Rough Cluster Quality Index Based on Decision Theory

Pawan Lingras, *Member*, *IEEE*, Min Chen, and Duoqian Miao

Abstract—Quality of clustering is an important issue in application of clustering techniques. Most traditional cluster validity indices are geometry-based cluster quality measures. This paper proposes a cluster validity index based on the decision-theoretic rough set model by considering various loss functions. Experiments with synthetic, standard, and real-world retail data show the usefulness of the proposed validity index for the evaluation of rough and crisp clustering. The measure is shown to help determine optimal number of clusters, as well as an important parameter called *threshold* in rough clustering. The experiments with a promotional campaign for the retail data illustrate the ability of the proposed measure to incorporate financial considerations in evaluating quality of a clustering scheme. This ability to deal with monetary values distinguishes the proposed decision-theoretic measure from other distance-based measures. The proposed validity index can also be extended for evaluating other clustering algorithms such as fuzzy clustering.

Index Terms—Cluster validity, decision theory, loss functions, rough-set-based clustering, k-means clustering.

1 Introduction

Clustering is one of the important techniques in data mining. Clustering categorizes unlabeled objects into several clusters such that the objects belonging to the same cluster are more similar than those belonging to different clusters. Conventional clustering assigns an object to exactly one cluster. Rough-set-based variation makes it possible to assign objects to more than one cluster [1], [11], [12], [17], [19], [21], [23], [25], [33]. Quality of clustering is an important issue in application of clustering techniques to real-world data. A good measure of cluster quality will help in deciding various parameters used in clustering algorithms. One such parameter that is common to most clustering algorithms is the number of clusters.

Many different indices of cluster validity have been proposed. In general, indices of cluster validity fall into one of three categories. Some validity indices measure partition validity to evaluate the properties of crisp structure imposed on the data by the clustering algorithm, such as Dunn indices [7] and Davies-Bouldin index [6]. These validity indices are based on similarity measure of clusters whose bases are the dispersion measure of a cluster and the cluster dissimilarity measure. In the case of fuzzy clustering algorithms, some validity indices such as partition coefficient [2] and classification entropy [2] use only the information of fuzzy membership grades to evaluate clustering results. The third category consists of validity indices that make use of not only the fuzzy membership grades but also the structure of the data. All these validity indices are essentially based on the geometric characteristics of the clusters. This paper proposes a decision-theoretic measure of cluster quality. Decision-theoretic framework has been helpful in providing a better understanding of classification models [14], [39], [41], [43], [44]. The decision theoretic rough set model considers various classes of loss functions. By adjusting loss functions, the decision-theoretic rough set model can also be extended to the multicategory problem. It is possible to construct a cluster validity index by considering various loss functions based on decision theory. Such a measure has an added advantage of being applicable to rough-set-based clustering.

This paper describes how to develop a cluster validity index from the decision-theoretic rough set model. Based on the decision theory, the proposed rough cluster validity index is taken as a function of total risk for grouping objects using a clustering algorithm. Since crisp clustering is a special case of rough clustering, the proposed index validity is applicable to both rough clustering and crisp clustering. Experiments with synthetic and real-world data show the usefulness of the proposed validity index for the evaluation of rough clustering and crisp clustering. The measure is shown to help determine optimal number of clusters. The experiments also show how the measure can be used to determine an important parameter called *threshold* in rough clustering.

The proposed measure is applied to a synthetic data set that is designed to highlight the usefulness of the proposed measure, especially for rough clustering. Experiments with standard data sets are also used to ensure that the proposed measure works well with clustering schemes that have been tested by data mining community [38]. Finally, the distinguishing character of the decision-theoretic measure is illustrated by using monetary profit and loss considerations in a real-world data set. The experiments show how the business-oriented data mining

- *P. Lingras is with the Department of Mathematics and Computing Science, Saint Mary's University, 923 Robie Street, Halifax, NS B3H3C3, Canada. E-mail: pawan.lingras@smu.ca.*
- *M. Chen and D. Miao are with the Department of Electronics and Information Engineering, Tongji University, Jiading Campus, 4800 Cao'an Highway, Shanghai 201804, PR China. E-mail: minchen2008halifax@yahoo.com, miaoduoqian@163.com.*

Manuscript received 14 May 2008; revised 10 Oct. 2008; accepted 1 Dec. 2008; published online 18 Dec. 2008.
Recommended for acceptance by D. Talia.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2008-05-0264.
Digital Object Identifier no. 10.1109/TKDE.2008.236.

could benefit from the proposed framework as opposed to traditional distance-based measures.

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. Assume 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 introduces some notations related to the rough set theory. The notion of rough set was proposed by Pawlak [27], [28], and Pawlak et al. [29]. Let P be an equivalence relation on X. The pair $apr = (X, P)$ 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 nonempty intersection with A. We call $bnd(A) = \overline{apr}(A) - \underline{apr}(A)$ the boundary region of A.

2.1 Crisp Clustering

K-means clustering is one of the most popular statistical clustering techniques [9], [22]. The objective is to assign n objects to k clusters. The process begins by randomly choosing k objects as the centroids of the k clusters. The objects are assigned to one of the k clusters based on the minimum value of the distance $d(\vec{x}_l, \vec{c}_i)$ between the object vector $\vec{x}_l$ and the cluster vector $\vec{c}_i$. The distance $d(\vec{x}_l, \vec{c}_i)$ can be the standard euclidean distance.

After the assignment of all the objects to various clusters, the new centroid vectors of the clusters are calculated as

$$\vec{c}_i = \frac{\sum_{\vec{x}_l \in \vec{c}_i} \vec{x}_l}{\mid \vec{c}_i \mid}, \ where \ 1 \leq i \leq k.$$

Here $\mid \vec{c}_i \mid$ is the cardinality of cluster $\vec{c}_i$. The process stops when the centroids of clusters stabilize, i.e., the centroid vectors from the previous iteration are identical to those generated in the current iteration.

2.2 Rough Clustering

The conventional clustering techniques mandate that an object must belong to precisely one cluster. Such a requirement is found to be too restrictive in many data mining applications [13]. 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 [30]. 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 [3], [15], [27], [28], [35]. 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 [11], [12], [23], [24], [25], [31], [32], [33]. Lingras [16], [17] described how a rough set theoretic classification scheme can be represented using a rough set genome. In subsequent publications [19], [21], modifications of K-means and *Kohonen Self-Organizing Maps* (SOMs) were proposed to create intervals of clusters based on rough set theory. Asharaf et al. [1] extended the approach further with Rough Support Vector Clustering.

Rough sets were originally proposed using equivalence relations. However, it is possible to define a pair of upper and lower bounds $(\underline{apr}(C), \overline{apr}(C))$ or a rough set for every set $C \subseteq X$ as long as the properties specified by Pawlak [27], [28] are satisfied. Yao [40] and Yao and Lin [42] described various generalizations of rough sets by relaxing the assumptions of an underlying equivalence relation. Such a trend toward generalization is also evident in rough mereology proposed by Polkowski and Skowron [34] and the use of information granules in a distributed environment by Skowron and Stepaniuk [36]. The present study uses such a generalized view of rough sets.

Let us consider a hypothetical classification scheme

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

that partitions the set X 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 $C_i \in X/P$ using its lower $\underline{apr}(C_i)$ and upper $\overline{apr}(C_i)$ bounds. We will use m-dimensional vector representations, $\vec{x}_t$ for objects and $\vec{c}_i$ for cluster C_i, whenever it is notationally convenient.

We are considering the upper and lower bounds of only a few subsets of X. Therefore, it is not possible to verify all the properties of the rough sets [28]. However, the family of upper and lower bounds of $\vec{c}_i \in X/P$ are required to follow some of the basic rough set properties such as:

(P1) An object $\vec{x}$ can be part of at most one lower bound,
(P2) $\vec{x} \in \underline{A}(\vec{c}_i) \Longrightarrow \vec{x} \in \overline{A}(\vec{c}_i)$,
(P3) An object $\vec{x}$ is not part of any lower bound
$\Longleftrightarrow$
$\vec{x}$ 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. 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 [45] provide a more detailed discussion on the important properties of rough sets, and positive, boundary, and negative regions.

Lingras and West incorporated rough set into k-means clustering, which requires the addition of the concept of lower and upper bounds [19]. This section describes a refined version of the original proposal [18], [21], [32]. The following equation is used to calculate the centroids of clusters that need 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

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

where $\omega_{low} + \omega_{bnd} = 1$ and $1 \leq i \leq k$. The parameters ω_{low} and ω_{bnd} 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, $\vec{x}_l (1 \leq l \leq n)$, let $d(\vec{x}_l, \vec{c}_i)$ be the distance between itself and the centroid of cluster $\vec{c}_i$. The ratio $d(\vec{x}_l, \vec{c}_j)/d(\vec{x}_l, \vec{c}_i)$, $1 \leq i, j \leq k$, are used to determine the membership of $\vec{x}_l$ [21], [32]. Let $d(\vec{x}_l, \vec{c}_i) = min_{1 \leq j \leq k} d(\vec{x}_l, \vec{c}_j)$ and $\{T_l = j\text{: } d(\vec{x}_l, \vec{c}_j)/d(\vec{x}_l, \vec{c}_i) \leq threshold \text{ and } i \neq j\}$. We will use $\vec{x}_l \rightarrow T_l$ to denote the fact that object $\vec{x}_l$ is similar to all the elements of set T_l.

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_{low}, ω_{bnd}, 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.

2.3 Cluster Quality

Several cluster validity indices have been proposed to evaluate cluster quality obtained by different clustering algorithms. An excellent summary of various validity measures can be found in Halkidi et al. [10]. Here, we introduce two classical cluster validity indices and one used for fuzzy clusters.

