Clustering Uncertain Data via Representative Possible Worlds with Consistency Learning

Han Liu[1] , **Xianchao Zhang**[2] , **Xiaotong Zhang**[1] , **Qimai Li**[1] , **Xiao-Ming Wu**[1]
[1]The Hong Kong Polytechnic University, Hong Kong
[2]Dalian University of Technology, China
liu.han.dut@gmail.com, xczhang@dlut.edu.cn,
zxt.dut@hotmail.com, {csqmli,csxmwu}@comp.polyu.edu.hk

Abstract

Clustering uncertain data is an essential task in data mining for the internet of things. Possible world based algorithms seem promising for clustering uncertain data. However, there are two issues in existing possible world based algorithms: (1) They rely on all the possible worlds and treat them equally, but some marginal possible worlds may cause negative effects. (2) They do not well utilize the consistency among possible worlds, since they conduct clustering or construct the affinity matrix on each possible world independently. In this paper, we propose a representative possible world based consistent clustering (RPC) algorithm for uncertain data. First, by introducing representative loss and using Jensen-Shannon divergence as the distribution measure, we design a heuristic strategy for the selection of representative possible worlds, thus avoiding the negative effects caused by marginal possible worlds. Second, we integrate a consistency learning procedure into spectral clustering to deal with the representative possible worlds synergistically, thus utilizing the consistency to achieve better performance. Experimental results show that our proposed algorithm performs better than the state-of-the-art algorithms.

1 Introduction

Most existing clustering algorithms focus on certain data. However, due to various reasons like randomness in data generation and collection, imprecision in physical measurement, privacy concern and data staling [Aggarwal and Yu, 2009], uncertain data is ubiquitous in many real applications, such as sensor networks, biomedical measurement, location tracking, meteorological forecasting and so on [Zhang *et al.*, 2017]. Uncertain data has posed a serious challenge to existing clustering algorithms.

Several algorithms have been proposed for clustering uncertain data. Partition-based algorithms, e.g., UK-means [Chau *et al.*, 2006] and UK-medoids [Gullo *et al.*, 2008], use expected distance or uncertain distance to extend traditional k-means or k-medoids to deal with uncertain data. However, they reduce complex probability distributions to a single probability distribution or a determinate value, thus cannot handle the uncertain information well [Zhang *et al.*, 2017]. Density-based algorithms, e.g., FDBSCAN [Kriegel and Pfeifle, 2005a] and FOPTICS [Kriegel and Pfeifle, 2005b], extend traditional DBSCAN [Ester *et al.*, 1996] or OPTICS [Ankerst *et al.*, 1999] for clustering uncertain data by use of probabilistic definitions. However, they suffer from the unreasonable independent distance assumption [Züfle *et al.*, 2014], thus are difficult to obtain satisfactory performance.

Different from partition-based and density-based algorithms, possible world based algorithms, e.g., SC [Volk *et al.*, 2009] and REP [Züfle *et al.*, 2014], employ multiple independent and identically distributed realizations of an uncertain dataset to deal with data uncertainty, thus reducing the loss of uncertain information and avoiding the independent distance assumption. However, they have two unaddressed issues: (1) They rely on all the possible worlds and treat them equally, but some marginal possible worlds may cause negative effects on the clustering result. (2) They do not well utilize the consistency among possible worlds, since they conduct clustering or construct the affinity matrix on each possible world independently. Nevertheless, the consistency is important since different possible worlds can utilize it to transfer useful information for improving the performance.

In this paper, we propose a representative possible world based consistent clustering (RPC) algorithm for uncertain data, which improves existing algorithms from the following two aspects: (1) To alleviate the negative effects caused by marginal possible worlds, we introduce the definition of representative loss, use Jensen-Shannon divergence as the distribution measure, and then design a heuristic strategy for the selection of representative possible worlds. This strategy can be used by any possible world based algorithm to improve the performance. (2) To utilize the consistency to achieve better performance, we integrate a consistency learning procedure into spectral clustering to deal with the representative possible worlds synergistically. Extensive experimental results demonstrate the superiority of the proposed algorithm over the existing ones.

2 Related Work

Traditional algorithms. (1) *Partition-based algorithms*: UK-means [Chau *et al.*, 2006] is the first partition-based algorithm for clustering uncertain data. It extends the traditional k-

means by using expected distance. To improve the efficiency of UK-means, [Kao *et al.*, 2008; Kao *et al.*, 2010; Ngai *et al.*, 2011] use various pruning techniques to avoid the computation of redundant expected distances. CK-means [Lee *et al.*, 2007] optimizes UK-means by resorting to the moment of inertia of rigid bodies. DUK-means [Zhou *et al.*, 2018] is an improved version of UK-means, which is specifically designed for distributed network environment. UK-medoids [Gullo *et al.*, 2008] employs uncertain distance to extend the traditional k-medoids. MMVar [Gullo *et al.*, 2010] uses a novel objective function which aims to minimize the variance of cluster mixture models. UCPC [Gullo and Tagarelli, 2012] introduces the notion of uncertain centroid and it is a local search based heuristic algorithm. All these algorithms can deal with uncertain data to some extent. However, they reduce complex probability distributions to a single probability distribution or a determinate value, thus cannot handle the uncertain information well [Zhang *et al.*, 2017]. (2) *Density-based algorithms*: FDBSCAN [Kriegel and Pfeifle, 2005a] and FOPTICS [Kriegel and Pfeifle, 2005b] are the first density-based and hierarchical density-based algorithms for clustering uncertain data respectively. They introduce a series of probabilistic definitions to extend the traditional DBSCAN or OPTICS. Zhang et al. [Zhang *et al.*, 2017] analyze the limitations in FDBSCAN and FOPTICS, and propose a novel density-based algorithm PDBSCAN for clustering uncertain data. However, these algorithms rely on the unreasonable independent distance assumption [Züfle *et al.*, 2014], thus are difficult to obtain satisfactory clustering results.

