A Similarity Between Uncertain Data Measurement Method Based on stochastic simulation

Yuan Cheng[1], Ronghua Chi[2] and Dapeng Lang[3]
{changuang7@sina.com[1], chironghua@126.com[2], langdapeng.heu@gmail.com[3]}

College of Computer Science and Technology, Harbin University of Science and Technology[1],
College of Computer and Information Engineering, Heilongjiang University of Science and Technology[2]
College of Computer Science and Technology, Harbin Engineering University[3]

Abstract. The distance measurement between uncertain data is an important basis for accurate clustering. Taking full advantage of the uncertainty characteristics of the object will help to represent the uncertain data more accurately and calculate its distance. Based on the probability distribution function to represent the characteristics of uncertainty distribution, this paper studies a method for measuring distance between uncertain objects based on stochastic simulation. The effectiveness of the proposed method is verified by experiments.

Keywords: uncertain data, distance measurement, probability density function, stochastic simulation.

1 Introduction

Under the background of the rapid development of the Internet+ and big data, data acquisition methods are more abundant, which may make data uncertain, including the uncertainties of data values caused by inaccurate raw data or the use of coarse-grained datasets, i.e. the uncertainties of attribute values. The existence of uncertainty increases the complexity of data, so that traditional data mining methods are difficult to directly use for mining knowledge.

In the clustering for uncertain data, due to the uncertainties in data, the distance measurement between data objects may be disturbed to also be uncertain, which may further affect the clustering results. Therefore, related researches mostly construct a data model capturing the uncertainty distribution of object, measure distance between uncertain data, and finally based on a traditional clustering algorithm, such as K-Means, DBSCAN, or OPTICS, form clustering algorithms, such as UK-Means [1], FDBSCAN [2], FOPTICS [3], etc., to cluster uncertain data [4-6]. These algorithms usually represent an uncertainty object as an area of uncertainty, and use the probability density function (Probability Density Function) to build a model that conforms to the actual probability distribution of uncertain data as much as possible. That is, the data model could describe the uncertainty characteristics of an uncertain object. When measuring the distance between uncertain objects, because of the existence of uncertainty, some studies using the expected distance method [1-4], which is calculated based on the sample points representing the uncertain area, and is linearly integrated into one scalar value. Its calculation efficiency is greatly affected by the number of sample points. Therefore, in order to improve the calculation efficiency, [1,6] reduce the amount of calculation when calculating the expected distance based on the pruning strategy, and improve the calculation efficiency of the expected distance by defining the cluster boundary. [7] calculates the expected distance based on information theory; and [8] improves the calculation efficiency of the expected distance based on the center boundary. The difference between uncertain objects is measured using the expected distance, and the distance described by probability can be linearly integrated into a scalar value. Another research idea is to calculate the similarity between objects based on uncertain data model. The uncertainty model represents the distribution characteristics of uncertainty, so the distance between uncertain objects represented by this distribution function can be obtained by comparing the differences between the distribution functions [9,10]. KL divergence is a common method for measuring the difference of probability distribution. However, a major problem of KL divergence is that it does not have symmetry, while the distance between objects in data mining needs to be symmetric. Therefore, the difference between the uncertainty distributions based on the KL divergence is not the most appropriate. We will propose a method to measure the distance between uncertain objects to solve these problems.

2 Similarity Measurement Based on Stochastic Simulation

2.1 Distance measurement between uncertain objects

The uncertain data model uses probability density function to represent the distribution characteristics of uncertainty. The difference between functions is mainly reflected as the difference of the area formed by the definition domain and the function curve. The stochastic simulation is an effective method to calculate such area. There are two probability density functions representing the uncertainties of two objects A and B in Figure 1. The difference between the two functions in the definition domain $[a,b]$ is shown as the shaded part in Figure 1, that is, the difference in area of non-intersecting parts formed by two curves and x axis in the definition domain. Therefore, we calculate the distance between uncertain data represented by the probability model based on stochastic simulation method.