2.3.1 Davies-Bouldin Index

This index [6] is a function of the ratio of the sum of within-cluster scatter to between-cluster separation. The scatter within the ith cluster, denoted by S_i, and the distance between cluster $\vec{c}_i$ and $\vec{c}_j$, denoted by d_{ij}, are defined as follows:

$$S_{i,q} = \left(\frac{1}{|\vec{c}_i|} \sum_{\vec{x} \in \vec{c}_i} \|\vec{x} - \vec{c}_i\|_2^q \right)^{1/q},$$

$$d_{ij,t} = \|\vec{c}_i - \vec{c}_j\|_t,$$

where $\vec{c}_i$ is the center of the ith cluster. $|\vec{c}_i|$ is the number of objects in $\vec{c}_i$. Integers q and t can be selected independently such that $q, t > 1$. The Davies-Bouldin index for a clustering scheme (CS) is then defined as

$$DB(CS) = \frac{1}{k} \sum_{i=1}^{k} R_{i,qt},$$

where $R_{i,qt} = \max_{1 \leq j \leq k, j \neq i} \{\frac{S_{i,q} + S_{j,q}}{d_{ij,t}}\}$.

The Davies-Bouldin index considers the average case of similarity between each cluster and the one that is most similar to it. Lower Davies-Bouldin index means a better clustering scheme.

2.3.2 Dunn Index

Dunn proposed another cluster validity index [7]. The index corresponding to a clustering scheme (CS) is defined by

$$D(CS) = \min_{1 \leq i \leq k} \left\{ \min_{1 \leq j \leq k, j \neq i} \left(\frac{\delta(\vec{c}_i, \vec{c}_j)}{\max_{1 \leq q \leq k} \Delta(\vec{c}_q)} \right) \right\},$$

where

$$\delta(\vec{c}_i, \vec{c}_j) = \min_{1 \leq i, j \leq k, i \neq j} \| \vec{c}_i - \vec{c}_j \|,$$

$$\Delta(\vec{c}_i) = \max_{\vec{x}_l, \vec{x}_t \in \vec{c}_i} \| \vec{x}_l - \vec{x}_t \| .$$

If a data set is well separated by a clustering scheme, the distance among the clusters, $\delta(\vec{c}_i, \vec{c}_j) (1 \leq i, j \leq k)$, is usually large and the diameters of the clusters, $\Delta(\vec{c}_i) (1 \leq i \leq k)$, are expected to be small. Therefore, a large value of $D(CS)$ corresponds to a good clustering scheme. The main drawback of the Dunn index is that the calculation is computationally expensive and the index is sensitive to noise.

2.3.3 Xie-Beni Index

This index is also called the compactness and separation validity function [37]. It is a representative cluster validity measure for fuzzy clustering. In fuzzy clustering, we assign a membership u_{it} for an object $\vec{x}_t$ to a cluster $\vec{c}_i$

$$S = \frac{\sum_{i=1}^{k} \sum_{t=1}^{n} u_{it} \times \|\vec{c}_i - \vec{x}_t\|}{n \times \min_{i,j} \|\vec{c}_i - \vec{c}_j\|}.$$

The numerator in the Xie-Beni index is a measure of cluster compactness, while the denominator reflects the separation of clusters. Xie et al. [38] modified the cluster validity measure to develop an improved fuzzy clustering algorithm.

Similar to the validity measures for crisp clustering, fuzzy clustering algorithms are based on the geometric distances. In the subsequent sections, we will discuss a decision-theoretic view of clustering. It will also help us evaluate clustering schemes based on monetary cost and benefit considerations.

2.4 Yao's Decision-Theoretic Framework

Yao proposed probabilistic rough set approximations in [39], which apply the Bayesian decision procedure for the

construction of probabilistic approximations. The classification of objects according to approximation operators in rough set theory can be easily fitted into the Bayesian decision-theoretic framework. Let $\Omega = \{A, A^c\}$ denote the set of states indicating that an object is in A and not in A, respectively. Let $A = \{a_1, a_2, a_3\}$ be the set of actions, where a_1, a_2, and a_3 represent the three actions in classifying an object, deciding $POS(A)$, deciding $NEG(A)$, and deciding $BND(A)$, respectively. The probabilities $P(A|[\vec{x}])$ and $P(A^c|[\vec{x}])$ are the probabilities that an object in the equivalence class $[\vec{x}]$ belongs to A and A^c, respectively. The expected loss $R(a_i|[\vec{x}])$ associated with taking the individual actions can be expressed as

$$R(a_1|[\vec{x}]) = \lambda_{11}P(A|[\vec{x}]) + \lambda_{12}P(A^c|[\vec{x}]),$$
$$R(a_2|[\vec{x}]) = \lambda_{21}P(A|[\vec{x}]) + \lambda_{22}P(A^c|[\vec{x}]),$$
$$R(a_3|[\vec{x}]) = \lambda_{31}P(A|[\vec{x}]) + \lambda_{32}P(A^c|[\vec{x}]),$$

where $\lambda_{i1} = \lambda(a_i|A)$, $\lambda_{i2} = \lambda(a_i|A^c)$, and $i = 1, 2, 3$. The Bayesian decision procedure leads to the following minimum-risk decisions:

If $R(a_1|[\vec{x}]) \leq R(a_2|[\vec{x}])$ and $R(a_1|[\vec{x}]) \leq R(a_3|[\vec{x}])$, decide $POS(A)$;

If $R(a_2|[\vec{x}]) \leq R(a_1|[\vec{x}])$ and $R(a_2|[\vec{x}]) \leq R(a_3|[\vec{x}])$, decide $NEG(A)$;

If $R(a_3|[\vec{x}]) \leq R(a_1|[\vec{x}])$ and $R(a_3|[\vec{x}]) \leq R(a_2|[\vec{x}])$, decide $BND(A)$.

Tie-breaking criteria should be added so that each object is classified into only one region. Since $P(A|[\vec{x}]) + P(A^c|[\vec{x}]) = 1$, the rules to classify any object in $[\vec{x}]$ can be simplified based on the probability $P(A|[\vec{x}])$ and the loss function λ_{ij} ($i = 1, 2, 3$; $j = 1, 2$).

3 Multicategory Decisions

The previous section described Yao et al.'s framework for a binary classification problem. In this section, we extend it to multicategory problems, where categorization may be either supervised or unsupervised.

3.1 Extending Yao's Model

Recently, Yao's decision-theoretic framework was extended to the multicategory problem [20]. Let $X = \{\vec{x}_1, \ldots, \vec{x}_n\}$ be a finite set of objects. Let $C = \{\vec{c}_1, \ldots, \vec{c}_k\}$ be a finite set of k states given that C is the set of categories and each category is represented by a vector $\vec{c}_i$ ($1 \leq i \leq k$). Furthermore, let C partition the set of objects X. For every object, $\vec{x}_l$, T_l represents a nonempty set of all the categories that are similar to $\vec{x}_l$. Clearly, $T_l \subseteq C$. We will use $\vec{x}_l \rightarrow T_l$ to denote the fact that object $\vec{x}_l$ is similar to all the elements of set T_l. Moreover, object $\vec{x}_l$ can be similar to one and only one T_l. Therefore, upper ($\overline{apr}$) and lower ($\underline{apr}$) approximations of each category $\vec{c}_i$ can be expressed as follows:

$$\overline{apr}(\vec{c}_i) = \{\vec{x}_l | \vec{x}_l \rightarrow T_l, \vec{c}_i \in T_l\},$$
$$\underline{apr}(\vec{c}_i) = \{\vec{x}_l | \vec{x}_l \rightarrow T_l, \{\vec{c}_i\} = T_l\}.$$

3.2 Loss Functions for Multicategory Problem

Following Yao [39], Lingras et al. [20] proposed a set of states and actions to describe the decision-theoretic framework for multicategory rough sets, given as follows.

States. *The states are essentially the set of categories* $C = \{\vec{c}_1, \ldots, \vec{c}_k\}$.

Actions. *Let* $B = \{B_1, \ldots, B_s\} = 2^C - \{\emptyset\}$ *be a family of nonempty subsets of* C, *where* $s = 2^k - 1$. *A set of actions* $b = \{b_1, \ldots, b_s\}$ *corresponds to set* B, *where* b_j *represents the action in assigning an object* $\vec{x}_l$ *to the set* B_j.

Note that some of the sets B_js will be the same as the set T_ls defined in previous sections. There will be a total of n T_ls, one for each object, and they may not be distinctly different from each other. That is, two objects may be similar to the same subset of C. On the other hand, there will be exactly $s = 2^k - 1$ distinct B_js. For simplicity, we will use b_j to refer to the action as well as the set B_j.

The Bayesian decision procedure for multicategory rough sets is described as follows.

Let $\lambda_{\vec{x}_l}(b_j|\vec{c}_i)$ denote the loss, or cost, for taking action b_j when an object $\vec{x}_l$ belongs to $\vec{c}_i$. Let $P(\vec{c}_i|\vec{x}_l)$ be the conditional probability of an object $\vec{x}_l$ being in state $\vec{c}_i$. Therefore, the expected loss $R(b_j|\vec{x}_l)$ associated with taking action b_j for an object $\vec{x}_l$ is given by

$$R(b_j|\vec{x}_l) = \sum_{i=1}^{k} \lambda_{\vec{x}_l}(b_j|\vec{c}_i)P(\vec{c}_i|\vec{x}_l).$$

For an object $\vec{x}_l$, if $R(b_j|\vec{x}_l) \leq R(b_h|\vec{x}_l)\ \forall\, h = 1, \ldots, s$, then decide b_j.

