Precision of Rough Set Clustering

Pawan Lingras[1], Min Chen[1,2], and Duoqian Miao[2]

[1] Department of Mathematics & Computing Science, Saint Mary's University, Halifax, Nova Scotia, B3H 3C3, Canada
[2] School of Electronics and Information Engineering, Tongji University, Shanghai, 201804, P.R. China
pawan.lingras@smu.ca

Abstract. Conventional clustering algorithms categorize an object into precisely one cluster. In many applications, the membership of some of the objects to a cluster can be ambiguous. Therefore, an ability to specify membership to multiple clusters can be useful in real world applications. Fuzzy clustering makes it possible to specify the degree to which a given object belongs to a cluster. In Rough set representations, an object may belong to more than one cluster, which is more flexible than the conventional crisp clusters and less verbose than the fuzzy clusters. The unsupervised nature of fuzzy and rough algorithms means that there is a choice about the level of precision depending on the choice of parameters. This paper describes how one can vary the precision of the rough set clustering and studies its effect on synthetic and real world data sets.

Keywords: Rough sets, K-means clustering algorithm, precision.

1 Introduction

In addition to clearly identifiable groups of objects, it is possible that a data set may consist of several objects that lie on the fringes. The conventional clustering techniques will mandate that such objects belong to precisely one cluster. Such a requirement is found to be too restrictive in many data mining applications. In practice, an object may display characteristics of different clusters. In such cases, an object should belong to more than one cluster, and as a result, cluster boundaries necessarily overlap. Fuzzy set representation of clusters, using algorithms such as fuzzy C-means, make it possible for an object to belong to multiple clusters with a degree of membership between 0 and 1 [11]. In some cases, the fuzzy degree of membership may be too descriptive for interpreting clustering results. Rough set based clustering provides a solution that is less restrictive than conventional clustering and less descriptive than fuzzy clustering.

Rough set theory has made substantial progress as a classification tool in data mining [1,14]. The basic concept of representing a set as lower and upper bounds can be used in a broader context such as clustering. Clustering in relation to rough set theory is attracting increasing interest among researchers [4,2,8,9,10,15,13]. Lingras [5] described how a rough set theoretic classification scheme can be represented using a rough set genome. In subsequent publications

C.-C. Chan et al. (Eds.): RSCTC 2008, LNAI 5306, pp. 369–378, 2008.

[6,7], modifications of K-means and *Kohonen Self-Organizing Maps* (SOMs) were proposed to create intervals of clusters based on rough set theory.

Clustering is an unsupervised learning process. That means there is no correct solution prescribed by an expert. For example, in a multidimensional space with a large number of objects, one cannot easily identify the number of clusters an algorithm should aim for. Researchers have proposed various cluster quality measures that make it possible to arrive at the appropriate number of clusters. The rough clustering has an additional issue that one needs to consider, namely, the precision of the clusters. Precision of the clusters refers to the number of objects that are precisely assigned to a cluster. An object in rough set clustering may be assigned to exactly one cluster or it may be assigned to multiple clusters. The objects that are assigned to multiple clusters are said to belong to the boundary region. Percentage of objects in boundary region is inversely proportional to the precision of rough clustering. This paper demonstrates how the size of boundary region can be varied with the help of *threshold* in rough set clustering. Experiments with a synthetic data set and a real world data set also suggest a procedure for choosing an appropriate precision.

2 Adaptation of Rough Set Theory for Clustering

Due to space limitations, some familiarity with rough set theory is assumed [14]. Rough sets were originally proposed using equivalence relations. However, it is possible to define a pair of upper and lower bounds $\left(\underline{A}(C), \overline{A}(C)\right)$ or a rough set for every set $C \subseteq U$ as long as the properties specified by Pawlak [14] are satisfied. Yao *et al.* [16] described various generalizations of rough sets by relaxing the assumptions of an underlying equivalence relation. Such a trend towards generalization is also evident in rough mereology proposed by Polkowski and Skowron [12] and the use of information granules in a distributed environment by Skowron and Stepaniuk. The present study uses such a generalized view of rough sets. If one adopts a more restrictive view of rough set theory, the rough sets developed in this paper may have to be looked upon as interval sets.

