A Framework of Three-Way Cluster Analysis

Hong Yu(✉)

Chongqing Key Laboratory of Computational Intelligence,
Chongqing University of Posts and Telecommunications,
Chongqing 400065, China
yuhong@cqupt.edu.cn

Abstract. A new framework of clustering is proposed inspired by the theory of three-way decisions, which is an alternative formulation different from the ones used in the existing studies. The novel three-way representation intuitively shows which objects are fringe to the cluster and it is proposed for dealing with uncertainty clustering. Instead of using two regions to represent a cluster by a single set, a cluster is represented using three regions through a pair of sets, and there are three regions such as the core region, fringe region and trivial region. A cluster is therefore more realistically characterized by a set of core objects and a set of boundary objects. In this paper, we also illustrate an algorithm for incomplete data by using the proposed evaluation-based three-way cluster model. The preliminary experimental results show that the proposed method is effective for clustering incomplete data which is one kind of uncertainty data. Furthermore, this paper reviews some three-way clustering approaches and discusses some future perspectives and potential research topics based on the three-way cluster analysis.

Keywords: Clustering · Three-way decision theory · Uncertainty · Three-way clustering

1 Introduction

Clustering is a method that uses unsupervised learning and it has been widely applied to many areas such as information retrieval, image analysis, bioinformatics, networks structure analysis and a number of other applications [16]. Often, there is uncertainty in the real world. To take the social networks services as an example, the user's interests are changing and the interest community is also varied. The study of artificial intelligence and cognitive science had observed a well recognized feature of human intelligence, that is, in the cognition and treatment of real world problems, human often observe and analyze the same problem from different levels or different granularity. The process of clustering just reflects the process of making decision in different levels. That is, clustering is a process of deciding whether an object belongs to a cluster or not on a certain granularity level.

L. Polkowski et al. (Eds.): IJCRS 2017, Part II, LNAI 10314, pp. 300–312, 2017.
DOI: 10.1007/978-3-319-60840-2_22

Let us take the objects in Fig. 1 as a universe. For the finest granularity clustering result, each object is taken as a single cluster. For a coarser granularity clustering result, the objects may be clustered in two classes. For the coarsest granularity clustering result, all objects are included in a large cluster. In the process of clustering, if the known information is enough, a certain clustering result corresponding to a granularity will be obtained; if the known information is not sufficient to judge whether an object belongs to a cluster, it needs further information to make decision.

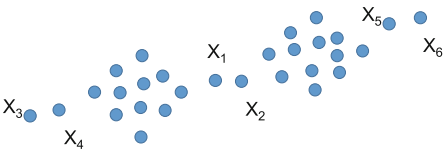

Fig. 1. Schematic diagram of a data set

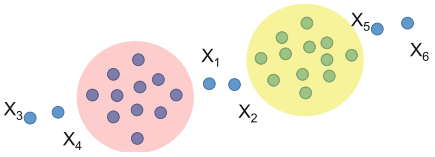

Fig. 2. Schematic diagram of clustering (Color figure online)

Let us observe Fig. 1 again. When we observe the universe in view of a granularity level, we see that there are two distinct clusters, the red one and the yellow one shown in Fig. 2. Then, let us observe $\mathbf{x}_1$ and $\mathbf{x}_2$, they might belong to the red cluster, but it is also possible that they belong to the yellow cluster. One of the solving strategies is that an object "determinately" belongs to different clusters. In view of this strategy, it is often referred to some terminologies such as soft clustering, fuzzy clustering, or an overlapping clustering; in other words, an object can belong to different clusters. We continue to observe $\mathbf{x}_3$ and $\mathbf{x}_4$. It is absolutely reasonable that we assign them into the red cluster. It is the same to $\mathbf{x}_5$ and $\mathbf{x}_6$. The results are shown in Fig. 3 and it is a typical two-way result of overlapping (soft) clustering. Actually, this kind of clustering strategy is a two-way decision result, namely, it decides that an object belongs to a certain cluster or not belongs to this certain cluster. At present, researches are basically based on the two-way decisions. However, the two-way result can not intuitively reveal the fact that $\mathbf{x}_3$ and $\mathbf{x}_4$ are the fringe objects of the red cluster, the same to $\mathbf{x}_5$ and $\mathbf{x}_6$. By contrast, Fig. 4 depicts a three-way clustering result, where $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ and $\mathbf{x}_4$ are assigned into the fringe regions of the red cluster.

