Uncertain Data Clustering in Distributed Peer-to-Peer Networks

Jin Zhou, *Member, IEEE*, Long Chen, *Member, IEEE*, C. L. Philip Chen, *Fellow, IEEE*, Yingxu Wang, and Han-Xiong Li, *Fellow, IEEE*

***Abstract*— Uncertain data clustering has been recognized as an essential task in the research of data mining. Many centralized clustering algorithms are extended by defining new distance or similarity measurements to tackle this issue. With the fast development of network applications, these centralized methods show their limitations in conducting data clustering in a large dynamic distributed peer-to-peer network due to the privacy and security concerns or the technical constraints brought by distributive environments. In this paper, we propose a novel distributed uncertain data clustering algorithm, in which the centralized global clustering solution is approximated by performing distributed clustering. To shorten the execution time, the reduction technique is then applied to transform the proposed method into its deterministic form by replacing each uncertain data object with its expected centroid. Finally, the attribute-weight-entropy regularization technique enhances the proposed distributed clustering method to achieve better results in data clustering and extract the essential features for cluster identification. The experiments on both synthetic and real-world data have shown the efficiency and superiority of the presented algorithm.**

***Index Terms*— Attribute weight entropy, distributed clustering, peer-to-peer (P2P) networks, uncertain data.**

I. Introduction

CLUSTERING has emerged as an essential data mining technique for statistical analysis, pattern recognition, and image segmentation [1]–[3]. It partitions the data into clusters according to the similarities between objects and helps in extraction of new information or discovering new patterns. In the past few decades, a large number of clustering algorithms have been proposed [4], in which the K-means algorithm [5] is one well-known clustering method. Then the variants of this algorithm are further discussed in [6] and [7], and the strong consistency of this method has been proved in [8] and [9].

However, in many real applications today, like sensor monitoring and location-based services [10], data mostly contains inherent uncertainty due to the random nature of the data generation, measurement inaccuracy, sampling discrepancy, data staling, and other errors. Generally, with uncertainty, the data object is no longer a single point in space but is represented by a probability density function (pdf) [11]. The traditional clustering algorithms are limited to considering geometric-distance-based similarity measures between certain data points, and cannot efficiently evaluate the difference between uncertain data objects. Lots of new clustering algorithms for uncertain data have been proposed to tackle this issue [12].

Early studies on uncertain data clustering are mainly various extensions of traditional clustering algorithms for certain data, by defining new similarity measurements between uncertain data objects, including the ED-based similarity [13], the density-based similarity [14], and the distribution-based similarity [15]. Chau *et al.* [13] propose the first ED-based clustering algorithm for uncertain data named the uncertain K-means (UK-means) algorithm. It enhances the traditional k-means algorithm with the use of a new distance-based similarity, i.e., the expected distance (ED), to handle the data uncertainty. Then, some improved algorithms are put forward to reduce the complexity of ED calculations by using some pruning tricks [16], [17] or by speeding up the ED calculation itself [18]. The work [19] reduces the UK-means algorithm to the certain K-means (CK-means) algorithm by replacing each uncertain data object with its expected centroid, thereby tremendously decreasing the computational complexity for ED calculation. For the density-based clustering, Kriegel and Pfeifle [14] define two fuzzy distance functions, i.e., the distance density function and the distance distribution function, to express the similarity between uncertain data objects, and they also integrate these new distance functions into the hierarchical clustering method [20]. Different from these two kinds of similarities above, the clustering algorithms with distribution-based similarity consider using divergences to measure the similarity between data objects. Most early researches usually utilize Kullback–Leibler (KL) divergence or Bregman divergence to cluster the object with known distribution [21], [22]. A recent work on uncertain data clustering is based on probability

Manuscript received December 2, 2015; revised June 21, 2016 and November 15, 2016; accepted February 25, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 61640218, Grant 61573166, Grant 61572540, and Grant 61673405, in part by the Project of Shandong Province Higher Educational Science and Technology Program under Grant J16LN07, and in part by the Macau Science and Technology Development under Grant 019/2015/A and UM Multiyear Research Grants.

J. Zhou and Y. X. Wang are with the Shandong Provincial Key Laboratory of Network-Based Intelligent Computing, University of Jinan, Jinan 250022, Shandong, China, and also with the School of Information Science and Engineering, University of Jinan, Jinan 250022, Shandong, China (e-mail: ise_zhouj@ujn.edu.cn; 1050352357@qq.com).

L. Chen is with the Faculty of Science and Technology, University of Macau, Macau 99999, China (e-mail: longchen@umac.mo).

C. L. Philip Chen is with the Faculty of Science and Technology, University of Macau, Macau 99999, China, and UMacau Research Institute, Zhuhai 519080, China, and also with Dalian Maritime University, Dalian 116026, China (e-mail: philip.chen@ieee.org).

H.-X. Li is with the Department of Systems Engineering and Engineering Management, City University of Hong Kong, Hong Kong (e-mail: mehxli@cityu.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2017.2677093

distribution similarity [15], in which the uncertain data object is modeled as a random variable following a probability distribution and then the KL divergence is used to directly compute the probability distribution similarity between uncertain data objects. All these methods have a common characteristic: It is that they are all based on centralized operation, i.e., data sets are of small manageable sizes, usually residing on one central site, and a single process performs clustering on the data.

In recent years, with the increasing number of real applications on distributed peer-to-peer (P2P) networks, uncertain data analysis in large dynamic networks is likely to garner increasing importance in the near future [23]. For example, in a hotel booking system, customers are asked to evaluate hotels through a series of indicators, such as facility information, sanitary condition, service quality, and location information. Each hotel can be scored by many customers. All evaluations to a hotel should be modeled as an uncertain object on the customer score space. In reality, hotels may be registered in the different distributed sites that provide reservation service. An important analysis work is to cluster the hotels from all sites according to customer's evaluation information. As another example, a city usually deploys multiple dispersed weather monitoring stations. Each station will monitor the daily weather conditions, such as temperature, humidity, wind speed, and so on. The daily weather data varies from day to day. A period of the continuous weather monitoring records (e.g., one month) can also be modeled as an uncertain object, represented by a pdf. Performing clustering on the weather condition data can reveal interesting insights on the weather correlation between different regions of the city in different months. In these new applications, data sources are distributed over a large network containing no special central control. The traditional centralized clustering approaches for uncertain data have shown the weaknesses: 1) raw information sharing is discouraged due to the confidentiality and security requirements in distributed P2P networks; 2) effective data collection from all peers to the central site is not guaranteed due to the energy or bandwidth limitations; and 3) high-computational complexity with large data sets. These motivate seeking a new clustering algorithm in distributed network environments for uncertain data, i.e., the distributed uncertain data clustering.

Actually, in the last decades, a great deal of attention has been paid to the distributed data clustering on P2P networks [24]. Datta *et al.* [25] propose one of the first distributed algorithms for P2P systems, named the P2P K-means algorithm. This method predetermines the same initial cluster prototypes at all peer nodes. Moreover, the update of cluster prototypes at each peer is just to calculate the mean of the data itself and the data of its neighbor peers, not considering the consensus constraint of cluster prototypes among neighboring peers. In [26], a good solution for distributed clustering in Wireless Sensor Networks (WSNs) is presented by recasting the global clustering to a set of smaller local clustering problems with consensus constraints. But the complex definition of the consensus constraint of the cluster prototypes among peers increases the computational complexity of the algorithm. Recently, Pedrycz [27], and Pedrycz and Rai [28] introduce a new distributed clustering architecture, named collaborative clustering, to operate on the separate subsets of data collaboration by exchanging information of local partition matrices. Hammouda [29] apply the collaborative pattern to the distributed document clustering, realizing the merge of peer documents into local clusters via the exchange of cluster keyphrase summaries. However, these collaborative approaches consider the fully connected network structure and show the limitation for applications with large, dynamic network.

One more important issue is that the existing distributed clustering algorithms on P2P networks are all implemented for certain data. To the best of our knowledge, this paper is the first to study uncertain data clustering on distributed network environments. We propose a novel distributed clustering algorithm for uncertain data, named the distributed UK-means (DUK-means) algorithm, which searches the global clusters by capitalizing on the consensus constraint formulation and the collaboration between neighboring peers. In this algorithm, the local clustering is performed independently at each peer with its optimization pursuits by integrating the local data objects and the cluster prototype findings exchanged with the neighbor peers, until reaching the global consensus of all peers. In addition, considering the computational complexity of the DUK-means algorithm caused by the calculation of the distance similarity between uncertain data objects, we reduce the DUK-means algorithm to its deterministic form by replacing each uncertain data object with its expected centroid. Moreover, the existing clustering algorithms for uncertain data often treat all features equally in deciding the cluster memberships of objects. This is not desirable in some applications, e.g., high-dimensions data clustering, where the cluster structure in the data set is often limited to a subset of features rather than the entire feature set. Subspace clustering provides an effective solution to discover the clusters in different subspaces within a data set [30]. With reference to [31] and [32], we apply the attribute-weight-entropy regularization technique to the DUK-means algorithm to achieve the ideal distribution of attribute weights. Better clustering results are obtained, and the essential features are exacted for cluster identification. The experiments on both synthetic and real-world data sets have shown the efficiency and superiority of the presented algorithms.

The rest of this paper is organized as follows. The DUK-means algorithm is presented in Section II. In Section III, we reduce the DUK-means algorithm to its deterministic form. In Section IV, the attribute-weight-entropy regularization technique is incorporated into the proposed DUK-means algorithm. Section V illustrates the experiment results of different clustering algorithms on synthetic and real world data sets. Finally, conclusions are drawn in Section VI.

II. Distributed Uncertain Data Clustering

A. Probability Density Function of Uncertain Data Object

Given the object set $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_N\}$, N is the number of uncertain data objects. Normally, each uncertain data object $\mathbf{o}_n (1 \leq n \leq N)$ is a random variable $\mathbf{x}_n$ following a probability distribution in a continuous M-dimensional space $\mathbf{R}^M$, which is described by a pdf f_n. Usually, the pdf f_n is unknown in real applications. Therefore, in this paper, the pdf f_n of the uncertain data object $\mathbf{o}_n$ is estimated by kernel density estimation [33], [34], which is the sum of kernel functions as

$$f_n(\mathbf{x}_n) = \frac{1}{|S_n|\,(2\pi)^{M/2}\prod_{m=1}^{M}\sigma_m}\sum_{\mathbf{e}\in S_n}\prod_{m=1}^{M} e^{-\frac{(x_{nm}-e_m)^2}{2\sigma_m^2}} \tag{1}$$

where the sample point $\mathbf{e} \in S_n$ is one observation of the random variable $\mathbf{x}_n$ and is denoted by a M-dimensional vector, i.e., $\mathbf{e} = [e_1, e_2, \ldots, e_M]$. S_n is a sample set that denotes all the samples observed for the random variable $\mathbf{x}_n$ and the size of the sample set is represented by $|S_n|$. With the assumption that the components of each data object are conditionally independent, each kernel function in (1) is the product of M Gaussian kernel functions. The *mth* Gaussian kernel function is centered at e_m with variance σ_m. σ_m is called the bandwidth and is set to $1.06 \times \delta_m |S_n|^{-(1/5)}$ according to the Silverman approximation rule [34], where δ_m is the standard deviation of the *mth* dimension of the sample points in S_n, i.e., $\delta_m = ((1/|S_n|)\sum_{\mathbf{e}\in S_n}(e_m - \mu_m)^2)^{1/2}$, where $\mu_m = \sum_{\mathbf{e}\in S_n} e_m$.

With the definition of pdf, we have

$$f_n(\mathbf{x}_n) > 0 \quad \forall \mathbf{x}_n \in R^M \tag{2}$$

and

$$\int_{\mathbf{x}_n\in R^M} f_n(\mathbf{x}_n)d\mathbf{x}_n = 1. \tag{3}$$

In the simulation part of this paper, even though we know the preset distribution of data objects to emulate the situation of real applications, we only randomly get some samples according to the preset distribution and use (1) to approximate the preset distribution. The number of samples in experiments is simply fixed to 1000, i.e., $|S_n| = 1000$.

