Three-Way Decisions Method for Overlapping Clustering

Hong Yu and Ying Wang

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

Abstract. Most of clustering methods assume that each object must be assigned to exactly one cluster, however, overlapping clustering is more appropriate than crisp clustering in a variety of important applications such as the network structure analysis and biological information. This paper provides a three-way decision strategy for overlapping clustering based on the decision-theoretic rough set model. Here, each cluster is described by an interval set that is defined by a pair of sets called the lower and upper bounds. Besides, a density-based clustering algorithm is proposed using the new strategy, and the results of the experiments show the strategy is effective to overlapping clustering.

Keywords: overlapping clustering, three-way decision, decision-theoretic rough set theory, data mining.

1 Introduction

In recent years, clustering has been widely used as a powerful tool to reveal underlying patterns in many areas such as data mining, web mining, geographical data processing, medicine and so on. Most of clustering methods assume that each object must be assigned to exactly one cluster. However, in a variety of important applications such as network structure analysis, wireless sensor networks and biological information, overlapping clustering is more appropriate[3].

Many researchers have proposed some overlapping clustering methods for different application background. For example, Takaki and Tamura et al. [10] propose a method of overlapping clustering for network structure analysis, Aydin and Naït-Abdesselam et al. [1] propose an overlapping clusters algorithm used in the mobile Ad hoc networks. Lingras and Bhalchandra et al. [5] compare crisp and fuzzy clustering in the mobile phone call dataset. Obadi and Dráždilová et al. [8] propose an overlapping clustering method for DBLP datasets based on rough set theory.

The rough set theory [9] approximates a concept by three regions, namely, the positive, boundary and negative regions, which immediately leads to the notion of three-way decision clustering approach. Three-way decisions constructed from

J.T. Yao et al. (Eds.): RSCTC 2012, LNAI 7413, pp. 277–286, 2012.

the three regions are associated with different actions and decisions. In fact, the three-way decision approach has been achieved in some areas as the email spam filtering [15], three-way investment decisions [6], and so on [4] [12].

To combat the overlapping clustering, this paper proposes a new three-way decision clustering strategy based on the decision-theoretic rough set model [13]. Yao and Lingras et al. [14] had represented each cluster by an interval set instead of a single set as the representation of a cluster. Chen and Miao [2] study the clustering method represented as interval sets, wherein the rough k-means clustering method is combined. Inspired by the representation, the cluster in our strategy is also represented by an interval set, which is defined by a pair of sets called the lower and upper bounds. Objects in the lower bound are typical elements of the cluster and objects between the upper and lower bounds are fringe elements of the cluster.

Furthermore, the solutions to obtain the lower and upper bounds are formulated based on the three-way decisions in this paper. Then, a density-based clustering algorithm is proposed, and we demonstrate the effectiveness of the algorithm through experiments.

2 Formulation of Clustering

2.1 Decision-Theoretic Rough Set Model

The decision-theoretic rough set model [13], DTRS shorted, applies the Bayesian decision procedure for the construction of probabilistic approximations.

Let $\Omega = \{A, A^c\}$ denote the set of states indicating that an object is in A and not in A, respectively. Let $Action = \{a_P, a_N, a_B\}$ be the set of actions, where a_P, a_N, and a_B represent the three actions in classifying an object, deciding $POS(A)$, deciding $NEG(A)$ and deciding $BND(A)$, respectively. Let $i = P, N, B$, and $\lambda_{iP}(a_i|A)$ and $\lambda_{iN}(a_i|A^c)$ denote the loss (cost) for taking the action a_i when the state is A, A^c, respectively. For an object with description $[x]$, suppose an action a_i is taken. The expected loss $R(a_i|[x])$ associated with taking the individual actions can be expressed as:

$$\begin{aligned} R(a_P|[x]) &= \lambda_{PP}P(A|[x]) + \lambda_{PN}P(A^c|[x]), \\ R(a_N|[x]) &= \lambda_{NP}P(A|[x]) + \lambda_{NN}P(A^c|[x]), \\ R(a_B|[x]) &= \lambda_{BP}P(A|[x]) + \lambda_{BN}P(A^c|[x]). \end{aligned}$$

where the probabilities $P(A|[x])$ and $P(A^c|[x])$ are the probabilities that an object in the equivalence class $[x]$ belongs to A and A^c, respectively.

