Expert Systems with Applications 38 (2011) 2923–2932

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Interval set clustering

Min Chen *, Duoqian Miao

Department of Computer Science and Technology, Tongji University, Shanghai 201804, China

ARTICLE INFO

Keywords:
Rough *k*-means clustering
Cluster validity
Decision theory
Risk

ABSTRACT

Rough *k*-means clustering describes uncertainty by assigning some objects to more than one cluster. Rough cluster quality index based on decision theory is applicable to the evaluation of rough clustering. In this paper we analyze rough *k*-means clustering with respect to the selection of the *threshold*, the value of risk for assigning an object and uncertainty of objects. According to the analysis, clusters presented as interval sets with lower and upper approximations in rough *k*-means clustering are not adequate to describe clusters. This paper proposes an interval set clustering based on decision theory. Lower and upper approximations in the proposed algorithm are hierarchical and constructed as outer-level approximations and inner-level ones. Uncertainty of objects in out-level upper approximation is described by the assignment of objects among different clusters. Accordingly, ambiguity of objects in inner-level upper approximation is represented by *local uniform factors of objects*. In addition, interval set clustering can be improved to obtain a satisfactory clustering result with the optimal number of clusters, as well as optimal values of parameters, by taking advantage of the usefulness of rough cluster quality index in the evaluation of clustering. The experimental results on synthetic and standard data demonstrate how to construct clusters with satisfactory lower and upper approximations in the proposed algorithm. The experiments with a promotional campaign for the retail data illustrates the usefulness of interval set clustering for improving rough *k*-means clustering results.

1. Introduction

Cluster analysis is a widely used technique in data mining and are applied to diverse areas. The main objective in cluster analysis is to categorizes unlabeled objects into several clusters such that the objects belonging to the same cluster are more similar than those belonging to different clusters. Sometimes, it is impossible to assign an object to exactly one cluster. Rough sets as one of soft computing methodology is capable of handling such challenge in data mining. In the past years, rough set based variations of *k*-means clustering (Asharaf, Shevade, & Murty, 2005; Hirano & Tsumoto, 2005; Lingras & West, 2004; Nguyen, 2007; Peters, Skowron, Suraj, Rzasa, & Borkowski, 2002) have been proposed.

In rough *k*-means clustering clusters are represented as interval sets with lower and upper approximations. The lower approximation is a subset of the upper approximation. The objects in the lower approximations belong certainly to the clusters. The objects in the upper approximations may belong to other clusters. Because an object that does not belong to any lower approximations is members of at least two upper approximations. However, some objects in the lower approximations may be farther from the centroid of the cluster they belong to than other objects in the same lower approximations.

Quality of clustering is an important issue in application of clustering techniques. Rough cluster quality index (Lingras, Chen, & Miao, 2009) is constructed by considering various loss functions based on decision theory. It is taken as a function of total risk for grouping objects using a clustering algorithm. Rough cluster quality index is applicable to both rough clustering and crisp clustering. Moreover, it helps determine optimal number of clusters, as well as an important parameter called *threshold* in rough clustering.

This paper analyzes the selection of the *threshold* in rough *k*-means clustering, the value of risk for assigning an object and uncertainty of objects. According to the analysis, an interval set clustering algorithm based on decision theory is proposed. Lower and upper approximations in the proposed algorithm are hierarchical and constructed as outer-level approximations and inner-level ones. Uncertainty of objects in out-level upper approximation is described by the assignment among different clusters. Accordingly, ambiguity of objects in inner-level upper approximation is represented by *local uniform factors*. In addition, interval set clustering can obtain a satisfactory clustering result with the optimal number of clusters, as well as optimal values of parameters, by taking advantage of the usefulness of rough cluster quality index in the evaluation of clustering.

The structure of the paper is as follows. In Section 2 we introduce rough *k*-means clustering and rough cluster quality index based on decision theory. In Section 3 we investigate the selection of the *threshold* in rough *k*-means clustering, the value of risk for

* Corresponding author.
E-mail address: tomatocm@163.com (M. Chen).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.06.052

assigning an object and uncertainty of objects. According to the analysis we propose an interval set clustering algorithm based on decision theory. The experiments on synthetic data, standard data and the retail data are presented in Section 5. The paper concludes with a summary in Section 6.

2. Literature review

First, we describe the notations that will appear in this section. Let $X = \{\vec{x}_1, \ldots, \vec{x}_n\}$ be a finite set of objects. Assuming that the objects are represented by m-dimensional vectors. A classifying scheme classifies n objects into k categories $C = \{\vec{c}_1, \ldots, \vec{c}_k\}$. We use the term *category* instead of class or cluster to emphasize the fact that it can be used in supervised and unsupervised learning. For a clustering scheme (CS), such as crisp clustering and rough clustering, C is the set of clusters. And each of the clusters $\vec{c}_i$ is represented by an m-dimensional vector, which is the centroid or mean vector for that cluster.

This section also introduce some notations related to rough set. The notion of rough set was proposed by Pawlak (1982), Pawlak (1984), Pawlak (1992) and Pawlak et al. (1988). Let E an equivalence relation on X. The pair $apr = (X, E)$ is called an approximation space. Any subset $A \subseteq X$ may be represented by its lower and upper approximations. The lower approximation $\underline{apr}(A)$ is the union of all the elementary sets which are subsets of A, and the upper approximation $\overline{apr}(A)$ is the union of all the elementary sets which have a non-empty intersection with A. We call $bnd(A) = \overline{apr}(A) - \underline{apr}(A)$ is the boundary region of A.

2.1. Rough k-means clustering

Lingras and West incorporated rough set into k-means clustering, which requires the addition of the concept of lower and upper bounds (Lingras & West, 2004). This section describes a refined version of the original proposal (Lingras, 2007; Lingras, Hogo, & Snorek, 2004; Peters, 2006). The following equation is used to calculate the centroids of clusters that needs to be modified to include the effects of lower as well as upper bounds. The modified centroid calculations for rough clustering are then given by:

$$\vec{c}_i = \begin{cases} \omega_l \times \frac{\sum_{\vec{x}_l \in \underline{apr}(\vec{c}_i)} \vec{x}_l}{|\underline{apr}(\vec{c}_i)|} + \omega_b \times \frac{\sum_{\vec{x}_l \in bnd(\vec{c}_i)} \vec{x}_l}{|bnd(\vec{c}_i)|} \\ \text{for } \underline{apr}(\vec{c}_i) \neq \phi \text{ and } bnd(\vec{c}_i) \neq \emptyset, \\ \frac{\sum_{\vec{x}_l \in \underline{apr}(\vec{c}_i)} \vec{x}_l}{|\underline{apr}(\vec{c}_i)|} \\ \text{for } \underline{apr}(\vec{c}_i) \neq \phi \text{ and } bnd(\vec{c}_i) = \emptyset, \\ \frac{\sum_{\vec{x}_l \in bnd(\vec{c}_i)} \vec{x}_l}{|bnd(\vec{c}_i)|} \\ \text{for } \underline{apr}(\vec{c}_i) = \phi \text{ and } bnd(\vec{c}_i) \neq \emptyset, \end{cases} \quad (1)$$

where $\omega_l + \omega_b = 1$ and $1 \leqslant i \leqslant n$. The parameters ω_l and ω_b correspond to the relative importance of lower and upper bounds. The next step is to design criteria to determine whether an object belongs to the upper and lower bound of a cluster. For any object vector, $\vec{x}_l (1 \leqslant l \leqslant n)$, let $d(\vec{x}_l, \vec{c}_i)$ be the distance between itself and the centroid of cluster $\vec{c}_i$. The ratio $d(\vec{x}_l, \vec{c}_i)/d(\vec{x}_l, \vec{c}_j),\ 1 \leqslant i,\ j \leqslant k$, are used to determine the membership of $\vec{x}_l$ (Lingras et al., 2004; Peters, 2006). Let $d(\vec{x}_l, \vec{c}_i) = min_{1 \leqslant j \leqslant k} d(\vec{x}_l, \vec{c}_j)$ and $T_l = \{j : d(\vec{x}_l, \vec{c}_i)/d(\vec{x}_l, \vec{c}_j) \geqslant threshold \text{ and } i \neq j\}$.