Let us consider a hypothetical classification scheme

$$U/P = \{C_1, C_2, \ldots, C_k\} \tag{1}$$

that partitions the set U based on an equivalence relation P. Let us assume due to insufficient knowledge that it is not possible to precisely describe the sets $C_i, 1 \leq i \leq k$, in the partition. Based on the available information, however, it is possible to define each set $C_i \in U/P$ using its lower $\underline{A}(C_i)$ and upper $\overline{A}(C_i)$ bounds. We will use m-dimensional vector representations, $\mathbf{u}, \mathbf{v}$ for objects and $\mathbf{c_i}$ for cluster C_i.

We are considering the upper and lower bounds of only a few subsets of U. Therefore, it is not possible to verify all the properties of the rough sets [14]. However, the family of upper and lower bounds of $\mathbf{c_i} \in \mathbf{U/P}$ are required to follow some of the basic rough set properties such as:

(P1) An object $\mathbf{x}$ can be part of at most one lower bound
(P2) $\mathbf{x} \in \underline{A}(\mathbf{c}_i) \Longrightarrow \mathbf{x} \in \overline{A}(\mathbf{c}_i)$
(P3) An object $\mathbf{x}$ is not part of any lower bound $\Longleftrightarrow$
$\mathbf{x}$ belongs to two or more upper bounds.

Property (P1) emphasizes the fact that a lower bound is included in a set. If two sets are mutually exclusive, their lower bounds should not overlap. Property (P2) confirms the fact that the lower bound is contained in the upper bound. Property (P3) is applicable to the objects in the boundary regions, which are defined as the differences between upper and lower bounds. The exact membership of objects in the boundary region is ambiguous. Therefore, property (P3) states that an object cannot belong to only a single boundary region. Their discussion can provide more insight into the essential properties for a rough set model. Note that (P1)-(P3) are not necessarily independent or complete. However, enumerating them will be helpful later in understanding the rough set adaptation of evolutionary, neural, and statistical clustering methods. In the context of decision-theoretic rough set model, Yao and Zhao [17] provide a more detailed discussion on the important properties of rough sets and positive, boundary, and negative regions.

3 Adaptation of K-Means to Rough Set Theory

Here, we refer readers to [3] for discussion on conventional K-means algorithm. Incorporating rough sets into K-means clustering requires the addition of the concept of lower and upper bounds. Calculation of the centroids of clusters from conventional K-Means needs to be modified to include the effects of these bounds. The modified centroid calculations for rough sets are then given by:

$$
\begin{aligned}
&\text{if} \quad \underline{A}(\mathbf{c}) \neq \emptyset \text{ and } \overline{A}(\mathbf{c}) - \underline{A}(\mathbf{c}) = \emptyset \\
&\qquad c_j = \frac{\sum_{\mathbf{x} \in \underline{A}(\mathbf{c})} x_j}{|\underline{A}(\mathbf{c})|} \\
&\text{else if} \quad \underline{A}(\mathbf{c}) = \emptyset \text{ and } \overline{A}(\mathbf{c}) - \underline{A}(\mathbf{c}) \neq \emptyset \\
&\qquad c_j = \frac{\sum_{\mathbf{x} \in (\overline{A}(\mathbf{c}) - \underline{A}(\mathbf{c}))} x_j}{|\overline{A}(\mathbf{c}) - \underline{A}(\mathbf{c})|} \\
&\text{else} \\
&\qquad c_j = w_{lower} \times \frac{\sum_{\mathbf{x} \in \underline{A}(\mathbf{c})} x_j}{|\underline{A}(\mathbf{c})|} + w_{upper} \times \frac{\sum_{\mathbf{x} \in (\overline{A}(\mathbf{c}) - \underline{A}(\mathbf{c}))} x_j}{|\overline{A}(\mathbf{c}) - \underline{A}(\mathbf{c})|},
\end{aligned}
\tag{2}
$$

where $1 \leq j \leq m$. Here, m is the dimensions of the vectors $\mathbf{c}$ and $\mathbf{x}$. The parameters w_{lower} and w_{upper} correspond to the relative importance of lower and upper bounds, and $w_{lower} + w_{upper} = 1$. If the upper bound of each cluster were equal to its lower bound, the clusters would be conventional clusters. Therefore, the boundary region $\overline{A}(\mathbf{c}) - \underline{A}(\mathbf{c})$ will be empty, and the second term in the

equation will be ignored. Thus, Eq. (2) will reduce to conventional centroid calculations.