2.2 Extend DTRS for Clustering

To define our framework, we will assume $\mathbf{C} = \{C_1, \cdots, C_k, \cdots, C_K\}$, where $C_k \subseteq U$, is a family of clusters of a universe $U = \{x_1, \cdots, x_n\}$.

In order to interpret clustering, let's extend the DTRS model firstly. The set of states is given by $\Omega = \{C, \neg C\}$, the two complement states indicate that an

object is in a cluster C and not in a cluster C, respectively. The set of action is given by $A = \{a_P, a_B, a_N\}$, where a_P, a_B and a_N represent the three actions in classifying an object, a_P represents that we will take the description of an object x into the domain of the cluster C; a_B represents that we will take the description of an object x into the boundary domain of the cluster C; a_N represents that we will take the description of an object x into the negative domain of the C.

Let λ_{PP} ,λ_{BP}, λ_{NP}, λ_{PN}, λ_{BN}, λ_{NN} denote the loss (cost) for taking the action a_P, a_B and a_N when the state is C, $\neg C$, respectively. For an object x with description $[x]$, suppose an action a_i is taken. According to Subsection 2.1, the expected loss associated with taking the actions can be expressed as:

$$\begin{array}{l} Risk(a_P|[x]) = \lambda_{PP}Pr(C|[x]) + \lambda_{PN}Pr(\neg C|[x]); \\ Risk(a_B|[x]) = \lambda_{BP}Pr(C|[x]) + \lambda_{BN}Pr(\neg C|[x]); \\ Risk(a_N|[x]) = \lambda_{NP}Pr(C|[x]) + \lambda_{NN}Pr(\neg C|[x]). \end{array} \tag{1}$$

Where $Pr(C|[x])$ represents the probability that an object x in the description $[x]$ belongs to the cluster C, and $Pr(C|[x]) + Pr(\neg C|[x]) = 1$. The Bayesian decision procedure leads to the following minimum-risk decision:

$$\begin{array}{l} (P) If\ Risk(a_P|[x]) \leq Risk(a_N|[x])\ and\ Risk(a_P|[x]) \leq Risk(a_B|[x]), \\ decide\ POS(C); \\ (B) If\ Risk(a_B|[x]) < Risk(a_P|[x])\ and\ Risk(a_B|[x]) < Risk(a_N|[x]), \\ decide\ BND(C); \\ (N) If\ Risk(a_N|[x]) \leq Risk(a_P|[x])\ and\ Risk(a_N|[x]) \leq Risk(a_B|[x], \\ decide\ NEG(C); \end{array} \tag{2}$$

Consider a special kind of loss functions with $\lambda_{PP} \leq \lambda_{BP} < \lambda_{NP}$ and $\lambda_{NN} \leq \lambda_{BN} < \lambda_{PN}$. That is, the loss of classifying an object x belonging to C into the positive region $POS(C)$ is less than or equal to the loss of classifying x into the boundary region $BND(C)$, and both of these losses are strictly less than the loss of classifying x into the negative region $NEG(C)$. The reverse order of losses is used for classifying an object x that does not belong to C, namely the object x is a negative instance of C. For this type of loss function, the above minimum-risk decision rules can be written as:

$$\begin{array}{l} (P) If\ Pr(C|[x]) \geq \alpha\ and\ Pr(C|[x]) \geq \gamma, decide\ POS(C); \\ (B) If\ Pr(C|[x]) < \alpha\ and\ Pr(C|[x]) > \beta, decide\ BND(C); \\ (N) If\ Pr(C|[x]) \leq \beta\ and\ Pr(C|[x]) \leq \gamma, decide\ NEG(C); \end{array} \tag{3}$$

