Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

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

A hybrid approach to outlier detection based on boundary region

Feng Jiang[a,*], Yuefei Sui[b], Cungen Cao[b]

[a] College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, PR China
[b] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, PR China

ARTICLE INFO

Article history:
Received 21 July 2010
Available online 11 September 2011
Communicated by R.C. Guido

Keywords:
Outlier detection
Rough sets
Boundary
Distance
KDD

ABSTRACT

In recent years, much attention has been given to the problem of outlier detection, whose aim is to detect outliers – objects who behave in an unexpected way or have abnormal properties. The identification of outliers is important for many applications such as intrusion detection, credit card fraud, criminal activities in electronic commerce, medical diagnosis and anti-terrorism, etc. In this paper, we propose a hybrid approach to outlier detection, which combines the opinions from boundary-based and distance-based methods for outlier detection (Jiang et al., 2005, 2009; Knorr and Ng, 1998). We give a novel definition of outliers – *BD* (*boundary and distance*)-*based outliers*, by virtue of the notion of boundary region in rough set theory and the definitions of distance-based outliers. An algorithm to find such outliers is also given. And the effectiveness of our method for outlier detection is demonstrated on two publicly available databases.

1. Introduction

Usually, the tasks of knowledge discovery in database (KDD) can be classified into four general categories: (a) dependency detection, (b) class identification, (c) class description, and (d) outlier/exception detection (Knorr and Ng, 1998). In contrast to most KDD tasks, outlier detection aims to find small groups of data objects that are exceptional when compared with the rest large amount of data, in terms of certain sets of properties. For many applications, such as fraud detection in E-commerce (Bolton and Hand, 2002), it is more interesting to find the rare events than to find the common ones, from a knowledge discovery standpoint. Studying the extraordinary behaviors of outliers can help us uncover the valuable information hidden behind them.

Recently, the detection of outlier has gained considerable interest in KDD. Many researchers have begun focusing on outlier detection and attempted to apply algorithms for finding outliers to tasks such as fraud detection (Bolton and Hand, 2002), identification of computer network intrusions (Lane and Brodley, 1999; Eskin et al., 2002), detection of employers with poor injury histories (Knorr et al., 2000), and peculiarity-oriented mining (Zhong et al., 2001, 2003).

With increasing awareness on outlier detection in literatures, more concrete meanings of outliers are defined for solving problems in specific domains. While there is no single, generally accepted, formal definition of an outlier, Hawkins' definition captures the spirit: "an outlier is an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism" (Hawkins, 1980; Knorr and Ng, 1998).

Outlier detection has a long history in statistics (Hawkins, 1980; Barnett and Lewis, 1994). Other researchers, beginning with the work by Knorr and Ng (1998), Knorr et al. (2000), have taken a non-parametric approach and proposed using an example's distance to its nearest neighbors as a measure of unusualness (Angiulli and Pizzuti, 2002; Ramaswamy et al., 2000; Knorr et al., 2000). Although distance is an effective non-parametric approach to detect outliers, the drawback is the amount of computation time required. Straightforward algorithms, such as those based on nested loops, typically require $O(N^2)$ distance computations. This quadratic scaling means that it will be very difficult to mine outliers as we tackle increasingly larger data sets.

As an extension of naive set theory, rough set theory is introduced by Pawlak (1982, 1991), for the study of intelligent systems characterized by insufficient and incomplete information. It is motivated by practical needs in classification and concept formation. The rough set philosophy is based on the assumption that with every objects of the universe there is associated a certain amount of information (data, knowledge), expressed by means of some attributes. Objects having the same description are indiscernible. In recent years, there has been a fast growing interest in rough set theory (Qian et al., 2008; Yin et al., 2009). Successful applications of the rough set model in a variety of problems have demonstrated its importance and versatility (Wang et al., 2008, 2011; Miao et al., 2009; Chen et al., 2010).

[*] Corresponding author. Tel./fax: +86 532 88959036.
E-mail addresses: jiangkong@163.net, jiangkong2002@163.com (F. Jiang), yfsui@ict.ac.cn (Y. Sui), cgcao@ict.ac.cn (C. Cao).

0167-8655/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2011.07.002

In a series of works, we have proposed different methods to detect outliers exploiting the framework of rough set theory (Jiang et al., 2005, 2008, 2009). In (Jiang et al., 2009), we introduced distance-based outlier detection to rough set theory and proposed the definitions of distance metrics for distance-based outlier detection in rough set theory. And in (Jiang et al., 2005), based on the notions in rough sets, we proposed a boundary-based method for outlier detection. Those objects in boundary regions are deemed as having more likelihood to be an outlier than objects in lower approximations (Chen et al., 2010).

Although distance-based method has been widely used, it still has some drawbacks, for instance, the distance-based method cannot give a degree of outlierness for each object, it is difficult to find the local outliers using the method, and the computation complexity of the method is usually too high, etc. On the other hand, although the basic idea of boundary-based method is intuitive and meaningful, the method does not have a good performance for outlier detection, which can be concluded from Jiang et al. (2008). In this paper, as an extension of our previous work, we combine the opinions from distance-based and boundary-based methods for outlier detection, to obtain a hybrid method for outlier detection, which aims to complement the advantages of the two previous methods and solve the problems of them together.