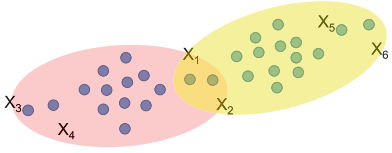

Fig. 3. The two-way clustering result (Color figure online)

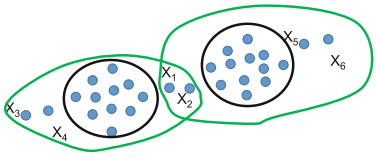

Fig. 4. The three-way clustering result

One usually makes a decision based on available information and evidence. However, the information acquisition is usually a dynamic process. Since the current information is not sufficient, we can produce another solution to the uncertain clustering problem. For those objects which are difficult to make decision at present, we can put forward a two-way decisions result after game playing under the existing knowledge system; we can also produce a three-way decisions result, which makes decisions exactly for these objects which have enough information and waits for new information to make further decisions for those objects whose information is not sufficient. This is a typical idea of three-way decisions.

The three-way decision method represents a concept using three regions instead of two. This three-way decisions scheme has not been considered explicitly in theories of machine learning and rule induction, although it has been studied in other fields. There are three relationships between an object and a cluster: (1) the object certainly belongs to the cluster, (2) the object certainly does not belong to the cluster, and (3) the object might or might not belong to the cluster. It is a typical three-way decision processing to decide the relationship between an object and a cluster. Such relationships will inspire us to introduce the three-way decisions into the cluster analysis problem in this paper.

2 Related Work

A common assumption underlying many cluster analysis methods is that a cluster can be represented by a single set, where the boundary of the cluster is crisp. The crisp boundary leads to easy analytical results but may be too restrictive for some practical applications. Several proposals have been made to reduce such a stringent assumption.

In the fuzzy cluster analysis, it is assumed that a cluster is represented by a fuzzy set that models a gradually changing boundary [6]. However, a fuzzy clustering provides a quantitative characterization of the unclear cluster boundary at the expense of losing the qualitative characterization that better shows the structures provided by a clustering. To resolve this problem, Lingras and his associates [12,13] studied rough clustering and interval set clustering. Yao et al. [20] represented each cluster by an interval set instead of a single set as the representation of a cluster. Chen and Miao [3] described a clustering method by incorporating interval sets in the rough k-means. The basic idea of these work is to derive and describe a cluster by a pair of lower and upper bounds. By describing a cluster in terms of a pair of crisp sets, one recovers the qualitative characterization of a cluster. Most of these algorithms are explained in rough set terminology and an equivalence relation that is needed for defining approximations is not explicitly referred to.

The main objective of this paper is to extend cluster analysis by representing a cluster with two sets. This leads to the introduction of three-way cluster analysis. Furthermore, the strategy of three-way cluster analysis does not require an equivalence relation. Objects in the core region are typical elements of the cluster and objects in the fringe region are fringe elements of the cluster. That

is, a cluster is more realistically characterized by a set of core objects and a set of fringe objects.

The essential ideas of three-way decisions are commonly used in everyday life and widely applied in many fields and disciplines including medical decision-making, social judgement theory, hypothesis testing in statistics, management sciences and peer review process. Therefore, Yao [17,18] introduced and studied the notion of three-way decisions, consisting of the positive, boundary and negative rules. Three-way decisions construct from three regions which are associated with different actions and decisions.

Recently, the three-way decisions approach has been achieved in some areas such as decision making [1,8–11], email spam filtering [31], clustering analysis [21, 22], and so on [2,7,19,26–28,30]. We also proposed some clustering approaches based on the three-way decisions [23–25]. In this paper, we first formalize the representation of a cluster with two sets, then we illustrate a clustering approach for incomplete data based on the proposed framework.

3 Framework of Three-Way Clustering

3.1 Representation of Three-Way Clustering

Let $U = \{\mathbf{x}_1, \cdots, \mathbf{x}_n, \cdots, \mathbf{x}_N\}$ be a finite set, called the universe or the reference set. $\mathbf{x}_n$ is an object which has D attributes, namely, $\mathbf{x}_n = (x_n^1, \cdots, x_n^d, \cdots, x_n^D)$. x_n^d denotes the value of the d-th attribute of the object $\mathbf{x}_n$, where $n \in \{1, \cdots, N\}$, and $d \in \{1, \cdots, D\}$. The result of clustering scheme $\mathbf{C} = \{C^1, \cdots, C^k, \cdots, C^K\}$ is a family of clusters of the universe, in which K means this universe is composed of K clusters.

According to Vladimir Estivill-Castro, the notion of a "cluster" cannot be precisely defined, which is one of the reasons why there are so many clustering algorithms [4]. There is a common denominator: a group of data objects. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).

