Neurocomputing 158 (2015) 81–89

Contents lists available at ScienceDirect

Neurocomputing

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

Large margin clustering on uncertain data by considering probability distribution similarity

Lei Xu [a], Qinghua Hu [b], Edward Hung [a], Baowen Chen [c], Xu Tan [c], Changrui Liao [d,*]

[a] Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China
[b] School of Computer Science and Technology, Tianjin University, Weijin Road No. 92, Nankai District, Tianjin, P.R. China
[c] Shenzhen Institute of Information Technology, Shenzhen, China
[d] Key Laboratory of Optoelectronic Devices and Systems of Ministry of Education and Guangdong Province, College of Optoelectronic Engineering, Shenzhen University, Shenzhen 518060, China

ARTICLE INFO

Article history:
Received 10 June 2014
Received in revised form
30 January 2015
Accepted 1 February 2015
Communicated by Dacheng Tao
Available online 7 February 2015

Keywords:
Clustering
Uncertain data
Probability density function
Large margin
Histogram intersection kernel

ABSTRACT

In this paper, the problem of clustering uncertain objects whose locations are uncertain and described by probability density functions (pdf) is studied. Though some existing methods (i.e. *K*-means, DBSCAN) have been extended to handle uncertain object clustering, there are still some limitations to be solved. *K*-means assumes that the objects are described by reasonably separated spherical balls. Thus, UK-means based on *K*-means is limited in handling objects which are in non-spherical shape. On the other hand, the probability density function is an important characteristic of uncertain data, but few existing clustering methods consider the difference between objects relying on probability density functions. Therefore, in this article, a clustering algorithm based on probability distribution similarity is proposed. Our method aims at finding the largest margin between clusters to overcome the limitation of UK-means. Extensively experimental results verify the performance of our method by effectiveness, efficiency and scalability on both synthetic and real data sets.

1. Introduction

In real applications, raw data (for example, in the case of sensor data) are usually not precise or accurate when they were collected or produced due to the limitation of instruments or measurement method. It is necessary to consider the uncertainty of data. For example, some temperature sensors are installed to monitor the temperature of a location. The temperature values observed by the sensor are not accurate due to the limitation of equipments. The real value of temperature at a given time cannot be exactly known and is uncertain. Another example for uncertain data, in an online shopping system (hotel booking system), different customers may give different scores to a hotel. The rating of a hotel is uncertain. Uncertain data can be represented by an exact value with margins of error, with or without a probability (density) distribution function (pdf). The result can also be in the form of a set of values or an interval, one of which maybe the real value. Thus, the methods will perform better when data uncertainty is considered.

Clustering uncertain data has been studied for many years [1–3]. For example, UK-means is a generalization of *K*-means which is extended to handle uncertain object clustering. The uncertain data are usually represented by probabilistic density functions (pdf) in multidimensional space. For example, in a marketing survey, travelers will be invited to evaluate the level of the hotels by scoring on various aspects, such as the location, the clear situation, the service, the safety, and so on. The level of a hotel can be modeled by an uncertain object in a multidimensional space.

Data uncertainty are usually classified into two types. One is existential uncertainty, which is caused by the fact that we are not sure the object or data tuple exists or not [4–7]; The other is value uncertainty caused by not knowing the value precisely. In this paper, we focus on the second case (value uncertainty). Recently, the relationship uncertainty to learn the order of the values on a dimension of the domain is proposed by Jiang et al. [8].

Though some existing methods have been extended to handle uncertain objects, i.e. UK-means and FDBSCAN, there are still some limitations to be solved to improve the performance of clustering. Probability density function (pdf) is an important characteristic of uncertain object, but few existing work focuses on comparing the difference between the pdfs of uncertain objects. Most previous work assumes that the uncertain objects with the same pdf and only relies on the geometric distance between uncertain objects.

In most previous work [1–3], each uncertain object is represented by hundreds or thousands of samples, which is more

* Corresponding author.

http://dx.doi.org/10.1016/j.neucom.2015.02.002
0925-2312/

complex than certain object. Thus, the similarity calculation between uncertain objects is hundreds or thousands times more complex than that between certain objects. Expensive cost on similarity computation is still a challenge for uncertain object clustering.

Our contribution in the work includes: (1) the probability distribution function (pdf) of each object is considered when measuring the similarity between uncertain objects. (2) Large margin clustering is used to improve the clustering quality. (3) The spectral clustering is more efficient than UK-means, because the similarity matrix can be obtained in advance.

In the rest of this paper, related work is briefly reviewed in Section 2. Section 3 introduces the preliminary knowledge on uncertain objects. Section 4 introduces the similarity calculation between uncertain objects and the clustering algorithm. Section 5 shows experimental evaluation on the performance of the our algorithm. Finally, we conclude our work in Section 6.