Possible world based algorithms. SC [Volk *et al.*, 2009] is the first possible world based algorithm for clustering uncertain data. It conducts clustering on each possible world independently and integrates the clustering results into one final result. REP [Züfle *et al.*, 2014] also conducts clustering on each possible world independently, but it selects the representative clustering result as the final result. Recently, [Liu *et al.*, 2018] tries to leverage the consistency principle for clustering uncertain data. It constructs the affinity matrix for each possible world independently and then learns a consensus affinity matrix for clustering uncertain data. However, the consistency learning method introduced in [Liu *et al.*, 2018] lacks the procedure of updating the affinity matrix of each possible world, thus reducing the ability of consistency learning. Possible world based algorithms avoid the issues in traditional algorithms and seem more promising for clustering uncertain data. However, as we point out hereinafter, there are some unaddressed issues in existing possible world based algorithms.

3 Preliminaries

3.1 Consistency Principle

Consistency principle is a common assumption, which has been widely used in many machine learning methods [Liu *et al.*, 2017; Ma *et al.*, 2017]. Its definition is as follows [Wang and Zhou, 2010].

Definition 1 *(Consistency principle). Given a dataset which has multiple representations, consistency principle refers to an assumption that the class labels and cluster structures of the multiple representations are consistent.*

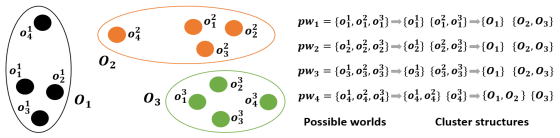


Figure 1: Consistency principle for possible world.

By using consistency principle to minimize the disagreement of different representations, we can improve the algorithm performance greatly. The detailed proof can refer to [Dasgupta *et al.*, 2001].

3.2 Uncertain Data and Possible World

Uncertain data can be considered at table, tuple or attribute level [Sarma *et al.*, 2009]. For uncertain data clustering, we mainly focus on attribute level uncertainty. That is to say, each uncertain object is represented as a random variable with a probability distribution, which is associated with the probability that the object appears at any position in a multidimensional space.

Possible world is an effective tool to model uncertain data [Dalvi and Suciu, 2007; Sarma *et al.*, 2009]. Its definition is as follows [Hua and Pei, 2011].

Definition 2 *(Possible world). Let* $UD = \{O_1, O_2, ..., O_n\}$ *be an uncertain dataset. A possible world* $pw = \{o_1, o_2, ..., o_n\}$ $(o_i \in O_i)$ *is a set of instances such that each instance is taken from each corresponding uncertain object. Let* PW *be the set of all the possible worlds,* $P(pw)$ *be the existence probability of* pw*, then* $\sum_{pw \in PW} P(pw) = 1$.

Possible world can be generated through the inversion method. Due to space limitation, more information and proofs can refer to [Devroye, 1986; Jampani *et al.*, 2008].

Consistency Principle for Possible World

According to the definition of possible world, different possible worlds come from the same uncertain dataset and they are a number of independent and identically distributed realizations of an uncertain dataset [Hua and Pei, 2011]. Therefore, if we treat each possible world as one representation of the uncertain dataset, by the concept of consistency principle, we can have the following consistency principle for possible world: *the class labels and cluster structures of different possible worlds are consistent*.

In general, the consistency principle for possible world conforms to the reality well, i.e., in most cases the class labels and cluster structures of different possible worlds are consistent. For example, in Figure 1, O_1, O_2, O_3 are uncertain objects, and o_1^i, o_2^i, o_3^i, o_4^i are the possible instances of O_i $(i \in \{1, 2, 3\})$. If we divide O_1, O_2, O_3 into two clusters, based on the geometric information, O_1 should belong to one cluster, O_2 and O_3 should belong to the other cluster. For the possible worlds pw_1, pw_2, pw_3 with their components shown in Figure 1, it is easy to find that their class labels and cluster structures are consistent.

However, the consistency principle for possible world is not absolute. In some cases, abnormal possible worlds violate the principle, and we call this kind of possible worlds as the

marginal ones. Formally, the definition of marginal possible world is as follows.

Definition 3 *(Marginal possible world). Let PW be the set of all the possible worlds, marginal possible world refers to the possible world whose class label and cluster structure have large differences with most possible worlds in PW.*

For example, in Figure 1, pw_4 is a possible world which consists of some abnormal instances. As the class label and cluster structure of pw_4 are very different from most possible worlds, pw_4 is a marginal possible world.