In the existing works, a cluster is usually represented by a single set, namely, $C^k = \{\mathbf{x}_1^k, \cdots, \mathbf{x}_i^k, \cdots, \mathbf{x}_{|C^k|}^k\}$, abbreviated as C without ambiguous. From the view of making decisions, the representation of a single set means that, the objects in the set belong to this cluster definitely, the objects not in the set do not belong to this cluster definitely. This is a typical result of two-way decisions. For hard clustering, one object just belong to one cluster; for soft clustering, one object might belong to more than one cluster. However, this representation cannot show which objects might belong to this cluster, and it cannot show the degree of the object influence on the form of the cluster intuitively. As discussed before, the use of three regions to represent a cluster is more appropriate than the use of a crisp set, which also directly leads to three-way decisions based interpretation of clustering.

In contrast to the general crisp representation of a cluster, we represent a three-way cluster C as a pair of sets:

$$C = \{Co(C), Fr(C)\}. \tag{1}$$

Here, $Co(C) \subseteq U$ and $Fr(C) \subseteq U$. Let $Tr(C) = U - Co(C) - Fr(C)$. Then, $Co(C)$, $Fr(C)$ and $Tr(C)$ naturally form the three regions of a cluster as Core Region, Fringe Region and Trivial Region respectively. That is:

$$\begin{array}{l} CoreRegion(C) = Co(C), \\ FringeRegion(C) = Fr(C), \\ TrivialRegion(C) = U - Co(C) - Fr(C). \end{array} \tag{2}$$

If $\mathbf{x} \in CoreRegion(C)$, the object $\mathbf{x}$ belongs to the cluster C definitely; if $\mathbf{x} \in FringeRegion(C)$, the object $\mathbf{x}$ might belong to C; if $\mathbf{x} \in TrivialRegion(C)$, the object $\mathbf{x}$ does not belong to C definitely.

These subsets have the following properties.

$$\begin{array}{l} U = Co(C) \cup Fr(C) \cup Tr(C), \\ Co(C) \cap Fr(C) = \emptyset, \\ Fr(C) \cap Tr(C) = \emptyset, \\ Tr(C) \cap Co(C) = \emptyset. \end{array} \tag{3}$$

If $Fr(C) = \emptyset$, the representation of C in Eq. (1) turns into $C = Co(C)$; it is a single set and $Tr(C) = U - Co(C)$. This is a representation of two-way decisions. In other words, the representation of a single set is a special case of the representation of three-way cluster.

Furthermore, according to Formula (3), we know that it is enough to represent expediently a cluster by the core region and the fringe region.

In another way, we can define a cluster scheme by the following properties:

$$\begin{array}{l} (i)\ Co(C^k) \neq \emptyset, 1 \leq k \leq K; \\ (ii)\ \bigcup Co(C^k) \bigcup Fr(C^k) = U, 1 \leq k \leq K. \end{array} \tag{4}$$

Property (i) implies that a cluster cannot be empty. This makes sure that a cluster is physically meaningful. Property (ii) states that any object of U must definitely belong to or might belong to a cluster, which ensures that every object is properly clustered.

With respect to the family of clusters, $\mathbf{C}$, we have the following family of clusters formulated by three-way representation as:

$$\mathbf{C} = \{\{Co(C^1), Fr(C^1)\}, \cdots, \{Co(C^k), Fr(C^k)\}, \cdots, \{Co(C^K), Fr(C^K)\}\}. \tag{5}$$

Obviously, we have the following family of clusters formulated by two-way decisions as:

$$\mathbf{C} = \{Co(C^1), \cdots, Co(C^k), \cdots, Co(C^K)\}. \tag{6}$$

3.2 An Evaluation-Based Three-Way Cluster Model

In this subsection, we will introduce an evaluation-based three-way cluster model, which produces three regions by using an evaluation function and a pair of thresholds on the values of the evaluation function. The model partially addresses the issue of trisecting a universal set into three regions.

Suppose there are a pair of thresholds (α, β) and $\alpha \geq \beta$. Although evaluations based on a total order are restrictive, they have a computational advantage. One can obtain the three regions by simply comparing the evaluation value with a pair of thresholds. Based on the evaluation function $v(\mathbf{x})$, we get the following three-way decision rules:

$$\begin{aligned} Co(C^k) &= \{x \in U | v(\mathbf{x}) > \alpha\}, \\ Fr(C^k) &= \{x \in U | \beta \leq v(\mathbf{x}) \leq \alpha\}, \\ Tr(C^k) &= \{x \in U | v(\mathbf{x}) < \beta\}. \end{aligned} \tag{7}$$

In fact, the evaluation function $v(\mathbf{x})$ can be a risk decision function, a similarity function and so on. In other words, the evaluation function will be specified accordingly when an algorithm is devised. We will give an algorithm as an example in Sect. 4 for clustering incomplete data, since incomplete data is a typical kind of uncertain data.