1. If $T_l \neq \emptyset,\ \vec{x}_l \in \overline{apr}(\vec{c}_j),\ \forall j \in T_l$. Furthermore, $\vec{x}_l$ is not part of any lower bound.
2. Otherwise, if $T_l = \emptyset,\ \vec{x}_l \in \underline{apr}(\vec{c}_i)$.

The rough k-means algorithm, described above, depends on three parameters ω_l, ω_b and *threshold*. It should be emphasized that approximation space apr 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. Though it is not possible to verify all the properties of rough set for rough k-means clustering, it can be easily shown that the resulting upper and lower approximations in fact follow important rough set theoretic properties as follows:

(P1) An object can be part of at most one lower approximation
(P2) $\vec{x}_l \in \underline{apr}(\vec{c}_i) \Rightarrow \vec{x}_l \in \overline{apr}(\vec{c}_i)$
(P3) An object $\vec{x}_l$ is not part of any lower approximation
$\Updownarrow$
$\vec{x}_l$ belongs to two or more upper approximations.

2.2. Rough cluster quality based on decision theory

So far most cluster validity indices are proposed to evaluate crisp and fuzzy clustering (Bezdek & Pal, 1995; Bezdek & Pal, 1998; Davies & Bouldin, 1979; Dunn, 1973; Dunn, 1974). There is an evaluation measure proposed by Lingras et al. (2009) for rough clustering at present. Under the decision theoretic framework (Lingras, Chen, & Miao, 2008; Yao, 2003; Yao, 2007), Lingras et al. proposed a cluster validity indices for rough clustering.

Rough cluster quality considered various loss functions for assigning an object and clustering scheme. The definitions are given as follows:

2.2.1. Cluster core

Let $core(\vec{c}_i)$ be the core of the cluster $\vec{c}_i$, which is used to calculate the centroid of the cluster. Any $\vec{x}_l \in core(\vec{c}_i)$ can not belong to other clusters. Therefore, $core(\vec{c}_i)$ can be considered the best representation of $\vec{c}_i$ to a certain extent.

2.2.2. Risk for assigning an object to clusters

For a given clustering scheme CS, let $b_j(CS, \vec{x}_l)$ be the action that assigns the object $\vec{x}_l$ to a cluster or a group of clusters. (Note that an object may not belong to a single cluster under rough clustering.) The risk associated with the assignment will then be given as $R(b_j(CS, \vec{x}_l)|\vec{x}_l)$. $R(b_j(CS, \vec{x}_l)|\vec{x}_l)$ is obtained assuming that the conditional probability $P(\vec{c}_i|\vec{x}_l)$ is proportional to the similarity between $\vec{x}_l$ and $core(\vec{c}_i)$.

2.2.3. Group risk for clustering scheme

Given a clustering scheme (CS) and a group of objects $\vec{c} = \{\vec{x}_1, \ldots, \vec{x}_g\}$, we define $R(CS, \vec{c})$ as the group risk for $\vec{c}$ under a clustering scheme, given by:

$$R(CS, \vec{c}) = \sum_{\vec{x}_l \in \vec{c}} R(b_j(CS, \vec{x}_l)|\vec{x}_l).$$

Therefore, the cluster validity indices for a clustering scheme (CS) can be taken as the function of group risk, defined as follows:

$$R(CS) = \sum_{i=1}^{k} R(CS, \vec{c}_i) = \sum_{l=1}^{n} R(b_j(CS, \vec{x}_l)|\vec{x}_l).$$

Obviously, the smaller the value of the total risk, the better a clustering scheme is. The objective is to minimize $R(CS)$ in order to obtain the optimal number of clusters for a clustering scheme (CS).

Crisp clustering is a special case of rough clustering. The proposed risk measure for crisp clustering can be expressed as follows:

$$R(CC) = \sum_{i=1}^{k} R(CC, core(\vec{c}_i)) = \sum_{i=1}^{k} R\left(CC, \underline{apr}(\vec{c}_i)\right) = \sum_{i=1}^{k} R(CC, \overline{apr}(\vec{c}_i)).$$

Rough cluster quality index is an evaluation measure for rough clustering as well as crisp clustering. This is the first measure that takes into account special features of rough clustering that allow for an object to belong to more than one cluster. The measure is shown to be useful in determining important aspects of a clustering exercise such as determining the appropriate number of clusters and size of boundary region (in case of rough clustering). Such a cluster validity measure can be useful in further theoretical development in clustering.

3. Comments on rough *k*-means clustering

Rough *k*-means clustering describes uncertainty of objects by assigning objects in the boundary region to more than one cluster. In order to have a close look at the assignments of objects and the value of risk for the assignments for rough *k*-means clustering, we analyze the selection of the *threshold*, uncertainty of objects in the boundary region and the value of risk for assigning an object, with the following example.

Example 1. Let $X = \{\vec{x}_1, \ldots, \vec{x}_{20}\}$ and $C = \{\vec{c}_1, \vec{c}_2\}$. The distribution of objects is shown in Fig. 1. For simplify the figure, we drop the notation $\vec{x}_i$ $(5 \leqslant i \leqslant 20)$ from Fig. 1, if no confusion arises. According to the distribution, sixteen objects $\vec{x}_i$ $(5 \leqslant i \leqslant 20)$ are expected to form two groups with different densities, but four objects $\vec{x}_i$ $(1 \leqslant i \leqslant 4)$ are comparatively far from these groups.

3.1. The selection of the threshold

There is limitation of the selection of the *threshold*. According to the definition of the T_l in rough *k*-means clustering, objects lie in between two clusters tend to be in the boundary region. If the *threshold* is at a very small value, such as 0.1, some objects close to the centroids of clusters are assigned to the boundary region by mistake. However, when the *threshold* is at a reasonable value, some objects in the lower approximation seem isolated comparing to other objects in the same lower approximation. Therefore, the *threshold* together with T_l is not adequate enough to decide the boundary region for rough *k*-means clustering. It can be explained by the following example.

Example 1(a). We performed rough *k*-means clustering on the objects in Fig. 1 with $k = 2$ and $\omega_l = 0.75$ for different *threshold*s. Fig. 2(a) presents the clustering result when the value of the *threshold* changes from 0.9 to 0.8. When the value of *threshold* decreases from 0.7 to 0.5 the clustering result is shown in Fig. 2(b). In the figures, the dashed line outlines the upper approximation of each cluster, and the solid line describes the lower approximation of each cluster. It can be seen from Fig. 2(a) that no objects in the boundary region when the value of the *threshold* changes from 0.9 to 0.8. Since the upper approximations are the same as the lower approximations. We can get a reasonable clustering result shown in Fig. 2(b) when the *threshold* changes from 0.7 to 0.5. However, $\vec{x}_1$ and $\vec{x}_2$ are isolated in the corresponding lower approximations they belong to. It is noted that the clustering result is unacceptable when *threshold* < 0.5. For example, two objects $\vec{x}_i$ and $\vec{x}_j$ $(5 \leqslant i, j \leqslant 20)$ are in the boundary region when *threshold* = 0.3.

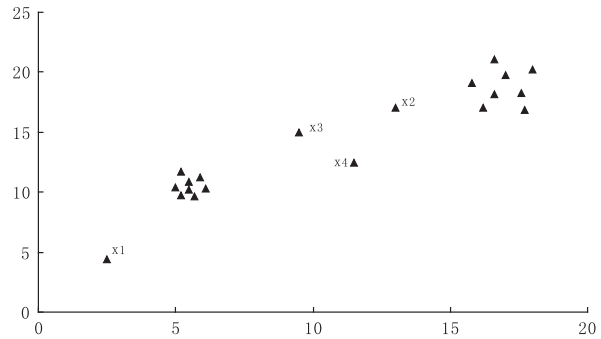

Fig. 1. Distribution of objects.