3.3 Unaddressed Issues

(1) Negative effects caused by marginal possible worlds. Existing possible world based algorithms rely on all the possible worlds and treat them equally. However, marginal possible worlds belong to the abnormal ones, their class labels and cluster structures have large differences with most possible worlds, which may disturb the integrating or selecting procedure of existing possible world based algorithms and cause negative effects on the clustering result. To solve this issue, we propose to select some representative possible worlds to filter out marginal possible worlds. By **representative possible worlds** we mean a subset of all the possible worlds which has a strong ability to represent all the possible worlds. As marginal possible worlds are abnormal and their representative ability is weak, we can filter out marginal possible worlds and avoid the negative effects by selecting representative possible worlds.

(2) Utilizing the consistency principle for possible world not well. The consistency principle makes it possible to transfer useful information among different possible worlds, which can potentially improve the clustering quality. However, existing possible world based algorithms conduct clustering or construct the affinity matrix on each possible world independently, thus cannot well utilize the consistency among possible worlds. To tackle this issue, we propose a consistent spectral clustering method which can update the eigenvector matrix of each possible world iteratively and minimize the disagreement of different possible worlds, thus better achieving the consistency learning and improving the performance.

4 The Proposed Algorithm

The proposed algorithm consists of two parts: selecting representative possible worlds and consistent spectral clustering.

Given an uncertain dataset $UD = \{O_1, O_2, ..., O_n\}$ in a d-dimensional independent space, $PW = \{pw_i | i = 1, 2, ..., M\}$, $PWR = \{pwr_j | j = 1, 2, ..., R\}$, $PWU = \{pwu_k | k = 1, 2, ..., M - R\}$ respectively denote the set of all the possible worlds, the representative possible world set and the unrepresentative possible world set. M, R and $M-R$ respectively denote the number of possible worlds in PW, PWR and PWU. Here $PW = PWR \cup PWU$.

4.1 Selecting Representative Possible Worlds

By selecting representative possible worlds, we can filter out marginal possible worlds and avoid the waste of time caused by redundant possible worlds. In order to select representative possible worlds, we introduce the definition of representative loss, use Jensen-Shannon divergence as the distribution measure, and then design a heuristic strategy for the selection of representative possible worlds.

Representative Loss

Intuitively, given any two possible worlds pw and pw', if we want to use pw to represent pw', then the smaller the difference between pw and pw', the less the loss that pw represents pw'. We aim to select PWR from PW to represent PW. As $PW = PWR \cup PWU$ and the loss that PWR represents PWR is equal to 0, then the loss that PWR represents PW is equal to the loss that PWR represents PWU. Based on these observations, we have the following definition.

Definition 4 *(Representative loss). Let PWR be the representative possible world set and $pwr_j \in PWR$, PWU be the unrepresentative possible world set and $pwu_k \in PWU$. If using PWR to represent PWU, then the representative loss, denoted by $L(PWR \to PWU)$, can be defined as:*

$$L(PWR \to PWU) = \sum_{k=1}^{M-R} \min_{pwr_j} \Phi(pwr_j, pwu_k), \quad (1)$$

where $\Phi(pwr_j, pwu_k)$ is the difference between pwr_j and pwu_k, $M - R$ is the number of possible worlds in PWU.

From this definition, it can be seen that if we know how to compute the difference between possible worlds, we can get the representative loss that PWR represents PWU, i.e., the representative loss that PWR represents PW.

Jensen-Shannon Divergence between Possible Worlds

As a possible world can be regarded as a probability distribution, we can compute the difference between possible worlds by Jensen-Shannon divergence [Lin, 1991]. Compared with KL divergence [Kullback and Leibler, 1951], Jensen-Shannon divergence is symmetric and finite, therefore it is more suitable as the representative loss measure.

Given any two possible worlds pw and pw', the Jensen-Shannon divergence between them can be defined as:

$$JSD(pw||pw') = \frac{1}{2}D(P_{pw}||H) + \frac{1}{2}D(P_{pw'}||H), \quad (2)$$

where P_{pw} and $P_{pw'}$ are the probability distributions of pw and pw' respectively, and $H = \frac{1}{2}(P_{pw} + P_{pw'})$. $D(P||Q)$ is the KL divergence between two probability distributions P and Q. For continuous probability distributions P and Q with a variable x in a domain $\mathbb{D}$, $D(P||Q)$ is defined as:

$$D(P||Q) = \int_{\mathbb{D}} f(x) log \frac{f(x)}{g(x)} dx, \quad (3)$$

where $f(x)$ and $g(x)$ are the probability density functions of P and Q. $D(P||Q)$ can also be expressed as:

$$D(P||Q) = E(log \frac{f(x)}{g(x)}), \quad (4)$$

where E denotes the expectation. According to the law of large numbers and Eq.(4), given a sample set S, $D(P||Q)$ can be estimated by:

$$D(P||Q) = \frac{1}{|S|} \sum_{x \in S} log \frac{f(x)}{g(x)}, \quad (5)$$

where $|S|$ denotes the number of objects in S.