B. Distributed Uncertain K-Means Clustering Algorithm

Consider a distributed P2P network with J peers, where each peer j $(1 \le j \le J)$ is allowed to communicate only with its one-hop neighbors $i \in NB_j$. The distributed P2P network is assumed connected, meaning that there is at least a multihop communication path between any two peers. In our research, the distributed P2P network is considered to collect the uncertain data objects and perform the clustering task. Each peer j consists of a set of N_j uncertain data objects $\{\mathbf{o}_{jn} | 1 \le n \le N_j\}$. Each object $\mathbf{o}_{jn}$ is a continuous random variable $\mathbf{x}_{jn}$ following a pdf f_{jn}. We assume each peer j has the data from each of the K clusters. Each cluster prototype is denoted by $\mathbf{c}_{jk} = [c_{jk1}, c_{jk2}, \ldots, c_{jkM}]$ for $1 \le k \le K$. The new ED between the uncertain object $\mathbf{o}_{jn}$ and the cluster prototype $\mathbf{c}_{jk}$ is defined as

$$\mathrm{ED}(\mathbf{o}_{jn}, \mathbf{c}_{jk}) = \int_{\mathbf{x}_{jn}\in R^M}\sum_{m=1}^{M}(x_{jnm} - c_{jkm})^2 f_{jn}(\mathbf{x}_{jn})d\mathbf{x}_{jn}. \tag{4}$$

Then with the goal of minimizing ED in a distributive mode, the distributed uncertain data clustering problem can be formulated as

$$\begin{aligned} \min\ F(\boldsymbol{U}, \boldsymbol{C}) &= \sum_{j=1}^{J}\sum_{n=1}^{N_j}\sum_{k=1}^{K} u_{jnk} ED(\mathbf{o}_{jn}, \mathbf{c}_{jk}) \\ &= \sum_{j=1}^{J}\sum_{n=1}^{N_j}\sum_{k=1}^{K} u_{jnk}\int_{\mathbf{x}_{jn}\in R^M} \\ &\quad \times \sum_{m=1}^{M}(x_{jnm} - c_{jkm})^2 f_{jn}(\mathbf{x}_{jn})d\mathbf{x}_{jn} \\ \text{s.t. } c_{jkm} &= c_{ikm}, \quad i \in NB_j \end{aligned} \tag{5}$$

where $\boldsymbol{U} = [u_{jnk}]$ is the membership degree matrix and $\sum_{k=1}^{K} u_{jnk} = 1, u_{jnk} \in \{0, 1\}$, $u_{jnk} = 1$ means that the nth data object is assigned to the kth cluster in the jth peer and vice versa. $\boldsymbol{C} = [c_{jkm}]$ is the cluster prototype matrix and c_{jkm} denotes the mth dimension of the kth cluster prototype in the jth peer.

In this new objective function, given the f_{jn} is a mixture of Gaussians, we can derive the integral directly. The consensus constraint $c_{jkm} = c_{ikm}$ ensures the agreement on the cluster prototypes obtained at all peers. Minimizing F is a constrained nonlinear optimization problem. The general solution is to consider the augmented Lagrangian of (5) as (6), which is an unconstrained minimization problem

$$\begin{aligned} \min\ & G(\boldsymbol{U}, \boldsymbol{C}, \boldsymbol{P}) \\ &= \sum_{j=1}^{J}\sum_{n=1}^{N_j}\sum_{k=1}^{K} u_{jnk}\int_{\mathbf{x}_{jn}\in R^M}\sum_{m=1}^{M}(x_{jnm} - c_{jkm})^2 \\ &\quad \times f_{jn}(\mathbf{x}_{jn})d\mathbf{x}_{jn} \\ &\quad + \sum_{j=1}^{J}\sum_{i\in NB_j}\sum_{k=1}^{K}\sum_{m=1}^{M} p_{jikm}(c_{jkm} - c_{ikm}) \end{aligned} \tag{6}$$

where $\boldsymbol{U} = [u_{\mathrm{jnk}}]$ and $\boldsymbol{C} = [c_{jkm}]$ are membership degree matrix and cluster prototype matrix, respectively, same as those defined in (5). $\boldsymbol{P} = [p_{jikm}]$ is the Lagrange multiplier matrix corresponding to the consensus constraint $c_{jkm} = c_{ikm}$, $i \in NB_j$.

Then the iterative optimization is performed to minimize G $(\boldsymbol{U}, \boldsymbol{C}, \boldsymbol{P})$ with respect to one variable of $\boldsymbol{U}$ or $\boldsymbol{C}$ with all other variables fixed, followed by a gradient descent step over the multipliers $\boldsymbol{P} = [p_{jikm}]$. Define t as iteration index, we have

$$\boldsymbol{U}(t+1) = \arg\min_{\boldsymbol{U}} G(\boldsymbol{C}(t), \boldsymbol{P}(t)) \tag{7}$$

$$\boldsymbol{C}(t+1) = \arg\min_{\boldsymbol{C}} G(\boldsymbol{U}(t+1), \boldsymbol{P}(t)) \tag{8}$$

$$\boldsymbol{P}(t+1) = \arg\min_{\boldsymbol{P}} G(\boldsymbol{U}(t+1), \boldsymbol{C}(t+1)). \tag{9}$$

The update of the membership degree matrix $\boldsymbol{U}$ and the cluster prototype matrix $\boldsymbol{C}$ are proved in the following theorems.

Theorem 1: Let $\boldsymbol{C}(t)$ and $\boldsymbol{P}(t)$ be fixed, $G(\boldsymbol{U}(t+1))$ is locally minimized if $\boldsymbol{U}(t+1)$ is given in

$$u_{jnk}(t+1) = \begin{cases} 1, & \text{if } \mathrm{ED}(\mathbf{o}_{jn}, \mathbf{c}_{jk}(t)) \le \mathrm{ED}(\mathbf{o}_{jn}, \mathbf{c}_{jr}(t)) \\ & \qquad \text{for } 1 \le r \le K \\ 0, & \text{other} \end{cases} \tag{10}$$

for $1 \le j \le J$, $1 \le n \le N_j$, $1 \le k \le K$.

The proof of Theorem 1 can be easily derived from [5] and [35]. Here, $u_{jnk} = 1$ means that the nth object is assigned to the kth cluster in the jth peer.

Theorem 2: Let $\boldsymbol{U}(t+1)$ be fixed, $G(\boldsymbol{C}(t+1), \boldsymbol{P}(t+1))$ is locally minimized if $\boldsymbol{C}(t+1)$ given via (11), as shown at the bottom of this page, which is followed by $\boldsymbol{P}(t+1)$, given via (12): for $1 \leq j \leq J, 1 \leq k \leq K, 1 \leq m \leq M$ and

$$p_{jikm}(t+1) = p_{jikm}(t) + \eta(c_{jkm}(t+1) - c_{ikm}(t+1)) \tag{12}$$

for $1 \leq j \leq J$, $1 \leq k \leq K$, $1 \leq m \leq M$, $i \in NB_j$.

Here, η is a positive parameter that affects the convergence speed of the clustering algorithm. In practice, a relatively small value of η is suggested to promote the clustering to converge.

Proof: If $\boldsymbol{U}(t+1)$ is fixed, by setting the gradient of $G(\boldsymbol{C}(t+1), \boldsymbol{P}(t))$ to zero with respect to $c_{jkm}(t+1)$, we obtain

$$\begin{aligned}
&\frac{\partial G(\boldsymbol{C}(t+1), \boldsymbol{P}(t))}{\partial c_{jkm}(t+1)} \\
&\quad = -2\sum_{n=1}^{N_j} u_{jnk}(t+1) \\
&\qquad \times \int_{\mathbf{x}_{jn}\in R^M} (x_{jnm} - c_{jkm}(t+1)) f_{jn}(\mathbf{x}_{jn}) d\mathbf{x}_{jn} \\
&\qquad + \sum_{i\in NB_j} p_{jikm}(t) - \sum_{i\in NB_j} p_{ijkm}(t) \\
&\quad = 0
\end{aligned} \tag{13}$$

for $1 \leq j \leq J$, $1 \leq k \leq K$, $1 \leq m \leq M$.

From (3) and (13), we have (14), as shown at the bottom of this page, for $1 \leq j \leq J$, $1 \leq k \leq K$, $1 \leq m \leq M$.

Note that the update of cluster prototypes is followed by a gradient descent step over the multipliers p_{jikm} as (12).

In the proposed algorithm, $p_{jikm}(0)$ and $p_{ijkm}(0)$ are initialized to zero. Then it holds $p_{jikm}(1) = p_{jikm}(0) + \eta(c_{jkm}(0) - c_{ikm}(0))$ and $p_{ijkm}(1) = p_{ijkm}(0) + \eta(c_{ikm}(0) - c_{jkm}(0))$. Therefore, we have $p_{jikm}(1) = -p_{ijkm}(1)$. Likewise, we have $p_{jikm}(t) = -p_{ijkm}(t)$ for $\forall t > 0$. Substituting it into (14), we obtain (11). This completes the proof.

The pseudocode of the DUK-means algorithm is summarized in Algorithm 1.

In this algorithm, there are two main phases, namely, the clustering on individual peer and the communication to exchange the cluster prototypes between neighbor peers. They intertwine and occur in a fixed sequence. Initially, each peer generates its initial cluster prototypes, and broadcasts its cluster prototypes to its neighbors. Then the clustering is performed in distributive manner at each peer using the local data and the cluster prototypes exchanged from its neighbor peers at this point of time. After one step of clustering, all peers are ready to start the communication phase. They broadcast the cluster prototypes to their neighbors and set up new conditions for the next new phase of the clustering. The overall optimization takes a finite number of iterations, which terminates once there is no further significant improvement in the cluster prototypes of all peers. In practice, when the variation of cluster prototypes in two consecutive iterations is smaller than a preset threshold, the peer will send the “convergence” message to its neighbors. The iteration of the algorithm will terminate if all peers achieve the convergence. The satisfying of the consensus constraint $c_{jkm} = c_{ikm}$ in the objective function through Theorem 2 ensure the agreement on the cluster prototypes of all peers is achieved.

Algorithm 1 Distributed Uncertain K-means Clustering Algorithm

```
1: For each peer $j$, randomly generate initial cluster prototypes $\boldsymbol{C}_j(0)$, initialize $p_{jikm}(0)$=0, broadcast its initial cluster prototypes to its neighboring peers, $t$ is set to 0;
2:   Each peer $j$ update the membership degree $\boldsymbol{U}_j(t+1)$ via (10) by using $\boldsymbol{C}_j(t)$;
3:   Each peer $j$ update the cluster prototypes $\boldsymbol{C}_j\ (t+1)$ via (11) by using $\boldsymbol{U}_j(t$+1) and $\boldsymbol{P}_j(t)$;
4:   Each peer $j$ broadcasts the updated cluster prototypes to its neighboring peers;
5:   Each peer $j$ update the multipliers $\boldsymbol{P}_j(t$+1) via (12) by using $\boldsymbol{C}_j(t$+1) and $\boldsymbol{P}_j(t)$;
6:   $t = t$ +1;
7: Repeat step2-step6 until the variation of cluster prototypes of all peers in two consecutive iterations is smaller than a preset threshold.
```

According to the definitions above, we have the computational complexity of the DUK-means algorithm is O ($tN_{\max}S_{\max}KM$), where t is the number of iterations required, K is the number of clusters, M is the number of data dimensions, $N_{\max}$ is the maximal number of uncertain data objects in all peers, and $S_{\max}$ is the maximal number of sample points in all uncertain data objects. Assume that the uncertain data objects are uniformly distributed on each peer, it means $N_{\max} \approx N/J$, where N is the total number of data objects and J is the number of peers. If the numberf of iterative calculations is the same, the DUK-means algorithm will be about J times faster than the UK-means algorithm. However, in practice, due to the incomplete connectivity of distributed peers, the exchange of cluster prototypes is limited between neighbor peers. Compared with the UK-means algorithm, the

$$c_{jkm}(t+1) = \frac{\sum_{n=1}^{N_j} u_{jnk}(t+1) \int_{\mathbf{x}_{jn}\in R^M} x_{jnm} f_{jn}(\mathbf{x}_{jn}) d\mathbf{x}_{jn} - \sum_{i\in NB_j} p_{jikm}(t)}{\sum_{n=1}^{N_j} u_{jnk}(t+1)} \tag{11}$$

$$c_{jkm}(t+1) = \frac{\sum_{n=1}^{N} u_{jnk}(t+1) \int_{\mathbf{x}_{jn}\in R^M} x_{jnm} f_{jn}(\mathbf{x}_{jn}) d\mathbf{x}_{jn} - \frac{1}{2}\sum_{i\in NB_j} (p_{jikm}(t) - p_{ijkm}(t))}{\sum_{n=1}^{N} u_{jnk}(t+1)} \tag{14}$$

DUK-means algorithm requires relatively more iterations (this cost is far less than J), which impede us from obtaining the maximum performance improvement. The subsequent experimental results in Section V also demonstrate the case.

