Available online at www.sciencedirect.com

ScienceDirect

Fuzzy Sets and Systems 270 (2015) 111–126

FUZZY sets and systems

www.elsevier.com/locate/fss

Fuzzy joint points based clustering algorithms for large data sets

Efendi Nasibov [a,b,*], Can Atilgan [a], Murat Ersen Berberler [a], Resmiye Nasiboglu [a]

[a] *Department of Computer Science, Dokuz Eylul University, Buca, 35160 Izmir, Turkey*
[b] *Institute of Control Systems, Azerbaijan National Academy of Sciences, AZ-1141 Baku, Azerbaijan*

Received 24 September 2013; received in revised form 16 June 2014; accepted 15 August 2014

Available online 27 August 2014

Abstract

The fuzzy joint points (FJP) method is one of the successful fuzzy approaches to density-based clustering. Besides the basic FJP method, there are other methods based on the FJP approach such as, Noise-Robust FJP (NRFJP), and Fuzzy Neighborhood DBSCAN (FN-DBSCAN). These FJP-based methods suffer from the low speed of the FJP algorithm, thus applications that deal with large databases cannot benefit from them. The Modified FJP (MFJP) method addresses this issue and achieves an improvement in speed, but it is not satisfactory from the point of applicability. In this work, we integrate various methods with FJP to establish an optimal-time algorithm. An even faster algorithm which uses the FJP approach in a somewhat supervised fashion is also proposed. Along with theoretic comparison, experimental results are presented to show the significant speed improvement, which will allow the FJP-based methods to be used on large data sets.

Keywords: Fuzzy neighborhood relation; Fuzzy Joint Points (FJP); Clustering; Optimal algorithm

1. Introduction

Clustering is one of the main tasks of data mining, which is a well-studied field of extracting information from large data sets. The goal of clustering algorithms is segmenting the entire data set into relatively homogeneous clusters, where the similarity of the records within the cluster is maximized, and the similarity to records outside this cluster is minimized. Clustering techniques are commonly categorized into hierarchical, model-based, grid-based, partitioning-based and density-based clustering [1,2].

Implementing fuzzy approaches for clustering usually offers more robust methods. Fuzzy C-Means (FCM) is the most popular fuzzy clustering algorithm, which falls into the partitioning-based category. Various algorithms have been developed by integrating FCM with other methods [3–7]. Despite their speed advantage, partitioning-based algorithms have some major disadvantages. One of these disadvantages is the need of number of clusters to be known beforehand. These algorithms are usually run multiple times with different cluster numbers and the best of the obtained results is then singled out. Another disadvantage is the dependency to the chosen metric. Shapes of the resulting

* Corresponding author at: Department of Computer Science, Dokuz Eylul University, Buca, 35160 Izmir, Turkey.
E-mail address: efendi.nasibov@deu.edu.tr (E. Nasibov).

http://dx.doi.org/10.1016/j.fss.2014.08.004
0165-0114/

clusters are actually determined by the distance function. This implies that clusters with different and irregular shapes cannot be discovered using partitioning-based algorithms. Specifying initial cluster centers and handling noise are the other problems with these algorithms.

Density-based clustering methods discover spatially connected components of data by investigating neighborhood relations. These methods typically deal with noise well and are designed to discover clusters with irregular shapes. Some well-known algorithms such as DBSCAN, OPTICS, DBCLASD and DENCLUE [8–11] embody the density-based approach. Algorithms like GDBSCAN can also work on geometric objects rather than merely points [12]. For clustering web-pages, an algorithm with relatively low time complexity was given in [13]. DBSCAN is sensitive to two parameter inputs, which are used to determine neighborhood relations. OPTICS addresses the parameter selection problem of DBSCAN by intelligently ordering the points with respect to a reachability property described in the work [9]. It does not explicitly extract clustering partitions, but maintains a cluster ordering that contains information equivalent to a wide range of parameter settings of DBSCAN. A robust density-based method that builds a hierarchy of partitions with different densities called HDS (or Auto-HDS) is given in [14]. HDS starts with clustering the entire data set (i.e. no point is considered as noise) and gradually reduces the number of data points it handles by a predefined fraction, resulting in a number of partitions with different densities. Each partition is then evaluated to construct a compact hierarchy. It also provides ranking criteria to choose the best clustering.

A density-based algorithm which uses fuzzy neighborhood relation is the Fuzzy Joint Points (FJP) algorithm [15]. It is a fuzzy relative of DBSCAN. FN-DBSCAN, Scalable FN-DBSCAN, NRFJP and MFJP are other algorithms based on FJP approach [16–20]. They were developed to improve the FJP algorithm in terms of noise handling, speed and overall clustering performance. Although FJP-like algorithms yield some notable advantages over DBSCAN, they are slower. The modified version of FJP, i.e. MFJP, provides speed improvement to some extent. However, the implementation of the MFJP algorithm was still slow, such that the computational tests could be conducted up to only 1336 data points with 2 dimensions in an acceptable amount of time.

In this study, we discuss how FJP-based methods can be further improved in terms of worst-case time complexity and running time performance, so that they can be used in practice on larger data sets. In Section 2, we look over the FJP and MFJP methods. Their time complexities are analyzed to reveal the bottlenecks, and improvements that lead to an optimal-time algorithm are introduced in Section 3. A novel algorithm based on the FJP approach, which has a considerable speed advantage is then presented in Section 4. We give theoretic and experimental comparisons of the discussed methods in Section 5. Finally, Section 6 concludes the paper.

2. The FJP and MFJP methods

The FJP and MFJP methods use Euclidean distance as their fundamental notion of distance between data points. There are clustering methods that use different distance functions, whereas the majority of the distance-based clustering methods use Euclidean distance. Euclidean distance between any points a and b of m-dimensional space E^m is defined then

$$d(a,b)=\sqrt{\sum_{i=1}^{m}(a_i-b_i)^2}.$$

Any distance function mentioned in this paper refers to Euclidean distance. Data points are handled as conical fuzzy points and fuzzy neighborhood relations are used to determine clusters. The definitions of the mathematical notions that are essential for the FJP and MFJP methods are introduced below.

Conical fuzzy point: A conical fuzzy point $P=(p,R)\in F(E^m)$ is a fuzzy set whose membership function is given by

$$\mu(x)=\begin{cases}1-\frac{d(x,p)}{R}, & d(x,p)\le R\\ 0, & \text{otherwise,}\end{cases}$$

where $p\in E^m$ is the center of the conical fuzzy point P, and $R\in E^1$ is the radius of the point's support set *supp* P, where