We employ the kernel density estimation method [Silverman, 1986] to obtain the probability density functions f_{pw} and $f_{pw'}$ of the probability distributions P_{pw} and $P_{pw'}$. Specifically, f_{pw} can be estimated as:

$$f_{pw}(x) = \frac{1}{|pw| \prod_{j=1}^{d} h_j} \sum_{o \in pw} \prod_{j=1}^{d} K(\frac{x.D_j - o.D_j}{h_j}). \quad (6)$$

In Eq.(6), o denotes an object in pw and it can be represented by $(o.D_1, o.D_2, ..., o.D_d)$, d denotes the total dimensionality, and $|pw|$ denotes the number of objects in pw. K denotes the kernel function, and we use the most common Gaussian kernel function. h_j denotes the bandwidth of the j-th dimension. For Gaussian kernel function, we set $h_j = 1.06 \times \delta_j |pw|^{-\frac{1}{5}}$ according to the Silverman's rule of thumb [Silverman, 1986], where δ_j is the standard deviation of the j-th dimension of the objects in pw.

By using Jensen-Shannon divergence as the distribution measure to compute the difference between possible worlds, i.e., replacing $\Phi(pwr_j, pwu_k)$ in Eq.(1) with $JSD(pwr_j||pwu_k)$, we can get the representative loss.

Selection Strategy

Our goal is to select a given number of possible worlds as the representative possible worlds. In general, a good representative possible world set should have a strong representative ability, i.e., its corresponding representative loss should be small. Inspired by this observation, we propose the following selection strategy:

Let PWR be the representative possible world set, and PWU be the unrepresentative possible world set. Now select a possible world pwu^* from PWU and move pwu^* to PWR, if we want the new representative possible world set $PWR \cup pwu^*$ to be the best, then the selection strategy should ensure the representative loss that $PWR \cup pwu^*$ represents $PWU \backslash pwu^*$ to be the minimum. Formally:

$$pwu^* = \arg\min_{pwu^*} L(PWR \cup pwu^* \rightarrow PWU \backslash pwu^*). \quad (7)$$

From Eq.(7), it can be seen that pwu^* should have a strong representative ability. Marginal possible worlds are the abnormal ones and their representative ability is poor, therefore this selection strategy can filter out marginal possible worlds.

Based on the selection strategy, we design a heuristic method to select the representative possible worlds, which is shown in Algorithm 1 (Part 1).

4.2 Consistent Spectral Clustering

We integrate a consistency learning procedure into spectral clustering to deal with the representative possible worlds synergistically.

Spectral Clustering

Assume that pwr_j is a possible world from the representative possible world set PWR and $PWR = \{pwr_j | j = 1, 2, ..., R\}$, where R denotes the number of possible worlds in PWR. $W^{(j)}$ is the similarity matrix of pwr_j, which is computed by the Gaussian kernel. $L^{(j)}$ is the normalized Laplacian matrix of pwr_j and $L^{(j)} = {D^{(j)}}^{-\frac{1}{2}} W^{(j)} {D^{(j)}}^{-\frac{1}{2}}$.

Algorithm 1 RPC

Input: Uncertain dataset $UD = \{O_1, O_2, ..., O_n\}$, the number of clusters k, the number of all the possible worlds M, the number of representative possible worlds R

Output: The clusters $C_1, C_2, ..., C_k$

Part 1: Selecting representative possible worlds (Lines 1-5)

1: Generate PW, initialize $PWR = \emptyset$ and $PWU = PW$, and calculate the JSD between any two possible worlds in PW
2: **Repeat**
3: Select a possible world pwu^* from PWU by Eq.(7)
4: $PWR \leftarrow PWR \cup pwu^*$, $PWU \leftarrow PWU \backslash pwu^*$
5: **Until** $|PWR| \geqslant R$, $|PWR|$ denotes the current number of possible worlds in PWR

Part 2: Consistent spectral clustering (Lines 6-12)

6: For $\forall pwr_j \in PWR$, compute $W^{(j)}$, $D^{(j)}$, $L^{(j)}$
7: For $\forall pwr_j \in PWR$, compute the k eigenvectors corresponding to the k largest eigenvalues of $L^{(j)}$ and use them to initialize the corresponding $U^{(j)}$
8: **Repeat**
9: Update U^* by solving Eq.(13)
10: Update each $U^{(j)}$ by solving Eq.(15)
11: **Until** Eq.(11) is convergent
12: Run k-means on U^* and get the clusters $C_1, C_2, ..., C_k$

$D^{(j)}$ is a diagonal matrix and $D^{(j)}(i,i) = \sum_{l=1}^{n} W^{(j)}(i,l)$, where n denotes the number of objects in pwr_j. For the representative possible world pwr_j, the objective function of spectral clustering is:

$$\max_{U^{(j)}} tr({U^{(j)}}^T L^{(j)} U^{(j)}), \ \ s.t. \ {U^{(j)}}^T U^{(j)} = I, \quad (8)$$

where $tr(\cdot)$ denotes the trace of a matrix, and the solution of $U^{(j)} \in \mathbb{R}^{n \times k}$ is composed by k eigenvectors corresponding to the k largest eigenvalues of $L^{(j)}$.