The basic idea of our method is as follows. Given an information system $IS = (U,A,V,f)$. For any $X \subseteq U$ $(X \neq \emptyset)$, $B \subseteq A$ and $x \in X$, we first divide X into three parts: the exceptional boundary $EB(X)$, the principal boundary $PB_B(X)$, and the lower approximation $\underline{X}_B$ of X under relation $IND(B)$. Then by virtue of the spirit of distance-based method, we calculate the distance between x and each object in $EB(X)$, $PB_B(X)$, and $\underline{X}_B$, respectively. However, differing from the traditional distance-based method, we adopt different attitudes to objects from different parts of X when detecting outliers in X. That is, since objects in $EB(X)$ have the most likelihood to be an outlier, the more the objects in $EB(X)$ that are a short distance from object x, the more the likelihood of x to be an outlier. On the other hand, since objects in $\underline{X}_B$ have the least likelihood to be an outlier, the more the objects in $\underline{X}_B$ that are a great distance from object x, the more the likelihood of x to be an outlier. And for objects in $PB_B(X)$, the likelihood to be an outlier is moderate. Therefore, the more the objects in $PB_B(X)$ that are at a distance from object x, the more the likelihood of x to be an outlier. In a word, when given a set of indiscernibility relations (or knowledge) on U, if object x is always close to most objects in $EB(X)$, and x is always far from most objects in $\underline{X}_B$. In addition, x always keeps a proper distance from most objects in $PB_B(X)$. Then we may consider object x as not behaving normally according to the given knowledge at hand. We call such objects *BD* (*boundary and distance*)*-based outliers* with respect to X.

The remainder of this paper is organized as follows. In the next section, we present some preliminaries of rough set theory that are relevant to this paper. In Section 3, we give some definitions concerning BD-based outliers in information systems of rough set theory. An example and an algorithm to find BD-based outliers are also given. Experimental results are given in Section 4. Finally, Section 5 concludes the paper.

2. Preliminaries

In rough set terminology, a data table is also called an *information system* (Wang et al., 2009, 2011). When the attributes are classified into decision attributes and condition attributes, a data table is also called a decision system. More formally, an information system is a quadruple $IS = (U,A,V,f)$, where:

(1) U is a non-empty finite set of objects;
(2) A is a non-empty finite set of attributes;
(3) V is the union of attribute domains, i.e., $V = \bigcup_{a \in A} V_a$, where V_a denotes the domain of attribute a;
(4) $f : U \times A \rightarrow V$ is an information function such that for any $a \in A$ and $x \in U$ $f(x,a) \in V_a$.

Each subset $B \subseteq A$ of attributes determines a binary relation $IND(B)$, called *indiscernibility relation*, defined as follows (Wang et al., 2009, 2011):

$$IND(B) = \{(x,y) \in U \times U : \forall a \in B\ (f(x,a) = f(y,a))\}. \quad (1)$$

It is obvious that $IND(B)$ is an equivalence relation on U.

Given any $B \subseteq A$, relation $IND(B)$ induces a *partition* of U, which is denoted by $U/IND(B)$, where an element from $U/IND(B)$ is called an *equivalence class* or *elementary set*. For every element x of U, let $[x]_B$ denote the equivalence class of relation $IND(B)$ that contains element x, called the equivalence class of x under relation $IND(B)$.

Let $B \subseteq A$ and $X \subseteq U$, the B-lower and B-upper approximation of X is defined respectively as follows

$$\underline{X}_B = \bigcup\{[x]_B \in U/IND(B) : [x]_B \subseteq X\}; \quad (2)$$

$$\overline{X}_B = \bigcup\{[x]_B \in U/IND(B) : [x]_B \cap X \neq \emptyset\}. \quad (3)$$

The set $BN_B(X) = \overline{X}_B - \underline{X}_B$ is called the *B-boundary region* of X. An element in the lower approximation $\underline{X}_B$ necessarily belongs to X, while an element in the upper approximation $\overline{X}_B$ possibly belongs to X. And an element in the boundary region $BN_B(X)$ cannot be unambiguously classified into X.

A set is said to be *rough* (respectively *crisp*) if the boundary region is non-empty (respectively empty). Consequently each rough set has *boundary-line* elements, i.e., objects, which cannot be with certainty classified neither as members of the set nor of its complement. Obviously crisp sets have no boundary-line elements at all. That means that boundary-line elements cannot be properly classified by employing the available knowledge. While the elements in the lower approximations can be properly classified by virtue of the available knowledge. Hence we may deem the boundary-line elements as behaving in an unexpected way or featuring abnormal properties when comparing with the objects in the lower approximation. And since the aim of outlier detection is to find the small groups of objects in domain U who behave in an unexpected way or have abnormal properties. Therefore, in this paper, we consider that the boundary-line elements have more likelihood to be an outlier than the objects in the lower approximation, and utilize the information contained within the boundary region for outlier detection. That is, we shall adopt different attitudes to objects in the boundary region and objects in the lower approximation when detecting outliers (Jiang et al., 2005).