C. Consistency Analysis of the Distributed Uncertain K-Means Clustering Algorithm

In the section above, we present the DUK-means algorithm for uncertain data clustering. Now, we ask: What are the results of the distributed clustering? In this section, we give the proof that the distributed clustering solution on accuracy of classification achieved by the DUK-means algorithm coincides with that by the centralized UK-means method.

Theorem 3: The cluster prototypes and the objective function obtained by the DUK-means algorithm are consistent with the ones of the centralized UK-means algorithm [13].

Proof: From (11), we have

$$\sum_{n=1}^{N_j} u_{jnk} c_{jkm} = \sum_{n=1}^{N_j} u_{jnk} \int_{\mathbf{x}_{jn} \in R^M} x_{jnm} f_{jn}(\mathbf{x}_{jn}) d\mathbf{x}_{jn} - \sum_{i \in NB_j} p_{jikm} \tag{15}$$

for $1 \leq j \leq J$, $1 \leq k \leq K$, $1 \leq m \leq M$.

Considering all peers, we have

$$\sum_{j=1}^{J} \sum_{n=1}^{N_j} u_{jnk} c_{jkm} = \sum_{j=1}^{J} \sum_{n=1}^{N_j} u_{jnk} \int_{\mathbf{x}_{jn} \in R^M} x_{jnm} f_{jn}(\mathbf{x}_{jn}) d\mathbf{x}_{jn} - \sum_{j=1}^{J} \sum_{i \in NB_j} p_{jikm} \tag{16}$$

for $1 \leq k \leq K, 1 \leq m \leq M$.

Let $\mathbf{c}'_k = [c'_{k1}, c'_{k2}, \ldots, c'_{kM}]$ for $1 \leq k \leq K$ be the consensus cluster prototypes reached by the DUK-means algorithm, from (16), we have

$$c'_{km} \sum_{j=1}^{J} \sum_{n=1}^{N_j} u_{jnk} = \sum_{j=1}^{J} \sum_{n=1}^{N_j} u_{jnk} \int_{\mathbf{x}_{jn} \in R^M} x_{jnm} f_{jn}(\mathbf{x}_{jn}) d\mathbf{x}_{jn} - \sum_{j=1}^{J} \sum_{i \in NB_j} p_{jikm} \tag{17}$$

for $1 \leq k \leq K, 1 \leq m \leq M$.

It follows (18), as shown at the bottom of this page, for $1 \leq k \leq K$, $1 \leq m \leq M$.

Let $p_{jikm}(0)$ and $p_{ijkm}(0)$ be initialized to zero, from (12), we have $p_{jikm}(t) = -p_{ijkm}(t)$ for $\forall t > 0$. Therefore, we have

$$\sum_{j=1}^{J} \sum_{i \in NB_j} p_{jikm} = 0 \tag{19}$$

for $1 \leq k \leq K, 1 \leq m \leq M$.

Substituting (19) into (18), we obtain

$$c'_{km} = \sum_{j=1}^{J} \sum_{n=1}^{N_j} u_{jnk} \int_{\mathbf{x}_{jn} \in R^M} x_{jnm} f_{jn}(\mathbf{x}_{jn}) d\mathbf{x}_{jn} \Bigg/ \sum_{j=1}^{J} \sum_{n=1}^{N_j} u_{jnk} \tag{20}$$

for $1 \leq k \leq K$, $1 \leq m \leq M$.

Note that if all objects are collected into one central unit from all peers, the cluster prototypes in (20) are consistent with the centralized ones obtained by the UK-means algorithm where $N = \sum_{j=1}^{J} N_j$ is the total number of objects of all peers.

Similar to the consistence proof of cluster prototypes, we obtain the objective function of the DUK-means algorithm with the consensus cluster prototypes as

$$\begin{aligned} F(\boldsymbol{U}, \boldsymbol{C}) &= \sum_{j=1}^{J} \sum_{n=1}^{N_j} \sum_{k=1}^{K} u_{jnk} \mathrm{ED}(\mathbf{o}_{jn}, \mathbf{c}_{jk}) \\ &= \sum_{j=1}^{J} \sum_{n=1}^{N_j} \sum_{k=1}^{K} u_{jnk} ED\big(\mathbf{o}_{jn}, \mathbf{c}'_k\big). \end{aligned} \tag{21}$$

As in the above, if all objects are collected into one central unit from all peers, the objective function in (21) is consistent with the centralized one obtained by the UK-means algorithm where $N = \sum_{j=1}^{J} N_j$ is the total number of objects of all peers. This completes the proof.

III. Optimization of the Distributed Uncertain K-Means Clustering Algorithm

In the actual iterative execution of the DUK-means algorithm, we need to calculate the ED between each object and the cluster prototype, which is not only a computationally expensive process but also one of most frequently executed operations. To improve efficiency, a reduction technique is used to optimize the DUK-means algorithm to the distributed K-means (DK-means) algorithm.

First, we define the expected centroid $\mathbf{z}_{jn}$ as (22) for each uncertain object $\mathbf{o}_{jn}$

$$\mathbf{z}_{jn} = \int_{\mathbf{x}_{jn} \in R^M} f_{jn}(\mathbf{x}_{jn}) \mathbf{x}_{jn} d\mathbf{x}_{jn}. \tag{22}$$

We have

$$\begin{aligned} \|\mathbf{x}_{jn} - \mathbf{c}_{jk}\|^2 &= \|\mathbf{x}_{jn} - \mathbf{z}_{jn}\|^2 + \|\mathbf{c}_{jk} - \mathbf{z}_{jn}\|^2 \\ &\quad - 2(\mathbf{c}_{jk} - \mathbf{z}_{jn})^T \cdot (\mathbf{x}_{jn} - \mathbf{z}_{jn}). \end{aligned} \tag{23}$$

$$c'_{km} = \frac{\sum_{j=1}^{J} \sum_{n=1}^{N_j} u_{jnk} \int_{\mathbf{x}_{jn} \in R^M} x_{jnm} f_{jn}(\mathbf{x}_{jn}) d\mathbf{x}_{jn} - \sum_{j=1}^{J} \sum_{i \in NB_j} p_{jikm}}{\sum_{j=1}^{J} \sum_{n=1}^{N_j} u_{jnk}} \tag{18}$$

Then we can calculate the difference between $\mathrm{ED}(\mathbf{o}_{jn}, \mathbf{c}_{jk})$ and $\mathrm{ED}(\mathbf{o}_{jn}, \mathbf{z}_{jn})$ as

$$\begin{aligned}&\mathrm{ED}(\mathbf{o}_{jn}, \mathbf{c}_{jk}) - \mathrm{ED}(\mathbf{o}_{jn}, \mathbf{z}_{jn})\\ &\quad= \int_{\mathbf{x}_{jn}\in R^M} (\|\mathbf{x}_{jn} - \mathbf{c}_{jk}\|^2 - \|\mathbf{x}_{jn} - \mathbf{z}_{jn}\|^2) f_{jn}(\mathbf{x}_{jn}) d\mathbf{x}_{jn}\\ &\quad= \int_{\mathbf{x}_{jn}\in R^M} (\|\mathbf{c}_{jk} - \mathbf{z}_{jn}\|^2 - 2(\mathbf{c}_{jk} - \mathbf{z}_{jn})^T \cdot (\mathbf{x}_{jn} - \mathbf{z}_{jn}))\\ &\qquad \times f_{jn}(\mathbf{x}_{jn}) d\mathbf{x}_{jn}\\ &\quad= \|\mathbf{c}_{jk} - \mathbf{z}_{jn}\|^2. \end{aligned} \tag{24}$$

Then the objective function of DUK-means algorithm is transformed into

$$\begin{aligned} F(\boldsymbol{U}, \boldsymbol{C}) = &\sum_{j=1}^{J} \sum_{n=1}^{N_j} \sum_{k=1}^{K} u_{jnk} \mathrm{ED}(\mathbf{o}_{jn}, \mathbf{z}_{jn})\\ &+ \sum_{j=1}^{J} \sum_{n=1}^{N_j} \sum_{k=1}^{K} u_{jnk} \|\mathbf{c}_{jk} - \mathbf{z}_{jn}\|^2. \end{aligned} \tag{25}$$

When the objects and their pdfs are fixed, the first term above is a constant. The objective function can be reduced to the minimization problem of the squared distance between object expected centroids and cluster prototypes

$$\begin{aligned} &\min\ F'(\boldsymbol{U}, \boldsymbol{C}) = \sum_{j=1}^{J} \sum_{n=1}^{N_j} \sum_{k=1}^{K} u_{jnk} \sum_{m=1}^{M} (c_{jkm} - z_{jnm})^2\\ &\text{s.t. } c_{jkm} = c_{ikm},\ i \in NB_j. \end{aligned} \tag{26}$$

Therefore, we have a new method to solve the problem of distributed clustering for uncertain data. First, we can preprocess the data via (22) to calculate the expected centroids of objects as the certain input data. Then the optimization problem (26) with constraints can be solved with its augmented Lagrangian as

$$\begin{aligned} \min G'(\mathbf{U}, \mathbf{C}, \mathbf{P}') = &\sum_{j=1}^{J} \sum_{n=1}^{N_j} \sum_{k=1}^{K} u_{jnk} \sum_{m=1}^{M} (c_{jkm} - z_{jnm})^2\\ &+ \sum_{j=1}^{J} \sum_{i\in NB_j} \sum_{k=1}^{K} \sum_{m=1}^{M} p'_{jikm} (c_{jkm} - c_{ikm}). \end{aligned} \tag{27}$$

Then the matrices $\boldsymbol{U}$ and $\boldsymbol{C}$ will be updated via Theorems 4 and 5.

Theorem 4: Let $\boldsymbol{C}(t)$ and $\boldsymbol{P}'(t)$ be fixed, $G'(\boldsymbol{U}(t+1))$ is locally minimized if $\boldsymbol{U}(t+1)$ is given by

$$\begin{aligned} &u_{jnk}(t+1)\\ &= \begin{cases} 1, & \text{if } \sum_{m=1}^{M} (c_{jkm}(t) - z_{jnm})^2 \le \sum_{m=1}^{M} (c_{jrm}(t) - z_{jnm})^2\\ & \qquad \text{for } 1 \le r \le K\\ 0, & \text{other} \end{cases} \end{aligned} \tag{28}$$

for $1 \le j \le J$, $1 \le n \le N_j$, $1 \le k \le K$.

Theorem 5: Let $\boldsymbol{U}(t+1)$ be fixed, $G'(\boldsymbol{C}(t+1), \boldsymbol{P}'(t))$ is locally minimized if $\boldsymbol{C}(t+1)$ given via the following equation is then followed by $\boldsymbol{P}'(t+1)$, given via (30):

$$c_{jkm}(t+1) = \frac{\sum_{n=1}^{N_j} u_{jnk}(t+1) z_{jnm} - \sum_{i\in NB_j} p'_{jikm}(t)}{\sum_{n=1}^{N_j} u_{jnk}(t+1)} \tag{29}$$

for $1 \le j \le J$, $1 \le k \le K$, $1 \le m \le M$, $i \in NB_j$ and

$$p'_{jikm}(t+1) = p'_{jikm}(t) + \eta(c_{jkm}(t+1) - c_{ikm}(t+1)) \tag{30}$$

for $1 \le j \le J$, $1 \le k \le K$, $1 \le m \le M$, $i \in NB_j$ where η is a positive scalar.

The proofs of these two theorems are very similar to those of Theorems 1 and 2.

