Revista Cubana de Ciencias Informáticas
Vol. 10, No. 4, Octubre-Diciembre, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 43- 67

Tipo de artículo: Artículo de revisión
Temática: Inteligencia Artificial
Recibido: 13/06/2016 | Aceptado: 28/09/2016

Una revisión sobre aprendizaje no supervisado de métricas de distancia

A brief review on unsupervised metric learning

Isabel Cristina Pérez Verona [1*], Leticia Arco García [2]

[1] Universidad "Máximo Gómez Báez" de Ciego de Ávila, Cuba. CP: 69450 isabelc@unica.cu
[2] Universidad Central "Marta Abreu" de Las Villas, Carretera a Camajuaní, km 5 ½. Santa Clara, Villa Clara, Cuba. CP: 54830 leticiaa@uclv.edu.cu

[*] Autor para correspondencia: isabelc@unica.cu

Resumen

Muchos de los métodos de aprendizaje automático dependen del cálculo de distancias en un espacio multidimensional para estimar la similitud entre dos ejemplos teniendo en cuenta la estructura de los datos. Está comprobado que se obtienen mejores resultados cuando la métrica se diseña específicamente para un contexto dado, pero esta es una tarea compleja. El aprendizaje de métricas de distancia consiste en aprender una métrica determinada respondiendo específicamente a las características de los datos históricos. En casos particulares donde no se conoce mucha información sobre los datos, se han obtenido buenos resultados utilizando algoritmos no supervisados de aprendizaje de distancias. Estos algoritmos no requieren información de etiqueta de clases, y se han utilizado principalmente para mejorar los resultados de métodos de agrupamiento. En este artículo se mencionan algunos de los aportes más recientes a los algoritmos no supervisados de aprendizaje de distancias, sus ventajas, desventajas y posibles aplicaciones.

Palabras clave: aprendizaje no supervisado, métrica, distancia, reducción de dimensión

Abstract

Several machine learning methods rely on the notion of distances in a multidimensional space, these distances are used for estimating the similarity between two objects, according to historical data. In such cases, when the metric is specifically designed to the context, better results are often obtained. However, designing a metric is a complex task. Metric learning automatically learns a distance metric according to the characteristics of the data. Unsupervised metric learning algorithms have achieved good results in cases where there is not available much information about the data. These algorithms do not require class label information, they are applied to improve unsupervised machine learning

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 4, Octubre-Diciembre, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 43- 67

methods, mainly for improving clustering results. Here we will mention some of the recent works done in this area, their advantages, disadvantages and applications.

Keywords: *unsupervised metric learning, dimensionality reduction, distance*

Introducción

El Aprendizaje Automatizado (*Machine Learning*; ML) es una rama de la inteligencia artificial, en gran parte inspirada en el razonamiento humano, que comprende el aprendizaje a partir de experiencia (Sammut y Webb 2011). El aprendizaje automático aborda, a su vez, una serie de problemáticas que tributan a problemas específicos, entre ellos: los problemas de clasificación, asociación, agrupamiento, y selección de rasgos (Sammut y Webb 2011). En el agrupamiento se parte de un conjunto de ejemplos el cual se desea organizar en grupos usualmente de acuerdo a una noción de similitud que generalmente es determinada por una función o métrica de distancia. La proximidad entre ejemplos determina la pertenencia o no a un grupo; por tanto, se estima que un elemento será más similar o tendrá mayores propiedades en común con los elementos de su grupo, que con respecto a los elementos de un grupo diferente.

Muchos de los métodos de aprendizaje automático dependen del cálculo de distancias para estimar la similitud entre dos ejemplos teniendo en cuenta la estructura de los datos. Este es el caso, por ejemplo, del algoritmo de los k vecinos más cercanos (k-*Nearest Neighbor*; kNN) para la comparación de las instancias entrantes con los datos conocidos (ejemplos de entrenamiento) o el k-medias (k-*means*) para calcular la distancia entre los objetos y su centro más cercano, entre otros. La aplicación $D: X \times X \rightarrow \Re^{+}$ sobre un espacio X se denomina métrica si $\forall\, x_i, x_j, x_k \in X$ se satisfacen las propiedades :

$D(x_i, x_j) + D(x_j, x_k) \geq D(x_i, x_k)$ (Desigualdad triangular)

$D(x_i, x_j) \geq 0$ (No negatividad)

$D(x_i, x_j) = D(x_j, x_i)$ (Simetría)

$D(x_i, x_j) = 0 \Leftrightarrow x_i = x_j$ (Distinguibilidad)

Si solo satisface las tres primeras condiciones, entonces la función es llamada pseudométrica. Si bien existen varias métricas generales como la similitud coseno, la distancia de Levenshtein, la distancia Euclidiana (Deza y Deza 2009), etc, la distancia Euclidiana es la más utilizada por los algoritmos de aprendizaje automatizado por su simplicidad y propiedades de generalización; sin embargo, esta distancia ignora cualquier tipo de regularidad estadística que pueda ser estimada a partir del conjunto de datos. Es posible transformar una métrica, y obtener una familia de métricas sobre

un espacio *X* calculando la distancia Euclidiana después de aplicar una transformación lineal sobre las instancias de entrada. Diversos autores han planteado que el uso de métricas que responden específicamente a las características de los datos influye positivamente en el desempeño de algoritmos basados en distancias (Yang y Jin 2006; Bellet, Habrard y Sebban 2013; Kulis 2012).

El aprendizaje de métricas de distancia (*metric learning*) tiene como principio la adaptación de una métrica de distancia a una aplicación específica, utilizando para ello información del conjunto de entrenamiento, como por ejemplo, modificar la métrica de distancia de un *k*NN para implementar la distancia de Mahalanobis (Bellet, Habrard y Sebban 2013), donde el objetivo final es inducir una métrica de distancia más potente a partir de los datos conocidos (Bellet, Habrard y Sebban 2013). Mahalanobis determina la similitud entre dos variables aleatorias multidimensionales y a diferencia de la distancia Euclidiana, tiene en cuenta la correlación entre las variables aleatorias. De ahí que gran parte de la rama del aprendizaje supervisado de métricas se basa en el aprendizaje de la métrica de Mahalanobis, el cual constituye la instancia más simple del problema de aprendizaje de distancias.

Los métodos de aprendizaje de métricas de distancia pueden clasificarse teniendo en cuenta varios criterios, uno de los más comunes es atendiendo a la disponibilidad de información (Bellet, Habrard y Sebban 2013). De acuerdo a esta clasificación pueden dividirse en tres categorías principales: los métodos que aprenden funciones de distancia de manera supervisada (*supervised metric learning*), los llamados métodos de aprendizaje de distancia semi-supervisados que son aquellos casos donde se conoce cierta cantidad de información y existe cierta supervisión en el proceso de aprendizaje (Bellet, Habrard y Sebban 2013) y aquellos que aprenden funciones de distancia de manera no supervisada (*unsupervised metric learning*) (Wang y Sun 2012). En este artículo abordaremos las características principales de los métodos de aprendizaje no supervisado de distancias.

Uno de los principales retos al trabajar con aprendizaje de distancias, lo constituye la alta dimensionalidad de los datos. Ciertos métodos de *manifold learning*[1] al aprender la distribución intrínseca de los datos han sido tratados como casos del aprendizaje no supervisado de métricas de distancias como es el caso del Análisis de Componentes Principales (*Principal Component Analysis*; PCA) (Abdi y Williams 2010), el escalado multidimensional (*Multidimensional Scaling*; MDS) (Torgerson 1952), y métodos no lineales como: el mapeo isométrico (ISOMAP) (Tenenbaum, Silva y Langford 2000), embebido local lineal (*Locally Linear Embedding*; LLE) y el mapeo de valores propios de Laplace (*Laplacian Eigenmaps*; LE) (Belkin y Niyogi 2001). Estos métodos de *manifold learning* no están sujetos a la información adicional en forma de restricciones que utilizan los métodos de aprendizaje de distancia, ya que obtienen

[1] Encuentra una representación de una dimensión *d* para representar datos de dimensión *D*, siendo *d*<*D*.

la información que necesitan a partir de los propios datos y la dimensión donde se encuentran embebidos. Resulta de gran importancia conocer las características principales de los métodos no supervisados para el aprendizaje de distancias, así como sus ventajas y desventajas, ya que estos pueden influir significativamente en la calidad de los métodos de agrupamiento basados en distancias. De ahí que el objetivo de este trabajo consiste en realizar una breve descripción del aprendizaje de métricas de distancias y sus clasificaciones; haciendo énfasis en las propiedades específicas de los métodos no supervisados de aprendizaje de distancias, sus ventajas, desventajas y dominio de aplicación.