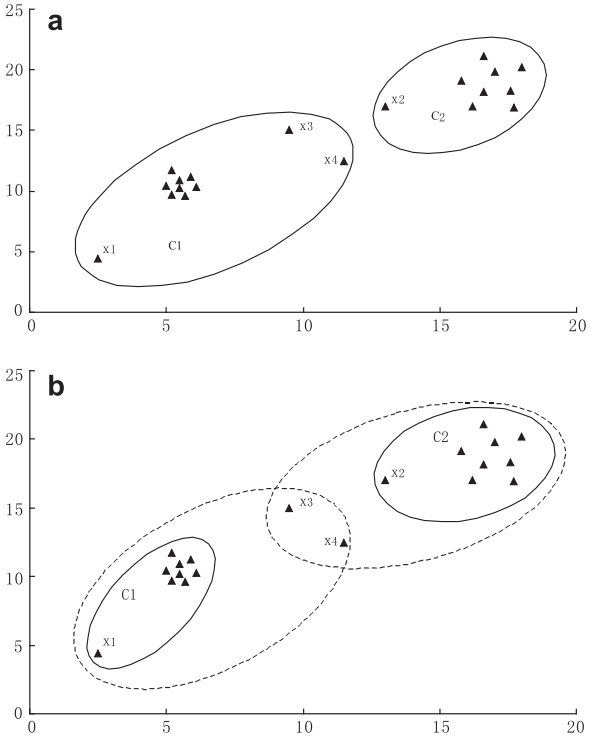

Fig. 2. Rough clustering with change in threshold.

3.2. Uncertainty of objects in the boundary region

First, let us look at definitions of lower approximation and upper approximation in rough set. Assuming that U is the universe and Y is a subset of U. Let R be an equivalence relation on U. The lower approximation and upper approximation of Y are denoted as $\underline{apr}(Y)$ and $\overline{apr}(Y)$, respectively. Accordingly, the boundary region of Y is defined as $BN_R(Y) = \overline{apr}(Y) - \underline{apr}(Y)$. Objects in $BN_R(Y)$ can be explained as these objects those do not definitely belong to Y or $\sim Y$ (namely $U - Y$) according to R. According to the definition, it is uncertain that an object in $BN_R(Y)$ definitely belong to Y even if the object only appear in $BN_R(Y)$ instead of $BN_R(Z)$ ($Z \subseteq U$ and $Z \neq Y$). Therefore, uncertainty in rough set can be explained as follows:

Objects in $BN_R(Y)$ do not definitely belong to Y.
⇕
(1) Objects belong to at least two subsets of U including Y.
(2) Objects do not belong to any other subset of U but the degree to Y objects belong is uncertain.

In rough *k*-means clustering, clusters are represented as interval sets with lower and upper approximations. Let $X = \{\vec{x}_1, \ldots, \vec{x}_n\}$ be the set of objects and $C = \{\vec{c}_1, \ldots, \vec{c}_k\}$ be the set of clusters. The lower approximation and upper approximation of $\vec{c}_i$ is denoted as $\underline{apr}(\vec{c}_i)$ and $\overline{apr}(\vec{c}_i)$, respectively. Accordingly, the boundary region of $\vec{c}_i$ is defined as $bnd(\vec{c}_i) = \overline{apr}(\vec{c}_i) - \underline{apr}(\vec{c}_i)$. It is noted that

Table 1
The value of risk for assigning an object in Fig. 2(b).

The value of risk	(0,0.13]	0.233	0.267
Object	$\vec{x}_i$ $(3 \leqslant i \leqslant 20)$	$\vec{x}_1$	$\vec{x}_2$

uncertainty of objects in the boundary region of clusters is described by assigning objects to more than one cluster. However, according to analysis of ambiguity in rough set, it is not adequate to describe uncertainty of objects for the boundary region in rough k-means clustering. Therefore, interval set representation of clusters in rough k-means clustering need to be improved. It can be explained by the following example.

Example 1(b). As can be seen from Fig. 2(b), $\vec{x_1}$ is in the lower approximation of $\vec{c_1}$. However, $\vec{x_1}$ is rather isolated comparing to other objects in the same lower approximation. Therefore, it is difficult to make sure that $\vec{x_1}$ definitely belong to $\vec{c_1}$ though $\vec{x_1}$ certainly do not belong to $\vec{c_2}$.

Therefore, similar to rough set, uncertainty of objects for rough clustering can be expressed as follows:

An object $\vec{x}_l$ in $bnd(\vec{c}_i)$ does not definitely belong to $\vec{c}_i$.
⇕
(U1) $\vec{x}_l$ belongs to at least two clusters including $\vec{c}_i$.
(U2) $\vec{x}_l$ does not belong to any other cluster but the degree to $\vec{c}_i$ $\vec{x}_l$ belongs is uncertain.

3.3. The value of risk for assigning an object

In practice, objects comparatively isolated in lower approximations have greater values of risk than other objects. It can be explained by Example 1(c) as follows:

Example 1(c). For the good clustering result shown in Fig. 2(b), the value of risk for assigning an object is presented in Table 1. The value of risk for assigning $\vec{x}_1$ or $\vec{x}_2$ is greater than that for assigning any other object. The bigger the value of risk for assigning an object, the worse the assignment is. Therefore, it is not a good decision to assign $\vec{x}_1$ and $\vec{x}_2$ to lower approximations of $\vec{c}_1$ and $\vec{c}_2$, respectively.

Therefore, lower and upper approximations in rough k-mean clustering are not adequate to describe uncertainty of objects. It suggests to discover a solution that is adequate to describe uncertainty of objects in rough k-means clustering.

4. Interval set clustering

Based on the analysis above, one-level lower and upper approximations for rough k-means clustering are not adequate enough to describe uncertainty of objects. According to uncertainty of objects for rough clustering analyzed above, we propose an interval set clustering. Since clusters are represented as interval sets with two-level lower and upper approximations, the proposed algorithm is adequate to describe uncertainty in categorizing objects.

4.1. Definitions and properties

For defining our framework we will assume existence of a hypothetical clustering scheme, CS, that partitions a set of objects $X = \{\vec{x}_1, \ldots, \vec{x}_n\}$ into clusters $CS = \{\vec{c}_1, \ldots, \vec{c}_k\}$ (Lingras et al., 2009) and each cluster is represented by the cluster centroid $\vec{c}_i$ $(1 \leqslant i \leqslant k)$. Clustering algorithms such as k-means approximate the actual clustering. For a given clustering scheme CS, let $b_j(CS, \vec{x}_l)$ be the action that assigns the object $\vec{x}_i$ to a cluster or a group of clusters. The risk associated with the assignment will then be given as $R(b_j(CS, \vec{x}_l)|\vec{x}_l)$. For simplicity, we will use $R(\vec{x}_l)$ as a shorthand for $R(b_j(CS, \vec{x}_l)|\vec{x}_l)$. We will start with formal definitions for interval set clustering.

4.1.1. Set of clusters similar to an object $\vec{x}_l$

For every object, $\vec{x}_l$, we define a non-empty set T_l of all the clusters that are similar to $\vec{x}_l$. Clearly, $T_l \subseteq CS$. We will use $\vec{x}_l \to T_l$ to denote the fact that object $\vec{x}_l$ is similar to all the elements of set T_l.

The definition of the similarity will depend on a given application. For an object $\vec{x}_l$, we can specify T_l given that CS is rough k-means clustering (RC) proposed by Lingras and West (2004) as follows:

$$d(\vec{x}_l, \vec{c}_i) = min_{1 \leqslant t \leqslant k} d(\vec{x}_l, \vec{c}_t), \tag{2}$$

$$T_l = \{t : d(\vec{x}_l, \vec{c}_i)/d(\vec{x}_l, \vec{c}_t) \geqslant threshold \text{ and } i \neq t\}. \tag{3}$$

4.1.2. Object and cluster outer similarity

Given T_l for object $\vec{x}_l$, the outer similarity between object $\vec{x}_l$ and cluster $\vec{c}_i$ is defined as follows:

$$Sim_{outer}(\vec{x}_l, \vec{c}_i) = \begin{cases} \frac{1}{|T_l|}, & \text{if } \vec{c}_i \in T_l, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

4.1.3. Outer upper and lower approximations

If an object $\vec{x}_l$ is assigned to a set T_l, then the object belongs to the outer upper approximations of all clusters $\vec{c}_i \in T_l$. If $-T_l- = 1$, then $\vec{x}_l$ belongs to the outer lower approximation of the only $\vec{c}_i \in T_l$. Please note that when $|T_l| = 1$, $\{\vec{c}_i\} = T_l$. Therefore, outer upper ($\overline{apr}_{outer}$) and outer lower ($\underline{apr}_{outer}$) approximation of each category $\vec{c}_i$ can be defined as follows:

$$\overline{apr}_{outer}(\vec{c}_i) = \{\vec{x}_l |\ \vec{x}_l \to T_l, \vec{c}_i \in T_l\}, \tag{5}$$

$$\underline{apr}_{outer}(\vec{c}_i) = \{\vec{x}_l |\ \vec{x}_l \to T_l, \{\vec{c}_i\} = T_l\}. \tag{6}$$

We cannot test all the properties of rough set theory. However, it can be easily shown that the outer upper and lower approximations in fact follow important rough set theoretic properties.