Consistency Learning

The eigenvector matrix $U^{(j)}$ can reflect the cluster structure of the representative possible world pwr_j. To meet the requirement of consistency, we assume that each eigenvector matrix $U^{(j)} \in \mathbb{R}^{n \times k}$ tends to a common eigenvector matrix $U^* \in \mathbb{R}^{n \times k}$. Then by minimizing the disagreement between each $U^{(j)}$ and U^*, we can achieve the consistency learning among different possible worlds. For the disagreement between $U^{(j)}$ and U^*, we use the squared Euclidean distance between the similarity matrices to measure it:

$$\begin{aligned} &Dis(U^{(j)}, U^*) = ||S_{U^{(j)}} - S_{U^*}||_F^2, \\ &s.t. \ {U^{(j)}}^T U^{(j)} = I, \ \ {U^*}^T U^* = I, \end{aligned} \quad (9)$$

where $S_{U^{(j)}}$ and S_{U^*} denote the similarity matrices of $U^{(j)}$ and U^*, and $||\cdot||_F$ denotes the Frobenius norm of the matrix.

Considering the feasibility of optimization, we use the commonly adopted inner product to compute the similarity matrix, i.e., $S_{U^{(j)}} = U^{(j)} {U^{(j)}}^T$. Then with some manipulations, minimizing Eq.(9) can be transformed as:

$$\begin{aligned} &\max_{U^{(j)}, U^*} tr(U^{(j)} {U^{(j)}}^T U^* {U^*}^T), \\ &s.t. \ {U^{(j)}}^T U^{(j)} = I, \ \ {U^*}^T U^* = I. \end{aligned} \quad (10)$$

Overall Objective Function and Optimization

By integrating the objective functions of spectral clustering and consistency learning, we can get the overall objective

function of consistent spectral clustering as follows:

$$\max_{U^{(j)},U^*} \sum_{j=1}^{R} (tr(U^{(j)^T} L^{(j)} U^{(j)}) + tr(U^{(j)} U^{(j)^T} U^* U^{*^T})),$$
$$s.t.\ U^{(j)^T} U^{(j)} = I,\ 1 \leqslant j \leqslant R,\ U^{*^T} U^* = I. \tag{11}$$

For Eq.(11), we can employ the alternative iteration method to solve it.

(1) Optimizing Eq.(11) with respect to U^*. Fix each $U^{(j)}$, then Eq.(11) becomes:

$$\max_{U^*} \sum_{j=1}^{R} tr(U^{(j)} U^{(j)^T} U^* U^{*^T}),\ s.t.\ U^{*^T} U^* = I. \tag{12}$$

Eq.(12) can be written as:

$$\max_{U^*} tr(U^{*^T} (\sum_{j=1}^{R} U^{(j)} U^{(j)^T}) U^*),\ s.t.\ U^{*^T} U^* = I. \tag{13}$$

It is easy to find that optimizing Eq.(13) is equivalent to solve the standard spectral clustering with a modified Laplacian matrix $\sum_{j=1}^{R} U^{(j)} U^{(j)^T}$, i.e., the solution of U^* is composed by k eigenvectors corresponding to the k largest eigenvalues of $\sum_{j=1}^{R} U^{(j)} U^{(j)^T}$.

(2) Optimizing Eq.(11) with respect to one of the $U^{(j)}$s. Fix the other $U^{(j)}$s and U^*, then Eq.(11) becomes:

$$\max_{U^{(j)}} tr(U^{(j)^T} L^{(j)} U^{(j)}) + tr(U^{(j)} U^{(j)^T} U^* U^{*^T}),$$
$$s.t.\ U^{(j)^T} U^{(j)} = I. \tag{14}$$

Eq.(14) can be written as:

$$\max_{U^{(j)}} tr(U^{(j)^T} (L^{(j)} + U^* U^{*^T}) U^{(j)}),$$
$$s.t.\ U^{(j)^T} U^{(j)} = I. \tag{15}$$

Optimizing Eq.(15) is similar with optimizing Eq.(13), therefore the solution of $U^{(j)}$ is composed by k eigenvectors corresponding to the k largest eigenvalues of $L^{(j)} + U^* U^{*^T}$.

The overall procedure of consistent spectral clustering is shown in Algorithm 1 (Part 2).

5 Experiments

5.1 Datasets

Real benchmark datasets. We conduct experiments on 6 real benchmark datasets. The details of the datasets are shown in Table 1. These datasets are originally established as collections of data with determinate values, we follow the method in [Züfle *et al.*, 2014; Zhang *et al.*, 2017] to generate the uncertainty of Gaussian distribution for these datasets.

Real world uncertain datasets. We also use 3 real world uncertain datasets: Movement (http://archive.ics.uci.edu/ml/), NBA (http://espn.go.com/nba/) and Weather (http://bcc.ncc-cma.net/) to perform experiments.

(1) Movement: it consists of 13197 radio signal records about 314 temporal sequences from a wireless sensor network deployed in real-world office environments. Each record has four dimensions which are respectively corresponding to four sensor nodes. According to user movement path, the dataset is divided into six classes. Each temporal sequence is treated as an uncertain object and each record of the temporal sequence is treated as a possible value of the uncertain object.