Where:

$$\begin{array}{l} \alpha = \frac{(\lambda_{PN} - \lambda_{BN})}{(\lambda_{PN} - \lambda_{BN}) + (\lambda_{BP} - \lambda_{PP})} = (1 + \frac{(\lambda_{BP} - \lambda_{PP})}{(\lambda_{PN} - \lambda_{BN})})^{-1} \\ \gamma = \frac{(\lambda_{PN} - \lambda_{NN})}{(\lambda_{PN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{PP})} = (1 + \frac{(\lambda_{NP} - \lambda_{PP})}{(\lambda_{PN} - \lambda_{NN})})^{-1} \\ \beta = \frac{(\lambda_{BN} - \lambda_{NN})}{(\lambda_{BN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{BP})} = (1 + \frac{(\lambda_{NP} - \lambda_{BP})}{(\lambda_{BN} - \lambda_{NN})})^{-1} \end{array} \tag{4}$$

In this paper, we consider that the cluster have the boundary, so we just discuss the relationship between thresholds α and β as $\alpha > \beta$. According to Eq.(4), it follows that $\alpha > \gamma > \beta$. After tie-breaking, the following simplified rules (P)-(N) are obtained:

$$\begin{array}{l}(P)\ If\ Pr(C|[x]) \geq \alpha, decide\ POS(C);\\ (B)\ If\ \beta < Pr(C|[x])) < \alpha, decide\ BND(C);\\ (N)\ If\ Pr(C|[x]) \leq \beta, decide\ NEG(C).\end{array} \tag{5}$$

Obviously, rules (P)-(N) give a three-way decision method for clustering. That is, an object belongs to a cluster definitely if it is in $POS(C)$ based on the available information; an object may be a fringe member if it is in $BND(C)$, we can decide whether it is in a cluster through further information. Clustering algorithms can be devised according to the rules (P)-(N).

On the other hand, according to the rough set theory [9] and the rules (P)-(N), for a subset $C \subseteq U$, we can define its lower and upper approximations as follows.

$$\begin{array}{l}\underline{apr}(C) = POS(C) = \{x|Pr(C|[x]) \geq \alpha\};\\ \overline{apr}(C) = POS(C) \cup BND(C) = \{x|Pr(C|[x]) > \beta\}.\end{array} \tag{6}$$

2.3 Re-formulation of Clustering Using Interval Set

Yao and Lingras et al.[14] had formulated the clustering using the form of interval sets. It is naturally that the region between the lower and upper bound of an interval set means the overlapping region.

Assume $\mathbf{C} = \{C_1, \cdots, C_k, \cdots, C_K\}$ is a family of clusters of a universe $U = \{x_1, \cdots, x_n\}$. Formally, we can define a clustering by the properties:

$$(i)\ C_k \neq \emptyset, 0 \leq k \leq K; \quad (ii)\ \bigcup_{C_k \in \mathbf{C}} C_k = U.$$

Property (i) requires that each cluster cannot be empty. Property (ii) states that every $x \in U$ belongs to at least one cluster. Furthermore, if $C_i \cap C_j = \emptyset, i \neq j$, it is a crisp clustering, otherwise it is an overlapping clustering.

As we have discussed, we may use an interval set to represent the cluster in $\mathbf{C}$, namely, C_k is represented by an interval set $[C_k^l, C_k^u]$. Combine the conclusion in the above subsection, we can represent the lower and upper bound of the interval set as the lower and upper approximate, that is, C_k is represented by an interval set $[\underline{apr}(C_k), \overline{apr}(C_k)]$.

Any set in the family $[\underline{apr}(C_k), \overline{apr}(C_k)] = \{X|\underline{apr}(C_k) \subseteq X \subseteq \overline{apr}(C_k)\}$ may be the actual cluster C_k. The objects in $\underline{apr}(C_k)$ may represent typical objects of the cluster C_k, objects in $\overline{apr}(C_k) - \underline{apr}(C_k)$ may represent fringe objects, and objects in $U - \overline{apr}(C_k)$ may represent the negative objects. With respect to the family of clusters $\mathbf{C} = \{C_1, \cdots, C_k, \cdots, C_K\}$, we have the following family of interval set clusters:

$$\mathbf{C} = [\underline{apr}(C_1), \overline{apr}(C_1)], \ldots, [\underline{apr}(C_k), \overline{apr}(C_k)], \ldots, [\underline{apr}(C_K), \overline{apr}(C_K)].$$

Corresponding to Property (i) and (ii), we adopt the following properties for a clustering in the form of interval set:

$$(i)\ \underline{apr}(C_k) \neq \emptyset, 0 \leq k \leq K; \quad (ii)\ \bigcup \overline{apr}(C_k) = U.$$

Property (i) requires that the lower approximate must not be empty. It implies that the upper approximate is not empty. It is reasonable to assume that each cluster must contain at least one typical object and hence its lower bound is not empty. In order to make sure that a clustering is physically meaningful, Property (ii) states that any object of U belongs to the upper approximate of a cluster, which ensures that every object is properly clustered.

