Information Sciences 360 (2016) 231–243

Contents lists available at ScienceDirect

Information Sciences

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

Density-based geodesic distance for identifying the noisy and nonlinear clusters

Jaehong Yu[a], Seoung Bum Kim[a,b,*]

[a] Department of Industrial Management Engineering, Korea University, 145 Anam-Ro, Seoungbuk-Gu, Anam-dong , Seoul 136-713, South Korea

[b] Center for Discrete Mathematics and Theoretical Computer Science, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

ARTICLE INFO

Article history:
Received 27 September 2015
Revised 1 April 2016
Accepted 26 April 2016
Available online 30 April 2016

Keywords:
Geodesic distance
Mutual neighborhood-based density coefficient
Noisy data clustering
Nonlinearity

ABSTRACT

Clustering analysis can facilitate the extraction of implicit patterns in a dataset and elicit its natural groupings without requiring prior classification information. For superior clustering analysis results, a number of distance measures have been proposed. Recently, geodesic distance has been widely applied to clustering algorithms for nonlinear groupings. However, geodesic distance is sensitive to noise and hence, geodesic distance-based clustering may fail to discover nonlinear clusters in the region of the noise. In this study, we propose a density-based geodesic distance that can identify clusters in nonlinear and noisy situations. Experiments on various simulation and benchmark datasets are conducted to examine the properties of the proposed geodesic distance and to compare its performance with that of existing distance measures. The experimental results confirm that a clustering algorithm with the proposed distance measure demonstrated superior performance compared to the competitors; this was especially true when the cluster structures in the data were inherently noisy and nonlinearly patterned.

1. Introduction

Clustering analysis can facilitate the discovery of inherent patterns from large datasets and reveal their natural groupings without using prior classification information. Clustering algorithms systematically partition a dataset by minimizing within-group variation and maximizing between-group variation [39,13]. Clustering analysis has been applied in various fields including information retrieval [28], text mining [44], bioinformatics [4], marketing management [6], and process control [25,43]. A number of clustering algorithms have been developed [23]. The most prominent of these are *k*-means [29], PAM (partitioning around medoids) [26,41], DBSCAN (density-based spatial clustering of applications with noise) [15], and modularity-based clustering [31,32].

For more appropriate groupings with these clustering algorithms, determining the proper distance measure is an important issue. In general, the majority of the existing clustering algorithms adopt the Euclidean distance and Manhattan distance as dissimilarity measures [39,10,22]. Euclidean distance and Manhattan distance are defined as the length of a straight line between two observations in Euclidean space and the sum of the differences between two observations in all features, respectively. Euclidean and Manhattan distances can be generalized as Minkowski distances. Cosine distance and Pearson correlation distance can also be used for clustering analysis. Cosine distance is defined as the cosine of the angle

* Corresponding author. Tel.: +82 232903397.
E-mail addresses: dreamer7744@korea.ac.kr (J. Yu), sbkim1@korea.ac.kr (S.B. Kim).

http://dx.doi.org/10.1016/j.ins.2016.04.032
0020-0255/

between two observations [38] and the Pearson correlation distance is quantified from a correlation coefficient between two observations [21]. These two distance measures fall between [0, 2] regardless of the scale or size of the datasets. They have been widely used for microarray clustering and document clustering tasks [11,9,38].

Although these distance measures render reasonable results within the situations for which they were designed, no consensus exists regarding the best all-around performer in real-life situations. Because the majority of these distance measures do not consider the shape of the data, they produce unsatisfactory results when the data exhibit nonlinear and (or) local patterns such as *S*-curves and Swiss roll shapes.

To overcome these limitations, Tenenbaum et al. [40] proposed a geodesic distance to capture intrinsic nonlinear patterns (manifold structures) in datasets. The geodesic distance is computed from the neighborhood graph. The weights of the graph can be represented by Euclidean distances between neighborhoods and the final geodesic distances between observations are defined as the sum of the weights in their shortest path. This geodesic distance can effectively reflect the topological structures of the dataset and thus, it can accommodate nonlinear patterns. To utilize this property, several clustering algorithms have adopt the geodesic distance [37,2,16]. However, the geodesic distance does not consider the density of the data and it is vulnerable to noise around the clusters, encountered in many real situations [5,8,3].