The loss function is generalized from the 0.5 probabilistic model [29] given by Yao [39] as follows:

$$\lambda_{\vec{x}_l}(b_j|\vec{c}_i) = \frac{|b_j - T_l|}{|b_j|}, \quad \text{if } \vec{c}_i \in b_j,$$
$$\lambda_{\vec{x}_l}(b_j|\vec{c}_i) = \frac{|b_j - \emptyset|}{|b_j|}, \quad \text{if } \vec{c}_i \notin b_j.$$

When $\vec{c}_i$ belongs to b_j, the loss for taking action b_j corresponds to the fraction of b_j that is not related to $\vec{x}_l$. Otherwise, the loss for taking action b_j will have the maximum value of 1.

The loss function described here can be further enhanced for business applications by using the actual dollar costs of making the decisions. We will consider such an enhancement when we experiment with data from a retail store.

4 Rough Cluster Quality Index Based on Decision Theory

Clustering is an unsupervised classification method when the only data available are unlabeled. Most clustering algorithms need to know the number of clusters. A cluster validity measure can provide us some information about the appropriate number of clusters. Cluster validity measures such as Davies-Bouldin [6] can help us assess whether a clustering method accurately presents the structure of the data set. There are several cluster validity indices to evaluate crisp and fuzzy clustering [2], [4], [5], [6], [7], [37]. However, there is no evaluation measure for rough clustering at present.

Decision-theoretic framework has been helpful in providing a better understanding of the classification model. The decision-theoretic rough set model considers various classes of loss functions as described above. The extension of the decision-theoretic rough set model to the multicategory problem and corresponding loss functions are also described in the previous section. It is possible to construct a cluster validity measure by considering various loss functions based on decision theory.

Within a given set of objects, there may be clusters such that objects in the same cluster are more similar than those in different clusters. The objective of clustering is to find the right groups or clusters for the given set of objects. However, to find the right clusters, we need exponential time comparisons and the problem has been proved to be NP-hard [8]. 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 $CS = \{\vec{c}_1, \ldots, \vec{c}_k\}$. Clustering algorithms such as k-means approximate the actual clustering. It is possible that each object may not necessarily belong to only one cluster. However, there will be a core corresponding to each cluster within our clustering scheme. We will start with formal definitions for the proposed validity indices based on the hypothetical cores. The centroid of the hypothetical core will be used in our definitions.

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 $\vec{x}_l \in core(\vec{c}_i)$ cannot belong to other clusters. Therefore, $core(\vec{c}_i)$ can be considered the best representation of $\vec{c}_i$ to a certain extent.*

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 $core(\vec{c}_i)$.*

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*

$$R(CS, \vec{c}) = \sum_{\vec{x}_l \in \vec{c}} R(b_j(CS, \vec{x}_l)|\vec{x}_l). \tag{2}$$

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

$$R(CS) = \sum_{i=1}^{k} R(CS, \vec{c}_i) = \sum_{l=1}^{n} R(b_j(CS, \vec{x}_l)|\vec{x}_l). \tag{3}$$

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

For rough clustering (RC), an object $\vec{x}_l$ may belong to more than one cluster. Moreover, each cluster $\vec{c}_i$ is represented by its lower approximation $\underline{apr}(\vec{c}_i)$ and upper approximation $\overline{apr}(\vec{c}_i)$. There also exists the boundary region $bnd(\vec{c}_i) = \overline{apr}(\vec{c}_i) - \underline{apr}(\vec{c}_i)$. Based on the definitions given above, we give the following definitions for rough clustering.

Risk for assigning an object to clusters under rough clustering. *For rough clustering, let $b_j(RC, \vec{x}_l)$ be the action that assigns the object $\vec{x}_l$ to a set $S \subseteq RC$ such that $\vec{c}_i \in S$. Since T_l is equal to $b_j(RC, \vec{x}_l)$, the loss function for $\vec{x}_l$ can be expressed as follows:*

$$\begin{aligned} &\lambda_{\vec{x}_l}(b_j(RC, \vec{x}_l)|\vec{c}_i) = 0 \quad if\ \vec{c}_i \in b_j(RC, \vec{x}_l); \\ &\lambda_{\vec{x}_l}(b_j(RC, \vec{x}_l)|\vec{c}_i) = 1 \quad if\ \vec{c}_i \notin b_j(RC, \vec{x}_l). \end{aligned} \tag{4}$$

The risk associated with the assignment will then be given as $R(b_j(RC, \vec{x}_l)|\vec{x}_l)$. Let $sim(\vec{x}_l, \vec{c}_i)$ be the similarity between $\vec{x}_l$ and $core(\vec{c}_i)$. Usually, sim will be inversely proportional to the distance between the two vectors. We will assume that $core(\vec{c}_i) = \underline{apr}(\vec{c}_i)$. If the lower bound is empty, then we will assume that the $core(\vec{c}_i)$ is the centroid of $\overline{apr}(\vec{c}_i)$.

$R(b_j(RC, \vec{x}_l)|\vec{x}_l)$ is obtained assuming that the conditional probability $P(\vec{c}_i|\vec{x}_l)$ is proportional to $sim(\vec{x}_l, \vec{c}_i)$, given by

$$P(\vec{c}_i|\vec{x}_l) = \frac{sim(\vec{x}_l, \vec{c}_i)}{\sum_{1 \le j \le k} sim(\vec{x}_l, \vec{c}_j)}, \tag{5}$$

$$\begin{aligned} &R(b_j(RC, \vec{x}_l)|\vec{x}_l) \\ &= \sum_{\substack{i=1,\ldots,k \\ \vec{c}_i \notin b_j(RC, \vec{x}_l)}} \lambda_{\vec{x}_l}(b_j(RC, \vec{x}_l)|\vec{c}_i) \times P(\vec{c}_i|\vec{x}_l). \end{aligned} \tag{6}$$

Risk for Lower Approximation. *For rough clustering, let $R(RC, \underline{apr}(\vec{c}_i))$ be the risk for a lower approximation. Since $\underline{apr}(\vec{c}_i)$ consists of a group of objects $\underline{apr}(\vec{c}_i) = \{\vec{x}_{i1}, \ldots, \vec{x}_{ig}\}$, $R(RC, \underline{apr}(\vec{c}_i))$ is given by*

$$R(RC, \underline{apr}(\vec{c}_i)) = \sum_{\vec{x}_l \in \underline{apr}(\vec{c}_i)} R(b_j(RC, \vec{x}_l)|\vec{x}_l). \tag{7}$$

Risk for Upper Approximation. *For rough clustering, let $R(RC, \overline{apr}(\vec{c}_i))$ be the risk for a upper approximation. Since $\overline{apr}(\vec{c}_i)$ consists of a group of objects $\overline{apr}(\vec{c}_i) = \{\vec{x}_{i1}, \ldots, \vec{x}_{ig}\}$, $R(RC, \overline{apr}(\vec{c}_i))$ is given by*

$$R(RC, \overline{apr}(\vec{c}_i)) = \sum_{\vec{x}_l \in \overline{apr}(\vec{c}_i)} R(b_j(RC, \vec{x}_l)|\vec{x}_l). \tag{8}$$

Risk for Boundary Area. *For rough clustering, let $R(RC, bnd(\vec{c}_i))$ be the risk for the boundary area of $\vec{c}_i$. Since $bnd(\vec{c}_i)$ consists of a group of objects $bnd(\vec{c}_i) = \{\vec{x}_{i1}, \ldots, \vec{x}_{ig}\}$, $R(RC, bnd(\vec{c}_i))$ is given by*

$$R(RC, bnd(\vec{c}_i)) = \sum_{\vec{x}_l \in bnd(\vec{c}_i)} R(b_j(RC, \vec{x}_l)|\vec{x}_l). \tag{9}$$

One can deduce the following properties for rough clustering:

- $R(RC, \overline{apr}(\vec{c}_i)) = R(RC, \underline{apr}(\vec{c}_i)) + R(RC, bnd(\vec{c}_i))$; (P1.1)
- $R(RC, \underline{apr}(\vec{c}_i)) \le R(RC, \overline{apr}(\vec{c}_i))$; (P1.2)
- $R(RC, \overline{apr}(\vec{c}_i) \cap \underline{apr}(\vec{c}_i)) = R(RC, \underline{apr}(\vec{c}_i))$; (P1.3)

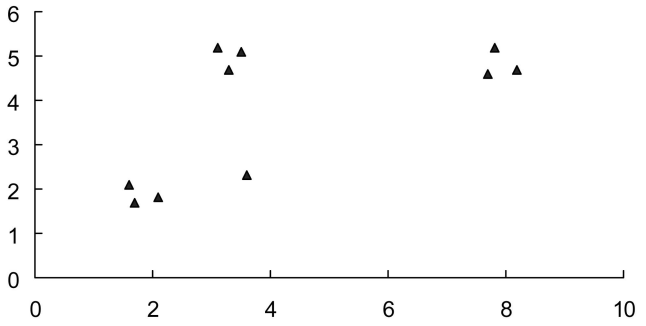

Fig. 1. Distribution of objects.