(P1) An object can be part of at most one outer lower approximation
(P2) $\vec{x}_l \in \underline{apr}_{outer}(\vec{c}_i) \Rightarrow \vec{x}_l \in \overline{apr}_{outer}(\vec{c}_i)$
(P3) An object $\vec{x}_l$ is not part of any outer lower approximation
⇕
$\vec{x}_l$ belongs to two or more outer upper approximations

Accordingly, the boundary region of $\vec{c}_i$ is defined as follows:

$$bnd_{outer}(\vec{c}_i) = \overline{apr}_{outer}(\vec{c}_i) - \underline{apr}_{outer}(\vec{c}_i). \tag{7}$$

It can be easily shown that uncertainty of objects in the outer boundary region in fact follow (U1) in Section 3.2. For an object $\vec{x}_l$, uncertainty of $\vec{x}_l$ with respect to an cluster $\vec{c}_i$, denoted as $U_{outer}(\vec{x}_l, \vec{c}_i)$, can be expressed as follows:

$$U_{outer}(\vec{x}_l, \vec{c}_i) \propto \frac{1}{Sim_{outer}(\vec{x}_l, \vec{c}_i)}. \tag{8}$$

$U_{outer}(\vec{x}_l, \vec{c}_i)$ describes the probability that $\vec{x}_l$ belongs to any other cluster instead of $\vec{c}_i$. However, $Sim_{outer}(\vec{x}_l, \vec{c}_i)$ defines the probability that $\vec{x}_l$ only belong to $\vec{c}_i$. Hence, the bigger the outer similarity between $\vec{x}_l$ and $\vec{c}_i$, the smaller uncertainty between $\vec{x}_l$ and $\vec{c}_i$ is.

Objects in the outer lower approximations of clusters are confined to belong to only one cluster. In order to have a close look at the degree to a cluster such an object belongs, we define the

inner upper and lower approximations for each cluster. A cluster with the inner upper and lower approximations make a clear line between objects that are totally similar to the cluster with those that are not. We begin with the notations of the *uniform objects*.

4.1.4. Uniform objects

Let $\vec{x}_l$ be an object, $\vec{x}_l$ is an uniform object if $\vec{x}_l$ belongs to at most one cluster. Given T_l for $\vec{x}_l$, uniform object $\vec{x}_l$ can be defined as follows:

$$\vec{x}_l \in \underline{apr}_{outer}(\vec{c}_i) \quad \text{and} \quad T_l = \{\vec{c}_i\}. \tag{9}$$

4.1.5. Uniform h-distance of an uniform object $\vec{x}_l$

Let $\vec{x}_l$ and $\vec{x}_t$ be two uniform objects and T_l is the set of all the clusters that are similar to $\vec{x}_l$. For any positive integer h, the uniform h-distance of $\vec{x}_l$, denoted as h-distance $(\vec{x}_l)$, is defined as the distance $d(\vec{x}_l, \vec{x}_t)$ between $\vec{x}_l$ and $\vec{x}_t$ such that:

(i) $\vec{x}_t \in \underline{apr}_{outer}(\vec{c}_i)$ and $T_l = \{\vec{c}_i\}$
(ii) For at least h objects $\vec{x}_t \in X$ $(1 \leqslant t \leqslant n \wedge t \neq i)$ it holds that $d(\vec{x}_i, \vec{x}_t) \leqslant d(\vec{x}_i, \vec{x}_j)$, and
(iii) For at most $h-1$ objects $\vec{x}_t \in X$ $(1 \leqslant t \leqslant n \wedge t \neq i)$ it holds that $d(\vec{x}_i, \vec{x}_t) < d(\vec{x}_i, \vec{x}_j)$.

4.1.6. Uniform h-distance neighborhood of an uniform object $\vec{x}_l$

Given the *h-distance* of an uniform object $\vec{x}_l$, the h-distance neighborhood of $\vec{x}_l$ contains every uniform object whose distance from $\vec{x}_l$ is not greater than h-distance $(\vec{x}_l)$, i.e., $N_{h-distance(\vec{x}_l)}(\vec{x}_l) = \{\vec{x}_t \in \vec{c}_i |\ d(\vec{x}_l, \vec{x}_t) \leqslant h - distance(\vec{x}_l) \wedge t \neq l \ \wedge\ T_l = \{\vec{c}_i\}\}$. These uniform objects $\vec{x}_t$ are called the h-nearest neighbors of $\vec{x}_l$.

Whenever no confusion arises, the notation $N_{h-distance(\vec{x}_l)}(\vec{x}_l)$ is simplified as $N_h(\vec{x}_l)$. Note that the $h - distance(\vec{x}_l)$ is well defined for any positive integer h. The cardinality of $N_h(\vec{x}_l)$ is no less than h. Based on the notations above, we give the following definitions to discover objects that are different from their local neighborhoods according to the values of risks.

4.1.7. Reachable risk of an uniform object $\vec{x}_l$ w.r.t. $N_h(\vec{x}_l)$

Let $R(\vec{x}_i)$ be the risk for assigning $\vec{x}_l$ to clusters. For any integer h, the *reachable risk* of object $\vec{x}_l$ with respect to $N_h(\vec{x}_l)$ is defined as

$$reach\text{-}risk_h(\vec{x}_l) = max\{R(\vec{x}_t), R(\vec{x}_l)\} \quad (\vec{x}_t \in N_h(\vec{x}_l)). \tag{10}$$

Note that the reachable risk of an object $\vec{x}_i$ is defined based on the risk for assigning an object in $N_h(\vec{x}_i)$. Because the risk for assigning an object is obtained assuming that the conditional probability $P(\vec{c}_t|\vec{x}_i)$ is proportional to the similarity between the object and $core(\vec{c}_t)$. We can get the following property:

Property 1. If two objects $\vec{x}_i$ and $\vec{x}_j$ are similar, the value of $R(\vec{x}_i)$ is close to that of $R(\vec{x}_j)$.

Since the higher the value of the risk, the worse the assignment is in the clustering scheme. If any object $\vec{x}_j$ in $N_h(\vec{x}_i)$ is far away from object $\vec{x}_i$, then the reachable risk of object $\vec{x}_i$ is quite different from that of $\vec{x}_j$. However, if they are sufficiently close, the reachable risk of object $\vec{x}_i$ is close to that of $\vec{x}_j$. The reason is that in doing so, the statistical fluctuations of risk for all the $\vec{x}_j$'s close to $\vec{x}_i$ can be significantly reduced. The strength of this smoothing effect can be controlled by the parameter h.

4.1.8. Local reachable risk of an object $\vec{x}_i$

Given a positive integer h, the *local reachability risk* of $\vec{x}_i$ based on h-nearest neighborhoods of $\vec{x}_i$ is defined as

$$lrr_h(\vec{x}_i) = \frac{\sum_{\vec{x}_j \in N_h(\vec{x}_i)} reach\text{-}risk_h(\vec{x}_j)}{|N_h(\vec{x}_i)|}. \tag{11}$$

The local reachable risk of an object $\vec{x}_i$ is the average reachable risk based on the h-nearest neighborhoods of $\vec{x}_i$. According to Property 1, the local reachable risk of an object $\vec{x}_i$ may be close to $R(\vec{x}_i)$. The more suitable the value of h, the more similar the reachable risks for objects within the same neighborhood are.

4.1.9. Local uniform factor of an object $\vec{x}_i$ based on reachable risk

Given a positive integer h, the *local uniform factor* of $\vec{x}_i$ based on reachable risk is defined as

$$LUF_h(\vec{x}_i) = \frac{\sum_{\vec{x}_j \in N_h(\vec{x}_i)} \frac{lrr_h(\vec{x}_j)}{lrr_h(\vec{x}_i)}}{|N_h(\vec{x}_i)|}.$$

The uniform factor of object $\vec{x}_i$ captures the degree to which $\vec{x}_i$ uniform with its neighborhoods based on the value of risk. It is the average of the ratio of the local reachable risk of $\vec{x}_i$ and those of $\vec{x}_i$'s h-nearest neighbors. It is easy to see that the lower $\vec{x}_i$'s local reachable risk is, and the higher the local reachable risk of $\vec{x}_i$'s h-nearest neighbors, the greater the LUF value of $\vec{x}_i$ is. We can get the following property of LUF.