According to Eq.(6), the family of clusters $\mathbf{C}$ give a three-way decision clustering. Namely, objects in $\underline{apr}(C_k)$ are decided definitely to belong to the cluster C_k, objects in $U - \overline{apr}(C_k)$ can be decided not to belong to the cluster C_k. Set $BND(C_k) = \overline{apr}(C_k) - \underline{apr}(C_k)$. Objects in the region $BND(C_k)$ may be belong to the cluster or not.

There exists $k \neq t$, it is possible that $\underline{apr}(C_k) \cap \underline{apr}(C_t) \neq \emptyset$, or $BND(C_k) \cap BND(C_t) \neq \emptyset$. In other words, it is possible that an object belongs to more than one cluster.

3 Clustering Algorithm Using Three-Way Decision

Density-based clustering analysis is one kind of clustering analysis methods that can discover clusters with arbitrary shape and is insensitive to noise data. Therefore, according to the three-way decision rules (P)-(N) in Subsection 2.2, a density-based clustering algorithm will be proposed in this section to combat the overlapping clustering.

Considering the discovery area, set the center is p and Rth is the radius, the number of points in the area is called the density of p relative to Rth, denoted by $Density(p, Rth)$. The concepts are defined as follows [7].

Reference points: For any node p, distance Rth and threshold mth in the space, if $Density(p, Rth) \leq mth$, then p is a reference point and mth is the density threshold value.

The reference points are fictional points, not the points in the dataset. Threshold value mth represents a reference number. When the density of p greater than mth, p is an intensive point, otherwise it is a sparse point.

Representing Region: Every reference point p is the representative of a circular area where the point is the center of the area and the radius is Rth, and the region is the representing region of the reference point p.

All points(objects) in the representing region of a reference point p are seen as an equivalence class. In order to cluster objects(points) in the space, we need to give the method to calculate the probability in Eq.(6).

Probability: $[x]$ is a description of an object x, the $Pr(C|[x])$ is:

$$Pr(C|[x]) = \frac{|C \cap [x]|}{|[x]|}. \tag{7}$$

General speaking, the equivalence class $[x]$ of an object x can be used as a description of the object. That is, Eq.(7) gives a computing method for the