According to Theorem 3, we also can give the proof that the cluster prototypes achieved by the DK-means algorithm coincides with the ones in the following equation obtained by the centralized CK-means algorithm [19], i.e., the classical centralized K-means algorithm [5]:

$$c_{km} = \sum_{n=1}^{N} u_{nk} z_{nm} \Big/ \sum_{n=1}^{N} u_{nk} \tag{31}$$

for $1 \le k \le K, 1 \le m \le M$, where $N = \sum_{j=1}^{J} N_j$ is the total number of all objects.

At this point, we can conclude that the DK-means algorithm provides an efficient solution for uncertain data clustering on accuracy of classification, which coincides with that of the centralized CK-means method. In fact, DK-means algorithm is locally optimal with respect to the DUK-means objective function by replacing each uncertain data object with its expected centroid. We also wish to point out that the consensus constraint $c_{jkm} = c_{ikm}$ ensures the agreement on the cluster prototypes obtained at all peers. Hence, the DK-means algorithm can achieve the global clustering solution similar to the one obtained by the P2P K-means algorithm [25].

IV. Attribute Weighted Distributed Uncertain Clustering

A. Attribute Weighted Distributed Uncertain K-Means Clustering Algorithm

For many real applications, especially the high-dimensional sparse data clustering, the cluster structure in the data set is often limited to a subset of features rather than the entire feature set. A better solution is to introduce the proper attribute weight into the clustering process according to the importance of different dimensions for cluster identification, which is referred to as subspace clustering [30]. With reference to the early research of entropy weighting K-means (EWKM) clustering method [31] and [32], we propose the attribute weighted DUK-means (WDUK-means) clustering algorithm, in which the attribute-weight-entropy regularization technique is considered to achieve the ideal distribution of attribute weights. The new ED is defined as the following equation and the new objective function is developed as (33):

$$\begin{aligned} &\mathrm{ED}(\mathbf{o}_{jn}, \mathbf{c}_{jk})\\ &\quad= \int_{\mathbf{x}_{jn}\in R^M} \sum_{m=1}^{M} w_{jkm} (x_{jnm} - c_{jkm})^2 f_{jn}(\mathbf{x}_{jn}) d\mathbf{x}_{jn} \end{aligned} \tag{32}$$

and

$$\begin{aligned}\min\ & F(\boldsymbol{U},\boldsymbol{C},\boldsymbol{W})\\ &= \sum_{j=1}^{J}\sum_{n=1}^{N_j}\sum_{k=1}^{K} u_{jnk}\int_{x_{jn}\in R^M}\sum_{m=1}^{M} w_{jkm}(x_{jnm}-c_{jkm})^2\\ &\quad\times f_{jn}(\mathbf{x}_{jn})d\mathbf{x}_{jn}\\ &\quad+\gamma\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{m=1}^{M} w_{jkm}\log w_{jkm}\\ &\text{s.t. } c_{jkm}=c_{ikm},\quad i\in NB_j\\ &\qquad w_{jkm}=w_{ikm},\quad i\in NB_j,\quad \sum_{m=1}^{M} w_{jkm}=1\\ &\qquad 0\le w_{jkm}\le 1\end{aligned}\tag{33}$$

where $\boldsymbol{U}=[u_{\text{jnk}}]$ and $\boldsymbol{C}=[c_{jkm}]$ are membership degree matrix and cluster prototype matrix respectively, same as defined in (5). $\boldsymbol{W}=[w_{jkm}]$ is the attribute weight matrix and w_{jkm} denotes the mth dimension of the kth cluster weight vector in the jth peer. γ is a positive scalar.

In this new objective function, the second term is the negative entropy of attribute weights that regularize the optimal distribution of all attribute weights according to the available data. $\gamma\,(\gamma>0)$ is a positive regularizing and adjustable parameter. With a proper choice of γ, we can balance the two terms to find the optimal solution.

The augmented Lagrangian technique is also applied to solve this constrained optimization problem as

$$\begin{aligned}\min\ & \boldsymbol{G}(\boldsymbol{U},\boldsymbol{C},\boldsymbol{P},\boldsymbol{W},\boldsymbol{Q})\\ &= \sum_{j=1}^{J}\sum_{n=1}^{N_j}\sum_{k=1}^{K} u_{jnk}\int_{\mathbf{x}_{jn}\in R^M}\sum_{m=1}^{M} w_{jkm}(x_{jnm}-c_{jkm})^2\\ &\quad\times f_{jn}(\mathbf{x}_{jn})d\mathbf{x}_{jn}\\ &\quad+\gamma\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{m=1}^{M} w_{jkm}\log w_{jkm}\\ &\quad+\sum_{j=1}^{J}\sum_{i\in NB_j}\sum_{k=1}^{K}\sum_{m=1}^{M} p_{jikm}(c_{jkm}-c_{ikm})\\ &\quad+\sum_{j=1}^{J}\sum_{i\in NB_j}\sum_{k=1}^{K}\sum_{m=1}^{M} q_{jikm}(w_{jkm}-w_{ikm})\\ &\quad-\sum_{j=1}^{J}\sum_{k=1}^{K}\gamma_{jk}\left(\sum_{m=1}^{M} w_{jkm}-1\right).\end{aligned}\tag{34}$$

The matrices $\boldsymbol{U}$, $\boldsymbol{C}$, and $\boldsymbol{W}$ are updated corresponding to (35)–(39), respectively

$$u_{jnk}(t+1)=\begin{cases}1, & \text{if } \mathrm{ED}(\mathbf{o}_{jn},\mathbf{c}_{jk}(t))\le \mathrm{ED}(\mathbf{o}_{jn},\mathbf{c}_{jr}(t))\\ & \qquad\text{for } 1\le r\le K\\ 0, & \text{other}\end{cases}\tag{35}$$

for $1\le j\le J$, $1\le n\le N_j$, $1\le k\le K$, and (36), shown at the bottom of this page, for $1\le j\le J$, $1\le k\le K$, $1\le m\le M$

$$p_{jikm}(t+1)=p_{jikm}(t)+\eta_1(c_{jkm}(t+1)-c_{ikm}(t+1))\tag{37}$$

for $1\le j\le J$, $1\le k\le K$, $1\le m\le M$, $i\in NB_j$ where η_1 is a positive scalar defined as η. Equation (38), as shown at the bottom of this page, holds for $1\le j\le J$, $1\le k\le K$, $1\le m\le M$, and

$$q_{jikm}(t+1)=q_{jikm}(t)+\eta_2(w_{jkm}(t+1)-w_{ikm}(t+1))\tag{39}$$

for $1\le j\le J$, $1\le k\le K$, $1\le m\le M$, $i\in NB_j$ where η_2 is a positive scalar defined as η.

Here, $\boldsymbol{P}=[p_{jikm}]$ and $\boldsymbol{Q}=[q_{jikm}]$ are two matrices containing the Lagrange multipliers corresponding to the consensus constraints $c_{jkm}=c_{ikm}$ and $w_{jkm}=w_{ikm}$, $i\in NB_j$. η_1 and η_2 are positive scalars. The iterative optimization analysis can be found in Appendix A. The pseudocode of the WDUK-means algorithm is summarized in Algorithm 2.

B. Optimization of the Attribute Weighted Distributed Uncertain K-Means Clustering Algorithm

With reference to Section III, we can also reduce the WDUK-means algorithm to its deterministic form, i.e., the attribute weighted distributed K-means (WDK-means) algorithm, as follows:

$$\begin{aligned}\min\ & F'(\boldsymbol{U},\boldsymbol{C},\boldsymbol{W})\\ &= \sum_{j=1}^{J}\sum_{n=1}^{N_j}\sum_{k=1}^{K} u_{jnk}\sum_{m=1}^{M} w_{jkm}(c_{jkm}-z_{jnm})^2\\ &\quad+\gamma\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{m=1}^{M} w_{jkm}\log w_{jkm}.\end{aligned}\tag{40}$$

Here, $\mathbf{z}_{jn}=\int_{\mathbf{x}_{jn}\in R^M} f_{jn}(\mathbf{x}_{jn})\mathbf{x}_{jn}d\mathbf{x}_{jn}$ is the centroid of the uncertain object $\boldsymbol{x}_{jn}$. The detailed derivation process of the reduction is shown in Appendix B.

$$c_{jkm}(t+1)=\frac{\sum_{n=1}^{N_j} u_{jnk}(t+1)w_{jkm}(t)\int_{\mathbf{x}_{jn}\in R^M} x_{jnm}f_{jn}(\mathbf{x}_{jn})d\mathbf{x}_{jn}-\sum_{i\in NB_j} p_{jikm}(t)}{\sum_{n=1}^{N_j} u_{jnk}(t+1)w_{jkm}(t)}\tag{36}$$

$$w_{jkm}(t+1)=\frac{\exp\left(\frac{-\sum_{n=1}^{N_j} u_{jnk}(t+1)\int_{\mathbf{x}_{jn}\in R^M}(x_{jnm}-c_{jkm}(t+1))^2 f_{jn}(\mathbf{x}_{jn})d\mathbf{x}_{jn}-2\sum_{i\in NB_j} q_{jikm}(t)}{\gamma}\right)}{\sum_{s=1}^{M}\exp\left(\frac{-\sum_{n=1}^{N_j} u_{jnk}(t+1)\int_{\mathbf{x}_{jn}\in R^M}(x_{jns}-c_{jks}(t+1))^2 f_{jn}(\mathbf{x}_{jn})d\mathbf{x}_{jn}-2\sum_{i\in NB_j} q_{jiks}(t)}{\gamma}\right)}\tag{38}$$

Algorithm 2 Attribute Weighted Distributed Uncertain K-Means Clustering Algorithm

```
1: For each peer $j$, randomly generate initial cluster prototypes, set the initial value of attribute weight to $1/M$, initialize $p_{jikm}(0) = 0$ and $q_{jikm}(0) = 0$, broadcast its initial cluster prototypes and attribute weights to its neighboring peers, $t = 0$;
2:    Each peer $j$ update the membership degree $\boldsymbol{U}_j(t+1)$ via (35) by using $\boldsymbol{C}_j(t)$;
3:    Each peer $j$ update the cluster prototypes $\boldsymbol{C}_j(t+1)$ via (36) by using $\boldsymbol{U}_j(t+1)$, $\boldsymbol{W}_j(t)$ and$\boldsymbol{P}_j(t)$;
4:    Each peer $j$ broadcasts the updated cluster prototypes to its neighboring peers;
5:    Each peer $j$ update the multipliers $\boldsymbol{P}_j(t+1)$ via (37) by using $\boldsymbol{C}_j(t+1)$ and $\boldsymbol{P}_j(t)$;
6:    Each peer $j$ update the attribute weights $\boldsymbol{W}_j(t+1)$ via (38) by using $\boldsymbol{U}_j(t+1)$, $\boldsymbol{C}_j(t+1)$ and $\boldsymbol{Q}_j(t)$;
7:    Each peer $j$ broadcasts the updated attribute weights to its neighboring peers;
8:    Each peer $j$ update the multipliers $\boldsymbol{Q}_j(t+1)$ via (39) by using $\boldsymbol{W}_j(t+1)$ and $\boldsymbol{Q}_j(t)$;
9:    $t = t$ +1;
10: Repeat step2-step9 until the variation of cluster prototypes of all peers in two consecutive iterations is smaller than a preset threshold.
```

Then the matrices ***U***, ***C,*** and ***W*** will be updated via (41)–(45) to solve the constrained optimization problem above. The pseudocode of the WDK-means algorithm can be referred to in Algorithm 2

$$u_{jnk}(t+1) = \begin{cases} 1, & \text{if } \sum_{m=1}^{M} (c_{jkm}(t) - z_{jnm})^2 \le \sum_{m=1}^{M} (c_{jrm}(t) - z_{jnm})^2 \\ & \quad \text{for } 1 \le r \le K \\ 0, & \text{other} \end{cases} \tag{41}$$

for $1 \le j \le J$, $1 \le n \le N_j$, $1 \le k \le K$

$$c_{jkm}(t+1) = \frac{\sum_{n=1}^{N_j} u_{jnk}(t+1) w_{jnk}(t) z_{jnm} - \sum_{i \in NB_j} p'_{jikm}(t)}{\sum_{n=1}^{N_j} u_{jnk}(t+1) w_{jnk}(t)} \tag{42}$$

for $1 \le j \le J$, $1 \le k \le K$, $1 \le m \le M$

$$p'_{jikm}(t+1) = p'_{jikm}(t) + \eta_1 (c_{jkm}(t+1) - c_{ikm}(t+1)) \tag{43}$$

for $1 \le j \le J$, $1 \le k \le K$, $1 \le m \le M$, $i \in NB_j$, where η_1 is a positive scalar. Equation (44), as shown at the bottom of this page, holds for $1 \le j \le J$, $1 \le k \le K$, $1 \le m \le M$, and

$$q'_{jikm}(t+1) = q'_{jikm}(t) + \eta_2 (w_{jkm}(t+1) - w_{ikm}(t+1)) \tag{45}$$

for $1 \le j \le J$, $1 \le k \le K$, $1 \le m \le M$, $i \in NB_j$, where η_2 is a positive scalar.