Property 2. The closer to 1 the value of $LUF(\vec{x}_i)$, the more the distribution of $\vec{x}_i$ is uniform with the objects in its neighborhood.

The interpretation of Property 2 is as follows. Let $\vec{c}_t$ be a cluster. Let *reach-risk-min* denote the minimum reachable risk of objects in $\vec{c}_t$, i.e., $reach\text{-}risk\text{-}min = min\{reach\text{-}risk(\vec{x}_i)|\vec{x}_i \in \vec{c}_t\}$. Similarly, let *reach-risk-max* denote the maximum reachable risk of objects in $\vec{c}_t$. Then the local reachable risk of $\vec{x}_i$, as the definition, is $\leqslant$ *reach-risk-max* and $\geqslant$ *reach-risk-max*. Let e be defined as (*reach-risk-max*/*reach-risk-min* -1). By the definition of LUF, we have $reach\text{-}risk\text{-}min/reach\text{-}risk\text{-}max \leqslant LUF(\vec{x}_i) \leqslant reach\text{-}risk\text{-}max/reach\text{-}risk\text{-}min$. Hence, for all objects $\vec{x}_i \in \vec{c}_t$, such that $1/(1+e) \leqslant LUF(\vec{x}_i) \leqslant (1+e)$. Let us consider that $\vec{c}_t$ is a tight cluster and $\vec{x}_i$ is deep inside the cluster. It forces the LUF of $\vec{x}_i$ to be close to 1.

According to Property 2, the value of $\|LUF_h(\vec{x}_i) - 1\|$ is used to capture the degree to which $\vec{x}_i$ uniform with its neighborhoods. To return to Example 1, summary of $\|LUF_4(\vec{x}_i) - 1\|$ for objects belong to the lower approximations in Fig. 2(b) is shown in Table 2. Both the value of $\|LUF_4(\vec{x}_1) - 1\|$ and that of $\|LUF_4(\vec{x}_2) - 1\|$ are comparatively high and greater than the value of $\|LUF_4(\vec{x}_i) - 1\|$ $(3 \leqslant i \leqslant 20)$. Hence it is not a good decision to take the action of assigning $\vec{x}_1$ and $\vec{x}_2$ to the lower approximations of $\vec{c}_1$ and $\vec{c}_2$, respectively.

LUF have the characteristic of focusing on the evaluation of the action of assigning an object to clusters in a clustering scheme. In comparison with rough k-means clustering, interval set clustering constructs satisfactory lower and upper approximations by taking advantage of the characteristic of LUF.

4.2. Interval set clustering

Based on the definitions above, interval set clustering proceeds as follows:

Table 2
Summary of $\|LUF_4(\vec{x}_i) - 1\|$ for objects in Fig. 2(b).

$\|LUF_4(\vec{x}_i) - 1\|$	(0,0.1]	(0.1,0.2]	(0.2,0.3]	(0.3,0.4]	(0.4,0.5]	(0.5,+∞)
Number of objects	7	6	2	1	0	2
Object	$\vec{x}_i$ $(3 \leqslant i \leqslant 20)$					$\vec{x}_1$, $\vec{x}_2$

Input: Data set X, the number of clusters (k), ω_l, *threshold*, h and $threshold_h$.

Step 1: Get original clusters with lower approximations and upper approximations denoted as follows:

$$OL(C) = \left\{\underline{apr'}(\vec{c}_1)\cdots\underline{apr'}(\vec{c}_k)\right\},$$
$$OU(C) = \{\overline{apr'}(\vec{c}_1)\cdots\overline{apr'}(\vec{c}_k)\}.$$

- Step 1.1: Perform rough k-means clustering for different selections of parameters.
- Step 1.2: Evaluate rough k-means clustering by rough cluster quality index for the best selection of parameters k, ω_l and *threshold*.

Step 2: Calculate LUF for objects in lower approximations according to the value of h.

Step 3: Determine the new lower approximations and upper approximations by considering LUF of objects in lower approximations where $threshold_h$ is a given value:

$$\underline{apr}(\vec{c}_i) = \left\{\vec{x}_i|\vec{x}_i \in \underline{apr'}(\vec{c}_i) \wedge LUF_h(\vec{x}_i) \leqslant threshold_h\right\},$$
$$\overline{apr}(\vec{c}_i) = \overline{apr'}(\vec{c}_i) \cup \left(\underline{apr'}(\vec{c}_i) - \underline{apr}(\vec{c}_i)\right).$$

Step 4: Obtain satisfactory clusters with lower approximations and upper approximations denoted as follows:

$$L(C) = \{\underline{apr}(\vec{c}_1)\cdots\underline{apr}(\vec{c}_k)\},$$
$$U(C) = \{\overline{apr}(\vec{c}_1)\cdots\overline{apr}(\vec{c}_k)\}.$$

Step 5: According to Eq. (1), we calculate the new centroids for clusters.

The properties of lower approximations and upper approximations in interval set clustering can be obtained as follows:

(P1) An object can be a member of at most one lower approximation.
(P2) $\vec{x}_l \in \underline{apr}(\vec{c}_i) \Rightarrow \vec{x}_l \in \overline{apr}(\vec{c}_i)$.
(P3) An object that does not belong to any lower approximation is member of at least one upper approximation.

Interval set clustering has the same properties (P1) and (P2) as rough k-means clustering. However, interval set clustering has the property (P3) that an object that does not belong to any lower approximation may belong to one or more than one upper approximation. It exactly match the properties of upper approximations in rough sets.

We perform interval set clustering on the data in Example 1. Assuming the clustering result of rough k-means clustering is shown in Fig. 2(b). The final clusters in interval set clustering with $h = 4$ and $threshold_h = 0.4$ are shown in Fig. 3. In the figure, the dashed line outlines the upper approximation of each cluster, and the solid line describes the lower approximation of each cluster. It shows that $\vec{x}_1$ and $\vec{x}_2$ are assigned to the upper approximations of $\vec{c}_1$ and $\vec{c}_2$ respectively in interval set clustering. However, both of the two objects are members of lower approximations in rough k-mean clustering. It matches the fact that $\vec{x}_1$ and $\vec{x}_2$ are farther away from the centroids of clusters they belong to than any other objects in the lower approximations of their corresponding clusters. Hence interval set clustering constructs more satisfactory lower and upper approximations than rough k-means clustering.

5. Study data and design of the experiment

We apply interval set clustering to three kinds of data sets, synthetic data set, a standard data set and a retail store's data set, to demonstrate how to construct clusters with more satisfactory lower and upper approximations. We design the experiments on each data set as follows:

Step 1. Perform crisp clustering with different number of clusters on the data set. Determine the optimal number of clusters according to the change in rough cluster quality index.
Step 2. Perform rough clustering with the optimal number of clusters with different *threshold* on the data set. Determine the optimal *threshold* according to the change in rough cluster quality index. Moreover, we discover how rough cluster quality index varies for different number of clusters with the optimal *threshold*. Note that we set ω_l at a value of 0.75 in the experiments.
Step 3. Present the number of objects for different $\|LUF_h(\vec{x}_i) - 1\|$ in the table to find the optimal values for h and $threshold_h$. Note that we repeat the experiment five times to get the average number of objects with the same parameters for each h.

Because rough cluster quality index is proved to find the optimal values for the number of clusters and *threshold* in rough clustering (Lingras et al., 2009). We focus on Step 3 to discover satisfactory lower and upper approximations in interval set clustering.

5.1. Synthetic data

In order to visualize the data, we limit the synthetic data to two-dimensional space. As can be seen from Fig. 4, there are 68 objects and three clusters can be identified in the figure. It seems that eight objects do not belong to any particular cluster. We performed crisp clustering and rough clustering on the synthetic data set.