Algorithm 1. Density-based Clustering Algorithm Using Three-way Decision

Input : a universe $U = \{x_1, \cdots, x_n\}$.
Output: the clustering result $\mathbf{C}$.
begin
Step 1. Initial: Set $UN = \emptyset$, $RF = \emptyset$, the possibility $Pr(C_k|RF_t) = 0$.
Step 2. Find all candidate reference points:
$RF_1 \leftarrow x_1$; $\mathbf{RF} = \mathbf{RF} \cup \{RF_1\}$;
for *every* x_i **do**
$temp = \min\limits_{RF_t} |RF_t - x_i|$; $k = arg(\min\limits_{RF_t} |RF_t - x_i|)$;
If $temp > Rth$ then { $RF_{T+1} \leftarrow x_i$; $\mathbf{RF} = \mathbf{RF} \cup RF_{T+1}$;}
Else alter RF_k based on object x_i;
end
Step 3. Choice the reference points and the noise points from the candidates:
for *every* $RF_t \in \boldsymbol{RF}$ **do**
For (every x_i) do { If $|RF_t - x_i| < Rth$ then $RF_t = RF_t \cup x_i$; }
If $|RF_t| < mth$ then { $UN = UN \cup RF_t$; $RF = RF - RF_t$; }
end
Step 4. Clustering the reference points according to three-way rules (P)-(N):
for *every* $\underline{apr}(C_k)$ **do**
$\underline{apr}(C_k) = RF_k$; $\overline{apr}(C_k) = RF_k$;
for *every* RF_t **do**
$Pr(C_k|RF_t) = \frac{|C_k \cap RF_t|}{|RF_t|}$ // Eq.(7)
If $Pr(C_k|RF_t) \geq \alpha$ then
$\underline{apr}(C_k) = \underline{apr}(C_k) \cup RF_t$; $\overline{apr}(C_k) = \overline{apr}(C_k) \cup RF_t$;
If $\beta < Pr(\overline{C_k}|RF_t) < \alpha$ then $\overline{apr}(C_k) = \overline{apr}(C_k) \cup RF_t$,
end
end
$\mathbf{C} = [\underline{apr}(C_1), \overline{apr}(C_1)], \ldots, [\underline{apr}(C_k), \overline{apr}(C_k)], \ldots, [\underline{apr}(C_K), \overline{apr}(C_K)]$.
for *every* $\underline{apr}(C_k)$ **do**
If $\underline{apr}(\overline{C_k}) \supseteq \underline{apr}(C_j)$ then $\overline{apr}(C_k) = \overline{apr}(C_k) \cup \overline{apr}(C_j)$; $\mathbf{C} = \mathbf{C} - C_j$;
end
Step 5. Clustering the noise points.
for *every* $\underline{apr}(C_k)$ **do**
for *every* UN_s **do**
If $UN_s \subseteq \underline{apr}(C_k)$ then
$\underline{apr}(C_k) = \underline{apr}(C_k) \cup UN_s$; $\overline{apr}(C_k) = \overline{apr}(C_k) \cup UN_s$;
Else $\{\mathbf{C} = \overline{\mathbf{C}} \cup UN_S;\}$
end
end
end

probability, then we can devise the algorithm based on three-way decision. In other words, the different algorithms can be developed based on the different approaches of computing probability.

In this paper, a density-based clustering algorithm using three-way decision is proposed as follows. Here, UN and $\mathbf{RF} = \{RF_1, \cdots, RF_t, \cdots, RF_T\}$ means the noise data set and the family of reference points sets, respectively.

In the above algorithm, Step 1 to Step 3 obtain an initial clustering result by choosing the reference points and representing regions according to the relative concepts. Step 4 modifies the clusters according to three-way decision rules (P) to (N).

4 Experiments

The new algorithm is performed by Visual C++. Firstly, some UCI datasets [11] are used to test the different thresholds such as the distance threshold value Rth, density threshold mth, α and β. Obviously, Rth and mth are decided by the characteristics of the dataset. However, there is an interesting result that the clustering result seems good when $\alpha = 0.8$ and $\beta = 0.4$ in most cases. Thus, the result is accepted in the later experiments. On the other hand, it enlightens us we should think a formal way to define the α and β in the further work.

4.1 Synthetic Data Set

The synthetic data set is tested to illustrate the ideas presented in the previous section. The two dimensions data set is depicted in Fig.1, which have 374 points, and Fig.2 gives the clustering result. Here, the thresholds are $Rth = 1.75$, $mth = 10$, $\alpha = 0.8$ and $\beta = 0.4$.

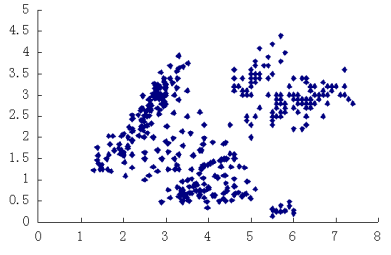

Fig. 1. A synthetic data set

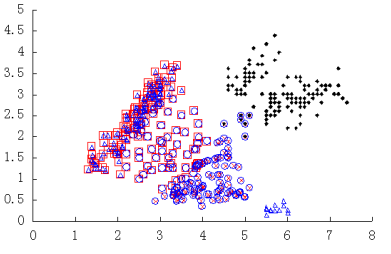

Fig. 2. The clustering result of the data set

From Fig.2, we can see that these points are clustered into three clusters. That is, the cross points means the lower and bounder regions of C_1, respectively; the circular points and tangle points means the lower and bounder regions of C_2, respectively; the dots means the lower regions of C_3. Here, the boundary of C_3 is empty.

Observe Fig.2, the lower approximations of cluster C_2 and C_3 are overlapping, and the number of the overlapping objects(points) is 6, which can be denoted by $\underline{apr}(C_2) \cap \underline{apr}(C_3) \neq \emptyset$, and $|\underline{apr}(C_2) \cap \underline{apr}(C_3)| = 6$. In addition, $\underline{apr}(C_1) \cap \underline{apr}(C_3) = \emptyset$, $|\underline{apr}(C_1) \cap \underline{apr}(C_2)| = 66$, and $\overline{apr}(C_1) \neq \overline{apr}(C_2)$.

The conclusions from Fig.2 are positive to clustering. For example, when the dataset represents the network structure, where the C_1 and C_2 have so many

overlapping users. Obviously, it looks reasonable to build a new cluster composed by the uniting $\underline{apr}(C_1) \cup \underline{apr}(C_2)$. Otherwise, since the number of the overlapping objects between C_2 and C_3 is 6, we needn't to unite the two clusters. How to formal the idea is our further work.

4.2 UCI Data Set

More experiments on some standard data sets from UCI repository [11] are tested in this subsection, and results are shown in Table 1. In order to measure the test's accuracy, both the precision and the recall of the test are considered, and the F-measure is extended as follows.

Assume there is a data set $U = \{x_1, \ldots, x_i, \ldots, x_n\}$, and the objects in U are clustered into $T = \{T_1, \ldots, T_m, \ldots, T_M\}$. On the other hand, the result of clustering by the clustering algorithm based on three-way decision is: $\mathbf{C} = \{[\underline{apr}(C_1), \overline{apr}(C_1)], \ldots, [\underline{apr}(C_k), \overline{apr}(C_k)], \ldots, [\underline{apr}(C_K), \overline{apr}(C_K)]\}$.

Table 1. The CPU time and Results of the Algorithm

Database	$\|\mathbf{U}\|$	$\|A\|$	*distance*	Thresholds			Results	
				α, β	*Rth*	*mth*	$F-measure$	CPU(S)
iris	150	4	2.53	0.8,0.4	1.53	30	0.778	0.015
Letter1	1655	16	11.3	0.8,0.4	10	200	0.609	0.5
Poker1	199	10	11.2	0.8,0.4	11	80	0.595	0.078
Poker2	1188	10	12	0.8,0.4	11	900	0.503	1.296
White	4535	11	53	0.8,0.4	52	500	0.601	2.734

Precision is the number of correct upper approximate results divided by the number of all returned upper approximate results. *Recall* is the number of correct lower approximate results divided by the number of results that should have been returned. The $F-measure$ can be interpreted as a weighted as a weighted as a weighted average of the precision and recall, where an F-measure reaches its best value at 1 and worst value at 0. That is, the $F-measure$ can be denoted as the following equation.

$$F-measure \;=\; 2 \;\times\; \frac{Precision \;\times\; Recall}{Precision \;+\; Recall} \tag{8}$$

In Table 1, the $|\mathbf{U}|$ and $|A|$ are the number of objects and the number of attributes in the data set, respectively. *distance* means the average distance among objects in the data set.

Here, we choose some clusters from Letter and Porker datasets in UCI repository to generate some datasets used in Tabel 1.

Letter1 Dataset, which is composed of 1665 objects from Letter data set. There are 567 objects belong to decision attribute 'A', 570 objects belong to decision attribute 'E', and 528 objects belong to decision attribute 'O'.

Poker1 Dataset, which has 199 objects from Poker-hand-training-true data set. There are 93 objects belong to decision attribute '4', 54 objects belong to

decision attribute '5', 36 objects belong to decision attribute '6', 6 objects belong to '7', 5 objects belong to '8', and 4 objects belong to '9'.

Poker2 Dataset, which concludes 1188 objects from Poker-hand-training-true. There are 403 objects belong to decision attribute '0', 568 objects belong to '1', 165 objects belong to '3', 36 objects belong to '6', 6 objects belong to '7', 5 objects belong to '8', and 4 objects belong to '9'.

From Table 1, we can see that the CPU runtime and the F-measure are accredited. Actually, the results of clustering are changed with the change of the parameters such as α, β, Rth and mth. Through the experiments, we find out that the result would be better when the value of Rth close to the average distance. However, the accuracy of the algorithm need to improve.

