Knowledge-Based Systems 126 (2017) 33–47

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Self-adapted mixture distance measure for clustering uncertain data

Han Liu[a,b,*], Xianchao Zhang[b], Xiaotong Zhang[a,b], Yi Cui[c]

[a] School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China
[b] School of Software, Dalian University of Technology, Dalian 116020, China
[c] School of Economics and Management, Beijing Jiaotong University, Beijing 100044, China

ARTICLE INFO

Article history:
Received 13 November 2016
Revised 9 April 2017
Accepted 10 April 2017
Available online 11 April 2017

Keywords:
Clustering
Uncertain data
Induced kernel distance
Jensen–Shannon divergence
Self-adapted mixture distance measure

ABSTRACT

Distance measure plays an important role in clustering uncertain data. However, existing distance measures for clustering uncertain data suffer from some issues. Geometric distance measure can not identify the difference between uncertain objects with different distributions heavily overlapping in locations. Probability distribution distance measure can not distinguish the difference between different pairs of completely separated uncertain objects. In this paper, we propose a self-adapted mixture distance measure for clustering uncertain data which considers the geometric distance and the probability distribution distance simultaneously, thus overcoming the issues in previous distance measures. The proposed distance measure consists of three parts: (1) The induced kernel distance: it can be used to measure the geometric distance between uncertain objects. (2) The Jensen–Shannon divergence: it can be used to measure the probability distribution distance between uncertain objects. (3) The self-adapted weight parameter: it can be used to adjust the importance degree of the induced kernel distance and the Jensen–Shannon divergence according to the location overlapping information of the dataset. The proposed distance measure is symmetric, finite and parameter adaptive. Furthermore, we integrate the self-adapted mixture distance measure into the partition-based and density-based algorithms for clustering uncertain data. Extensive experimental results on synthetic datasets, real benchmark datasets and real world uncertain datasets show that our proposed distance measure outperforms the existing distance measures for clustering uncertain data.

1. Introduction

Clustering is a fundamental task in machine learning and data mining. Traditional clustering algorithms mainly focus on certain data. However, due to a variety of reasons like randomness in data generation and collection, imprecision in physical measurement and data staling, uncertain data arises naturally from many real applications such as biomedical measurement [1], sensor networking [2], meteorological forecasting and so on [3]. Uncertain data brings new challenges to traditional clustering algorithms.

Distance measure plays a key role in clustering uncertain data. Existing uncertain data clustering algorithms extend traditional clustering algorithms for dealing with uncertain data by using geometric distance measure or probability distribution distance measure. Partition-based algorithms like UK-means [4], UK-medoids [5] use the geometric distance measure like expected distance, uncertain distance to extend the traditional clustering algorithms *k*-means and *k*-medoids. Density-based algorithms like FDBSCAN [6], FOPTICS [7] employ the geometric distance based probability definitions to extend the traditional clustering algorithms DBSCAN [8] and OPTICS [9]. These algorithms rely on geometric distance measure, thus can not identify the difference between uncertain objects with different distributions heavily overlapping in locations. Recently, Jiang et al. [10] have proposed to use the probability distribution distance measure Kullback–Leibler (KL) divergence to measure the difference between uncertain objects, however, the KL divergence can not distinguish the difference between different pairs of completely separated uncertain objects and it suffers from asymmetry and infinity.

In this paper, we firstly point out the limitations of the existing distance measures, then propose a self-adapted mixture distance measure for clustering uncertain data which considers the geometric distance and the probability distribution distance simultaneously. The proposed distance measure consists of three parts: (1) The induced kernel distance: this part employs an induced kernel distance to measure the geometric distance between uncertain objects. In contrast to the previous geometric distance measures, it is finite and can deal with the non-linear data better. (2) The

* Corresponding author.
E-mail addresses: liu.han.dut@gmail.com (H. Liu), xczhang@dlut.edu.cn (X. Zhang).

http://dx.doi.org/10.1016/j.knosys.2017.04.002
0950-7051/

Jensen–Shannon divergence: this part uses the Jensen–Shannon divergence to measure the probability distribution difference between uncertain objects. In contrast to the KL divergence, it is symmetric and finite. (3) The self-adapted weight parameter: this part utilizes the location overlapping information of the dataset to adjust the importance degree of the induced kernel distance and the Jensen–Shannon divergence. It provides a reasonable weight parameter choice for these two distance measures. The proposed self-adapted mixture distance measure is symmetric, finite and parameter adaptive. Furthermore, we integrate it into the typical partition-based algorithm *k*-medoids and density-based algorithm DBSCAN for clustering uncertain data. Experimental results on synthetic datasets, real benchmark datasets and real world uncertain datasets demonstrate the superiority of our proposed self-adapted mixture distance measure over the existing distance measures for clustering uncertain data.