Objects in $Co(C^k)$ definitely belong to the cluster C^k, objects in $Tr(C^k)$ definitely do not belong to the cluster C^k, and objects in the region $Fr(C^k)$ might or might not belong to the cluster. For the objects in $Fr(C^k) \neq \emptyset$, we need more information to make decisions.

Under the representation, we can formulate the soft clustering and hard clustering as follows. For a clustering, if there exists $k \neq t$, such that

$$\begin{aligned} &(1)\ Co(C^k) \cap Co(C^t) \neq \emptyset,\ or \\ &(2)\ Fr(C^k) \cap Fr(C^t) \neq \emptyset,\ or \\ &(3)\ Co(C^k) \cap Fr(C^t) \neq \emptyset,\ or \\ &(4)\ Fr(C^k) \cap Co(C^t) \neq \emptyset, \end{aligned} \tag{8}$$

we call it is a soft clustering; otherwise, it is a hard clustering.

As long as one condition from Eq. (8) is satisfied, there must exist at least one object belonging to more than one cluster.

4 An Algorithm for Incomplete Data Using the Three-Way Cluster Model

4.1 To Measure Distance Between Incomplete Objects

In this paper, we suppose we have the attribute significance degrees in advance. Of course, it is another interesting research issue, which is not discussed here for sake of space. Thus, we set the descending order of attribute importance degree to be $A = \{a_1, \cdots, a_D\}$, D is the number of attributes. Set $W = \{w_1, w_2, \cdots, w_d, \cdots, w_D\}$ be the set of attribute weights, and $w_1 \geq w_2 \geq \cdots \geq w_k \geq \cdots \geq w_D$.

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. So how to measure the distance or similarity between objects is a key problem in cluster analysis. However, some common methods for computing similarity could not be used to calculate the similarity between incomplete data directly because of the missing values. The partial Euclidean distance formula [5,14,29] is used to measure the distance between the two incomplete data. But the formula only considers non-missing attributes and ignores the impact of missing values on similarity. Besides, Euclidean distance is not conducive to find the spherical structure.

Therefore, we proposed a new similarity measurement between incomplete data by improving the existing partial Euclidean distance formula. The proposed method considers the influence on similarity from the attribute importance as well as the missing rate. Let us consider the following situation, there are two incomplete data in far away distance in fact. The attribute values are similar on non-important attributes but different on important attributes. When the two objects miss a great deal of important attributes, the distance computed by the previous formula will be much less than the actually distance because the result might come from some non-important attributes. The inaccurate distance could seriously affect the effect of the clustering algorithm. In order to avoid this situation, the missing rate and the sum of missing attribute weight are added to the weighted partial Euclidean distance formula. Thus, the improved formula will drastically enlarge the distance when missing lots of important values. Similarly, the improved formula just increases the distance slightly when missing a small account of non-important values. Then, the improved partial Euclidean distance formula is given as follows:

$$Dist(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{\sum_{d=1}^{D} I_d w_d} \frac{\left(\sum_{d=1}^{D}(x_i^d - x_j^d)^2 I_d w_d^2\right)^{1/2}}{\left(\sum_{d=1}^{D}(x_i^d I_d w_d)^2\right)^{1/2} + \left(\sum_{d=1}^{D}(x_j^d I_d w_d)^2\right)^{1/2}} + W_{miss} \times MR, \tag{9}$$

where $I_d = \begin{cases} 1, x_d \neq * \wedge x_j^d \neq * \\ 0, else \end{cases}$, $*$ means the value is missing, and W_{miss} is the sum of attribute weights which are missing on $\mathbf{x}_i$ or $\mathbf{x}_j$, the formula is as follows:

$$W_{miss} = \sum_{x_i^d = * \vee x_j^d = *} w_d. \tag{10}$$

MR is the joint miss rate of object $\mathbf{x}_i$ and $\mathbf{x}_j$. It is the proportion of the number of missing attributes on the total number of attributes as follows: $MR = \frac{\sum_{d=1}^{D} MI_d}{D}$, where $MI_d = \begin{cases} 1, x_i^d = * \vee x_j^d = * \\ 0, else \end{cases}$.

