2,P3,P4,P5

Definición 2.2.4. Una **similaridad** se define como una función $s : X \times X \to \mathbb{R}$ sobre X si se cumple $\forall x_i, x_j \in X$:

- P1. $s(x_i, x_j) \geqslant 0 \, \forall \, x_i, x_j \in X$ (No negatividad)
- P2. $s(x_i, x_j) = s(x_j, x_i)$ (Simetría)
- P3. $Si \; s(x_i, x_j) \leqslant s(x_i, x_j) \; \forall x_i, x_j \in X$ y $s(x_i, x_j) = 0 \Leftrightarrow x_i = x_j$

Al contrario que las distancias que tienen valor mínimo cero y no están acotadas superiormente, las medidas de similaridad suelen tomar valores en el intervalo $[0, 1]$. Por tanto, una función de similitud está inversamente relacionada con una función de distancia: cuanto mayor es el valor de una similaridad, más parecidos son los objetos en la comparación. En las distancias ocurre al contrario: cuanto menor es el valor, más semejantes son los objetos.

En ocasiones, conviene convertir las similaridades en disimilaridades o vicerversa. Algunas de las tranformaciones más habituales tal y como indican Gower y Legendre (1986), son:

- **De disimilaridades a similaridades** Si δ es una distancia, δ se puede convertir en una similaridad mediante las conversiones proporcionadas por $exp(\frac{-\delta(x_i,x_j)}{\sigma})$ o $\frac{1}{1+\delta(x_i,x_j)}$.
- **De similaridades a disimilaridades** Asumiendo que $s \in [0, 1]$, la transformación más habitual para obtener una distancia δ a partir de una similaridad s es $\delta = 1 - s$. También se suelen hacer las transformaciones $\frac{1-s}{s}$, $\sqrt{1-s}$, entre otras.

Observación 2.2.2. A lo largo de este trabajo, se abusará de la terminología y no se hará distinción entre métrica y pseudométrica. Por otro lado, también se hará referencia de manera indistinta al término distancia, disimilaridad o métrica.

2.2.2. Medidas de distancia entre vectores de características

En la literatura existen numerosas distancias y similaridades para calcular la separación entre los dos vectores que describen a un par de objetos. La distancia más común es la distancia Euclídea, pero también existen otras funciones que son ampliamente utilizadas en función del tipo de variables y del ámbito de aplicación. Autores como Gower (1985), Cuadras (1989) y Deza y Deza (2009) exponen en sus trabajos un exhaustivo resumen de diferentes tipos de coeficientes y medidas de los que se ofrece un resumen en esta sección de aquellos que son empleados en los anexos A, B y C.

2.2.2.1. Distancias Métricas

De entre todas las posibles distancias que se pueden calcular entre dos objetos, la familia de **distancias de Minkowski** es ampliamente utilizada.

Dados dos objetos $x_i, x_j \in \mathbb{R}^d$ con coordenadas $\mathbf{x_i} = \{x_{i1}, ..., x_{id}\}$ y $\mathbf{x_j} = \{x_{j1}, ..., x_{jd}\}$ respectivamente, la distancia de Minkowski responde a la siguiente expresión:

$$\delta_p(x_i, x_j) = \left(\sum_{g=1}^{d} |\mathbf{x}_{ig} - \mathbf{x}_{jg}|^p \right)^{1/p} = \|\mathbf{x}_i - \mathbf{x}_j\|_p \tag{2.1}$$

Estas distancias también son conocidas como normas L_p y para valores del parámetro $p \in [1, \infty[$ se tiene que L_p es una métrica que cumple las propiedades P1 a P5 vistas en la sección anterior.

Casos particulares de la distancia de Minkowski

1. Cuando $p = 1$, se tiene el caso particular de la **distancia ciudad o de Manhattan**, que representa la distancia entre dos puntos mediante la suma de las diferencias absolutas de sus coordenadas:

$$\delta_1(x_i, x_j) = \sum_{g=1}^{d} |\mathbf{x}_{ig} - \mathbf{x}_{jg}| \tag{2.2}$$

2. Una modificación de la distancia de Manhattan es la **métrica de Canberra** que es invariante frente a cambios de escala:

$$\delta_C(x_i, x_j) = \sum_{g=1}^{d} \frac{|\mathbf{x}_{ig} - \mathbf{x}_{jg}|}{|\mathbf{x}_{ig}| - |\mathbf{x}_{jg}|} \tag{2.3}$$

3. Cuando el valor del parámetro $p = 2$, se tiene el caso particular de la **distancia Euclídea** que responde a la siguiente expresión y coincide

con la definición de la norma 2. La distancia Euclídea corresponde con la noción habitual empleada de distancia entre dos puntos del plano.

$$\delta_2(x_i, x_j) = \sqrt{\sum_{g=1}^{d}(\mathbf{x}_{ig} - \mathbf{x}_{jg})^2} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)'} = \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \tag{2.4}$$

4. Una expresión límite de la distancia de Manhattan es la llamada **distancia máxima**, que determina que la distancia entre dos puntos viene dada por el valor máximo de la mayor diferencia entre sus coordenadas:

$$\delta_\infty(x_i, x_j) = \text{máx}\{|\mathbf{x}_{ig} - \mathbf{x}_{jg}|\} \tag{2.5}$$

Se puede ver gráficamente en la figura 2.3[3] las diferencias en la distancia de Manahattan y la distancia Euclídea a la hora de calcular la distancia entre dos puntos.

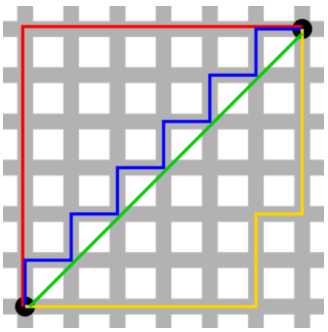


Figura 2.3: Distancia Euclídea (color verde) y Manhattan (rojo, azul y amarillo)

En la geometría Euclidiana la forma de los círculos es la que conocemos, sin embargo esta forma cambia cuando la distancia es determinada por una métrica diferente a la Euclídea. En la figura 2.4[4] se ilustran algunos de los casos de la familia L_p de distancias de Minkowski en un espacio vectorial bidimensional en los que todos los puntos de las distintos círculos unidad tienen distancia 1 al punto central.

En el caso en el que $p = 1$, el círculo unidad tiene forma de diamante, mientras que cuando $p > 2$, su representación es un cuadrado con esquinas redondeadas. A medida que el valor de p aumenta, el círculo unidad se convierte en un cuadrado. En el caso de $p < 1$, el círculo unidad se convierte en

[3] Imagen tomada de `https://en.wikipedia.org/wiki/Taxicab_geometry`
[4] Imagen tomada de Bellet, Habrard y Sebban (2015)

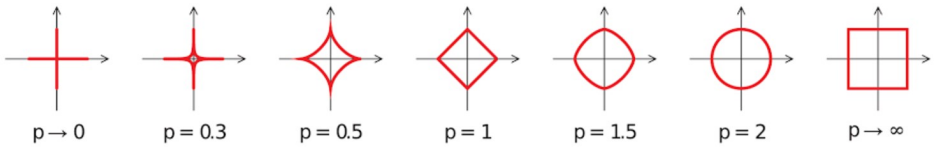


Figura 2.4: Distancias de Minkowski: círculos unidad para varios valores del parámetro p

convexo y por tanto no se cumple la desigualdad triangular, con lo que no se satisfacen las propiedades de las métricas tal y como se expone en la sección 2.2.2.2.

La distancia Euclídea es la más utilizada, pero presenta una serie de inconvenientes tal y como indica Cuadras (1989):

- no está acotada.
- no es invariante frente a cambios de escala de las variables.
- presupone que los datos son incorrelacionados y de varianza unidad.

En general, la invarianza frente a cambios de escala se resuelve dividiendo por un término que elimine este efecto. Esto conduce a la denominada **distancia Euclídea ponderada**, que se define por

$$\delta_M(x_i, x_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)' M (\mathbf{x}_i - \mathbf{x}_j)} = \|\mathbf{x}_i - \mathbf{x}_j\|_M \tag{2.6}$$

donde M es una matriz diagonal que se utiliza para estandarizar los datos y hacer la medida invariante ante cambios de escala. El efecto que tiene esta matriz diagonal es indicar qué dimensiones son más o menos importantes que otras.

La expresión obtenida δ_M se conoce también como **distancia de forma cuadrática** y la distancia Euclídea ponderada es un caso particular. Si la matriz M es la diagonal unitaria, la distancia cuadrática se convierte en una distancia Euclídea ordinaria .

Tal y como se apuntaba antes, la distancia Euclídea y en general las distancias de Minkowski asumen que no existe correlación entre los datos (Cuadras, 1989). La distancia de la forma cuadrática permite modelar la similitud en espacios vectoriales donde las características de los objetos tienen algún tipo de correlación. Un ejemplo representativo se tiene en el trabajo de Skopal (2007) en el que se considera el caso concreto de que los objetos sean imágenes en un espacio de color RGB (Red Green Blue). En esta situación, la componente verde estará más correlacionada con el azul que con el rojo y por tanto, es necesaria una distancia que pueda tener en cuenta la

relación existente entre las componentes de los vectores. En la expresión δ_M, la matriz M refleja esta relación.

En el caso particular de tomar $M = \Sigma$, donde Σ es la matriz de covarianzas de X, se obtiene la **distancia de Mahalanobis**[5] que se define como:

$$\delta_M(x_i, x_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)'\Sigma^{-1}(\mathbf{x}_i - \mathbf{x}_j)} \tag{2.8}$$

En general, se realiza un abuso de la terminología y se suele utilizar el término distancia de Mahalanobis para hacer referencia a las distancias cuadráticas δ_M generalizadas (Bellet et al., 2015) parametrizadas por una matriz $M \in \mathbb{S}_+^d$, donde $\mathbb{S}_+^d$ es el cono de las matrices simétricas Semidefinidas Positivas (SDP) d x d con valores reales. $M \in \mathbb{S}_+^d$ asegura que se satisfagan las propiedades de una pseudo-distancia.

Observación 2.2.3. En el texto se utiliza d_M para hacer referencia a la distancia de Mahalanobis y distinguirla de la distancia de forma cuadrática generalizada.

2.2.2.2. Distancias No Métricas y similaridades

Las medidas no métricas y de similaridad se han utilizado en muchas áreas de estudio como por ejemplo en la recuperación de información o búsqueda de similitud entre objetos en bases de datos multimedia.

En esta sección se exponen algunas de las distancias no métricas y similaridades más conocidas y empleadas.

Las distancias de Minkowski cumplen las propiedades de una métrica cuando el parámetro $p \geq 1$. Sin embargo, si $p \in]0, 1[$ la expresión $\delta_p(x_i, x_j) = \left(\sum_{g=1}^{d} |\mathbf{x}_{ig} - \mathbf{x}_{jg}|^p\right)^{1/p} = \|\mathbf{x}_i - \mathbf{x}_j\|_p$ se conoce como **distancia fraccionaria** L_p (Aggarwal et al., 2001), pero no se trata de una métrica, sino de una semimétrica ya que no cumple la desigualdad triangular (P5).

Una distancia que se utiliza muy a menudo es la **distancia del Coseno** que mide el coseno del ángulo que forman dos vectores de entrada. Su definición se basa en la medida del coseno que es una similaridad cuya expresión es

[5] La versión original de la distancia de Mahalanobis no mide la distancia entre dos puntos en un espacio vectorial, sino que hace referencia a la distancia entre un punto x_i de una determinada distribución y la media μ de dicha distribución con matriz de covarianza Σ como indica la siguiente expresión:

$$\delta_M(x_i, \mu) = \sqrt{(\mathbf{x}_i - \mu)'\Sigma^{-1}(\mathbf{x}_i - \mu)} \tag{2.7}$$

$$s_{cos}(x_i, x_j) = \frac{\mathbf{x}_i' \cdot \mathbf{x}_j}{\sqrt{(\mathbf{x}_i \cdot \mathbf{x}_i')(\mathbf{x}_j \cdot \mathbf{x}_j')}} \tag{2.9}$$

La medida del coseno es ampliamente utilizada en el campo de la recuperación textual (Baeza-Yates y Ribeiro-Neto, 1999), entre otros. Aunque s_{cos} es una medida de similaridad, mediante el cambio descrito en la sección 2.2.1 se puede convertir en la distancia del coseno con la tranformación $\delta_{cos}(x_i, x_j) = 1 - s_{cos}(x_i, x_j)$ que, como en el caso anterior, es una semimétrica.

Muchos de los trabajos de aprendizaje de similitudes se centran en aprender una función de similitud definida por $S_M = S_M(x_i, x_j) = \mathbf{x}_i' M \mathbf{x}_j / N(x_i, x_j)$ donde M es una matriz y $N(x_i, x_j)$ es un término de normalización.

En el caso de que $N(x_i, x_j) = \sqrt{(\mathbf{x}_i \cdot \mathbf{x}_i')(\mathbf{x}_j \cdot \mathbf{x}_j')}$, se tiene la función generalizada de similaridad del coseno que se ha definido antes. Si $N(x_i, x_j) = 1$, $S_M(x_i, x_j) = \mathbf{x}_i' M \mathbf{x}_j$ es la denominada **función bilinear de similitud** parametrizada por una matriz M que no es necesario que sea semidefinida positiva ni simétrica. De hecho, una de las ventajas es que al contrario que con las distancias de Minkowski, la distancia de Mahalanobis y la similaridad del coseno, se puede emplear como una medida de similaridad entre instancias de distinta dimensión mediante la elección de una matriz M no cuadrada.

Otro concepto muy empleado es el de **kernel**, que es una función de similaridad $k_e : X \times X \to \mathbb{R}$ que satisface $k_e(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$ donde $\phi : \mathbf{x}_i \to \phi(\mathbf{x}_i) \in F$ es un mapeo (*mapping*) de X a un espacio F dotado de un producto interno.

Las funciones de kernel más utilizadas son:

- Lineal: $k_e(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) = \mathbf{x}_i{}' \cdot \mathbf{x}_j$
- Polinomial de grado q: $k_e(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i{}' \cdot \mathbf{x}_j + c)^q$, donde c es un parámetro positivo.
- Función de base radial (RBF): $k_e(\mathbf{x}_i, \mathbf{x}_j) = exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$

Otro concepto empleado en la literatura es el de **divergencia** que surge como un concepto más débil que una distancia porque no satisface obligatoriamente la propiedad simétrica y/o la desigualdad triangular. Una divergencia se define en términos de una función estrictamente convexa y diferenciable, propiedades importantes para muchos algoritmos de aprendizaje automático.

Se pueden encontrar en la literatura ejemplos de divergencias muy empleadas como la de Bregman, Kullback-Leibler o Jeffrey. Las divergencias se

utilizan habitualmente en trabajos relacionados con imágenes de las que se puede consultar un resumen en el trabajo de Rubner et al. (2001).

Para una mayor descripción sobre médidas no métricas aplicadas en distintos dominios, se recomienda al lector el trabajo de Skopal y Bustos (2011).

2.3. Representación de Características

Encontrar una buena representación de los datos con los que se trabaja que se asemeje a la "ideal" se conoce como el *problema de representación de datos o características*.

Dado que la representación de los datos desempeña un papel clave en el rendimiento de los métodos de aprendizaje automático, su transformación en características útiles y significativas, se ha convertido en un tema de gran interés en la comunidad científica. De hecho, autores como Hertz (2006) indican que de disponer de la representación óptima de los datos, se podría eliminar la necesidad de agruparlos o clasificarlos, ya que tal representación ideal mapearía directamente cada instancia de entrada en su clúster o etiqueta de clase correspondiente.

Otro aspecto importante a considerar es que en las últimas décadas, las nuevas tecnologías han permitido el acceso a grandes volúmenes de datos, donde la calidad no suele ser óptima. Habitualmente a este tipo de datos se le denomina "información en bruto" o *raw data*, término que puede utilizarse para hacer referencia tanto a la representación original de los objetos adquirida directamente a través de sensores u otros dispositivos, como a una base de datos bien definida a partir de la cual se quiere obtener una representación de características mejor y más informativa. Para ello, es importante antes de llevar a cabo cualquier análisis o tarea, planificar una fase de preprocesado para eliminar datos ruidosos, irrelevantes o redundantes. Por otro lado, resulta interesante valorar si es posible aportar a los datos algún tipo de enriquecimiento mediante la combinación o fusión de distintos descriptores u otras características empleando fuentes de información externa o realizando alguna transformación al formato de los datos disponibles.

2.3.1. Combinación de Características

Existen múltiples técnicas de combinación o fusión de características. En su forma más general, la fusión de información (*Information Fusion*) se puede considerar como un proceso en el que los datos disponibles se combinan para diversos fines, tanto para la mejora de la comprensión del dominio de la aplicación como para encontrar representaciones de mayor calidad de los datos, como el perfeccionamiento en la toma de decisiones. La fusión se utiliza en un gran número de áreas de investigación: sistemas biométricos (Ross y Jain, 2003), sistemas de recuperación de la información (Kludas et al., 2008)

o métodos que trabajan con la confidencialidad y privacidad de los datos (Navarro-Arribas y Torra, 2012), por citar algunas.

El concepto de Fusión de la Información surgió en 1987 y se definió como un modelo multinivel para describir el proceso de fusión de los datos (White, 1987). Desde entonces múltiples autores han investigado en este campo [p.e., Hall y Llinas (1997); Torra y Narukawa (2007); Castanedo (2013)]. Una definición más actual es que la fusión de la información es un área muy amplia de investigación que estudia métodos para combinar datos o información suministrada por múltiples fuentes o combinar información de una única fuente, pero obtenida en diferentes instantes de tiempo. La información se puede fusionar en diferentes niveles: sensor, característica, puntuación (*match score*) o nivel de decisión (*decision level*), tal y como indican Villegas y Paredes (2009).

Debido a que la Fusión de Información es un campo de estudio que involucra a muchas áreas de conocimiento, resulta complicado establecer una única clasificación de las técnicas y métodos relacionados como señala Castanedo (2013). De hecho, el concepto de Fusión de la Información se puede interpretar desde distintas perspectivas y por tanto, adquiere diferentes significados en función del campo de aplicación (Chen y Meer, 2003).

Gran parte de la literatura identifica principalmente la fusión de información con la combinación de datos provenientes de múltiples sensores. En este problema un aspecto muy importante a considerar, son los distintos formalismos de representación de los datos (numéricos, categóricos, conjuntos fuzzy, etc.) para los que se han desarrollado en las últimas décadas técnicas adecuadas de agregación o integración como la media aritmética, la media ponderada, la Integral de Choquet o los operadores de la media ponderada ordenada, más conocidos como OWA (*Ordered Weighted Averaging*) . Esto queda fuera del alcance de esta Tesis Doctoral y se remite al lector para ampliar más información en este área a los trabajos de Yager (1988) o Torra et al. (2010).

En el área de la recuperación de la información, los objetos (una señal de audio, una imagen o un documento de texto, por ejemplo) como muestra la figura 2.5, quedan descritos por la concatenación de varios descriptores diferentes, pero complementarios (Faria et al., 2014) produciéndose habitualmente espacios de alta dimensión donde el aprendizaje es computacionalmente muy costoso debido a la maldición de la dimensión (*curse of dimensionality*) (Bellman y Bellman, 1961; Bishop, 1995).

Si se toma una base de datos en la que los objetos son imágenes, el contenido de cada una de ellas se puede definir matemáticamente mediante una representación vectorial que haga referencia a alguna característica como el color. Sin embargo, es complicado representar toda la información relevante de una imagen empleando una única característica y por tanto se suelen utilizar distintos descriptores visuales al tener un potencial mucho mayor para

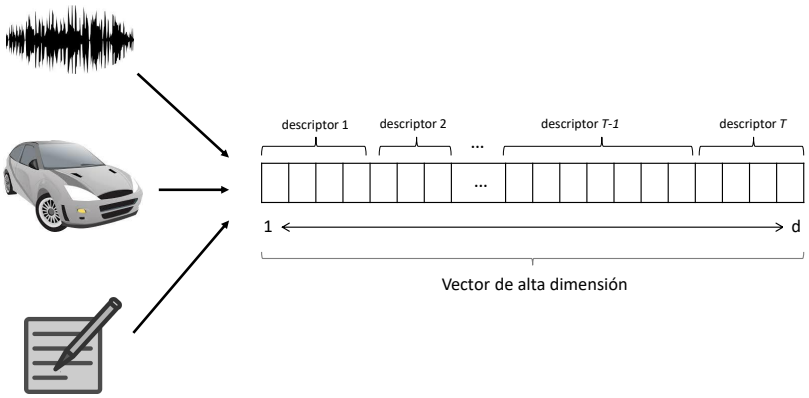


Figura 2.5: Vector de características agrupado por descriptores.

caracterizar una imagen como indican en su trabajo Datta et al. (2008).

Como consecuencia, un vector de características que describa una imagen puede estar formado por una combinación de distintos descriptores relacionados con aspectos como la forma (descriptores de Fourier y Zernike), el color (histograma de color, correlograma de color o vector de coherencia de color, por ejemplo) o la textura (matrices de coocurrencias, filtros de Gabor o características de Tamura).

En particular en el área de la Recuperación de la Información Basada en Contenido (CBIR) (Piras y Giacinto, 2017), a lo largo de los años, se han propuesto varias técnicas para combinar y fusionar los diferentes descriptores que caracterizan a una imagen (da Silva Torres et al., 2005; Yue et al., 2011; Bhowmik et al., 2014).

Cuando se tienen varios descriptores, se pueden emplear distintos esquemas de fusión de la información que han sido categorizados como "fusión temprana" (*early fusion*) o "fusión tardía" (*late fusion*) (Snoek et al., 2005; Zhang y Ye, 2009; Piras y Giacinto, 2017). Mientras que la primera opción utiliza una única medida o *score* que fusiona todos los descriptores, la segunda calcula múltiples medidas o *scores* sobre los descriptores por separado que luego combina. Entre las técnicas de combinación de *scores* más sencillas y habituales, figuran las basadas en la suma o el producto conocidas como *CombSum* o *CombProd*, entre otras, y que fueron introducidas por Shaw y Fox (1994).

En general, el uso de descriptores y medidas de distancia, relaciona el problema de aprendizaje de *scores* de similitud, con su combinación. Es por ello, que en los artículos de la primera contribución se emplean como comparativa al método propuesto, como se puede comprobar en el anexo A y en el anexo B, donde se realiza una síntesis de métodos relacionados con la fusión de *scores*.

2.3.2. Extracción de Características

El problema de seleccionar una representación adecuada de los objetos sigue siendo una cuestión compleja, para la que no es sencillo encontrar una solución adecuada. Entre las muchas formas posibles de representaciones de datos en el estado del arte, algunos formatos y/o combinaciones son más efectivos que otros para una determinada tarea a realizar. Por otro lado es importante considerar que las características en un dominio específico, no generalizan bien a otros dominios. Además, el uso de distintos descriptores tiene inconvenientes como el hecho de que alguno de los descriptores puede ser más relevante que otro o la alta dimensión que aportan al formato de entrada.

Por todo ello, se han sugerido numerosos métodos que se ocupan de cualquier transformación de los datos originales a las características que finalmente conformarán la entrada de los algoritmos de aprendizaje automático. El objetivo será mapear los datos (trasladarlos) en un espacio de características en el que estén mejor representados y convertirlos en un formato más informativo y significativo para la tarea a realizar. En muchos casos, las transformaciones realizadas también se pueden usar para reducir la dimensión del espacio de entrada, seleccionando únicamente las características relevantes.

Este tipo de transformaciones se conoce como *Extracción de Características* (Guyon et al., 2006) y se definen como un preproceso esencial en Reconocimiento de Patrones y Aprendizaje Automático que se descompone en dos partes: Selección de Características (*Feature Selection*) y Construcción de Características (*Feature Construction*).

En los últimos años se han desarrollado muchos métodos y dado que es un área activa de investigación, existe abundante literatura referente al problema de la representación de datos y la extracción de características [p.e., Bengio et al. (2013); García et al. (2015); Storcheus et al. (2015)]. En las siguientes secciones se expone un resumen de los principales métodos.

2.3.2.1. Selección de Características

Una de las técnicas más estudiadas en la extracción de características es la Selección de Características (*Feature Selection*), que se ocupa de seleccionar el mejor subconjunto de características del conjunto original (Guyon y Elisseeff, 2003), para reducir la dimensión del espacio de características y que los algoritmos de aprendizaje automático sean más rápidos.

Teniendo en cuenta todos esos aspectos, se puede definir la selección de características como un proceso en el que se escoge el subconjunto óptimo de características basado en algún criterio acorde a un determinado propósito (p.e., mejorar el rendimiento de un método en términos de rapidez, poder predictivo, simplicidad del modelo, etc.) (García et al., 2015).

La selección de características puede ser estudiada desde muchos enfoques. El más habitual consiste en categorizar las diferentes aproximaciones de la selección de características como métodos de filtrado (*filter methods*), métodos "envolventes" (*wrapper methods*) y métodos "incrustados" (*embedded methods*). A continuación se expone una breve síntesis de los aspectos más importantes de los distintos métodos tomando como referencia los trabajos de Guyon y Elisseeff (2003) y Storcheus et al. (2015).

Los procedimientos de filtrado seleccionan un subconjunto de características basado en algún criterio o puntuación que se calcula independientemente del algoritmo de aprendizaje. Algunos ejemplos de medidas de filtrado son la Puntuación de Fisher (*Fisher Score*) o la Ganancia de la Información (*Information Gain*).

Los métodos envolventes (*wrapper*) seleccionan el mejor conjunto de características en función del rendimiento del algoritmo de aprendizaje. Las estrategias más utilizadas por estos métodos son la Selección Secuencial hacia Adelante (*Sequential Forward Selection, SFS*) y la Eliminación Secuencial hacia Atrás (*Sequential Backward Elimination, SBE*). Un popular método wrapper para máquinas de soporte vectorial (SVMs) basado en la estrategia SBE fue propuesto por Guyon et al. (2002) y se conoce como SVM-RFE (*SVM-Recursive Feature Elimination*).

Las técnicas relacionadas con los métodos "embebidos" (*Embeddings*) extraen las características que mejor describen los datos mediante la proyección de los datos de entrada en un subespacio de menor dimensión (Cunningham y Ghahramani, 2015). Estos métodos son similares a los *wrappers* en el sentido de que las características se seleccionan específicamente para un determinado algoritmo de aprendizaje y durante el proceso de aprendizaje.

2.3.2.2. Construcción de Características

Otro enfoque particular de la extracción de características es la creación de nuevas características a partir de los datos de entrada, usualmente conocido como *Construcción de Características* (Liu y Motoda, 1998; Guyon et al., 2006). Las nuevas características construidas están definidas a partir de las originales para mejorar el poder discriminatorio de éstas últimas, por tanto, es muy importante la selección de características relevantes. En el caso de características numéricas, son muy utilizados los operadores algebraicos como la resta, la suma, el producto, el cociente o la media en la construcción de características (Liu y Motoda, 1998; García et al., 2015).

Un ejemplo de la aplicación de estos operadores se tiene en Hertz et al. (2004) en el que para representar un par de puntos, se emplea la concatenación de la suma y la diferencia de sus vectores como formato de entrada de los datos a un árbol de decisión empleando el algoritmo C4.5.

En la misma línea de trabajo, Woznica y Kalousis (2010) se enfrentan a

un problema de clasificación binaria donde las instancias de aprendizaje son las diferencias absolutas de pares de instancias originales. Otros autores en el área de verificación de rostros para descubrir si dos caras coinciden o no, presentan transformaciones de los datos originales utilizando las diferencias absolutas y los productos calculados elemento a elemento entre los vectores de características como instancias de entrenamiento. Se pueden consultar los trabajos de Kumar et al. (2009, 2011); Berg y Belhumeur (2012).

Recientemente ha surgido un enfoque en el que se emplean redes neuronales convolucionales (*CNN, Convolutional Neural Networks*) para obtener representaciones más significativas de los objetos conocidas como características aprendidas o *learned features* (Bengio et al., 2013; Antipov et al., 2015). Una de las ventajas es que es posible automatizar el paso de la extracción de características a partir de datos de entrada "crudos" (*raw data*) de modo que no se necesite la supervisión o ingeniería humana para la elaboración manual de las características (Bengio y Courville, 2013).

En muchos casos, con el objetivo de mejorar la expresión de los datos, se opta por una *Expansión de características*, que consiste en la ampliación de las características originales mediante las nuevas características construidas a través de las distintas tranformaciones u operaciones que se pueden realizar sobre los datos.