With reference to Appendix A, we can give the iterative optimization analysis for the equations.

In addition, we also give the consistency analysis of the WDK-means algorithm shown in Appendix C. It proves that the distributed clustering solution obtained by the WDK-means algorithm coincides with that by the centralized EWKM algorithm [31], including cluster prototypes and attribute weights.

V. Experiments

To evaluate the performance of proposed algorithms (the DUK-means algorithm, the DK-means algorithm, the WDUK-means algorithm, and the WDK-means algorithm), a series of experiments are conducted with synthetic and real-world data. Two centralized clustering algorithms for uncertain data, including the UK-means algorithm [13] and the CK-means algorithm [19], are chosen for the comparative analysis. All experiment data are normalized to the interval [0, 1]. For all the clustering algorithms, the stopping threshold is uniformly set to 10^{-6}.

Four clustering performance metrics are considered in our experiments, including the classification rate (CR) [36], the normalized mutual information (NMI) [37], the adjusted rand index (ARI) [38], and the CPU time (CT) [19]. Because the proposed algorithms are all K-means type clustering methods, the clustering results are very sensitive to the initial cluster prototypes. To achieve the convincing clustering results, we let each algorithm be executed on each data set 100 times (the cluster prototypes are randomly initialized at each time) and calculate the average of CR (ACR), the average of NMI (ANMI), the average of ARI (AARI), and the average of CT (ACT). Note that compared with the clustering operations, the communication cost is very small, in microseconds per message. We do not include it in the CT.

A. Synthetic Data Sets

This experiment works on a distributed P2P network in which peers are distributed uniformly over a 500 m × 500 m region. The communication range of each peer R is set to 100 m, i.e., each peer only exchanges information with its immediate topological neighbors in the communication range. Fig. 1 illustrates an example of such a kind of distributed P2P network with 50 peers.

The synthetic data are generated in a continuous 2-D space ($M = 2$). Assume each peer contains 150 uncertain objects

$$w_{jkm}(t+1) = \frac{\exp\left(-\gamma^{-1} \sum_{n=1}^{N_j} u_{jnk}(t+1)(c_{jkm}(t+1) - z_{jnm})^2 - 2\gamma^{-1} \sum_{i \in NB_j} q'_{jikm}(t)\right)}{\sum_{l=1}^{M} \exp\left(-\gamma^{-1} \sum_{n=1}^{N_j} u_{jnk}(t+1)(c_{jkl}(t+1) - z_{jnl})^2 - 2\gamma^{-1} \sum_{i \in NB_j} q'_{jikl}(t)\right)} \tag{44}$$

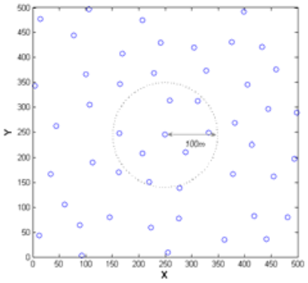


Fig. 1. Distribution of 50 peers in a P2P network.

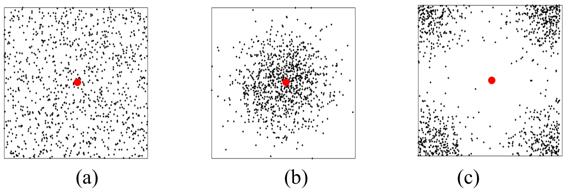


Fig. 2. Three types of distributions. (a) Uniform. (b) Gaussian. (c) Inverse Gaussian.

belonging to three clusters ($K = 3$). Each cluster has 50 uncertain objects. For each uncertain object, a centroid point is first generated from a mixture of three bivariate Gaussian densities given by

$$
\begin{aligned}
0.33 * N & \left[\begin{pmatrix} 5.0 \\ 5.0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0.0 \\ 0.0 & \sigma^2 \end{pmatrix}\right] \\
& + 0.33 * N \left[\begin{pmatrix} 5.0 \\ 10.0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0.0 \\ 0.0 & \sigma^2 \end{pmatrix}\right] \\
& + 0.34 * N \left[\begin{pmatrix} 10.0 \\ 10.0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0.0 \\ 0.0 & \sigma^2 \end{pmatrix}\right]
\end{aligned}
$$

where σ is the bandwidth. Then centered at this centroid point, a total of $|S| = 1000$ sample points are observed at random following one of the three types of distributions, which include the uniform distribution, the Gaussian distribution, and the inverse Gaussian distribution, as shown in Fig. 2. Here, the red solid circle indicates the centroid point. In the synthetic data experiments, we assume each cluster corresponds to one type of distribution.

Then three data sets are created as follows: E1 contains $30 \times 150 = 4500$ uncertain objects with 30 peers ($N = 4500$, $J = 30$), E2 contains $50 \times 150 = 7500$ uncertain objects with 50 peers ($N = 7500$, $J = 50$), and E3 contains $70 \times 150 = 10\,500$ uncertain objects with 70 peers ($N = 10\,500$, $J = 70$). For different data set E1, E2, and E3, σ^2 is set to different values with 1.5, 2.0, and 2.5. Fig. 3 illustrates the distribution of 150 centroid points of one peer in E2.

Fig. 4 lists the clustering results of different uncertain clustering algorithms on three synthetic data sets E1, E2, and E3. They draw similar conclusions. The difference is because of the different values of σ^2 and the different sizes of data sets. On the one hand, the larger σ^2 indicates more overlap of the data in different clusters and lower CR and vice versa. On the other hand, the bigger size of data set needs more execution time for clustering and vice versa. From Fig. 4(a), an important finding is that the proposed distributed algorithms (DUK-means, DK-means, WDUK-means and WDK-means) are about 10–40 times faster than their corresponding centralized approaches (UK-means and CK-means). Particularly, DK-means and WDK-means with the reduction technique achieve further enormous time savings (about 99%) compared with DUK-means and WDUK-means as shown in Fig. 4(b). More importantly, in Fig. 4(c)–(e), the results of the proposed distributed algorithms (DUK-means, DK-means, WDUK-means, and WDK-means) in ACR and ANMI are very close to the ones obtained by the corresponding centralized approaches (UK-means and CK-means). This verifies the consistency of the clustering results between our distributed algorithms and the traditional centralized methods.

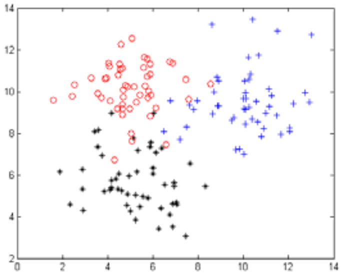


Fig. 3. Distribution of 150 object centroids of one peer in E2.

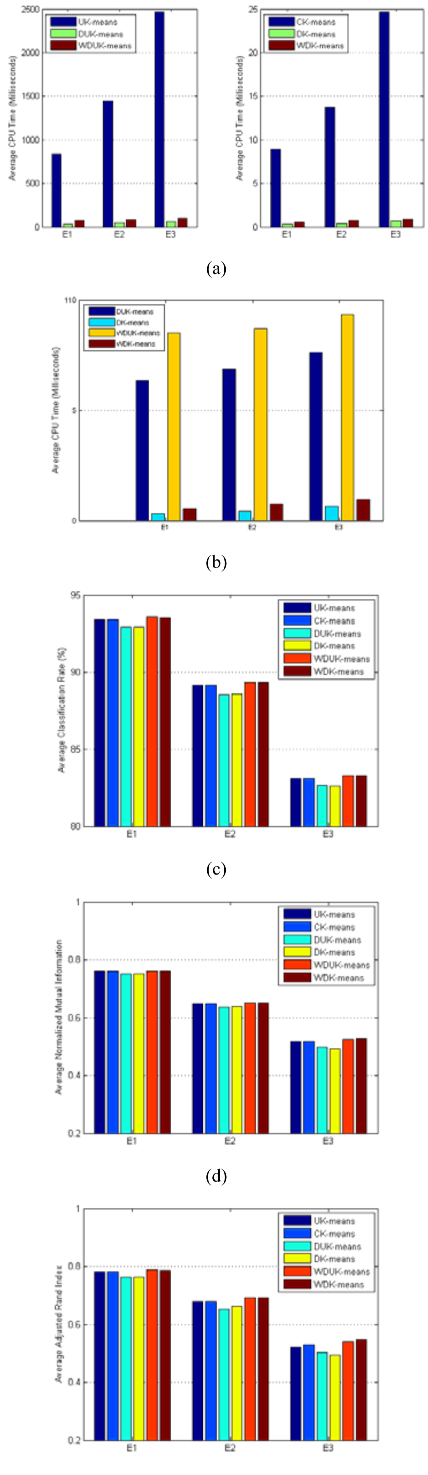
Another finding is that the proposed distributed algorithms with attribute-weight-entropy regularization technique (WDUK-means and WDK-means) do not provide significant improvements in ACR and ANMI compared with other approaches. That is because the data dimension is only two in this experiment. The feature weighting has small impact on the clustering results. We will further discuss this issue in the following experiments.

As a short summary of the observations above, when the centralized clustering approaches are discouraged by the technical constraints like the volume size of data and the privacy and security problems like no full data transmission is permitted, the distributed clustering approaches proposed in this paper are a good selection. In these methods, great savings in execution time will be achieved without affecting CR in clustering. Moreover, the reduction technique for uncertain clustering can further significantly shorten the execution time of the algorithm.

B. UCI Machine Learning Data Sets

In this section, we consider our algorithms on some uncertain data generated by extending the certain data sets in UCI machine learning repository [39]. For each data point in a certain data set, we design a Gaussian distribution centered at this data point and randomly samples 1000 data according to the Gaussian distribution. We take the 1000 samples as the observed values for the uncertain data. As mentioned earlier, we only use (1) to approximate the distribution of uncertain data and emulate the situation of real applications in which we only have some observations of uncertain data.

Seven real-world certain data sets from UCI machine learning repository are chosen to be transformed for our experiments, which include Iris ($N = 150$, $K = 3$, $M = 4$),



(a)

(b)

(c)

(d)

(e)

Fig. 4. Three statistics of different uncertain clustering algorithms on the synthetic data sets in terms of (a) and (b) ACT(Milliseconds), (c) ACR(%), (d) ANMI, and (e) AARI.

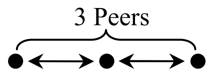


Fig. 5. Simple linear network architecture.