If there is no missing value on the two objects, the proposed formula is the tradition Euclidean distance formula.

4.2 The Algorithm Based on Three-Way Cluster Framework

So far, Formula 9 can be used as the evaluation function in applications. However, we find that the property of clustering is not good enough as required. Thus, we proposed to divide the incomplete data into four types such as sufficient data, valuable data, inadequate data and invalid data according to the concept of complete degree in [23]. In this paper, we continue the sort thought except further working on the measurement of similarity as described in the last subsection.

The proposed three-way clustering algorithm for incomplete data is depicted in Algorithm 1: the three-way clustering algorithm for incomplete data, shorted by TWD-ID. We first divide the data set to four subset, i.e., sufficient data, valuable data, inadequate data and invalid data. Generally speaking, sufficient data have more information. Thus, we find the center of K clusters in the sufficient data set. In fact, there are a bunch of clustering approaches to determine the center. In our experiments, we adopt the outstanding density peaks clustering method in the reference [15]. So, the left work is to decide the left objects where to go. Step 4 describes how to decide the left objects in sufficient data, and Step 5 describes how to decide the other types data.

Algorithm 1. the three-way clustering algorithm for incomplete data

Input: U, $W = \{w_1, w_2, \cdots, w_D\}$, K, α, β, R_{th1};
Output: $\mathbf{C} = \{\{Co(C^1), Fr(C^1)\}, \cdots, \{Co(C^K), Fr(C^K)\}\}$.
Step 1: divide the incomplete data set into four subsets according to the concept of complete degree in [25];
Step 2: compute the distance matrix between objects using Eq. (9);
Step 3: obtain the K center of clusters using the method [17] in the sufficient data subset;
Step 4: compute the local density for each remaining sufficient data point and sort the local densities in descending order; and assign the remaining sufficient data to the core region of the cluster which is its nearest neighbor of highest density;
Step 5: decide the rest of objects to the core region or fringe region of the corresponding cluster according to the three-way decision rules [112].

There could be many missing values in important attributes in the valuable data, inadequate data and invalid data, it is often that the common strategy of filling values may cause new uncertainty. Thus, it is more reasonable that we assign the incomplete data to the fringe regions of clusters waiting more information to help further decision than assign them arbitrarily to the core region or trivial region, when decision information is insufficient or the object just meet the divided condition to the fringe region.

In order to make decisions on these data, we find the neighbors $X_{i-Neighbor}$ within the neighbor radius R_{th} of the object $\mathbf{x}_i$ first, where $X_{i-Neighbor} = \{\mathbf{x}_j | Dist(\mathbf{x}_i, \mathbf{x}_j) \leqslant R_{th}\}$. Then, the object $\mathbf{x}_i$ is assigned to the core region or fringe region of the corresponding clusters according to the proportion of each

cluster in the neighbor objects set $X_{i-Neighbor}$. That is, the proportion is defined as follows:

$$P(X_{i-Neighbor}|C^k) = \frac{\left|\left\{\mathbf{x}_j|\mathbf{x}_j \in X_{i-Neighbor} \wedge \mathbf{x}_j \in C^k\right\}\right|}{|X_{i-Neighbor}|} \tag{11}$$

According to the above formula, the three-way decision rules are given as follows:

$$\begin{array}{l} if\ P(X_{i-Neighbor}|C^k) \geq \alpha,\ the\ object\ is\ decided\ to\ Co(C^k); \\ if\ \beta < P(X_{i-Neighbor}|C^k) < \alpha,\ the\ object\ is\ decided\ to\ Fr(C^k); \\ if\ P(X_{i-Neighbor}|C^k) \leqslant \beta,\ the\ objecis\ decided\ to\ Tr(C^k). \end{array} \tag{12}$$

How to decide the threshold α and β automatically is still an unsolved problem. We can decide the thresholds by experience or through active learning method.

4.3 Experimental Results

In this subsection, we validate the proposed method TWD-ID on three UCI repository [32] data sets with some classical clustering strategies for incomplete data such as WDS-FCM, PDS-FCM, OCS-FCM, NPS-FCM [5] and NNI-FCM [29]. All the experiments are performed on a 3.2 GHz computer with 4 GB memory, and all algorithms are programmed in C++. The quality of the final clustering is evaluated by the traditional indices such as the Accuracy and F-measure, where the objects in fringe regions are deemed to be core regions to fit these common formulae.