- $R(RC, \overline{apr}(\vec{c}_i) \cup \underline{apr}(\vec{c}_i)) = R(RC, \overline{apr}(\vec{c}_i))$; (P1.4)
- $R(RC, \underline{apr}(\vec{c}_i) \cap \underline{apr}(\vec{c}_j)) = R(RC, \emptyset) = 0$; (P1.5)
- $R(RC, \underline{apr}(\vec{c}_i) \cup \underline{apr}(\vec{c}_j)) = R(R\overline{C, \underline{apr}(\vec{c}_i)) + R(RC, \underline{apr}(\vec{c}_j))}$; (P1.6)
- $R(RC, \overline{apr}(\vec{c}_i) \cup \overline{apr}(\vec{c}_j)) \leq R(RC, \overline{apr}(\vec{c}_i)) + R(RC, \overline{apr}(\vec{c}_j))$; (P1.7)
- $R(RC) \leq \sum_{i=1}^{k} R(RC, \overline{apr}(\vec{c}_i))$; (P1.8)
- $R(RC) \geq \sum_{i=1}^{k} R(RC, \underline{apr}(\vec{c}_i))$. (P1.9)

Property (P1.1) follows from the fact that $\underline{apr}(\vec{c}_i)$ and $bnd(\vec{c}_i)$ are disjoint and their union is equal to $\overline{apr}(\vec{c}_i)$. The fact that $\underline{apr}(\vec{c}_i)$ is a subset of $\overline{apr}(\vec{c}_i)$ can be used to derive properties (P1.2), (P1.3), and (P1.4). Since $\underline{apr}(\vec{c}_i) \cap \underline{apr}(\vec{c}_j) = \emptyset$, we get properties (P1.5), and (P1.6). However, $\overline{apr}(\vec{c}_i) \cup \overline{apr}(\vec{c}_j)$ may not be an empty set. Hence, we get the properties (P1.7) and (P1.8). Finally, property (P1.9) can be derived using the knowledge that the unions of all the lower bounds is an improper subset of X.

Properties (P1.8) and (P1.9) tell us that we cannot calculate the risk for a rough clustering scheme by simply summing up risks for either the lower bounds or upper bounds of the clusters.

Crisp clustering is a special case of rough clustering. Using the definitions given above, we can obtain the risk for crisp clustering (CC) as follows. The core of a cluster is, in fact, the cluster $\vec{c}_i$ obtained from clustering.

- $core(\vec{c}_i) = \underline{apr}(\vec{c}_i) = \overline{apr}(\vec{c}_i)$; (P2.1)
- $R(CC, core(\vec{c}_i)) = R(CC, \underline{apr}(\vec{c}_i)) = R(CC, \overline{apr}(\vec{c}_i))$; (P2.2)

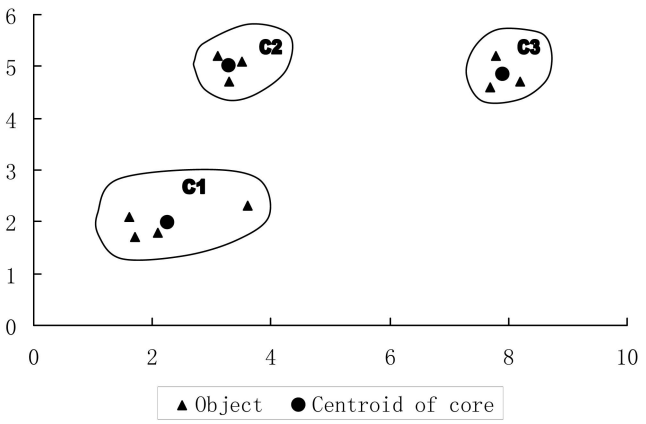

Fig. 2. Crisp clustering.

TABLE 1
Risk Calculations for Objects in Crisp Clustering (CC)

$\vec{x}_l$	$R(b_j(CC, \vec{x}_l)\|\vec{x}_l)$	b_j	$\vec{T}_l$
$\vec{x}_1 = (1.7, 1.7)$	0.417	$\{\vec{c}_1\}$	$\{\vec{c}_1\}$
$\vec{x}_2 = (2.1, 1.8)$	0.446		
$\vec{x}_3 = (1.6, 2.1)$	0.443		
$\vec{x}_4 = (3.6, 2.3)$	0.568		
$\vec{x}_5 = (3.5, 5.1)$	0.524	$\{\vec{c}_2\}$	$\{\vec{c}_2\}$
$\vec{x}_6 = (3.1, 5.2)$	0.508		
$\vec{x}_7 = (3.3, 4.7)$	0.559		
$\vec{x}_8 = (7.7, 4.6)$	0.391	$\{\vec{c}_3\}$	$\{\vec{c}_3\}$
$\vec{x}_9 = (7.8, 5.2)$	0.383		
$\vec{x}_{10} = (8.2, 4.7)$	0.357		

- $R(CC, core(\vec{c}_i) \cap core(\vec{c}_j)) = R(CC, \emptyset) = 0$; (P2.3)
- $R(CC, core(\vec{c}_i) \cup core(\vec{c}_j)) = R(CC, core(\vec{c}_i)) + R(CC, core(\vec{c}_j))$. (P2.4)

Therefore, the proposed risk measure for crisp clustering can be expressed as follows:

$$\begin{aligned} R(CC) &= \sum_{i=1}^{k} R(CC, core(\vec{c}_i)) \\ &= \sum_{i=1}^{k} R(CC, \underline{apr}(\vec{c}_i)) \\ &= \sum_{i=1}^{k} R(CC, \overline{apr}(\vec{c}_i)). \end{aligned} \tag{10}$$

Let us illustrate the proposed risk measure for two different clustering schemes, crisp clustering and rough clustering, with the following example. Since crisp clustering is assumed to approximate the actual clustering, we can evaluate rough clustering by comparing the value of the proposed risk measure to that of crisp clustering.

Example 1. Let $X = \{\vec{x}_1, \ldots, \vec{x}_{10}\}$ and $C = \{\vec{c}_1, \vec{c}_2, \vec{c}_3\}$. The distribution of objects is shown in Fig. 1. According to the distribution, nine objects $\vec{x}_i$ $(1 \leq i \leq 10,\ i \neq 4)$ are expected to form three groups, but one object denoted by $\vec{x}_4 = (3.6, 2.3)$ is far from these groups.

For crisp clustering, we get the three clusters as shown in Fig. 2. Risk of each object, as well as b_j and T_l for each object, are presented in Table 1. Table 2 shows $core(\vec{c}_i)$ and group risk of each cluster $\vec{c}_i$.

For rough clustering, we set $k = 3$ and $\omega_{low} = 0.8$. We also adjust the *threshold* to obtain the results presented in Fig. 3. In the figure, the dashed line outlines the upper approximation of each cluster, and the solid line describes

TABLE 2
Risk for Clusters in Crisp Clustering (CC)

$\vec{c}_i$	$core(\vec{c}_i)$	$objects$	$R(CC, \vec{c}_i)$
$\vec{c}_1$	$(2.25, 1.975)$	$\{\vec{x}_1, \vec{x}_2, \vec{x}_3, \vec{x}_4\}$	1.874
$\vec{c}_2$	$(3.3, 5)$	$\{\vec{x}_5, \vec{x}_6, \vec{x}_7\}$	1.59
$\vec{c}_3$	$(7.9, 4.833)$	$\{\vec{x}_8, \vec{x}_9, \vec{x}_{10}\}$	1.131

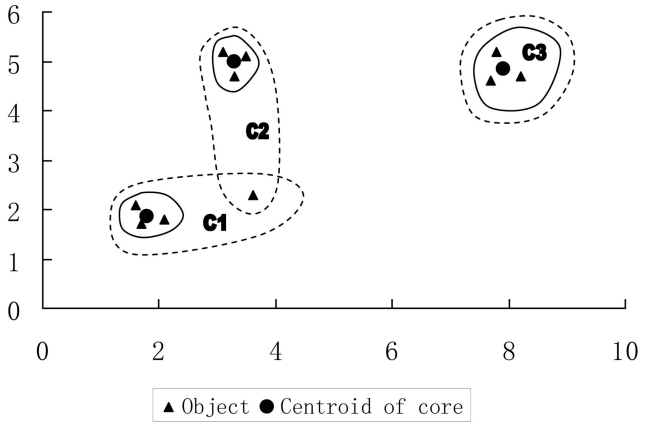

Fig. 3. Rough clustering.

TABLE 3
Risk Calculations for Objects in Rough Clustering (RC)

$\vec{x}_l$	$R(b_j(RC,\vec{x}_l)\|\vec{x}_l)$	b_j	$\vec{T}_l$
$\vec{x}_1=(1.7,1.7)$	0.417		
$\vec{x}_2=(2.1,1.8)$	0.446	$\{\vec{c}_1\}$	$\{\vec{c}_1\}$
$\vec{x}_3=(1.6,2.1)$	0.443		
$\vec{x}_4=(3.6,2.3)$	0.2	$\{\vec{c}_1,\vec{c}_2\}$	$\{\vec{c}_1,\vec{c}_2\}$
$\vec{x}_5=(3.5,5.1)$	0.501		
$\vec{x}_6=(3.1,5.2)$	0.488	$\{\vec{c}_2\}$	$\{\vec{c}_2\}$
$\vec{x}_7=(3.3,4.7)$	0.529		
$\vec{x}_8=(7.7,4.6)$	0.38		
$\vec{x}_9=(7.8,5.2)$	0.373	$\{\vec{c}_3\}$	$\{\vec{c}_3\}$
$\vec{x}_{10}=(8.2,4.7)$	0.347		

the lower approximation of each cluster. Since the distance between $\vec{x}_4$ and $core(\vec{c}_1)$ is close to that between $\vec{x}_4$ and $core(\vec{c}_2)$, $\vec{x}_4$ belongs to the upper approximations of $\vec{c}_1$ and $\vec{c}_2$. Risk, as well as b_j and T_l for each object, is presented in Table 3. Table 4 shows centroid and group risk for each cluster $\vec{c}_i$. Because of the existence of the upper approximations, the centroids of $core(\vec{c}_1)$ and $core(\vec{c}_2)$ go toward the center much more than those in crisp clustering.

According to (3), we obtain the risk for crisp clustering and that for rough clustering, shown in Table 5. Since we have one object $\vec{x}_4$ that should not belong to a single cluster, rough clustering provides a more reasonable representation. This fact is confirmed by the risk for rough clustering, which is smaller than that for crisp clustering.

5 Study Data and Experiments

We use three data sets, synthetic data set to highlight various features of the proposal, a standard data set to compare our results with other researchers [26], [38], and a retail store's data set to show the unique ability of our proposal to consider monetary costs and benefits while analyzing cluster quality.