Habitualmente, la expansión conlleva el aumento de la dimensión de las características originales, sin embargo, distintos trabajos relacionados con la aplicación de CNN como el de Wu et al. (2016) o Park et al. (2016), señalan la obtención de mejores resultados mediante la expansión o combinación de las características extraídas manualmente con otras aprendidas a través de una red neuronal en determinados contextos como el de identificación de personas. Más ejemplos de este tipo de amplificación o expansión de características se pueden encontrar en Simonyan y Zisserman (2014) o Ng et al. (2015).

Otros ejemplos de uso de este enfoque se tienen como aplicación en métodos generales en el reconocimiento de patrones (Yao et al., 2003; Tsai et al., 2011) y en contextos específicos, como la recuperación de texto (Dalton et al., 2014), la detección de intrusos (Guo et al., 2014) y el análisis de sentimientos (Jotheeswaran y Koteeswaran, 2015).

En particular, Tsai et al. (2011) y Guo et al. (2014) emplean una expansión de características basándose en el uso de distancias para formar nuevas características y mejorar el rendimiento de un algoritmo de clasificación. En estos trabajos, se extienden los vectores de características originales mediante el cálculo de distancias de cada muestra de datos a un determinado número de centroides localizados a través un algoritmo de agrupamiento (*clustering*).

2.4. Relación de la representación de Características con el aprendizaje de métricas

Después de aplicar los métodos de representación de características, y en concreto los métodos de construcción de características, puede que los objetos estén representados por vectores de características de alta dimensión, lo que conlleva un alto coste computacional. Para hacer frente a este problema, es común utilizar métodos de reducción de la dimensión, cuyo objetivo es proyectar los objetos en un espacio de menor dimensión que el original donde se preserve la geometría y/o la estructura de los datos originales.

Existen múltiples métodos, lineales y no lineales, que se han empleado en muchos estudios y en distintas áreas de aplicación como el Análisis de Componentes Principales (Pearson, 1901), el Análisis de Componentes Independentes (*Independent Component Analysis, ICA*) (Hyvärinen y Oja, 2000), el Análisis lineal discriminante (*Linear Discriminant Analysis, LDA*) (Fisher, 1938), mapeo isométrico (*Isomap*) (Tenenbaum et al., 2000) o el mapeo de valores propios de Laplace (*Laplacian Eigenmaps, LE*) (Belkin y Niyogi, 2002), entre otros.

En los métodos de aprendizaje de métricas, por otro lado, la métrica aprendida puede utilizarse para proyectar los datos en un nuevo espacio de características, proporcionando una nueva representación de los vectores de características en un nuevo espacio con menor dimensión, relacionando de esta manera el aprendizaje de métricas con las técnicas de extracción de características (Globerson y Roweis, 2005), además de con los métodos de reducción de la dimensión.

Un ejemplo de la conexión entre el aprendizaje de métricas y la representación de los datos se puede comprobar al considerar la relación de la distancia de Mahalanobis con las transformaciones lineales (Perez-Suay y Ferri, 2008). Aunque la distancia de Mahalanobis se define originalmente para una matriz de covarianzas relacionada con el problema que se esté tratando, realmente se puede definir esta distancia a partir de cualquier matriz simétrica y semidefinida positiva M que utilizando la descomposición de Cholesky, puede expresarse como el producto de una determinada matriz por su transpuesta (Peña, 2002; Bellet et al., 2015), es decir, $M = W'W$, donde $W \in \mathbb{R}^{r \times d}$ y r es el rango de M.

Así, se tiene que la distancia de Mahalanobis se puede escribir como:

$$\begin{aligned} d_M(x_i, x_j) &= \sqrt{(\mathbf{x}_i - \mathbf{x}_j)' M (\mathbf{x}_i - \mathbf{x}_j)} = \\ &= \sqrt{(\mathbf{x}_i - \mathbf{x}_j)' W' W (\mathbf{x}_i - \mathbf{x}_j)} = \\ &= \sqrt{(W\mathbf{x}_i - W\mathbf{x}_j)' (W\mathbf{x}_i - W\mathbf{x}_j)} = d_2(W\mathbf{x}_i, W\mathbf{x}_j) \end{aligned} \tag{2.10}$$

Por tanto, la distancia de Mahalanobis M es equivalente a encontrar una transformación lineal W sobre el espacio original de características y luego usar la métrica de distancia Euclídea sobre el espacio transformado. La transformación W induce una métrica que generalmente se aprende de modo que los valores de las distancias entre ejemplos similares sean menores en comparación con los ejemplos de clases disimilares.

Cuando la métrica de Mahalanobis M considerada no es de rango completo, es decir $rango(M) = r < d$, resulta equivalente a realizar una proyección lineal de los datos en un espacio de menor dimensión r, con lo que se obtiene una representación de los datos más compacta y un menor coste computacional al calcular las distancias, especialmente cuando el espacio original de características es de alta dimensión.

En el área de Aprendizaje de métricas, han surgido un gran número de publicaciones en las últimas décadas desde el trabajo seminal propuesto por Xing et al. (2002). Otros autores como Bar-Hillel et al. (2003); Davis et al. (2007a); Weinberger y Saul (2009) o Köstinger et al. (2012) han realizado distintas aportaciones de los que se dan más detalles en las contribuciones de la Tesis en los anexos B y C.

La conexión de los métodos de aprendizaje de distancias métricas con la extracción de características y la reducción de la dimensión, nos llevó a plantear una serie de cuestiones que se abordan en esta Tesis: ¿existe algún método alternativo cuando no son necesarias las propiedades de las métricas para establecer un ranking de similitud entre pares de objetos? ¿se pueden combinar estos métodos con la extracción de características y el uso de distancias? ¿pueden las distancias aportar información "*per se*" a la vez que ayudar en la reducción de la dimensión del formato de características? ¿la combinación de distintos tipos de distancias en una nueva representación de las características tiene efecto sobre el rendimiento de clasificadores?

En este contexto, haciendo uso de la configuración habitual de los datos de entrada en problemas de aprendizaje de métricas dada por pares de objetos etiquetados como similares o disimilares, esta Tesis propone una nueva representación de las instancias originales que describen a los objetos basándose en el uso de un conjunto de distancias métricas y no métricas. Este formato de entrada se utiliza como las instancias de entrenamiento para un clasificador cuyo *score* se empleará para construir un ranking de similitud entre pares de objetos.

En las publicaciones derivadas de la Tesis que se presentan en los tres anexos y de las que se dan todos los detalles en el siguiente capítulo, se estudia en primer lugar la mejora de esta representación de los datos basada en distancias frente al tradicional uso de la representación basada en la concatenación de características. A continuación, se investiga si este formato basado en distancias se puede combinar con técnicas de extracción de características con las que obtener un mejor rendimiento del clasificador.

Parte II

Contribuciones y conclusiones

Capítulo 3

Contribuciones

Resumen:
Este capítulo presenta los detalles de los trabajos publicados durante la realización de la Tesis Doctoral y resume las contribuciones derivadas de los mismos.

3.1. Publicaciones

El trabajo que conforma la presente Tesis Doctoral se ha publicado en revistas indexadas en el *Journal Citation Report (JCR)* y se ha presentado en distintos Congresos internacionales en coautoría con los directores de la Tesis.

Los detalles de los artículos publicados en revistas son:

- Emilia López-Iñesta, Francisco Grimaldo and Miguel Arevalillo-Herráez
  Título: Combining feature extraction and expansion to improve classification based similarity learning.
  Año de publicación: 2017
  Revista: *Pattern Recognition Letters, 93*, 95-103.
  Factor de impacto: 1.586 (Q2).
  DOI: 10.1016/j.patrec.2016.11.005.

- Emilia López-Iñesta, Francisco Grimaldo and Miguel Arevalillo-Herráez
  Título: Learning Similarity Scores by using a family of distance functions in multiple feature spaces.
  Año de publicación: 2017
  Revista: *International Journal of Pattern Recognition and Artificial Intelligence, 31*(8), 1750027, 21 pages.
  Factor de impacto: 0.915 (Q3).
  DOI: 10.1142/S0218001417500276.

- Emilia López-Iñesta, Francisco Grimaldo and Miguel Arevalillo-Herráez
  Título: Classification similarity learning using feature-based and distance-based representations: A comparative study.
  Año de publicación: 2015
  Revista: *Applied Artificial Intelligence, 29*(5), 445-458.
  Factor de impacto: 0.54 (Q4).
  DOI: 10.1080/08839514.2015.1026658

Las comunicaciones que se expusieron en Congresos internacionales fueron:

- Emilia López-Iñesta, Miguel Arevalillo-Herráez and Francisco Grimaldo
  Título: *Boosting Classification Based Similarity Learning by using Standard Distances.*
  Año: 2015
  Congreso: International Conference of the Catalan Association for Artificial Intelligence, Valencia, October 2015.

- Emilia López-Iñesta, Francisco Grimaldo and Miguel Arevalillo-Herráez
  Título: *Comparing feature-based and distance-based representations for classification similarity learning.*
  Año: 2014
  Congreso: International Conference of the Catalan Association for Artificial Intelligence, Barcelona, October 2014.

- Emilia López-Iñesta, Miguel Arevalillo-Herráez and Francisco Grimaldo
  Título: *Classification-based multimodality fusion approach for similarity ranking.*
  Año: 2014
  Congreso: 17th International Conference on Information Fusion, Salamanca, July 2014.

Las publicaciones derivadas de este trabajo doctoral pueden agruparse en dos contribuciones como muestra el mapa conceptual de la figura 3.1, en el que se hace una breve síntesis de los principales conceptos y áreas de estudio en las que se ha centrado esta Tesis Doctoral y que han sido expuestos en el planteamiento del problema en la sección 2.1. En el texto que sigue, se explica el contenido de las contribuciones y se indican en negrita los conceptos que configuran los nodos del mapa conceptual.

Tal y como indica el mapa conceptual, el cálculo de la similitud (**Similarity measurement**) entre pares de objetos puede ser abordado tanto desde el aprendizaje supervisado de distancias métricas (**Metric Learning**) como

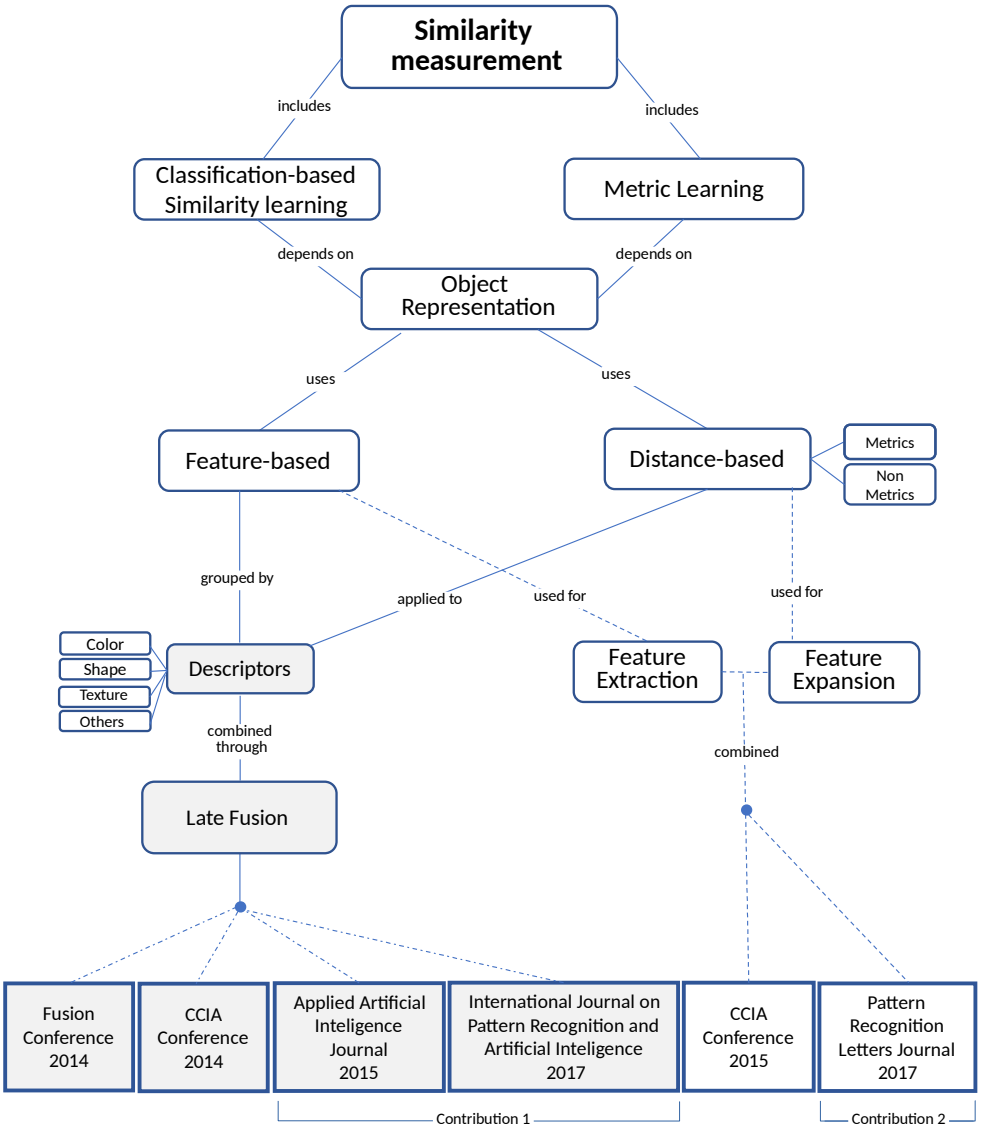


Figura 3.1: Principales conceptos, áreas de estudio, publicaciones y contribuciones de la Tesis Doctoral

de los métodos de clasificación para el aprendizaje de medidas de similitud (**Classification-based Similarity Learning**). Ambos planteamientos hacen uso de un enfoque común empleando ejemplos de entrenamiento compuestos, por ejemplo, de pares de objetos etiquetados como similares o disimilares.

Muchos de los trabajos sobre el aprendizaje de métricas han reportado

resultados consistentemente mejores que el aprendizaje de similitudes empleando métodos de clasificación en una variedad de contextos (Köstinger et al., 2012).

Sin embargo, a través de una cuidadosa selección del formato de entrada de los objetos (**Object Representation**), los clasificadores pueden obtener resultados que superen a los métodos de aprendizaje de métricas más avanzados.

3.1.1. Primera Contribución

La primera contribución de la Tesis está formada por los dos artículos siguientes:

- "Classification similarity learning using feature-based and distance-based representations: A comparative study". *Applied Artificial Intelligence, 29*(5), 445-458. (2015).
- "Learning Similarity Scores by using a family of distance functions in multiple feature spaces". *International Journal of Pattern Recognition and Artificial Intelligence, 31*(8). (2017).

En estos trabajos se plantea el uso de métodos de clasificación para la obtención de puntuaciones con las que establecer un ranking de similitud entre pares de objetos empleando la configuración habitual en problemas de aprendizaje de métricas dada por pares de objetos etiquetados como similares o disimilares.

A la hora de emplear un clasificador, uno de los aspectos más importantes que se ha de tener en cuenta para su rendimiento óptimo, tal y como se ha expuesto en la motivación de este trabajo, es el formato de datos de entrada proporcionado al clasificador.

En ambos artículos se presenta un enfoque general que es válido en cualquier contexto en el que los objetos estén representados por vectores numéricos (**Feature-based**) agrupados por varios descriptores (**Descriptors**) que se relacionan con diferentes características. Por ejemplo, los coeficientes cepstrales derivados de la escala Mel (*MFCC, Mel-Frequency Cepstral Coefficients*) y otras características relacionadas con el timbre y el tempo en el tratamiento digital de señales de voz y audio o el histograma de color o descriptores relacionados con la textura en el caso de las imágenes.

Es importante que estos vectores numéricos sean lo suficientemente representativos como para contener la información relevante. Es por ello que a menudo se utilizan varios descriptores para tener una buena y completa representación del objeto, lo que permite al clasificador trabajar con toda la información disponible. Sin embargo, esta opción comporta en múltiples ocasiones un problema debido a la alta dimensión de los vectores de características. Otra opción alternativa consiste en utilizar una representación de los

objetos basada en distancias (**Distance-based**). De hecho, es una práctica frecuente el uso de diferentes medidas de similitud, cada una actuando sobre cada descriptor que posteriormente se combinan para producir un valor de puntuación de similitud.

Teniendo en cuenta esta idea, la propuesta presentada en esta contribución emplea como función de (di)similitud la puntuación obtenida por un clasificador suave al que se le trata de proporcionar un resumen más informativo de los datos originales y con una menor dimensión.

En particular, se presenta una función de transformación que actúa como una capa de preprocesado que opera en el espacio de características original y que se basa en una familia de funciones de distancias δ_h (métricas y no métricas). La capa de preprocesado define un nuevo vector de características para cada par de objetos (x_i, x_j), tal y como refleja la figura 3.2. Estos son proporcionados como datos de entrenamiento al clasificador, que siguiendo un esquema de fusión tardía (**Late Fusion**) ejerce un papel de combinador flexible de las distancias y produce una puntuación que servirá para predecir la similitud de cualquier par de objetos sin etiquetar.

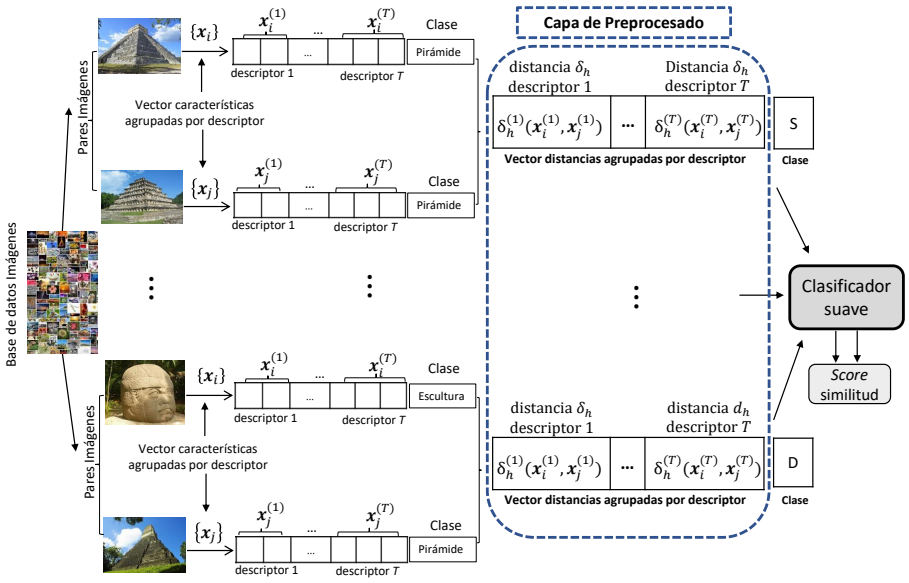


Figura 3.2: Capa de preprocesado basada en distancias calculadas por descriptor que se emplean en un esquema de *Late Fusion*

Cabe recalcar que las distancias son calculadas de manera independiente en cada espacio de representación de los T distintos descriptores que caracterizan a los objetos. En la figura anterior (3.2) se representa con $\mathbf{x}_i^{(t)}$ y $\mathbf{x}_j^{(t)}$, $t = 1, ..., T$ al conjunto de características que corresponde al descriptor t de $\mathbf{x}_i$ y $\mathbf{x}_j$ respectivamente.

Para ilustrar el método propuesto en los dos artículos (en anexos A y B) que conforman esta primera contribución se han utilizado bases de datos donde los objetos son imágenes, aunque el método propuesto es válido para cualquier tipo de objetos caracterizados por varios descriptores tal y como se expone en el segundo artículo. Entre las bases de datos empleadas en estos dos trabajos se incluye *Corel*[1], uno de los conjuntos más empleados en el ámbito de recuperación de imágenes.

A pesar de que existe una pérdida implícita e inevitable de información cuando las características originales se resumen con un valor de distancia, el proceso de agrupación supone una inyección de conocimiento adicional debido a que el uso de la medida de distancia, supone un emparejamiento implícito entre las características de dos imágenes y además, el uso de descriptores aporta información sobre la asociación de las características. Por otra parte, generalmente existe una correlación entre similitud semántica y valores pequeños de distancia.

En el primer artículo, el objetivo principal es comparar el rendimiento obtenido a partir de las representaciones basadas en características y las que emplean múltiples distancias cuando se aplican a un entorno de aprendizaje de similitud mediante métodos de clasificación, así como analizar la influencia de diferentes tamaños de datos de entrenamiento. Por lo tanto, el objetivo es doble: por un lado, se quiere estudiar la capacidad de un clasificador (una máquina de soporte vectorial) para hacer frente a la alta dimensión de las características a medida que el tamaño del entrenamiento crece; por otro lado, se quiere probar bajo qué circunstancias la reducción de la dimensión conduce a mejores resultados que el tratamiento de los objetos en su totalidad.

En este primer trabajo se emplea una combinación de distancias de Minkowski para la representación multidistancia de los objetos debido a que son medidas ampliamente utilizadas en la literatura. Teniendo en cuenta que el método propuesto puede ser considerado desde la perspectiva de un enfoque de fusión tardía de distintas puntuaciones, se incluyen métodos de combinación de *scores* en la comparación con la propuesta realizada.

En el segundo artículo, se emplean nuevas distancias para buscar una combinación que aporte una mayor diversidad de información al clasificador y que mejore su rendimiento. En concreto, se utilizan las distancias de Minkowski empleadas en el primer artículo con la distancia coseno y la de Mahalanobis.

Además, se incluyen técnicas representativas del estado del arte de aprendizaje de distancias métricas como nuevos métodos de comparación con la propuesta presentada. Asimismo, con el objeto de proporcionar un método de comparación con el mismo poder de representación que el presentado en esta contribución, parece natural la inclusión de una combinación lineal

[1] `https://archive.ics.uci.edu/ml/datasets/corel+image+features`

ponderada de las distancias calculadas entre descriptores.

Por último, con el fin de evaluar el comportamiento y rendimiento de los métodos y analizar si sus diferencias son estadísticamente significativas, también se ha realizado una prueba no paramétrica de Friedman, así como test a posteriori para cada base de datos analizada.

Las conclusiones de esta primera contribución muestran que el uso de las distintas distancias tiene un doble efecto. Por un lado, la reducción de la dimensión beneficia la aplicabilidad de los métodos de clasificación, reduciendo la sobrecarga y aumentando la generalización. Por otro lado, la consideración conjunta de varias distancias en cada subespacio puede proporcionar información complementaria al clasificador sobre el concepto semántico de similitud entre pares de objetos.

Los resultados obtenidos muestran que el método propuesto supera al resto de técnicas subrayando los beneficios de utilizar múltiples distancias en cada espacio de representación.

3.1.2. Segunda Contribución

El artículo "Combining feature extraction and expansion to improve classification based similarity learning" publicado en la revista *Pattern Recognition Letters* en 2017 conforma la segunda contribución de la Tesis Doctoral.

Como en los trabajos anteriores, se parte de la idea de que una de las maneras más sencilla y habitual en la que se pueden proporcionar los datos de entrenamiento de pares de objetos a un clasificador es mediante la concatenación de los vectores de características que describen los objetos y una etiqueta que indique si los objetos son similares o disimilares.

Para denotar a los pares se utilizará $p_k = (x_{k_1}, x_{k_2})$ y para la etiqueta se usará $l_k \in \{similar, disimilar\}$ donde $x_{k_1}, x_{k_2} \in X$ y $k = 1, \ldots, K$. Los datos de entrenamiento en este caso tienen la estructura $\{\mathbf{p_k}, l_k\}$, donde $\mathbf{p_k} = \mathbf{x_{k_1}}||\mathbf{x_{k_2}}$ y $||$ denota el operador concatenación.

Otra manera de afrontar este problema de aprendizaje, es considerando los vectores concatenados $\{\mathbf{p_k}\}$ como si fueran datos en bruto sin procesar (*raw data*) y planteando un nuevo problema de extracción de características sobre las características disponibles. Esto es, en lugar de utilizar el conjunto original de vectores $\{\mathbf{p_k}\}$, la idea es utilizarlos para construir un nuevo conjunto de características, $\{\mathbf{p'_k}\}$, que sea más preciso para tareas de clasificación.

El objetivo sería por tanto, llevar a cabo una transformación de las características originales para proporcionar un formato de entrada de datos al clasificador más informativo, de manera que se obtuviera una mejor puntuación de la similitud entre pares de objetos independientemente del método de clasificación empleado.

Esta segunda contribución de la Tesis presenta un enfoque híbrido que

combina la extracción de características (**Feature Extraction**) y la ampliación de las mismas (**Feature Expansion**). En una primera fase, se produce una transformación de las variables originales que permite obtener una sustancial ventaja sobre el uso tradicional que se hace de las características originales concatenadas y se demuestran las ventajas de esta trasformación sobre los clasificadores.

A continuación de la fase de extracción, se realiza una ampliación de funciones en la que se utilizan distancias, cuyo empleo se inspira en la primera contribución de esta Tesis, aunque su uso es significativamente diferente ya que las distancias calculadas se utilizan para ampliar, y no reemplazar la información que aparece en $\mathbf{p_k}$ como se observa en la figura 3.3.

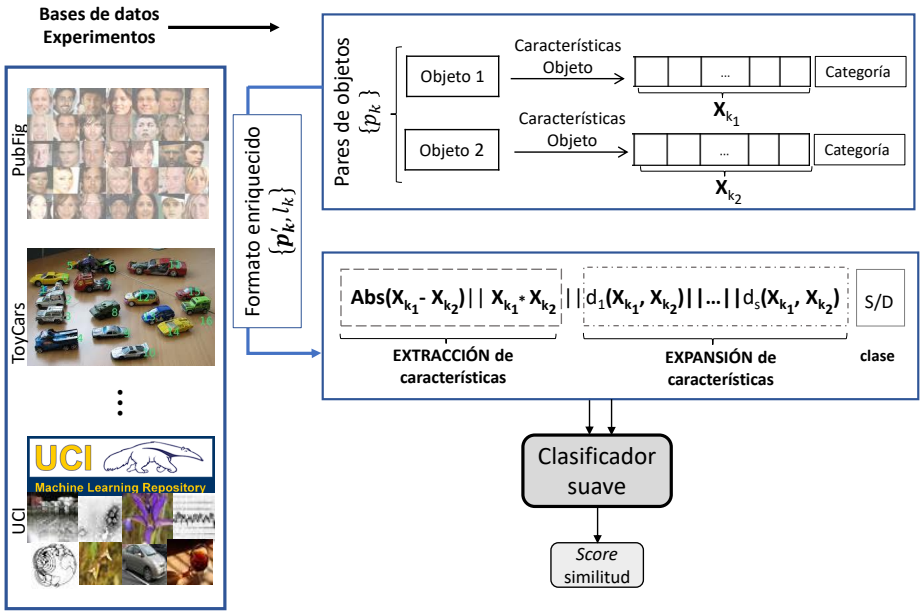


Figura 3.3: Nuevo formato enriquecido de los datos de entrada al clasificador después de las fases de Extracción y Expansión de características

El hecho de emplear las distancias en la fase de expansión implica un aumento en la dimensión original de los datos, objetivo contrario a la finalidad de la primera contribución en donde parte del interés del uso de las distancias radicaba en reducir la dimensión de entrada a un clasificador. En esta segunda contribución, el hecho de si los datos están agrupados o no por descriptores resulta irrelevante para el método, consiguiéndose un enfoque más general.

El método se compara con técnicas de extracción de características del estado del arte y métodos de aprendizaje de métricas. Los resultados obtenidos muestran rendimientos comparables en favor del método propuesto en

distintos contextos como el reconocimiento de objetos o la verificación facial.

Para mostrar la aplicación general del método propuesto, en esta segunda contribución, además de conocidas bases de datos relacionadas con imágenes (LFW[2], PubFig[3], ToyCars[4]) asociadas con tareas como la verificación de rostros o el reconocimiento de objetos, se emplean una variedad de conjuntos de datos del repositorio UCI[5] que ilustran múltiples contextos de uso.

Una vez expuestas las contribuciones de este trabajo doctoral, en el siguiente capítulo se detallan las conclusiones generales obtenidas a través de las publicaciones realizadas, así como las futuras líneas de investigación que se pueden plantear.

[2] http://vis-www.cs.umass.edu/lfw/
[3] http://www.cs.columbia.edu/CAVE/databases/pubfig/
[4] http://lear.inrialpes.fr/people/nowak/similarity/index.html
[5] http://archive.ics.uci.edu/ml

Capítulo 4

Conclusions

In this last chapter of the Thesis, the conclusions obtained from this Doctoral Dissertation are presented, as well as ideas that can be developed as future work.

