Available online at www.sciencedirect.com

ScienceDirect

Fuzzy Sets and Systems 286 (2016) 157–172

FUZZY sets and systems

www.elsevier.com/locate/fss

Geodesic distance based fuzzy c-medoid clustering – searching for central points in graphs and high dimensional data

András Király [a,*], Ágnes Vathy-Fogarassy [b], János Abonyi [a]

[a] University of Pannonia, Department of Process Engineering, P.O. Box 158, Veszprém H-8200, Hungary
[b] University of Pannonia, Faculty of Information Technology, P.O. Box 158, Veszprém H-8200, Hungary

Received 5 July 2013; received in revised form 14 May 2015; accepted 26 June 2015

Available online 30 June 2015

Abstract

Clustering high dimensional data and identifying central nodes in a graph are complex and computationally expensive tasks. We utilize k-nn graph of high dimensional data as efficient representation of the hidden structure of the clustering problem. Initial cluster centers are determined by graph centrality measures. Cluster centers are fine-tuned by minimizing fuzzy-weighted geodesic distances. The shortest-path based representation is parallel to the concept of transitive closure. Therefore, our algorithm is capable to cluster networks or even more complex and abstract objects based on their partially known pairwise similarities.

The algorithm is proven to be effective to identify senior researchers in a co-author network, central cities in topographical data, and clusters of documents represented by high dimensional feature vectors.

Keywords: Clustering; Fuzzy c-medoid; Centrality; Geodesic distance

1. Introduction

Cluster is a group of objects that are more similar to one another than to members of other clusters. In metric spaces, similarity is often defined by means of a distance norm. Distance can be measured among the data vectors themselves, or as a distance from a prototypical object of the cluster. In case of high-dimensional data, classical Euclidean distance-based methods like k-means, fuzzy c-means or fuzzy c-medoid do not perform well and it is difficult to design and validate cluster prototypes, that are able to represent the distribution of the data. One approach to handle this problem is to assume, that the data of interest lie on an embedded non-linear manifold within the higher-dimensional space.

In brief, our main idea is to reveal the hidden, complex structure of high dimensional data by constructing and clustering the k-nn graph of the data. Our concept can be seen in Fig. 1. Shortest path distances among the objects (nodes) can be considered as the approximation of geodesic distances defined on the low dimensional manifold within

* Corresponding author.
E-mail address: kiralya@fmt.uni-pannon.hu (A. Király).

http://dx.doi.org/10.1016/j.fss.2015.06.022
0165-0114/

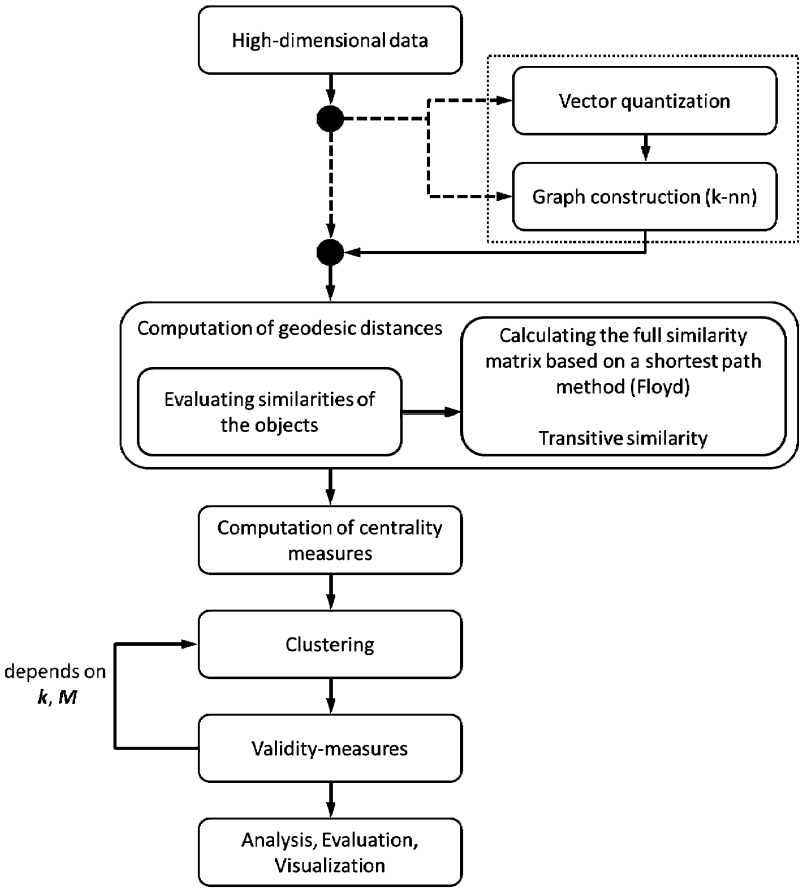

Fig. 1. Block scheme describing the stages of the proposed method.

the higher-dimensional space. Centrality measures of the nodes of the weighted undirected k-nn graph can be used to initialize the cluster centers. The classical fuzzy c-medoid algorithm is able to find clusters using the obtained initial values and distance matrix. The clustered graph (distance matrix) is visualized by a distance-preserving mapping to give more insight to the problem.

It should be noted, that the proposed shortest-path based approach is parallel to the concept of transitive closure. Therefore, our algorithm is capable to cluster networks or even more complex and abstract objects based their on partially known pairwise similarities.

The rest of the paper is organized as follows. In the next section, we give a short review of the most relevant literature. We present the algorithm in Section 3.1. The details of the graph construction are given in Section 3.2. Section 3.3 clarifies the connection between geodesic distances and transitive similarities. In Section 3.4 we discuss the most important aspects of the utilized centrality measures, while the details are given in Appendix A. Appendix B presents some cluster validity indices used to prove that it is worth using the proposed techniques. Section 4 demonstrates that the proposed approach works well on benchmark data sets and high dimensional problems, like document clustering. Comparison with other graph-based clustering techniques and two real-life examples are also presented. Section 5 concludes our work.

2. Literature review

Clustering is a very important research topic in machine learning and data mining. The classical k-means algorithm is capable to solve most of the practical segmentation problems [1]. Although k-means is an almost fifty years old method, the existence of its several variants, like the k-medoid [2] or the fuzzy c-means [3] algorithm prove the high relevance of this approach. Cluster prototypes may be vectors of the same dimension as the data objects, but they