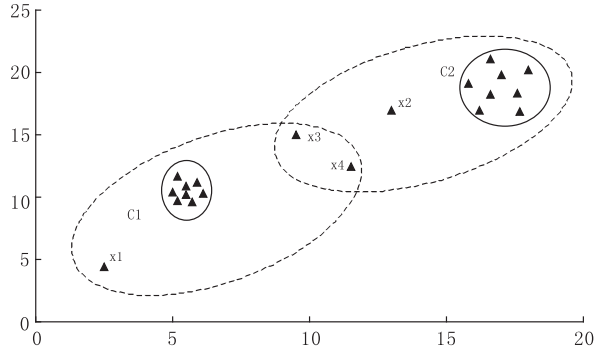

Fig. 3. Interval set clustering.

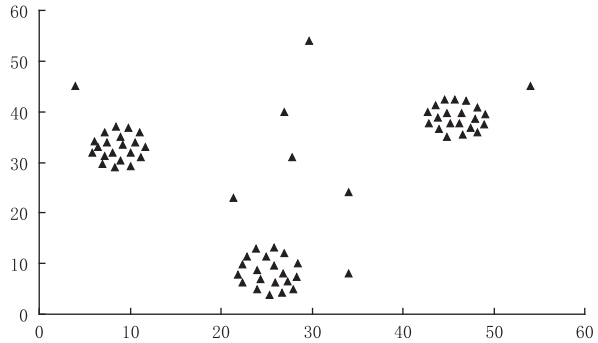

Fig. 4. Synthetic data.

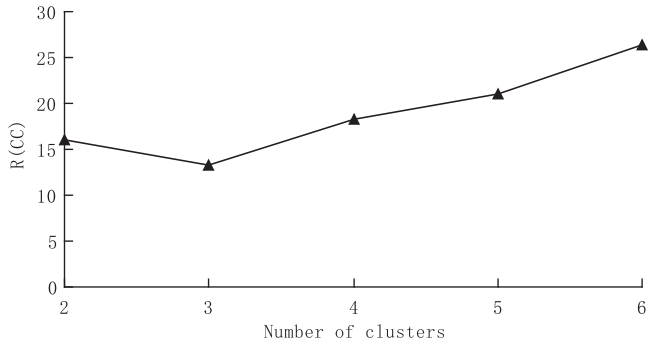

Fig. 5. Synthetic data: cluster index for crisp clustering for different number of clusters.

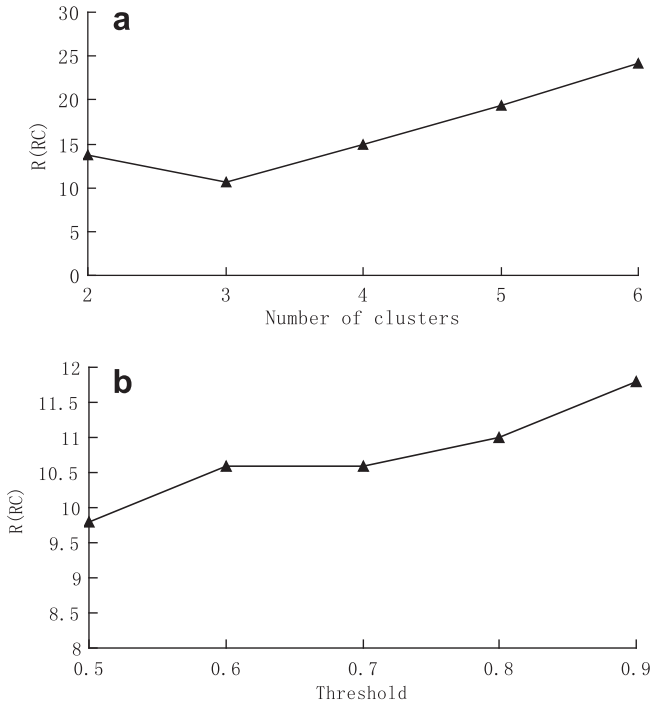

Fig. 6. (a) Synthetic data: cluster index for rough clustering for different number of clusters. (b) Synthetic data: change in cluster index for rough clustering with threshold.

We perform crisp clustering and rough clustering on the synthetic data set. Fig. 5 shows how the cluster index varies for different number of clusters for crisp clustering. Similar trend can also be found for rough clustering in Fig. 6(a) with *threshold* = 0.7. Fig. 6(b) shows how changing the value of *threshold* can affect the index of clustering with $k = 3$. Similar to the analysis in Lingras et al. (2009), it is reasonable to use the value of *threshold* = 0.7 and set $k = 3$. A satisfactory rough clustering with $k = 3$, $\omega_l = 0.75$ and *threshold* = 0.7 is presented in Fig. 7(a).

In order to discover the optimal values for h and $Threshold_h$, the change in the number of objects in lower approximations for $\|LUF_h(\vec{x}_i) - 1\|$ with different h is presented in Table 3. Because the value of h can not be too high or too small, the h values were changed from 2 to 10 in the table. When the h reaches a value of 5 or 6, three objects in lower bounds have very high values of $\|LUF_h(\vec{x}_i) - 1\|$ that is greater than 0.4. However, the values of $\|LUF(\vec{x}_i) - 1\|$ for the other 60 objects in lower approximations are small and not greater than 0.2. It indicates a great change in the number of objects in lower approximations with different $\|LUF_h(\vec{x}_i) - 1\|$ when $h = 5$ or $h = 6$. Comparatively, more objects

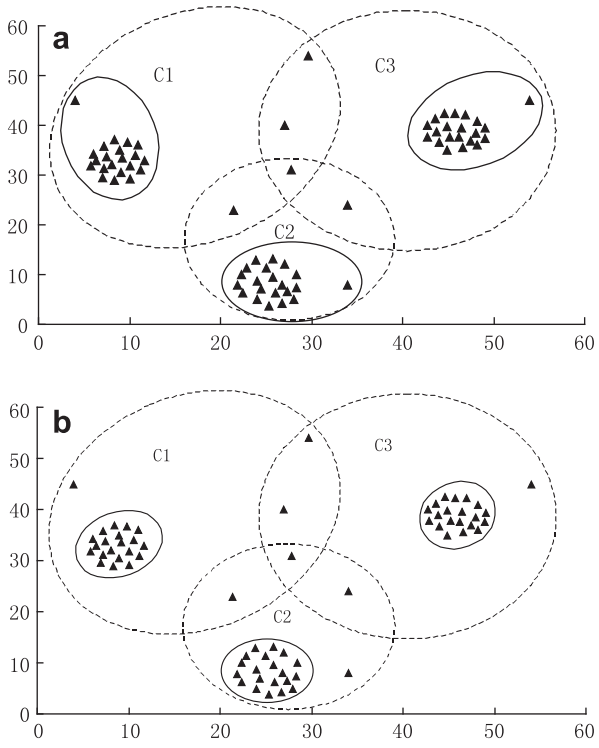

Fig. 7. (a) Synthetic data: rough clustering. (b) Synthetic data: interval set clustering.

Table 3
Synthetic data: the number of objects in lower approximations for $\|LUF_h(\vec{x}_i) - 1\|$ with $k = 3$ and *threshold* = 0.7.

h	$\|LUF_h(\vec{x}_i) - 1\|$					
	(0,0.1]	(0.1,0.2]	(0.2,0.3]	(0.3,0.4]	(0.4,0.5]	(0.5,+∞)
2	32	15	4	3	1	8
3	36	20	0	2	1	4
4	44	12	2	0	1	3
5	53	7	0	0	1	2
6	50	10	0	0	1	2
7	54	6	0	1	1	1
8	51	8	1	1	1	1
9	54	6	0	1	1	1
10	53	7	0	1	1	1

have the value of $\|LUF_h(\vec{x}_i) - 1\|$ that is close to 0 when $h = 5$. Therefore, the optimal values for h is 5 and it is reasonable to use a value of $threshold_h$ between 0.2 and 0.4. The interval set clustering with $k = 3$, $h = 5$ and $threshold_h = 0.3$ is presented in Fig. 7(b). It shows that three objects were changed into the members of lower approximations of the clusters they belong to.

5.2. Wisconsin breast cancer data

In this section, we use a standard real world data set that is tested for clustering by other researchers such as Xie, Raghavan, Dhatric, and Zhao (2005). Wisconsin breast cancer databases were obtained from the University of Wisconsin Hospitals, Madison by Mangasarian and Wolberg (1990). This data set contains 699 instances that fall into two classes: benign (458 instances) and malignant (241 instances). Each instance is represented by nine attributes, all of which are scaled to a [1,10] range. However, 16 instances have attributes that have missing values. After eliminating the 16 instances, the number of instance was 683.