Through the stochastic simulation method, we perform a large number of random samplings on the definition domain of the probability density function, and perform the integral operation on the difference of the corresponding function values to obtain an approximate difference with higher accuracy. For uncertain objects x_i x_i and y_i, the probability density functions describing their uncertainty distribution characteristics are g_{x_i} $\widehat{f}_{x_i}$ and g_{y_i}. Then we generate l random numbers: $s_j \in I_{x_i} \cup I_{y_i} \ (1 \leq j \leq l)$, in the union of their definition domains. After the random numbers are generated, we can obtain the integral of the difference between the probability density function in the specific definition domain, which represents the difference between the probability of the values of uncertain objects. The difference is the distance between uncertain data objects, which is shown in equation (1).

$$d(x_i, y_i) = (b-a) \cdot \int \left| g_{x_i}(s_j) - g_{y_i}(s_j) \right| ds \qquad (1)$$

The above method is described as follows:

```
DBU($g_{x_i}$, $g_{y_i}$, $I_{x_i}$, $I_{y_i}$, $l$)
Input: $g_{x_i}$, $g_{y_i}$: the probability density functions of uncertain objects $x_i$ $x_i$ and $y_i$;
       $I_{x_i}$, $I_{y_i}$: the definition domains of $x_i$ $x_i$ and $y_i$;
       $l$: the number of generated random numbers;
Output: $d$: the distance between $x_i$ and $y_i$.
Steps:
1) Define $[a,b] = I_{x_i} \cup I_{y_i}$;
2) Let $d = 0$;
3) for $s$ from 1 to $l$:
```

```
4)    generate $s$ random numbers in $I_{x_i} \cup I_{y_i}$;
5)    $d+=(b-a)\times\left|g_{x_i}(s)-g_{y_i}(s)\right|$;
6) end for
7) return $d$.
```

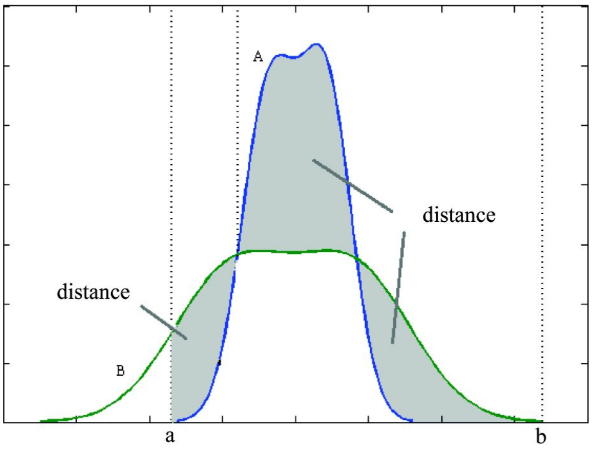

Fig.1. Distance measurement between uncertain objects

2.2 Distance measurement on clusters

In different clustering processes, in addition to measuring the distance between objects, dividing the clusters that objects belong to also involves measuring the distances about clusters, including: ① the distance between uncertain data object and cluster, ② the distance between the clusters.

In the uncertainty-oriented clustering method, the objects in the clusters are described by the uncertain data model. It is difficult to measure distance directly based on the set of uncertain models represented by the probability density function. Therefore, we can firstly extract the cluster characteristics based on uncertain data model, and then use the above DBU method to measure the distance about clusters represented by the data model.

Figure 2 shows a set of probability density functions of objects within a cluster. Firstly, based on the stochastic simulation method, l random numbers are generated in definition domain, and the values of each random number on each probability density function are obtained to the set $s_j(1\le j\le l)$. If they are regarded as the samples, we can get its uncertain data model g_{c_i} based on [11] for cluster c_i. Therefore, the clusters consisting of uncertain data can be expressed as a form of uncertain data model based on stochastic simulation, which is used to represent the characteristic of the cluster.

The distance between uncertain data object and the cluster can be converted into a distance between the uncertain data object and an uncertain model representing the cluster characteristics. While the distance between clusters can be transformed into the distance between two uncertain data models representing cluster characteristics. Then according to the above DBU method, the distance between the probability density functions, i.e. uncertain data model, could be calculated based on the random simulation.

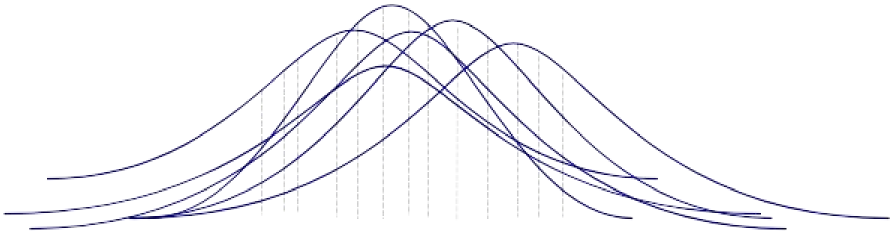

Fig.2. Extracting the probability density function of one cluster

3 Experiments

This paper will compare the effectiveness of distance measurement methods on 5 UCR datasets: Beef, Coffee, Fish, OliveOil, Trace [12], with the other two methods: expected distance [1], and KL divergence based method [10]. The accuracy of the clustering results could indicate the effectiveness of the proposed distance measurement method. We use Entropy and Purity [13] to measure the clustering accuracy. The smaller the value of the former is, the better the clustering is. The latter is the opposite. With the same clustering method, the more accurate the clustering is, the better the distance measurement is.

There is no uncertainties in the original UCR datasets, we need to construct uncertainties based on the method mentioned in [14] first. The uncertainties can be described by the samples representing the possible values, so we choose Gaussian distribution to generate s samples for each object o in dataset D. There is a parameter $\sigma \in [0, e]$ that is the unified standard deviation when using Gaussian distribution, where e represents the uncertainty. Generate a center point $g = o$, and then generate the other $s - 1$ samples based on Gaussian distribution with a σ value and $\mu = \varepsilon$ for any object o in D. If we choose different parameters, we could obtain different probability density functions for uncertainties of each object o in D. Then we can get uncertain objects with different uncertainty distribution characteristics.

Therefore, we will use the above method to generate the uncertainties for the UCR datasets, and verify the effectiveness of our proposed method, comparing with the other two distance measurement methods, based on the idea of K-Means clustering algorithm.

Figure 3 shows the comparison results of clustering accuracy. Compared with the other two distance measurement methods, it can be seen that the proposed method could obtain more accurate clustering results on the datasets. This is because the construction of uncertain data models and the distance measurement between uncertain objects both make full use of the characteristics of uncertainty.

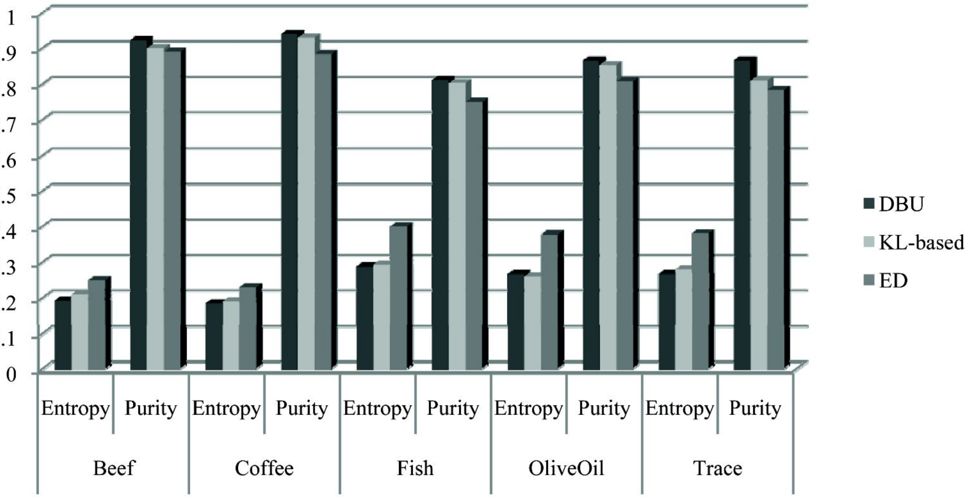

Fig.3. Comparison of clustering accuracy on UCR datasets

Figure 4 shows the comparison of the execution time of the clustering process when the number of sample points are gradually increased. The results indicate that on several datasets,

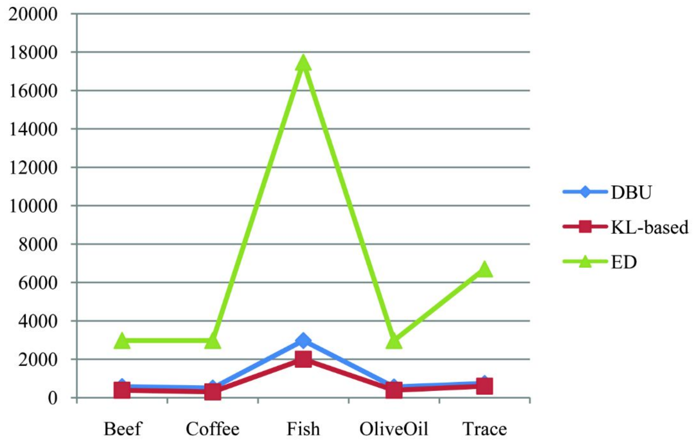

Fig.4. Comparison of clustering execution time on UCR datasets

the DBU method performs better than the expected distance, and is close to the execution time of the method based on the KL divergence. It shows that the proposed method could not only help to obtain more accurate clustering results, but also have higher calculation efficiency.

Conclusion

This study is oriented to data with attribute uncertainty. To solve the problems in the calculating distance between uncertain objects, a distance measurement method based on stochastic simulation is proposed based on the uncertainty distribution characteristics represented by probability density function. The proposed method could not only ensure the accuracy of the metric due to the calculation based on uncertainty distribution characteristics, but also improve the efficiency of the distace calculation used stochastic simulation. The validity and accuracy of the proposed method are verified through experiments on UCR datasets.

Acknowledgments. This work is sponsored by the Natural Science Foundation of Heilongjiang Province under grant number F2017015, and Training Program For Young Innovators in Heilongjiang General Institutes of Higher Education under grant number UNPYSCT-2017079.

References

[1]Ngai W, Kao B, Chui C, et al. Efficient clustering of uncertain data. Sixth International Conference on Data Mining, IEEE, 2006: 436-445.

[2]Kriegel H, Pfeifle M. Density-based clustering of uncertain data. Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, ACM, 2005: 672-677.

[3]Cormode G, McGregor A. Approximation algorithms for clustering uncertain data. Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, ACM, 2008: 191-200.

[4]Tomar D, Sathappan S. A method for handling clusturing of uncertain data. International Conference on Advances in Human Machine Interaction. IEEE, 2016: 1-5.

[5]Zhang X, Gao L, Yu H. Constraint Based Subspace Clustering for High Dimensional Uncertain Data. Pacific-Asia Conference on Knowledge Discovery and Data Mining, Auckland, New Zealand. Springer International Publishing, 2016: 271-282.

[6]Liu C, Niu Z, Liao K. Mechanisms to improve clustering uncertain data with UKmeans. Data & Knowledge Engineering, 2018, 116: 61-79.

[7]Gullo F, Ponti G, Tagarelli A, et al. An information-theoretic approach to hierarchical clustering of uncertain data. Information Sciences, 2017, 402: 199-215.

[8]Liao K T, Liu C M. An Effective Clustering Mechanism for Uncertain Data Mining Using Centroid Boundary in UKmeans. Computer Symposium (ICS), 2016 International. IEEE, 2016: 300-305.

[9]Xu L, Hu Q, Hung E, et al. Large margin clustering on uncertain data by considering probability distribution similarity. Neurocomputing, 2015, 158: 81-89.

[10]Jiang B, Pei J, Tao Y, et al. Clustering uncertain data based on probability distribution similarity. Knowledge and Data Engineering, IEEE Transactions on, 2013, 25(4): 751-763.

[11]Yuan Cheng, Ronghua Chi, Suxia Zhu, et al.: An Uncertain Data Model Construction Method Based on Nonparametric Estimation. Proceedings of 2016 IEEE International Conference on Electronic Information and Communication Technology, 2016, Harbin. USA: Institute of Electrical and Electronics Engineers Inc,pp. 384-389 (2016)

[12]Chen Y, Keogh E, Hu B, et al. The ucr time series classification archive[J]. 2015.

[13]Zhao Y, Karypis G. Criterion Functions for Document Clustering: Experiment and Analysis. Technical Report TR01-40, Department of Computer Science, University of Minnesota, USA, 2001.

[14]Züfle A, Emrich T, Schmid K A, et al. Representative clustering of uncertain data. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014: 243-252.