TABLE I

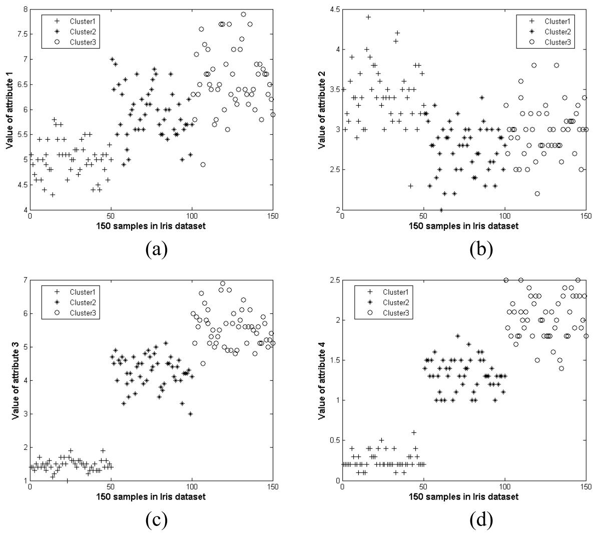
STATISTICS OF DIFFERENT UNCERTAIN CLUSTERING ALGORITHMS ON THE UCI MACHINE LEARNING DATA SETS IN TERMS OF ACT (ms), ACR(%), ANMI, AND AARI

Data sets	UK-means	CK-means	DUK-means	DK-means	WDUK-means	WDK-means
Iris	39.22 87.06 0.7340 0.6639	0.72 87.06 0.7340 0.6639	19.51 85.96 0.6884 0.6318	0.27 86.11 0.6891 0.6402	26.18 95.02 0.8592 0.7920	0.68 95.65 0.8645 0.8341
Haberman	51.48 56.22 0.0108 0.0127	0.84 56.38 0.0112 0.0135	17.50 53.72 0.0013 0.0009	0.25 55.86 0.0047 0.0044	33.73 61.50 0.0204 0.1207	0.32 61.87 0.0237 0.1433
Heart	1389.24 64.82 0.0852 0.1042	5.80 64.89 0.0887 0.1051	373.67 62.91 0.0624 0.0411	1.86 63.23 0.0799 0.0543	497.55 66.57 0.1008 0.1640	2.67 67.81 0.1126 0.2173
Glass	384.20 35.64 0.4243 0.1726	0.30 35.41 0.4207 0.1664	75.49 33.76 0.3926 0.1300	0.09 34.60 0.3959 0.1497	136.02 36.32 0.4529 0.2049	0.14 37.45 0.4727 0.2263
Ionosphere	805.32 70.43 0.1246 0.1618	4.12 70.54 0.1257 0.1690	381.68 70.20 0.1209 0.1251	1.53 69.68 0.1189 0.0946	560.10 71.65 0.1492 0.2402	2.29 71.40 0.1464 0.2177
Wine	191.08 94.67 0.8160 0.8459	1.26 94.44 0.8125 0.8414	62.85 93.09 0.7822 0.7837	0.35 93.40 0.7860 0.7918	84.79 94.51 0.8143 0.8430	0.86 95.38 0.8498 0.8864
Wdbc	1221.02 92.21 0.6114 0.7109	5.26 92.17 0.6086 0.7082	273.88 91.65 0.6027 0.6187	1.28 91.67 0.6056 0.6510	608.42 92.51 0.6271 0.7206	3.20 92.46 0.6233 0.7163

Glass ($N = 214$, $K = 6$, $M = 9$), Ionosphere ($N = 351$, $K = 2$, $M = 33$), Haberman ($N = 306$, $K = 2$, $M = 3$), Heart ($N = 267$, $K = 2$, $M = 44$), Wine ($N = 178$, $K = 3$, $M = 13$), and Wdbc ($N = 569$, $K = 2$, $M = 30$). Because the number of data objects in these data sets is relatively small, only three peers ($J = 3$) are considered to be deployed in the distributed P2P networks, and the simple linear network structure shown in Fig. 5 is adopted.

The clustering results are listed in Table I. Without any surprise, most experiments show the similar conclusion with the experiment above. This demonstrates the efficiency of the proposed distributed uncertain data clustering algorithms. It is worth pointing out that the proposed WDUK-means algorithm and WDK-means algorithm have shown excellent performance in ACR, ANMI, and AARI with the cost of a little more execution time on the Iris data set (about 10% promotion in ACR), which attributes to the introduction of attribute-weight-entropy regularization technique.

In order to have an intuitive understanding of the inherent properties of this technique, we further investigate the distribution of four attributes/dimensions of original Iris data, as shown in Fig. 6. Note that these data points are actually regarded as the centroid of the corresponding uncertain objects. We can clearly see that attributes 3 and 4 are more compact in each cluster. They should be more important and contribute much more than the other two attributes in clustering, so that higher weights should be assigned to these



(a) (b) (c) (d)

Fig. 6. Distribution of four attributes of the Iris data set. (a) Attribute 1. (b) Attribute 2. (c) Attribute 3. (d) Attribute 4.

TABLE II

ATTRIBUTE WEIGHT ASSIGNMENT OBTAINED BY THE WDUK-MEANS ALGORITHM AND THE WDK-MEANS ALGORITHM ON THE IRIS DATA SET

WDUK-means				
	Attribute 1	Attribute 2	Attribute 3	Attribute 4
Cluster1	0.081	0.062	**0.652**	**0.205**
Cluster2	0.069	0.117	**0.398**	**0.416**
Cluster3	0.166	0.129	**0.412**	**0.293**
WDK-means				
	Attribute 1	Attribute 2	Attribute 3	Attribute 4
Cluster1	0.013	0.024	**0.721**	**0.242**
Cluster2	0.055	0.086	**0.428**	**0.431**
Cluster3	0.079	0.107	**0.496**	**0.318**

two attributes. This can be verified by Table II, in which attributes 3 and 4 have higher weights than attributes 1 and 2 for each attribute weighted distributed clustering algorithm. All these demonstrate the efficiency of the attribute-weight-entropy regularization technique for data clustering with different distributions of each dimension.

VI. CONCLUSION

This paper focuses on uncertain data clustering problem and proposes a distributed clustering algorithm in P2P networks. The centralized clustering solution is obtained in a distributive mode at each peer by collaborating with the neighboring peers only. Based on the reduction technique, the distributed uncertain data clustering algorithm actually turns out to be equivalent to the deterministic clustering, which greatly shortens the execution time of the algorithm. The attribute-weight-entropy regularization technique is applied in the distributed clustering method to achieve ideal distribution of attribute weights, which ensures the good clustering results. Experiments on several synthetic and real-world data sets have demonstrated the good performance of the proposed algorithms. The results of this paper provide some valuable directions for future work. The proposed algorithm is of great generality and could be further applied in uncertain data clustering research in distributed environments.

Currently, most of the study on clustering set the number of clusters as a user-defined parameter, which is difficult to specify. In the research of centralized clustering, some scholars have tried to determine the most appropriate number of clusters through the cluster validation techniques [40], [41]. However, in the distributed clustering, this is still a hard problem because it is sometime difficult to collect all the data or the data membership information due to the privacy and security concerns or the technical constraints brought by distributive environments. In this paper, the number of clusters is predetermined. Supervising the distributed clustering algorithm to optimize the number of clusters and the clustering performance together will be considered as a good direction of future work. In addition, we assume that different dimensions of the data are independent and their pdfs can be approximated by multiple kernel functions separately. To consider a nontrivial covariance structure or dependent attributes/dimensions is a tough problem in the distributed environment and will be another good future research direction.

APPENDIX A
ITERATIVE OPTIMIZATION ANALYSIS OF THE WDUK-MEANS ALGORITHM

Theorem A.1: Let $\boldsymbol{C}(t)$ and $\boldsymbol{W}(t)$ be fixed, $G(\boldsymbol{U}(t+1))$ is locally minimized if $\boldsymbol{U}(t+1)$ is given via (35).

The proof of Theorem A.1 can be easily derived from [5] and [35].

Theorem A.2: Let $\boldsymbol{U}(t+1)$ and $\boldsymbol{W}(t)$ be fixed, $G(\boldsymbol{C}(t+1), \boldsymbol{P}(t+1))$ is locally minimized if $\boldsymbol{C}(t+1)$ given via (36) is followed by $\boldsymbol{P}(t+1)$ given via (37).

Proof: If $\boldsymbol{U}(t+1)$ and $\boldsymbol{W}(t)$ are fixed, by setting the gradient of $G(\boldsymbol{C}(t+1), \boldsymbol{P}(t))$ to zero with respect to $c_{jkm}(t+1)$, we obtain

$$\begin{aligned}\frac{\partial G(\boldsymbol{C}(t+1), \boldsymbol{P}(t))}{\partial c_{jkm}(t+1)} &= -2\sum_{n=1}^{N_j} u_{jnk}(t+1)\int_{\mathbf{x}_{jn}\in R^M} w_{jkm}(t)(x_{jnm}-c_{jkm}(t+1))\\ &\quad \times f_{jn}(\mathbf{x}_{jn})d\mathbf{x}_{jn}\\ &\quad + \sum_{i\in NB_j} p_{jikm}(t) - \sum_{i\in NB_j} p_{ijkm}(t)\\ &= 0 \end{aligned} \tag{A.1}$$

for $1 \le j \le J$, $1 \le k \le K$, $1 \le m \le M$.

From (A.1), we have (A.2), as shown at the top of the next page, for $1 \le j \le J$, $1 \le k \le K$, $1 \le m \le M$.

Note that the update of cluster prototypes is followed by a gradient descent step over the multipliers $p_{jikm}(t)$ as $p_{jikm}(t) = p_{jikm}(t-1)+\eta_1(c_{jkm}(t)-c_{ikm}(t))$ for $1 \le j \le J$, $1 \le k \le K$, $1 \le m \le M$, $i \in NB_j$, where η_1 is a positive scalar. If $p_{jikm}(0)$ and $p_{ijkm}(0)$ are initialized to zero, we have $p_{jikm}(t) = -p_{ijkm}(t)$ for $\forall t > 0$. Substituting it into (A.2), we obtain (36). This completes the proof.

$$c_{jkm}(t+1) = \frac{\sum_{n=1}^{N} u_{jnk}(t+1) w_{jkm}(t) \int_{\mathbf{x}_{jn} \in R^M} x_{jnm} f_{jn}(\mathbf{x}_{jn}) d\mathbf{x}_{jn} - \frac{1}{2} \sum_{i \in NB_j} (p_{jikm}(t) - p_{ijkm}(t))}{\sum_{n=1}^{N} u_{jnk}(t+1) w_{jkm}(t)} \tag{A.2}$$

Theorem A.3: Let $\boldsymbol{U}(t+1)$ and $\boldsymbol{C}(t+1)$ be fixed, $G(\boldsymbol{W}(t+1), \boldsymbol{Q}(t+1))$ is locally minimized if $\boldsymbol{W}(t+1)$ given via (38) is followed by $\boldsymbol{Q}(t+1)$ given via (39).

Proof: If $\boldsymbol{U}(t+1)$ and $\boldsymbol{C}(t+1)$ are fixed, by setting the gradient of $G(\boldsymbol{W}(t+1), \boldsymbol{Q}(t), \boldsymbol{\Lambda})$ to zero with respect to λ_{jk} and $w_{jkm}(t+1)$, we obtain

$$\frac{\partial G(\boldsymbol{W}(t+1), \boldsymbol{Q}(t), \boldsymbol{\Lambda})}{\partial \lambda_{jk}} = -\left(\sum_{m=1}^{M} w_{jkm}(t+1) - 1\right) = 0 \tag{A.3}$$

$$\begin{aligned}\frac{\partial G(\boldsymbol{W}(t+1), \boldsymbol{Q}(t), \boldsymbol{\Lambda})}{\partial\, w_{jkm}(t+1)} &= \sum_{n=1}^{N_j} u_{jnk}(t+1) \int_{\mathbf{x}_{jn} \in R^M} (x_{jnm} - c_{jkm}(t+1))^2 f_{jn}(\mathbf{x}_{jn}) d\mathbf{x}_{jn} \\ &\quad + \gamma\,(\log w_{jkm}(t+1) + 1) - \lambda_{jk} + \sum_{i \in NB_j} q_{jikm}(t) \\ &\quad - \sum_{i \in NB_j} q_{ijkm}(t) \\ &= 0\end{aligned} \tag{A.4}$$

for $1 \le j \le J$, $1 \le k \le K$, $1 \le m \le M$.

From (A.4), we have

$$\begin{aligned} w_{jkm}(t+1) &= \exp\left(-\gamma^{-1} \sum_{n=1}^{N_j} u_{jnk}(t+1) \int_{\mathbf{x}_{jn} \in R^M} (x_{jnm} - c_{jkm}(t+1))^2 \right. \\ &\quad \left. \times f_{jn}(\mathbf{x}_{jn}) d\mathbf{x}_{jn} - \gamma^{-1}\left(\sum_{i \in NB_j} q_{jikm}(t) - \sum_{i \in NB_j} q_{ijkm}(t)\right)\right) \\ &\quad \times \exp(-1 + \gamma^{-1}\lambda_{jk}) \end{aligned} \tag{A.5}$$

for $1 \le j \le J$, $1 \le k \le K$, $1 \le m \le M$.