4.1. General conclusions

This Doctoral Thesis presents a method of learning scores using supervised classification techniques to establish a ranking of pairs of objects based on their similarity that is proposed as an alternative to the use of metric learning techniques.

In both approaches, the representation format for the pairs of examples is very important to obtain good results, since it is the easiest way to represent them by concatenating the feature vectors in a given multidimensional space. This leads to a very common problem related to the excessive dimensionality of the data, which makes it difficult to learn both metric learning methods and machine learning algorithms.

In general, metric learning methods have shown a superior performance compared to the classification techniques. However, in view of the results obtained in this Doctoral Thesis, it has been demonstrated after an exhaustive experimentation that through a certain transformation of the input data format, the methods based on classification can obtain better results than metric learning techniques (Köstinger et al., 2012).

Taking as a starting point a characterization of a set of objects through several descriptors, the first contribution of this Thesis (annexes A and B) shows the potential of using a representation of pairs of objects based on distances rather than the regular use of the concatenation of feature vectors when using a soft classifier to establish a similarity ranking between the pairs.

In this contribution, the distances are computed independently in each descriptor space, turning the original expression of each pair of objects into a set of distance values. This causes that part of the original information to be lost, however, in return, it allows dimensionality to be reduced. On the other hand, the simultaneous use of different distance functions makes it possible to recover some of the information that is lost by summarizing the original features using a single distance.

Another important aspect is that the distances provide an enrichment in the form of auxiliary information that comes from the available data, offering details of the relationships between each pair of objects and the usual correlation between small distance values and semantic similarity (López-Iñesta et al., 2014b, 2015b).

It should also be noted that the combination of different metric and non-metric distances allows emphasizing different types of relationship and information of the objects and classes to be distinguished, thus contributing to the computation of a similarity score that reflects how similar two objects are. The results of the experiments carried out in López-Iñesta et al. (2014a), López-Iñesta et al. (2015a) and López-Iñesta et al. (2017b) show that it is the joint use of the distances which enhances the performance of the classifiers, rather than the use of a particular distance measure.

It has also been observed that a key aspect for the classifier performance is related to the balance of the training set classes. In this regard, in the first contribution the effect of class imbalance was studied and the desirability of finding an adequate proportion of similar and dissimilar pairs. Although the choice of the optimal percentage will always depend on the database and the application domain, the use of a training set with balanced classes has had a positive effect on the 40 databases analyzed in the published articles.

In the second article, López-Iñesta et al. (2017b), it is observed that metric learning methods do not show good performance with respect to the proposal based on classification because of the high dimensionality of the input. The results show that the simultaneous use of all descriptors does not provide an advantage over the use of a single descriptor. While a single descriptor may work relatively well, performance deteriorates when all descriptors are employed under an early fusion scheme, which leads us to confirm the use of late fusion schemes in this Thesis.

The use of classifiers to combine distances allows us to learn robust similarity measures that work in a homogeneous way regardless of the semantic category under which objects are labeled in a database. The proposed technique always obtains the best performance when applied to the majority categories and only encounters some difficulty when facing those with a very small number of objects (e.g. the *Dracs* category of the *Small* Database, that only has 7 objects).

Also, in this first contribution, a complete evaluation of the method is

carried out by comparing it with other state-of-the-art techniques about score combination and metric learning, confirming the benefits of the integration of several distance functions in the methods which follow a late fusion scheme using a soft classifier.

In the second contribution in annex C, we propose the design of a new representation of objects that combines both feature extraction and feature expansion in a classification approach for similarity learning. The conclusions show that the extraction phase achieves a change in the characterization of the objects that improves the separability of the classes and therefore the performance of the classifier.

On the basis of the good results of the first contribution of this Thesis, the expansion phase is based on the use of a set of standard distances in order to enrich the format of the objects obtained after the extraction phase. While it is true that the dimensionality of the data of the extraction phase increases in a set of values proportional to the number of distances employed, the use of those brings an additional injection of knowledge due to the implicit pairing between the feature vectors of each pair of objects as indicated above.

A relevant result is that the addition of distances works positively on the performance of both linear and non-linear classifiers. This is reflected in the results of the article López-Iñesta et al. (2017a), which indicate that both in specific contexts as object recognition or more general applications (UCI database), the performance of classifiers using only the data format corresponding to the expansion of distances ("EXPAN") always has a performance superior to the concatenation of the vectors of characteristics of pairs of objects (format "RAW"), regardless of whether the classifier is linear or non-linear.

In addition, it is proven in this contribution that, since descriptions of the objects are not necessary through descriptors, the proposed method becomes more general and can be applied in multiple contexts, with higher performances than those obtained by the metric learning methods.

As a global conclusion of this Doctoral Dissertation, it can be confirmed that the change of representation of the original instances that describe a set of objects through a process of extraction of characteristics combined with an expansion of these characteristics using a set of different metric and non metric distances becomes a key aspect when using a soft classifier to define a similarity ranking between pairs of objects as well as using it as an alternative method to metric learning techniques.

4.2. Future work

In light of the contributions presented in this Doctoral Thesis, a number of issues are considered that suggest further research and will define future research lines.

As has been indicated throughout the Thesis, classification algorithms strongly depend on the quality of the data, which sometimes include noisy or non-representative cases that are detrimental to its performance. Therefore, the selection of instances is suggested to discard useless or harmful examples in the training set and to improve the accuracy of the classification model.

Taking into account the classifier used, the use of multiple kernels can be considered, instead of using a single one. The different kernels can correspond to the use of different notions of similarity or can use information from multiple sources (both different representations and different subsets of characteristics).

Regarding the use of features, the proposed method can be applied to all types of vectorial representations. Consequently, it may be interesting to apply the feature extraction and feature expansion to the specific case of having available features extracted by means of deep learning methods such as CNN (*Convolutional Neural Networks*).

Finally, another line of work is the study of other distance functions combinations using aggregation operators such as OWA (*Ordered Weighted Averaging*) to check if additional gains can be achieved.

Parte III

Anexos

Apéndice A

Classification similarity learning using feature-based and distance-based representations: a comparative study

Autores:
Emilia López-Iñesta, Francisco Grimaldo, Miguel Arevalillo-Herráez

Revista:
Applied Artificial Intelligence: An International Journal

Año 2015, volumen 29, número 5

ISSN: 0883-9514

DOI
http://www.tandfonline.com/doi/full/10.1080/08839514.2015.1026658

A.1. Notation

Notation	Description
m	Feature space dimensionality
$X = \{x_i\}$	Objects collection
$\mathbb{F}$	Feature space
$\mathbb{F}^(t)$	Feature space related to descriptor t
x_i	Generic object
$\mathbf{x}_i$	Feature vector of a generic object x_i
S	Similar pairs of objects
D	Dissimilar pairs of objects
d	Distance
$\hat{d}$	Normalized distance
p	Minkowski distance parameter
N	Family of distances size
T	Descriptors size
L_p	Multidistance representation based on Minkowski distances
L_1	Multidistance representation based on Minkowski distance $p = 1$
w	Transformation function from original data to $N \cdot T$ distances
$d_n^{(t)}$	Distance between two descriptors objects $\mathbf{x}_i^{(t)}$ and $\mathbf{x}_j^{(t)}$
w'	Transformation function from original data to m distances
γ	Parameter Gaussian Radial Basis Function (RBF)
C	Parameter Gaussian Radial Basis Function (RBF)

Classification similarity learning using feature-based and distance-based representations: a comparative study

A.2. Abstract

Automatically measuring the similarity between a pair of objects is a common and important task in the machine learning and pattern recognition fields. Being an object of study for decades, it has lately received an increasing interest from the scientific community. Usually, the proposed solutions have used either a feature-based or a distance-based representation to perform learning and classification tasks. This paper presents the results of a comparative experimental study between these two approaches for computing similarity scores using a classification-based method. In particular, we use the Support Vector Machine, as a flexible combiner both for a high dimensional feature space and for a family of distance measures, to finally learn similarity scores. The approaches have been tested in a Content-Based Image Retrieval context, using three different repositories. We analyze both the influence of the different input data formats and the training size on the performance of the classifier. Then, we found that a low dimensional multidistance-based representation can be convenient for small to medium-size training sets whereas it is detrimental as the training size grows.

Keywords: similarity learning, distance-based representation, training size

A.3. Introduction

Learning a function that measures the similarity between a pair of objects is a common and important task in applications such as classification, information retrieval, machine learning and pattern recognition. The Euclidean distance has been widely used since it provides a simple and mathematically convenient metric on raw features, even when dealing with a small training set, but it is not always the optimal solution for the problem being tackled (McFee y Lanckriet, 2010). This has led to the development of numerous similarity learning techniques (Bar-Hillel et al., 2003; Davis et al., 2007b) aimed to build a model or function that, from pairs of objects, produces a numeric value that indicates some kind of conceptual or semantic similarity and also allows to rank objects in descending or ascending order according to this score.

Some studies have have put their attention into automatically learning a similarity measure that satisfies the properties of a metric distance (Xing et al., 2002; Globerson y Roweis, 2005) from the avalaible data (e.g. in the form of pairwise constraints obtained from the original labeled information) and have turned supervised metric learning into a topic of great interest (Bellet et al., 2012). Under this scope, when the properties of a metric are not required, a similar setting can also be used to train a classifier to decide whether a new pair of unlabeled objects is similar or not, an approach that is named as classification similarity learning.

Classification similarity learning has traditionally represented the annotated objects in the training set as numeric vectors in a multidimensional feature space. It is also known that the performance of many classification algorithms is largely dependent on the size and the dimensionality of the training data. Hence, a question that arises is what the ideal size and dimension should be to obtain a good classification performance, considering that greater values generally yield to a better classification but at the cost of increasing the computational load and the risk of overfitting (Liu y Zheng, 2006).

To deal with this issue, the dimensionality of the training data has been commonly reduced by using distance-based representations such as pairwise distances or (dis)similarities (Lee et al., 2010). It is also a frequent practice to use different (dis)similarity measures, each acting on distinct subsets of the muldimensional available features, that are lately combined to produce a similarity score value. A number of combination techniques then exists, under the name of fusion schemes, that have been categorised either as early or late fusion (Zhang y Ye, 2009). While the first one uses a unified measure that merges all the features, the second one computes multiple feature measures on a separate basis and then combines them to obtain the similarity between two objects.

Inspired by the late fusion scheme, in this paper we use a multidistance representation that transforms the original feature space in a distance space resulting from the concatenation of several distance functions computed between pairs of objects. This kind of input data involves an additional knowledge injection to the classifier, because the use of a distance measure is an implicit match between the characteristics of two objects and also because of the usual correlation between semantic similarity and small values of distance. It is worth mentioning that this multidistance space is related to the dissimilarity space defined in Duin y Pekalska (2012). Nevertheless, it differs from it in that the space transformation is carried out at a feature level between freely selected pairs of objects instead of using a fixed representation set.

The aim of this paper is to compare the performance obtained from the feature-based and the multidistance-based representations when applied to

a classification similarity learning setting as well as to analyze the influence of different training data sizes. Thus, our goal is twofold: on the one hand, we want to study the ability of a classifier to deal with a high feature dimensionality when the training size grows; on the other hand, we want to test under which circumstances the reduction in dimensionality leads to better results than treating objects in their wholeness.

The proposed experimentation concerns the problem of Content-Based Image Retrieval (CBIR), where image contents are characterized by multidimensional vectors of visual features (e.g. shape, color or texture). By considering pairs of images labeled as similar or dissimilar as training instances, we face a binary classification problem that can be solved through a soft classifier that provides the probability of belonging to each class. This probability value can be considered as the score determining the degree of similarity between the images and it can be used for ranking purposes. In particular, the Support Vector Machine classification algorithm has been selected and we use different values for the Minkowski distance to construct two multidistance-based representations. Additionally, we use as baseline for our comparison the performances obtained from the global Euclidean distance and two other traditional score-based normalization methods: the standard Gaussian normalization and the Min-max normalization.

The rest of the paper is organized as follows: Section A.4 formulates the problem and describes the multidistance-based representations into detail; Section A.5 presents the experimental setting and analyzes the obtained results; finally, Section A.6 states the conclusions and discusses future work.

A.4. Problem Formulation

Let us assume we have a collection of images $X = \{x_i\}, i = 1, 2 \ldots$, which are conveniently represented in an m-dimensional feature space $\mathbb{F}$. Let us also assume that, when needed, this feature space can be defined as the Cartesian product of the vector spaces related to T different visual descriptors such as color, texture or shape (see Eq.(A.1)).

$$\mathbb{F} = \mathbb{F}^{(1)} \times \ldots \times \mathbb{F}^{(t)} \times \ldots \times \mathbb{F}^{(T)} \tag{A.1}$$

Hence, we can represent as $\mathbf{x}_i^{(t)}$ the set of features that correspond to descriptor t in $\mathbf{x}_i$. Let us finally consider a classical similarity learning setup (Xing et al., 2002; Globerson y Roweis, 2005), where k training pairs (x_i, x_j) are available that are accordingly labeled as similar (S) or dissimilar (D). In classification-based learning, these pairs are used to train a classifier that can later be able to classify new sample pairs. Thus, when it comes to using a soft classifier, its output will provide a score that may be used to judge the similarity between objects.

A straigthforward approach that fits this scheme is to concatenate the feature vectors of the objects and use the resulting double-size vector as the input to the classifier (see the arrow labeled "Feature-based representation" in Figure A.1). However, by following this approach, the learning problem size highly depends on the dimensionality of the feature space $\mathbb{F}$, which is usually rather large. This situation might be specially critical for small sample datasets, which unfortunately are often the case. The dimensionality of the input data can then be reduced by using feature reduction techniques such as Principal or Independent Component Analysis. Another way of tackling this problem is by applying a similarity-based spatial transformation (Duin y Pekalska, 2012). In this paper we evaluate the performance of two distance-based representations, namely: Multidistance L_p and Multidistance L_1. Both representations result from a preprocessing layer that acts before passing the training data to an SVM (see the arrow labeled "Multidistance-based representation" in Figure A.1).

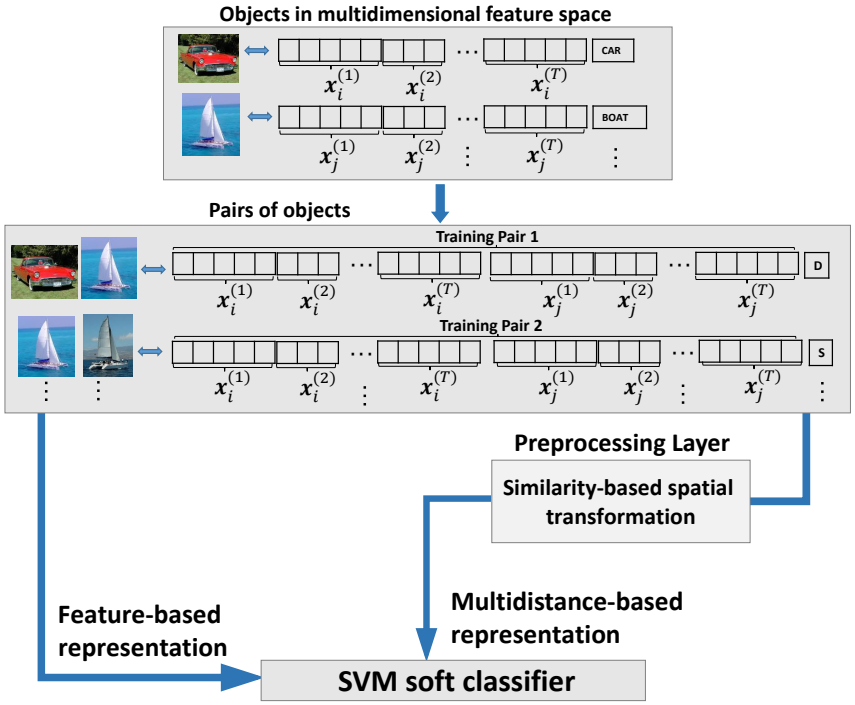


Figura A.1: Feature-based and multidistance-based classification similarity learning approaches.

On the one hand, the Multidistance L_p representation derives from computing a family of N distance functions (e.g. Euclidean, cosine or Mahalanobis) for every training pair. Each distance function is defined in each descriptor vector space as in Eq. (A.2).

$$d_n^{(t)} : \mathbb{F}^{(t)} \times \mathbb{F}^{(t)} \longrightarrow \mathbb{R} \tag{A.2}$$

Thus, we define a transformation function w, as indicated in Eq. (A.3), that constructs a tuple of values $\langle d_1^{(1)}, \ldots, d_N^{(1)}, d_1^{(2)}, \ldots, d_N^{(2)}, \ldots, d_1^{(T)}, \ldots, d_N^{(T)} \rangle$ from the feature-based representation of two images x_i and x_j, where $d_n^{(t)}$ denotes the distance between $\mathbf{x}_i^{(t)}$ and $\mathbf{x}_j^{(t)}$.

$$w : \mathbb{F} \times \mathbb{F} \longrightarrow \mathbb{R}^{N \cdot T} \tag{A.3}$$

The choice of the most suitable distance function depends on the task at hand and affects the performance of a retrieval system (Papa et al., 2009). This has led different authors to analyze the performance of several distance measures for specific tasks (Aggarwal et al., 2001; Howarth y Rüger, 2005). Therefore, rather than choosing the most appropriate distance for a task, the proposed multidistance representation aims to boost performance by combining several distance functions simultaneously. This operation transforms the original data into a labeled set of $N \cdot T$-tuples, where each element refers to a distance value, calculated on a particular subset of the features (i.e. the corresponding descriptor).

In order to increase the accuracy of the classification (Vert et al., 2004) by placing equal emphasis on each descriptor space (Ali y Smith-Miles, 2006), we normalize the labeled tuples $\langle \hat{d}_1^{(1)}, \ldots, \hat{d}_N^{(1)}, \hat{d}_1^{(2)}, \ldots, \hat{d}_N^{(2)}, \ldots, \hat{d}_1^{(T)}, \ldots, \hat{d}_N^{(T)} \rangle$ through a simple linear scaling operation into range $[0, 1]$. A complete schema of the input data transformation done by the preprocessing layer can be seen in Figure A.2.

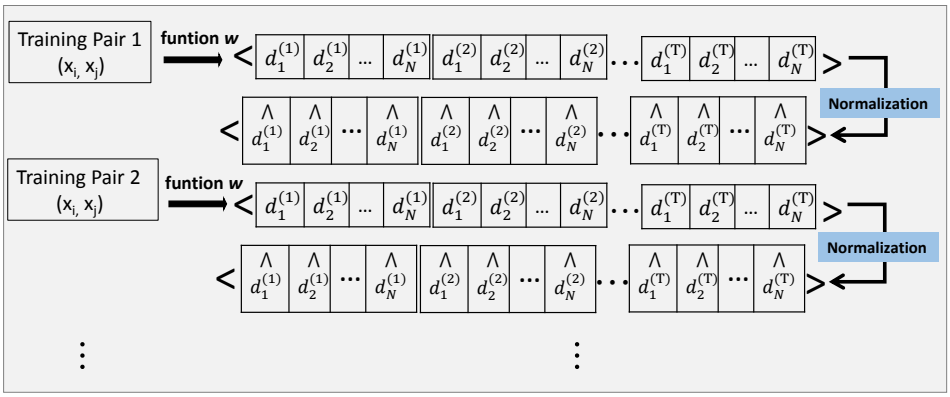


Figura A.2: Scheme of the preprocessing layer in the Multidistance L_p representation.

On the other hand, the Multidistance L_1 representation is constructed by

separately calculating the L_1 distance function (i.e. the Manhattan distance) for every feature of each pair of images in the training set. Hence, the distance function can be defined as in Eq. (A.4). Note that this representation is more simple and generic than the previous one, since it does not require the images to be described by a set of visual descriptors.

$$d : \mathbb{F} \times \mathbb{F} \longrightarrow \mathbb{R}^m \tag{A.4}$$

In turn, we define a transformation function $w^{'}$ as indicated in Eq. (A.5) that, given the feature-based representation of two images x_i and x_j, constructs a tuple of values $\langle d_1, \ldots, d_m \rangle$. Again, we normalize the labeled tuples $\langle \hat{d}_1, \ldots, \hat{d}_m \rangle$ by means of a simple linear scaling operation into range $[0, 1]$ as shown in Figure A.3.

$$w^{'} : \mathbb{F} \times \mathbb{F} \longrightarrow \mathbb{R}^m \tag{A.5}$$

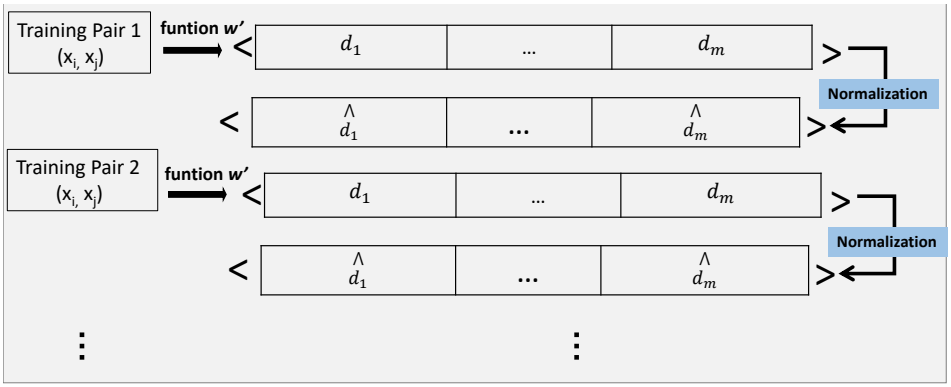


Figura A.3: Scheme of the preprocessing layer in the Multidistance L_1 representation.

Once the SVM soft classifier has been trained, it can be used to provide a score value that can be treated as a similarity estimation between images. For any new pair (x_i, x_j), function w or $w^{'}$ is applied to convert the original features into a tuple of distances (using the same distance functions as for training). After normalization, the resulting vector is used as input to the classifier, that provides a confidence estimation that the pair belongs to any of the classes. This estimate can be used directly for ranking purposes, or converted into a probability value by using the method in Platt (1999).

A.5. Evaluation

A.5.1. Databases

To analyze the performance of the SVM classifier for the feature-based and the multidistance-based training data formats, a number of experiments have been carried out using three different data sets, which are representative of a range of different situations and have also been used in other previous studies e.g., Arevalillo-Herráez y Ferri (2013); Arevalillo-Herráez et al. (2008b). All datasets follow a similar structure where each entry contains a feature vector and a label. Accordingly, each feature vector corresponds to the feature representation of an image through a set of visual descriptors. The label refers to the semantic concept that the image represents, according to a manual classification. The details about the databases are as follows:

- A database called "Small" containing 1508 pictures, some of which were extracted from the web and others were taken by the members of the research group. These have been manually classified as belonging to 29 different semantic concepts such as flowers, horses, paintings, skies, textures, ceramic tiles, buildings, clouds, trees, etc. This database and corresponding labels have also been used in de Ves et al. (2006); León et al. (2007), where further details can be found. In this case, the descriptors include a 10×3 HS color histogram and texture information in the form of two granulometric cumulative distribution functions (Soille, 2003).

- A subset of 5476 images named "Art" from the large commercial collection called "Art Explosion" composed of a total of 102894 royalty free photographs and distributed by the company Nova Development (`http://www.novadevelopment.com`). The images from the original repository, organized in 201 thematic folders, have been classified into 62 categories where images have been carefully selected so that the ones in the same category represent a similar semantic concept. The features which have been computed for these pictures are the same as those extracted for the "Small" database.

- A subset of the Corel database used in Giacinto y Roli (2004). This is composed of 30000 images which were manually classified into 71 categories. The four image descriptors used are a 9-dimensional vector with the mean, standard deviation and skewness for each hue, saturation and value in the HSV color space; a 16-dimensional vector with the co-occurrence in horizontal, vertical and the two diagonal directions; a 32-dimensional vector with the 4×2 color HS histograms for each of the resulting subimages after one horizontal and one vertical split; and a 32-dimensional vector with the HS histogram for the entire image.

A summary of the main characteristics of the three databases is given in Table A.1. The interested reader can find more details about them in `http://www.uv.es/arevalil/dbImages/` and `http://kdd.ics.uci.edu/databases/CorelFeatures`.

Tabla A.1: A summary of the three data sets used in the experiments.

Data set	Size	Categories	Total dimension	Descriptor sizes
Small	1 508	29	50	30,10,10
Art	5 476	63	104	30,12,7,3,10,10,4,11,11,6
Corel	30 000	71	89	9,16,32,32

A.5.2. Experimental setting

The classification similarity learning approaches have been also compared with three other traditional score methods, that have been used as baseline. The first one is the global Euclidean distance applied on the entire feature vectors. The second one is the standard Gaussian normalization as described in Iqbal y Aggarwal (2002), that consists of a mapping function $d_2^{(t)} \rightarrow (d_2^{(t)} - \mu)/3\sigma$, where μ and σ represent the mean and the standard deviation of the Euclidean distance on each descriptor vector space ($d_2^{(t)}$). The third one is the Min-max normalization, that performs a linear transformation on data computing the minimum and the maximum of the distance $d_2^{(t)}$. The last two approaches are both applied to the individual visual image descriptors and will be referred to as Gaussian normalization and Min-max normalization, respectively. The experiments were run 50 times each and the results were averaged.

To evaluate the influence of the training size, the experiments were run over six training sets with increasing sizes. The smallest training set had 250 pairs, the next one had 500 pairs, while the remaining training sets increased their size sequentially from 1000 up to 4000 with steps of 1000 pairs. In each training set the pairs were labeled as similar (S) when the labels associated with the vectors were the same, and as dissimilar (D) otherwise.

After the training phase, if any, the ranking performance of each algorithm was assessed on a second different and independent test set composed of 5000 pairs randomly selected from each repository. To this end, the Mean Average Precision (MAP), one commonly used evaluation measure in the context of information retrieval (Thomee y Lew, 2012), was used. The MAP value corresponds to a discrete computation of the area under the precision-recall curve. Thus, by calculating the mean average precision we had a single overall measure that provided a convenient trade-off between precision and recall along the whole ranking.

For the Multidistance L_p representation we have considered a pool com-

posed of four Minkowski distances, L_p norms, with values $p = 0.5, 1, 1.5, 2$. These are widely used dissimilarity measures that have shown relatively large differences in performance on the same data (Aggarwal et al., 2001; Howarth y Rüger, 2005), and hence suggest that may be combined to obtain improved results. Fractional values of p have been included because they have been reported to provide more meaningful results for high dimensional data, both from the theoretical and empirical perspective Aggarwal et al. (2001), a result that has also been confirmed in a CBIR context (Howarth y Rüger, 2005). Note also that the Multidistance L_1 representation just uses the Minkowski distance with parameter value $p = 1$.

In addition, the kernel chosen for the SVM has been a Gaussian Radial Basis Function (RBF) . The parameters γ and C have been tuned by using an exhaustive grid search on a held out validation set composed of a 30 % partition of the training data ($C \in \{10^{-6}, 10^{-5}, ..., 10^{0}, 10^{1}\}$ and $\gamma \in \{10^{-2}, 10^{-1}, ..., 10^{4}, 10^{5}\}$). To compensate the SVM sensitiveness to unbalanced data sets (Köknar-Tezel y Latecki, 2011), we also conducted an initial study to determine the most adequate proportion of similar/dissimilar pairs in the training set. This was done for each database and the best performance in the Small and Art databases is achieved when the percentage of similar pairs in the training set is around 30 %. In the Corel database this percentage raises to 50 % due to the bigger size of this database.

A.5.3. Results

In this section we compare the performance of the feature-based and multidistance-based representations presented in this paper to carry out classification similarity learning. Figures A.4, A.5 and A.6 plot the average MAP as a function of the training size for each database, also including the MAP values obtained for the baseline methods, where no learning is involved.

The plots reveal that, in general terms, the multidistance-based representations outperform the feature-based representation in the presence of small to medium-size training sets. On the other hand, when the number of training pairs is big enough, using the original image features leads to higher MAP values. The location of the cut-off point as well as other particularities depend on the characteristics of each database (i.e. size, categories and visual descriptors) as described below.

For the Small database, the Multidistance L_1 representation performs better than the feature-based representation when using less than 1000 training pairs (see Figure A.4). Though, it is worth noting that this representation is outperformed by all the baseline methods in the extreme case of not having enough information for training the classifier (i.e. less than about 300 pairs). The Multidistance L_p representation shows a limited learning capacity in this database, since the MAP values obtained before reaching the cut-off point with the feature-based representation (i.e. around 600 pairs)

are lower than those of the baseline methods. When we allow the training size to become large enough (i.e. beyond 1000 pairs), the classical feature-based representation improves the results that can be obtained from the rest of the algorithms. Under this circumstance, the information loss incurred by the multidistance-based representations is detrimental since it limits the learning capabilities of the SVM classifier.

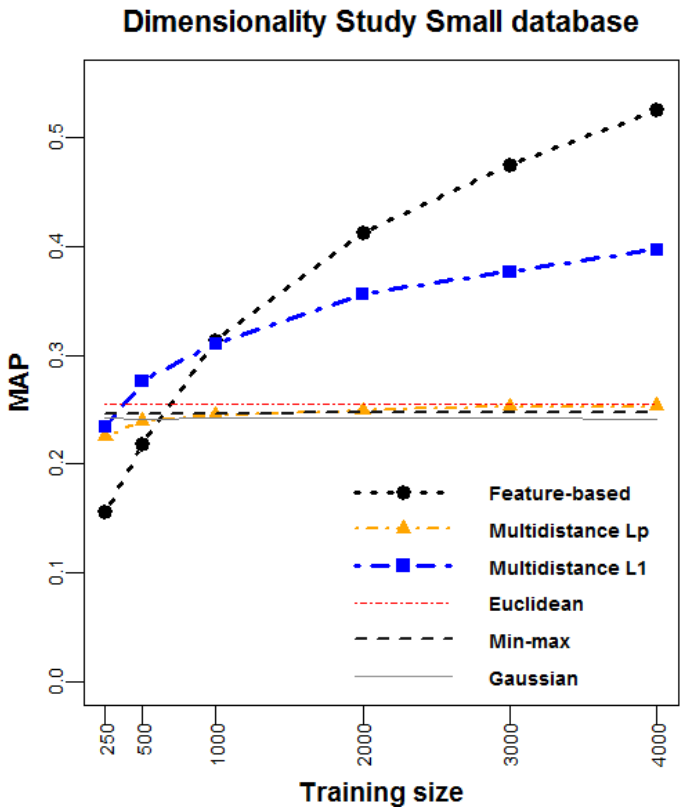


Figura A.4: Average MAP vs training set size for the Small database.

The Art database shows better results for the Multidistance L_p representation, which is the most interesting option for training sets composed of up to 500 pairs (see Figure A.5). The Multidistance L_1 representation produces the highest MAP values and outperfoms the feature-based representation in the interval ranging from 500 to, approximately, 2400 training pairs. Then again, the feature-based representation casts the best results when dealing with training sets of higher dimensionality. Overall, in this dataset, all the proposed classification similarity learning techniques perform better than the baseline methods.

Finally, the results for the Corel database show a similar ordering when focusing on the feature-based and the multidistance-based representations

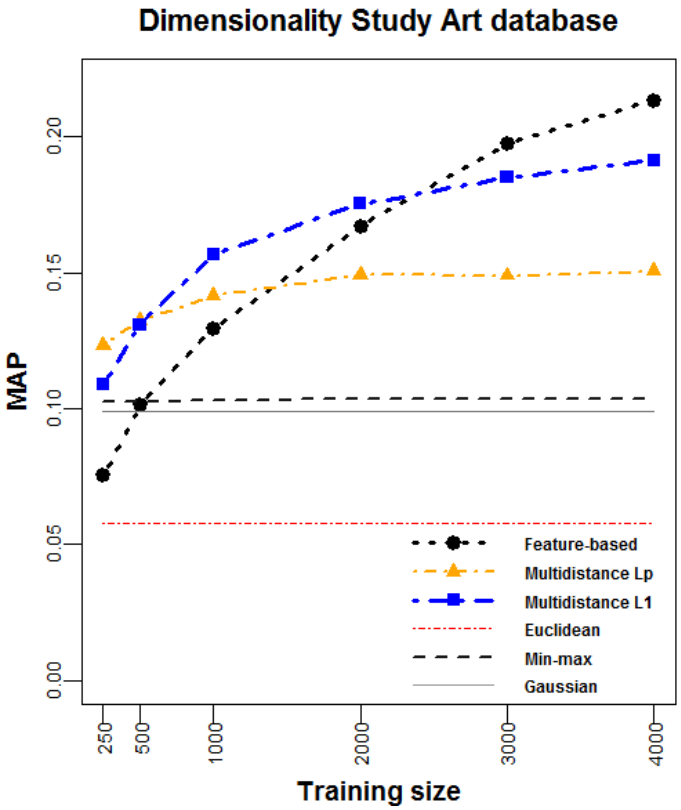


Figura A.5: Average MAP vs training set size for the Art database.

(see Figure A.6). That is, the Multidistance L_p representation marginally outperforms the Multidistance L_1 representation for training sizes smaller than 2000 pairs. Then, the Multidistance L_1 representation works better than the feature-based representation while not reaching the cuf-off point that, in this case, falls out of the plot. However, if we rightly consider the high MAP values obtained from the baseline methods, we see how difficult it is to learn a classification model in this database without having a big amount of training pairs. Apart from the bigger size of this database, the main reason behind such a behaviour is the higher subjectivity of its classification, where different criteria have been applied (also by different people) that results in a classification that considers similar concepts under different labels. Just as an example, the Corel database includes Insects I and Insects II as two different categories and, as a result, our experimental setting considers images in these two groups as dissimilar even though they have very similar visual descriptors.

In spite of this singularity, all in all, the results obtained demonstrate that the training size has an important effect on the classification performance

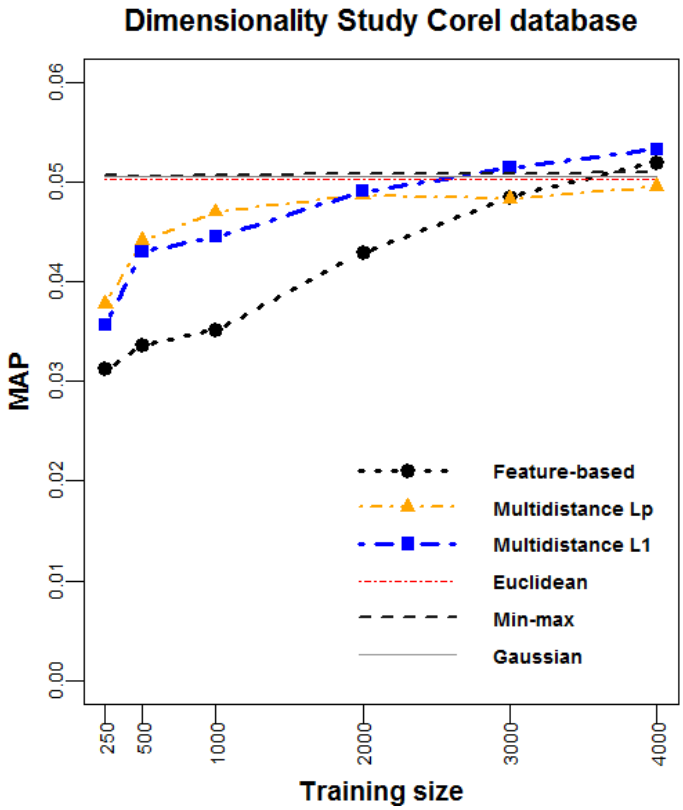


Figura A.6: Average MAP vs training set size for the Corel database.

when we adopt differents strategies for the representation of input data.

A.6. Conclusion

In this paper we have conducted an experimental study comparing two approaches for learning similarity scores in a multidimensional feature space using a classification-based method as the SVM. The difference between these approaches is based on the representation format followed by the sample dataset that is used to train the classifier. On the one hand, a feature-based representation of objects can have as drawback the high dimensionality of the learning problem that it poses to the classifier. On the other hand, a multidistance-based representation can reduce the dimensionality by transforming the original multidimensional space in a distance space constructed as the concatenation of distance functions.

A series of performance patterns have been extracted from the analysis of the different input data formats and the training size. We found that a low dimensional multidistance-based representation can be convenient for small

to medium-size training sets whereas it is detrimental as the training size grows. The dimensionality reduction (e.g. in the form of distances relations and its combitation) supposes additional information to the classifier and boosts its performance. For large training sets, though, a higher dimensional feature-based representation provides better results for the data base considered. This results can be of value when designing future systems that need to automatically capture the similarity of pairs of objects.

Future work will extend this study by including further investigation considering more distance combinations as well as other suitable techniques to reduce the dimensionality of the training set and to finally improve the performance of classifiers.

Apéndice B

Learning Similarity Scores by using a family of distance functions in multiple feature spaces

Autores:
Emilia López-Iñesta, Francisco Grimaldo, Miguel Arevalillo-Herráez

Revista:
International Journal of Pattern Recognition and Artificial Intelligence

Año 2017, volumen 31, número 8

ISSN: 0218-0014

DOI
http://www.worldscientific.com/doi/abs/10.1142/S0218001417500276

B.1. Notation

Notation	Description
d	Feature space dimensionality
d_M	Mahalanobis distance
$X = \{x_i\}$	Objects collection
$\mathbb{F}$	Feature space
$\mathbb{F}^{(t)}$	Feature space related to descriptor t
x_i	Generic object
$\mathbf{x}_i$	Feature vector of a generic object x_i
x_i'	Transpose vector x_i
M'	Transpose matrix M
W	Linear transformation matrix
S	Similar pairs of objects
D	Dissimilar pairs of objects
d	Distance
$\hat{d}$	Normalized distance
k	Index for pairs of objects $k = 1, ..., K$
p	Minkowski distance parameter
D_n	Familiy of distances, $n = 1, ..., N$
N	Family of distances size
T	Descriptors size
L_p	Minkowski distances
w	Transformation function from original features to a tuple of $N \cdot T$ distances
$d_n^{(t)}$	Distance between two descriptors objects $\mathbf{x}_i^{(t)}$ and $\mathbf{x}_j^{(t)}$
$\mathbf{d_{i,j}}$	tuple of $N \cdot T$ distances
w'	Transformation function from original features to a tuple of distances
s	Degree of similarity between an object pair
γ	Parameter Gaussian Radial Basis Function (RBF)
C	Parameter Gaussian Radial Basis Function (RBF)
V	Cut-off value
p_{adj}	Adjusted p-values
ITML D_n	ITML method used on features of the n-th descriptor

Learning Similarity Scores by using a family of distance functions in multiple feature spaces

B.2. Abstract

There exists a large number of distance functions that allow one to measure similarity between feature vectors and thus can be used for ranking purposes. When multiple representations of the same object are available, distances in each representation space may be combined to produce a single similarity score. In this paper, we present a method to build such a similarity ranking out of a family of distance functions. Unlike other approaches, that aim to select the best distance function for a particular context, we use several distances and combine them in a convenient way. To this end, we adopt a classical similarity learning approach, and face the problem as a standard supervised machine learning task. As in most similarity learning settings, the training data is composed of a set of pairs of objects that have been labeled as similar/dissimilar. These are first used as an input to a transformation function, that computes new feature vectors for each pair by using a family of distance functions in each of the available representation spaces. Then, this information is used to learn a classifier. The approach has been tested using three different repositories. Results show that the proposed method outperforms other alternative approaches in high dimensional spaces, and highlight the benefits of using multiple distances in each representation space.

Keywords: Classification similarity learning, Metric Learning, distance combination, similarity ranking, multiple features

B.3. Introduction

The accuracy of many classification algorithms is critically dependent on a distance defined over the input space. The Euclidean and other standard distances are rarely optimal for the problem being tackled (McFee y Lanckriet, 2010), but they are mathematically convenient and generally yield reasonable results in the absence of training data.

When training data is available, this information can be used to build a more convenient similarity score that takes the nature of the data into consideration. In metric learning, attention is placed on learning a distance that satisfies the properties of a metric (Bellet et al., 2015; Xing et al., 2002; Globerson y Roweis, 2005; Davis et al., 2007a; Weinberger y Saul, 2009) . These methods are specially useful when the measure is to be used within

popular algorithms that assume the metric properties on the underlying distance functions, such as k-means clustering or nearest neighbor classification. In classification similarity learning, though, the training information is used to learn a classifier that is able to decide whether a new pair of objects is similar or not (Domeniconi et al., 2005; Weinberger y Saul, 2009). These methods yield a score, rather than a metric, that is specially useful when the objective is to rank elements according to their similarity. In any case, it is a common approach to use training data composed of a set of pairs of objects that are known as similar/dissimilar (Bellet et al., 2015).

Assuming a dimensionality d of the data, metric learning methods work by computing a $d \times d$ positive definite transformation matrix that rotates and scales the feature space. This matrix takes into account both correlation between features and the relative importance of each feature, and it is used to linearly project the data into a new space where the similarity constraints imposed by the training data are satisfied better. Overall, these methods work particularly well and show an excellent generalization performance with low dimensional data, but do not scale well to high dimensional problems. The computation of the $d \times d$ matrix can be considered as a statistical inference procedure over $O(d^2)$ parameters, and learning under such conditions in high dimensional settings is prone to overfitting and requires a very large amount of training data (Davis y Dhillon, 2008). One option to deal with this problem is to use the training data to learn a set of weights, rather than a transformation matrix. This makes the number of parameters grow linearly with the dimensionality, and increases the applicability of the methods when a relatively small set of training data is available. The use of different type of regression methods over the label has been common in this context (Ksantini et al., 2007, 2008; Caenen y Pauwels, 2002).

When multiple representations of the same object are available, it is also common to reduce the dimensionality by applying late fusion methods (Zhang y Ye, 2009). This is a frequent setting in e.g. object detection, face recognition, audio classification or Content Based Image Retrieval (CBIR), where objects are usually represented by several descriptors that relate to different characteristic e.g. the color histogram or texture-related descriptors in the case of images and the Mel-Frequency Cepstral Coefficients (MFCCs) or low-level signal parameters in the case of audio signal. In this situation, it is possible to compute a concrete distance in each descriptor space, and rely on distance values to compute a combined score that (hopefully) represents better the similarity between objects. The choice of the most suitable distance function depends on the context, and may significantly affect the performance of a retrieval system (Papa et al., 2009). This has led different authors to analyze the performance of several distance measures for specific tasks e.g. Aggarwal et al. (2001) or Howarth y Rüger (2005). However, and independently from the distance function used, there is an implicit and una-

voidable loss of information when the original features are summarized with a distance value. Depending on the particular application, this loss may be compensated by a dimensionality reduction that makes the problem more tractable and approachable from a classification perspective.

In this paper, we aim at providing a more informative summarization of the original features, at the cost of a lower reduction of the dimensionality. In essence, the classification framework is identical to the one used by typical late fusion methods based on classification, but we attempt to boost performance by considering several distance functions simultaneously rather than choosing the most appropriate distance for the task. In particular, a transformation function that operates in the original feature space has been designed. This transformation is based on a family of distance functions, which are applied in each descriptor space to yield a new feature vector for each pair of objects provided as training data. The resulting feature vectors are then used to train a soft classifier, which is finally used to predict similarity between any unseen pair of objects. Our proposal is evaluated and compared to existing representative methods in both the metric learning and regression literature, showing a significantly better performance. The reasons for the improvement are also studied in detail. In particular, it is shown that the use of multiple distances in each representation space helps recover part of the original information that is lost when replacing the original features by a distance value, boosting retrieval results.

The remainder of the paper is organized as follows: Section B.4 deals with the related work; section B.5 formulates the problem and describes the proposed supervised learning technique; section B.6 presents the databases used in this work, the experimental setting and the results obtained that support our main conclusions, which are outlined in the last section (B.7) along with some proposals for future work.

B.4. Related work

The notion of a distance underpins the functioning of standard methods in machine learning, pattern recognition and data mining e.g. nearest neighbor, k-means. Standard functions are widely used because of their simplicity e.g. Euclidean distance, Jaccard Similarity, Cosine Similarity. However, their performance depend on the particularities of the problem at hand, and do not benefit from additional information about the context that may potentially become available.

Metric learning approaches do make use of this information, that is used to feed the methods with a set of constraints over the distance values. These methods try to learn a generalized Mahalanobis distance metric $d_M = d_M(x_i, x_j) = (\mathbf{x}_i - \mathbf{x}_j)'M(\mathbf{x}_i - \mathbf{x}_j)$ parametrized by a Positive Semidefinite (PSD) matrix M that can be decomposed as $M = W'W$. This is

equivalent to computing a linear projection of the data into a new space and then using the Euclidean distance to compare the sample data.

In general, metric learning can be formulated as an optimization problem. The different methods and algorithms proposed in the literature are characterised by its different objective functions and constraints, that involve either optimization in the space of PSD matrices, or learning W (the projection matrix).

Xing et al. (2002) presented a first seminal work in this field, formulating Metric Learning as a convex optimization problem given a supervised data framework. The goal was to maximize the sum of distances between all pairs of dissimilar instances and minimize the distances for all similar pairs by using Semidefinite Programming with a projected gradient descent algorithm. The learned metric was used to improve the performance of the k-Nearest Neighbors algorithm (k-NN).

Another algorithm, Relevant Component Analysis (RCA) (Bar-Hillel et al., 2003), tries to find a global linear projection from the original feature space to a lower dimensional space assigning large weights to relevant features and low weights to the irrelevant ones. To estimate the relevant features, the authors introduced the idea of *chunklets*, that is, small subsets of points that are known to belong to the same but maybe unknown class. These chunklets are obtained from equivalence relations by a transitive closure. An important detail is that RCA only uses positive equivalence constraints. The reason is that, although negative equivalence constraints contain useful information, they are less informative than positive ones.

In Information Theoretic Metric Learning (ITML) (Davis et al., 2007a), an information-theoretic measure is used and the authors translate the problem of learning an optimal distance metric to that of learning the optimal Gaussian with respect to an entropic objective. ITML considers simple distance constraints enforcing that similar instances have a distance lower than a given upper bound and dissimilar instances be further than a specific lower bound. The optimization method computes Bregman projections and no Semidefinite Programming is required.

Other distinguished methods for metric learning are Learning from Relative Comparisons (Schultz y Joachims, 2003), Large Margin Nearest Neighbour (LMNN) (Weinberger y Saul, 2009) or Logistic discriminant Metric Learning (LDML) (Guillaumin et al., 2009). One major difficulty of these methods relates to the difficulty associated with learning and using the transformation matrix in high dimensional settings. To overcome this burden, other structured methods that search for a lower number of parameters have recently been proposed (Davis y Dhillon, 2008).

When the available features come from different descriptors, another way to face the dimensionality problem is by using late fusion methods (Dong et al., 2014). In this case, similarity scores are independently computed in each

representation space, and fused into a single value at a later stage, generally using standard classification methods. This is alternative to concatenating all features coming from the various representation spaces into a larger vector.

The most simple late fusion methods are the CombSum and CombProd approaches (McDonald y Smeaton, 2005). In these methods, the combined similarity score is computed as the sum/product of the distance values, respectively. The CombSum approach is equivalent to a linear combination assuming that all visual descriptors have the same importance, but it can be conceptually extended to a weighted linear combination, to give a different relevance to each descriptor. In this direction, some works have focused on the use of different techniques to compute fixed weights or coefficients for such a linear combination. For example, weights are determined by using a Genetic Algorithm in da Silva Torres et al. (2005); and by using ranks in Giacinto y Roli (2004).

Other probabilistic approaches e.g. Ksantini et al. (2007, 2008), have boosted the retrieval performance of CBIR systems by using regression methods on training data to compute the weights of a pseudo-metric and obtain the relative relevance of the feature vectors. Several other proposals in the literature also allow for a non linear combination of scores (da Silva Torres et al., 2009) to merge the individual information coming from each image feature. An illustrative case is the multi-objective optimization technique used in Zhang y Izquierdo (2006) to define a global measure as an optimal linear combination of partial similarity functions. In a different line of work, a series of combination strategies attempt to pre-normalize the data by appropriately scaling the values obtained in each descriptor space before the fusion.

A non-linear scaling is presented in Iqbal y Aggarwal (2002), where color, texture and structure distances are pre-processed using a technique based on Gaussian normalization, and a global measure is defined as a weighted linear combination of the normalized distances.

A more advanced non-linear normalization that uses training information has been presented in Arevalillo-Herráez et al. (2008a), this time considering a probabilistic framework to map the similarity value in each descriptor space to the probability that the elements being compared be similar.

The use of Kernel methods for assessing similarity and combining multiple similarity measures has also been common in the literature (Vert et al., 2004). In particular, Support Vector Machines (SVM) have demonstrated their capacities in pattern recognition (Vapnik, 1995) and have been used for classification in certain tasks, such as Multimedia Semantic Indexing (Ayache et al., 2007). In Zhang y Ye (2009) the combination of multiple distances was formulated as a classification problem and solved by an SVM. Likewise, authors in Dimitrovski et al. (2010) applied a linear combination to fuse the different image features in a feature aggregation scheme and used an

SVM-based method to improve rankings. SVMs have also been successfully used in other fields to combine multiple similarity metrics. An application example can be found in the MARLIN system (Bilenko y Mooney, 2003), which learns a function that improves linkage accuracy and identification of duplicates among records in a database. An extension of these approaches is to combine a set of kernel functions by following some fusion rule for similarity measuring (Lee et al., 2007). Each different kernel captures a different notion of similarity and their combination allows the merging of information from different spaces into a unified similarity space (McFee y Lanckriet, 2011).

The suitability of early or late fusion methods can only be decided within a specific context. Late fusion approaches inherently reduce the dimensionality of the original data, by summarizing the original information into a series of scores that are used as inputs to the classification methods. From a performance perspective, this has a double effect. On the one hand, the dimensionality reduction benefits the applicability of classification methods, by reducing overfitting and increasing generalization. On the other hand, the amount of information in the input is also reduced. Hence, and to keep this reduction to a minimum, it is critical to choose the most adequate distance in each subspace. This is because a concrete distance may be more accurate than others at capturing a particular aspect of the high level semantic concept of similarity. Based on this same reasoning, the joint consideration of several distances in each subspace may provide the method with a more accurate representation of the similarity concept being searched, providing complementary information about different aspects of this concept. This issue is deeply explored in the work presented in this paper.

B.5. System description

Let us assume that we have a collection of objects $X = \{x_i\}, i = 1, 2 \ldots$, which are conveniently represented in a multidimensional feature space $\mathbb{F}$. Let us also assume that this feature space is defined as the Cartesian product of the vector spaces related to t different descriptors e.g. color, texture or shape in the case of images ($\mathbb{F} = \mathbb{F}^{(1)} \times \ldots \times \mathbb{F}^{(t)} \times \ldots \times \mathbb{F}^{(T)}$), and let us represent as $\mathbf{x}_i$ the entire feature vector for the object x_i and as $\mathbf{x}_i^{(t)}$ the part of the feature vector that correspond to descriptor t in $\mathbf{x_i}$.

Let us also consider training information composed of k pairs of objects (x_i, x_j), conveniently labeled as similar (S) or dissimilar (D). This formulation has been adopted in many other works (Xing et al., 2002; Globerson y Roweis, 2005), and it is a representative setting for a typical two-class classification problem.

Under this setting, let us consider a set of N distance functions $D_1 \ldots D_N$ (such as Euclidean, cosine or Mahalanobis) and a transformation function

$w : \mathbb{F} \times \mathbb{F} \longrightarrow \mathbb{R}^{N \cdot T}$ that, given the feature-based representation of two objects x_i and x_j, computes a tuple of $N \cdot T$ values $\mathbf{d_{i,j}} = w(\mathbf{x}_i, \mathbf{x}_j) = \langle d_1^{(1)}, \ldots, d_N^{(1)}, d_1^{(2)}, \ldots, d_N^{(2)}, \ldots, d_1^{(T)}, \ldots, d_N^{(T)} \rangle$, where $d_n^{(t)}$ denotes the distance between $\mathbf{x}_i^{(t)}$ and $\mathbf{x}_j^{(t)}$, according to D_n (see Figure B.1).

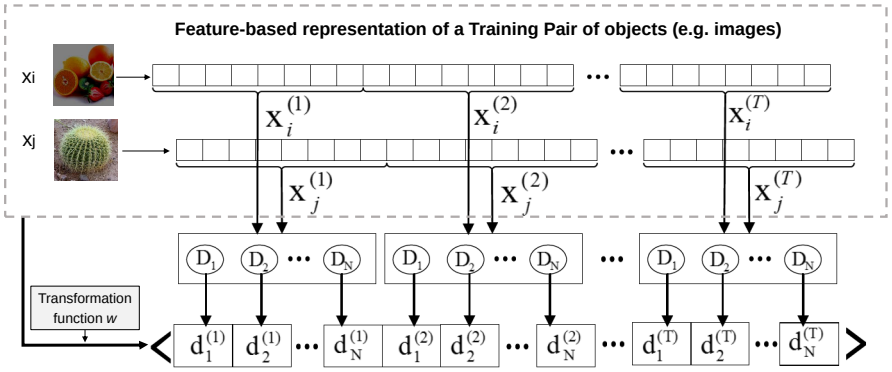


Figura B.1: Illustrative scheme of the transformation function w

Our objective is to learn a new function $s : \mathbb{R}^{N \cdot T} \longrightarrow [0, 1]$ that maps the $N \cdot T$-dimensional vector of distances into a single score value in the bounded interval $[0, 1]$, which represents the degree of similarity between an object pair according to an initially unknown and subjective criteria. This function shall allow to compute the similarity between any two unseen objects x_i and x_j as $s(\mathbf{d_{i,j}})$

To this end, the battery of distance functions $D_1 \ldots D_N$ is used on each training pair $(\mathbf{x_i}, \mathbf{x_j})$, to compute the value of $\mathbf{d_{i,j}} = w(\mathbf{x_i}, \mathbf{x_j})$. This operation transforms the original data into a labeled set of $N \cdot T$-tuples, where each element refers to a distance value, calculated on a particular subset of the features (descriptor). To increase the accuracy of the classification as (Vert et al., 2004), the tuples $\langle d_1^{(1)}, \ldots, d_N^{(1)}, d_1^{(2)}, \ldots, d_N^{(2)}, \ldots, d_1^{(T)}, \ldots, d_N^{(T)} \rangle$ are normalized using a simple linear scaling operation into the range $[0, 1]$, so that equal emphasis is placed on each descriptor space (Ali y Smith-Miles, 2006).

As a final step, a Support Vector Machine (SVM) soft classifier is trained by using the normalized tuples $\langle \hat{d}_1^{(1)}, \ldots, \hat{d}_N^{(1)}, \hat{d}_1^{(2)}, \ldots, \hat{d}_N^{(2)}, \ldots, \hat{d}_1^{(T)}, \ldots, \hat{d}_N^{(T)} \rangle$, along with the available labels for the pairs. Figure B.2 shows an illustrated explanation of the training phase.

Once the SVM soft classifier has been trained, it can be used to provide a score value that can be treated as a similarity estimation between objects. For any new pair $(\mathbf{x_i}, \mathbf{x_j})$, function w is applied to convert the original features into a the tuple of distances $\mathbf{d_{i,j}}$ (using the same battery of functions as for training). After normalization, the resulting vector is used as an input to

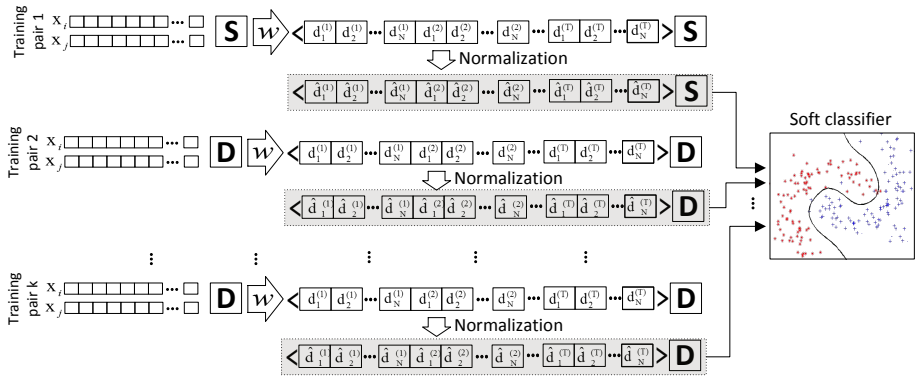


Figura B.2: Training process. The original feature vectors from the training pairs are transformed by using the function w, and normalized to range $[0, 1]$. The resulting tuples are used to train an SVM soft classifier.

the classifier. This yields a convenient confidence estimation that the pair belongs to any of the classes. This estimate can be used directly for ranking purposes, or converted into a probability value by using the method in Platt (1999).

B.6. Evaluation

B.6.1. Databases

To assess the relative merits of the proposed method with regard to other alternative techniques, a number of experiments have been carried out. These have been run on three different data sets of images, which have also been used in a large number of previous studies e.g., Arevalillo-Herráez y Ferri (2013) or van Gemert et al. (2006) and are representative of a range of different situations. All datasets follow a similar structure and each entry contains a hand-crafted feature vector and a label. Each feature vector corresponds to the feature representation of an image through a set of visual descriptors. The label refers to the semantic concept that the image represents, according to a manual classification. These databases are:

- A small database composed of 1508 pictures (**Small**), some of which were extracted from the web and others were taken by the members of the research group. These have been manually classified as belonging to 29 different semantic concepts such as flowers, horses, paintings, skies, textures, ceramic tiles, buildings, clouds, trees, etc. The number of images in each of these categories varies from 24 to 300. This database and corresponding labels have also been used in de Ves et al. (2006); León et al. (2007), where further details can be found. The 10

x 3 HS color histogram and nine texture features have been computed for this repository, namely Gabor Convolution Energies (Smith y Burns, 1997), Gray Level Co-occurrence Matrix (Conners et al., 1984), Gaussian Random Markov Fields (Chellappa y Chatterjee, 1985), two versions of the Spatial Size distribution (Ayala y Domingo, 2001) (one using a horizontal segment and another with a vertical segment) and the coefficients of fitting the granulometry distribution that, in turn, include features representing statistical measures, values of the granulometry size distribution function F for even values of pixels in [0, 20], density values in the same points and the B-splines coefficients that approximate the function F (Chen y Dougherty, 1994).

- A subset of 5476 images, extracted from a larger commercial collection called "Art Explosion" composed of a total of 102894 royalty free photographs and distributed by the company Nova Development (`http://www.novadevelopment.com`) (**Art**). The images from the original repository, organized in 201 thematic folders, have been classified into 63 categories where images have been carefully selected so that the ones in the same category represent a similar semantic concept. The features which have been computed for these pictures are the same as those extracted for the "Small" database.

- A subset of the Corel photography collection (**Corel Small**). This is a large commercial database frequently used in interactive retrieval (Calumby et al., 2016), especially in the CBIR scientific community consisting of 800 photo CDs, each one with approximately 100 images belonging to a certain category (Müller et al., 2002). Unfortunately, images are copyrighted and are not free of charge. To be able to use this set for the evaluation of CBIR systems, authors have created their own databases by manually classifying a subset of the images contained in the collection. In this work we have used the one in the UCI KDD repository (`http://kdd.ics.uci.edu`), that has also been used in Giacinto y Roli (2005). This is composed of 19510 images grouped into 43 categories. The four image descriptors used are: a 9-dimensional vector with the mean, standard deviation and skewness for each hue, saturation and value in the HSV color space; a 16-dimensional vector with the co-occurrence in horizontal, vertical and the two diagonal directions; a 32-dimensional vector with the 4×2 color HS histograms for each of the resulting subimages after one horizontal and one vertical split; and a 32-dimensional vector with the HS histogram for the entire image.

A summary of the main characteristics of the three databases is given in Table B.1. More details about the *Corel Small* subset and the *Small*

and *Art* databases can be found in `http://kdd.ics.uci.edu/databases/CorelFeatures` and `http://www.uv.es/arevalil/dbImages/`, respectively.

Tabla B.1: A summary of the three data sets used in the experiments.

Data set	Size	Categories	Total dimension	Descriptor sizes
Small	1 508	29	104	30,12,7,3,10,10,4,11,11,6
Art	5 476	63	104	30,12,7,3,10,10,4,11,11,6
Corel Small	19 510	43	89	9,16,32,32

B.6.2. Experimental setting

Apart from the proposed approach, six other methods have been considered in the comparison, namely:

- Information Theoretic Metric Learning (**ITML**) Davis et al. (2007a) and Relevant Component Analysis (**RCA**) (Bar-Hillel et al., 2003). These methods are representative of the metric learning paradigm. They use the full vector of features and do not group by descriptor.

- The Bayes Logistic Regression Model presented in Ksantini et al. (2007, 2008), in representation of weight-based methods that group by descriptor and search for the most suitable set of weights for each distance. Two versions of this method are considered. The first one (**Bayes-Logit**) uses the same input as our method. In the second (**Bayes-Logit-Euclidean**), only the Euclidean distances in each subspace have been computed and provided as input to the regression. This allows to discern between the relative merits associated with using the enriched input with multiple distances and the ones related to replacing the regression by a classification mechanism.

- The application of the Euclidean distance on the entire feature vectors **Euclidean**, and the **CombSum** and **CombProd** methods in McDonald y Smeaton (2005), which have been adopted as baselines.

All experiments presented have followed the same scheme, except for the **RCA** method that requires a different input. In this case, 1700 chunklets of three elements have been provided, considering that a chunklet holds as much information as three pairs (Bar-Hillel et al., 2005). In the rest of the cases, a set composed of 5000 pairs of images was used as training information. These pairs were labeled as similar (S) when the labels associated with the images were the same, and as dissimilar (D) when this was not the case. After training, the ranking performance of each algorithm was assessed on a second different and independent set of 5000 pairs, randomly selected from

the corresponding repository. To yield a reliable comparison, all experiments have been repeated 100 times, and results have been averaged. Performance has been compared by using recall and mean average precision (MAP). These are two commonly used evaluation measures in the context of information retrieval (Thomee y Lew, 2012).

Recall represents a measure of the ability of a system to retrieve relevant pairs. It is usually measured at a cut-off value V, and hence evaluates the performance of the method on the first V elements of the ranking. In our context, $recall(V)$ represents the proportion of the similar image pairs in the ranking that can be found in between the top V pairs. This is computed as the number of similar pairs in the first V elements of the ranking, divided by the total number of similar pairs in the entire test set. Average values across the 100 repetitions have been considered. Average precision (AP) is a global measure that evaluates the entire ranking as a whole, and corresponds to a discrete computation of the area under the precision-recall curve. This is defined as the average value of precision at all different recall values, understanding $precision(V)$ as the proportion of similar pairs, measured on the first V elements in the ranking. AP is computed as

$$AP = \frac{\sum_{i=1}^{r\_length} precision(i) \cdot similar(i)}{n\_similar}$$

where $similar(V)$ is a function that equals 1 when the image pair at position V of the ranking is similar and 0 otherwise; $n\_similar$ represents the number of similar pairs in the ranking; and r_length represents the total length of the ranking (5000). The value of MAP corresponds to the mean value of the average precision across all 100 repetitions of the experiment. The combination of recall and MAP allows for a comprehensive performance comparison and visualization. On the one hand, a plot of the recall values for all possible values of V (1 to 5000) provides an appropriate and simple visual comparison of the ranking performance at all possible cut-off values. On the other hand, the mean average precision is a single overall measure that provides a convenient trade-off between precision and recall for all possible cut-off values. In both cases, the higher the value, the better the performance.

For the method proposed, we have considered a pool composed of four Minkowski distances (L_p norms), for values $p = 0.5, 1, 1.5, 2$ and the Cosine and Mahalanobis distances. These are widely used dissimilarity measures that have shown relatively large differences in performance on the same data (Aggarwal et al., 2001; Howarth y Rüger, 2005), and hence suggest that may be combined to obtain improved results. Fractional values of p have been included because they have been reported to provide more meaningful results for high dimensional data, both from the theoretical and empirical perspective (Aggarwal et al., 2001; Howarth y Rüger, 2005). In turn, the

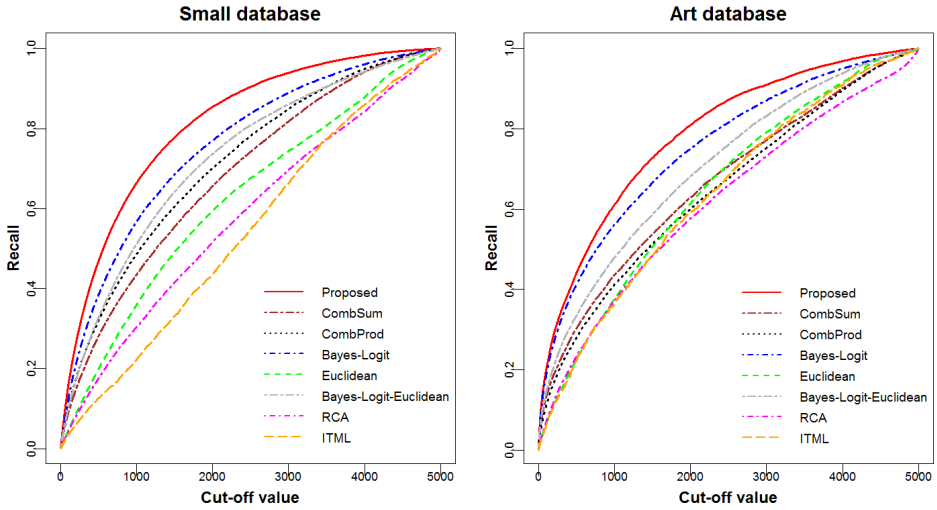
Cosine distance considers the features as vectors and focuses on the angle between them, whereas the Mahalanobis distance takes into account the covariance of data. In addition, the kernel chosen for the SVM has been a Gaussian radial basis function (RBF). The parameters γ and C have been tuned by using an exhaustive grid search on a held out validation set composed of a 30 % partition of the training data ($C \in \{10^{-6}, 10^{-5}, ..., 10^{0}, 10^{1}\}$ and $\gamma \in \{10^{-2}, 10^{-1}, ..., 10^{4}, 10^{5}\}$). We have analysed the performance of the proposal also in the linear case and using another non-linear SVM with a third degree polynomial kernel on all databases considered. The results using the linear and the RBF kernel were almost the same, slightly in favour of the RBF. However, the polynomial kernel obtained worse results than the linear and RBF kernels in all datasets.

To compensate the SVM sensitiveness to unbalanced data sets (Köknar-Tezel y Latecki, 2011), we also conducted an initial study to determine the most adequate proportion of similar/dissimilar pairs in the training set. This was done for each database and the best performance in the *Small* and *Art databases* is achieved when the percentage of similar pairs in the training set is around 30 %. In *Corel Small*, this percentage raises to 50 % due to the larger size of this database.

B.6.3. Results

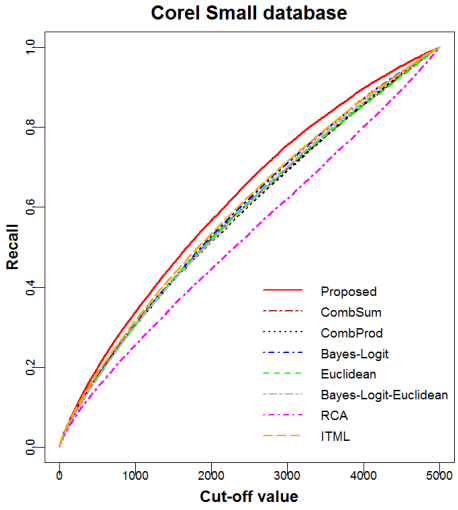
Experiment 1. As a first experiment, we have run a full comparison of the proposed methods. Recall plots for the proposed approach and all other methods in the comparison are presented in Figure B.3, for the three image repositories. It can be observed that the proposed method outperforms the others along the entire ranking in all databases, although the magnitude of the improvement depends on the database being considered. Improvements are larger in *Art* and *Small*, and less noticeable in *Corel Small*. Notice that results in this later repository are generally poor as reported by previous work (Müller et al., 2002), and the plot shows results that are close to the diagonal for all methods. This implies that the results are close to a random selection in all cases, and suggests that the selected features are not very informative for the task at hand. This database allows for a more exhaustive evaluation of the behavior of the methods, providing a specific sample of a more general classification case in which a number of features with a relatively low but positive correlation with the value of the label are available.

In order to further assess the behavior of the methods, evaluate their performance across the whole ranking, and analyze whether the differences are statistically significant, an analysis of the MAP scores has also been performed. To this end, the MAP score has been computed separately for each of the 100 repetitions of the experiment. With all these measures, a non-parametric Friedman test (Garcia y Herrera, 2008) has been run for each database. Results are reported in Table B.2. The resulting average ranks are



(a)

(b)



(c)

Figura B.3: Comparative performance between the proposed method and the other approaches in a) the Small database; b) the Art database; and c) the Corel Small database.

shown in the second column for each database, where the relative position of each method with respect to the rest has also been specified in brackets. Significant differences between the first positioned method (i.e. Proposed) and the remaining ones are identified by the adjusted p-values (see p_{adj} columns) computed by the post-hoc Holm test at a significance level $\alpha = 0.05$. These results confirm the analysis of the recall graphs in Figure B.3. The proposed method obtains the highest MAP values in all databases. In addition, the performance differences between the proposed method and the rest are statistically significant across all databases.

The performance of the methods is stable in *Small* and *Art*. The proposed method with the multidistance input format is consistently ranked first in all databases, and obtains the best performance both when features are informative for the task at hand (*Small* and *Art*) and when they are not (*Corel Small*). In *Art* and *Small*, methods *Bayes-logit* and *Bayes-Logit-Euclidean* occupy the second and third positions in the raking, but they do not perform well in the third repository, when less informative features are used. With regard to the metric learning methods, it can also be observed that they present a relatively poor performance. *RCA* consistently leads to one of the three worst results in all databases. Although *ITML* scores second in *Corel Small*, results in the repository are all very close and the method behaves the worst in the other two databases. Overall, these results lead to the conclusion that the use of a distance-based classification approach in the distance space is a source of improvement. But the combination of several distances also has a positive performance impact and yields further gains.

Experiment 2. To complement the results provided in the first experiment, we have also carried out a deeper analysis to evaluate the results obtained with the methods across the different classes in the dataset. For clarity reasons, table B.3 shows the results of the experiments for the three largest classes, in the three repositories and in terms of MAP. The proposed method performs first for the largest classes of all databases. Overall, the proposed method has scored first in 18 out of the 29 classes in *Small*; and second in 3 more cases. In *Art*, it has scored first in 27 out of the 63 classes; and second in 25 more classes. In *Corel Small*, it has scored first in 21 out of the 43 classes; and second in 5 more classes.

Experiment 3. At the light of the poor performance exhibited by the two metric learning methods in the *Small* and *Art* repositories, we have run a third experiment to analyze the potential causes, including the sensibility of the method to the dimensionality of the data. In particular, we have compared the performance of *ITML* as the dimensionality of the data grows. To this end, we have applied the method independently in each descriptor space, and also in the joint space generated by concatenating the features in

each subspace. Illustrative results have been plotted in Figure B.4, for the *Small* repository. Each curve included in the legend as *ITML Dn* represents the results of the method by using only the features of the n-th descriptor. The one with the legend *ITML* corresponds to the use of all features simultaneously. The results obtained with the proposed method have been included as a reference.

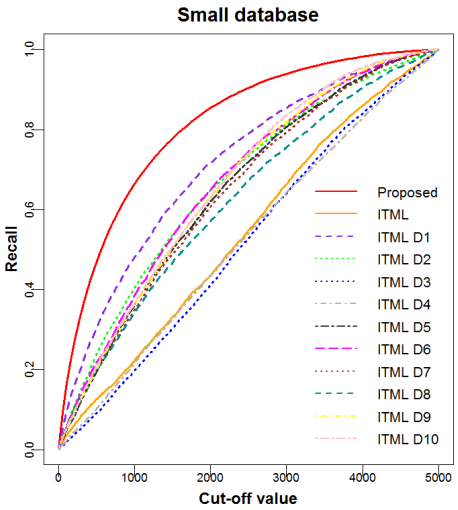


Figura B.4: Comparative performance of the ITML method, depending on the features used.

As it can be observed, the simultaneous use of all features does not provide an advantage with regard to using a single descriptor. In fact, it is just the opposite. While there exist features e.g. *ITML D3* that behave relatively well on their own, their result becomes worse and close to random (the diagonal) when they are all combined under an early fusion scheme. These results support the use of late fusion schemes such as the one proposed in the paper, to better capture the essence of the similarity concept being modeled.

Experiment 4. One can also doubt about whether the actual reasons behind the performance improvements reported in the first experiment are due to the distance-based classification approach, the combination of several distances or the use of a concrete distance that performs particularly well. To answer this question, we have further evaluated the approach by comparing the proposed method to six degenerated versions of it that use just a single distance measure. This simplified scheme is depicted in Figure B.5, where a single distance D per descriptor has been used. We refer to this simplified scheme as **single-distance** and we have tested it for each of the different distances used in this work, i.e. L_p norms, Cosine distance and Mahalanobis.

Figure B.6 and table B.4 offer a performance comparison between the

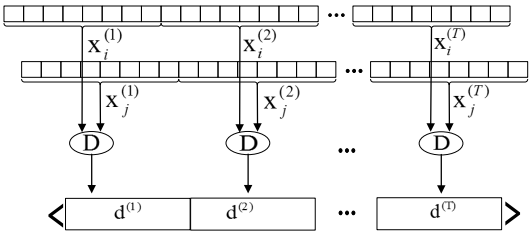


Figura B.5: Simplified scheme. The family of distances in Figure B.1 is replaced by a single distance D.

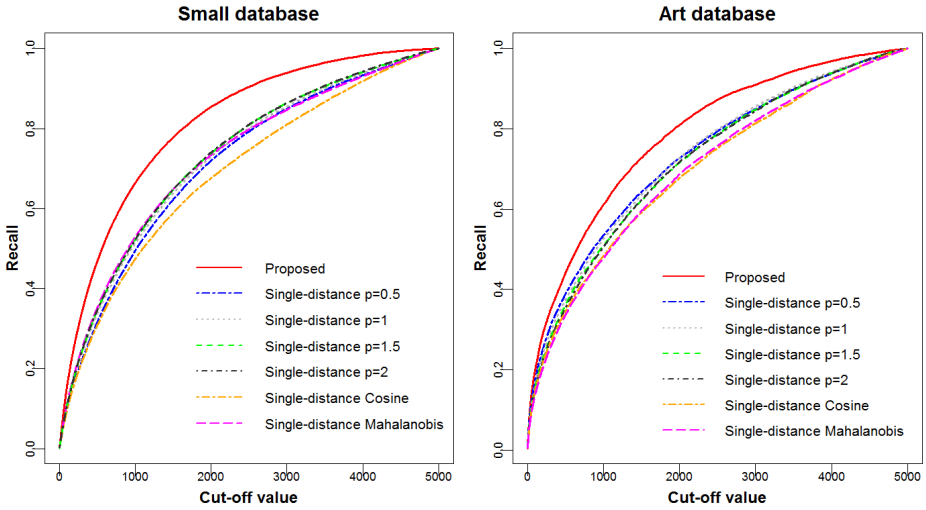
proposed method and the six single-distance versions. It can be observed that the performance of the different single-distance schemes depends on the database, and no particular distance (L_p norm, Cosine or Mahalanobis) consistently works better than the others. This can be clearly observed from the data in Table B.4 (see column Avg. Rank). For instance, the fractional norm $p = 0.5$ occupies the second position in the rank for the Art and Corel Small databases but it is at the fifth one in the Small database. However, the proposed method (using multiple distances) always scores first. In addition, all differences are found to be statistically significant.

These results suggest that gains are mainly due to using several distances as in input. This conclusion is also supported by the relative performance reported for the methods *Bayes-logit* and *Bayes-logit-Euclidean* in the first experiment (see Figure B.3 and Tables B.2 and B.3).

B.7. Conclusion

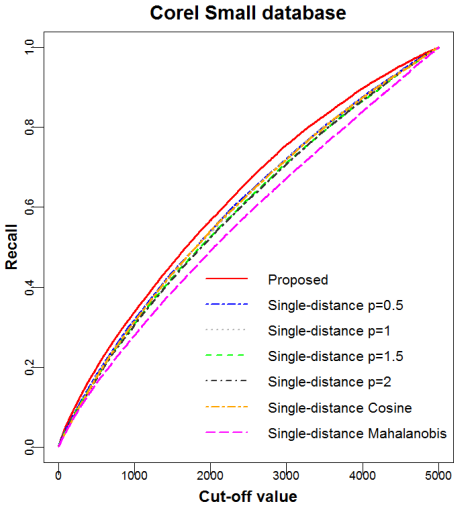
In this paper, we have presented a classification-based late fusion method for computing similarity scores. The method uses the same training information as classical metric learning methods, which is composed of pairs of objects labeled as similar/dissimilar; and it is based on a transformation function that converts the original features into a distance space by using a battery of pre-defined distance functions. The use of various distance function aims to recover part of the information which is lost when summarizing the original features by using a single distance, at the cost of a dimensionality increase that is proportional to the number of distances used. The complementary information provided by the different distance functions with regard to different aspects of the similarity concept being modeled has shown to compensate for the dimensionality increase, causing a positive effect on the retrieval results.

An exhaustive evaluation of the method, also in comparison with other state-of-the-art techniques, has led us to suggest the integration of several distance functions in late fusion classification methods. The benefits of this



(a)

(b)



(c)

Figura B.6: Comparative performance between single-distance and multi-distance input data formats in a) the Small database; b) the Art database; and c) the Corel Small database.

approach have been successfully tested in both the proposed method with an SVM classifier and by using the Bayes Logistic Regression Model presented in Ksantini et al. (2007, 2008).

Although the SVM and Logistic Regression have been used to develop a proof of concept, the method is by no means restricted to the use of this particular classifier. On the contrary, the framework presented is open to the use of alternative classification methods.

Future lines of work include a study of the most suitable combination of distance functions; the integration of learned metrics into the proposed framework; and an analysis of the properties of distances spaces, to further exploit the proposed transformation and explore the possibility to yield scores that satisfy the properties of a metric. This would allow to use the framework to produce scores that seamlessly integrate with well accepted classification methods, such as k-means clustering or nearest neighbor classification.

Tabla B.2: MAP, average ranks and adjusted p-values corresponding to a Holm post-hoc test when comparing each method to the proposed one, in all repositories.

Method	MAP	Avg. Rank	p_{adj}
Proposed	**0.379**	**1.00 (1)**	–
Bayes-logit	0.274	2.14 (2)	$< 10^{-4}$
RCA	0.114	6.84 (7)	$< 10^{-64}$
ITML	0.078	7.99 (8)	$< 10^{-90}$
CombSum	0.192	4.99 (5)	$< 10^{-31}$
CombProd	0.222	3.63 (4)	$< 10^{-14}$
Euclidean	0.128	5.99 (6)	$< 10^{-50}$
Bayes-logit-Euclidean	0.231	3.23 (3)	$< 10^{-10}$

(a) *Small* repository

Method	MAP	Avg. Rank	p_{adj}
Proposed	**0.192**	**1.14 (1)**	–
Bayes-logit	0.166	1.97 (2)	0.01
RCA	0.059	6.64 (6)	$< 10^{-57}$
ITML	0.055	7.10 (8)	$< 10^{-66}$
CombSum	0.110	3.92 (4)	$< 10^{-16}$
CombProd	0.084	5.16 (5)	$< 10^{-31}$
Euclidean	0.057	6.92 (7)	$< 10^{-62}$
Bayes-logit-Euclidean	0.128	3.13 (3)	$< 10^{-9}$

(b) *Art* repository

Method	MAP	Avg. Rank	p_{adj}
Proposed	**0.192**	**1.14 (1)**	–
Bayes-logit	0.166	1.97 (2)	0.01
RCA	0.059	6.64 (6)	$< 10^{-57}$
ITML	0.055	7.10 (8)	$< 10^{-66}$
CombSum	0.110	3.92 (4)	$< 10^{-16}$
CombProd	0.084	5.16 (5)	$< 10^{-31}$
Euclidean	0.057	6.92 (7)	$< 10^{-62}$
Bayes-logit-Euclidean	0.128	3.13 (3)	$< 10^{-9}$

(c) *Corel Small* repository

Tabla B.3: MAP results by method and by top-3 classes in all databases

Method	All classes	Class I	Class II	Class III
Proposed	**0.379**	**0.785**	**0.630**	**0.715**
Bayes-Logit	0.274	0.714	0.512	0.631
Bayes-Logit-Euclidean	0.231	0.686	0.336	0.629
RCA	0.114	0.467	0.210	0.493
ITML	0.078	0.314	0.407	0.310
Combsum	0.192	0.614	0.327	0.521
Combprod	0.222	0.683	0.317	0.568
Euclidean	0.128	0.508	0.217	0.495

(a) *Small* repository

Method	All classes	Class I	Class II	Class III
Proposed	**0.193**	**0.576**	**0.479**	**0.660**
Bayes-Logit	0.166	0.502	0.477	0.600
Bayes-Logit-Euclidean	0.128	0.357	0.474	0.594
RCA	0.059	0.315	0.401	0.610
ITML	0.055	0.366	0.403	0.633
Combsum	0.110	0.270	0.446	0.554
Combprod	0.084	0.256	0.443	0.631
Euclidean	0.057	0.435	0.383	0.531

(b) *Art* repository

Method	All classes	Class I	Class II	Class III
Proposed	**0.070**	**0.594**	**0.650**	**0.612**
Bayes-Logit	0.063	0.568	0.648	0.558
Bayes-Logit-Euclidean	0.063	0.562	0.614	0.524
RCA	0.056	0.497	0.624	0.471
ITML	0.068	0.462	0.455	0.396
Combsum	0.065	0.579	0.644	0.498
Combprod	0.066	0.559	0.636	0.515
Euclidean	0.064	0.591	0.649	0.503

(c) *Corel Small* repository

Tabla B.4: MAP, average ranks and adjusted p-values corresponding to a Holm post-hoc test when comparing the proposed method to a simplified single-distance scheme, in all repositories.

Method	MAP	Avg. Rank	p_{adj}
Proposed	**0.379**	**1.00 (1)**	–
Single-Cosine	0.209	6.34 (7)	$< 10^{-68}$
Single-Mahalanobis	0.250	3.27 (2)	$< 10^{-13}$
Single-distance (p=0.5)	0.221	5.82 (6)	$< 10^{-56}$
Single-distance (p=1)	0.236	4.59 (5)	$< 10^{-32}$
Single-distance (p=1.5)	0.244	3.51 (4)	$< 10^{-16}$
Single-distance (p=2)	0.244	3.47 (3)	$< 10^{-16}$

(a) *Small* repository

Method	MAP	Avg. Rank	p_{adj}
Proposed	**0.192**	**1.21 (1)**	–
Single-Cosine	0.120	5.61 (6)	$< 10^{-47}$
Single-Mahalanobis	0.119	5.76 (7)	$< 10^{-50}$
Single-distance (p=0.5)	0.158	2.57 (2)	$< 10^{-6}$
Single-distance (p=1)	0.148	3.32 (3)	$< 10^{-12}$
Single-distance (p=1.5)	0.138	4.60 (5)	$< 10^{-28}$
Single-distance (p=2)	0.135	4.93 (4)	$< 10^{-34}$

(b) *Art* repository

Method	MAP	Avg. Rank	p_{adj}
Proposed	**0.070**	**2.01 (1)**	–
Single-Cosine	0.061	4.44 (6)	$< 10^{-15}$
Single-Mahalanobis	0.056	6.06 (7)	$< 10^{-40}$
Single-distance (p=0.5)	0.062	3.16 (2)	$< 10^{-4}$
Single-distance (p=1)	0.062	3.87 (3)	$< 10^{-9}$
Single-distance (p=1.5)	0.062	4.06 (4)	$< 10^{-11}$
Single-distance (p=2)	0.062	4.40 (5)	$< 10^{-15}$

(c) *Corel Small* repository

Apéndice C

Combining feature extraction and expansion to improve classification based similarity learning

Autores:
Emilia López-Iñesta, Francisco Grimaldo, Miguel Arevalillo-Herráez

Revista:
Pattern Recognition Letters

Año 2016, volumen 93

ISSN: 0167-8655

DOI
http://doi.org/10.1016/j.patrec.2016.11.005

C.1. Notation

Notation	Description
$X = \{x_i\}$	Objects collection
n	Objects collection size
d	Feature space dimensionality
d_M	Mahalanobis distance
x_i	Generic object
$\mathbf{x}_i$	Feature vector of a generic object x_i
x_i'	Transpose vector x_i
M'	Transpose matrix M
W	Linear transformation matrix
$\mathbb{S}_+^d$	Cone of simmetric PSD matrices
S	Similar pairs of objects
D	Dissimilar pairs of objects
R	Triplets of objects
$L(M)$	Function in metric learning optimization problem
$l(M, S, D, R)$	Loss term
$R(M)$	Regularizer term
λ	Trade-off between regularizer and loss terms
$d_s(\cdot, \cdot)$	Set of standard distances,
p	Minkowski distance parameter
L_p	Minkowski distances
$\|\|$	Concatenation symbol
k	Index for pairs of objects $k = 1, ..., K$
p_k	Pairs of objects
l_k	Similar/disimilar label for pairs of objects p_k
$\mathbf{p_k}$	Concatenated feature vectors pairs of objects
$\mathbf{p_k'}$	Concatenated feature vectors pairs of objects after transformation
γ	Parameter Gaussian Radial Basis Function (RBF)
C	Parameter Gaussian Radial Basis Function (RBF)

Combining feature extraction and expansion to improve classification based similarity learning

C.2. Abstract

Metric learning has been shown to outperform standard classification based similarity learning in a number of different contexts. In this paper, we show that the performance of classification similarity learning strongly depends on the data format used to learn the model. We then present an enriched classification similarity learning method that follows a hybrid approach that combines both feature extraction and feature expansion. In particular, we propose a data transformation and the use of a set of standard distances to supplement the information provided by the feature vectors of the training samples. The method is compared to state-of-the-art feature extraction and metric learning approaches, using linear learning algorithms in both a classification and a regression experimental setting. Results obtained show comparable performances in favour of the method proposed.

Keywords: Classification similarity learning, Metric Learning, Feature extraction, Feature expansion

C.3. Introduction

Comparisons are an essential task in many Pattern Recognition and Machine Learning methods. Given a collection of objects $X = \{x_1, x_2, ..., x_n\}$, with associated representations in a multidimensional vector space, $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\} \subseteq \mathbb{R}^d$, the simplest way to compare objects is by using standard similarity/distance functions on the feature-based representation e.g. Euclidean, Mahalanobis or cosine, to name a few. However, similarity measures are context dependent and they do not necessarily yield the best results.

An alternative and more sophisticated approach consists of learning the similarity function from training data, using a set of known comparison results. These training results may be pairwise (x_i and x_j are similar/dissimilar) or relative (x_i is closer to x_j than to x_k). Metric Learning (ML) and Classification Similarity Learning (CSL) are two common approaches to tackle this problem.

In metric learning, the goal is to define a distance $d_M = d_M(x_i, x_j) = (\mathbf{x}_i - \mathbf{x}_j)' M (\mathbf{x}_i - \mathbf{x}_j) \, \forall x_i, x_j \in X$, where M is a positive semi-definite (PSD) matrix learned from the training data by minimizing (or maximizing) some

criteria related to the performance of the function d_M. A major advantage of these methods is that the distance d_M is a pseudometric. Hence d_M can be seamlessly integrated into existing classification approaches that assume pseudometric spaces e.g. nearest neighbor.

In classification based similarity learning, a set of pairwise relations is used to learn a classifier. Once the classifier has been trained, it is used to yield scores that are related to the similarity between objects. Although the resulting values do not satisfy the properties of a pseudometric, these methods are a competitive alternative approach when these properties are not needed e.g. ranking purposes. In its simplest form, each pairwise relation is represented by concatenating the raw features of the two objects. During training, this is given together with a binary label that represent whether the objects are similar or not. However, the feature based representation is generally a major factor that affects classification performance, as this determines to what extent classes are separable. It is hence expected that an appropriate selection of the input format may yield a more accurate score.

In this paper, we focus on the data input of classification based similarity learning methods. The main contribution of this work is the proposal of a novel combination of feature extraction and feature expansion to perform a transformation of the original features that improves the similarity score independently of the classification method employed. In a first extraction stage, the concatenation-based representation described above is transformed into a space where data can be separated easier by following Kumar et al. (2011). In a second expansion phase, the data input is enriched by concatenating additional features that gather different relations between the raw features of the two objects as in López-Iñesta et al. (2014a). In particular, the values of a new set of distinct and standard distance measures have been considered. Under this setting, the results obtained are above state-of-the-art metric learning methods that were reported to perform consistently better than classification similarity learning approaches (Köstinger et al., 2012).

The rest of the paper is organized as follows. Section C.4 describes related work dealing with similarity learning, including both ML and CSL approaches. Section C.5 explains the proposed classification-based similarity learning framework. The experimental setting and the analysis of the results are presented in Sections C.6 and C.7. Section C.8 shows the implications of extending the proposed method to the non-linear case. Finally, Section C.9 states the conclusions and discusses future work.

C.4. Related work

A number of methods in classification, computer vision and pattern recognition rely on the application of a similarity function on data samples. The relatively strong performance dependence between the methods and

the similarity function has motivated an extensive research on approaches that attempt to learn the function from example data, in order to produce a customized function that is more adequate for the problem at hand. In this context, it is a common setting to learn from a set of pairs of objects, conveniently labeled as similar/dissimilar; or from a set of explicit relations between objects.

C.4.1. Metric Learning

Metric learning methods use this training information to search for a transformation matrix M that allows to compute distances as $d_M(x_i, x_j) = (\mathbf{x}_i - \mathbf{x}_j)'M(\mathbf{x}_i - \mathbf{x}_j)$ where $M \in \mathbb{S}_+^d$ and $\mathbb{S}_+^d$ is the cone of symmetric PSD d x d real-valued matrices. $M \in \mathbb{S}_+^d$ ensures that d_M satisfies the properties of a pseudo-distance and parametrizes a Mahalanobis distance family. Taking into consideration that any PSD matrix can be decomposed as $M = W'W$, it can easily be shown the above distance is equivalent to computing a linear projection of the data into a new space where constraints are satisfied better, and then using the Euclidean distance to compare the samples. In the absence of this projection ($M = W = I$), d_M is the Euclidean distance.

A first seminal work in metric learning was presented by Xing et al. (2002). In this work, they formulated Metric Learning as a convex optimization problem, using the training information to define the constraints (also known as must-link/cannot link). The goal was to maximize the sum of distances between all pairs of dissimilar instances and minimize the distances for all similar pairs by using Semidefinite Programming with a projected gradient descent algorithm. The metric learn was used to improve the performance of the k-Nearest Neighbors algorithm (k-NN).

In general, metric learning can be formulated as an optimization problem that has the following general form (Bellet et al., 2015):

$$\min_M L(M) = \ell(M, S, D, R) + \lambda R(M) \tag{C.1}$$

where $L(M)$ is a loss function. The first term (loss term) applies a penalty when constraints are not fulfilled; $R(M)$ is a regularizer term on the parameters M of the learned metric; and λ is a trade-off between the regularizer and the loss.

The different methods in the literature are characterised by using different loss functions, regularizers on M and constraints. In this section, we concentrate on some well-performing learning algorithms that motivate the approach presented in this paper, and refer the reader to the complete surveys by Kulis (2013) and Bellet et al. (2015). The following ML methods presented in this section have been used in the extensive performance comparisons presented in section C.6.

A popular algorithm used for k-NN classification is the Large Margin Nearest Neighbour (LMNN) approach (Weinberger y Saul, 2009). In this case, the authors inspired their work on neighborhood component analysis (Goldberger et al., 2004), and introduced the concept of *target neighbors* for an instance x_i as the k nearest neighbors with the same label y_i that belong to a local neighborhood defined by a sphere of some radius. They also established a safety perimeter to push away instances with different labels (*impostors*). LMNN's formulation tries to increase similarity to target neighbours, while reducing it to impostors lying within the k-NN region. To estimate the solution matrix M, they use gradient descent on the objective function. Despite that LMNN performs very well in practice, it is sometimes prone to over-fitting.

In Information Theoretic Metric Learning (ITML) (Davis et al., 2007b), an information-theoretic measure is used and the LogDet divergence is introduced as a regularizer term in the optimization problem to avoid over-fitting. The authors translate the problem of learning an optimal distance metric to that of learning the optimal Gaussian with respect to an entropic objective. ITML considers simple distance constraints enforcing that similar instances have a distance lower than a given upper bound $d_M(x_i, x_j) \leq u$; and dissimilarity instances be further than a specific lower bound $d_M(x_i, x_j) \geq \nu$. The optimization method computes Bregman projections and no Semidefinite Programming is required.

Logistic Discriminant Metric Learning (LDML) (Guillaumin et al., 2009) presents an approach for the particular context of Face Identification. The authors model the probability that two images (x_i, x_j) depict the same person as $p_{ij} = p(y_i = y_j | x_i, x_j; M, b) = \sigma(b - d_M(x_i, x_j))$, where $\sigma(z) = (1 + exp(-z))^{-1}$ is the sigmoid function that maps the distance to class probability and b is a bias term that acts as the optimal distance threshold value and is learned together with metric parameters. As $d_M(x_i, x_j)$ is linear with respect to the elements of M, it is possible to rewrite $p_{ij} = \sigma(bW'X_{ij})$ where W is the vector containing the elements of M and X_{ij} the entries of $(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)$ so the model p_{ij} appears as a standard linear logistic discriminant model $\sum_{i,j} t_{ij} ln(p_{ij}) + (1 - t_{ij}) ln(1 - p_{ij})$ where $t_{ij} = 1$ denotes the equality of labels y_i and y_j. The matrix M is estimated using the maximum likelihood by projected gradient ascent in an iterative manner.

A more recent approach, namely "Keep It Simple and Straightforward MEtric" (KISSME) has more recently been proposed by Köstinger et al. (2012). In this case, the distance metric is learned from equivalence constraints (S and D) applied to Face verification and person re-identification. This method tackles the problem from a statistical inference perspective and, in contrast to others, does not pose a complex optimization problem. Hence, it does not require computationally expensive iterations and it is orders of magnitudes faster than other comparable techniques.

C.4.2. Classification Similarity Learning

In CSL the learning of the metric matrix M is replaced by the use of a classifier. Although the properties of a pseudo-metric do not hold in this case, the approach is still valid when the pairwise similarity measure learned does not need to be integrated in other methods whose theoretical formulation is based on the use of a pseudo-metric e.g. algorithms like k-NN. A typical scope of application is the construction of similarity-based rankings, which are necessary in a wide range of applications e.g. multimedia retrieval.

Different CSL methods generally share the same classification framework, and differ in the format of the data input and/or the classification mechanism used. As data representation plays a key role in the performance of machine learning methods, the transformation of raw data into a more meaningful and informative input has become an active topic of research (Bengio et al., 2013; Arevalillo-Herráez et al., 2008a). In a machine learning context, it is commonly known as *Feature Extraction* (Guyon et al., 2006), and refers to methods that deal with any transformation or combination from input data to features.

In the last years, the many forms of feature extraction have been studied through a large number of diverse methods from different machine learning areas (Storcheus et al., 2015). One of the most well-studied techniques is *Feature Selection*, which is concerned with choosing the best feature subset from the original input feature set (Guyon y Elisseeff, 2003). It can also be considered as a special case of a more general feature weighting approach, and includes filter methods, wrapper methods, and embedded methods. Another particular feature extraction approach is the creation of new features from input data, usually known as *Feature Construction* (Guyon et al., 2006). As representative examples, Woznica y Kalousis (2010) face a binary classification problem where the learning instances are the pairwise absolute differences of the original instances; and this approach is extended in a face verification context, by adding element-wise products between the feature vectors (Berg y Belhumeur, 2012).

It has also been common to use the computed features to expand the original vectors, in a so called *Feature Expansion* setting. The use of these approaches has been common as general methods in pattern recognition (Yao et al., 2003; Tsai et al., 2011) and in specific contexts, such as text retrieval (Dalton et al., 2014), intrusion detection (Guo et al., 2014) and sentiment analysis (Jotheeswaran y Koteeswaran, 2015). For example, Guo et al. (2014) and Tsai et al. (2011) extend the original feature vectors by computing new features that are related to the distances from each data sample to a number of centroids found by a clustering algorithm.

C.5. Proposed method

Given a number of n labeled pairs of objects $p_k = (x_{k_1}, x_{k_2})$ $k = 1, \ldots, K$, $x_{k_1}, x_{k_2} \in X$ and their corresponding labels $l_k \in \{similar, dissimilar\}$ $k = 1, \ldots, K$, the simplest way to train and predict with a classifier is by using the feature based representation of the objects. The training data in this case is given as information-label pairs $\{\mathbf{p_k}, l_k\}$, where $\mathbf{p_k} = \mathbf{x_{k_1}}||\mathbf{x_{k_2}}$ and $||$ denotes the concatenation operator.

Another way to face this learning problem is by considering the vectors $\{\mathbf{p_k}\}$ as if they were the raw data, and posing a new feature extraction problem over the already extracted features. This is, rather than using the original set of vectors $\{\mathbf{p_k}\}$, the idea is to use them to construct a more meaningful and informative set of features, $\{\mathbf{p'_k}\}$, that is more adequate for classification tasks.

This type of approach has previously been used in other contexts, achieving significant improvements in the results. For example, a simple and practical transformation in a face verification context was presented in Kumar et al. (2009, 2011), and later used in Köstinger et al. (2012). In this case, the term $\mathbf{p_k}$ in the information-label pairs was replaced by $\mathbf{p'_k} = \mathbf{abs}(\mathbf{x_{k_1}} - \mathbf{x_{k_2}})||\mathbf{x_{k_1}} * \mathbf{x_{k_2}}$, where abs, $-$ and $*$ denote the absolute value, subtraction and multiplication element-wise operations, respectively. This straight forward yet effective transformation is based on the arguments that a) the differences between the features will be small if elements are similar; and b) that the sign of the multiplication is important to separate samples around 0. Its potential has also been recently validated in an image retrieval context (López-Iñesta et al., 2015b), showing a significantly higher performance in the low sample case as compared to using the original feature vectors.

This kind of transformation is specially relevant when linear classifiers are used, as the original features $\mathbf{p_k} = \mathbf{x_{k_1}}||\mathbf{x_{k_2}}$ yield a non-separable problem and hence are not well suited as an input. To illustrate this issue, lets consider a simple toy example with a single feature per object, in the interval [0,1]. Lets call a the feature from the first object, b the feature from the second and consider the function in Fig. C.1(a), that relates the absolute difference between the features to the probability that the two objects are similar. Fig. C.1(b) plots a distribution of similar (red crosses) and non-similar (blue bullets) instances generated at random under this setting, which cannot be linearly separated. Similar instances tend to lay near the diagonal, where the value of both features is close and hence the distance approaches zero. Although dissimilar instances stay far from the diagonal, they distribute at both sides, hindering separation. On the other hand, Fig. C.1(c) shows the result of using the proposed input $\mathbf{p'_k} = \mathbf{abs}(\mathbf{x_{k_1}} - \mathbf{x_{k_2}})||\mathbf{x_{k_1}} * \mathbf{x_{k_2}}$. In this toy example, this is equivalent to replacing the original features a and b by

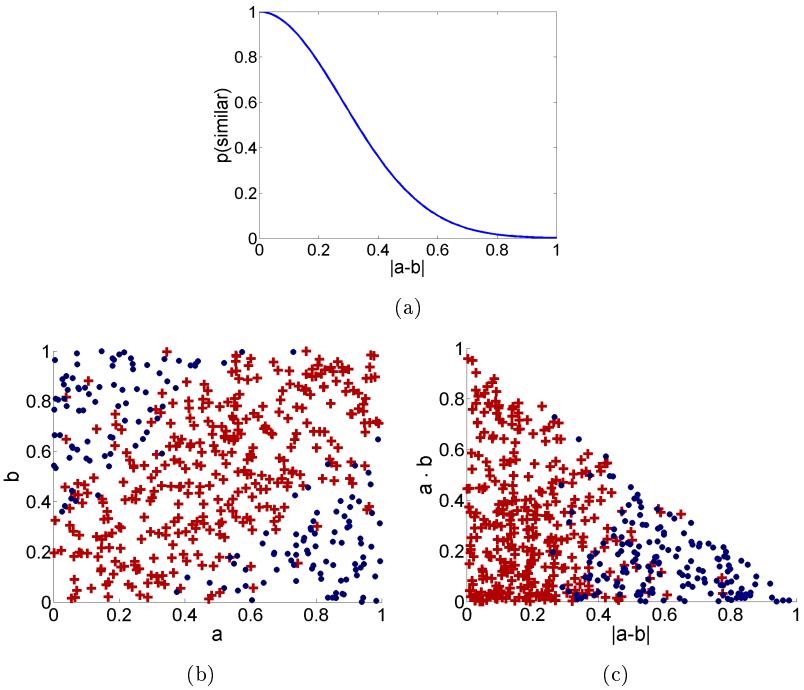


Figura C.1: Effect of the data transformation proposed in Kumar et al. (2011).

$|a-b|$ and $a \cdot b$. Data in this figure is clearly more suited for linear classification than the ones shown in Fig. C.1(b).

When the set of vectors $\{\mathbf{p_k}\}$ contain information related to different descriptors e.g. color, texture, etc., a feature extraction process that considers the implicit relation between the features of a same descriptor has been shown to lead to additional benefits. In particular, the use of several standard distances independently computed in each descriptor space has shown to boost performance in López-Iñesta et al. (2014a). In this case, a pool composed of four Minkowski distances (Lp norms for $p \in \{0.5, 1, 1.5, 2\}$) was used to transform the original feature vector p_k into a new one composed of distance values defined on the multiple descriptor spaces. With m descriptors, this leads to a new set of feature vectors, each composed of $4m$ features. Despite the partial loss of information as compared to using the original features, the intrinsic dimensionality reduction associated with the method has a compensatory effect and leads to higher precision rates.

In this paper, we build on the methods presented in Kumar et al. (2009, 2011) and López-Iñesta et al. (2014a), by using a hybrid approach that combines both feature extraction (EXTRAC) and feature expansion (EXPAN) (see Fig. C.2). First, the transformation from the original features to the

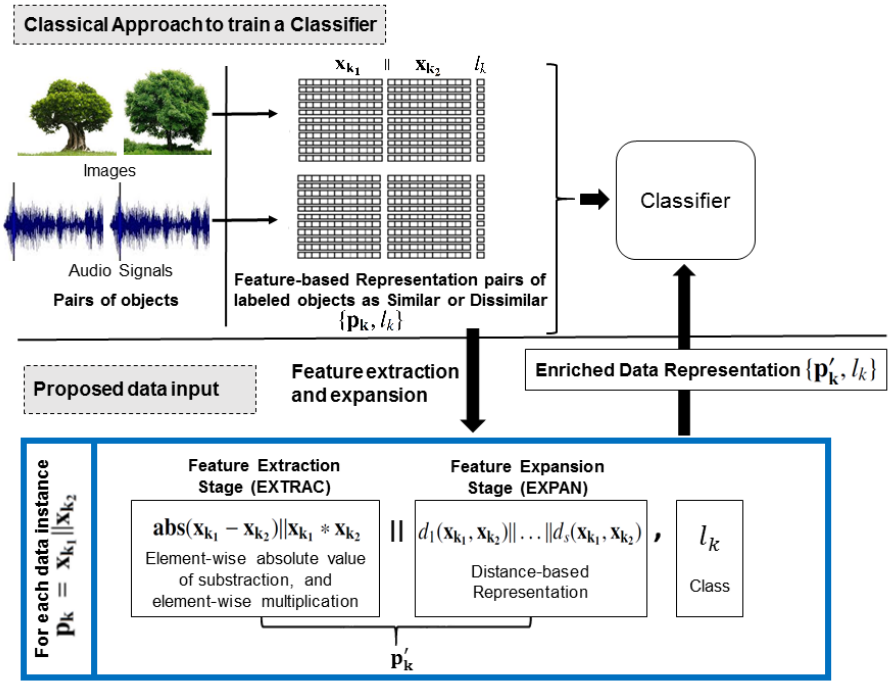


Figura C.2: Enriched Classification Similarity Learning data format.

element-wise absolute value of the subtraction and multiplication has been adopted. This is shown to provide a substantial advantage over the original features in the linear case, which has been reported in the experimental section. Then, a set of distance values between the vectors $\mathbf{x_{k_1}}$ and $\mathbf{x_{k_2}}$ is used to expand the original vectors and supplement the information contained therein. In particular, the proposed input format are the information-label pairs $\{\mathbf{p'_k}, l_k\}$, with $\mathbf{p'_k}$ defined according to Eq. C.2

$$\mathbf{p'_k} = \mathbf{abs}(\mathbf{x_{k_1}} - \mathbf{x_{k_2}}) || \mathbf{x_{k_1}} * \mathbf{x_{k_2}} || d_1(\mathbf{x_{k_1}}, \mathbf{x_{k_2}}) || \ldots || d_s(\mathbf{x_{k_1}}, \mathbf{x_{k_2}}) \quad \text{(C.2)}$$

where $d_1(\cdot,\cdot) \ldots d_s(\cdot,\cdot)$ represent a set of standard distance functions defined on the original feature space.

This feature expansion is inspired by previous works in López-Iñesta et al. (2014a), but the use in this context is significantly different. On the one hand, the computed distances are used to expand, and not replace, the information contained in $\mathbf{p_k}$. This implies a dimensionality increase compared to the original data dimension, contrary to the purpose in López-Iñesta et al. (2014a). In addition, whether the data is group or not by descriptors is irrelevant to the method. This makes it a more general approach and allows for an explicit comparison to state-of-the-art techniques in metric learning.

The rationale behind this enlarged input format is twofold. On the one hand, the distinct nature of each distance function may contribute to the learning by catching a different relation between the features in each image pair. For example, while the cosine distance considers the features as vectors and focuses on the angle between them, the Euclidean distance considers them as points and measures the straight line distance between the points. On the other hand, the feature-based input format constructed in the extraction stage avoids the implicit loss of similarity-related information that otherwise occurs when the original features are transformed into distance values. While the inclusion of the distances implies a slight dimensionality increase, this is, in general, proportionally small when compared to dimensionality of the original data.

Results reported in the experimental section clearly show that the proposed hybrid approach helps the classifier learn a more informed similarity score function and leads to an improved Enriched Classification Similarity Learning (ECSL) method. As a further advantage of the proposed representation, it should be noted that it guarantees that the resulting score is symmetric, a property that does not hold when the original set of vectors $\{\mathbf{p_k}\}$ are used.

C.6. Experimental setting

An extensive experimentation has been carried out to compare the performance of the proposed Enriched Classification Similarity Learning method to that of other state-of-the-art feature extraction and metric learning approaches. In particular, we have evaluated our ECSL method against: a) using a linear SVM with the input format used by Kumar et al. (2009, 2011) ($\mathbf{p_k} = \mathbf{x_{k_1}}||\mathbf{x_{k_2}}$), and b) a set of representative ML methods, namely: Information-Theoretic Metric Learning (ITML) (Davis et al., 2007b), Large-Margin Nearest Neighbors (LMNN)(Weinberger y Saul, 2009), Linear Discriminant Metric Learning (LDML) (Guillaumin et al., 2009) and KISS MEtric Learning (KISSME) (Köstinger et al., 2012).

To test the validity of the proposed input as a framework, we have tested it in both a classification and a regression setting. For the classification setting, we have used the Linear Support Vector Machine from the LibSVM implementation (Chang y Lin, 2011), setting the cost parameter to the default value of 1, as in Köstinger et al. (2012). For the regression setting, we have used Logistic Regression (LR) (Bishop, 2006) as a well-known binary classification method whose prediction values are probabilities to belong to the positive (similar) class. In this section we focus on algorithms that produce linear models so as to perform a fair comparison with the existing metric learning techniques. A discussion on the effects of ECSL in the non-linear case can be found in Section C.8.

For the feature expansion carried out by the proposed method (see the standard distances added in Equation C.2), we have considered a pool composed of four Minkowski distances (L_p norms), for values $p = 0.5, 1, 1.5, 2$, plus the Cosine and Mahalanobis distances. These are widely used dissimilarity measures that have shown relatively large differences in performance on the same data (Aggarwal et al., 2001; Howarth y Rüger, 2005), and hence suggest that may be combined to obtain improved results. Fractional values of p have been included because they have been reported to provide more meaningful results for high dimensional data, both from the theoretical and empirical perspective. In turn, the Cosine distance considers the features as vectors and focuses on the angle between them, whereas the Mahalanobis distance takes into account the covariance of data.

The performance of all methods has been tested in up to 34 standard datasets of diverse nature coming from two sources: 4 state-of-the-art image datasets for face verification and object recognition and 30 representative general purpose datasets from the UCI Machine Learning Repository [2]. For face verification and object recognition we selected four different well-known and challenging datasets that contain images with important variations in illumination, poses or scale. The main characteristics of these datasets are as follows (a summary can be found in Table C.1):

- The ***Labeled Faces in the Wild (LFW)*** (Huang et al., 2007) consists of 13233 face images of 5749 people taken from the Yahoo! News Web. We use the image restricted test protocol on LFW where the only available information is whether each pair of training images are from the same subject or not. Out of all the possible feature vectors available for the LFW dataset, we use the Scale-Invariant Feature Transform (SIFT) based feature vectors (Guillaumin et al., 2009) of LFW and the high-level face representation obtained in (Kumar et al., 2009). These are referred to as LFW-SIFT and LFW-Attr, respectively.

- The ***PubFig database*** (Kumar et al., 2009) is a large, real-world face dataset consisting of 58797 images of 200 people collected from Google and Flickr. Its image attributes provide high-level semantic features indicating the presence or absence of visual face traits (such as hair, glass, age, race, smiling and so on) and allow a semantic description that is more robust against large image variations and that can lead to good verification performance.

- The ***ToyCars*** (Nowak y Jurie, 2007) dataset contains 256 image crops of 14 different toy cars and trucks, which is used in a classification task that consists of determining the corresponding vehicle to a new unseen image.

[2] `http://archive.ics.uci.edu/ml/`

Tabla C.1: A summary of the four datasets for face verification and object recognition.

Dataset	Size	Dim	Classes	Train	Test
ToyCars	256	50	14	8515	7381
LFW-SIFT	13233	100	5479	5400	600
LFW-Attr	13233	65	5479	5351	596
PubFig	58797	65	200	18000	2000

To make the above datasets tractable for the distance metric learning algorithms, Köstinger et al. (2012) performed a dimensionality reduction by Principal Component Analysis (PCA) in a pre-processing step that we inherit in our comparative study. The LFW-SFIT dataset is projected onto a 100 dimensional subspace, LFW-Attr and Pubfig to a 65 dimensional subspace and the ToyCars database to a 50 dimensional subspace. As indicated by the authors, the influence of the PCA dimensionality in these datasets is not critical and does not lead to significant changes in the ranked comparative performance of the tested methods.

The details about the 30 datasets being tested from the UCI Machine Learning Repository can be found in Table C.4. The domains covered by these datasets are also long-familiar and address from medical pathologies (e.g. Arrhythmia, Hepatitis) to car evaluations (e.g. Auto, Car). Our selection includes a broad range of datasets, showing a high diversity with regard to the number of instances, the data dimension and the number of classes.

As for the previous datasets, instances of the databases from the UCI repository have been respectively projected by PCA onto lower dimensional subspaces that account for at least 99 % of the total variance of the data. We also run a sensitivity analysis to test the effects of varying the PCA variance coverage parameter in the range (0.9, 1). Although results have not been included due to space limitations, ECSL was mostly ranked in the same position for all databases and, thus, the PCA dimensionality does not influence the interpretation of the results shown in this paper.

For the sake of comparison, we inherit the experimental framework presented in Köstinger et al. (2012), that allows for the evaluation of the methods according to their performance at ranking a number of pairs according to their estimated similarity. To this end, each repository is divided into 10 folds of disjoint objects, and a cross validation approach is used (only in the case of the ToyCars dataset we used 2 folds).

For each experiment, one fold is chosen for test and the remainder ones are used for training. Training and test sets are generated at random, according to the class information available. Results on each fold are appropriately combined to produce a ROC curve for each method.

The number of pairs in the training and test sets used for face verification

and object recognition are the same as in Köstinger et al. (2012) and are summarized in Table C.1. In turn, the training and test sets used for the datasets from the UCI repository were constructed from a set of 6000 random pairs of instances divided into 10 folds.