The next step in the modification of the K-means algorithms for rough sets is to design criteria to determine whether an object belongs to the upper or lower bound of a cluster given as follows. For each object vector $\mathbf{x}$, let $d(\mathbf{x}, \mathbf{c}_j)$ be the distance between itself and the centroid of cluster $\mathbf{c}_j$. Let $d(\mathbf{x}, \mathbf{c}_i) = \min_{1 \leq j \leq k} d(\mathbf{x}, \mathbf{c}_j)$. The ratio $d(\mathbf{x}, \mathbf{c}_i)/d(\mathbf{x}, \mathbf{c}_j)$, $1 \leq i, j \leq k$, are used to determine the membership of $\mathbf{x}$. Let $T = \{j : d(\mathbf{x}, \mathbf{c}_i)/d(\mathbf{x}, \mathbf{c}_j) \leq threshold \text{ and } i \neq j\}$.

1. If $T \neq \emptyset$, $\mathbf{x} \in \overline{A}(\mathbf{c}_i)$ and $\mathbf{x} \in \overline{A}(\mathbf{c}_j), \forall j \in T$. Furthermore, $\mathbf{x}$ is not part of any lower bound. The above criterion guarantees that property (P3) is satisfied.
2. Otherwise, if $T = \emptyset$, $\mathbf{x} \in \underline{A}(\mathbf{c}_i)$. In addition, by property (P2), $\mathbf{x} \in \overline{A}(\mathbf{c}_i)$.

It should be emphasized that the approximation space A is not defined based on any predefined relation on the set of objects. The upper and lower bounds are constructed based on the criteria described above.

4 Refinements of Rough Set Clustering

Rough clustering is gaining increasing attention from researchers. The rough K-means approach, in particular, has been a subject of further research. Peters [15] discussed various deficiencies of Lingras and West's original proposal [6]. The first set of independently suggested alternatives by Peters are similar to the Eq. (2). Peters also suggest the use of ratios of distances as opposed to differences between distances similar to those used in the rough set based Kohonen algorithm described in [7]. The use of ratios is a better solution than differences. The differences vary based on the values in input vectors. The ratios, on the other hand, are not susceptible to the input values. Peters [15] have proposed additional significant modifications to rough K-means that improve the algorithm in a number of aspects. The refined rough K-means algorithm simplifies the calculations of the centroid by ensuring that lower bound of every cluster has at least one object. It also improves the quality of clusters as clusters with empty lower bound have a limited basis for its existence. Peters tested the refined rough K-means for various datasets. The experiments were used to analyze the convergence, dependency on the initial cluster assignment, study of Davies-Boulden index, and to show that the boundary region can be interpreted as a security zone as opposed to the unambiguous assignments of objects to clusters in conventional clustering. Despite the refinements, Peters concluded that there are additional areas in which the rough K-means needs further improvement, namely in terms of selection of parameters.

By its very definition, unsupervised learning is an exercise with no known solution. Clustering is one of the primary examples of unsupervised clustering, which attempts to find groups of objects with similar characteristics. There are a number of unknowns involved in the process. The appropriate number of groups

is not known apriori. Measures such as Davies-Boulden index have been used to identify the most appropriate number of clusters. As mentioned previously, even if there were clearly identifiable clusters of objects, it is quite often likely that some of the objects may be straying from these clusters. In that case, the next issue is how to decide what percentage of objects are straying from the neatly formed clusters. These stray objects will then be assigned to boundary regions of multiple clusters using the rough K-means algorithm. This paper experiments with the issue of determining the appropriate number of boundary region objects using two data sets. The first data set is a two dimensional set of objects artificially created with clearly identifiable clusters and stray objects. Since we can visualize the appropriate rough set clustering, we can test the behavior of the rough K-means algorithm for different values of *threshold*. The *threshold* parameter helps us control the size of the boundary region. We define the percentage of boundary region as a ratio of cardinality of the union of all the boundary regions divided by the total number of objects expressed as percentages, given by:

$$BoundarySize = \frac{\| \bigcup_{\mathbf{c} \in \mathbf{U/P}} (\overline{A}(\mathbf{c}) - \underline{A}(\mathbf{c})) \|}{\|U\|} \times 100 \tag{3}$$

The following section studies the variation in *BoundarySize* along with qualitative analysis of changing memberships to suggest a procedure for identifying appropriate value of the *threshold* in the rough K-means algorithm.