5 Conclusion

In many applications such as network structure analysis, wireless sensor networks and biological information, an object should belong to more than one cluster, and as a result, cluster boundaries necessarily overlap. Three-way decisions rules constructed from the decision-theoretic rough set model are associated with different regions. This paper provides a three-way decision strategy for overlapping clustering. Here, each cluster is described by an interval set that is defined by a pair of sets called the lower and upper bounds. In addition, a density-based clustering algorithm is proposed and tested by using the new strategy. The analysis of the example indicates the strategy is effective to overlapping clustering. How to use less parameters and improve the accuracy of the algorithm is the further work.

Acknowledgments. This work was supported in part by the China NSFC grant(No.61073146) and the Chongqing CSTC grant (No.2009BB2082).

References

1. Aydin, N., Naït-Abdesselam, F., Pryyma, V., Turgut, D.: Overlapping Clusters Algorithm in Ad hoc Networks. In: 2010 IEEE Global Telecommunications Conference (2010)
2. Chen, M., Miao, D.Q.: Interval set clustering. Expert Systems with Application 38, 2923–2932 (2011)
3. Fu, Q., Banerjee, A.: Multiplicative Mixture Models for Overlapping Clustering. In: IEEE International Conference on Data Mining, pp. 791–797 (2003)
4. Herbert, J.P., Yao, J.T.: Learning Optimal Parameters in Decision-Theoretic Rough Sets. In: Wen, P., Li, Y., Polkowski, L., Yao, Y., Tsumoto, S., Wang, G. (eds.) RSKT 2009. LNCS, vol. 5589, pp. 610–617. Springer, Heidelberg (2009)
5. Lingras, P., Bhalchandra, P., Khamitkar, S., Mekewad, S., Rathod, R.: Crisp and Soft Clustering of Mobile Calls. In: Sombattheera, C., Agarwal, A., Udgata, S.K., Lavangnananda, K. (eds.) MIWAI 2011. LNCS, vol. 7080, pp. 147–158. Springer, Heidelberg (2011)

6. Liu, D., Yao, Y.Y., Li, T.R.: Three-way investment decisions with decision-theoretic rough sets. International Journal of Computational Intelligence Systems 4(1), 66–74 (2011)
7. Ma, S., Wang, T.J., Tang, S.W., Yang, D.Q., Gao, J.: A Fast Clustering Algorithm Based on Reference and Density. Journal of Softwar. 14(6), 1089–1095 (2003) (in Chinese)
8. Obadi, G., Dráždilová, P., Hlaváček, L., Martinovič, J., Snášel, V.: A Tolerance Rough Set Based Overlapping Clustering for the DBLP Data. In: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 57–60 (2010)
9. Pawlak, Z.: Rough sets. International Journal of Computer and Information Science 11(5), 341–356 (1982)
10. Takaki, M.: A Extraction Method of Overlapping Cluster based on Network Structure Analysis. In: IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, pp. 212–217 (2007)
11. UCIrvine Machine Learning Repository, `http://archive.ics.uci.edu/ml/`
12. Yao, Y.Y.: The-Superiority of Three-way Decisions in Probablistic Rough Set Models. Information Sciences 181, 1080–1096 (2011)
13. Yao, Y.Y., Wong, S.K.M.: A decision theoretic framework for approximating concepts. International Journal of Man-machine Studies 37(6), 793–809 (1992)
14. Yao, Y.Y., Lingras, P., Wang, R.Z., Miao, D.Q.: Interval Set Cluster Analysis: A Re-formulation. In: Sakai, H., Chakraborty, M.K., Hassanien, A.E., Ślęzak, D., Zhu, W. (eds.) RSFDGrC 2009. LNCS (LNAI), vol. 5908, pp. 398–405. Springer, Heidelberg (2009)
15. Zhou, B., Yao, Y.Y., Luo, J.G.: A Three-Way Decision Approach to Email Spam Filtering. In: Farzindar, A., Kešelj, V. (eds.) Canadian AI 2010. LNCS, vol. 6085, pp. 28–39. Springer, Heidelberg (2010)