Desarrollo

Aprendizaje de métricas de distancias

En 2002, utilizando información adicional para el agrupamiento, los autores del artículo (Xing et al. 2002) formalizaron el problema del aprendizaje de distancias como un problema de optimización convexo. El objetivo del aprendizaje de distancias es adaptar una función de distancia basada en pares y evaluada en los reales a un problema específico utilizando la información proporcionada por ejemplos de entrenamientos. Estos métodos, en esencia, intentan resolver un problema de optimización, estimando los parámetros de la métrica para adecuarse mejor a las características de los datos. La clasificación de los métodos de aprendizaje de distancia está profundamente relacionada con la disponibilidad de información. Muchos de estos métodos en lugar de tener acceso a instancias, etiquetas y clases, realizan el aprendizaje de la matriz positiva semidefinida M en d_M de una forma débilmente supervisada a través de información dada por conjuntos de restricciones (Bellet, Habrard y Sebban 2013). Estas restricciones, que caracterizan la similitud entre dos ejemplos, pueden clasificarse en dos grupos de restricciones: restricción por pares como se muestra en las expresiones (1) y (2) y restricción relativa como se muestra en la expresión (3).

Estas restricciones actúan como información de la clase, solo que en lugar de conocerse la etiqueta específica de la clase del objeto se tiene la información en forma de restricciones de similitud. En la expresión (1) se establece una relación de similitud directa o equivalente (*must-link*), donde, si los objetos x_i y x_j son similares entonces están contenidos en el mismo espacio inducido por la métrica, mientras que la expresión (2) representa aquellos objetos x_i y x_j que no son similares, y por tanto no deben estar incluidos en el mismo espacio (*cannot-link*). Es deseable que la función de distancia a utilizar facilite la cercanía de los pares similares, mientras que separe a los elementos que no son similares entre sí. Por ejemplo, en (Xing et al. 2002), la función de distancia es explícitamente aprendida para reducir al mínimo la distancia entre ejemplos de datos similares y maximizar la distancia entre puntos de datos disimilares.

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 4, Octubre-Diciembre, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 43- 67

Restricciones por pares (*must-link*, *cannot-link*):

$$\mathcal{S} = \{(x_i, x_j): x_i \text{ y } x_j \text{ son similares}\} \quad (1)$$

$$\mathcal{D} = \{(x_i, x_j): x_i \text{ y } x_j \text{ no son similares}\} \quad (2)$$

Restricción relativa (*training triplets*):

$$\mathcal{R} = \{(x_i, x_j, x_z): x_i \text{ es más parecido a } x_j \text{ que a } x_z\} \quad (3)$$

$$\min_M \ell(M, \mathcal{S}, \mathcal{D}, \mathcal{R}) + \lambda R(M) \quad (4)$$

El algoritmo de aprendizaje de distancias estima los parámetros de la métrica que se adapten al conjunto de restricciones, tal como ilustra la expresión (4). Este problema se puede modelar como un problema de optimización continuo donde el objetivo es minimizar la función de costo sobre los parámetros $\ell(M, \boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{D}}, \boldsymbol{\mathcal{R}})$, $\lambda R(M)$ actúa como regularizador de los parámetros de la matriz de distancias M (Bellet, Habrard y Sebban 2013). La función resultante de este proceso se utiliza para mejorar los algoritmos basados en métricas, como se muestra en la Figura 1.

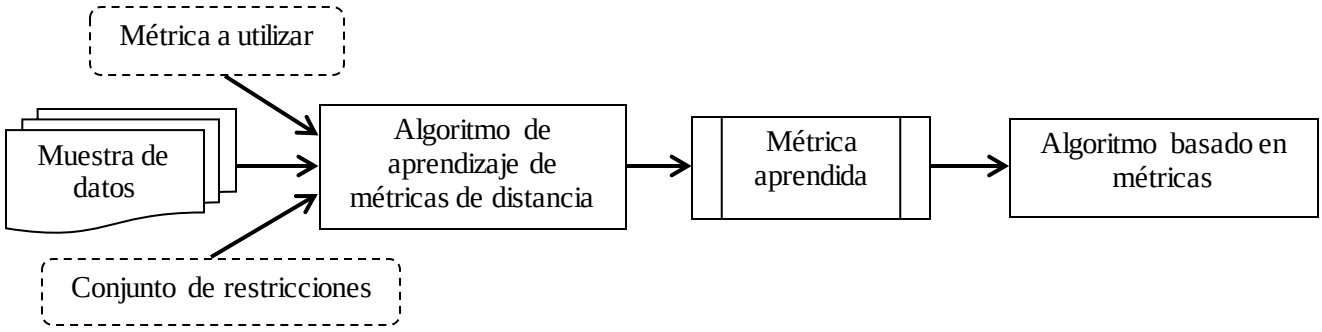

Figura 1. Algoritmos que utilizan el aprendizaje de métricas para mejorar su función de distancia.

El problema de la escalabilidad o redimensión es una problemática en todas las áreas de aprendizaje automático debido a la cantidad de datos disponibles que se incrementan rápidamente. El aprendizaje de distancias es interpretado como el aprendizaje de una matriz $d \times d$, esta clase de algoritmos usualmente es adaptable de acuerdo a la dimensión d de los datos, por lo cual una cualidad deseable de un algoritmo de aprendizaje de métricas es establecer la dimensión de acuerdo a la cantidad de ejemplos de entrenamiento n (o restricciones). Sin embargo, es difícil diseñar algoritmos que se redimensionen coherentemente de acuerdo a esta cantidad de datos (Bellet, Habrard y Sebban 2013).

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 4, Octubre-Diciembre, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 43- 67

Los algoritmos utilizados en el aprendizaje de distancias pueden clasificarse de acuerdo a propiedades tales como: paradigma de aprendizaje, forma de la métrica, carácter de la solución, si realiza o no reducción de dimensión, adaptabilidad, entre otros, como se muestra en la Figura 2. Estos factores determinan qué algoritmo utilizar según un problema específico.

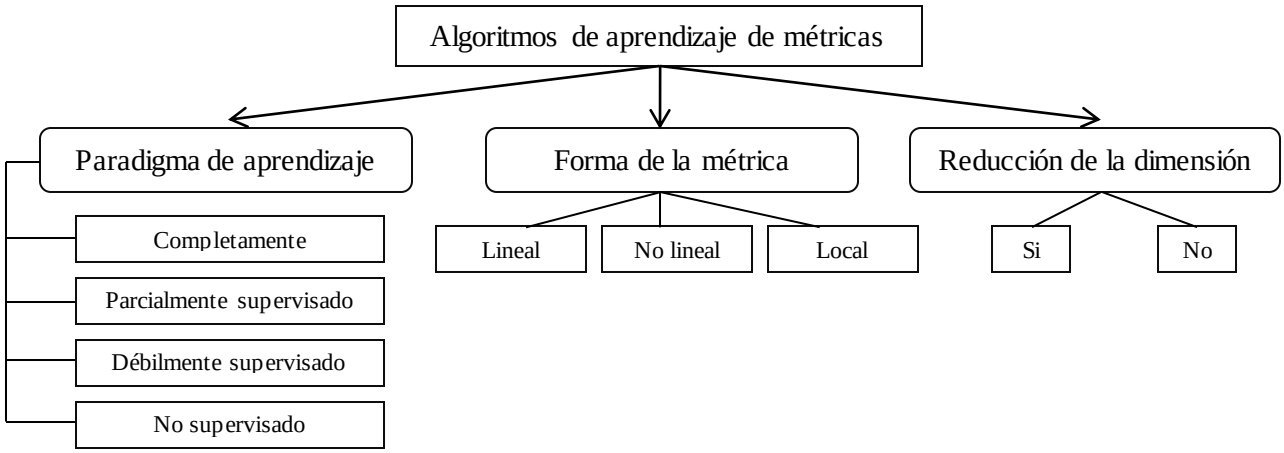

Figura 2. Clasificación de los algoritmos de aprendizaje de distancias.

Según el tipo de aprendizaje, estos algoritmos se clasifican en completamente supervisados, parcialmente supervisados, débilmente supervisados y no supervisados. A diferencia de la mayoría de los algoritmos de aprendizaje supervisado donde las instancias de entrenamiento son etiquetadas a partir de sus clases, en los algoritmos supervisados de aprendizaje de funciones de distancia, las instancias de entrenamiento se convierten en conjuntos de restricciones. El algoritmo tiene acceso a un conjunto de instancias de entrenamiento etiquetadas $\{z_i = (x_i, y_i)\}_{i=1}^{n}$ donde cada ejemplo de entrenamiento $z_i \in \mathcal{Z} = X \times Y$ está compuesto por una instancia de $x_i \in X$ y una etiqueta o clase $y_i \in Y$. La información contenida en las etiquetas se utiliza para generar conjuntos de pares o ternas de restricciones específicas S, D, R, para una instancia basada en un criterio de vecindad (Yang, Huang et al. 2013). En el aprendizaje débilmente supervisado de métricas de distancia el algoritmo no tiene acceso a las etiquetas de las instancias de entrenamiento individuales y en el aprendizaje parcialmente supervisado o semi-supervisado se tiene acceso a un gran conjunto de instancias no etiquetadas acerca de las cuales no existe suficiente información disponible, en estos dos casos solo se conoce la información en forma de conjuntos de restricciones S, D, R (esta información es proporcionada por los datos de manera indirecta, por ejemplo la retroalimentación implícita del usuario al hacer clic en un motor de búsqueda). Los métodos no supervisados solamente operan con los conjuntos de instancias no etiquetadas.

No solo es importante atender a la manera en la que se realiza el aprendizaje para decidir qué método utilizar de acuerdo al problema, la elección de la métrica es claramente un factor clave en el desempeño del método e influye considerablemente en los resultados (Bellet, Habrard y Sebban 2013). Métricas lineales como la distancia de Mahalanobis poseen un poder expresivo limitado pero son sencillas de optimizar ya que usualmente conllevan a formulaciones convexas y son menos propensas al sobre-ajuste, las métricas no lineales como la distancia cuadrática del histograma, a menudo resultan en formulaciones no convexas (sujetas a un óptimo local) pero resultan buenas capturando las variaciones no lineales en los datos, las métricas locales debido a la cantidad de parámetros que aprenden, son muy utilizadas, por ejemplo, en problemas de aprendizaje simultáneo.

La elección de la técnica de optimización también depende del tipo de problema. En (Cong, Pérez y Morell 2015) se mencionan varias técnicas de optimización populares para contextos específicos como: el gradiente descendiente (*gradient descent*) (Boyd and Vandenberghe 2009) para problemas de optimización con restricciones basadas en matrices, el gradiente proyectado (*projected gradient descent*) (Goldstein 1964) para modelos convexos en los que se desea preservar la convexidad, la proyección de Bregman (*Bregman proyections*) (Bregman 1967) y el gradiente descendiente estocástico (*stochastic gradient descent*) (Bottou 1998) para realizar cambios en las restricciones modificando una única restricción en cada iteración (lo cual es provechoso en casos en los que resulta costoso calcular todo el gradiente de la función de costo debido a la cantidad de restricciones), entre otros. Aunque también puede darse el caso de que los autores utilicen una técnica de optimización personalizada para modelar su problema individualmente.

La optimización, en el aprendizaje de distancias, también puede efectuarse a través de la descomposición de valores propios. Las técnicas de optimización basadas en este principio son utilizadas generalmente para descubrir las transformaciones lineales del espacio de entrada y son usadas en métodos como PCA, MDS, análisis discriminante lineal (*Linear Discriminant Analysis*; LDA) (Scholkopft and Mullert 1999), y análisis de componentes relevantes (*Relevant Component Analysis*; RCA) (Shental, Hertz et al. 2002).

Principales resultados de la revisión bibliográfica

Para comprender realmente en qué situación se encuentra el aprendizaje no supervisado de distancias, se realizó una búsqueda prestando particular interés en los últimos cinco años. La mayor parte de los datos se extrajo de las búsquedas realizadas en el sistema de indexado de SCOPUS. En la Figura 3 se muestran los resultados obtenidos de la búsqueda de los términos: "*unsupervised metric-learning*" en los campos título (TITLE), resumen (ABS) o palabras claves (KEY) a partir del año 2005 hasta mayo de 2016, en la base de datos SCOPUS.

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 4, Octubre-Diciembre, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 43- 67

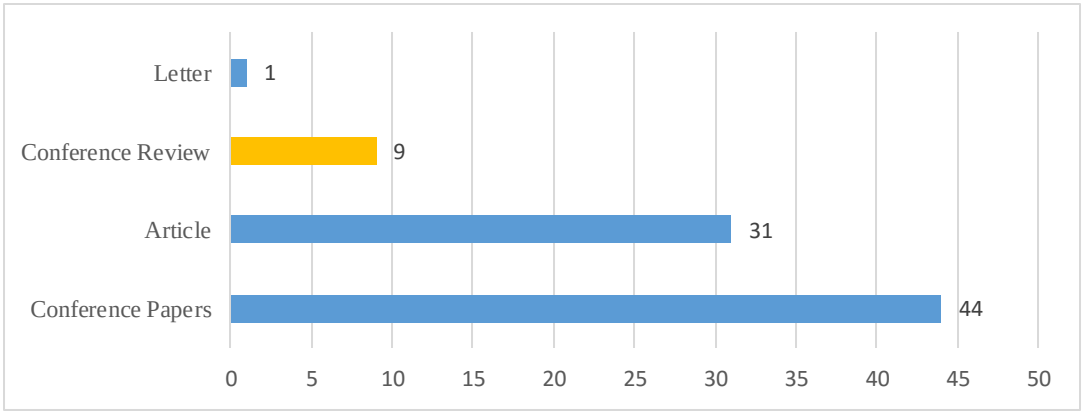

Figura 3. Resultados de la búsqueda TITLE-ABS-KEY (unsupervised metric-learning) AND PUBYEAR > 2005 en SCOPUS atendiendo al tipo de publicación.

Las consultas realizadas fueron:

TITLE-ABS-KEY (unsupervised metric-learning) AND PUBYEAR > 2005

KEY (unsupervised metric-learning) AND DOCTYPE (ar OR re) AND PUBYEAR > 2010

TITLE-ABS-KEY (unsupervised- metric-learning) AND PUBYEAR > 2005

Se realizaron también búsquedas en otras bases de datos, enfocando la búsqueda en los términos "*unsupervised metric-learning*" dedicando principal interés a los documentos de tipo *review* o *survey*. Los resultados obtenidos se analizaron para filtrar aquellos que no estuviesen directamente relacionados con el tema y sin embargo aparecieron en la búsqueda por hacer mención a los términos utilizados. En la Tabla 1 se recogen los principales datos de los artículos de revisión que abordan el tema del aprendizaje no supervisado de métricas de distancia.

Tabla 1. Artículos de revisión que abordan el tema del aprendizaje no supervisado de métricas de distancia.

Autores	**Título**	**Año**	**Publicación**	**Afiliación**
Yang, L., & Jin, R	Distance Metric Learning: a comprehensive survey	2006	Technical report, arXiv	Department of Computer Science and Engineering, Michigan State University
Yang, L.	An overview in Distance Metric Learning	2008	Proceedings of the Computer Vision and Pattern Recognition Conference	Visual Information Processing and Learning (VIPL). China
Kulis, B.	Metric learning: A survey	2012	Journal of Foundations and Trends in Machine Learning	Department of Computer Science, Boston Univ.

Cao, Qiong and Ying, Yiming and Li, Peng	Distance metric learning revisited	2012	European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)	College of Engineering, Mathematics and Physical Sciences (CEMPS), University of Exeter
Wang F., Sun J.	Distance Metric Learning in Data Mining	2012	2012 International Conference on Data Mining of the Society for Industrial and Applied Mathematics (SIAM)	IBM TJ Watson Research Center
Wang F., Sun J.	Survey on distance metric learning and dimensionality reduction in data mining	2014	Journal of Data Mining and Knowledge Discovery	IBM T. J. Watson Research Center; School of Computational Science and Engineering, College of Computing, Georgia Institute of Technology

Como se observa en la Tabla 1 y en la Figura 3, el aprendizaje de métricas de distancias está siendo investigado en la actualidad. La mayoría de los artículos de revisión sobre el tema hacen énfasis en los métodos que siguen un aprendizaje completamente supervisado de la métrica de distancia. Sin embargo, el aprendizaje de métricas resulta muy importante para obtener buenos resultados en los métodos de agrupamiento sobre datos no etiquetados. En tal caso no es posible aplicar métodos completamente supervisados. Por tanto, es de interés en este trabajo de revisión hacer énfasis en aquellos métodos que realizan el aprendizaje de distancias de manera no supervisada; es decir, aquellos métodos que parten de datos que no están etiquetados y de los cuales no se cuenta con información adicional.

Principales métodos para el aprendizaje no supervisado de distancias

Muchos algoritmos no supervisados para la reducción de dimensionalidad realizan un aprendizaje no supervisado de métricas de distancias utilizando información de los propios datos o de la dimensión donde se encuentran representados. Este grupo de métodos se pueden clasificar en métodos no lineales y lineales. Los algoritmos de reducción no lineales consideran que cada uno de los datos de alta dimensionalidad puede ser descrito a través de una función compuesta por los parámetros más relevantes y los datos son vistos como extractos de una dimensión subyacente embebida en la dimensión original del espacio. El objetivo es embeber datos que originalmente se encuentran en una dimensión en otra dimensión reducida, al mismo tiempo que se preservan las características principales de los datos. Para cada espacio dimensional debe existir intrínsecamente un espacio reducido; y por tanto, es posible acceder a los datos reducidos a través de algoritmos que interpreten o preserven la naturaleza de los datos embebidos (Cayton 2005). Entre los métodos más utilizados de este tipo se encuentran ISOMAP, el cual busca un sub-espacio que preserve mejor las distancias

geodésicas entre dos puntos de datos y los métodos LLE y LE, que se enfocan en la preservación de las estructuras de las vecindades locales.