5 Study Data and Experimental Analysis

We use two kinds of data, synthetic data and real data, to demonstrate how to choose an appropriate *threshold* for rough clustering.

5.1 Synthetic Data

The synthetic data set has been developed to study how the *BoundarySize* varies with *threshold* for rough clustering. In order to visualize the data set, we restrict it to two dimensions as can be seen in Fig. 1. There are a total of 65 objects. It is obvious that there are three distinct clusters, denoted by C_1, C_2 and C_3. However, five objects, identified as x_i $(1 \leq i \leq 5)$, do not belong to any particular cluster. We performed rough clustering on the synthetic data set for different values of *threshold*.

Fig. 2 shows how changing the value of *threshold* can affect the *BoundarySize* of rough clustering with $k = 3$ and $w_{lower} = 0.75$. In the inset figure, we can see a slow increase in the *BoundarySize* until the *threshold* reaches a value of 1.4, since the higher values lead to larger boundary regions. While the *threshold* values were changed from 1.4 to 2, the *BoundarySize* remained constant at 7.7%. However, the re-distribution of objects in the boundary region did occur.

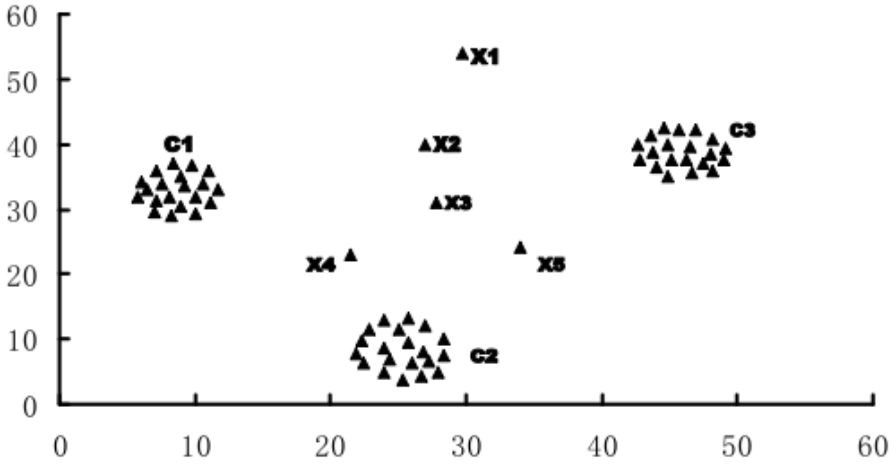

Fig. 1. Synthetic data

For example, $\mathbf{x}_5$, which was in the boundary region of $\mathbf{c}_2$ and $\mathbf{c}_3$, was also added to the boundary region of $\mathbf{c}_1$, when *threshold* changed the value from 1.6 into 1.7. It is obvious from Fig. 1 that $\mathbf{x}_5$ should only belong to the boundary region of $\mathbf{c}_2$ and $\mathbf{c}_3$. That means increasing the value of *threshold* beyond a certain value can lead to unreasonable addition of some objects to boundary regions of some of the clusters. Moreover, one should not increase the boundary region too much as it will lead to fairly indecisive and uninformative rough clustering. Fig. 2 shows a sudden and sharp increase in the *BoundarySize* after *threshold* reaches a value of 2. The *BoundarySize* goes up to a value of more than 50% when *threshold* reaches the value of 2.5. Therefore, it is reasonable to consider $threshold = 1.4$ as an appropriate value in terms of the variance in *BoundarySize*. This value of *threshold* can be identified by the fact that further number of increases in *threshold* do not lead to net change in *BoundarySize*.