(2) NBA: it consists of 2197 records about the top 300 players in ESPN 2015 rank. Each record has five dimensions: points, rebounds, assists, steals and blocks. According to season average performance, they are divided into three classes: star/key/role player. Each player is treated as an uncertain object and each season average performance of the player is treated as a possible value of the uncertain object.

(3) Weather: it consists of 18360 records about 153 weather stations around China. Each station contains the monthly average weather condition from 2006 to 2015. Each record has two dimensions: average temperature and average precipitation. Each station is labeled with a climate type. We have three types of climates: temperate continental climate, temperate monsoon climate and tropical/subtropical monsoon climate. The stations with the same label are considered to be in the same class. Each station is treated as an uncertain object and each monthly average weather condition of the station is treated as a possible value of the uncertain object.

Table 1: Real benchmark datasets.

Dataset	#Objects	#Attributes	#Classes
Wine	178	13	3
Ecoli	327	7	5
Image	2310	19	7
Libras	360	90	15
USPS	929	256	10
Waveform	5000	21	3

5.2 Experimental Setup

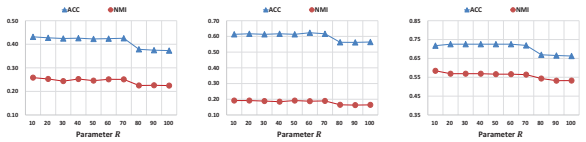
Baselines. We compare RPC with the state-of-the-art clustering algorithms for uncertain data, including UK-means (UKM), CK-means (CKM), UK-medoids (UKMD), MMVar (MMV), UCPC, FDBSCAN (FDB), FOPTICS (FOP), PDBSCAN (PDB), SC and REP. We also compare with the improved versions of SC and REP, which use our proposed selection strategy to select the representative possible worlds and then perform the original SC and REP on the representative possible worlds, and we call them RP-SC and RP-REP.

Settings. For UK-means, CK-means, UK-medoids, MMVar, UCPC and RPC, the sets of initial centroids or partitions are randomly selected. To avoid that the clustering results are affected by random chance, we average the results over 10 different runs. For FDBSCAN, FOPTICS, PDBSCAN, SC, REP, RP-SC and RP-REP, since these algorithms are sensitive to parameters, we adjust the parameters continuously until the performance of each method becomes the best and stable. The methods of determining the parameters can refer to [Kriegel and Pfeifle, 2005a; Kriegel and Pfeifle, 2005b; Zhang *et al.*, 2017; Volk *et al.*, 2009; Züfle *et al.*, 2014].

Evaluation metrics. We adopt two widely used evaluation metrics [Manning *et al.*, 2008]: clustering accuracy (ACC) and normalized mutual information (NMI) to evaluate the clustering performance.

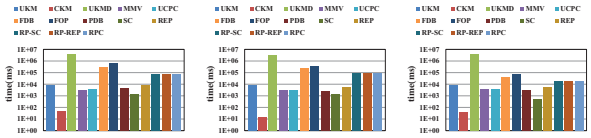
Table 2: Clustering results in terms of effectiveness.

Dataset	Metric	UKM	CKM	UKMD	MMV	UCPC	FDB	FOP	PDB	SC	REP	RP-SC	RP-REP	RPC
Wine	ACC	0.8343	0.8213	0.8163	0.8056	0.8444	0.7247	0.7528	0.7303	0.7079	0.7416	0.7360	0.8034	**0.9663**
	NMI	0.7091	0.6435	0.6880	0.6419	0.6795	0.5562	0.6277	0.6195	0.4817	0.5460	0.5434	0.5823	**0.8782**
Ecoli	ACC	0.6321	0.6300	0.6352	0.6309	0.5765	0.5260	0.6667	0.6575	0.6728	0.6667	0.7034	0.7278	**0.8055**
	NMI	0.6102	0.6362	0.5912	0.5569	0.5588	0.2040	0.5917	0.5536	0.4973	0.5124	0.5858	0.5860	**0.6871**
Image	ACC	0.6639	0.6425	0.6980	0.5945	0.5819	0.5494	0.7177	0.7299	0.5870	0.5636	0.6576	0.6545	**0.8350**
	NMI	0.7115	0.6601	0.6818	0.6070	0.5933	0.6849	0.7464	0.7647	0.6182	0.5871	0.6925	0.6854	**0.7838**
Libras	ACC	0.5322	0.5053	0.5294	0.4211	0.4414	0.2528	0.3417	0.3222	0.2611	0.3167	0.3083	0.3778	**0.6006**
	NMI	0.6583	0.6555	0.6314	0.5490	0.5752	0.4814	0.5742	0.5637	0.4997	0.5752	0.5292	0.6100	**0.7056**
USPS	ACC	0.6220	0.6245	0.6499	0.5107	0.5269	0.4101	0.4769	0.4833	0.4101	0.4456	0.4327	0.4639	**0.7658**
	NMI	0.6797	0.6539	0.6574	0.5338	0.5326	0.4741	0.5622	0.5782	0.4939	0.5386	0.5242	0.5639	**0.8082**
Waveform	ACC	0.8381	0.8382	0.7080	0.6569	0.6542	0.3428	0.3294	0.5938	0.4366	0.4386	0.4956	0.4524	**0.9618**
	NMI	0.6667	0.6697	0.4554	0.4397	0.4046	0.0602	0.0667	0.2975	0.0931	0.0489	0.1061	0.1465	**0.8400**
Movement	ACC	0.3490	0.3341	0.3478	0.3427	0.3494	0.2834	0.2643	0.3121	0.2548	0.2866	0.2866	0.3153	**0.4315**
	NMI	0.2133	0.1935	0.2172	0.1837	0.1985	0.0445	0.0791	0.1170	0.0688	0.0975	0.1350	0.1361	**0.2584**
NBA	ACC	0.5463	0.5457	0.5403	0.5257	0.5473	0.5667	0.5067	0.5867	0.5133	0.5433	0.5667	0.5700	**0.6133**
	NMI	0.1591	0.1648	0.1558	0.1647	0.1690	0.1443	0.0918	0.1759	0.0671	0.1446	0.1563	0.1571	**0.1919**
Weather	ACC	0.5869	0.6144	0.5961	0.6105	0.6033	0.5882	0.5163	0.6993	0.5294	0.5490	0.6340	0.6405	**0.7176**
	NMI	0.4892	0.4690	0.4486	0.4183	0.3747	0.1937	0.4575	0.5277	0.2714	0.2761	0.3133	0.3724	**0.5842**