Fig. 8 shows the variation in rough cluster quality index as we change the number of clusters in the k-means crisp clustering. Sim-

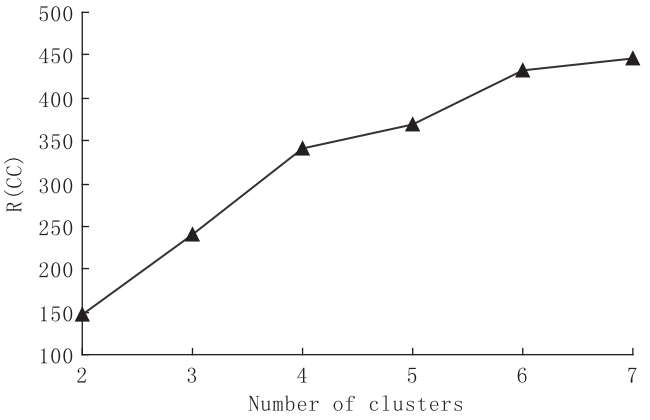

Fig. 8. Breast cancer data: cluster index for crisp clustering for different number of clusters.

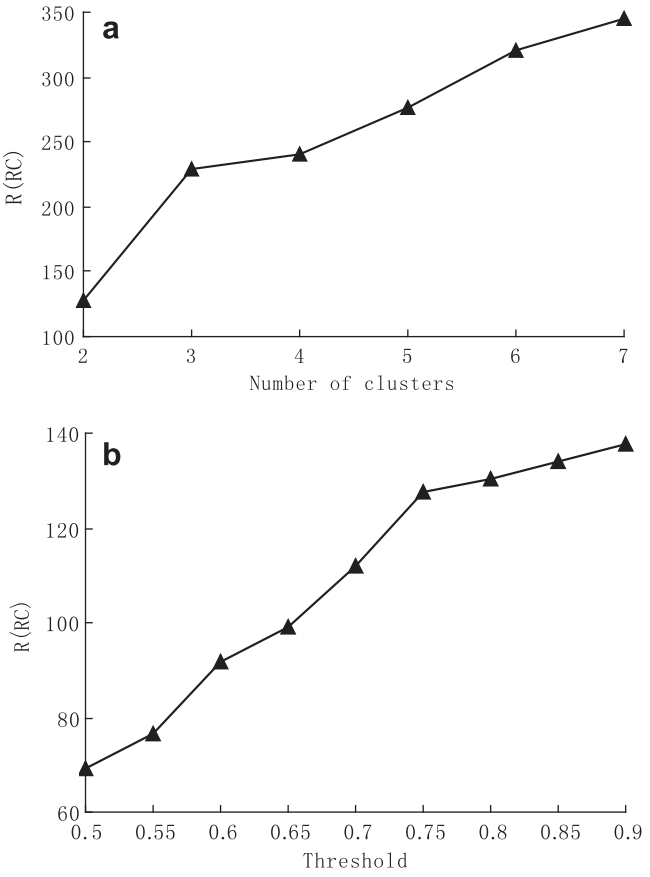

Fig. 9. (a) Breast cancer data: cluster index for rough clustering for different number of clusters. (b) Breast cancer data: change in cluster index for rough clustering with threshold

ilar to the analysis in Lingras et al. (2009), it suggests that two can be an optimal value for k. The variation in quality index for different values of threshold is shown in Fig. 9(b) with $k = 2$. There is a sharp drop in quality index when the *threshold* is reduced from 0.75 to 0.5. Therefore, *threshold* = 0.75 can be used as an appropriate value. The changes in quality index for different number of clusters with *threshold* = 0.75 is shown in Fig. 9(a). It shows similar trend with Fig. 8.

As discussed above, 683 instances are grouped into two clusters. The *threhsold* in rough clustering was set at 0.75. In order to discover the optimal values of h and $Threshold_h$, we have a close look at the values of $\|LUF_h(\vec{x}_i) - 1\|$ for instances in lower approximations. The change in the number of objects in lower approximations for $\|LUF(\vec{x}_i) - 1\|$ with different h is presented in Table 4. Because the value of h can not be too high or too small, the h values were changed from 5 to 30 in the table. We can see a rapid decline in the number of objects when the value of $\|LUF_h(\vec{x}_i) - 1\|$ increases from the values in the range of $(0,0.1]$ to those in the range of $(0.1,0.2]$. The rapid decline is obvious when the value of h is greater than 5. However, there is not too much difference on the change of the number of clusters with different h values. Therefore, the selection of $threshold_h$ is flexible. We can choose a value that is greater than 5, such as 10, for h. $threshold_h$ can be set at a value in the range of $(0.2,0.5]$ depending on the requirements. An interval set clustering result with $h = 10$ and $Threshold_h = 0.1$ is presented in Table 5. It shows the number of objects in lower approximation, upper

Table 4
Breast cancer data: the number of objects in lower approximations for $\|LUF_h(\vec{x}_i) - 1\|$ with $k = 2$ and *threshold* = 0.75.

h	$\|LUF_h(\vec{x}_i) - 1\|$					
	$(0,0.1]$	$(0.1,0.2]$	$(0.2,0.3]$	$(0.3,0.4]$	$(0.4,0.5]$	$(0.5,+\infty)$
5	492	104	42	8	3	0
10	482	83	63	17	3	1
15	475	91	48	26	7	2
20	470	94	47	22	14	2
25	477	82	55	18	14	3
30	481	74	57	17	17	3

Table 5
Breast cancer data: the number of objects in interval set clustering with $h = 10$ and $threshold_h = 0.1$.

Area	$\vec{c}_1$	$\vec{c}_2$
Lower approximation	286	196
Upper approximations	475	242
Re-assigned objects in upper approximations	155	12
Boundary region	189	46

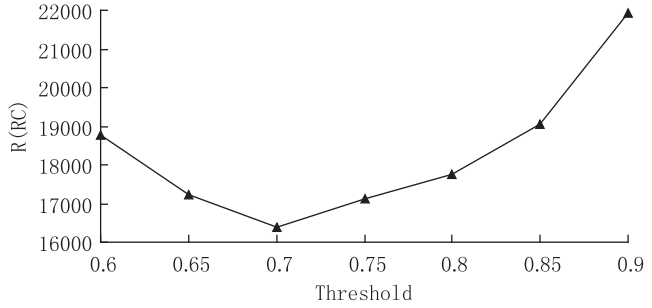

Fig. 10. Retail data: change in cluster index for interval set clustering with threshold.

Table 6
The retail data: the number of objects in lower approximations for $\|LUF_h(\vec{x}_i) - 1\|$ with $k = 7$ and *threshold* = 0.7.

h	$\|LUF_h(\vec{x}_i) - 1\|$					
	$(0,0.1]$	$(0.1,0.2]$	$(0.2,0.3]$	$(0.3,0.4]$	$(0.4,0.5]$	$(0.5,+\infty)$
5	4645	29	4	4	0	3
10	4622	43	4	3	0	1
15	4621	41	5	2	2	1
20	4636	50	7	1	1	1
25	4476	36	4	3	0	1
30	4570	45	5	2	0	1

Table 7
The retail data: the number of objects in interval set clustering with $h = 10$ and $threshold_h = 0.1$.

Area	$\vec{c}_1$	$\vec{c}_2$	$\vec{c}_3$	$\vec{c}_4$	$\vec{c}_5$	$\vec{c}_6$	$\vec{c}_7$
Lower approximations	3323	754	332	99	66	35	12
Upper approximations	3899	1515	673	233	201	101	34
Re-assigned objects in upper approximations	6	9	5	6	13	5	7

approximation and boundary region for each cluster. The re-assigned objects illustrates the number of objects in each cluster just belong to one upper approximation. Therefore, 155 objects are just in the upper approximation of $\vec{c}_1$ though they partly belong to $\vec{c}_1$. And it is also difficult to make sure that 12 objects definitely belong to $\vec{c}_2$ though they just in the upper approximation of $\vec{c}_2$.