From (A.5) and (A.3), we have (A.6), as shown at the top of the next page, for $1 \le j \le J$, $1 \le k \le K$, $1 \le m \le M$.

Similar to the cluster prototypes, the update of attribute weights is followed by a gradient descent step over the multipliers $q_{jikm}(t)$ as $q_{jikm}(t) = q_{jikm}(t-1) + \eta_2(w_{jkm}(t) - w_{ikm}(t))$ for $1 \le j \le J$, $1 \le k \le K$, $1 \le m \le M$, $i \in NB_j$, where η_2 is a positive scalar. If $q_{jikm}(0)$ and $q_{ijkm}(0)$ are initialized to zero, we have $q_{jikm}(t) = -q_{ijkm}(t)$ for $\forall t > 0$. Substituting this into (A.6), we obtain (38). This completes the proof.

Appendix B
Reduction of the WDUK-Means Algorithm to the WDK-Means Algorithm

Let $D_{jk} = \text{diag}(w_{jk1}, w_{jk2}, \ldots, w_{jkM})$ denote the weight matrix corresponding to the attribute weight vector $\mathbf{w}_{\text{jk}}$, $\mathbf{z}_{jn} = \int_{\mathbf{x}_{jn} \in R^M} f_{jn}(\mathbf{x}_{jn}) \mathbf{x}_{jn} d\mathbf{x}_{jn}$, we have

$$\begin{aligned}\sum_{m=1}^{M} w_{jkm}(x_{jnm} - c_{jkm})^2 &= \sum_{m=1}^{M} w_{jkm}(x_{jnm} - z_{jnm})^2 + \sum_{m=1}^{M} w_{jkm}(c_{jkm} - z_{jnm})^2 \\ &\quad - 2(\mathbf{c}_{jk} - \mathbf{z}_{jn})^T \cdot D_{jk} \cdot (\mathbf{x}_{jn} - \mathbf{z}_{jn}).\end{aligned} \tag{B.1}$$

According to (32), calculate the difference between $\text{ED}(\mathbf{x}_{jn}, \mathbf{c}_{\text{jk}})$ and $\text{ED}(\mathbf{x}_{jn}, \mathbf{z}_{jn})$ as

$$\text{ED}(\mathbf{o}_{jn}, \mathbf{c}_{jk}) - \text{ED}(\mathbf{o}_{jn}, \mathbf{z}_{jn}) = \sum_{m=1}^{M} w_{jkm}(c_{jkm} - z_{jnm})^2. \tag{B.2}$$

Then the objective function of the WDUK-means algorithm is transformed into

$$\begin{aligned} F(\boldsymbol{U}, \boldsymbol{C}, \boldsymbol{W}) &= \sum_{j=1}^{J} \sum_{n=1}^{K} \sum_{k=1}^{K} u_{jnk} \int_{\mathbf{x}_{jn} \in R^M} w_{jkm}(x_{jnm} - z_{jnm})^2 f_{jn}(\mathbf{x}_{jn}) d\mathbf{x}_{jn} \\ &\quad + \sum_{j=1}^{J} \sum_{n=1}^{K} \sum_{k=1}^{K} u_{jnk} \sum_{m=1}^{M} w_{jkm}(c_{jkm} - z_{jnm})^2 \\ &\quad + \gamma \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{m=1}^{M} w_{jkm} \log w_{jkm}. \end{aligned} \tag{B.3}$$

When the objects and their pdfs are fixed, the first term above is a constant. The objective function is reduced to its deterministic form as (40).

Appendix C
Consistency Analysis of the WDK-Means Algorithm

In this Appendix, we give the proof that the distributed clustering solution achieved by the WDK-means algorithm coincides with that by the centralized EWKM algorithm [31].

Theorem C.1: The cluster prototypes obtained by the WDK-means algorithm are consistent with that by the centralized clustering method.

Proof: From (42), we have

$$\sum_{n=1}^{N_j} u_{jnk} w_{jnk} c_{jkm} = \sum_{n=1}^{N_j} u_{jnk} w_{jnk} z_{jnm} - \sum_{i \in NB_j} p_{jikm} \tag{C.1}$$

for $1 \le j \le J$, $1 \le k \le K$, $1 \le m \le M$.

$$w_{jkm}(t+1) = \frac{\exp\left(-\gamma^{-1}\sum_{n=1}^{N_j} u_{jnk}(t+1)\int_{\mathbf{x}_{jn}\in R^M}(x_{jnm}-c_{jkm}(t+1))^2 f_{jn}(\mathbf{x}_{jn})dx_{jn} - \gamma^{-1}\sum_{i\in NB_j}(q_{jikm}(t)-q_{ijkm}(t))\right)}{\sum_{l=1}^{M}\exp\left(-\gamma^{-1}\sum_{n=1}^{N_j} u_{jnk}(t+1)\int_{\mathbf{x}_{jn}\in R^M}(x_{jnl}-c_{jkl}(t+1))^2 f_{jn}(\mathbf{x}_{jn})d\mathbf{x}_{jn} - \gamma^{-1}\sum_{i\in NB_j}(q_{jikl}(t)-q_{ijkl}(t))\right)} \tag{A.6}$$

Considering all peers, we have

$$\sum_{j=1}^{J}\sum_{n=1}^{N_j} u_{jnk}w_{jnk}c_{jkm} = \sum_{j=1}^{J}\sum_{n=1}^{N_j} u_{jnk}w_{jnk}z_{jnm} - \sum_{j=1}^{J}\sum_{i\in NB_j} p_{jikm} \tag{C.2}$$

for $1 \le k \le K$, $1 \le m \le M$.

Let $\mathbf{c}_k^{'} = [c_{k1}^{'}, c_{k2}^{'}, \ldots, c_{kM}^{'}]$ for $1 \le k \le K$ be the consensus cluster prototypes reached by the WDK-means algorithm, from (C.2), we have

$$c_{km}^{'}\sum_{j=1}^{J}\sum_{n=1}^{N_j} u_{jnk}w_{jnk} = \sum_{j=1}^{J}\sum_{n=1}^{N_j} u_{jnk}w_{jnk}z_{jnm} - \sum_{j=1}^{J}\sum_{i\in NB_j} p_{jikm} \tag{C.3}$$

for $1 \le k \le K$, $1 \le m \le M$.

It follows that:

$$c_{km}^{'} = \frac{\sum_{j=1}^{J}\sum_{n=1}^{N_j} u_{jnk}w_{jnk}z_{jnm} - \sum_{j=1}^{J}\sum_{i\in B_j} p_{jikm}}{\sum_{j=1}^{J}\sum_{n=1}^{N_j} u_{jnk}w_{jnk}} \tag{C.4}$$

for $1 \le k \le K$, $1 \le m \le M$.

Let $p_{jikm}(0)$ and $p_{ijkm}(0)$ be initialized to zero, from (43), we have $p_{jikm}(t) = -p_{ijkm}(t)$ for $\forall t > 0$. Therefore, we have

$$\sum_{j=1}^{J}\sum_{i\in NB_j} p_{jikm} = 0 \tag{C.5}$$

for $1 \le k \le K$, $1 \le m \le M$.

Substituting (C.5) into (C.4), we obtain

$$c_{km}^{'} = \sum_{j=1}^{J}\sum_{n=1}^{N_j} u_{jnk}z_{jnm} \Bigg/ \sum_{j=1}^{J}\sum_{n=1}^{N_j} u_{jnk} \tag{C.6}$$

for $1 \le k \le K$, $1 \le m \le M$.

Note that if all objects are collected into one central unit from all peers, the cluster prototypes in (C.6) are consistent with the centralized ones obtained by the EWKM algorithm [31]. This completes the proof.

Theorem C.2: The attribute weights obtained by the WDK-means algorithm are consistent with that by the centralized clustering method.

Proof: Let $\mathbf{w}_k^{'} = [w_{k1}^{'}, w_{k2}^{'}, \ldots, w_{kM}^{'}]$ for $1 \le k \le K$ be the consensus attribute weights reached by the WDK-means algorithm, from (44), we have (C.7), as shown at the bottom of this page, for $1 \le k \le K$, $1 \le m \le M$.

It follows that (C.8), as shown at the bottom of this page, for $1 \le k \le K$, $1 \le m \le M$.

According to the constraint of attribute weights, we have

$$\sum_{m=1}^{M} w_{km}^{'} = 1 \tag{C.9}$$

for $1 \le k \le K$.

Then we have (C.10), as shown at the top of the next page.

From (C.8), as shown at the bottom of this page, and (C.10), we obtain (C.11), as shown at the top of the next page.

Let $q_{jikm}(0)$ and $q_{ijkm}(0)$ be initialized to zero, from (45), we have $q_{jikm}(t) = -q_{ijkm}(t)$ for $\forall t > 0$. Then we have

$$\sum_{j=1}^{J}\sum_{i\in NB_j} q_{jikm} = 0 \tag{C.12}$$

for $1 \le k \le K$, $1 \le m \le M$.

Substituting (C.12) into (C.11), we obtain

$$w_{km}^{'} = \frac{\sqrt[J]{\exp\left(-\gamma^{-1}\sum_{j=1}^{J}\sum_{n=1}^{N_j} u_{jnk}(c_{jkm}-z_{jnm})^2\right)}}{\sum_{l=1}^{M}\sqrt[J]{\exp\left(-\gamma^{-1}\sum_{j=1}^{J}\sum_{n=1}^{N_j} u_{jnk}(c_{jkl}-z_{jnl})^2\right)}} \tag{C.13}$$

for $1 \le k \le K$, $1 \le m \le M$.

Similar to the consistence proof of cluster prototypes, by comparing (C.13) with the attribute weights (C.14) obtained by the centralized EWKM algorithm [31], we know that (C.14)

$$(w_{km}^{'})^J = \frac{\prod_{j=1}^{J}\exp\left(-\gamma^{-1}\sum_{n=1}^{N_j} u_{jnk}(c_{jkm}-z_{jnm})^2 - 2\gamma^{-1}\sum_{i\in NB_j} q_{jikm}\right)}{\prod_{j=1}^{J}\sum_{l=1}^{M}\exp\left(-\gamma^{-1}\sum_{n=1}^{N_j} u_{jnk}(c_{jkl}-z_{jnl})^2 - 2\gamma^{-1}\sum_{i\in NB_j} q_{jikl}\right)} \tag{C.7}$$

$$w_{km}^{'} = \frac{\sqrt[J]{\exp\left(-\gamma^{-1}\sum_{j=1}^{J}\sum_{n=1}^{N_j} u_{jnk}(c_{jkm}-z_{jnm})^2 - 2\gamma^{-1}\sum_{j=1}^{J}\sum_{i\in NB_j} q_{jikm}\right)}}{\sqrt[J]{\prod_{j=1}^{J}\sum_{l=1}^{M}\exp\left(-\gamma^{-1}\sum_{n=1}^{N_j} u_{jnk}(c_{jkl}-z_{jnl})^2 - 2\gamma^{-1}\sum_{i\in NB_j} q_{jikl}\right)}} \tag{C.8}$$

$$\frac{\sum_{m=1}^{M} \sqrt[J]{\exp\left(-\gamma^{-1}\sum_{j=1}^{J}\sum_{n=1}^{N_j} u_{jnk}(c_{jkm}-z_{jnm})^2 - 2\gamma^{-1}\sum_{j=1}^{J}\sum_{i\in NB_j} q_{jikm}\right)}}{\sqrt[J]{\prod_{j=1}^{J}\sum_{l=1}^{M}\exp\left(-\gamma^{-1}\sum_{n=1}^{N_j} u_{jnk}(c_{jkl}-z_{jnl})^2 - 2\gamma^{-1}\sum_{i\in NB_j} q_{jikl}\right)}} = 1 \tag{C.10}$$

$$w'_{km} = \frac{\sqrt[J]{\exp\left(-\gamma^{-1}\sum_{j=1}^{J}\sum_{n=1}^{N_j} u_{jnk}(c_{jkm}-z_{jnm})^2 - 2\gamma^{-1}\sum_{j=1}^{J}\sum_{i\in NB_j} q_{jikm}\right)}}{\sum_{l=1}^{M}\sqrt[J]{\exp\left(-\gamma^{-1}\sum_{j=1}^{J}\sum_{n=1}^{N_j} u_{jnk}(c_{jkl}-z_{jnl})^2 - 2\gamma^{-1}\sum_{j=1}^{J}\sum_{i\in NB_j} q_{jikl}\right)}} \tag{C.11}$$

is a special case of (C.13) in which all the data are collected in one node

$$w_{km} = \frac{\exp\left(-\gamma\sum_{n=1}^{N} u_{nk}(c_{km}-z_{nm})^2\right)}{\sum_{l=1}^{M}\exp\left(-\gamma\sum_{n=1}^{N} u_{nk}(c_{kl}-z_{nl})^2\right)}. \tag{C.14}$$

This completes the proof.