In addition to the geodesic distance, a path-based distance based on the minimum spanning tree structure was proposed to accommodate the nonlinearity [17,18]. However, this distance measure still suffers from the noises around the nonlinear clusters [7]. To handle the noises, Sajama and Orlitsky [36] proposed a density-based distance that can be calculated with the graph structure whose weights are defined by kernel density estimators. However, this method relatively complicated because it requires determination of several parameters before its full construction including types of kernel functions. Moreover, it has been known that the kernel density estimation scheme is vulnerable to high dimensional datasets [5,36].

To address the limitations of the existing distance measures, we propose a density-based geodesic distance that is especially useful for grouping data exhibiting noisy and nonlinear patterns. The proposed distance measure uses not only the neighborhood graph for nonlinearity, but also the density for robustness against the noise. Although the neighborhood graph is beneficial for describing the nonlinear patterns, the noise around the clusters prevents the graph structure from accommodating the nonlinearities. To achieve robustness against this noise, the proposed distance measure employs a density calculation scheme that scales the weights in the neighborhood graph. This makes the distance measure between sparse observations much greater and vice versa for dense observations. Therefore, the proposed distance measure can produce improved performance compared to the original geodesic distance when there is significant noise around the nonlinear groups.

The remaining of this paper is organized as follows. Section 2 introduces the proposed distance measure. Section 3 presents a simulation study to demonstrate the advantages of the proposed distance measure over the existing measures. Section 4 describes the results of experiments with real data to examine the properties of the proposed distance measure and to compare it with existing distance measures. Section 5 contains our concluding remarks.

2. Density-based geodesic distance

The proposed density-based geodesic distance is calculated with three main steps: (1) The first is to represent the data as a *k*-nearest neighbor graph. In this graph, all observations are represented as nodes and each observation is connected to its neighborhood with an edge. A nearest neighbor graph is widely used to reflect nonlinear patterns in a dataset [37,14]. (2) Then, each observation is described as a density. For more effective clustering analysis, we propose a novel density measure, computed from the *k*-nearest neighbor graph and mutual neighborhood relationships between observations. This density measure is called a mutual neighborhood-based density coefficient and is used to scale the distances between the observations. (3) In the final step, the distance between the observations is computed with the shortest path in the scaled graph. All weights are scaled based on the mutual neighborhood-based density coefficient and the distance between the observations is defined as a sum of weights in their shortest path.

2.1. Constructing the k-nearest neighbor graph

The first step of computing the proposed distance is to represent the data as a neighborhood graph structure. Several types of neighborhood graph structures exist, including the ε-neighbor graph and the *k*-nearest neighbor graph [30]. Among these, the *k*-nearest neighbor graph has been widely used in practice because of its easy construction and effective description of local properties [20,40]. Hence, in this study, we use the *k*-nearest neighbor graph construction scheme. To construct the *k*-nearest neighbor graph, the *k*-nearest neighborhood of each observation should be defined. The *k*-nearest neighborhoods of observation *i*, $K(i)$, is defined as follows:

$$K(x_i) = \{x_j | \|x_i - x_j\|_2 \le d_i^k\}, \tag{1}$$

where $\|x_i - x_j\|_2$ denotes the Euclidean distance between x_i and x_j and d_i^k is the kth smallest Euclidean distance from the observation x_i to the other observations.

From the definition of the *k*-nearest neighborhood, a symmetric neighborhood set of observation x_i, $\Omega(x_i)$, can be defined as follows:

$$\Omega(x_i) = \{x_j | x_j \in K(x_i) \text{ or } x_i \in K(x_j)\}. \tag{2}$$