PCA haya el sub-espacio que mejor preserve la varianza de los datos, MDS encuentra la proyección que mejor preserve la distancia de punto a punto dada por la matriz de distancias, el análisis de componentes independientes (*Independent Component Analysis*; ICA) (Langlois, Chartier y Gosselin 2010) busca una transformación lineal, con el objetivo de maximizar la independencia estadística de los datos, lo cual puede resultar útil en la interpretación de impulsos eléctricos para electroencefalogramas (Vega-Hernández y Valdés-Sosa 2009) y análisis de series de tiempo (González-Piedra 2011), entre otras. Otro ejemplo de métodos no supervisados para la reducción de dimensión que pueden aprender distancias de forma no supervisada son la preservación de proyecciones locales (*Locality Preserving Projections*; LPP) (Niyogi 2004) y la preservación de vecindades embebidas (*Neighborhood Preserving Embedding*; NPE) (He et al. 2005). Estos dos métodos (aproximaciones lineales a LE y LLE) pueden realizar aprendizaje supervisado de distancias si la información de la etiqueta es utilizada para construir la matriz de pesos.

En ocasiones, los datos de alta dimensionalidad se encuentran representados en espacios o variedades complejas (*manifolds*). Puntos que se encuentran alejados en la variedad donde se encuentran los datos reducidos, a primera vista, en el espacio de alta dimensionalidad original podrían parecer cercanos, como se muestra en la Figura 4 (A), y ser tratados de esta manera al utilizar medidas de distancia que no captan irregularidades de este tipo (como la distancia Euclidiana). El método ISOMAP busca un espacio reducido embebido en el espacio original que mantenga las distancias geodésicas[2] entre todos los puntos de coordenadas, con lo cual consigue caracterizar las vecindades presentes en la variedad. El uso de la distancia geodésica resulta mucho más expresivo y captura la distribución real de los datos, como es posible apreciar en la Figura 4 (B).

[2] La distancia de mínima longitud que une dos puntos en una superficie dada, y está contenida en esta superficie; por tanto, la línea más recta posible.

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 4, Octubre-Diciembre, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 43- 67

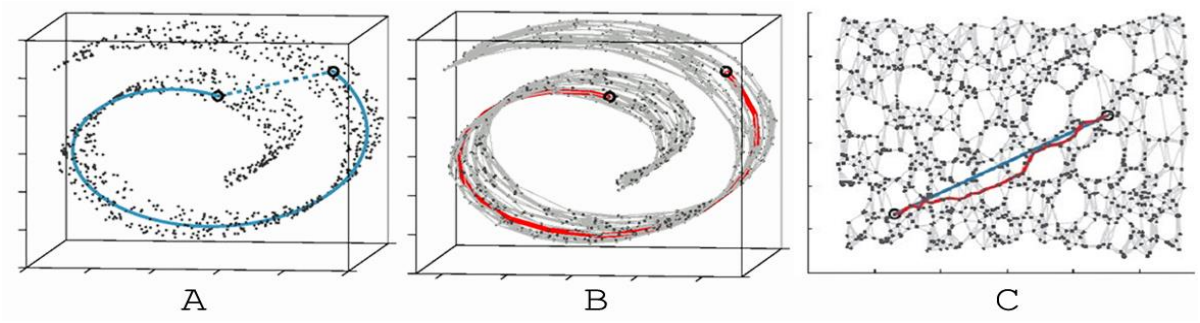

Figura 4. ISOMAP explorando los caminos geodésicos en el conjunto de datos Swiss roll para relacionar los objetos luego de realizar la reducción de dimensión. En A se localizan puntos representados en la variedad que son aparentemente cercanos, si se utiliza una medida de distancia que no capte las irregularidades de la variedad, como la distancia Euclidiana. En B se calcula la distancia real entre los puntos, utilizando las distancias geodésicas. En C se representan los puntos luego de la reducción de dimensión. Fuente: (Tenenbaum, Silva y Langford 2000).

ISOMAP

Entrada: $x_1, \ldots, x_n \in \mathbb{R}^D, k$

1. Crear el grafo de k vecinos más cercanos, ponderado con las distancias euclideanas $W_{ij} = \|x_i - x_j\|$, donde x_i y x_j son vecinos.
2. Aplicar un algoritmo de camino mas cercano como Djikstra o Ford y guardar las distancias cuadradas en W.

Salida: $Y \coloneqq cMDS(W)$

Algoritmo 1. Algoritmo ISOMAP. Fuente: (Cayton 2005).

El método ISOMAP establece relaciones de vecindad en la variedad basado en evaluaciones de las distancias geodésicas en las entradas y luego busca una representación Euclidiana, exacta o aproximada, que coincida con las evaluaciones geodésicas previas. ISOMAP comienza estimando las distancias geodésicas entre los puntos en las entradas utilizando las distancias más cercanas en el grafo de los vecinos más cercanos del conjunto de datos. Para ello, construye un grafo ponderado de los vecinos más cercanos utilizando la distancia Euclidiana y lo recorre utilizando un algoritmo para calcular el camino mínimo (Dijkstra o Ford), produciendo como salida las distancias geodésicas.

El algoritmo ISOMAP consta de dos etapas (Tenenbaum, Silva y Langford 2000):

- En la primera se establecen las relaciones de vecindad existentes en el sub-espacio variedad $\mathcal{M}$ basándose en las distancias Euclidianas $d_{\mathbf{X}}(i,j)$ obtenidas en el espacio original de entrada X. Las relaciones halladas se representan en un grafo ponderado $\mathcal{G}$, donde los pesos $d_{\mathrm{X}}(i,j)$ son asignados a los correspondientes bordes, como es posible observar en la Figura 4 (B).
- Luego se estiman los pares $d_{\mathcal{M}}(i,j)$ en $\mathcal{M}$ y las distancias geodésicas son definidas como las menores distancias

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 4, Octubre-Diciembre, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 43- 67

$d_G(i,j)$ en $\mathcal{G}$.

MDS

Entrada: $W \in R^{n \times n}$

1. Calcular la matriz de distancias B siguiendo la expresión $B = X^T X = -\frac{1}{2} HWH$, donde H es la matriz centrada y se calcula según la siguiente expresión $H = I - \frac{1}{N} 11^T$ y I $\in R^m$ es un vector.
2. Descomponer $B = V^{mds} \Lambda^{mds} (V^{mds})^T$ de forma tal que $[V^{mds}, \Lambda^{mds}] = eig(B)$, donde la matriz de coordenadas $X = V^{mds} (\Lambda^{mds})^{\frac{1}{2}}$.
3. Extraer las m proyecciones de Y cercanas a X en un rango de mayor a menor donde $Y = V_m (\Lambda^{mds})^{\frac{1}{2}}$. V_m^{mds} son los primeros vectores propios y Λ_m^{mds} los primeros valores propios.

Salida: X

Algoritmo 2. Representación de los pasos de MDS. Fuente: (Cayton 2005).

Una vez que se dispone de las distancias geodésicas que preservan la naturaleza de los datos a través de las rutas geodésicas antes representadas con curvas en la variedad, estas se transforman en líneas rectas en la dimensión Y. ISOMAP realiza este proceso aplicando el método MDS a la matriz de distancias geodésicas D_G. MDS busca una representación Euclidiana, exacta o aproximada D_G, donde $[D_G]_{ij} = d_G(i,j)$ en un espacio euclidiano Y de dimensión m que preserve la geometría intrínseca de $\mathcal{M}$, como se observa en la Figura 4 (C).

De los métodos que realizan reducción de dimensión basados en geometría, ISOMAP es uno de los más utilizados. Su popularidad viene dada por la expresividad que ofrece la información topológica a partir de las distancias geodésicas. Otros métodos de reducción de dimensión; sin embargo, se enfocan en la preservación de las estructuras locales de vecindad, tal es el caso de LLE. El método LLE visualiza la variedad como una colección de parches, o puntos de coordenadas que se solapan entre sí. Si la variedad es lo suficientemente uniforme y las vecindades son pequeñas, entonces el algoritmo considera estas zonas $\mathbb{R}^D$ como lineales. El objetivo es identificar cada uno de esos parches y caracterizar la geometría interna en ellos, para luego construir las vecindades (Castellanos Domínguez et al. 2011).

El principio de LLE es preservar la relación de orden local de los datos tanto en el espacio embebido como en el espacio original. Cada muestra x_i en el espacio de observación se intenta representar como una combinación ponderada y convexa de sus vecinos más cercanos. La matriz de pesos W se utiliza como sustituto de la geometría local de los parches, donde W_i representa la distribución de los puntos alrededor de x_i. La reconstrucción de los pesos se representa en el paso 1a del Algoritmo 3, donde C es la matriz de covarianza local y W_i es la caracterización de la geometría local que rodea al punto x_i en la variedad. En el segundo paso el algoritmo calcula una configuración en la dimensión reducida d.

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 4, Octubre-Diciembre, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 43- 67

Locally Linear Embedding

Entrada: $x_1, \ldots, x_n \in \mathbb{R}^D, d, k$

1. Calcular la reconstrucción de pesos para cada punto x_i:
   a. $W_i := \frac{\sum_k c_{jk}^{-1}}{\sum_{lm} c_{lm}^{-1}}$
2. Calcular el embebido en la dimensión reducida:
   a. Sea U una matriz donde las columnas son los vectores propios de $(I - W)\intercal(I - W)$
   b. $Y := [U]_{n \times d}$

Salida: Y

Algoritmo 3: Pasos del algoritmo LLE. Fuente: (Roweis, Saul y Roweis 2008).

Tanto ISOMAP como LLE, tienen como parámetro el número de vecinos k. Siendo $N(i)$ el conjunto de vecinos más cercanos a x_i, los pesos se seleccionan procurando minimizar el error cuadrático para cada i: $\left\| x_i - \sum_{j \in N(i)} W_{ij} x_j \right\|^2$ (Cayton 2005). LLE preserva las vecindades locales entre los objetos en la variedad de alta dimensionalidad, y a su vez conserva esta estructura en la representación reducida (Goldberg et al. 2008). En la Figura 5 se aprecia en código de colores para representar el resultado de utilizar LLE en el conjunto de datos Swiss Roll.

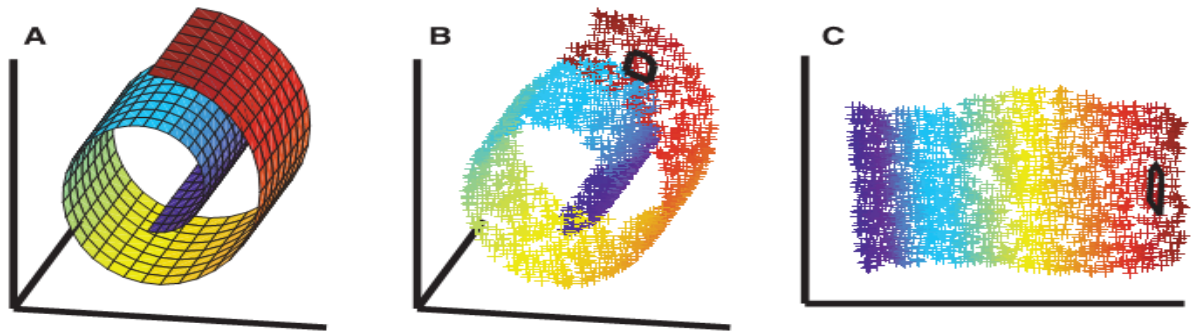

Figura 5. Resultado de utilizar LLE en el conjunto de datos Swiss Roll. En A se reconocen las coordenadas internas embebidas en la variedad, los colores representan la preservación de las vecindades locales, y la región marcada en B muestra la vecindad para un punto específico. En C puede verse como la vecindad es preservada después de realizar la reducción de dimensión. Fuente: (Tenenbaum, Silva y Langford 2000).

LLE ha sido muy utilizado en problemas con alta dimensionalidad, en los que se desean preservar las vecindades locales de los objetos (Tang et al. 2014; Liu et al. 2013; Karbauskait, Kurasova y Dzemyda 2015; Ziegelmeier, Kirby y Peterson 2012; Zhang y Wang 2006). Técnicas como ISOMAP y LLE se encuentran definidas solamente partiendo de los datos de entrenamiento; sin embargo, LLP (He y Niyogi 2003) puede ser aplicado a cualquier punto en el espacio de representación reducido. LLP construye un grafo incorporando información de las vecindades del conjunto de datos.

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 4, Octubre-Diciembre, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 43- 67

LLP halla la aproximación lineal óptima a las funciones propias del operador Laplace Beltrami en la variedad (el cual está dado por la relación de adyacencia entre puntos).

Locality Preserving Projections
Entrada: $x_1, \dots, x_n \in \mathbb{R}^D$ 1. Construir el grafo de adyacencia y seleccionar los pesos: $\omega_{ij} = exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$ si x_i está en la vecindad de x_j o viceversa 2. Calcular vectores propios $XLX^T a = \lambda XDX^T a$ 3. Embeber datos en la dimensión a reducir $\boldsymbol{x}_i \rightarrow \boldsymbol{y}_i = A^T \boldsymbol{x}_i, A = (a_0, \dots, a_{l-1})$ Salida: Y

Algoritmo 4: Pasos del algoritmo LPP. Fuente: (Niyogi 2004).

El algoritmo LPP recibe como entrada $x_1, \dots, x_n \in \mathbb{R}^D$ y primeramente construye el grafo de adyacencia G donde m es el número de nodos. Dos nodos i y j están conectados si existe una relación de cercanía entre ellos. La relación de cercanía puede establecerse utilizando uno de los siguientes criterios:

a. Usando el método e-vecindades (ϵ-*neigborhood*) donde $\epsilon \in \mathbb{R}$, dos nodos están conectados si $\|x_i - x_j\|^2 < \epsilon$ donde se utiliza la norma Euclidiana en $\mathbb{R}^n$.

b. Utilizando k-vecinos más cércanos ($k < N$) dos nodos i y j están conectados si i está entre los vecinos más cercanos de j o viceversa.

Posteriormente realiza la selección de pesos. Sea W una matriz simétrica dispersa de dimensión $m \times m$, donde W_{ij} representa el peso de la arista entre los nodos i y j. El valor del peso de la arista puede definirse de manera convencional, siendo 1 si los nodos están conectados y 0 si los nodos no están conectados; o bien utilizar una función núcleo (Belkin y Niyogi 2001) de forma tal que si dos nodos i y j están conectados entonces el peso de la arista se calcula de forma que $W_{ij} = e^{-\frac{\|x_i - x_j\|^2}{t}}$ donde $t \in \mathbb{R}$.

Una vez calculadas las vecindades, LPP procede al mapeo de la variedad calculando los vectores propios y los valores propios para el problema. Se calcula la expresión (5), donde D es una matriz diagonal tal que $D_{ii} = \sum_j W_{ij}$, $\boldsymbol{x}_i$ es el i-ésimo elemento de la matriz $\boldsymbol{X}$ y L es una matriz de Laplace, tal que $L = D - W$.

$$XLX^T a = \lambda XDX^T a \quad (5)$$

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 4, Octubre-Diciembre, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 43- 67

Suponiendo que $a_0, \ldots, a_{l-1}$ es el vector solución en la ecuación (5) y que este está ordenado de acuerdo a sus valores propios $\lambda_0 < \cdots < \lambda_{l-1}$, entonces el proceso de embebido sería $\boldsymbol{x}_i \rightarrow \boldsymbol{y}_i = A^T \boldsymbol{x}_i, A = (a_0, \ldots, a_{l-1})$, siendo $\boldsymbol{y}_i$ un vector de dimensión l y A una matriz de $n \times l$. El algoritmo simplificado LPP, a diferencia de ISOMAP y LLE, es lineal, lo cual facilita el trabajo con aplicaciones reales (Zhang, Qiao y Chen 2010). Para aprender una métrica en un entorno no supervisado la mayoría de los métodos de aprendizaje de distancia proyectan los datos observados en una variedad reducida donde las relaciones geométricas (como las distancias entre pares) sean preservadas. Este principio puede ser extendido al caso no lineal utilizando una función núcleo para mapear los datos. Basados en este principio, en (Wang, Zhao y Zhang 2011) se propuso un método no supervisado para maximizar las proyecciones de márgenes (*Unsupervised Maximun Margin Projections*; UMMP) (Wang, Zhao y Zhang 2011) asumiendo que el conjunto de datos está dividido en dos clases, y el objetivo es encontrar un hiperplano que maximice la distancia entre ambas, y se repite el procedimiento para cada grupo, tal como se aprecia en la Figura 6. Para cada posible etiquetado, se construye una máquina de vectores de soporte (*Support Vector Machine*; SVM) que maximice el margen entre las dos clases. UMPP permite buscar la etiqueta donde la SVM construida alcance el máximo valor de margen.

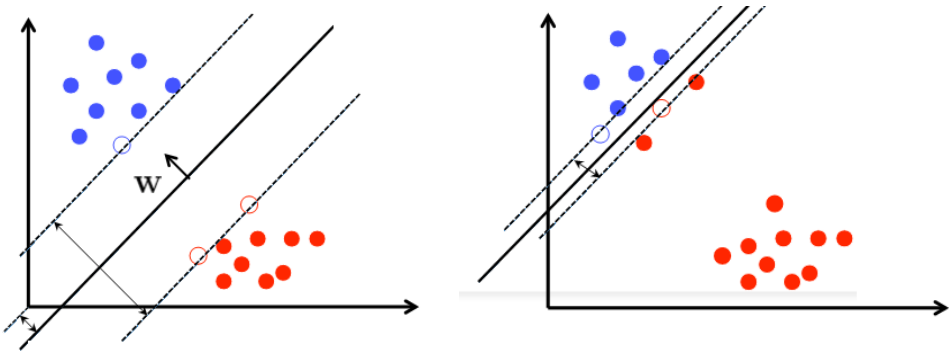

Figura 6. Visualización de UMMP. Fuente: (Wang y Sun 2012).

UMMP depende de la geometría del límite de decisión óptimo y no de la distribución de los puntos de datos alejados de este. Este algoritmo encuentra la proyección de los hiperplanos que maximice la separabilidad de los grupos. Inicialmente las etiquetas de las clases (para diferenciar los objetos) son asignadas arbitrariamente y se entrena una SVM con un margen por defecto. El objetivo es encontrar el etiquetado asociado al margen máximo obtenido por las SVM entrenadas en cada escenario de etiquetado posible.

UMMP busca direcciones en las cuales se maximice la separación entre los grupos, otro criterio para aprender la dirección de proyección es preservando la geometría de los datos en el espacio original de proyección. Si bien el método es no supervisado, en aplicaciones reales donde se conocen algunas restricciones a priori es posible incorporar objetivos

adicionales a la maximización del margen de proyecciones (*Maximun Margin Projections*; MMP). Leves modificaciones de este método han demostrado ser competitivas en aplicaciones reales como reconocimiento facial y clasificación de textos (Wang, Zhao y Zhang 2011).

En el contexto del aprendizaje de distancias comienza a verse la tendencia de aprovechar características específicas de los algoritmos existentes y adaptarlas a nuevos métodos. Métodos como SMLgb (*Sparse Metric Learning via Smooth Optimization*) (Ying, Huang y Campbell 2009) introducen el espectro disperso para aprender una matriz de datos de menor dimensión a la par que realizan la reducción de dimensión. SMLgb estima transformaciones lineales (equivalentes a la matriz de distancias) tales que combinen y retengan las ventajas de algoritmos de aprendizaje de distancias (Ying 2012). Este modelo tiene dos pautas fundamentales. La primera, la elección de una buena matriz de distancias M que preserve la estructura de distancia; es decir, la cercanía entre vecinos. La segunda, comprende la capacidad de la matriz de distancias de eliminar el ruido mientras conduce a la reducción de dimensión.

Jiang y Wang se inspiraron en los algoritmos de propagación de etiqueta y los mapas de difusión (Goldberg et al. 2008) para proponer un enfoque basado en difusión para mejorar una matriz de similitud (*Self-Smoothing Operator*; SSO) (Jiang y Wang 2011). El proceso de difusión propaga la masa de similitud en la variedad en la que están representados los datos. El resultado es el aprendizaje de una métrica global de similitud obtenida a través de la propagación de la similitud, donde esta propagación se realiza a través de un operador de auto suavizado.

Self-smoothing operator
Entrada: W, t 1. Calcular el núcleo de suavizado: $P = D^{-1}W$ donde D es una matriz diagonal con $D(i,i) = \sum_{k=1}^{n} W(i,k)$. 2. Realizar el suavizado t veces donde: $W_t = WP^t$. 3. Realizar una auto-normalización de la matriz de similitud $W^* = \Delta^{-1}W^t$ donde Δ es una matriz diagonal con $\Delta(i,i) = W_t(i,i)$. Salida: W^*

Algoritmo 5. Secuencia de pasos de SSO. Fuente: (Jiang y Wang 2011).

En SSO, se parte de un grafo $G = (\Omega, W)$ que representa los datos y la similitud entre ellos, donde $\Omega = \{x_i, i = 1, \dots, n\}$ es el espacio finito para nodos que representan ejemplos o datos, y W es una matriz de similitud donde $W(i,j) \in [0,1]$ representa la similitud entre x_i y x_j. Normalmente, esta matriz de similitud se obtiene de aplicar un núcleo gaussiano G a una matriz de distancia $W(i,j) = \exp\{-d^2(i,j)/k\sigma^2\}$, donde $d(i,j)$ representa la distancia entre x_i y x_j y la amplitud del núcleo se controla mediante k y σ. Un proceso de difusión estocástico en G permite la propagación de las similitudes locales a lo largo de la geometría de la variedad, sin tener que explícitamente construir la variedad. El núcleo de

suavizado P se induce a partir de la matriz de similitud W, expresado en un núcleo gaussiano de suavizado $P = D^{-1}W$ donde D es una matriz diagonal con $D(i,i) = \sum_{k=1}^{n} W(i,k)$. Este proceso de suavizado utiliza el núcleo t veces sobre la matriz de similitud W, de forma tal que $W_t = WP^T$, con el fin de garantizar que la diagonal de la matriz diagonal sea 1 y realiza un proceso de auto-normalización.

En realidad, el parámetro decisivo en SSO (si se cuenta con la matriz de similitud), lo constituye t. Una idea para comprender el funcionamiento del proceso de suavizado es utilizar una analogía con el procesado de imágenes con ruido. Aplicar un suavizado en una imagen con ruido, aumenta la razón de señal/ruido, lo que contribuye a reducir la información que resulta desconocida o difícil de interpretar debido al ruido. Sin embargo, si se suaviza demasiado la imagen, esto puede llevar a la pérdida de información relevante. En el contexto de SSO, la idea tras el parámetro t es similar al suavizado de imágenes con ruido; t debe ser un valor que permita la propagación de la similitud en la variedad sin corromper la información. Si bien la definición del valor apropiado de este parámetro no es una cuestión trivial (ya que de él depende en gran medida el método), los autores del método proponen un rango efectivo de $500 \sim 1000$.

A diferencia de otros métodos de propagación de etiqueta, o basados en consulta, SSO induce una métrica global que influye directamente en la calidad de la matriz de similitud, sin necesidad de introducir nociones adicionales de métricas de distancia. Este método ha tenido buenos resultados en áreas de recuperación de imágenes, agrupamiento, segmentación y clasificación (Jiang y Wang 2011).

Discusión

Los algoritmos de *manifold learning* no lineales para la reducción de dimensión extraen y aprenden la métrica del propio conjunto de datos, como un subproceso de la reducción de dimensionalidad, por ello, han sido considerados por muchos como métodos de aprendizaje no supervisado de distancias (Yang y Jin 2006; Kulis 2012; Wang y Sun 2015). A partir de algunos métodos de aprendizaje de distancia no supervisados y utilizando cierta información de los datos, se han modifican dichos métodos para obtener mejores resultados y con ello han surgido varios métodos híbridos (Wang et al. 2012; Fu 2014; Cinbis et al. 2011).

Uno de los métodos más completos y que responde a la esencia de la definición de aprendizaje no supervisado de distancia es SSO (Jiang y Wang 2011). Este método, a través de un proceso de propagación de la similitud utilizando un operador de suavizado, calcula una métrica de distancia. SSO ha obtenido buenos resultados en el área de

identificación de imágenes y video, y recuperación de imágenes. Sin embargo, uno de los puntos a mejorar sigue siendo la condición de la estimación del mejor valor para el parámetro que controla el operador de suavizado.

De los métodos de *manifold learning* mencionados, ISOMAP, LLE y LE son métodos no paramétricos, lo cual significa que la técnica no especifica un mapeo directo hacia el espacio reducido. Los métodos no paramétricos tienen como desventaja que no es posible generalizarlos para nuevos datos sin realizar nuevamente el proceso de reducción de dimensión. Otra desventaja es que no es posible delimitar cuanta información de la dimensión original es retenida en el espacio reducido al reconstruir los datos desde la dimensión reducida y midiendo el error entre los datos originales y los datos reconstruidos.

Las técnicas que utilizan el espectro disperso tienen varias desventajas identificables, una de ellas es un punto flaco en su función de costo. Estos métodos también se ven afectados por la maldición de la dimensionalidad, el número de datos que es requerido para caracterizar la variedad apropiadamente crece a medida que crece la dimensionalidad intrínseca de la variedad. Esta susceptibilidad a la dimensionalidad es una debilidad fundamental en los métodos de aprendizaje local. Otra de las susceptibilidades de este tipo de métodos según (Van Der Maaten, Postma y Van Den Herik 2009) es la predisposición al sobre entrenamiento (lo cual ha sido solucionado parcialmente con métodos adaptativos de vecindad o *e-neighboors*); la condición de linealidad local asume que las variedades no contienen discontinuidades y la sensibilidad al trabajar con variedades que no son isométricas a un espacio Euclidiano.

De los métodos de enfoque disperso, uno de los más populares lo constituye LLE, el cuál ha sido ampliamente aplicado en varias áreas (Ziegelmeier, Kirby y Peterson 2012; Yang, Xiang y Shi 2013; Liu et al. 2013). En los últimos años algunos autores han profundizado en la selección del parámetro de vecinos más cercanos obteniendo buenos resultados (Castellanos Domínguez et al. 2011; Karbauskait, Kurasova y Dzemyda 2015). Sin embargo, LLE es débil ante las irregularidades de la variedad al ser un método local, y debido a la simplicidad de la restricción de covarianza en su solución es propenso a redimensiones en la variedad en el proceso de embebido.

Si bien la distancia geodésica podría resultar útil para añadir expresividad a los datos y explorar mayor información en conjuntos de datos complejos; ISOMAP tiene desventajas, entre ellas, la inestabilidad topológica (Balasubramanian 2010) que provoca que construya conexiones erróneas en el grafo de vecindad, lo que podría afectar su desempeño, la presencia de espacios "vacíos" en la variedad, o la susceptibilidad a variedades no convexas. Sin embargo, varios autores han utilizado este método o variaciones del mismo debido a la expresividad que facilita la información topológica que proveen las distancias geodésicas (Hu, Lu y Xu 2012; Hauberg, Freifeld y Black 2012; Wang, Yuen y Feng 2012).

El método de aprendizaje adaptativo no lineal de métricas de distancia (*Non-linear adaptative metric learning*; NAML) (Chen et al. 2007) realiza agrupamiento y aprendizaje de distancias simultáneamente. Primero mapea los datos a un espacio de mayor dimensión a través de una función núcleo, y luego aplica una proyección lineal para encontrar una variedad de menor dimensión donde se maximice la separabilidad de los datos, en ese espacio es donde se realiza el agrupamiento. Este algoritmo ha tenido buenos resultados en comparación con otros métodos del estado del arte como LLE y LE. Los métodos de aprendizaje no supervisado de distancias en un espacio de núcleo compuesto (*Unsupervised distance metric learning in composite kernel space*; CKS-EWFC-K, CKS-EWFC-F) (Wang et al. 2016) se combinan en una plataforma de desarrollo de agrupamiento difuso y aprendizaje de métricas de distancia. Los algoritmos obtienen la función de distancia usada para el cálculo de la similitud a través de un proceso de aprendizaje no supervisado durante el proceso de agrupamiento del sub-espacio suavizado. Tanto NAML como los recientes CKS-EWFC-K y CKS-EWFC-F pueden adaptarse a los datos para aprender funciones de distancia acordes a los conjuntos de datos durante el proceso de agrupamiento. Sin embargo, aún se encuentran en fase de estudio experimental para el ajuste de los parámetros y las guías de identificación de los mismos.

Los métodos de aprendizaje no supervisado de métricas de distancias tienen variados campos de aplicación (Farenzena et al. 2010; Ma et al. 2012a; Liu et al. 2012). En (Farenzena et al. 2010) trabajaron en la acumulación de rasgos locales basado en simetría (*Symmetry-Driven Accumulation of Local Features*; SDALF) para explotar la propiedad de simetría en la identificación de imágenes de pederastas. En (Ma, Su y Jurie 2012a) desarrollaron un descriptor BiCov que utiliza filtros Gabor y un descriptor de varianza para manejar los cambios de iluminación y las variaciones del fondo en imágenes y utilizaron un vector de Fisher para codificar estadísticas de alto orden en características locales. En (Fu 2014), los autores reconstruyeron una métrica semántica latente a través de aprendizaje multi-vista para video. Este método multi-vista alcanza un balance entre la separación de los grupos y la similitud a las métricas originales, utilizando un algoritmo de optimización eficiente. Básicamente los autores realizan una combinación óptima de múltiples métricas, óptimo definido por el intercambio entre el margen máximo entre los grupos obtenidos usando la métrica y la similitud entre la métrica aprendida y las métricas originales.

Siguiendo la línea de reconocimiento facial y visión por computadora, en (Cinbis et al. 2011) realizan aprendizaje no supervisado de métricas de distancia para el reconocimiento facial en videos de TV (*Unsupervised face metric learning*). La identificación de rostros es un factor clave en televisión, sobre todo para programas de acople de subtítulos o guiones, una identificación precisa hace efectiva la transferencia supervisada de los datos dispersos basados en textos a otros rostros. En (Cinbis et al. 2011) se aprende una métrica de distancia que obtiene buenos resultados sin necesidad de etiquetar manualmente los ejemplos, este proceso de aprendizaje se realiza utilizando pares de rostros que aparecen

juntos en un fragmento como ejemplos positivos, y como ejemplos negativos, pares de rostros de personas diferentes que aparecen juntos en un fragmento de video. La métrica es aprendida sobre los pares seleccionados utilizando nueve descriptores faciales, y está orientada a la aparición de personajes específicos en un video.

Conclusiones

El aprendizaje de métricas de distancia resulta de gran importancia para mejorar los resultados de varios algoritmos de aprendizaje automático en la solución de disímiles problemas. Se ha trabajado intensamente en el desarrollo de métodos que realizan el aprendizaje completamente supervisado de métricas de distancia. Estos métodos requieren un conjunto de entrenamiento donde cada instancia esté etiquetada. Algunas veces resulta muy costoso obtener las etiquetas de las instancias, o debido a la naturaleza del problema, los objetos no tienen etiquetas asignadas, aunque sí se cuenta con un conjunto de restricciones que ofrecen información adicional de los datos. En tales casos se han desarrollado métodos de aprendizaje parcial y débilmente supervisado de métricas de distancia. Desafortunadamente, existen problemas para los cuales solo se cuenta con conjuntos de datos sin etiquetar y no se tiene información adicional. En estos casos es necesario aplicar métodos no supervisados para el aprendizaje de métricas. Estos métodos tienen gran importancia para mejorar la calidad de resultados de técnicas de aprendizaje no supervisado, por ejemplo, el agrupamiento. Es por ello que en este artículo de revisión se describieron los principales métodos de aprendizaje de métricas que permiten trabajar en tales condiciones, de ahí que se hizo énfasis en los métodos no supervisados para el aprendizaje de métricas.

Los métodos ISOMAP, LLE, LE, PCA, MDS, ICA, LPP, NPE son considerados métodos para la reducción de la dimensionalidad que realizan un aprendizaje no supervisado de métricas de distancias utilizando información de los propios datos o de la dimensión donde se encuentran representados. Estos métodos logran embeber datos que originalmente se encuentran en una dimensión en otra dimensión reducida, al mismo tiempo que se preservan las características principales de los datos. De estos métodos, uno de los más utilizados es ISOMAP, ya que busca un sub-espacio que preserve mejor las distancias geodésicas entre dos puntos de datos. El uso de la distancia geodésica resulta mucho más expresivo y captura la distribución real de los datos. Los métodos LLE y LE también han sido ampliamente utilizados porque se enfocan en la preservación de las estructuras de las vecindades locales. Tanto ISOMAP como LLE requieren que se les especifique como parámetro el número de vecinos a considerar y solo pueden aplicarse partiendo de los datos de entrenamiento. Si se desea partir de cualquier punto de un espacio de representación reducido, entonces se debe aplicar LLP, que construye un grafo incorporando información de las vecindades del conjunto de datos y permite

partir tanto del conjunto de entrenamiento original como de un espacio reducido. Además, LLP, a diferencia de ISOMAP y LLE, es lineal, por lo cual se sugiere su uso en el trabajo con aplicaciones reales.

SSO es otro método relevante en el aprendizaje de métricas de distancias y que no requiere contar con información adicional de los datos. SSO parte de un grafo que representa los datos y la similitud entre ellos, e induce una métrica global que influye directamente en la calidad de la matriz de similitud, sin necesidad de introducir nociones adicionales de métricas de distancia. Se sugiere utilizar este método en el aprendizaje de métricas para tributar después a un mejor resultado de técnicas de agrupamiento, segmentación y clasificación, aunque su utilidad mayor es en el agrupamiento por no necesitar información adicional de los datos.

Indudablemente queda mucho trabajo por hacer, y esta es un área relativamente joven del aprendizaje de distancias y aunque se han hecho avances significativos, el objetivo a perseguir sigue siendo crear métodos de aprendizaje de distancias que sean menos dependientes de información adicional y puedan extraer la información necesaria para aprender la métrica del conjunto de datos.

Referencias

ABDI, H. y WILLIAMS, L.J., Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2010, vol. 2, no. 4, pp. 433-459.

AMORES, J., SEBE, N. y RADEVA, P. Boosting the distance estimation: Application to the K-Nearest Neighbor classifier. *Pattern Recognition Letters*. 2006., vol. 27, no. 3, p. 201-209.

BALASUBRAMANIAN, M. The Isomap Algorithm and Topological Stability. *Science*, 2002, vol. 295, no. 5552, p. 7-7.

BELKIN, M. y NIYOGI, P., Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. En 2001 *Advances in Neural Information Processing Systems (NIPS)*, 2001, p. 585-591.