2. Related work

2.1. Algorithms based on geometric distance

UK-means [4] is one of the earliest algorithms which attempt to deal with uncertain data. It extends the traditional clustering algorithm *k*-means by using expected distance as the distance measure. The methods proposed in [11–15] use different pruning techniques and theories to save the computation time of the redundant expected distances. CK-means [16] is the most optimized version of UK-means. It can avoid a large amount of expected distance calculation and get the equivalent result with UK-means. UK-medoids [5] uses uncertain distance as the distance measure and employs the *k*-medoids clustering scheme. FDBSCAN [6] and FOPTICS [7] respectively introduce some new geometric distance based probability definitions to extend the density-based algorithm DBSCAN [8] and the hierarchical density-based algorithm OPTICS [9] for clustering uncertain data. Zhang et al. [17] also extend the traditional density-based algorithm DBSCAN for clustering uncertain data by providing the probabilistic definitions. Züfle et al. [18] use the density-based algorithm as the basic algorithm and propose a possible world based clustering framework for uncertain data. All these algorithms can deal with uncertain data to some extent. However, no matter expected distance, uncertain distance or geometric distance based probability definitions, the nature behind them relies on the geometric locations of uncertain objects, they do not take into account the probability distributions of uncertain objects, thus are difficult to identify the difference between uncertain objects with different distributions heavily overlapping in locations.

2.2. Algorithms based on probability distribution distance

Jiang et al. [10] propose to use the probability distribution distance measure KL divergence to measure the difference between uncertain objects for clustering uncertain data. Compared with geometric distance measures, the KL divergence can find the difference between uncertain objects with different distributions heavily overlapping in locations. However, it can not distinguish the difference between different pairs of completely separated uncertain objects. In addition, it suffers from asymmetry and infinity.

2.3. Other algorithms

MMVar [19] introduces the notion of uncertain prototype and then proposes an objective function which aims to minimize the variance of cluster mixture models. UCPC [20] proposes the notion of uncertain centroid and then designs a local search-based heuristic algorithm for clustering uncertain data. These two algorithms belong to partition-based algorithms and they do not clearly state what distance measures they depend on. In the clustering procedure, both MMVar and UCPC utilize the geometric location characteristic (expected distance) and the probability distribution characteristic (variance) of uncertain data, i.e., they simultaneously consider the geometric location and the probability distribution of uncertain data in an implicit mode. But as these characteristics seem a little simple for the complicated uncertain data, thus MMVar and UCPC are still difficult to obtain the satisfactory clustering performance. Yang et al. [21] propose a dynamic density-based clustering algorithm for uncertain data streams and it can find arbitrary shaped clusters in uncertain data streams. Tu et al. [22] propose a density grid-based clustering algorithm for uncertain data streams and it introduces an outlier detection mechanism to improve clustering performance. Both Günnemann et al. [23] and Zhang et al. [24] extend the density-based methods for subspace clustering of uncertain data. According to the search technique, they respectively belong to the bottom-up algorithm and top-down algorithm. These density-based algorithms focus on solving the issues in some special domains like clustering uncertain data streams and subspace clustering of uncertain data, and they do not solve the existing issues from the viewpoint of distance measures.

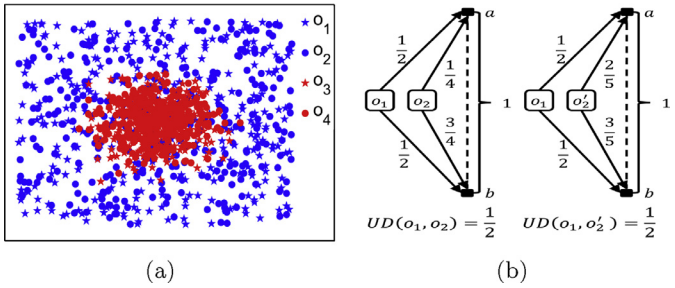

Fig. 1. Limitations of geometric distance measure.

3. Limitations of previous distance measures

3.1. Geometric distance measure

In UK-means [4] and its improved algorithms, they define the expected distance between an uncertain object o_i and a cluster center c_j as $ED(o_i, c_j) = \int d(x, c_j) f_i(x) dx$, where x is the uncertain dimensionality of o_i, $f_i(x)$ is the probability density function of o_i, d is the Euclidean distance. In CK-means [16], it proves that $ED(o_i, c_j) = d(C(o_i), c_j) + V(o_i)$, where $C(o_i)$ is the centroid of o_i and $C(o_i) = \int x f_i(x) dx$, $V(o_i)$ is the variance of o_i. Thus, in the clustering procedure, o_i will be assigned to the cluster center c_j which satisfies $\arg\min_{c_j}\{ED(o_i, c_j)\} = \arg\min_{c_j}\{d(C(o_i), c_j)\}$. This means that these expected distance based algorithms only consider the centroids of uncertain objects, thus they are difficult to identify the difference between uncertain objects with different distributions heavily overlapping in locations. For example in Fig. 1(a), o_1, o_2, o_3, o_4 are uncertain objects (each one is represented by a set of sample points), assume o_1, o_2 are uncertain objects with uniform distribution, o_3, o_4 are uncertain objects with Gaussian distribution, o_1, o_2, o_3, o_4 overlap heavily in locations and they have the same centroid. From the viewpoint of probability distribution, o_1, o_2 should belong to one cluster, o_3, o_4 should belong to another cluster. However, as o_1, o_2, o_3, o_4 have the same centroid, thus just using the expected distance is difficult to divide these uncertain objects into the right clusters.

In UK-medoids [5], it defines the uncertain distance between uncertain objects o_i and o_j as $UD(o_i, o_j) = \int\int d(x, y) f_i(x) f_j(y) dx dy$, where x and y are the uncertain dimensionalities of o_i and o_j