5.2 Real Data

This section reports experiments with a real world data set belonging to a small retail chain. The data consists of all the customer transactions in 2006. There were a total of 68716 transactions, one transaction per item purchased. 40260 of these transactions can be associated with 5878 identified customers. The objective of the experiment is to cluster the customers based on their spending habits. Each customer is represented by his monthly spending patterns. The monthly spending pattern gives a better understanding of a customer's spending habits than total spending. A customer who spends $100 regularly may be a little more loyal than one who spends $1000 during a single visit. The chronological ordering of spending does not help us understand the propensity of a customer to spend. For example, a person spending $100, $200, $300 in three months will look different from the one

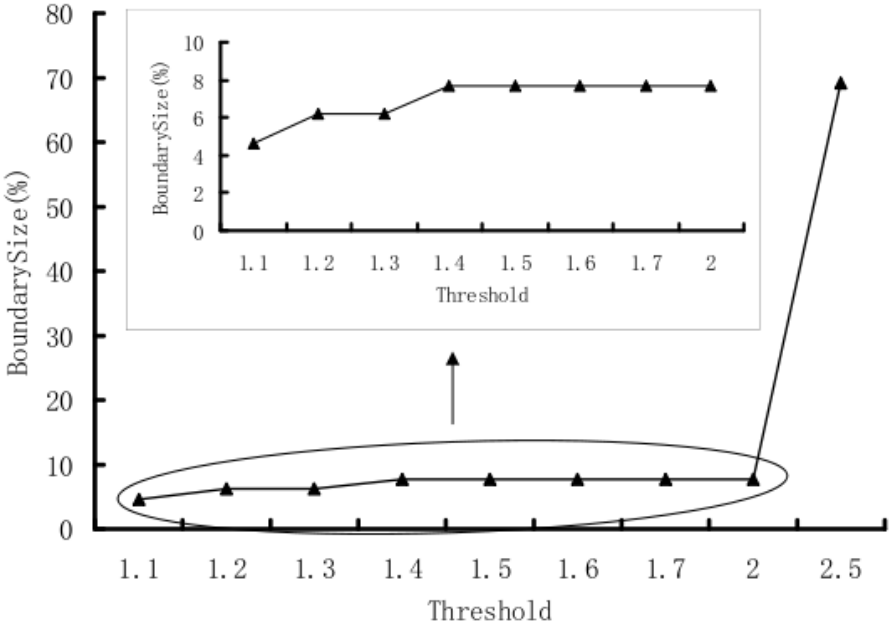

Fig. 2. Synthetic data: Change in *BoundarySize* with *threshold*

who spends \$300, \$100, \$200 during the same three months. Therefore, we sort the spending values, which makes the two customers identical in terms of their revenue generation potential. Instead of using twelve monthly spending and visit values, which may be too detailed for the purpose of grouping, we will represent the patterns using the lowest, highest and average spending. However, in some cases, lowest and highest values can be outliers. Therefore, we use second highest, second lowest and median values as a representative of the pattern.

313 customers visited in only one month. These customers were termed as infrequent customers. It was decided that there was no further need for grouping these customers. After eliminating the 313 customers, the number of customers was 5565. After experimenting with different number of clusters we set $k = 5$. w_{lower} was set at 0.75.

Fig. 3 describes the *BoundarySize* changes with the *threshold*, which is similar to the one found for the synthetic data. The *BoundarySize* goes up a little slowly until the *threshold* reaches a value of 1.4, where there is a marked increase. This suggests that 1.4 may be an appropriate value for the *threshold*. We can also see a sudden and sharp jump at *threshold* = 2.5. This reinforces our earlier observation that high values of *threshold* may lead to inconclusive rough clustering. Fig. 4 presents the rough centroids as the representative patterns for each cluster. Cluster $\mathbf{c}_1$ is the largest cluster consisting of moderate spenders who spend \$0 to \$52 in a month. The next cluster, $\mathbf{c}_2$, is about the quarter the size of $\mathbf{c}_3$ with spending ranging from \$0 to \$100. Third cluster ($\mathbf{c}_3$) is even smaller

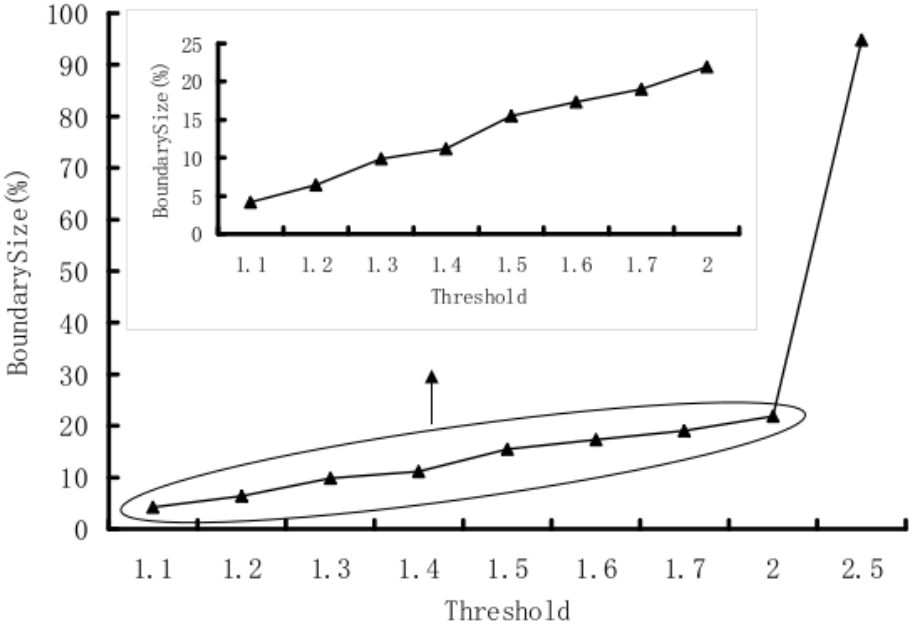

Fig. 3. Real data: Change in *BoundarySize* with *threshold*

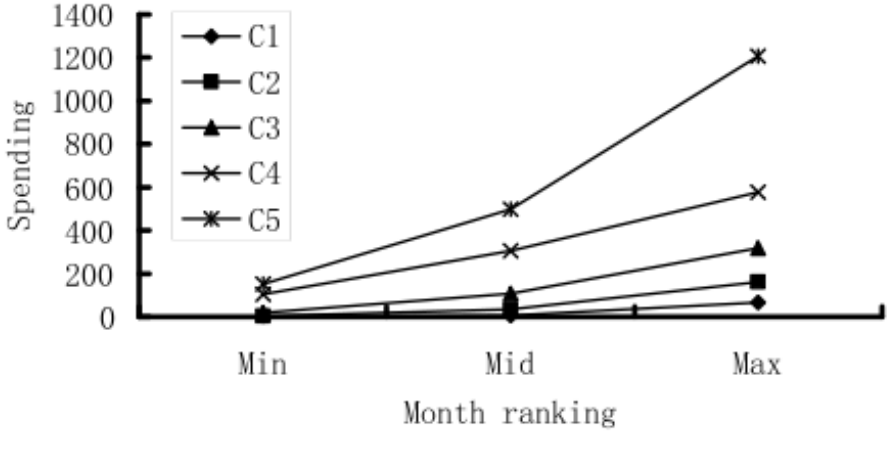

Fig. 4. Rough centroids for the retail data

with spending ranging from \$10 to \$250. Fourth cluster has approximately 70 to 100 customers who spend \$120 to \$500. The last cluster is the smallest with spending ranging from \$137 to \$1330. The overlap between different clusters for *threshold* = 1.4 and *threshold* = 2 are shown in Table 1. It can be seen in Table 1(a) that the intermediate clusters, i.e. $\mathbf{c}_2$, $\mathbf{c}_3$, and $\mathbf{c}_4$ have overlaps with two clusters on either side. For example, $\mathbf{c}_2$ overlaps with $\mathbf{c}_1$ and $\mathbf{c}_3$, while $\mathbf{c}_3$

Table 1. The number of objects in the intersection of clusters

	C1	C2	C3	C4	C5
C1	–	403	0	0	0
C2	403	–	177	0	0
C3	0	177	–	41	0
C4	0	0	41	–	9
C5	0	0	0	9	–

(a)threshold=1.4

	C1	C2	C3	C4	C5
C1	–	809	81	11	8
C2	809	–	388	28	9
C3	81	388	–	163	18
C4	11	28	163	–	59
C5	8	9	18	59	–

(b)threshold=2.0

overlaps with $\mathbf{c}_2$ and $\mathbf{c}_4$, and $\mathbf{c}_4$ overlaps with $\mathbf{c}_3$ and $\mathbf{c}_5$. Clusters $\mathbf{c}_1$ and $\mathbf{c}_5$ have overlap with only one cluster: $\mathbf{c}_1$ with $\mathbf{c}_2$ and $\mathbf{c}_5$ with $\mathbf{c}_4$. When the *threshold* is raised to 2.0, we can see from Table 1(b) that each cluster overlaps with other four clusters. That means many objects have now moved to boundary regions of all the clusters. This makes any conclusion about their membership impossible.