2. Related work

Clustering on certain data is a classic task in data mining, which has been studied for many years in pattern recognition, bioinformatics, and so on [9–11]. However, the work of clustering on uncertain data is just in a preliminary stage. The measurement of similarity between uncertain objects is different from that of certain objects. Most previous studies rely on geometric distance, which will be reviewed in Section 2.1. Another distance metric based on probability distribution is mentioned in Section 2.2.

2.1. Clustering based on geometric distance

UK-means [1] is first proposed to handle uncertain objects based on *K*-means. The only difference between UK-means and *K*-means is distance calculation method between an object and a cluster representative. Expected Euclidean distance is calculated in UK-means while *K*-means uses Euclidean distance. For arbitrary pdfs, the bottleneck of UK-means is expected distance calculation, which is computationally expensive. Thus, pruning techniques, such as MinMax and VDBi [2,3] are proposed to remove the candidate clusters from consideration, which are certainly not closest to the object, reducing a large amount of expected distance calculations. However, pruning UK-means is still undesirably slow due to the additional cost on pruning. In addition, both UK-means and pruning UK-means are based on *K*-means with assuming the clusters being separated spherical balls.

*F*DBSCAN [12] and *F*OPTICS [13] are density-based clustering methods on uncertain data based on DBSCAN [14] and OPTICS [15] respectively. Fuzzy distance function is used in *F*DBSCAN. The similarity between two uncertain objects is represented by distance probability functions and it lies in a range specified by the user. In DBSCAN, clusters are formed based on the concepts of core objects and reachability. In *F*OPTICS, a similar approach with probabilities is used to modify the OPTICS algorithm to handle uncertain data. Both UK-means and density-based clustering methods only consider the geometric distance between uncertain objects. Different from their work, the probability distribution distance is used to calculate the similarity between uncertain objects in our work.

Table 1
Notations used on uncertain object.

UK-means	uncertain *K*-means
$UD(o_i)$	uncertain domain of object o_i
$ED(o_i, o_j)$	expected distance between o_i and o_j
$f_i(x)$	the probability density function of object o_i (x is the possible location of o_i)
$F_i(x)$	the probability distribution function of object o_i (x is the possible location of o_i)
$\overline{o_i}$	the mean vector of object o_i
$\overline{p}_{c_j}$	the mean vector of the cluster representative p_{c_j}
$s_{i,t}$	the t-th sample of o_i

2.2. Clustering based on probabilistic distribution similarity

The distribution similarity appears in the task of document clustering in information retrieval. Each document is modeled as a multinomial distribution in the language model [16,17]. For example, Xu et al. [18] uses *K*-means clustering to measure the similarity between the multinomial distributions of documents by KL divergence. KL divergence is computed by using the number of occurrences of terms in documents. Other clustering methods which use KL divergence are also discussed [19,20], but none of them can handle uncertain objects. To our knowledge, Jiang et al. [21] first propose a framework of *k*-medoids and DBSCAN for clustering uncertain objects by integrating KL divergence. Different from their work, we measure the similarity based on probability distribution function between uncertain objects by using histogram intersection kernel (HIK) [22]. Moreover, large margin clustering based on spectral clustering is integrated in our work.

2.3. Fuzzy clustering

Fuzzy clustering is another related research field, which has been long studied in fuzzy logic [23–25]. In fuzzy clustering, an object can belong to more than one class with different degrees. Fuzzy *c*-means is a widely used fuzzy clustering method to generate fuzzy clusters. Other fuzzy clustering methods are also developed to handle fuzzy clustering problem. However, our work focuses on hard clustering based on uncertain objects, in which objects can only belong to one class. Fuzzy clustering methods can just process normal data while the uncertainty of objects is considered in our work. In this paper, we also demonstrate that our method can get better clustering quality than fuzzy *c*-means.

3. Preliminary knowledge

In this section, the model of uncertain objects and expected distance calculation used in UK-means are introduced in detail. And then UK-means is shown in Section 3.3. Table 1 lists the important notations used in this section.

3.1. Model of uncertain objects

In most work, an uncertain object is represented as a (finite or infinite) region, which covers the possible locations of the object (especially the number of possible locations is not finite). The region of the uncertain domain of object o_i is denoted as $UD(o_i)$. A pdf (probability density function), f_i, is used to indicate the probability density of each possible location within the region, i.e., $\int_{UD(o_i)} f_i(x)\,dx = 1$. In our work, we consider the problem of clustering uncertain objects in a multidimensional space. The multidimensional space is produced by the domains of all attributes.

A certain object is usually represented as a point in a multidimensional space. During data collection phase, the uncertain nature or system limitation is likely to cause imperfect data quality, which will lead uncertain attribute values of an object. It is reasonable to use a set of points to represent an uncertain object. Each point represents a possible location of the uncertain