BELLET, A., HABRARD, A. y SEBBAN, M., A Survey on Metric Learning for Feature Vectors and Structured Data, *arXiv 1306.6709*, 2013, p. 57.

CASTELLANOS DOMÍNGUEZ, G., ÁLVAREZ-MESA, A., VALENCIA-AGUIRRE, J. y DAZA-SANTACOLOMA, G., Global and Local Choice of the Number of Nearest Neighbors in Locally Linear Embedding. 2011. *Pattern Recognition Letters*, vol. 32, no. 16, p. 2171-2177.

CAYTON, L., Algorithms for manifold learning. *Univ. of California at San Diego Tech. Rep*, 2005, p. 1-17.

CHEN, J., ZHAO, Z., YE, J. y LIU, H., Adaptive Distance Metric Learning for Clustering. En 2007 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), IEEE, 2007, p. 1-7.

CINBIS, R.G., VERBEEK, J., SCHMID, C. y KUNTZMANN, L.J., Unsupervised Metric Learning for Face Identification in TV Video. En 2011 *International Conference on Computer Vision*, IEEE, 2011, p. 1559–1566.

CONG, B.N., PÉREZ, J.L.R. y MORELL, C. Aprendizaje supervisado de funciones de distancia: estado del arte. *Revista Cubana de Ciencias Informáticas*, 2015, vol. 9, no. 2, p. 14-28.

DEZA, M.M. y DEZA, E. *Encyclopedia of distances*. Springer Berlin Heidelberg. 2009, p. 1-583.

FARENZENA, M., BAZZANI, L., PERINA, A., MURINO, V. y CRISTANI, M., Person re-identification by symmetry-driven accumulation of local features. En 2010 *Conference on Computer Vision and Pattern Recognition (CVPR), IEEE*, 2010, p. 2360-2367.

FU, Y., Multi-view Metric Learning for Multi-view Video Summarization. *arXiv1405.6434*. 2014.

GOLDBERG, Y., ZAKAI, A., KUSHNIR, D. y RITOV, Y., Manifold Learning: The Price of Normalization. *Journal of Machine Learning Research*, 2008, vol. 9 no. Aug, p. 1909–1939.

GONZÁLEZ-PIEDRA, E. *Independent Component Analysis for Time Series*. Tesis Doctoral. Universidad Carlos III de Madrid, 2011.

HAUBERG, S., FREIFELD, O. y BLACK, M.J., A geometric take on metric learning. En 2012 *Advances in Neural Information Processing Systems (NIPS)*, p. 2024-2032.

HE, X., CAI, D., YAN, S. y ZHANG, H.-J., Neighborhood preserving embedding. En *Tenth IEEE International Conference on Computer Vision*, 2005, p. 1208-1213.

HE, X. y NIYOGI, P., Locality Preserving Projections. En 2004 *Advances in Neural Information Processing Systems (NIPS)*, MIT, vol 16, p. 153.

HU, Z.-P., LU, L. y XU, C.-Q., L1 Norm Sparse Distance Metric Learning for One-class Classifier [J]. *Acta Electronica Sinica*, 2012, vol. 1, p. 23.

JIANG, J. y WANG, B., Unsupervised Metric Learning by Self-Smoothing Operator. En *2011 International Conference on Computer Vision*. IEEE, 2011. p. 794-801.

KARBAUSKAIT, R., KURASOVA, O. y DZEMYDA, G., Selection of the number of neighbors of each data point for the locally linear embedding algorithm. *Information Technology and Control*, 2015, vol. 36, no. 4.

KULIS, B., Metric Learning : A Survey. *Foundations and Trends in Machine Learning*, 2012 vol. 5, no. 4, p. 287–364.

LANGLOIS, D., CHARTIER, S. y GOSSELIN, D., An Introduction to Independent Component Analysis : InfoMax and FastICA Algorithms. *Tutorials in Quantitative Methods for Psychology*. 2010, vol. 6, no. 1, p. 31-38.

LIU, C., GONG, S., LOY, C. y LIN, X., Person re-identification: What features are important? En 2012 European Conference on Computer Vision (*ECCV)*, Berlín, p. 341-401.

LIU, X., TOSUN, D., WEINER, M.W., SCHUFF, N., INITIATIVE, A.D.N. y OTHERS, Locally linear embedding (LLE) for MRI based Alzheimer's disease classification. *NeuroImage*, 2013, vol. 83, p. 148-157.

MA, B., SU, Y. y JURIE, F., Bicov: a novel image representation for person re-identification and face verification. En *British Machine Vision Conference*, 2012, p. 11.

MA, B., SU, Y. y JURIE, F., Local descriptors encoded by fisher vectors for person re-identification. En *European Conference on Computer Vision*. Springer Berlin Heidelberg, 2012. p. 413-422.

NIYOGI, X., Locality preserving projections. En *Advances in Neural Information Processing Systems (NIPS)*. MIT, 2004. p. 153.

ROWEIS, S.T., SAUL, L.K. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science, 2000*, vol. 290, no. 5500, p. 2323-2326.

SAMMUT, C. y WEBB, G.I. *Encyclopedia of machine learning*. Springer Science & Business Media. 2011.

TANG, B., SONG, T., LI, F. y DENG, L., Fault diagnosis for a wind turbine transmission system based on manifold learning and Shannon wavelet support vector machine. *Renewable Energy*, 2014, vol. 62, p. 1-9.

TENENBAUM, J.B., SILVA, V. De y LANGFORD, J.C., A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 2000, vol. 290, no. 5500, p. 2319-2323.

TORGERSON, W.S., Multidimensional scaling: I. Theory and method. *Psychometrika*, 1952, vol. 17, no. 4, pp. 401-419.

VAN DER MAATEN, L.J.P., POSTMA, E.O. y VAN DEN HERIK, H.J., *Dimensionality reduction: A comparative review*. Tilburg, Netherlands: Tilburg Centre for Creative Computing, Tilburg University, Technical Report: 2009-005, 2009.

VEGA-HERNANDEZ, M. y VALDES-SOSA, P.A. EEG Source Imaging With Spatio-Temporal Tomographic Nonnegative Independent Component Analysis. *Human brain mapping*, 2009, vol. 30, no 6, p. 1898-1910.

WANG, B., JIANG, J., WANG, W., ZHOU, Z. y TU, Z,. Unsupervised Metric Fusion by Cross Diffusion. En *Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2012. p. 2997-3004.

WANG, F. y SUN, J., Distance Metric Learning in Data Mining (Part II). En *SIAM 2012 International Conference on Data Mining*, 2012.

WANG, F. y SUN, J., Survey on distance metric learning and dimensionality reduction in data mining. *Data Mining and Knowledge Discovery*, 2015, vol. 29, no 2, p. 534-564.

WANG, F., ZHAO, B. y ZHANG, C., Unsupervised maximum margin projection. En 2011 *IEEE Transactions on Neural Networks (TNN)*, 2011, vol. 22, no. 9, p. 1446-1456.

WANG, J., DENG, Z., CHOI, K., JIANG, Y., LUO, X., CHUNG, F. y WANG, S., Distance metric learning for soft subspace clustering in composite kernel space. *Pattern Recognition*, 2016, vol. 52, p. 113-134.

WANG, Q.Y., YUEN, P.C. y FENG, G.C., Semi-supervised metric learning via topology representation. En *20th European Signal Processing Conference (EUSIPCO)*. IEEE, 2012. p. 639-643.

XING, E.P., NG, A.Y., JORDAN, M.I. y RUSSELL, S., Distance Metric Learning with Application to Clustering with Side-Information. En *Advances in Neural Information Processing Systems (NIPS)*, 2002, vol. 15, p. 505-512.

YANG, B., XIANG, M. y SHI, L. Feature reduction using locally linear embedding and distance metric learning. En *Emerging Technologies for Information Systems, Computing, and Management*. Springer New York, 2013. p. 537-544.

YANG, L. y JIN, R., Distance metric learning: A comprehensive survey. *Department of Computer Science and Engineering, Michigan State University*. 2016, vol 2.

YING, Y., Distance Metric Learning with Eigenvalue Optimization. *Journal of Machine Learning Research*. 2012, vol. 13, no Jan, p. 1-26.

YING, Y., HUANG, K. y CAMPBELL, C., Sparse Metric Learning via Smooth Optimization. En *Advances in Neural Information Processing Systems (NIPS)*, 2009, p. 2214-2222.

ZHANG, L., QIAO, L. y CHEN, S., Graph-optimized locality preserving projections. *Pattern Recognition*, 2010, vol. 43, no. 6, p. 1993-2002.

ZHANG, Z. y WANG, J., MLLE: Modified locally linear embedding using multiple weights. En *Advances in Neural Information Processing Systems (NIPS)*. 2006, p. 1593-1600.

ZIEGELMEIER, L., KIRBY, M. y PETERSON, C., Locally Linear Embedding Clustering Algorithm for Natural Imagery. *arXiv:1202.4387*. 2012.