TABLE 5
Risk as the Quality Indices for Example 1

Clustering scheme	Risk
crisp clustering	4.596
rough clustering	4.124

5.1 Synthetic Data

The synthetic data set has been developed to test some of the salient features of both crisp and rough clustering in relation to the proposed risk measure. In order to visualize the data set, we restrict it to two dimensions as can be seen in Fig. 4. There are a total of 65 objects. It is obvious that there are three distinct clusters. However, five objects do not belong to any particular cluster. We performed crisp clustering and rough clustering on the synthetic data set for different numbers of clusters.

Fig. 5 shows how the cluster index varies for different number of clusters for crisp clustering. As shown in Fig. 5, the risk decreases as the number of clusters is reduced from seven to three. However, we see a sudden and sharp increase when the number of clusters is reduced from three to two. The risk reaches the minimum value when objects are grouped into three clusters. That means the risk measure proposed in this study correctly indicates that the right number of clusters equals three. A similar trend can also be found for rough clustering in Fig. 6. The minimum risk can be found for the number of clusters equal to three. The value of *threshold* used in Fig. 6 is 1.4. The *threshold* is an important parameter in rough clustering that determines the size of boundary regions. The risk measure was used to arrive at the appropriate value of *threshold*.

Fig. 7 shows how changing the value of *threshold* can affect the risk of clustering. The number of clusters was kept constant at three, while the *threshold* values were changed from 1.1 to 1.7. The higher values led to larger boundary regions and lower risks. However, one should not increase the boundary region too much as it will lead to fairly indecisive and uninformative clustering scheme. Fig. 7 shows a rapid decline in risk until the *threshold* reaches a value of 1.4. One can see that the decline slows down after that point. Therefore, it is reasonable to use the value of $threshold = 1.4$.

The results so far seem to indicate that the risk measure can be used to determine important features of a clustering

TABLE 4
Risk for Clusters in Rough Clustering (RC)

$\vec{c}_i$	centroid	$objects$	$R(RC,\underline{apr}(\vec{c}_i))$	$R(RC,\overline{apr}(\vec{c}_i))$
$\vec{c}_1$	$(1.8,1.867)$	$\{\vec{x}_1,\vec{x}_2,\vec{x}_3,\vec{x}_4\}$	1.306	1.506
$\vec{c}_2$	$(3.3,5)$	$\{\vec{x}_4,\vec{x}_5,\vec{x}_6,\vec{x}_7\}$	1.518	1.718
$\vec{c}_3$	$(7.9,4.833)$	$\{\vec{x}_8,\vec{x}_9,\vec{x}_{10}\}$	1.1	1.1

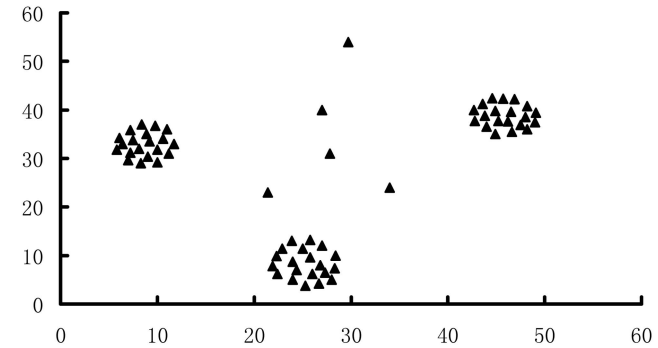

Fig. 4. Synthetic data.

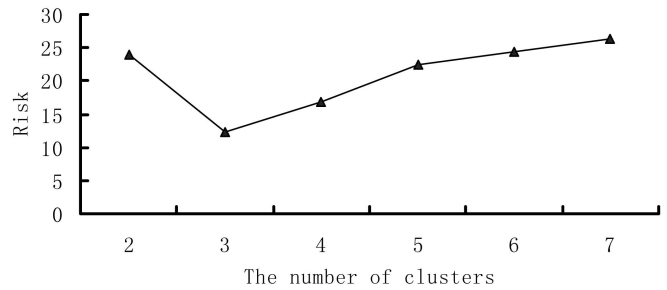

Fig. 5. Synthetic data: risk for crisp clustering for different number of clusters.

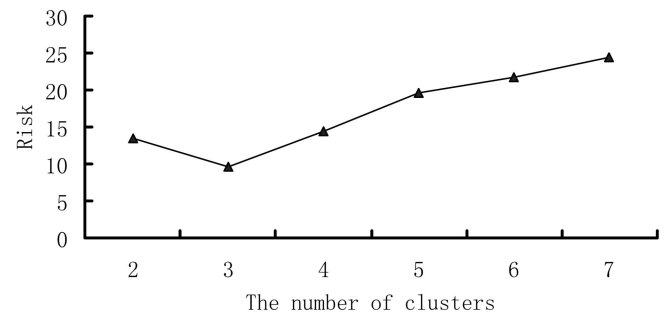

Fig. 6. Synthetic data: risk for rough clustering for different number of clusters.

scheme such as the number of clusters and the size of boundary region.

We can have a closer look at the detailed risk values that help us understand the difference between crisp and rough clustering. Fig. 8 presents the crisp clustering results. A satisfactory rough clustering with $k = 3$, $\omega_{low} = 0.75$, and $threshold = 1.4$ is presented in Fig. 9. Summary of the comparison between risks from various regions in rough and crisp clustering can be found in Table 6. In crisp clustering, the five objects that do not seem to belong to any particular cluster are forced to go to one of the clusters. This results in the centroids of the clusters to be shifted from the centroid of the cores of the clusters. That is why the risk of assigning objects to these cores tends to be higher in comparison to those for rough clustering. The difference in risk for crisp and rough clustering seems to be the highest for the boundary region consisting of the five objects. This large difference makes sense, since rough clustering clearly represents the ambivalent assignments of these five objects

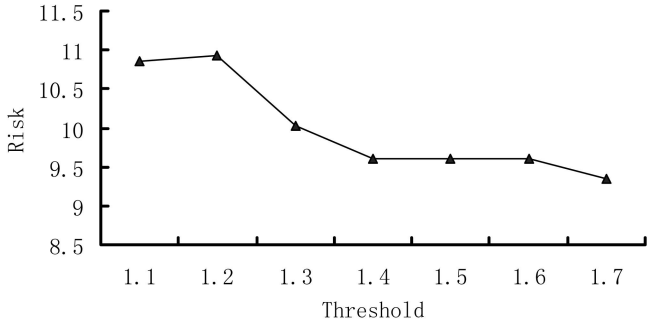

Fig. 7. Change in risk for rough clustering with *threshold* for synthetic data.

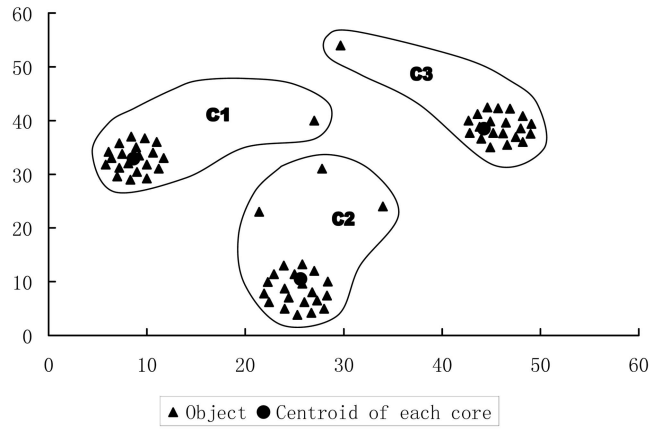

Fig. 8. Synthetic data: crisp clustering.

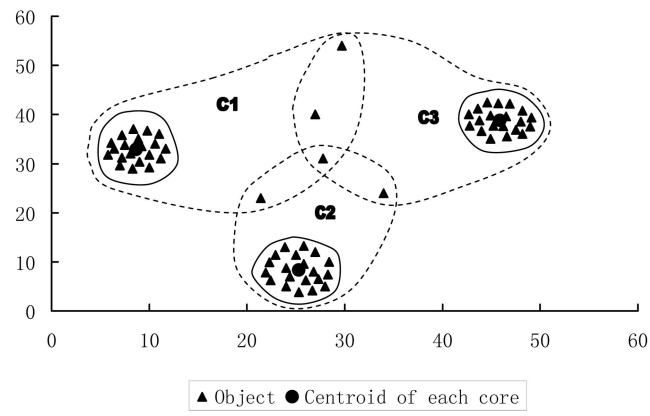

Fig. 9. Synthetic data: rough clustering.

to more than one cluster, while crisp clustering is forced to assign these objects to a single cluster.

The shape of the neighborhood of the cluster can vary depending on the exponent used in the distance function. Most studies use the euclidean distance, where the exponent of the difference is 2. It would be interesting to study the impact on the value of the number of clusters and *threshold* by using distance functions that use different values of exponents. We used the absolute values of distances in determining the closeness. Fig. 10 shows the change in risk for different values of *threshold* for three clusters. The results show that for all the values of exponents, the risk experiences an identifiable dip for $threshold = 1.4$. Fig. 11 shows the change in risk for

TABLE 6
Risk Comparison between Rough and Crisp Clustering of Synthetic Data

Clustering Scheme	$core(\vec{c}_1)$	$core(\vec{c}_2)$	$core(\vec{c}_3)$	boundary area	quality index
CC	2.998	3.626	2.72	3.03	12.372
RC	2.82	3.28	2.58	0.92	9.6

$threshold = 1.4$ and different number of clusters. The results show that for all the values of exponents other than 1, the risk is minimum for three clusters. Based on these results, one can say that the euclidean distance is adequate for our clustering process.

5.2 Wisconsin Breast Cancer Data

The previous section used synthetic data that was designed to highlight and test salient features of the proposed risk-based measure. In this section, we use a standard real-world data set that is tested for clustering by other researchers such as Xie et al. [38]. The testing for such a standard data set makes it possible to compare the proposed approach with some of the previous clustering results.

Wisconsin breast cancer databases were obtained from the University of Wisconsin Hospitals, Madison, by Dr. William H. Wolberg [26]. 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.

Fig. 12 shows the variation in risk as we change the number of clusters in the K-means crisp clustering. Corresponding changes in risk with the use of rough K-means algorithm is shown in Fig. 13. For rough clustering, we set $\omega_{low} = 0.75$ and $threshold = 1.4$. It can be seen from both the figures that the risk of clustering is minimum for two clusters and then continuously rises. This is an encouraging sign, since we want to group the objects into two categories: benign and malignant. The appropriate number of clusters obtained here also corresponds to those obtained by Xie et al. [38] when they tested their modified fuzzy clustering algorithm with Xie-Beni validity measure for fuzzy clustering.

The variation in risk for different values of *threshold* is shown in Fig. 14. The risk seems to decline from threshold value of 1.1 to 1.7. However, there is a sharp drop in risk when the threshold is reduced from 1.3 to 1.4. Therefore, *threshold* of 1.4 can again be used as an appropriate value.

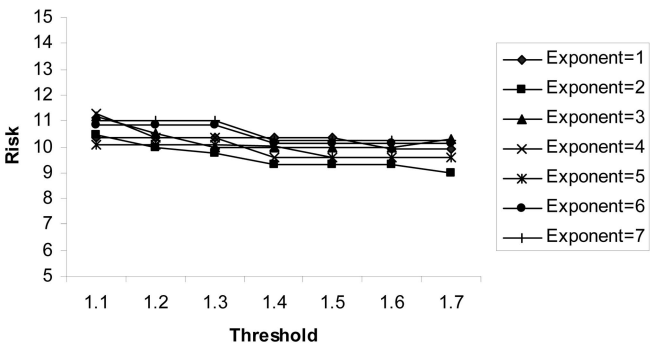

Fig. 10. Effect of distance functions.

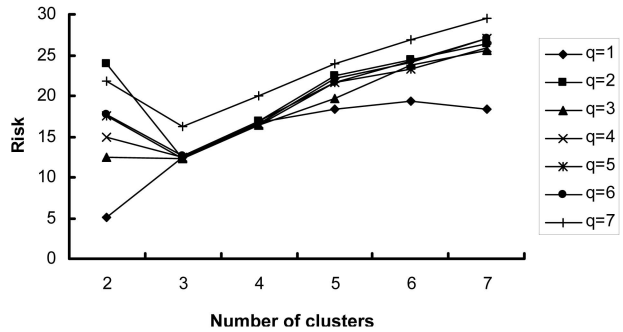

Fig. 11. Effect of distance functions.

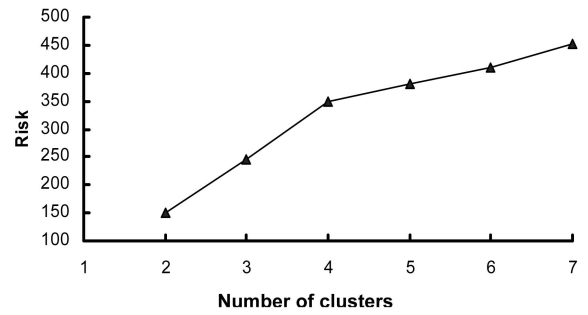

Fig. 12. Breast cancer data: crisp clustering.

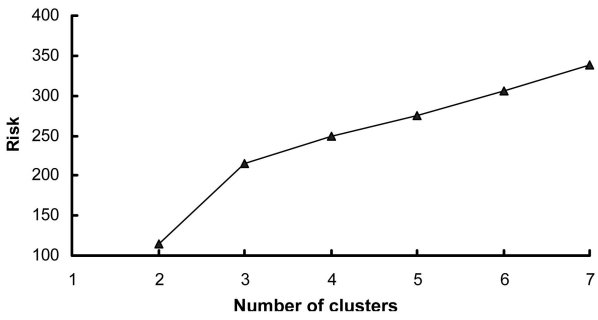

Fig. 13. Breast cancer data: rough clustering.

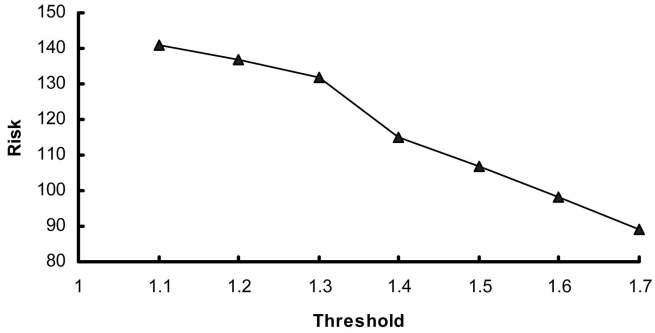

Fig. 14. Breast cancer data: threshold in rough clustering.

5.3 Monetary Loss Function for Retail

This section reports experiments with a real-world data set belonging to a small retail chain. The experiments highlight the contributions of decision-theoretic cluster quality measure. The loss function is enhanced with monetary cost and benefit considerations of a clustering scheme.

The data consists of all the customer transactions in 2006. There were a total of 68,716 transactions, one transaction per item purchased. 40,260 of these transactions can be associated with 5,878 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 every month may be a little more loyal than one who spends \$1,200 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 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 12-monthly spending and visit values, which may be too detailed for the purpose of grouping, we 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.

Three hundred and thirteen 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 5,565. After experimenting with different values, w_l was set at 0.75.

As mentioned before, we can enhance our loss function using dollar amounts. In our case, the dollar amounts will be the profits. We can look at profits as negative losses in our formulation. Let S_l be total annual sales of a customer $\vec{x}_l$. Assuming 30 percent profit margin, our profits will be $S_l \times 0.3$. Loss will be $-S_l \times 0.3$.

Let us consider a promotional campaign targeted at relatively high spenders. Let us assume that it is a two-tier campaign that will be aimed at the top two clusters of customers. The first-tier promotion will be directed at the customers in the highest spending cluster $\vec{c}_k$. It will cost \$100 and will result in 10 percent increase in sales. That means the increase in sales will be $S_l \times 0.1$. We have to subtract the cost of promotion in calculating the increase in profits, so the profits will be $S_l \times 0.1 \times 0.3 - \$100$. Since the dollar loss is the negative of profits, it will be $\$100 - S_l \times 0.1 \times 0.3$. We can now modify the cost for all the actions b_j such that $\vec{c}_k \in b_j$, since these actions possibly assign a customer to the highest spending cluster $\vec{c}_k$. The modified loss function for such an action will be given as

$$\lambda_{\vec{x}_l}(b_j|\vec{c}_i) = (\$100 - S_l \times 0.1 \times 0.3) \times \frac{|b_j - T_l|}{|b_j|} \quad \text{if} \quad \vec{c}_i \in b_j \wedge \vec{c}_k \in b_j, \tag{11}$$

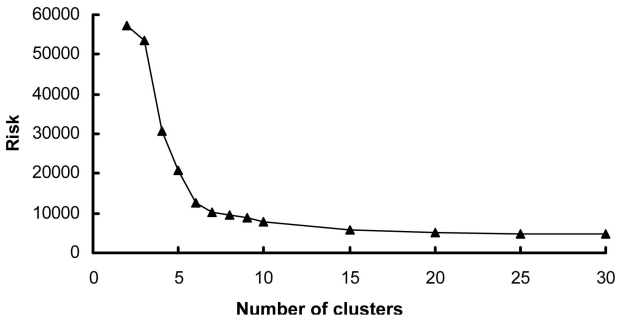

Fig. 15. Monetary risk for crisp clustering for two-tier retail promotion.

$$\lambda_{\vec{x}_l}(b_j|\vec{c}_i) = (\$100 - S_l \times 0.1 \times 0.3) \times \frac{|b_j - \emptyset|}{|b_j|} \quad \text{if} \quad \vec{c}_i \notin b_j \wedge \vec{c}_k \in b_j. \tag{12}$$

The second-tier promotional campaign will be directed at the second largest spending cluster $\vec{c}_{k-1}$. It will cost \$50 and is expected to lead to a 5 percent increase in sales. That means the increase in sales will be $S_l \times 0.05$. We have to subtract the cost of promotion in calculating increase in the profits, so the profits will be $S_l \times 0.05 \times 0.3 - \$50$. Since the dollar loss is the negative of profits, it will be $\$50 - S_l \times 0.05 \times 0.3$. We will exclude all the customers who have already been a target of tier-1 campaign. That means we need to modify the cost for all the actions b_j such that $\vec{c}_{k-1} \in b_j$ and $\vec{c}_k \notin b_j$. The modified loss function for such an action will be given as

$$\lambda_{\vec{x}_l}(b_j|\vec{c}_i) = (\$50 - S_l \times 0.05 \times 0.3) \times \frac{|b_j - T_l|}{|b_j|} \quad \text{if } \vec{c}_i \in b_j \wedge \vec{c}_k \notin b_j \wedge \vec{c}_{k-1} \in b_j, \tag{13}$$

$$\lambda_{\vec{x}_l}(b_j|\vec{c}_i) = (\$50 - S_l \times 0.05 \times 0.3) \times \frac{|b_j - \emptyset|}{|b_j|} \quad \text{if } \vec{c}_i \notin b_j \wedge \vec{c}_k \notin b_j \wedge \vec{c}_{k-1} \in b_j. \tag{14}$$

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.

Fig. 15 shows the changes in risks for two-tier promotional campaign with K-means crisp clustering. There is a decline in risk with increase in number of clusters. However, the rate of decline slows down after seven clusters. Therefore, it is reasonable to assume that the customers fall into seven clusters.

To illustrate the flexibility of the proposed decision-theoretic framework, we will further apply it to a three-tier campaign. The first-tier promotions will be aimed at customers that possibly belong to $\vec{c}_k$. It will cost \$200 and result in sales increase of 20 percent. The second-tier promotions will be aimed at customers that possibly belong to $\vec{c}_{k-1}$. It will cost \$100 and result in sales increase of 10 percent. The third-tier promotions will be aimed at customers that possibly belong to $\vec{c}_{k-2}$. It will cost \$50 and result in sales increase of 5 percent. The loss functions for actions that assign customers to either $\vec{c}_k$, $\vec{c}_{k-1}$, or $\vec{c}_{k-2}$ are modified using formulas similar to those used for two-tier campaign.

Fig. 16 shows the changes in risks for three-tier promotional campaign with K-means crisp clustering. As with the

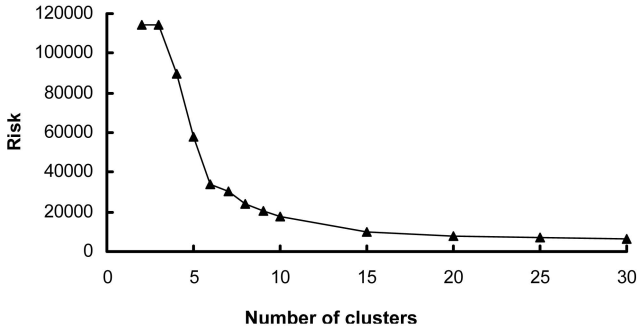

Fig. 16. Monetary risk for crisp clustering for three-tier retail promotion.

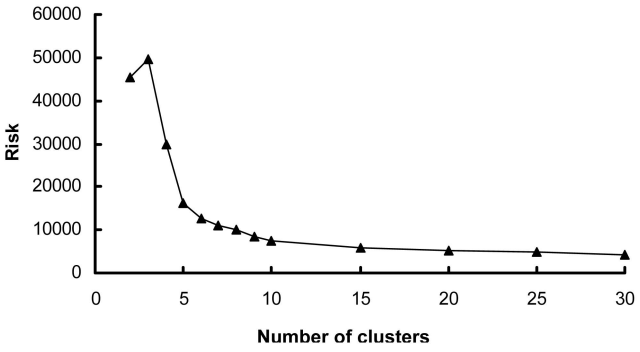

Fig. 17. Monetary risk for rough clustering for the two-tier retail promotion.

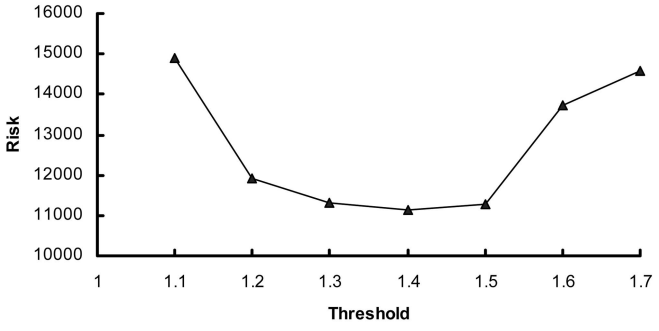

Fig. 18. Monetary risk for different thresholds in rough clustering.

two-tier campaign, there is a decline in risk with increase in number of clusters. It is interesting to note that the risk for two clusters and three clusters is exactly the same. This makes sense since a three-tier campaign should separate the customers into four categories: one for each promotion and one for no promotion. The rate of decline slows down a little later for the three-tier campaign compared to two-tier campaign. Since the rate of decline slows down significantly after 10 clusters, it may be reasonable to cluster the customers into 10 clusters. Again, a little higher number of clusters for the three-tier promotion makes sense compared to the two-tier promotion.

Fig. 17 shows the changes in risk for rough clustering for the two-tier promotional campaign. Except for a curious jump in risk from two clusters to three clusters, the pattern is the same as the one obtained for the two-tier crisp clustering. The rate of decrease in the risk seems to slow down after five clusters, and there is a further decline in rate after seven clusters.

TABLE 7
The Number of Objects in Lower and Upper Bounds of Five Clusters

area	C1	C2	C3	C4	C5
lower	3620	948	292	69	10
upper	4023	1526	510	117	19

TABLE 8
The Number of Objects in Lower and Upper Bounds of Seven Clusters

area	C1	C2	C3	C4	C5	C6	C7
lower	3167	1048	273	153	104	44	8
upper	3599	1658	535	333	207	73	11

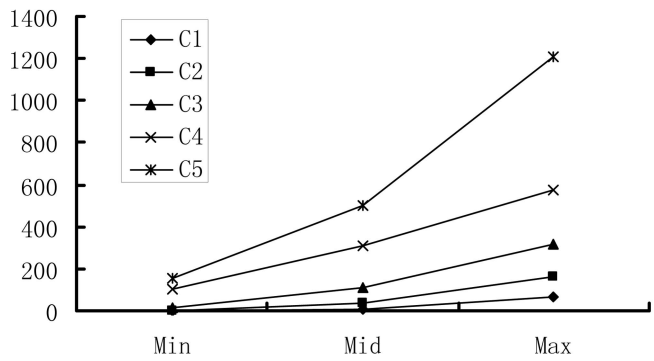

Fig. 19. Five rough centroids for the retail data.

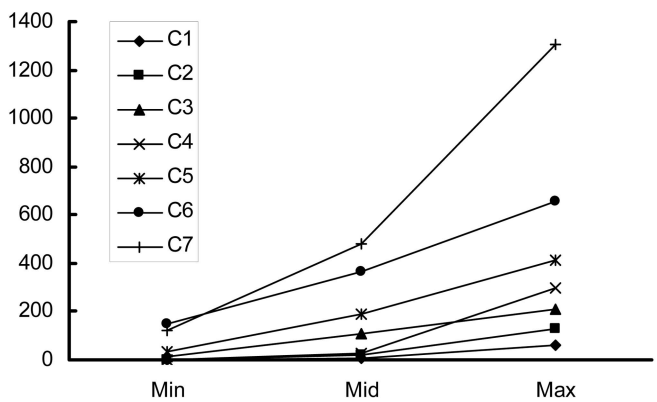

Fig. 20. Seven rough centroids for the retail data.

Therefore, we will set the number of clusters to be between five and seven. Fig. 18 shows the variation in risk as *threshold* changes for $k = 7$. Interestingly, similar to the synthetic and Wisconsin breast cancer data, when *threshold* is at a value of 1.4, there is a local minima suggesting that 1.4 is again a reasonable value. Similar local minima was also found for $k = 5$.

Tables 7 and 8 show the sizes of the upper and lower bounds of each cluster for $k = 5$ and $k = 7$, respectively. The representative patterns for each cluster can be found in Fig. 19 for $k = 5$ and Fig. 20 for $k = 7$. One can see that if we set $k = 7$, there is a marginal gain in lower risk, but the two-tier campaign will be directed at only 73 clusters. With $k = 5$, the campaign will be directed at a total of 117 customers. The store owner may choose to include a larger customer base in

TABLE 9
The Number of Objects in the Intersection of Clusters for Five 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	–

the campaign and choose the clustering scheme with $k = 5$. Moreover, five clusters are a little easy to analyze than the seven clusters. Therefore, let us look at the five clusters in little greater detail.

Cluster C1 is the largest cluster consisting of moderate spenders who spend \$0 to \$52 in a month. The next cluster, C2, is about the quarter the size of C1 with spending ranging from \$1 to \$163. Third cluster (C3) is even smaller with spending ranging from \$16 to \$330. Fourth cluster has approximately 69 to 117 customers with spending ranging from \$109 to \$594. Please note that these values are monthly spending. The average annual spending for these customers is in excess of \$3,500. When the store spends \$50 on promotion for these customers, it will likely receive additional annual profits exceeding $\$3{,}500 \times 0.05 \times 0.3 = \$52.50$. The profits can be increased by focusing on the lower bound of the cluster. The last cluster is the smallest with spending ranging from \$137 to \$1,330. The average annual spending for these customers is in excess of \$6,000. When the store spends \$100 on promotion for these customers, it will likely receive additional annual profits exceeding $\$6{,}000 \times 0.1 \times 0.3 = \$180$. Again, the profits can be increased by focusing on the lower bound of this cluster as well. The objects in the upper bound could be the target of the second-tier \$50 promotional campaign. The overlaps between different clusters is shown in Table 9. It can be seen that the intermediate clusters, i.e., C2, C3, and C4 have overlaps with two clusters on either side. For example, C2 overlaps with C1 and C3, while C3 overlaps with C2 and C4, and C4 overlaps with C3 and C5. On the other hand, clusters C1 and C5 have overlap with only one cluster: C1 with C2 and C5 with C4.

6 Summary and Conclusions

This paper describes a cluster quality index based on decision theory. The proposal uses a loss function to construct the quality index. Therefore, the cluster quality is evaluated by considering the total risk of categorizing all the objects. Such a decision-theoretic representation of cluster quality may be more useful in business-oriented data mining than traditional geometry-based cluster quality measures. In addition to evaluating crisp clustering, the proposal is an evaluation measure for rough 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). The application of the measure to synthetic data with known number of clusters and boundary region provides credence to the proposal. The measure is also tested with a standard data set that is used by other researchers for testing clustering schemes and cluster validity measures. The proposed measure gave comparable results to the previous studies.

A real advantage of the decision-theoretic cluster validity measure is its ability to include monetary considerations in evaluating a clustering scheme. Use of the measure to derive an appropriate clustering scheme for a promotional campaign in a retail store highlighted its unique ability to include cost and benefit considerations in commercial data mining. We can also extend it to evaluating other clustering algorithms such as fuzzy clustering. Such a cluster validity measure can be useful in further theoretical development in clustering. Results of such development will be reported in future publications.

Acknowledgments

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

References

[1] S. Asharaf, S.K. Shevade, and N.M. Murty, "Rough Support Vector Clustering," *Pattern Recognition,* vol. 38, no. 10, pp. 1779-1783, 2005.
[2] J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms.* Plenum Press, 1981.
[3] M. Banerjee, S. Mitra, and S.K. Pal, "Rough Fuzzy MLP: Knowledge Encoding and Classification," *IEEE Trans. Neural Networks,* vol. 9, no. 6, pp. 1203-1216, Nov. 1998.
[4] J.C. Bezdek and N.R. Pal, "Some New Indexes of Cluster Validity," *IEEE Trans. Systems, Man, and Cybernetics, Part B,* vol. 28, no. 3, pp. 301-315, June 1998.
[5] R.B. Calinski and J. Harabasz, "A Dendrite Method for Cluster Analysis," *Communications in Statistics—Theory and Methods,* vol. 3, pp. 1-27, 1974.
[6] D.L. Davies and D.W. Bouldin, "A Cluster Separation Measure," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 1, no 2, pp. 224-227, Apr. 1979.
[7] J.C. Dunn, "Well Separated Clusters and Optimal Fuzzy Partitions," *J. Cybernetics,* vol. 4, pp. 95-104, 1974.
[8] E. Falkenauer, *Genetic Algorithms and Grouping Problems.* John Wiley & Sons, 1998.
[9] J.A. Hartigan and M.A. Wong, "Algorithm AS136: A K-Means Clustering Algorithm," *Applied Statistics,* vol. 28, pp. 100-108, 1979.
[10] M. Halkidi, Y. Batistakis, and M. Vazirgianni, "Clustering Validity Checking Methods: Part II," *ACM SIGMOD Conf. Record,* vol. 31, no. 3, pp. 19-27, 2002.
[11] S. Hirano and S. Tsumoto, "On Constructing Clusters from Non-Euclidean Dissimilarity Matrix by Using Rough Clustering," *Proc. Japanese Soc. for Artificial Intelligence (JSAI) Workshops,* pp. 5-16, 2005.
[12] T.B. Ho and N.B. Nguyen, "Nonhierarchical Document Clustering by a Tolerance Rough Set Model," *Int'l J. Intelligent Systems,* vol. 17, no. 2, pp. 199-212, 2002.
[13] A. Joshi and R. Krishnapuram, "Robust Fuzzy Clustering Methods to Support Web Mining," *Proc. ACM SIGMOD Workshop Data Mining and Knowledge Discovery,* pp. 1-8, June 1998.
[14] Y. Li, C. Zhang, and J.R. Swan, "An Information Filtering Model on the Web and Its Application in JobAgent," *Knowledge-Based Systems,* vol. 13, no. 5, pp. 285-296, 2000.
[15] Y. Li, S.C.K. Shiu, S.K. Pal, and J.N.K. Liu, "A Rough Set-Based Case-Based Reasoner for Text Categorization," *Int'l J. Approximate Reasoning,* vol. 41, no. 2, pp. 229-255, 2006.
[16] P. Lingras, "Unsupervised Rough Set Classification Using GAs," *J. Intelligent Information Systems,* vol. 16, no. 3, pp. 215-228, 2001.

[17] P. Lingras, "Rough Set Clustering for Web Mining," *Proc. 2002 IEEE Int'l Conf. Fuzzy Systems,* pp. 12-17, 2002.
[18] P. Lingras, "Applications of Rough Set Based K-Means, Kohonen, GA Clustering," *Trans. Rough Sets,* vol. 7, pp. 120-139, 2007.
[19] P. Lingras and C. West, "Interval Set Clustering of Web Users with Rough K-Means," *J. Intelligent Information System,* vol. 23, no. 1, pp. 5-16, 2004.
[20] P. Lingras, M. Chen, and D.Q. Miao, "Rough Multi-Category Decision Theoretic Framework," *Rough Sets and Knowledge Technology,* pp. 676-683, Springer, 2008.
[21] P. Lingras, M. Hogo, and M. Snorek, "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 Int'l Journal,* vol. 2, no. 3, pp. 217-230, 2004.
[22] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," *Proc. Fifth Berkeley Symp. Mathematical Statistics and Probability,* vol. 1, pp. 281-297, 1967.
[23] S. Mitra, "An Evolutionary Rough Partitive Clustering," *Pattern Recognition Letters,* vol. 25, pp. 1439-1449, 2004.
[24] S. Mitra, H. Bank, and W. Pedrycz, "Rough-Fuzzy Collaborative Clustering," *IEEE Trans. Systems, Man, and Cybernetics,* vol. 36, no. 4, pp. 795-805, Aug. 2006.
[25] H.S. Nguyen, "Rough Document Clustering and the Internet," *Handbook on Granular Computing.* John Wiley & Sons, 2007.
[26] O.L. Mangasarian and W.H. Wolberg, "Cancer Diagnosis via Linear Programming," *SIAM News,* vol. 23, no. 5, 1990.
[27] Z. Pawlak, "Rough Sets," *Int'l J. Information and Computer Sciences,* vol. 11, pp. 145-172, 1982.
[28] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data.* Kluwer Academic Publishers, 1992.
[29] Z. Pawlak, S.K.M. Wong, and W. Ziarko, "Rough Sets: Probabilistic Versus Deterministic Approach," *Int'l J. Man-Machine Studies,* vol. 29, pp. 81-95, 1988.
[30] W. Pedrycz and J. Waletzky, "Fuzzy Clustering with Partial Supervision," *IEEE Trans. Systems, Man, and Cybernetics,* vol. 27, no. 5, pp. 787-795, Sept. 1997.
[31] G. Peters, "Outliers in Rough k-Means Clustering," *Proc. First Int'l Conf. Pattern Recognition and Machine Intelligence,* pp. 702-707, 2005.
[32] G. Peters, "Some Refinements of Rough k-Means," *Pattern Recognition,* vol. 39, no. 8, pp. 1481-1491, 2006.
[33] J.F. Peters, A. Skowron, Z. Suraj, W. Rzasa, M. Borkowski, "Clustering: A Rough Set Approach to Constructing Information Granules," *Proc. Sixth Int'l Conf. Soft Computing and Distributed Processing,* pp. 57-61, 2002.
[34] L. Polkowski and A. Skowron, "Rough Mereology: A New Paradigm for Approximate Reasoning," *Int'l J. Approximate Reasoning,* vol. 15, no. 4, pp. 333-365, 1996.
[35] S. Saha, C.A. Murthy, and S.K. Pal, "Rough Set Based Ensemble Classifier for Web Page Classification," *Fundamenta Informaticae,* vol. 76, nos. 1/2, pp. 171-187, 2007.
[36] A. Skowron and J. Stepaniuk, "Information Granules in Distributed Environment," *New Directions in Rough Sets, Data Mining, and Granular-Soft Computing,* N. Zhong, A. Skowron, and S. Ohsuga, eds., vol. 1711, pp. 357-365, Springer-Verlag, 1999.
[37] X. Xie and G. Beni, "A Validity Measure for Fuzzy Clustering," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 13, no. 8, pp. 841-847, Aug. 1991.
[38] Y. Xie, V.V. Raghavan, P. Dhatric, and X. Zhao, "A New Fuzzy Clustering Algorithm for Optimally Finding Granular Prototypes," *Int'l J. Approximate Reasoning,* vol. 40, pp. 109-124, 2005.
[39] Y.Y. Yao, "Decision-Theoretic Rough Set Models," *Lecture Notes in Computer Science,* vol. 4481, pp. 1-12, 2007.
[40] Y.Y. Yao, "Constructive and Algebraic Methods of the Theory of Rough Sets," *Information Sciences,* vol. 109, pp. 21-47, 1998.
[41] Y.Y. Yao, "Information Granulation and Approximation in a Decision-Theoretical Model of Rough Sets," *Rough-Neuro Computing: Techniques for Computing with Words,* pp. 491-516, Springer, 2003.
[42] Y.Y. Yao and T.Y. Lin, "Generalization of Rough Sets Using Modal Logic," *Intelligent Automation and Soft Computing,* vol. 2, no. 2, pp. 103-120, 1996.
[43] Y.Y. Yao and S.K.M. Wong, "A Decision Theoretic Framework for Approximating Concepts," *Int'l J. Man-Machine Studies,* vol. 37, pp. 793-809, 1992.
[44] Y.Y. Yao, S.K.M. Wong, and P. Lingras, "A Decision-Theoretic Rough Set Model," *Methodologies for Intelligent Systems,* vol. 5, pp. 17-24, 1990.
[45] Y.Y. Yao and Y. Zhao, "Attribute Reduction in Decision-Theoretic Rough Set Models," *Information Sciences,* vol. 178, no. 17, pp. 3356-3373, 2008.

Pawan Lingras is a graduate of IIT Bombay with graduate studies from University of Regina. He is currently a professor at Saint Mary's University, Halifax, Canada. His research interests include data mining, Web intelligence, information retrieval, and rough set theory. He is a member of the IEEE.

Min Chen received the MSc degree from Hefei University of Technology, Anhui. She is currently working toward the PhD degree at Tongji University, Shanghai, and is a researcher at the Key Laboratory of Embedded System and Service Computing, Ministry of Education, PR China. Her research interests focus on machine learning, rough sets, knowledge discovery, and data mining.

Duoqian Miao received the PhD degree from the Institute of Automation, Chinese Academy of Sciences, Beijing. He is a professor at Tongji University, Shanghai. His research interests focus on machine learning, rough set theory and applications, knowledge discovery and data mining, Chinese language processing and handwriting, and Chinese character recognition.