6 Conclusions

Rough set clustering makes it possible to assign stray objects - that may not belong to a precise cluster - to boundary regions of two or more clusters. This aspect of rough set clustering adds a degree of imprecision to the clustering scheme. The degree of imprecision is an additional unknown in the unsupervised learning based on rough set theory. The experiments with a synthetic data set and a real world data set show that it is important to choose a right balance between rough and precise cluster assignments. The paper describes a procedure that can be used to control the imprecision in rough set clustering for the rough K-means algorithm by varying the *threshold* parameter. The results presented here lay foundations for a more comprehensive study of the quality of rough set clustering, which will be presented in a subsequent publication.

Acknowledgement

The authors would like to thank China Scholarship Council and NSERC Canada for their financial support.

References

1. Banerjee, M., Mitra, S., Pal, S.K.: Rough fuzzy MLP: knowledge encoding and classification. IEEE Transactions on Neural Networks 9(6), 1203–1216 (1998)
2. Ho, T.B., Nguyen, N.B.: Nonhierarchical Document Clustering by a Tolerance Rough Set Model. International Journal of Intelligent Systems 17(2), 199–212 (2002)
3. Hartigan, J.A., Wong, M.A.: Algorithm AS136: A K-Means Clustering Algorithm. Applied Statistics 28, 100–108 (1979)

4. Hirano, S., Tsumoto, S.: Rough Clustering and Its Application to Medicine. Information Sciences 124, 125–137 (2000)
5. Lingras, P.: Unsupervised Rough Set Classification using GAs. Journal Of Intelligent Information Systems 16(3), 215–228 (2001)
6. Lingras, P., West, C.: Interval Set Clustering of Web Users with Rough K-means. Journal of Intelligent Information Systems 23(1), 5–16 (2004)
7. Lingras, P., Hogo, M., Snorek, M.: Interval Set Clustering of Web Users using Modified Kohonen Self-Organizing Maps based on the Properties of Rough Sets. Web Intelligence and Agent Systems: An International Journal 2(3), 217–230 (2004)
8. Mitra, S.: An evolutionary rough partitive clustering. Pattern Recognition Letters 25, 1439–1449 (2004)
9. Mitra, S., Bank, H., Pedrycz, W.: Rough-Fuzzy Collaborative Clustering. IEEE Trans. on Systems, Man and Cybernetics 36(4), 795–805 (2006)
10. Nguyen, H.S.: Rough Document Clustering and the Internet. Handbook on Granular Computing (2007)
11. Pedrycz, W., Waletzky, J.: Fuzzy Clustering with Partial Supervision. IEEE Trans. on Systems, Man and Cybernetics 27(5), 787–795 (1997)
12. Polkowski, L., Skowron, A.: Rough Mereology: A New Paradigm for Approximate Reasoning. International Journal of Approximate Reasoning 15(4), 333–365 (1996)
13. Peters, J.F., Skowron, A., Suraj, Z., Rzasa, W., Borkowski, M.: Clustering: A rough set approach to constructing information granules. In: Proceedings of 6th International Conference on Soft Computing and Distributed Processing, Rzeszow, Poland, June 24-25, 2002, pp. 57–61 (2002)
14. Pawlak, Z.: Rough Sets: Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht (1992)
15. Peters, G.: Some Refinements of Rough k-Means. Pattern Recognition 39(8), 1481–1491 (2006)
16. Yao, Y.Y.: Constructive and algebraic methods of the theory of rough sets. Information Sciences 109, 21–47 (1998)
17. Yao, Y.Y., Zhao, Y.: Attribute reduction in decision-theoretic rough set models. Information Sciences (to appear, 2008)