Expert Systems With Applications 51 (2016) 151–160

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

# Clustering using PK-D: A connectivity and density dissimilarity

Ariel E. Bayá*, Mónica G. Larese, Pablo M. Granitto

*CIFASIS, French Argentine International Center for Information and Systems Sciences, UNR-CONICET (Argentina), Bv. 27 de Febrero 210 Bis, Rosario 2000, Argentina*

ARTICLE INFO

*Keywords:*
Clustering
Dimensionality reduction

ABSTRACT

We present a new dissimilarity, which combines connectivity and density information. Usually, connectivity and density are conceived as mutually exclusive concepts; however, we discuss a novel procedure to merge both information sources. Once we have calculated the new dissimilarity, we apply MDS in order to find a low dimensional vector space representation. The new data representation can be used for clustering and data visualization, which is not pursued in this paper. Instead we use clustering to estimate the gain from our approach consisting of dissimilarity + MDS. Hence, we analyze the partitions' quality obtained by clustering high dimensional data with various well known clustering algorithms based on density, connectivity and message passing, as well as simple algorithms like *k*-means and Hierarchical Clustering (HC). The quality gap between the partitions found by *k*-means and HC alone compared to *k*-means and HC using our new low dimensional vector space representation is remarkable. Moreover, our tests using high dimensional gene expression and image data confirm these results and show a steady performance, which surpasses spectral clustering and other algorithms relevant to our work.



## 1. Introduction

Clustering algorithms can be used to uncover unknown relations existing in a set of unlabeled data. These algorithms can be divided into families according to their characteristics, for example there are partitional and hierarchical algorithms (Jain, Murty, & Flynn, 1999; Xu & Wunsch II, 2005). Similarly, hierarchical algorithms can be divided into agglomerative and divisive methods. Thus, we could make a taxonomy to categorize all clustering methods. This classification of algorithms into "families" shows a particular approach to clustering, one requiring many different algorithms for different kinds of data. On the contrary, we aspire to solve many clustering tasks using a reduced number of simple algorithms. Moreover, our main goal is to use the most simple algorithms available. As a result, we direct our interest into clustering methods involving kernels (Dhillon, Guan, & Kulis, 2004; Mika et al., 1999) or more general representations based on dissimilarity matrix (Pekalska & Duin, 2008; Pekalska, Paclik, & Duin, 2002; Schölkopf, 2001). These types of methods are able to simplify the clustering procedure by using a low dimensional vector representation derived from the kernel or dissimilarity matrix. There is a rich bibliography describing both groups of methods including also a crossover area discussing the relation between kernels and dissimilarities, i.e. a formal discussion explaining when a dissimilarity can be treated as a kernel and how to proceed when it cannot (Pekalska & Duin, 2008; Schölkopf, 2001; Williams, 2002). The dissimilarity proposed in this work does not qualify as a Mercer Kernel, hence, its decomposition leads to a non-Euclidean space. To find an Euclidean representation approximating the original data we only consider the positive spectra of the dissimilarity. Moreover, we only use a small subset of the eigenvectors of the centered dissimilarity. However, we can accurately represent datasets formed by arbitrary shaped clusters or high dimensional noisy data, even if the clusters do not have spherical shape or Gaussian distribution.

As we stated above, our motivation is to reduce the complexity of a clustering problem by improving the representation of the data. We have pursued this goal in a previous work and, as a result, we developed a penalized metric (Bayá & Granitto, 2011) that permitted us to cluster with a simple algorithm data having arbitrary shapes and high dimensionality. However, this metric could not overcome many of the limitations from methods based on connectivity. The solution proposed in the present paper aims at: (i) finding a lower dimensional representation of the original data and (ii) overcoming some of the limitations known to exist in connectivity approaches (Bayá & Granitto, 2011). Thus, we build a new dissimilarity combining connectivity and density information as an improvement to methods based solely on connectivity. Next, we apply MDS to the dissimilarity to find a new representation of the

* Corresponding author. Tel.: +54 341 4815569x326; fax: +54 341 4821771.
*E-mail addresses:* baya@cifasis-conicet.gov.ar (A.E. Bayá), larese@cifasis-conicet.gov.ar (M.G. Larese), granitto@cifasis-conicet.gov.ar (P.M. Granitto).

http://dx.doi.org/10.1016/j.eswa.2015.12.037
0957-4174/

original data. The representation found by MDS allows us to use any clustering algorithm without restricting us to those relying on dissimilarity matrix. Finally, we use a simple clustering algorithm to find groups in the new representation and compare the quality of them with other similar algorithms.

This manuscript has the following structure: Section 2 first considers previous works related to our ideas and goals. It also describes our two dissimilarity variants, the merging strategy and other related matters. Section 3 discusses first how to set up the parameters of our dissimilarity and then it shows the results of our experiments on real data. Finally, Section 4 presents some conclusions and considers ideas for future work.

## 2. Finding a new data representation

### 2.1. Related work

The idea of developing a method able to model the complex relations between the patterns of a dataset is not new. There are many dimensionality reduction methods (Belkin & Niyogi, 2003; Roweis & Saul, 2000; Tenenbaum, De Silva, & Langford, 2000), kernel methods (Mika et al., 1999) and spectral methods (Luxburg, 2007; Nadler & Galun, 2007) trying to accomplish this. Dimensionality reduction target is to find a new representation using fewer dimensions that preserves the relations existing in the original data. The outcome from these methods can be used for visualization, clustering or classification. However, applying clustering or classification after dimensionality reduction might not have always the desired effect. Preserving the ties between the new representation and the original data might not always be helpful to find "good" partitions. For example, Principal Components Analysis (James, Witten, Hastie, & Tibshirani, 2014) finds a representation preserving ties by retaining the components with highest standard deviation, however, the components dividing data into groups might not be those with highest standard deviation. Analogously, the cost function used by ISOMAP (Tenenbaum et al., 2000) or LLE (Roweis & Saul, 2000) to find a lower data representation does not emphasize in preserving the natural differences within the data. Our dissimilarity, on the contrary, is specially designed for clustering rather than visualization, hence, it emphasizes natural differences within the data. Therefore, after applying MDS we find a representation making the subsequent clustering step easier.