(a) Movement (b) NBA (c) Weather

Figure 2: The performance of RPC with different R on real world uncertain datasets.



(a) Movement (b) NBA (c) Weather

Figure 3: Clustering results in terms of efficiency.

5.3 Parameter Investigation for RPC

(1) For parameter k, we follow the common practice to set k to the true number of classes in the datasets. (2) For parameter M, the investigation results in previous possible world based methods show that setting $M = 100$ is enough to obtain satisfactory results [Volk *et al.*, 2009; Züfle *et al.*, 2014], so we set $M = 100$. (3) For parameter R, Figure 2 shows the performance of RPC with different R on real world uncertain datasets. From the results, it can be seen that when R is within 10∼70, the clustering performance is always good and stable. When the parameter R is larger than 70, the clustering performance will be affected seriously, which is because that the remaining 30 possible worlds contain many marginal ones. As selecting too many representative possible worlds will result in a waste of time to some extent, in this paper we set $R = 10$ and report the corresponding results.

5.4 Clustering Results

Effectiveness. From the effectiveness results in Table 2, it can be seen that RPC performs the best. RP-SC and RP-REP respectively perform better than SC and REP, but not as well as RPC. This is because that compared with SC and REP, RP-SC and RP-REP select the representative possible worlds, thus avoiding the negative effects caused by marginal possible worlds. However, compared with RPC, RP-SC and RP-REP do not make use of the consistency principle among different possible worlds. UK-means, CK-means, UK-medoids, MMVar and UCPC perform worse than RPC. The reason is that these algorithms reduce complex probability distributions to a single probability distribution or a determinate value, which may cause the loss of uncertain information. FDBSCAN, FOPTICS and PDBSCAN also perform worse than RPC. The reason is that they rely on the unreasonable independent distance assumption. All in all, in terms of effectiveness, RPC performs much better than the compared algorithms.

Efficiency. Due to space limit, we only report the efficiency results (in milliseconds) on real world uncertain datasets. Other datasets have the similar trend. From the results in Figure 3, it can be seen that UK-medoids is the slowest. RPC runs faster than FDBSCAN and FOPTICS, but slower than UK-means, CK-means, MMVar, UCPC and PDBSCAN. Among possible world based algorithms, RP-SC, RP-REP and RPC perform almost identically, and they are slower than SC and REP. The reason is that when selecting representative possible worlds, the computation process of Jensen-Shannon divergence is a little complex.

6 Conclusion

In this paper, we propose a representative possible world based consistent clustering algorithm for uncertain data. By selecting representative possible worlds, it avoids the negative effects caused by marginal possible worlds. By consistent spectral clustering, it makes use of the consistency principle to achieve better performance. Experimental results show that the proposed algorithm outperforms the state-of-the-art algorithms in effectiveness. For future work, we will extend the idea to uncertain data stream clustering and classification.

7 Acknowledgments

This work was supported by National Science Foundation of China (No. 61876028) and the grants 1-ZVJJ and G-YBXV funded by the Hong Kong Polytechnic University.

References

[Aggarwal and Yu, 2009] Charu C. Aggarwal and Philip S. Yu. A survey of uncertain data algorithms and applications. *TKDE*, 21(5):609–623, 2009.

[Ankerst *et al.*, 1999] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander. OPTICS: Ordering points to identify the clustering structure. In *SIGMOD*, pages 49–60, 1999.

[Chau *et al.*, 2006] Michael Chau, Reynold Cheng, Ben Kao, and Jackey Ng. Uncertain data mining: An example in clustering location data. In *PAKDD*, pages 199–204, 2006.

[Dalvi and Suciu, 2007] Nilesh N. Dalvi and Dan Suciu. Management of probabilistic data: Foundations and challenges. In *PODS*, pages 1–12, 2007.

[Dasgupta *et al.*, 2001] Sanjoy Dasgupta, Michael L. Littman, and David A. McAllester. PAC generalization bounds for co-training. In *NIPS*, pages 375–382, 2001.