In order to reflect the effect of the missing rate on the performance of algorithms, the incomplete data set is constructed randomly according to the 10%, 15% and 20% missing rate. To avoid the effect by the distribution of missing data, we test 10 times by generating different incomplete data sets randomly for each UCI data set. The mean and standard deviation of the results for 10 times under each missing rate are recorded in the following tables, where $\alpha = 0.7$ and $\beta = 0.45$ (Tables 1, 2 and 3).

Table 1. Experimental results on the iris data set

Algorithm	Miss rate					
	10%		15%		20%	
	Accuracy	F-measure	Accuracy	F-measure	Accuracy	F-measure
TWD-ID	$\mathbf{0.914 \pm 0.031}$	$\mathbf{0.913 \pm 0.035}$	$\mathbf{0.917 \pm 0.038}$	$\mathbf{0.915 \pm 0.041}$	$\mathbf{0.893 \pm 0.019}$	$\mathbf{0.888 \pm 0.020}$
WDS-FCM	0.583 ± 0.020	0.586 ± 0.022	0.468 ± 0.036	0.476 ± 0.036	0.464 ± 0.069	0.447 ± 0.092
PDS-FCM	0.898 ± 0.006	0.897 ± 0.006	0.892 ± 0.012	0.891 ± 0.012	0.889 ± 0.008	0.888 ± 0.007
OCS-FCM	0.883 ± 0.015	0.882 ± 0.014	0.858 ± 0.073	0.846 ± 0.108	0.867 ± 0.026	0.866 ± 0.027
NPS-FCM	0.869 ± 0.020	0.868 ± 0.021	0.845 ± 0.024	0.844 ± 0.024	0.807 ± 0.067	0.800 ± 0.076
NNI-FCM	0.900 ± 0.014	0.899 ± 0.014	0.889 ± 0.020	0.889 ± 0.019	0.811 ± 0.073	0.802 ± 0.083

Table 2. Experimental results on the page blocks data set

Algorithm	Miss rate					
	10%		15%		20%	
	Accuracy	F-measure	Accuracy	F-measure	Accuracy	F-measure
TWD-ID	**0.825 ± 0.064**	**0.827 ± 0.033**	**0.802 ± 0.063**	**0.811 ± 0.037**	**0.810 ± 0.069**	**0.810 ± 0.069**
WDS-FCM	0.607 ± 0.006	0.688 ± 0.004	0.710 ± 0.051	0.755 ± 0.032	0.772 ± 0.061	0.790 ± 0.036
PDS-FCM	0.690 ± 0.013	0.743 ± 0.007	0.689 ± 0.006	0.743 ± 0.004	0.691 ± 0.014	0.744 ± 0.008
OCS-FCM	0.652 ± 0.020	0.720 ± 0.013	0.613 ± 0.017	0.693 ± 0.012	0.583 ± 0.023	0.671 ± 0.018
NPS-FCM	0.668 ± 0.028	0.729 ± 0.017	0.656 ± 0.045	0.721 ± 0.029	0.648 ± 0.048	0.716 ± 0.031
NNI-FCM	0.717 ± 0.005	0.758 ± 0.003	0.697 ± 0.033	0.746 ± 0.021	0.692 ± 0.024	0.743 ± 0.014

Table 3. Experimental results on the pendigits data set

Algorithm	Miss rate					
	10%		15%		20%	
	Accuracy	F-measure	Accuracy	F-measure	Accuracy	F-measure
TWD-ID	**0.753 ± 0.035**	**0.731 ± 0.043**	**0.746 ± 0.033**	**0.727 ± 0.041**	**0.737 ± 0.037**	**0.717 ± 0.048**
WDS-FCM	0.331 ± 0.023	0.280 ± 0.024	0.323 ± 0.021	0.242 ± 0.023	0.319 ± 0.030	0.242 ± 0.026
PDS-FCM	0.663 ± 0.031	0.623 ± 0.033	0.689 ± 0.025	0.660 ± 0.036	0.676 ± 0.028	0.641 ± 0.041
OCS-FCM	0.539 ± 0.081	0.465 ± 0.099	0.464 ± 0.050	0.385 ± 0.061	0.369 ± 0.050	0.268 ± 0.055
NPS-FCM	0.630 ± 0.024	0.575 ± 0.028	0.581 ± 0.034	0.518 ± 0.044	0.530 ± 0.056	0.465 ± 0.066
NNI-FCM	0.489 ± 0.046	0.412 ± 0.051	0.481 ± 0.043	0.408 ± 0.051	0.421 ± 0.050	0.324 ± 0.064

The experiment results show that the proposed method is appropriate for clustering uncertainty data such as incomplete data. Besides, the accuracy and F-measure of the proposed algorithm are higher than the compared algorithms in the experiments.