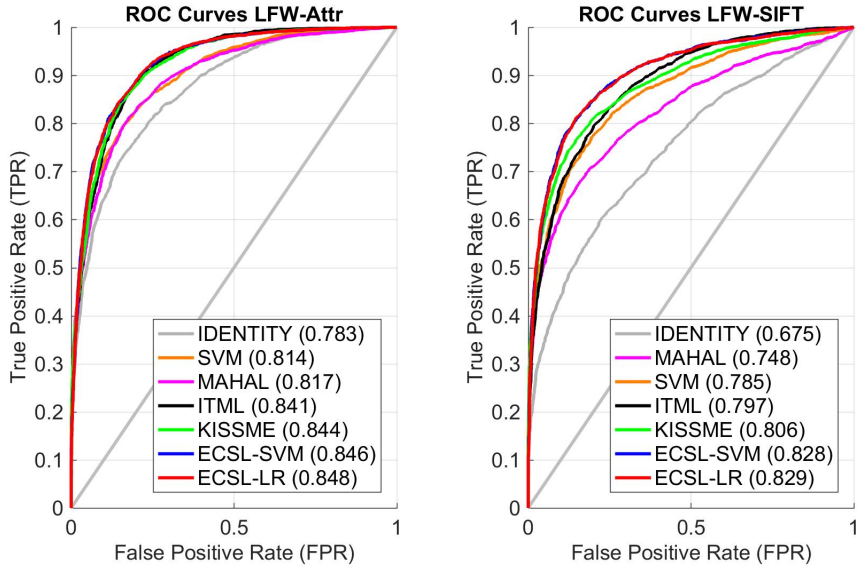
C.7. Results

We start our analysis of results with the performance obtained for the datasets dealing with face verification and object recognition. Figure C.3 shows comparative ROC curves for the databases LFW-Attr, LFW-SIFT and PubFig. Throughout this section we use the acronyms ECSL-SVM and ECSL-LR to refer, respectively, to the proposed Enriched Classification Similarity Learning using either a linear SVM or logistic regression as the core classification method. For the sake of clarity, plots presented in this section only include two metric learning methods, namely KISSME and ITML. The first has been chosen because it yields consistently the best results across all metric learning methods in all repositories. The second because it is a method frequently used in the literature for comparison purposes in metric learning contexts (Cao et al., 2013; Jain et al., 2008). We have also included the results of the Mahalanobis (MAHAL) and Euclidean distances (IDENTITY) as baselines. The Equal Error Rate (EER) is provided in brackets as part of the legend, and also shown in Table C.2 for all methods, including those not plotted in Figure C.3.

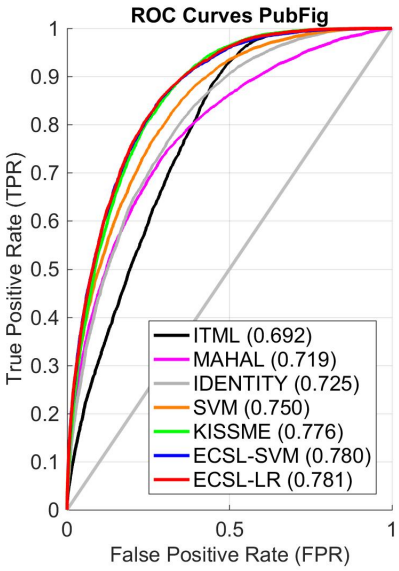
Tabla C.2: Equal Error Rate for ECSL-LR, ECSL-SVM and the rest of the methods in all databases. Best results in each dataset are marked in bold.

Method	LFW-Attr	LFW-SIFT	Pub Fig	Toy Cars	Toy Cars*
ECSL-LR	**0.848**	**0.829**	**0.781**	**0.986**	0.965
ECSL-SVM	0.846	0.828	0.780	0.902	**0.981**
KISSME	0.844	0.806	0.776	0.934	0.970
SVM	0.814	0.785	0.750	0.811	0.864
MAHAL	0.817	0.748	0.719	0.898	0.944
ITML	0.841	0.797	0.692	0.706	0.896
LDML	0.834	0.796	0.776	0.716	0.720
LMNN	0.831	0.785	0.735	0.805	0.919
IDENTITY	0.783	0.675	0.725	0.716	0.720

From the plots in Figure C.3 we can observe that the classification performance of the proposed method (in its two variants ECSL-LR and ECSL-SVM) outperforms the best of the metric learning techniques in the PubFig and LFW (Attributes) repositories. In LFW-SIFT, ECSL clearly dominates all others learning algorithms included in the comparison. In addition, a substantial performance increase can be observed with respect to the use of



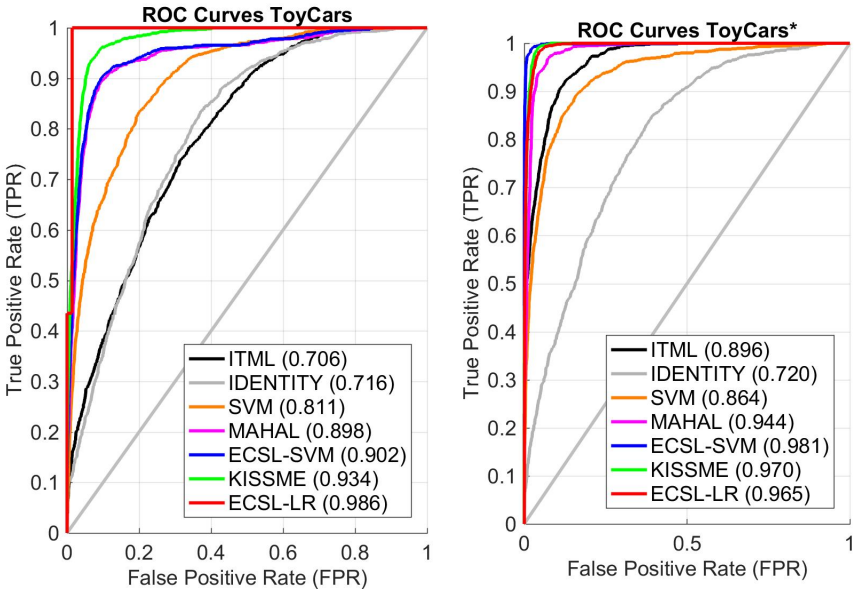
(a) (*b*)



(*c*)

Figura C.3: Comparative performance between the method proposed (ECSL-LR and ECSL-SVM), KISSME, ITML, MAHAL, SVM, IDENTITY in a) LFW-Attr; b) LFW-SIFT; and c) PubFig.

an SVM without the added distances in LFW-SIFT, LFW-Attr and PubFig. In fact, the standard SVM method consistently obtains worst results than any of the two metric learning approaches shown in the plots. The only ex-



(a) (b)

Figura C.4: Comparative performance between the method proposed (ECSL-LR and ECSL-SVM) and the comparative methods in the ToyCars database a) using pairs from disjoint sets in training and test; b) using disjoint sets of random pairs for training and test (ToyCars*)

ception is in PubFig, where the standard SVM method appears below the KISS method but performs better than ITML. In Table C.2, it can also be observed that LDML and LMNN methods are ranked different depending on the database but they are always outperformed by ECSL and KISSME.

The ToyCars repository behaves differently depending on whether ECSL is applied to a classification or regression setting. As shown in Figure C.4(b), KISSME outperforms ECSL-SVM, whereas ECSL-LR consistenly ranks first as in the rest of databases. The low performance of ITML is also noticeable in this case, scoring below the Euclidean distance. Yet the improvement of ECSL-SVM with respect to the standard SVM is remarkable in this database. To further study the performance of ECSL in this database, we have run a second experiment (see ToyCars* in Table C.2). Now, instead of selecting pairs from disjoint sets of objects for training and test, pairs have been randomly chosen (taking care that no pair is simultaneously used for training and test). These two experiments are representative of different scenarios. In the first one, the similarity function is learned from a set of objects, and then it is applied on a different set of never seen objects. This set-up is useful when we have a collection of classified objects that can be used to generate the similar and dissimilar pairs. In ToyCars*, the function is learned from a selection of pairs extracted from the same repository, as typically happens in

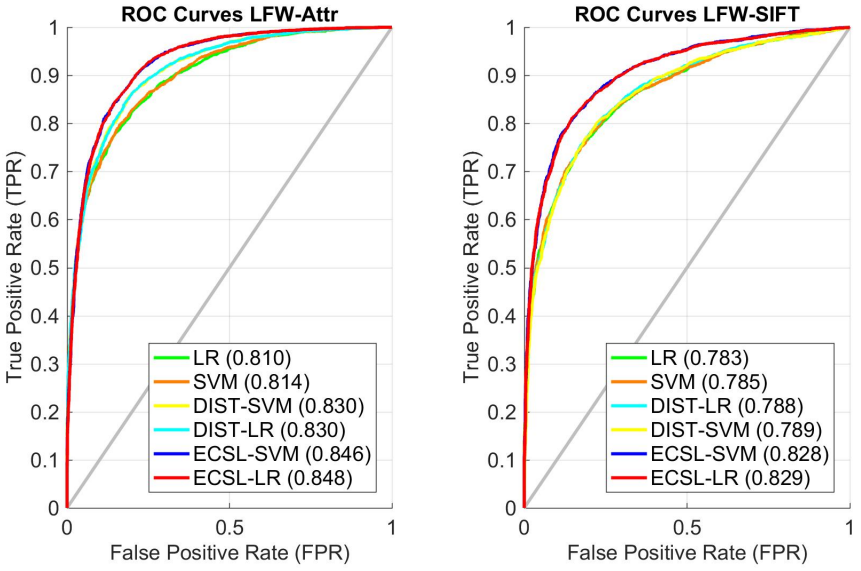
retrieval problems where it is the user who judges the similarity. The results for this second experiment are shown in Figure C.4(b). In this case, ECSL-SVM scores the best, but the performance of ECSL-LR is reduced so that now it is ranked after KISSME.

To test whether the improvement achieved is due to a particularly good behavior of one of the selected standard distances and whether feature expansion using a set of well-known distance values makes any benefit, we compare the performance of the proposed method (ECSL-LR and ECSL-SVM) to that obtained by using each of the added distances in isolation (see Table C.3). In all cases, we can observe that the performance of ECSL stays close to that of the best of the single distances (usually the Mahalanobis distance) but, when all distances are incorporated to the training data format, the performance is significantly boosted. Figure C.5 graphically shows the boosting effect obtained from feature expansion with respect to either learning a classifier (SVM or LR) using the state-of-the-art data input format (Kumar et al., 2011) or training the classifier with just the pool of considered distances (DIST-SVM or DIST-LR) , thus taking it as a distance combination method (López-Iñesta et al., 2014b). In all databases, adding distances to the original feature space resulted in a significant improvement in performance.

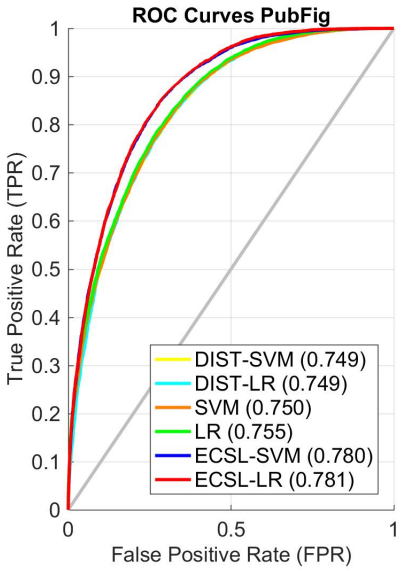
Tabla C.3: Equal Error Rate for ECSL-LR, ECSL-SVM and the considered standard distances (both combined and in isolation) for all image databases. Best results in each dataset are marked in bold.

Method	LFW-Attr	LFW-SIFT	Pub Fig	Toy Cars	Toy Cars*
ECSL-LR	**0.848**	**0.829**	**0.781**	**0.986**	0.965
ECSL-SVM	0.846	0.828	0.780	0.902	**0.981**
DIST-SVM	0.830	0.789	0.749	0.901	0.966
DIST-LR	0.830	0.788	0.749	0.936	0.966
SVM	0.814	0.785	0.750	0.811	0.864
LR	0.810	0.783	0.755	0.539	0.556
MAHAL	0.817	0.748	0.719	0.898	0.944
COSINE	0.782	0.702	0.725	0.723	0.730
IDENTITY	0.783	0.675	0.725	0.716	0.720
MINKOWSKI-1.5	0.794	0.692	0.731	0.728	0.719
MINKOWSKI-1	0.789	0.698	0.712	0.736	0.695
MINKOWSKI-0.5	0.753	0.688	0.669	0.707	0.653

Overall, the same results have been reproduced for the 30 datasets from the UCI Machine Learning Repository reported in this paper. Table C.4 shows the performances obtained for ECSL and KISSME, as this was the best alternative from the rest of the methods in the evaluation. In up to 21 databases one variant of the proposed method, ECSL-SVM or ECSL-LR,



(a) (b)



(c)

Figura C.5: Comparative performance between the method proposed (ECSL-LR and ECSL-SVM) and standard distances in a) LFW-Attr; b) LFW-SIFT; and c) PubFig.

performed the best. In 9 databases ECSL ranked second after KISSME but, even in these cases, the differences in performance where less than 0.01 (with the only exception of the Kimia database). On the contrary, when ECSL

ranks first differences are often larger (e.g. in 12 databases were greater than 0.01).

Tabla C.4: Equal Error Rate for ECSL-LR, ECSL-SVM and KISSME in databases from UCI. Best results in each dataset are marked in bold.

Dataset	Size x Dim - Classes	ECSL -LR	ECSL -SVM	KISSME
Soybean2	136 x 35 - 4	**0.816**	0.775	0.800
Iris	150 x 4 - 3	**0.925**	0.924	0.870
Hepatitis	155 x 19 - 2	0.754	0.756	**0.765**
Wine	178 x 13 - 3	0.848	0.844	**0.858**
Imox	192 x 8 - 4	**0.651**	0.649	0.644
Sonar	208 x 60 - 2	0.593	0.593	**0.597**
Glass	214 x 9 - 4	0.607	**0.613**	0.594
Ovarian	216 x 100- 2	0.829	0.828	**0.837**
Kimia	216 x 4096 - 18	0.870	0.814	**0.930**
Soybean1	288 x 35 - 15	**0.886**	0.875	0.839
Malaysia	291 x 8 - 20	**0.785**	**0.785**	0.782
Heart	297 x 13 - 2	**0.702**	0.701	0.692
Ecoli	336 x 7 - 8	0.583	**0.851**	0.845
Ionosphere	351 x 34 - 2	**0.784**	0.776	0.779
Auto	398 x 6 - 2	**0.762**	0.761	0.753

Dataset	Size x Dim - Classes	ECSL -LR	ECSL -SVM	KISSME
Arrhythmia	420 x 278 - 12	0.857	0.841	**0.866**
Diabetes	768 x 8 - 2	0.511	0.513	**0.520**
Breast	699 x 9 - 2	**0.917**	0.913	0.848
Car	1728 x 6 - 4	**0.562**	0.558	0.537
Mfeat Mor	2000 x 6 - 10	0.788	0.788	**0.789**
Mfeat kar	2000 x 64 - 10	0.708	0.717	**0.720**
Mfeat fou	2000 x 76 - 10	**0.816**	0.814	0.808
Mfeat fac	2000 x 216 - 10	**0.762**	0.760	0.714
Mfeat pix	2000 x 240 - 10	0.517	**0.609**	0.588
Mfeat Zer	2000 x 240 - 10	**0.611**	0.608	0.586
Mfeat	2000 x 649 - 10	0.598	**0.736**	0.701
Twonorm	7400 x 20 - 2	0.947	**0.949**	0.887
Ringnorm	7400 x 20 - 2	**0.706**	0.705	0.645
Cbands	12000 x 30 - 24	**0.541**	0.506	0.522
Texturel	81920 x 7 - 5	**0.875**	0.866	0.870

In order to further assess the behavior of these methods in all databases, evaluate their performance across the whole ranking, and analyze whether the differences are statistically significant, an analysis of the EER scores has also been performed. With these measures, a non-parametric Friedman test (Garcia y Herrera, 2008) has been run for all databases, with the null hy-

Tabla C.5: Friedman Ranking and p-values for the best three methods across all different databases.

Method	ECSL-LR	ECSL-SVM	KISSME
Friedman Average Ranking	**1.64**	2.07	2.26
p-values	-	0.072	0.019

pothesis that all methods perform equivalently. The resulting average ranks for the three best algorithms are shown in the first row of Table C.5. To test if there are significant differences between the first positioned method (ECSL-LR) and the remaining ones, we have also carried out a post-hoc Holm test at a significance level $\alpha = 0.05$. The p-values computed through the statistics of Friedman and the Iman-Davenport extension are shown in the second row of Table C.5. They suggest that differences are statistically significant with respect to the KISSME method ($p = 0.019 < 0.05$).

C.8. The non-linear case

To extend our study and analyse the performance of the proposal also in the non-linear case, we have repeated the entire experimentation using a non-linear SVM. Both Radial Basis Function (RBF) and third degree polynomial kernels (POLY) have been considered. Table C.6 shows a comparison between the previously introduced linear (LR and SVM-LIN suffixes) and non-linear models (SVM-RBF and SVM-POLY suffixes). Performance is compared when using three distinct representations, namely: the raw representation $\mathbf{p_k} = \mathbf{x_{k_1}}||\mathbf{x_{k_2}}$ (RAW prefix), when extending it with the distances as $\mathbf{p'_k} = \mathbf{x_{k_1}}||\mathbf{x_{k_2}}||d_1(\mathbf{x_{k_1}}, \mathbf{x_{k_2}})|| \dots ||d_s(\mathbf{x_{k_1}}, \mathbf{x_{k_2}})$ (EXPAN prefix) and when applying the proposed combination of feature extraction and expansion according to Eq. C.2 (ECSL prefix). For the sake of readability, this table shows a sufficiently representative selection of databases that covers different combinations of number of instances, features and classes.

The Equal Error Rates for each combination of classifier and data representation are reported in Table C.6. Best results for each database are in bold. In general, the ECSL representation behaves the best, but one size does not fit all and one can always find databases where the best performance is obtained for other method combinations, such as EXPAN-SVM-RBF in Ionosphere or Breast. Even in these cases, though, the feature expansion stage is beneficial for the non-linear classifiers, already casting a good performance. In fact, results indicate that the addition of the new distance-based features (EXPAN) has a positive impact on the performance in comparison to RAW representation in all cases, both using linear or non-linear models.

With regard to the use of the EXTRACT transformation proposed in Kumar et al. (2011), better results are not guaranteed in the non-linear case. This can be deduced by comparing the EXPAN and ECSL results in the last two rows for each data representation, which correspond to the SVMs that use non-linear kernels. As it can be observed, there is not a clear bias towards any of the methods as it happens in the linear case (first two rows) in favour of the ECSL data representation. This is in part because the transformation does not necessarily help the separability of the data in the non-linear case, as the argument provided in section C.4.2 and illustrated in Fig. C.1 only holds in the linear case.

Finally, with reference to the comparative performance of linear vs non-linear methods, this depends upon the particular database and no rule of thumb can be given in this respect.

Tabla C.6: Comparison of Equal Error Rate for RAW, ECSL and EXPAN approaches in several databases. Best results in each dataset are marked in bold.

Method	LFW-Attr	LFW-SIFT	PubFig	ToyCars
RAW-LR	0.549	0.533	0.521	0.509
RAW-SVM-LIN	0.548	0.633	0.518	0.509
RAW-SVM-RBF	0.698	0.600	0.497	0.688
RAW-SVM-POLY	0.749	0.705	0.488	0.700
EXPAN-LR	0.829	0.796	0.898	0.754
EXPAN-SVM-LIN	0.821	0.756	0.709	0.750
EXPAN-SVM-RBF	0.842	0.804	0.897	0.769
EXPAN-SVM-POLY	0.827	0.797	0.869	0.743
ECSL-LR	**0.848**	**0.829**	**0.986**	**0.781**
ECSL-SVM-LIN	0.846	0.828	0.902	0.780
ECSL-SVM-RBF	0.836	0.812	0.905	0.769
ECSL-SVM-POLY	0.834	0.811	0.890	0.752

Method	Auto	Twonorm	Ionosphere	Breast
RAW-LR	0.665	0.503	0.704	0.750
RAW-SVM-LIN	0.665	0.503	0.703	0.698
RAW-SVM-RBF	0.491	0.863	0.904	0.950
RAW-SVM-POLY	0.505	0.893	0.879	0.925
EXPAN-LR	0.761	0.799	0.779	0.894
EXPAN-SVM-LIN	0.757	0.795	0.759	0.812
EXPAN-SVM-RBF	0.527	0.868	**0.909**	**0.954**
EXPAN-SVM-POLY	0.506	0.896	0.906	0.938
ECSL-LR	**0.762**	0.947	0.784	0.916
ECSL-SVM-LIN	0.761	**0.949**	0.776	0.914
ECSL-SVM-RBF	0.518	0.932	0.886	0.930
ECSL-SVM-POLY	0.495	0.919	0.887	0.922

C.9. Conclusion

Some recent methods in the Metric Learning literature have reported a significantly higher performance than Classification based Similarity Learning approaches e.g. KISSME (Köstinger et al., 2012). In this paper, we show that this situation can be reverted by choosing a more appropriate data format. In particular, for each pair of objects p_k, we combine the information contained in their feature vectors as shown in Equation C.2, considering the result of a set of distance functions on the original feature space. With the transformation presented, classification similarity learning turns into a more competitive method, showing comparable performances in favour of the method proposed.

Although the SVM and logistic regression have been used to develop a proof of concept, the method is by no means restricted to the use of this particular classifier. On the contrary, the framework presented is open to the use of alternative classification methods and/or meta-estimators.

Another important aspect not considered in this research is the robustness of the methods to variations of the training size. Responses to small sample size situation are specially relevant when the number of examples is scarce (e.g. Content Based Image Retrieval). In this context, previous work e.g. López-Iñesta et al. (2015b) has already outlined the potential of integrating standard distance values into a classification approach for similarity learning.

Bibliografía

Aggarwal, C., Hinneburg, A. y Keim, D. On the surprising behavior of distance metrics in high dimensional space. En *Database Theory ICDT* (editado por J. Bussche y V. Vianu), vol. 1973 de *Lecture Notes in Computer Science*, páginas 420–434. Springer Berlin Heidelberg, 2001. (Citado en páginas 17, 24, 59, 63, 72, 81 y 104.)

Ali, S. y Smith-Miles, K. Improved support vector machine generalization using normalized input space. En *AI 2006: Advances in Artificial Intelligence*, vol. 4304 de *Lecture Notes in Computer Science*, páginas 362–371. Springer Berlin Heidelberg, 2006. (Citado en páginas 59 y 77.)

Antipov, G., Berrani, S.-A., Ruchaud, N. y Dugelay, J.-L. Learned vs. hand-crafted features for pedestrian gender recognition. En *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, páginas 1263–1266. ACM, New York, NY, USA, 2015. (Citado en páginas 14 y 31.)

Arevalillo-Herráez, M., Domingo, J. y Ferri, F. J. Combining similarity measures in content-based image retrieval. *Pattern Recognition Letters*, vol. 29(16), páginas 2174–2181, 2008a. (Citado en páginas 13, 75 y 99.)

Arevalillo-Herráez, M. y Ferri, F. J. An improved distance-based relevance feedback strategy for image retrieval. *Image and Vision Computing*, vol. 31(10), páginas 704 – 713, 2013. (Citado en páginas 61 y 78.)

Arevalillo-Herráez, M., Zacarés, M., Benavent, X. y de Ves, E. A relevance feedback cbir algorithm based on fuzzy sets. *Image Communication*, vol. 23(7), 2008b. (Citado en página 61.)

Ashton, K. That «Internet of Things» Thing. *RFID Journal*, 2009. (Citado en página 3.)

Ayache, S., Quénot, G. y Gensel, J. Classifier fusion for svm-based multimedia semantic indexing. En *Proceedings of the 29th European Conference on IR Research*, ECIR, páginas 494–504. Springer-Verlag, Berlin, Heidelberg, 2007. (Citado en página 75.)

AYALA, G. y DOMINGO, J. Spatial size distributions: Applications to shape and texture analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23(12), páginas 1430–1442, 2001. (Citado en página 79.)

BAEZA-YATES, R. A. y RIBEIRO-NETO, B. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999. (Citado en página 25.)

BAR-HILLEL, A., HERTZ, T., SHENTAL, N. y WEINSHALL, D. Learning distance functions using equivalence relations. En *Proceedings of the 20th International Conference on Machine Learning*, páginas 11–18. 2003. (Citado en páginas 33, 55, 74 y 80.)

BAR-HILLEL, A., HERTZ, T., SHENTAL, N. y WEINSHALL, D. Learning a mahalanobis metric from equivalence constraints. *Journal of Machine Learning Research*, vol. 6, páginas 937–965, 2005. (Citado en página 80.)

BELKIN, M. y NIYOGI, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, vol. 15, páginas 1373–1396, 2002. (Citado en página 32.)

BELLET, A., HABRARD, A. y SEBBAN, M. Similarity learning for provably accurate sparse linear classification. En *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. 2012. (Citado en página 56.)

BELLET, A., HABRARD, A. y SEBBAN, M. *Metric Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2015. (Citado en páginas 10, 11, 22, 24, 32, 71, 72 y 97.)

BELLMAN, R. y BELLMAN, R. *Adaptive Control Processes: A Guided Tour*. Princeton Legacy Library. Princeton University Press, 1961. (Citado en página 27.)

BENGIO, Y. y COURVILLE, A. Deep learning of representations. En *Handbook on Neural Information Processing* (editado por M. Bianchini, M. Maggini y L. C. Jain), páginas 1–28. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. (Citado en páginas 13 y 31.)

BENGIO, Y., COURVILLE, A. y VINCENT, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35(8), páginas 1798–1828, 2013. (Citado en páginas 13, 29, 31 y 99.)

BERG, T. y BELHUMEUR, P. N. Tom-vs-pete classifiers and identity-preserving alignment for face verification. En *British Machine Vision Conference, BMVC*, páginas 1–11. 2012. (Citado en páginas 31 y 99.)

BHOWMIK, N., GONZALEZ, V. R., GOUET-BRUNET, V., PEDRINI, H. y BLOCH, G. Efficient fusion of multidimensional descriptors for image retrieval. En *IEEE International Conference on Image Processing, ICIP 2014, Paris, France, October 27-30, 2014*, páginas 5766–5770. 2014. (Citado en página 28.)

BILENKO, M. y MOONEY, R. J. Adaptive duplicate detection using learnable string similarity measures. En *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, páginas 39–48. 2003. (Citado en página 76.)

BISHOP, C. M. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, 1995. (Citado en página 27.)

BISHOP, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. (Citado en página 103.)

BRUNNER, C., FISCHER, A., LUIG, K. y THIES, T. Pairwise support vector machines and their application to large scale problems. *Journal of Machine Learning Research*, vol. 13(Aug), páginas 2279–2292, 2012. (Citado en página 11.)

BURHANUDDIN, I. A., BAJAJ, P., SHEKHAR, S., MUKHERJEE, D., RAJ, A. y SANKAR, A. Similarity learning for product recommendation and scoring using multi-channel data. En *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, páginas 1143–1152. 2015. (Citado en página 11.)

BUSTOS, B. y SKOPAL, T. Dynamic similarity search in multi-metric spaces. En *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, MIR '06, páginas 137–146. ACM, New York, NY, USA, 2006. (Citado en páginas 17 y 18.)

CAENEN, G. y PAUWELS, E. J. Logistic regression model for relevance feedback in content-based image retrieval. En *Storage and Retrieval for Media Databases*, páginas 49–58. 2002. (Citado en página 72.)

CALUMBY, R. T., GONÇALVES, M. A. y DA SILVA TORRES, R. On interactive learning-to-rank for ir: Overview, recent advances, challenges, and directions. *Neurocomputing*, vol. 208, páginas 3 – 24, 2016. (Citado en página 79.)

CAO, Q., YING, Y. y LI, P. Similarity metric learning for face recognition. En *The 16th IEEE International Conference on Computer Vision (ICCV)*, páginas 2408–2415. 2013. (Citado en página 106.)

CASTANEDO, F. A Review of Data Fusion Techniques. *The Scientific World Journal*, vol. 2013, páginas 19+, 2013. (Citado en página 27.)

CHANG, C.-C. y LIN, C.-J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, vol. 2, páginas 27:1–27:27, 2011. (Citado en página 103.)

CHÁVEZ, E., NAVARRO, G., BAEZA-YATES, R. y MARROQUÍN, J. L. Searching in metric spaces. *ACM Comput. Surv.*, vol. 33(3), páginas 273–321, 2001. (Citado en página 19.)

CHECHIK, G., SHARMA, V., SHALIT, U. y BENGIO, S. Large scale online learning of image similarity through ranking. *J. Mach. Learn. Res.*, vol. 11, páginas 1109–1135, 2010. (Citado en página 11.)

CHELLAPPA, R. y CHATTERJEE, S. Classification of textures using Gaussian Markov random fields. *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33(4), páginas 959–963, 1985. (Citado en página 79.)

CHEN, H. y MEER, P. Robust fusion of uncertain information. En *2003 Conference on Computer Vision and Pattern Recognition Workshop*, vol. 6, páginas 64–64. 2003. (Citado en página 27.)

CHEN, Y. y DOUGHERTY, E. R. Gray-scale morphological granulometric texture classification. *Opt. Eng.*, vol. 33(8), páginas 2713–2722, 1994. (Citado en página 79.)

CHEN, Y., GARCIA, E. K., GUPTA, M. R., RAHIMI, A. y CAZZANTI, L. Similarity-based classification: Concepts and algorithms. *J. Mach. Learn. Res.*, vol. 10, páginas 747–776, 2009. (Citado en página 11.)