5.3. The retail data

The real data comes from a real world data set belonging to a small retail chain. Lingras et al. described the real data in detail in Lingras et al. (2009). Use of rough cluster quality index measure to derive an appropriate clustering scheme for a promotional campaign in a retail store was also proposed in Lingras et al. (2009). Moreover, a two-tier promotional campaign targeted at the customers in the first and second highest spending clusters is provided in detail based on the modified loss function for the action b_j. The loss function for all the actions b_j such that $\vec{c}_k \in b_j$ is as follows:

$$\lambda_{\vec{x}_i}(b_k|\vec{c}_i) = (\$100 - S_l \times 1.1 \times 0.3) \times \frac{|b_k - T_l|}{|b_k|} \quad \text{if } \vec{c}_i \in b_k; \tag{12}$$

$$\lambda_{\vec{x}_i}(b_k|\vec{c}_i) = (\$100 - S_l \times 1.1 \times 0.3) \times \frac{|b_k - \emptyset|}{|b_k|} \quad \text{if } \vec{c}_i \notin b_k. \tag{13}$$

The loss function for all the actions b_j such that $\vec{c}_{k-1} \in b_j$ and $\vec{c}_k \notin b_j$ is modified as follows:

$$\lambda_{\vec{x}_i}(b_{k-1}|\vec{c}_i) = (\$50 - S_l \times 1.05 \times 0.3) \times \frac{|b_{k-1} - T_l|}{|b_{k-1}|} \quad \text{if } \vec{c}_i \in b_{k-1}; \tag{14}$$

$$\lambda_{\vec{x}_i}(b_{k-1}|\vec{c}_i) = (\$50 - S_l \times 1.05 \times 0.3) \times \frac{|b_{k-1} - \emptyset|}{|b_{k-1}|} \quad \text{if } \vec{c}_i \notin b_{k-1}. \tag{15}$$

The loss functions for the remaining actions b_j that do not assign customers to either $\vec{c}_k$ or $\vec{c}_{k-1}$ remain unchanged.

According to the experiments in Lingras et al. (2009), it is reasonable to set the number of clusters to be between five and seven. The number of clusters is set at the value of seven in our experiments. Because we use the *threshold* that is no more than one in rough clustering, the optimal value of *threshold* need to be re-decided in our experiments. Fig. 10 shows the variation in rough cluster quality index as *threshold* changes from 0.6 to 0.9 for $k = 7$. When *threshold* is at a value of 0.7, there is a local minima suggesting that 0.7 is a reasonable value.

The change in the number of objects in lower approximations for $\|LUF(\vec{x}_i) - 1\|$ with different h is presented in Table 6. The h values were changed from 5 to 30 in the table. We can see a rapid decline in the number of objects when the value of $\|LUF_h(\vec{x}_i) - 1\|$ increases from the values in the range of $(0, 0.1]$ to those in the range of $(0.1, 0.2]$. The rapid decline suggests that 0.1 is a reasonable value for $threshold_h$. However, there is not too much difference on the change of the number of objects for different values of h. Therefore, the selection of h is flexible. We can choose a value that is greater than 5, such as 10, for h. An interval set clustering result with $h = 10$ and $Threshold_h = 0.1$ is presented in Table 7. The re-assigned objects illustrates the number of objects in each cluster just belong to one upper approximation. Therefore, interval set clustering improves rough k-means clustering with more satisfactory interval set representation of clusters.

6. Summary and conclusions

This paper describes an interval set clustering algorithm based on rough k-means clustering and rough cluster quality index. The proposal generates original clusters with lower approximations and upper approximations by taking rough clustering as the first step. Rough cluster quality index is then used to adjust lower approximation and upper approximation for each cluster by selecting the optimal number of clusters and the appropriate value for *threshold*. Because LUF have the characteristic of focusing on the evaluation of the action of assigning an object to clusters in a clustering scheme. Interval set clustering adjusts the assignment of objects in lower approximations by considering the LUF for each objects. An objects in the lower approximation with the LUF value that is not close to one may be not a member of the lower approximation. We assign such an object to the boundary region instead of the lower approximation of the cluster. Therefore, interval set clustering can obtain satisfactory clusters with lower and upper approximations so that an object that does not belong to any lower approximation may be a member of only one upper approximation. The proposal was successfully test in three experiments (synthetic data, a standard data and the retail data). Further work will focus on the evaluation of interval set clustering and its comparison with rough clustering and other clustering scheme.

Acknowledgment

I am greatly thankful to Dr. Lingras for his support and guidance during my research and writing this article.

References

Asharaf, S., Shevade, S. K., & Murty, N. M. (2005). Rough support vector clustering. *Pattern Recognition, 38*(10), 1779–1783.

Bezdek, J. C., & Pal, N. R. (1995). Cluster validation with generalized Dunn's indices. In N. Kasabov & G. Coghill (Eds.), *Proceedings of the 1995 second NZ international two-stream conference on ANNES*. Piscataway, NJ: IEEE Press.

Bezdek, J. C., & Pal, N. R. (1998). Some new indexes of cluster validity. *IEEE Transactions on Systems, Man, and Cybernetics, Part-B, 28*, 301–315.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 1*, 224–227.

Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics, 3*, 32–57.

Dunn, J. C. (1974). Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics, 4–1*, 95–104.

Hirano, S., & Tsumoto, S. (2005). On constructing clusters from non-euclidean dissimilarity matrix by using rough clustering. In *JSAI workshops, June 13–14, Kitakyushu city, Japan* (pp. 5–16).

Lingras, P. (2007). Applications of rough set based *K*-means, Kohonen, GA clustering, *Transactions on rough sets VII* (pp. 120–139).

Lingras, P., Chen, M., & Miao, D. Q. (2008). Rough multi-category decision theoretic framework. *Rough Sets and Knowledge Technology*.

Lingras, P., Chen, M., & Miao, D. Q. (2009). Rough cluster quality index based on decision theory. *IEEE Transactions on Knowledge and Data Engineering, 21*(7), 1014–1026.

Lingras, P., Hogo, M., & Snorek, M. (2004). 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.

Lingras, P., & West, C. (2004). Interval set clustering of web users with rough *K*-means. *Journal of Intelligent Information System, 23*(1), 5–16.

Mangasarian, O. L., & Wolberg, W. H. (1990). Cancer diagnosis via linear programming. *SIAM News, 23*(5).
Nguyen, H. S. (2007). Rough document clustering and the internet. *Handbook on Granular Computing*.
Pawlak, Z. (1982). Rough sets. *International Journal of Information and Computer Sciences, 11*, 145–172.
Pawlak, Z. (1984). Rough classification. *International Journal of Man–Machine Studies, 20*, 469–483.
Pawlak, Z. (1992). *Rough sets: Theoretical aspects of reasoning about data*. Kluwer Academic Publishers.
Pawlak, Z., Wong, S. K. M., & Ziarko, W. (1988). Rough sets: Probabilistic versus deterministic approach. *International Journal of Man–Machine Studies, 29*, 81–95.
Peters, G. (2006). Some refinements of rough *k*-means. *Pattern Recognition, 39*(8), 1481–1491.
Peters, J. F., Skowron, A., Suraj, Z., Rzasa, W., & Borkowski, M. (2002). Clustering: A rough set approach to constructing information granules. In *Proceedings of sixth international conference soft computing and distributed processing* (pp. 57–61).
Xie, Y., Raghavan, V. V., Dhatric, P., & Zhao, X. (2005). A new fuzzy clustering algorithm for optimally finding granular prototypes. *International Journal of Approximate Reasoning, 40*, 109–124.
Yao, Y. Y. (2003). Information granulation and approximation in a decision-theoretical model of rough sets. In L. Polkowski, S. K. Oal, & A. Skowron (Eds.), *Rough-neuro computing: Techniques for computing with words* (pp. 491–516). Berlin: Springer.
Yao, Y. Y. (2007). Decision-theoretic rough set models. In J. T. Yao, P. Lingras, W. Z. Wu, M. S. Szczuka, N. Cercone, & D. Slezak (Eds.), *RSKT 2007. LNCS* (Vol. 4481, pp. 1–12). Heidelberg: Springer.

Further Reading

Breunig, M. M., Kriegel, H., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on management of data, SIGMOD00, Dallas, Texas, United States, May 15–18, 2000* (pp. 93–104). New York: ACM Press.