5 Discussions

This paper aims at presenting an interpretation of three-way clustering for uncertainty clustering. The existing work usually represents a cluster with a single set and it is a typical result of two-way decisions. That is, objects in the set belong to the cluster definitely and objects not in the set do not belong to the cluster definitely. There are two regions to describe a cluster. In the proposed framework, we use three regions to represent a cluster inspired by the theory of three-way decisions. Objects in the core region belong to the cluster definitely, objects in the trivial region do not belong to the cluster definitely and objects in fringe region are the boundary elements of the cluster. The representation not only shows which objects just belong to this cluster but also shows which objects might belong to the cluster intuitively.

Through the further work on the fringe region, we can know the degree of an object influences on the form of the cluster intuitively, which is very helpful in some practical applications. Furthermore, an evaluation-based three-way cluster model and an algorithm for clustering incomplete data based on the proposed model are introduced.

In the following paper, I will summarize and conclude the paper with listing some important issues and research trends about the three-way clustering.

- Representation of three-way clustering. There are some work had been proposed in view of interval sets, decision-theoretic rough sets [22]. We can also represent the model of three-way clustering by using fuzzy set, shadow set and other models. Different interpretations of three-way clustering could give different solutions to different kinds of clustering problems.
- How to get the three-way clustering. It is a good way to extend from the classical two-way decision clustering approaches. The following properties are important to the efficiency and effectiveness of a novel algorithm: how to decide the thresholds, how to know the truth number of clusters.
- Developing new clustering approaches for more uncertainty situations such as dynamic, incomplete data or multi-source data. For example, we had done some preliminary work [23,25].
- Application of three regions. We can put forward the three-way clustering strategy to the application fields such as social network services, cyber marketing, E-commerce, recommendation service and other fields.

Acknowledgments. I am grateful to Professor Yiyu Yao for his suggestions, and I would like to thank Ms. Ting Su for her help to complete the experimental work. In addition, this work was supported in part by the National Natural Science Foundation of China under grant No. 61379114 and No. 61533020.

References

1. Azam, N., Yao, J.T.: Analyzing uncertainties of probabilistic rough set regions with game-theoretic rough sets. Int. J. Approx. Reason. **55**(1), 142–155 (2014)
2. Chen, H.M., Li, T.R., Luo, C., Horng, S., Wang, G.Y.: A decision-theoretic rough set approach for dynamic data mining. IEEE Trans. Fuzzy Syst. **99**(1) (2015). doi:10.1109/TFUZZ.2014.2387877
3. Chen, M., Miao, D.Q.: Interval set clustering. Expert Syst. Appl. **38**(4), 2923–2932 (2011)
4. Estivill-Castro, V.: Why so many clustering algorithms: a position paper. ACM SIGKDD Explor. Newsl. **4**(1), 65–75 (2002)
5. Hathaway, R.J., Bezdek, J.C.: Fuzzy C-means clustering of incomplete data. IEEE Trans. Syst. Man Cybern. Part B (Cybern.) **31**(5), 735–744 (2001)
6. Höppner, F., Klawonn, F., Kruse, R., Runkler, T.: Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition. Wiley, Chichester (1999)
7. Li, H.X., Zhou, X.Z.: Risk decision making based on decision-theoretic rough set: a three-way view decision model. Int. J. Comput. Intell. Syst. **4**(1), 1–11 (2011)
8. Li, Y., Zhang, Z., Chen, W.B., Min, F.: TDUP: an approach to incremental mining of frequent itemsets with three-way-decision pattern updating. Int. J. Mach. Learn. Cybern. **8**(1), 441–453 (2015)
9. Liang, D.C., Xu, Z.S., Liu, D.: Three-way decisions with intuitionistic fuzzy decision-theoretic rough sets based on point operators. Inf. Sci. **375**, 183–201 (2017)
10. Liang, D.C., Liu, D.: A novel risk decision-making based on decision-theoretic rough sets under hesitant fuzzy information. J. IEEE Trans. Fuzzy Syst. **23**(2), 237–247 (2015)