CHUI, M., LÖFFLER, M. y ROBERTS, R. The internet of things. *McKinsey Quarterly*, vol. 2(2010), páginas 1–9, 2010. (Citado en página 3.)

CONNERS, R. W., TRIVEDI, M. M. y HARLOW, C. A. Segmentation of a high-resolution urban scene using texture operators. *Computer Vision, Graphics, and Image Processing*, vol. 25(3), páginas 273–310, 1984. (Citado en página 79.)

CUADRAS, C. M. Distancias estadísticas. *Estadística Española*, vol. 30(119), páginas 295–357, 1989. (Citado en páginas 19, 21 y 23.)

CUNNINGHAM, J. y GHAHRAMANI, Z. Linear dimensionality reduction: survey, insights, and generalizations. *Journal of Machine Learning Research*, 2015. (Citado en página 30.)

DALTON, J., DIETZ, L. y ALLAN, J. Entity query feature expansion using knowledge base links. En *Proceedings of the 37th International ACM SIGIR Conference on Research Development in Information Retrieval*, SIGIR '14, páginas 365–374. ACM, New York, NY, USA, 2014. (Citado en páginas 31 y 99.)

DATTA, R., JOSHI, D., LI, J. y WANG, J. Z. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, vol. 40(2), páginas 5:1–5:60, 2008. (Citado en página 28.)

DAVIS, J. V. y DHILLON, I. S. Structured metric learning for high-dimensional problems. En *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, KDD '08. 2008. (Citado en páginas 72 y 74.)

DAVIS, J. V., KULIS, B., JAIN, P., SRA, S. y DHILLON, I. S. Information-theoretic metric learning. En *Proceedings of the 24th International Conference on Machine Learning*, vol. 227 de *ICML*, páginas 209–216. ACM, New York, NY, USA, 2007a. (Citado en páginas 11, 33, 71, 74 y 80.)

DAVIS, J. V., KULIS, B., JAIN, P., SRA, S. y DHILLON, I. S. Information-theoretic metric learning. En *Proceedings of the 24th International Conference Machine Learning (ICML)* (editado por Z. Ghahramani), vol. 227, páginas 209–216. ACM, 2007b. (Citado en páginas 55, 98 y 103.)

DEZA, M. M. y DEZA, E. *Encyclopedia of Distances*. Springer Berlin Heidelberg, 2009. (Citado en páginas 19 y 21.)

DIMITROVSKI, I., LOSKOVSKA, S. y CHORBEV, I. Efficient content-based image retrieval using support vector machines for feature aggregation. En *Innovations in Computing Sciences and Software Engineering* (editado por T. Sobh y K. Elleithy), páginas 319–324. Springer Netherlands, 2010. (Citado en página 75.)

DOMENICONI, C., GUNOPULOS, D. y PENG, J. Large margin nearest neighbor classifiers. *IEEE Transactions on Neural Networks*, vol. 16(4), páginas 899–909, 2005. (Citado en páginas 11 y 72.)

DONG, Y., GAO, S., TAO, K., LIU, J. y WANG, H. Performance evaluation of early and late fusion methods for generic semantics indexing. *Pattern Analysis and Applications*, vol. 17(1), páginas 37–50, 2014. (Citado en página 74.)

DUIN, R. P. W. y PEKALSKA, E. The dissimilarity space: Bridging structural and statistical pattern recognition. *Pattern Recognition Letters*, vol. 33(7), páginas 826–832, 2012. (Citado en páginas 56 y 58.)

EVANS, D. The internet of everything: How more relevant and valuable connections will change the world. *Cisco IBSG*, páginas 1–9, 2012. (Citado en página 4.)

FARIA, F. A., DOS SANTOS, J. A., ROCHA, A. y TORRES, R. D. S. A framework for selection and fusion of pattern classifiers in multimedia recognition. *Pattern Recogn. Lett.*, vol. 39, páginas 52–64, 2014. (Citado en páginas 13 y 27.)

FISHER, R. A. The statistical utilization of multiple measurements. *Annals of Eugenics*, vol. 8(4), páginas 376–386, 1938. (Citado en página 32.)

GARCIA, S. y HERRERA, F. An extension on statistical comparisons of classifiers over multiple data sets for all pairwise comparisons. *Journal of Machine Learning Research*, vol. 9, páginas 2677–2694, 2008. (Citado en páginas 82 y 111.)

GARCÍA, S., LUENGO, J. y HERRERA, F. *Data Preprocessing in Data Mining*, vol. 72 de *Intelligent Systems Reference Library*. Springer, 2015. (Citado en páginas 29 y 30.)

VAN GEMERT, J. C., GEUSEBROEK, J. M., VEENMAN, C. J., SNOEK, C. G. M. y SMEULDERS, A. W. M. Robust scene categorization by learning image statistics in context. En *CVPR Workshop on Semantic Learning Applications in Multimedia*. 2006. (Citado en página 78.)

GIACINTO, G. y ROLI, F. Nearest-prototype relevance feedback for content based image retrieval. En *Proceedings of 17th International Conference on the Pattern Recognition, (ICPR), Volume 2*, vol. 2, páginas 989–992. IEEE Computer Society, Washington, DC, USA, 2004. (Citado en páginas 61 y 75.)

GIACINTO, G. y ROLI, F. Instance-based relevance feedback for image retrieval. En *Advances in Neural Information Processing Systems 17 (NIPS)*, páginas 489–496. MIT Press, 2005. (Citado en página 79.)

GLOBERSON, A. y ROWEIS, S. T. Metric learning by collapsing classes. En *Advances in Neural Information Processing Systems 18 (NIPS)*, páginas 451–458. MIT Press, 2005. (Citado en páginas 32, 56, 57, 71 y 76.)

GOLDBERGER, J., ROWEIS, S. T., HINTON, G. E. y SALAKHUTDINOV, R. Neighbourhood components analysis. En *Advances in Neural Information Processing Systems 17 (NIPS)*, páginas 513–520. MIT Press, 2004. (Citado en página 98.)

GOWER, J. C. Measures of similarity, dissimilarity and distance. *Encyclopedia of statistical sciences*, vol. 5(3), páginas 397–405, 1985. (Citado en página 21.)

GOWER, J. C. y LEGENDRE, P. Metric and euclidean properties of dissimilarity coefficients. *Journal of Classification*, vol. 3(1), páginas 5–48, 1986. (Citado en página 20.)

GUBBI, J., BUYYA, R., MARUSIC, S. y PALANISWAMI, M. Internet of things (iot): A vision, architectural elements, and future directions. *Future Gener. Comput. Syst.*, vol. 29(7), páginas 1645–1660, 2013. (Citado en página 3.)

GUILLAUMIN, M., VERBEEK, J. y SCHMID, C. Is that you? metric learning approaches for face identification. En *The 12th IEEE International Conference on Computer Vision (ICCV)*, páginas 498–505. 2009. (Citado en páginas 74, 98, 103 y 104.)

GUO, C., ZHOU, Y., PING, Y., ZHANG, Z., LIU, G. y YANG, Y. A distance sum-based hybrid method for intrusion detection. *Applied Intelligence*, vol. 40(1), páginas 178–188, 2014. (Citado en páginas 16, 31 y 99.)

GUYON, I. y ELISSEEFF, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, vol. 3, páginas 1157–1182, 2003. (Citado en páginas 29, 30 y 99.)

GUYON, I., GUNN, S., NIKRAVESH, M. y ZADEH, L. A. *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*. Springer-Verlag New York, Inc., 2006. (Citado en páginas 13, 29, 30 y 99.)

GUYON, I., WESTON, J., BARNHILL, S. y VAPNIK, V. Gene selection for cancer classification using support vector machines. *Machine Learning*, vol. 46(1), páginas 389–422, 2002. (Citado en página 30.)

HALL, D. y LLINAS, J. An introduction to multisensor data fusion. *Proceedings of the IEEE*, vol. 85(1), páginas 6–23, 1997. (Citado en página 27.)

HERTZ, T. *Learning distance functions: Algorithms and applications. PhD thesis*. Hebrew Univ. Jerusalem, 2006. (Citado en página 26.)

HERTZ, T., BAR-HILLEL, A. y WEINSHALL, D. Learning distance functions for image retrieval. En *In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, páginas 570–577. 2004. (Citado en página 30.)

HOWARTH, P. y RÜGER, S. Fractional distance measures for content-based image retrieval. En *Proceedings of the 27th European conference on Advances in Information Retrieval Research (ECIR)*, páginas 447–456. Springer-Verlag, Berlin, Heidelberg, 2005. (Citado en páginas 17, 59, 63, 72, 81 y 104.)

HUANG, G. B., RAMESH, M., BERG, T. y LEARNED-MILLER, E. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Informe Técnico 07-49, University of Massachusetts, Amherst, 2007. (Citado en página 104.)

HYVÄRINEN, A. y OJA, E. Independent component analysis: Algorithms and applications. *Neural Netw.*, vol. 13(4-5), páginas 411–430, 2000. (Citado en página 32.)

IBBA, A., DUIN, R. P. W. y LEE, W.-J. A Study on Combining Sets of Differently Measured Dissimilarities. En *Proceedings of the 20th International Conference on Pattern Recognition*, páginas 3360–3363. IEEE Computer Society, 2010. (Citado en página 17.)

IQBAL, Q. y AGGARWAL, J. K. Combining structure, color and texture for image retrieval: A performance evaluation. En *16th International Conference on Pattern Recognition (ICPR)*, páginas 438–443. 2002. (Citado en páginas 62 y 75.)

JACOBS, D. W., WEINSHALL, D. y GDALYAHU, Y. Classification with nonmetric distances: Image retrieval and class representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22(6), páginas 583–600, 2000. ISSN 0162-8828. (Citado en página 17.)

JAIN, P., KULIS, B., DHILLON, I. S. y GRAUMAN, K. Online metric learning and fast similarity search. En *Advances in Neural Information Processing Systems 21 (NIPS)*, páginas 761–768. MIT Press, 2008. (Citado en página 106.)

JOTHEESWARAN, J. y KOTEESWARAN, S. A weighted semantic feature expansion using hyponymy tree for feature integration in sentiment analysis. En *International Conference on Green Computing and Internet of Things (ICGCIoT)*, páginas 289–293. 2015. (Citado en páginas 31 y 99.)

KELLEY, J. *General topology*. D. Van Nostrand Company, Inc., Toronto-New York-London, 1955. (Citado en página 18.)

KLUDAS, J., BRUNO, E. y MARCHAND-MAILLET, S. Information fusion in multimedia information retrieval. En *Adaptive Multimedia Retrieval: Retrieval, User, and Semantics: 5th International Workshop, AMR 2007, Paris, France, July 5-6, 2007 Revised Selected Papers* (editado por N. Boujemaa, M. Detyniecki y A. Nürnberger), páginas 147–159. Springer Berlin Heidelberg, 2008. (Citado en página 26.)

KÖKNAR-TEZEL, S. y LATECKI, L. J. Improving SVM classification on imbalanced time series data sets with ghost points. *Knowledge and Information Systems*, vol. 28(1), páginas 1–23, 2011. (Citado en páginas 63 y 82.)

KÖSTINGER, M., HIRZER, M., WOHLHART, P., ROTH, P. M. y BISCHOF, H. Large scale metric learning from equivalence constraints. En *International Conference on Computer Vision and Pattern Recognition (CVPR)*, páginas 2288–2295. 2012. (Citado en páginas 11, 33, 40, 47, 96, 98, 100, 103, 105, 106 y 114.)

KSANTINI, R., ZIOU, D., COLIN, B. y DUBEAU, F. Logistic regression models for a fast CBIR method based on feature selection. En *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, páginas 2790–2795. 2007. (Citado en páginas 72, 75, 80 y 88.)

KSANTINI, R., ZIOU, D., COLIN, B. y DUBEAU, F. Weighted pseudometric discriminatory power improvement using a bayesian logistic regression model based on a variational method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30(2), páginas 253–266, 2008. (Citado en páginas 72, 75, 80 y 88.)

KULIS, B. Metric learning: A survey. *Foundations and Trends in Machine Learning*, vol. 5(4), páginas 287–364, 2013. (Citado en páginas 10 y 97.)

KUMAR, N., BERG, A. C., BELHUMEUR, P. N. y NAYAR, S. K. Attribute and simile classifiers for face verification. En *The 12th IEEE International Conference on Computer Vision (ICCV)*, páginas 365–372. 2009. (Citado en páginas 31, 100, 101, 103 y 104.)

KUMAR, N., BERG, A. C., BELHUMEUR, P. N. y NAYAR, S. K. Describable visual attributes for face verification and image search. En *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 33, páginas 1962–1977. 2011. (Citado en páginas XX, 31, 96, 100, 101, 103, 109 y 113.)

LEE, W.-J., DUIN, R. P. W., IBBA, A. y LOOG, M. An experimental study on combining euclidean distances. En *Proceedings 2nd International Workshop on Cognitive Information Processing (14-16 June, 2010 Elba Island, Tuscany - Italy)*, páginas 304–309. 2010. (Citado en páginas 15 y 56.)

LEE, W.-J., VERZAKOV, S. y DUIN, R. P. W. Kernel combination versus classifier combination. En *MCS* (editado por M. Haindl, J. Kittler y F. Roli), vol. 4472 de *Lecture Notes in Computer Science*, páginas 22–31. Springer, 2007. (Citado en páginas 17 y 76.)

LEÓN, T., ZUCCARELLO, P., AYALA, G., DE VES, E. y DOMINGO, J. Applying logistic regression to relevance feedback in image retrieval systems. *Pattern Recognition*, vol. 40(10), páginas 2621–2632, 2007. (Citado en páginas 61 y 78.)

LIAO, S., HU, Y., ZHU, X. y LI, S. Z. Person re-identification by local maximal occurrence representation and metric learning. En *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015. (Citado en página 11.)

LIU, H. y MOTODA, H. *Feature Extraction, Construction and Selection: A Data Mining Perspective*. Kluwer Academic Publishers, Norwell, MA, USA, 1998. (Citado en página 30.)

LIU, W., MU, C., JI, R., MA, S., SMITH, J. R. y CHANG, S. Low-rank similarity metric learning in high dimensions. En *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, páginas 2792–2799. 2015. (Citado en página 14.)

LIU, Y., ZHANG, H. H. y WU, Y. Hard or soft classification? large-margin unified machines. *Journal of the American Statistical Association*, vol. 106(493), páginas 166–177, 2011. (Citado en página 12.)

LIU, Y. y ZHENG, Y. F. FS_SFS: A novel feature selection method for support vector machines. *Pattern Recognition*, vol. 39(7), páginas 1333–1345, 2006. (Citado en páginas 14 y 56.)

LÓPEZ-IÑESTA, E., AREVALILLO-HERRÁEZ, M. y GRIMALDO, F. Classification-based multimodality fusion approach for similarity ranking. En *17th International Conference on Information Fusion (FUSION)*, páginas 1–6. 2014a. (Citado en páginas 15, 48, 96, 101 y 102.)

LÓPEZ-IÑESTA, E., AREVALILLO-HERRÁEZ, M. y GRIMALDO, F. Boosting classification based similarity learning by using standard distances. En *Artificial Intelligence Research and Development - Proceedings of the 18th International Conference of the Catalan Association for Artificial Intelligence*, páginas 153–162. 2015a. (Citado en páginas 15 y 48.)

LÓPEZ-IÑESTA, E., GRIMALDO, F. y AREVALILLO-HERRÁEZ, M. Comparing feature-based and distance-based representations for classification similarity learning. En *Artificial Intelligence Research and Development - Proceedings of the 17th International Conference of the Catalan Association for Artificial Intelligence*, páginas 23–32. 2014b. (Citado en páginas 15, 48 y 109.)

LÓPEZ-IÑESTA, E., GRIMALDO, F. y AREVALILLO-HERRÁEZ, M. Classification similarity learning using feature-based and distance-based representations: A comparative study. *Applied Artificial Intelligence*, vol. 29(5), páginas 445–458, 2015b. (Citado en páginas 15, 48, 100 y 114.)

LÓPEZ-IÑESTA, E., GRIMALDO, F. y AREVALILLO-HERRÁEZ, M. Combining feature extraction and expansion to improve classification based

similarity learning. *Pattern Recognition Letters*, vol. 93, páginas 95–103, 2017a. (Citado en páginas 16 y 49.)

LÓPEZ-IÑESTA, E., GRIMALDO, F. y AREVALILLO-HERRÁEZ, M. Learning similarity scores by using a family of distance functions in multiple feature spaces. *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 31(08), página 1750027, 2017b. (Citado en páginas 15 y 48.)

MCDONALD, K. y SMEATON, A. F. A comparison of score, rank and probability-based fusion methods for video shot retrieval. En *Proceedings of the 4th international conference on Image and Video Retrieval (CIVR)*, páginas 61–70. Springer-Verlag, Berlin, Heidelberg, 2005. (Citado en páginas 75 y 80.)

MCFEE, B., BARRINGTON, L. y LANCKRIET, G. Learning content similarity for music recommendation. *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20(8), páginas 2207–2218, 2012. (Citado en página 11.)

MCFEE, B. y LANCKRIET, G. Learning multi-modal similarity. *Journal of Machine Learning Research*, vol. 12, páginas 491–523, 2011. (Citado en página 76.)

MCFEE, B. y LANCKRIET, G. R. G. Metric learning to rank. En *Proceedings of the 27th International Conference on Machine Learning (ICML)*, páginas 775–782. Omnipress, 2010. (Citado en páginas 5, 55 y 71.)

MÜLLER, H., MARCHAND-MAILLET, S. y PUN, T. The truth about Corel - evaluation in image retrieval. En *Proceedings of the International Conference on Image and Video Retrieval (CIVR)*, páginas 38–49. 2002. (Citado en páginas 79 y 82.)

NAVARRO-ARRIBAS, G. y TORRA, V. Information fusion in data privacy: A survey. *Inf. Fusion*, vol. 13(4), páginas 235–244, 2012. (Citado en página 27.)

NG, J. Y.-H., HAUSKNECHT, M. J., VIJAYANARASIMHAN, S., VINYALS, O., MONGA, R. y TODERICI, G. Beyond short snippets: Deep networks for video classification. En *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015. (Citado en página 31.)

NOWAK, E. y JURIE, F. Learning visual similarity measures for comparing never seen objects. En *International Conference on Computer Vision and Pattern Recognition (CVPR)*, páginas 1–8. 2007. (Citado en página 104.)

PAPA, J. P., FALCÃO, A. X. y SUZUKI, C. T. N. Supervised pattern classification based on optimum-path forest. *Int. J. Imaging Syst. Technol.*, vol. 19(2), páginas 120–131, 2009. (Citado en páginas 17, 59 y 72.)

PARK, E., HAN, X., BERG, T. L. y BERG, A. C. Combining multiple sources of knowledge in deep cnns for action recognition. En *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, páginas 1–8. 2016. (Citado en páginas 14 y 31.)

PEARSON, K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, vol. 2, páginas 559–572, 1901. (Citado en página 32.)

PEKALSKA, E. y DUIN, R. P. W. *The Dissimilarity Representation for Pattern Recognition: Foundations And Applications (Machine Perception and Artificial Intelligence)*. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2005. (Citado en página 15.)

PEKALSKA, E., HAROL, A., DUIN, R., SPILLMANN, B. y BUNKE, H. *Non-Euclidean or non-metric measures can be informative*, vol. 4109 de *Lecture Notes in Computer Science*, páginas 871–880. Springer Verlag, Germany, 2006. (Citado en páginas 12 y 17.)

PEÑA, D. *Análisis de datos multivariantes*. McGraw-Hill, 2002. (Citado en página 32.)

PEREZ-SUAY, A. y FERRI, F. J. Scaling up a metric learning algorithm for image recognition and representation. En *Proceedings of the 4th International Symposium on Advances in Visual Computing, Part II*, ISVC '08, páginas 592–601. Springer-Verlag, Berlin, Heidelberg, 2008. (Citado en página 32.)

PIRAS, L. y GIACINTO, G. Information fusion in content based image retrieval: A comprehensive overview. *Information Fusion*, vol. 37, páginas 50 – 60, 2017. (Citado en página 28.)

PLATT, J. C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. En *Advances in large margin classifiers*, páginas 61–74. MIT Press, 1999. (Citado en páginas 60 y 78.)

ROSS, A. y JAIN, A. Information fusion in biometrics. *Pattern Recognition Letters*, vol. 24(13), páginas 2115–2125, 2003. (Citado en página 26.)

RUBNER, Y., PUZICHA, J., TOMASI, C. y BUHMANN, J. M. Empirical evaluation of dissimilarity measures for color and texture. *Computer Vision and Image Understanding*, vol. 84(1), páginas 25–43, 2001. (Citado en página 26.)

SANTINI, S. y JAIN, R. Similarity measures. *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21(9), páginas 871–883, 1999. (Citado en página 18.)

SCHEIRER, W. J., WILBER, M. J., ECKMANN, M. y BOULT, T. E. Good recognition is non-metric. *Pattern Recognition*, vol. 47(8), páginas 2721 – 2731, 2014. (Citado en página 17.)

SCHULTZ, M. y JOACHIMS, T. Learning a distance metric from relative comparisons. En *Advances in Neural Information Processing Systems 16 (NIPS)*. MIT Press, 2003. (Citado en página 74.)

SHAW, J. A. y FOX, E. A. Combination of multiple searches. En *The 2nd Text REtrieval Conference (TREC)*, páginas 243–252. 1994. (Citado en página 28.)

DA SILVA TORRES, R., FALCÃO, A. X., GONÇALVES, M. A., PAPA, J. P., ZHANG, B., FAN, W. y FOX, E. A. A genetic programming framework for content-based image retrieval. *Pattern Recognition*, vol. 42(2), páginas 283–292, 2009. (Citado en página 75.)

DA SILVA TORRES, R., FALCÃO, A. X., ZHANG, B., FAN, W., FOX, E. A., GONÇALVES, M. A. y CALADO, P. A new framework to combine descriptors for content-based image retrieval. En *Conference on Information and Knowledge Management (CIKM)*, páginas 335–336. 2005. (Citado en páginas 28 y 75.)

SIMONYAN, K. y ZISSERMAN, A. Two-stream convolutional networks for action recognition in videos. En *NIPS*. 2014. (Citado en página 31.)

SKOPAL, T. Unified framework for fast exact and approximate search in dissimilarity spaces. *ACM Trans. Database Syst.*, vol. 32(4), 2007. (Citado en página 23.)

SKOPAL, T. y BUSTOS, B. On nonmetric similarity search problems in complex domains. *ACM Comput. Surv.*, vol. 43(4), página 34, 2011. (Citado en páginas 17, 19 y 26.)

SMITH, G. y BURNS, I. Measuring texture classification algorithms. *Pattern Recognition Letters*, vol. 18, páginas 1495–1501, 1997. (Citado en página 79.)

SNOEK, C. G. M., WORRING, M. y SMEULDERS, A. W. M. Early versus late fusion in semantic video analysis. En *Proceedings of the 13th Annual ACM International Conference on Multimedia*, MULTIMEDIA '05, páginas 399–402. ACM, New York, NY, USA, 2005. (Citado en página 28.)

SOILLE, P. *Morphological Image Analysis: Principles and Applications*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2 edición, 2003. (Citado en página 61.)

STORCHEUS, D., ROSTAMIZADEH, A. y KUMAR, S. A survey of modern questions and challenges in feature extraction. En *Proceedings of The 1st International Workshop on Feature Extraction: Modern Questions and Challenges, NIPS*, páginas 1–18. 2015. (Citado en páginas 29, 30 y 99.)

TENENBAUM, J. B., DE SILVA, V. y LANGFORD, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science*, vol. 290(5500), página 2319, 2000. (Citado en página 32.)

THOMEE, B. y LEW, M. S. Interactive search in image retrieval: a survey. *International Journal of Multimedia Information Retrieval*, vol. 1(2), páginas 71–86, 2012. (Citado en páginas 62 y 81.)

TORRA, V. y NARUKAWA, Y. *Modeling decisions - information fusion and aggregation operators.*. Springer, 2007. (Citado en página 27.)

TORRA, V., NAVARRO-ARRIBAS, G. y ABRIL, D. Supervised learning for record linkage through weighted means and owa operators. *Control and Cybernetics*, vol. 39(4), páginas 1011–1026, 2010. (Citado en página 27.)

TSAI, C.-F., LIN, W.-Y., HONG, Z.-F. y HSIEH, C.-Y. Distance-based features in pattern classification. *EURASIP J. Adv. Sig. Proc.*, vol. 2011, página 62, 2011. (Citado en páginas 16, 31 y 99.)

TVERSKY, A. Features of similarity. *Psychological Review*, vol. 84(4), páginas 327–352, 1977. (Citado en páginas 12 y 17.)

TVERSKY, A. y GATI, I. Similarity, separability, and the triangle inequality. *Psychological review*, vol. 89(2), página 123, 1982. (Citado en página 17.)

VAPNIK, V. N. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995. (Citado en página 75.)

VERT, J., TSUDA, K. y SCHÖLKOPF, B. *A Primer on Kernel Methods*, páginas 35–70. MIT Press, Cambridge, MA, USA, 2004. (Citado en páginas 59, 75 y 77.)

DE VES, E., DOMINGO, J., AYALA, G. y ZUCCARELLO, P. A novel bayesian framework for relevance feedback in image content-based retrieval systems. *Pattern Recognition*, vol. 39(9), páginas 1622–1632, 2006. (Citado en páginas 61 y 78.)

VILLEGAS, M. y PAREDES, R. Score fusion by maximizing the area under the roc curve. En *Pattern Recognition and Image Analysis: 4th Iberian Conference, IbPRIA 2009 Póvoa de Varzim, Portugal, June 10-12, 2009 Proceedings* (editado por H. Araujo, A. M. Mendonça, A. J. Pinho y M. I. Torres), páginas 473–480. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. (Citado en página 27.)

WAN, J., WANG, D., HOI, S. C. H., WU, P., ZHU, J., ZHANG, Y. y LI, J. Deep learning for content-based image retrieval: A comprehensive study. En *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, páginas 157–166. ACM, New York, NY, USA, 2014. (Citado en página 14.)

WEINBERGER, K. Q. y SAUL, L. K. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, vol. 10, páginas 207–244, 2009. (Citado en páginas 33, 71, 72, 74, 98 y 103.)

WHITE, F. E. Data Fusion Lexicon, Joint Directors of Laboratories. Informe técnico, Naval Ocean Systems Center, 1987. (Citado en página 27.)

WOZNICA, A. y KALOUSIS, A. A new framework for dissimilarity and similarity learning. En *PAKDD (2)*, vol. 6119 de *Lecture Notes in Computer Science*, páginas 386–397. Springer, 2010. (Citado en páginas 12, 30 y 99.)

WU, S., CHEN, Y. C., LI, X., WU, A. C., YOU, J. J. y ZHENG, W. S. An enhanced deep feature representation for person re-identification. En *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, páginas 1–8. 2016. (Citado en páginas 14 y 31.)

XING, E. P., NG, A. Y., JORDAN, M. I. y RUSSELL, S. J. Distance metric learning with application to clustering with side-information. En *Advances in Neural Information Processing Systems 15 (NIPS)*, páginas 505–512. MIT Press, 2002. (Citado en páginas 11, 33, 56, 57, 71, 74, 76 y 97.)

YAGER, R. R. On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Trans. Syst. Man Cybern.*, vol. 18(1), páginas 183–190, 1988. (Citado en página 27.)

YAO, K., LU, W., ZHANG, S., XIAO, H. y ANDA LI. Feature expansion and feature selection for general pattern recognition problems. En *Proceedings of the 2003 International Conference on Neural Networks and Signal Processing*, vol. 1, páginas 29–32 Vol.1. 2003. (Citado en páginas 31 y 99.)

YUE, J., LI, Z., LIU, L. y FU, Z. Content-based image retrieval using color and texture fused features. *Mathematical and Computer Modelling*, vol. 54(3–4), páginas 1121 – 1127, 2011. (Citado en página 28.)

ZEZULA, P., AMATO, G., DOHNAL, V. y BATKO, M. *Similarity Search: The Metric Space Approach (Advances in Database Systems)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005. (Citado en páginas 18 y 19.)

ZHANG, J. y YE, L. Local aggregation function learning based on support vector machines. *Signal Processing*, vol. 89(11), páginas 2291–2295, 2009. (Citado en páginas 28, 56, 72 y 75.)

ZHANG, Q. y IZQUIERDO, E. Optimizing metrics combining low-level visual descriptors for image annotation and retrieval. En *The 31st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, páginas 405–408. 2006. (Citado en página 75.)