[Devroye, 1986] Luc Devroye. *Non-uniform Random Variate Generation*. Springer Press, 1986.

[Ester *et al.*, 1996] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, pages 226–231, 1996.

[Gullo and Tagarelli, 2012] Francesco Gullo and Andrea Tagarelli. Uncertain centroid based partitional clustering of uncertain data. In *VLDB*, pages 610–621, 2012.

[Gullo *et al.*, 2008] Francesco Gullo, Giovanni Ponti, and Andrea Tagarelli. Clustering uncertain data via K-medoids. In *SUM*, pages 229–242, 2008.

[Gullo *et al.*, 2010] Francesco Gullo, Giovanni Ponti, and Andrea Tagarelli. Minimizing the variance of cluster mixture models for clustering uncertain objects. In *ICDM*, pages 839–844, 2010.

[Hua and Pei, 2011] Ming Hua and Jian Pei. *Ranking Queries on Uncertain Data*. Advances in Database Systems. Springer Press, 2011.

[Jampani *et al.*, 2008] Ravi Jampani, Fei Xu, Mingxi Wu, Luis Leopoldo Perez, Christopher Jermaine, and Peter J Haas. MCDB: A Monte Carlo approach to managing uncertain data. In *SIGMOD*, pages 687–700, 2008.

[Kao *et al.*, 2008] Ben Kao, Sau Dan Lee, David W. Cheung, Wai-Shing Ho, and K. F. Chan. Clustering uncertain data using Voronoi diagrams. In *ICDM*, pages 333–342, 2008.

[Kao *et al.*, 2010] Ben Kao, Sau Dan Lee, Foris K. F. Lee, David Wai-Lok Cheung, and Wai-Shing Ho. Clustering uncertain data using Voronoi diagrams and R-tree index. *TKDE*, 22(9):1219–1233, 2010.

[Kriegel and Pfeifle, 2005a] Hans-Peter Kriegel and Martin Pfeifle. Density-based clustering of uncertain data. In *KDD*, pages 672–677, 2005.

[Kriegel and Pfeifle, 2005b] Hans-Peter Kriegel and Martin Pfeifle. Hierarchical density-based clustering of uncertain data. In *ICDM*, pages 689–692, 2005.

[Kullback and Leibler, 1951] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[Lee *et al.*, 2007] Sau Dan Lee, Ben Kao, and Reynold Cheng. Reducing UK-means to K-means. In *ICDM Workshops*, pages 483–488, 2007.

[Lin, 1991] Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

[Liu *et al.*, 2017] Xinwang Liu, Miaomiao Li, Lei Wang, Yong Dou, Jianping Yin, and En Zhu. Multiple kernel k-means with incomplete kernels. In *AAAI*, pages 2259–2265, 2017.

[Liu *et al.*, 2018] Han Liu, Xianchao Zhang, and Xiaotong Zhang. Possible world based consistency learning model for clustering and classifying uncertain data. *Neural Networks*, 102:48–66, 2018.

[Ma *et al.*, 2017] Fan Ma, Deyu Meng, Qi Xie, Zina Li, and Xuanyi Dong. Self-paced co-training. In *ICML*, pages 2275–2284, 2017.

[Manning *et al.*, 2008] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[Ngai *et al.*, 2011] Wang Kay Ngai, Ben Kao, Reynold Cheng, Michael Chau, Sau Dan Lee, David W. Cheung, and Kevin Y. Yip. Metric and trigonometric pruning for clustering of uncertain data in 2D geometric space. *Information Systems*, 36(2):476–497, 2011.

[Sarma *et al.*, 2009] Anish Das Sarma, Omar Benjelloun, Alon Y. Halevy, Shubha U. Nabar, and Jennifer Widom. Representing uncertain data: Models, properties, and algorithms. *VLDB Journal*, 18(5):989–1019, 2009.

[Silverman, 1986] Bernard W Silverman. *Density Estimation for Statistics and Data Analysis*. CRC Press, 1986.

[Volk *et al.*, 2009] Peter Benjamin Volk, Frank Rosenthal, Martin Hahmann, Dirk Habich, and Wolfgang Lehner. Clustering uncertain data with possible worlds. In *ICDE*, pages 1625–1632, 2009.

[Wang and Zhou, 2010] Wei Wang and Zhi-Hua Zhou. A new analysis of co-training. In *ICML*, pages 1135–1142, 2010.

[Zhang *et al.*, 2017] Xianchao Zhang, Han Liu, and Xiaotong Zhang. Novel density-based and hierarchical density-based clustering algorithms for uncertain data. *Neural Networks*, 93:240–255, 2017.

[Zhou *et al.*, 2018] Jin Zhou, Long Chen, C. L. Philip Chen, Yingxu Wang, and Han-Xiong Li. Uncertain data clustering in distributed peer-to-peer networks. *TNNLS*, 29(6):2392–2406, 2018.

[Züfle *et al.*, 2014] Andreas Züfle, Tobias Emrich, Klaus Arthur Schmid, Nikos Mamoulis, Arthur Zimek, and Matthias Renz. Representative clustering of uncertain data. In *KDD*, pages 243–252, 2014.