11. Liu, D., Liang, D.C., Wang, C.C.: A novel three-way decision model based on incomplete information system. Knowl. Based Syst. **91**, 32–45 (2016)
12. Lingras, P., Yan, R.: Interval clustering using fuzzy and rough set theory. In: Proceedings of the 2004 IEEE Annual Meeting of the Fuzzy Information, Banff, Alberta, pp. 780–784 (2004)
13. Lingras, P., West, C.: Interval set clustering of web users with rough K-means. J. Intell. Inf. Syst. **23**(1), 5–16 (2004)
14. Lu, C., Song, S., Wu, C.: K-nearest neighbor intervals based AP clustering algorithm for large incomplete data. Math. Probl. Eng. **2015** (2015). http://dx.doi.org/10.1155/2015/535932
15. Rodriguez, A., Laio, A.: Clustering by fast search and find of density peaks. Science **344**(6191), 1492–1496 (2014)
16. Xu, R., Wunsch, D.: Survey of clustering algorithms. IEEE Trans. Neural Netw. **16**(3), 645–678 (2005)
17. Yao, Y.: An outline of a theory of three-way decisions. In: Yao, J.T., Yang, Y., Słowiński, R., Greco, S., Li, H., Mitra, S., Polkowski, L. (eds.) RSCTC 2012. LNCS, vol. 7413, pp. 1–17. Springer, Heidelberg (2012). doi:10.1007/978-3-642-32115-3_1
18. Yao, Y.: Three-way decisions and cognitive computing. Cogn. Comput. (2016). doi:10.1007/s12559-016-9397-5
19. Yao, Y.: Interval sets and three-way concept analysis in incomplete contexts. Int. J. Mach. Learn. Cybern. **8**(1), 3–20 (2017)
20. Yao, Y.Y., Lingras, P., Wang, R.Z., Miao, D.Q.: Interval set cluster analysis: a re-formulation. Rough Sets. Fuzzy Sets, Data Mining and Granular Computing, pp. 398–405. Springer, Berlin Heidelberg (2009). doi:10.1007/978-3-642-10646-0_48
21. Yu, H., Jiao, P., Yao, Y.Y., Wang, G.Y.: Detecting and refining overlapping regions in complex networks with three-way decisions. Inf. Sci. **373**, 21–41 (2016)
22. Yu, H., Liu, Z.G., Wang, G.Y.: An automatic method to determine the number of clusters using decision-theoretic rough set. Int. J. Approx. Reason. **55**, 101–115 (2014)
23. Yu, H., Su, T., Zeng, X.: A three-way decisions clustering algorithm for incomplete data. In: Miao, D., Pedrycz, W., Ślęzak, D., Peters, G., Hu, Q., Wang, R. (eds.) RSKT 2014. LNCS, vol. 8818, pp. 765–776. Springer, Cham (2014). doi:10.1007/978-3-319-11740-9_70
24. Yu, H., Wang, Y.: Three-way decisions method for overlapping clustering. In: Yao, J.T., Yang, Y., Słowiński, R., Greco, S., Li, H., Mitra, S., Polkowski, L. (eds.) RSCTC 2012. LNCS, vol. 7413, pp. 277–286. Springer, Heidelberg (2012). doi:10.1007/978-3-642-32115-3_33
25. Yu, H., Zhang, C., Wang, G.Y.: A tree-based incremental overlapping clustering method using the three-way decision theory. Knowl.-Based Syst. **91**, 189–203 (2016)
26. Yu, H., Wang, G.Y., Li, T.R., Liang, J.Y., Miao, D.Q., Yao, Y.Y.: Three-way Decisions: Methods and Practices for Complex Problem Solving. Science Press, Beijing (2015). (in Chinese)
27. Zhang, H.R., Min, F., Shi, B.: Regression-based three-way recommendation. Inf. Sci. **378**, 444–461 (2017)
28. Zhang, Y., Zou, H., Chen, X., Wang, X., Tang, X., Zhao, S.: Cost-sensitive three-way decisions model based on CCA. In: Cornelis, C., Kryszkiewicz, M., Ślęzak, D., Ruiz, E.M., Bello, R., Shang, L. (eds.) RSCTC 2014. LNCS, vol. 8536, pp. 172–180. Springer, Cham (2014). doi:10.1007/978-3-319-08644-6_18

29. Zhang, L., Li, B., Zhang, L., Li, D.: Fuzzy clustering of incomplete data based on missing attribute interval size. In: 2015 IEEE 9th International Conference on Anticounterfeiting, Security, and Identification (ASID), pp. 101–104. IEEE (2015)
30. Zhang, Y., Yao, J.T.: Gini objective functions for three-way classifications. Int. J. Approx. Reason. **81**, 103–114 (2017)
31. Zhou, B., Yao, Y., Luo, J.G.: Cost-sensitive three-way email spam filtering. J. Intell. Inf. Syst. **42**, 19–45 (2013)
32. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml