An Uncertain Tuple Density Clustering (UTDC) Algorithm for Uncertain Data

Chunhui Li¹, Linjiang Zheng²
College of Computer Science, Chongqing University, Chongqing, 400044, China
m13087109609@163.com

Abstract. The massive uncertain data generated by network applications has potential value, and it is of great significance to carry out clustering analysis of uncertain data. However, the uncertainty of data brings a serious challenge to traditional clustering algorithms. There are some problems in the existing clustering algorithms for uncertain data. (1) Some algorithms have a lot of meaningless distance calculations when calculating the distance of uncertain objects, and then the calculation complexity of algorithms increases. (2) The data model in clustering algorithms results in the loss of data distribution information, and then the accuracy of the algorithms decreases. In this paper, we propose an uncertain tuple density clustering (UTDC) algorithm for uncertain data. Firstly, we extract the data distribution features of uncertain instances and construct uncertain tuples by introducing the cloud model, which realizes the pruning of clustering objects. Secondly, we apply the EW distance to the traditional density clustering algorithm DBSCAN, which completes the density clustering for uncertain data. The experimental results show that comparing with UK-Means and FDBSCAN, UTDC algorithm effectively reduces the computational complexity and improves the accuracy.

1. Introduction
The existing clustering algorithms mainly focus on certain data. However, due to various reasons such as the inaccuracy of sensors in the data acquisition process or desensitization of privacy data [1], a large number of data is uncertain in practical applications. Uncertain data uses probability function to describe the possibility of event attributes, and it cannot give a specific value. This results in the fact that the accuracy of clustering algorithms is lower with uncertain data as the clustering object, then the algorithm cannot get a satisfactory clustering effect. Therefore, it is necessary to study the clustering algorithms for uncertain data.

In recent years, people have carried out related research on clustering algorithms for uncertain data. Someone proposes extending traditional clustering algorithms to realize the clustering task for uncertain data. For example, k-means and k-medoids are extend to UK-Means [2] and UK-Medoids [3] by introducing the expected distance or the uncertain distance, so as to achieve clustering for uncertain data. Kao B et al. [4] introduces Voronoi to prune clustering objects, which optimizes the performance of UK-Means. DUK-Means [5] algorithm is an improved algorithm for the UK-Means algorithm to adapt to the distributed network environment. However, these improved algorithms based on partition idea are not ideal in dealing with spherical clusters. It is difficult to find clusters of arbitrary shape for them.

And some traditional density-based clustering algorithms have also been extended to uncertain data clustering. FDBSCAN and FOPTICS [6] are extensions of the traditional density clustering algorithms DBSCAN and OPTICS, which use the probability density function to cluster uncertain data. Jin [7]
offers an efficient summary data structure for clustering data with two uncertainties. Dallachiesa M [8] raises an uncertain data clustering algorithm based on density grid, which can find clusters of arbitrary shape. To reduce the impact of uncertainty on clustering results, Han [9] proposes the concept of uncertainty, and gives different treatment methods for uncertain data with different distribution ranges. These improved density-based clustering algorithms can find clusters of arbitrary shape, but they usually calculate the distance density function among uncertain objects by sampling, which may lead to the loss of uncertain information. It may cause the accuracy of the algorithm to decrease.

In addition, Volk [10] uses the possible world model to cluster uncertain data. And REP [11] algorithm introduces the information theory to calculate the distance between possible worlds. Jiang [12] applies KL divergence as the similarity measure to clustering algorithms for uncertain data. However, these algorithms generate a large number of possible worlds in the clustering process and take up too much computing resources, which results in poor timeliness of the algorithm. Meanwhile, these algorithms cannot provide stable clustering results due to the insufficient consideration of uncertain data distribution.

In view of the existing problems of uncertain clustering algorithm, we propose an uncertain tuple density clustering (UTDC) algorithm for uncertain data, the algorithm mainly carries out the following work:

• In order to avoid the loss of uncertain data distribution information, we introduce the cloud model to construct uncertain tuples that are used to describe the distribution features of uncertain instances in the dataset. And the transformation from uncertain instances to uncertain tuples greatly reduces the number of initial clustering objects in the algorithm which ensures the timeliness of UTDC algorithm.

• In order to calculate the distance between uncertain data, we modify the $EW$ distance and integrate it into the traditional density clustering algorithm DBSCAN. It effectively calculates the distance density between uncertain objects, and then the accuracy of UTDC algorithm is guaranteed.

The experiment compares the computational complexity and accuracy between UTDC algorithm and other uncertain clustering algorithms. The performance of UTDC is verified by comparing the running time with UK-Means and FDBSCAN under different datasets, then the complexity of the algorithm is reduced. Then the effectiveness of UTDC is verified by comparing the precision and recall with UK-Means and FDBSCAN, then the accuracy of the algorithm is improved.

2. Preliminaries

2.1. Related definition

In most current studies, the uncertainty of data is divided into existence uncertainty and value uncertainty [13]. The data object of this paper is value uncertainty data characterized by discrete Probability Density Function. It is expressed as the fact that for any instance $x_{ij}$ in an uncertain dataset $D$, its probability is represented as $f(x_{ij}) > 0$, and $\int_{x_{ij} \in X_i} f(x) dx = 1$.

**Definition 1. (Uncertain tuple)** Given an uncertain dataset $D = \{X_1,X_2,...,X_n\}$. $X_i$ is the $i$th uncertain tuple of the set $D$.

**Definition 2. (Uncertain instance)** Uncertain tuple $X_i = \{(x_{ij},p_{ij})\} (j = 1,2,...,s)$ contains $s$ instances, which are represented as binary tuples $(x_{ij},p_{ij})$. Where $p_{ij} \in (0,1)$, which means the probability of the value $x_{ij}$ is $p_{ij}$, and $\sum_{j=1}^{s} p_{ij} = 1$.

**Definition 3. (Uncertain instance dimension)** The value $x_{ij} = (x_{ij}^1,x_{ij}^2,...,x_{ij}^d)$ of the uncertain instance $(x_{ij},p_{ij})$ is a $d$-dimensional vector, and $x_{ij}^d$ denotes the value of the current vector in the $d$th dimension.

2.2. Unaddressed issues

The data model results in the loss of uncertain data distribution information. Different from the clustering algorithms for certain data, the clustering algorithms for uncertain data need to fully
consider the distribution features of objects. As shown in table 1, uncertain tuples $X_1$ and $X_2$ consisting of 10 two-dimensional discrete instances. Their centroid is (0.5,0.5), but the interval distributions of discrete instances are $X_1 = ([0.43,0.57], [0.43,0.57])$ and $X_2 = ([0.46,0.54], [0.48,0.52])$ respectively. If the data model only considers the centroid of uncertain tuples and ignores the interval distribution of uncertain instances, $X_1$ and $X_2$ will be regarded as the same tuple for uncertain data clustering, which results in the result accuracy deviation.

**Meaningless distance calculation increases the algorithm complexity.** For example, the computational complexity of DBSCAN algorithm for certain data is $O(n \log n)$, where $n$ is the number of clustering objects. Considering the multi-instance of uncertain data, the complexity of the algorithm will increase significantly. Taking the dataset $D$ with $n$ tuples that each tuple consists of $s$ instances as an example. If uncertain instances are directly used as clustering objects for the algorithm, the complexity of the algorithm will increase to $O((n \times s)\log(n \times s))$. And some meaningless distance calculation of instances in this process obviously has a negative impact on the computational complexity of the algorithm.

Due to the above defects in clustering algorithms for uncertain data, we need to propose a more efficient clustering algorithm.

**Table 1. Uncertain tuples with the same centroid and different instance intervals.**

| Uncertain tuples $X_1$ | Uncertain instances $X_1$ |
|------------------------|--------------------------|
| $(x_{11},x_{12})$      | $(0.50,0.50)$            |
| $(x_{12},x_{13})$      | $(0.44,0.43)$            |
| $(x_{13},x_{14})$      | $(0.43,0.56)$            |
| $(x_{14},x_{15})$      | $(0.57,0.44)$            |
| $(x_{15},x_{16})$      | $(0.56,0.57)$            |

| Uncertain tuples $X_2$ | Uncertain instances $X_2$ |
|------------------------|--------------------------|
| $(x_{22},x_{23})$      | $(0.46,0.48)$            |
| $(x_{23},x_{24})$      | $(0.53,0.48)$            |
| $(x_{24},x_{25})$      | $(0.50,0.50)$            |
| $(x_{25},x_{26})$      | $(0.54,0.52)$            |
| $(x_{26},x_{27})$      | $(0.47,0.52)$            |

3. **UTDC algorithm**

3.1. **The framework of UTDC**

The uncertain tuple density clustering (UTDC) algorithm for uncertain data consists of two parts: **Building uncertain tuples** and **Integration of the density algorithm**. Firstly, we extract the distribution features of uncertain instances in all uncertain regions, and then build an uncertain tuple that can represent the uncertain region based on the cloud model. In this process, the uncertain tuples constructed based on the cloud model complete the transformation from qualitative concepts to quantitative data, which completes the pruning of clustering objects. It effectively reduces the initial input quantity of subsequent clustering processes and ensures the efficiency of the algorithm. Secondly, we integrate the $EW$ distance into the density-based clustering algorithm—DBSCAN. The traditional clustering algorithm DBSCAN can calculate the distance density between certain data, but it can't calculate the distance between uncertain tuples. So we integrate $EW$ distance into DBSCAN to calculate the distance density between uncertain objects, which ensure the accuracy of DBSCAN algorithm on uncertain datasets.

3.2. **Building uncertain tuples**

In order to make full use of the distribution features of uncertain data in the clustering process, we propose a binary group $(E(X_i), W(X_i))$ based on the idea of the cloud model [14] to describe uncertain tuples that represent the attributes of events. And $E(X_i)$ represents the basic certainty of the data attributes, which is the spatial distribution expectation of the uncertain tuple $X_i$, i.e., it is the most representative point of the current attribute. And $W(X_i)$ represents the uncertainty measure of data attributes, which is the range feature of the uncertain tuple $X_i$, i.e., it reflects the random distribution of the current attribute. For the uncertain tuple $X_i = \{(x_{i1}, p_{i1}), ..., (x_{ij}, p_{ij})\} (j = 1, 2, ..., s)$, its binary group $(E(X_i), W(X_i))$ is calculated as follows:
\[ E(X_i) = \sum_{m=1}^{s} (x_{im} * p_{im}) \]

\[ W(X_i) = \left\{ \left( \frac{p(x_{im} - x_{in})}{\max |x_{im}, x_{in} \in X_i, p \geq 1} \right) \right\} \]

In equation (1), \(x_{im}\) and \(p_{im}\) represent the uncertain instances and their corresponding probabilities under the uncertain tuple \(X_i\) respectively. The average value of uncertain instances is taken as the representative value of the current uncertain region by multiplying all instance values and corresponding probabilities and then summing them. Equation (2) calculates the range value of \(X_i\) that represents the current uncertain region, i.e., it is the maximum distance between uncertain instances belonging to the same tuple. Then we use the interval width to describe the instances distribution features in the uncertain region.

Based on the uncertain data represented by the cloud model, we design a selection strategy for building uncertain tuples (as Algorithm 1). The algorithm can find the outliers in the uncertain region, so as to ensure the validity of the uncertain tuples in the subsequent clustering process.

**Algorithm 1: Building Uncertain Tuples Algorithm**

**Input:** \(D\): uncertain data set, \(d\): distance threshold, \(P_i\): probability screening threshold

**Output:** uncertain tuples representing uncertain regions

**Method:**
1. do
2. Randomly select an uncertain region that is marked as unvisited;
3. for Each uncertain instance \((x_{ij}, p_{ij})\) in \(i\)
4. calculate the distance from \((x_{ij}, p_{ij})\) to other uncertain instances in \(i\);
5. if The minimum distance of \((x_{ij}, p_{ij})\) \(d(x_{ij})_{\min} > d\)
6. Add \((x_{ij}, p_{ij})\) to the candidate outlier set \(C\);
7. if The probability of \((x_{ij}, p_{ij})\) in \(C\) \(p_{ij} < P_i\)
8. \((x_{ij}, p_{ij})\) is an outlier, which is not included in the uncertain tuple \(X_i\);
9. end for
10. Calculate \(E(X_i)\) and \(W(X_i)\) of all uncertain instances except outliers in the region \(i\);
11. **Output** the tuple \(X_i\) representing the region \(i\), which is a binary group \((E(X_i), W(X_i))\);
12. Mark \(i\) is visited;
13. until No uncertain regions marked unvisited.

### 3.3. Integration of the density algorithm

DBSCAN algorithm usually uses the **Euclidean** distance to calculate the distance density between certain data. But the binary group \((E(X_i), W(X_i))\) contains the distribution features of uncertain data, and the **Euclidean** distance cannot calculate the distance density between uncertain tuples. In order to solve the above problem, we introduce the interval distance—**EW** distance, which can calculate the distance density of the uncertain data represented by the binary group \((E(X_i), W(X_i))\). Given uncertain tuples \(X_i = \{(x_{i1}, p_{i1}), ..., (x_{is}, p_{is})\}\) and \(X_j = \{(x_{j1}, p_{j1}), ..., (x_{js}, p_{js})\}\) the **EW** distance [15] between \(X_i\) and \(X_j\) is:

\[ \text{dist}^P_{\text{EW}}(X_i, X_j) = \sqrt{\left| E(X_i) - E(X_j) \right|^p + \frac{1}{3} \left| W(X_i) - W(X_j) \right|^p}, p \geq 1 \]

In equation (3), \(E(X_i) = \sum_{m=1}^{s} (x_{im} * p_{im})\) and \(E(X_j) = \sum_{n=1}^{s} (x_{jn} * p_{jn})\) are the expected values of \(X_i\) and \(X_j\) respectively, then \(W(X_i) = \left\{ \left( \frac{p(x_{im} - x_{in})}{\max |x_{im}, x_{in} \in X_i, p \geq 1} \right) \right\} \) and \(W(X_j) = \left\{ \left( \frac{p(x_{jm} - x_{jn})}{\max |x_{jm}, x_{jn} \in X_j, p \geq 1} \right) \right\} \) are the range values of \(X_i\) and \(X_j\) respectively.
Therefore, the steps of In-DBSCAN (Integrated DBSCAN Algorithm) based on EW distance are as Algorithm 2. Because uncertain tuples described by the expectation and distribution features of uncertain instances are used as clustering objects for In-DBSCAN algorithm, the algorithm can effectively reduce the computational complexity (from \(O((n \times s)\log(n \times s))\) to \(O(n \log n)\). It improves the efficiency of the uncertain data clustering algorithm.

Algorithm 2: Integrated DBSCAN Algorithm

**Input:** \(T\): uncertain tuples in the uncertain dataset, \(r\): maximum radius of neighbor, \(MinPts\): neighbor density threshold

**Output:** density-based cluster sets

**Method:**
1. Mark all uncertain tuples as unvisited;
2. do
3. Randomly select a tuple \(X_i\) that is marked as unvisited;
4. Mark \(X_i\) is visited;
5. Calculate the distance density of \(X_i\) based on EW distance;
6. if The \(r\)-neighbor of \(X_i\) has at least \(MinPts\) tuples
7. Create a new cluster \(C\) and add \(X_i\) to \(C\);
8. Let \(N\) be the set of tuples in the \(r\)-neighbor of \(X_i\);
9. for Each uncertain tuple \(X_i'\) in \(N\)
10. if \(X_i'\) is unvisited
11. Mark \(X_i'\) is visited;
12. Calculate the distance density of \(X_i'\) based on EW distance;
13. if The \(r\)-neighbor of \(X_i'\) has at least \(MinPts\) tuples
14. Add these tuples to \(N\);
15. if \(X_i'\) is not a member of any cluster yet
16. Add \(X_i'\) to \(C\);
17. end for
18. Output \(C\);
19. else Mark \(X_i\) as the noise point
20. until No uncertain tuples marked unvisited.

### 4. Experiments

#### 4.1. Datasets

The original dataset used in this paper is *Facebook Live Sellers in Thailand Data Set*. Based on the original dataset, uncertain datasets are synthesized as experimental datasets. The synthesis strategy is as follows: Read a record \(x_i = (x_{i1}^1, x_{i2}^2, ..., x_{id}^d)\) in the original dataset, where \(d\) represents the dimension of the record. For \(x_{ik}^k\) of the original record \(x_i\), it is perturbed by three parameters \(\alpha, \beta, \gamma\). That is, the uncertain value \(x_{ij}^k\) corresponding to the \(k\)-dimension value \(x_{ik}^k\) satisfies: \(x_{ij}^k = x_{ik}^k \times (1 + \gamma \times \alpha + \beta), \gamma \in [-1,1]\). Perform the same perturbation on the other dimensions of \(x_i\), and construct \(s\) uncertain instances \(x_{ij}\). Then, add a probability dimension \(p_{ij} (\Sigma_{j=1}^s p_{ij} = 1)\) to each perturbed uncertain instance \(x_{ij}\), representing the probability of the current uncertain instance. Finally, generate the corresponding uncertain tuple \(X_i = \{(x_{i1}, p_{i1}), ..., (x_{ij}, p_{ij})\} (j = 1, 2, ..., s)\) by perturbing the original record \(x_i\), and all the uncertain tuples constitute the uncertain dataset \(D = \{X_1, X_2, ..., X_i, ..., X_n\}\).

#### 4.2. Performance Analysis

The experiment verifies the performance of UTDC from two aspects: computational complexity and accuracy. And we use running time to evaluate the computational complexity, then we use precision and recall to evaluate the accuracy of the algorithm. For the above indexes, we compare UTDC with
UK-Means and FDBSCAN. The experimental results show that UTDC algorithm has certain advantages.

Computational complexity. Figure 1 and figure 2 show the running time comparison between UTDC algorithm with UK-Means algorithm and FDBSCAN algorithm on uncertain datasets. Figure 1 shows the running time comparison of various algorithms in case where the number of uncertain tuples $X_i$ is 7050 and the number of uncertain instances $s$ is increased from 1 to 5. It can be seen from the figure 1 that the increase in the number of uncertain instances does not have a significant impact on the running time of UTDC algorithm, and the running time is relatively stable. But the running time of UK-Means and FDBSCAN algorithm shows an increasing trend. This is because the uncertain tuples based on cloud model effectively reduce the initial input of UTDC algorithm, and uncertain instance quantity has less effect on the running time of the algorithm. UK-Means algorithm and FDBSCAN algorithm take uncertain instances as clustering objects, so the number of uncertain instances directly affects the efficiency of the algorithm, and the running time of the algorithm is relatively increased.

![Figure 1. Running time (amount=7050).](image1)

![Figure 2. Running time (amount=100000).](image2)

And the number of uncertain tuples $X_i$ is increased to 100000 and the number of uncertain instances $s$ is increased from 1 to 5, then the running time of three algorithms is shown in figure 2. The overall running time of UK-Means algorithm and FDBSCAN algorithm is higher than that of UTDC algorithm, and their growth trends are more obvious. The running time of UTDC algorithm is still less affected by uncertain instance quantity, and it is relatively stable.

Accuracy. Figure 3 and figure 4 show precision and recall between UTDC algorithm with UK-Means algorithm and FDBSCAN algorithm on uncertain datasets. As shown in figure 3 and figure 4, the precision and recall of UTDC algorithm are higher than that of UK-Means algorithm. UTDC algorithm uses $EW$ distance to calculate the distance density between the uncertain tuples including data distribution information of instances, which effectively reduces the impact of uncertainty on the clustering process, so the clustering quality is higher. However, UK-Means algorithm and FDBSCAN algorithm take uncertain instances as the clustering objects, and the utilization rate of data distribution information is low, which affects the clustering quality.

![Figure 3. Precision.](image3)

![Figure 4. Recall.](image4)

5. Conclusion
Aiming at how to effectively utilize the data distribution features contained in the uncertain dataset to complete the uncertain data clustering, we build an uncertain tuple model based on the cloud model, which reduces the negative impact of uncertain instances with random distribution features. Then we introduce the $EW$ distance to integrate the traditional density clustering algorithm DBSCAN, so that it
can calculate the distance density of uncertain tuples through the binary group \((E(X_i), W(X_i))\), and then complete the density clustering of uncertain data. Experimental results show that UTDC algorithm can effectively reduce the computational complexity, and the accuracy of clustering results is improved compared with other algorithms. In the future, we will carry out the relevant research on density clustering for uncertain data streams.

**Acknowledgments**
The paper was supported by the National Key R&D Program of China (No. 2017YFC0805208), the National Key R&D Program of China (No. 2016YFC0801707).

**References**
[1] Charu C. Aggarwal., Philip S. Yu. (2009) A survey of uncertain data algorithms and applications. IEEE Transactions on Knowledge and Data Engineering. 21(5): 609–623.
[2] Michael Chau., Reynold Cheng., Ben Kao., and Jackey Ng. (2006) Uncertain data mining: An example in clustering location data. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. Singapore, pp. 199–204.
[3] Francesco Gullo., Giovanni Ponti., and Andrea Tagarelli. (2008) Clustering uncertain data via K-medoids. In: Proceeding SUM ’08 Proceedings of the 2nd international conference on Scalable Uncertainty Management. Naples, pp. 229–242.
[4] Kao B., Lee S D., Cheng D W., et al. (2008) Clustering uncertain data using Voronoi diagrams. In: Proceedings of the 8th IEEE International Conference on Data Mining. Pisa, pp. 333-342.
[5] Jin Zhou., Long Chen., C. L. Philip Chen., Yingxu Wang., and Han-Xiong Li. (2018) Uncertain data clustering in distributed peer-to-peer networks. IEEE Trans Neural Netw Learn Syst. 29(6): 2392–2406.
[6] Hans-Peter Kriegel., Martin Pfeifle. (2005) Density-based clustering of uncertain data. In: Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, pp. 672–677.
[7] Jin Cheqing., Yu J X., Zhou Aoying., et al. (2014) Efficient clustering of uncertain data streams. Knowledge and Information Systems. 40(3): 509-539.
[8] Dallachiesa M., Jacques-Silva G., Gedik B., et al. (2015) Sliding windows over uncertain data streams. Knowledge and Information Systems. 45(1): 159-190.
[9] Han Donghong., Song Ming., Zhang Hongliang., et al. (2017) Algorithm for clustering uncertain data streams based on density. Journal of Tsinghua University (Science and Technology). 57(8): 884-891.
[10] Peter Benjamin Volk., Frank Rosenthal., Martin Hahmann., Dirk Habich., and Wolfgang Lehner. (2009) Clustering uncertain data with possible worlds. In: Proceedings of the 25th International Conference on Data Engineering. Shanghai, pp. 1625–1632.
[11] Züflie, A., Emrich, T., Schmid, K. A., Mamoulis, N., Zimek, A., and Renz, M. (2014) Representative clustering of uncertain data. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, pp. 243–252.
[12] Jiang, B., Pei, J., Tao, Y., and Lin, X. (2013) Clustering uncertain data based on probability distribution similarity. IEEE Transactions on Knowledge and Data Engineering. 25(4): 751–763.
[13] Zhou Ao-ying., Jin Che-qing., Wang Guo-ren., et al. (2009) Summary of research on uncertain data management technology. Chinese Journal of Computers. 32(1): 1-16.
[14] Yang Jie., Wang Guoyin., Liu Qun., et al. (2018) Retrospect and Prospect of Research of Normal Cloud Model. Chinese Journal of Computers. 41(3): 724-744.
[15] Bao Yu-e., Peng Xiaoqin., Zhao Bo. (2013) The Interval Number Distance and Completeness Based on the Expectation and Width. Fuzzy Systems and Mathematics. 27(6): 133-139.