References

[1] V. N. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley, 1998.

[2] L. Wang, B. Yang, Y. Chen, X. Zhang, and J. Orchard, "Improving neural-network classifiers using nearest neighbor partitioning," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2016.2580570, 2016.

[3] L. Chen, C. L. P. Chen, and M. Lu, "A multiple-kernel fuzzy C-means algorithm for image segmentation," *IEEE Trans. Syst., Man B, Cybern.*, vol. 41, no. 5, pp. 1263–1274, Oct. 2011.

[4] R. Xu and D. Wunsch, II, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, May 2005.

[5] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, vol. 1. 1967, pp. 281–297.

[6] T. Zhang, L. Chen, and C. L. P. Chen, "Clustering algorithm based on spatial shadowed fuzzy C-means and I-ching operators," *Int. J. Fuzzy Syst.*, vol. 18, no. 4, pp. 609–617, Aug. 2016.

[7] L. Chen, J. Zou, and C. L. P. Chen, "Kernel spatial shadowed C-Means for image segmentation," *Int. J. Fuzzy Syst.*, vol. 16, no. 1, pp. 46–56, Mar. 2014.

[8] D. Pollard, "Strong consistency of K-means clustering," *Ann. Statist.*, vol. 9, no. 1, pp. 135–140, Jan. 1981.

[9] V. Lyzinski, D. L. Sussman, M. Tang, A. Athreya, and C. Priebe, "Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding," *Electron. J. Statist.*, vol. 8, no. 2, pp. 2905–2922, Mar. 2014.

[10] R. Cheng, D. V. Kalashnikov, and S. Prabhakar, "Querying imprecise data in moving object environments," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 9, pp. 1112–1127, Sep. 2004.

[11] A. D. Sarma, O. Benjelloun, A. Halevy, S. Nabar, and J. Widom, "Representing uncertain data: Models, properties, and algorithms," *VLDB J.*, vol. 18, no. 5, pp. 989–1019, May 2009.

[12] C. C. Aggarwal and C. K. Reddy, *Data Clustering: Algorithms and Applications*. Boca Raton, FL, USA: CRC Press, Sep. 2013.

[13] M. Chau, R. Cheng, B. Kao, and J. Ng, "Uncertain data mining: An example in clustering location data," in *Proc. 10th Pacific–Asia Conf. Knowl. Discovery Data Mining*, vol. 3918. Apr. 2006, pp. 199–204.

[14] H. P. Kriegel and M. Pfeifle, "Density-based clustering of uncertain data," in *Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2005, pp. 672–677.

[15] B. Jiang, J. Pei, Y. Tao, and X. Lin, "Clustering uncertain data based on probability distribution similarity," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 4, pp. 751–763, Apr. 2013.

[16] W. K. Ngai, B. Kao, C. K. Chui, R. Cheng, M. Chau, and K. Y. Yip, "Efficient clustering of uncertain data," in *Proc. 6th IEEE Int. Conf. Data Mining*, Dec. 2006, pp. 436–445.

[17] B. Jiang, J. Pei, X. Lin, and Y. Yuan, "Probabilistic skylines on uncertain data: Model and bounding-pruning-refining methods," *J. Intell. Inf. Syst.*, vol. 38, no. 1, pp. 1–39, Feb. 2012.

[18] L. Xiao and E. Hung, "An efficient distance calculation method for uncertain objects," in *Proc. IEEE Symp. Comput. Intell. Data Mining*, Mar. 2007, pp. 10–17.

[19] S. D. Lee, B. Kao, and R. Cheng, "Reducing UK-means to K-means," in *Proc. 7th IEEE Int. Conf. Data Mining Workshops*, Oct. 2007, pp. 483–488.

[20] H. P. Kriegel and M. Pfeifle, "Hierarchical density-based clustering of uncertain data," in *Proc. 5th IEEE Int. Conf. Data Mining*, Nov. 2005, pp. 689–692.

[21] X. Liu and W. B. Croft, "Cluster-based retrieval using language models," in *Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 1999, pp. 254–261.

[22] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *J. Mach. Learn. Res.*, vol. 6, no. 4, pp. 1705–1749, 2005.

[23] H. Kargupta and K. Sivakumar, *Data Mining: Next Generation Challenges and Future Directions*. Cambridge, MA, USA: MIT Press, 2004.

[24] J. Zhou, C. L. P. Chen, L. Chen, and H. X. Li, "A collaborative fuzzy clustering algorithm in distributed network environments," *IEEE Trans. Fuzzy Syst.*, vol. 22, no. 6, pp. 1443–1456, Dec. 2014.

[25] S. Datta, C. Giannella, and H. Kargupta, "K-means clustering over a large, dynamic network," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2006, pp. 153–164.

[26] P. A. Forero, A. Cano, and G. B. Giannakis, "Distributed clustering using wireless sensor networks," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 4, pp. 707–724, Aug. 2011.

[27] W. Pedrycz, "Collaborative fuzzy clustering," *Pattern Recognit. Lett.*, vol. 23, no. 14, pp. 1675–1686, Dec. 2002.

[28] W. Pedrycz and P. Rai, "Collaborative clustering with the use of fuzzy c-means and its quantification," *Fuzzy Sets Syst.*, vol. 159, no. 18, pp. 2399–2427, Sep. 2008.

[29] K. M. Hammouda, "Distributed document clustering and cluster summarization in peer-to-peer environments," Ph.D. dissertation, Dept. Elect. Comput. Eng., Univ. Waterloo, Waterloo, ON, Canada, 2007.

[30] H. P. Kriegel, P. Kröger, and A. Zimek, "Clustering high-dimensional data: a survey on subspace clustering, pattern-based clustering, and correlation clustering," *ACM Trans. Knowl. Discovery Data*, vol. 3, no. 1, pp. 1–58, Mar. 2009.

[31] L. Jing, M. K. Ng, and J. Z. Huang, "An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 8, pp. 1026–1041, Aug. 2007.

[32] J. Zhou, L. Chen, C. L. P. Chen, Y. Zhang, and H.-X. Li, "Fuzzy clustering with the entropy of attribute weights," *Neurocomputing*, vol. 198, pp. 125–134, Jul. 2016.

[33] D. W. Scott, *Multivariate Density Estimation: Theory, Practical, and Visualization*. Hoboken, NJ, USA: Wiley, 1992.

[34] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. London, U.K.: Chapman & Hall, 1986.

[35] J. C. Bezdek, "A convergence theorem for the fuzzy ISODATA clustering algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI–2, no. 1, pp. 1–8, Jan. 1980.

[36] D. Graves and W. Pedrycz, "Kernel-based fuzzy clustering and fuzzy clustering: A comparative experimental study," *Fuzzy Sets Syst.*, vol. 161, no. 4, pp. 522–543, Feb. 2010.

[37] H. C. Huang, Y. Y. Chuang, and C. S. Chen, "Multiple kernel fuzzy clustering," *IEEE Trans. Fuzzy Syst.*, vol. 20, no. 1, pp. 120–134, Feb. 2012.

[38] L. Hubert and P. Arabie, "Comparing partitions," *J. Classification*, vol. 2, no. 1, pp. 193–218, 1985.

[39] K. Bache and M. Lichman, "UCI machine learning repository," School Inform. Comput. Sci., Univ. California, Irvine, CA, USA, Tech. Rep., 2013. [Online]. Available: http://archive.ics.uci.edu/ml/

[40] P. Guo, C. L. P. Chen, and M. R. Lyu, "Cluster number selection for a small set of samples using the Bayesian Ying-Yang model," *IEEE Trans. Neural Netw.*, vol. 13, no. 3, pp. 757–763, May 2002.

[41] M. J. Li, M. K. Ng, Y.-M. Cheung, and J. Z. Huang, "Agglomerative fuzzy K-means clustering algorithm with selection of number of clusters," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 11, pp. 1519–1534, Nov. 2008.

Jin Zhou (M'16) received the B.S. degree in computer science and technology and the M.S. degree in software engineering from Shandong University, Jinan, China, in 1998 and 2001, respectively, and the Ph.D. degree in software engineering from the University of Macau, Macau, China, in 2014.

He is currently an Associate Professor with the School of Information Science and Engineering, University of Jinan, Jinan. His current research interests include computational intelligence, and other machine learning techniques and their applications.

Long Chen (M'11) received the B.S. degree in information sciences from Peking University, Beijing, China, in 2000, the M.S.E. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2003, the M.S. degree in computer engineering from the University of Alberta, Edmonton, AB, Canada, in 2005, and the Ph.D. degree in electrical engineering from the University of Texas at San Antonio, San Antonio, TX, USA, in 2010.

From 2010 to 2011, he was a Post-Doctoral Fellow at the University of Texas at San Antonio. He is currently an Assistant Professor with the Department of Computer and Information Science, University of Macau, Macau, China. His current research interests include computational intelligence, Bayesian methods, and other machine learning techniques and their applications.

Mr. Chen has been involved in publication matters for many IEEE conferences, and was the Publications Co-Chair of the IEEE International Conference on Systems, Man, and Cybernetics in 2009.

C. L. Philip Chen (S'88–M'88–SM'94 –F'07) received the M.S. degree in electrical engineering from the University of Michigan, Ann Arbor, MI, USA, in 1985, and the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, IN, USA, in 1988.

He was a tenured professor, department head, and an associate dean with two different departments and universities in the US. He is currently a Chair Professor with the Department of Computer and Information Science and the Dean of the Faculty of Science and Technology with the University of Macau, Macau, China. His current research interests include computational intelligence, systems, and cybernetics.

Dr. Chen is a fellow of the American Association for the Advancement of Science, and HKIE. He was the President of the IEEE Systems, Man, and Cybernetics Society, from 2012 to 2013. In addition, he has served on different committees such as the IEEE Fellows Committee and Conference Integrity Committee. He is currently an Accreditation Board of Engineering and Technology Education Program Evaluator for computer engineering, electrical engineering, and software engineering programs, and the Editor-in-Chief of the IEEE SMC-Systems.

Yingxu Wang received the B.S. degree in electronics and information engineering from Beihang University, Beijing, China, in 2011. He is currently pursuing the M.S. degree with the School of Information Science and Engineering, University of Jinan, Jinan, China.

His current research interests include machine learning, and image segmentation and their applications.

Han-Xiong Li (S'94–M'97–SM'00–F'11) received the B.E. degree in aerospace engineering from the National University of Defense Technology, Changsha, China, the M.E. degree in electrical engineering from the Delft University of Technology, Delft, The Netherlands, and the Ph.D. degree in electrical engineering from the University of Auckland, Auckland, New Zealand.

He has involved in different fields, including military service, industry, and academia. He is currently a Professor with the Department of Systems Engineering and Engineering Management, the City University of Hong Kong, Hong Kong. He has authored over 140 SCI journal papers with an H-index 25. His current research interests include system intelligence and control, integrated process design and control, and distributed parameter systems with applications to electronics packaging.

Dr. Li received the Distinguished Young Scholar (overseas) by the China National Science Foundation in 2004, a Chang Jiang professor by the Ministry of Education, China in 2006, and a national professorship in the China Thousand Talents Program in 2010. He serves as an Associate Editor of the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, PART-B, the IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, and also serves as a Distinguished Expert for Hunan Government and the China Federation of Returned Overseas.