In a previous work we developed a distance called PKNNG (Bayá & Granitto, 2011), which we successfully used to cluster arbitrary shaped clusters, high dimensional noisy data and data embedded in a manifold. However, we were restricted to combine it with a small subset of clustering algorithms since PKNNG transformed the original data into a dissimilarity matrix. This dissimilarity is based on a graph of neighbors, hence, it relies only on connectivity. The components from the neighbor graph are connected by penalized edges joining the closest components with a single edge. Finally, the geodesic distance is calculated between all pairs using Dijkstra's algorithm. Since PKNNG relies on connectivity there are cases that are too complicated or not possible to solve, for example, a pair of overlapped spherical clusters with Gaussian distribution. We explore the use of density information as a possible solution to some of the limitations of PKNNG. There is a similarity known as Evidence Accumulation by multiple Clustering (EAC) (Fred & Jain, 2005), which is used for clustering. However, there is a range of values for which EAC does not behave as a good similarity because it fragments the information to the point of rendering it useless. Fred and Jain (2005) discuss this issue in great detail. Yet, we have found out that fragmented information provides us with interesting insight about the density among neighbors. Our method aggregates this information to PKNNG in an effort to overcome previous limitations.

Our dissimilarity is used in combination with Classical MDS to find a more simple representation of the original data. There are several methods that have already explored this topic, for instance, Xu, Hancock, and Wilson (2014) used Ricci flows to remove artifacts rendering dissimilarity non-Euclidean. Later they tested their corrected dissimilarity in classification problems. Solving classification problems in vector spaces derived from dissimilarities has been properly introduced by Pękalska and Duin (Pekalska & Duin, 2008; Pekalska et al., 2002). There are some results under particular circumstances showing that classification using dissimilarity based feature spaces can be better than the ones obtained based on kernels (Kim & Duin, 2010). However, it should be noted that the data supporting this conclusion is reduced. There are other contributions related to our work pursuing visualization rather than clustering. Isomap (Tenenbaum et al., 2000) aims to find a lower dimensional representation from a dataset by using connectivity, connections through the shortest path and geodesic distance. Both Isomap and PKNNG share some features, however, the penalization scheme from PKNNG opposes to the idea of preserving geometrical relations. Instead, PKNNG is intended for clustering, hence it penalizes non-neighboring distances. There are methods pursuing the objective of ISOMAP but using different strategies to achieve it, among the most relevant visualization/dimensionality reduction methods we can name: Local Liner Embedding (LLE) (Roweis & Saul, 2000), Laplacian Eigenmaps (Belkin & Niyogi, 2003) and stochastic neighbor embedding (SNE) (Hinton & Roweis, 2003).

Our objective is to construct a dissimilarity with discriminative properties by aggregating two sources of information: density and connectivity. The discriminative properties emphasize the dissimilarity between non-neighboring samples in order to simplify the search for clusters. We use three methods as basic blocks to build our function. The first two blocks are used to measure density by one of two methods: (1) EAC, which is a method based on ensembles of $k$-means (Forgy, 1965) and (2) a method based on k nearest neighbors (Mitchell, 1997) (k-nn). The k-nn ensembles mimics the behavior of EAC but using an unsupervised k-nn algorithm instead. As a result, the first and second block originate each a dissimilarity variant, which estimates density in a different way. The third building block is the PKNNG distance. Finally, after having our dissimilarity we apply classical MDS (Cox & Cox, 2000) to find a lower dimensional data representation, which we use to find clusters. Fig. 1 shows a diagram of the proposed pipeline and Section 2.3 provides a thorough description of our approach.

### 2.2. Finding a low dimensional data representation

Assuming there is a generic dissimilarity ($D \in \mathcal{R}^{n\times n}$) we would like to find a new vector space representation based on $D$. We define $S = D^2$ and $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$, where $S$ is a squared dissimilarity matrix ($s_{ij} = d_{ij}^2$), $I$ is the identity matrix, $\mathbf{1}$ is a $n \times 1$ vector of ones and $H$ is the centering matrix. We use these elements to find a new vector representation $X$:

$$\begin{aligned} B &= \frac{-HSH}{2} \\ &= V\Lambda V^T = XX^T, \end{aligned} \tag{1}$$

where $\Lambda$ is a diagonal matrix containing the eigenvalues of $B$ and $V$ is an orthogonal eigenvectors matrix. When $B$ is not a semidefinite matrix there will be negative eigenvalues in $\Lambda$. A discussion about this issue and the full derivation of the previous equation can be found in Pekalska et al. (2002), Williams (2002) and Schölkopf (2001). Mercer's Theorem (Cristianini and Shawe-Taylor, 2000, Section 3.3.1) relates the eigenvalues from $\Lambda$ to squared norms in the new space representation having $V$ as a base. Hence, the existence of negative eigenvalues amounts to negative squared distances, which contradicts Euclidean geometry. We solve this problem by