3. Boundary and distance-based outliers

By now, rough set theory has been found to have many interesting applications. The rough set approach seems to be of fundamental importance to AI and cognitive sciences, especially in the areas of machine learning and KDD (Skowron and Rauszer, 1992; Pawlak et al., 1995). However, in rough set community, there are few concerns on the problem of outlier detection. Therefore in this section, we discuss the issues of outlier definition and detection in information systems of rough sets. In the following subsection, we first give some definitions concerning BD-based outliers. Next an example and an algorithm to find BD-based outliers are presented.

3.1. Definitions

Our definition for BD-based outliers in an information system follows the spirit of Hawkins' definition for outliers (Hawkins, 1980). That is, given an information system $IS = (U,A,V,f)$ and

$X \subseteq U$ ($X \neq \emptyset$), for any $x \in X$, if x has some characteristics that differ greatly from those of other objects in X, in terms of attributes in A, we may call x an outlier with respect to X in IS.

Especially, our definition for BD-based outliers has a characteristic that is ignored by most current definitions for outliers. That is, for a given data set (universe) U, we do not have to detect outliers just in U by checking all elements of U. In fact we may consider detecting outliers with respect to any subset X of U, where X can be a particular subset of U which we are interested in or anything else which we are willing to separate from other elements of U.

Next, in order to be used in outlier detection, we define a new notion – *inner boundary*, which is based on the basic notion of rough sets – boundary region.

Definition 3.1 (*Inner boundary*). Let $IS = (U,A,V,f)$ be an information system, $X \subseteq U$ and $X \neq \emptyset$. For any $B \subseteq A$, the *inner boundary* $IB_B(X)$ of X under relation $IND(B)$ is defined as

$$IB_B(X) = \{x \in X : [x]_B \not\subseteq X\}, \tag{4}$$

where $[x]_B = \{u \in U : \forall a \in B(f(u,a) = f(x,a))\}$ denotes the indiscernibility class of relation $IND(B)$ that contains element x.

Proposition 3.1. *Given an information system $IS = (U,A,V,f)$. For any $X \subseteq U$ ($X \neq \emptyset$), let $IB_B(X)$ and $\underline{X}_B$ be the inner boundary and lower approximation of X under relation $IND(B)$, respectively. Then $IB_B(X) = X - \underline{X}_B$.*

Since the proof of Proposition 3.1 is straightforward, we omit it here.

From the above proposition, we can see that for any $B \subseteq A$, we can partition set $X \subseteq U$ into two parts: the inner boundary $IB_B(X)$ and the lower approximation $\underline{X}_B$ of X. Moreover, we can further partition the inner boundary $IB_B(X)$ of X into two parts: the *exceptional boundary* and the *principal boundary* of X.

Definition 3.2 (*Exceptional boundary*). Let $IS = (U,A,V,f)$ be an information system, where $A = \{a_1, a_2, \ldots, a_m\}$. For any $X \subseteq U$ ($X \neq \emptyset$), and $a_i \in A$, $1 \leqslant i \leqslant m$, let $IB_{\{a_i\}}(X)$ be the inner boundary of X under relation $IND(\{a_i\})$. The *exceptional boundary* $EB(X)$ of X in IS can be defined as

$$EB(X) = \bigcap_{i=1}^{m} IB_{\{a_i\}}(X). \tag{5}$$

Definition 3.3 (*Principal boundary*). Let $IS = (U,A,V,f)$ be an information system, where $A = \{a_1, a_2, \ldots, a_m\}$. For any $X \subseteq U$ ($X \neq \emptyset$), and $B \subseteq A$, let $IB_B(X)$ be the inner boundary of X under relation $IND(B)$, $EB(X)$ be the exceptional boundary of X. The *principal boundary* $PB_B(X)$ of X under relation $IND(B)$ is defined as

$$PB_B(X) = IB_B(X) - EB(X). \tag{6}$$

Thereby, given an information system $IS = (U,A,V,f)$, for any $X \subseteq U$ and $B \subseteq A$, we can divide X into three parts: the exceptional boundary $EB(X)$, the principal boundary $PB_B(X)$, and the lower approximation $\underline{X}_B$ of X. And when we detect outliers in X below, we shall adopt different attitudes to objects from different parts of X. That is, the objects in $EB(X)$ have the most likelihood to be an outlier, since they always lie in the boundary region with respect to the given knowledge in IS and objects in the boundary region cannot be properly classified by employing the available knowledge. And the objects in $PB_B(X)$ have more likelihood to be an outlier than objects in $\underline{X}_B$, since the objects in $PB_B(X)$ cannot be characterized with certainty as belonging or not to X, using the given knowledge $IND(B)$, but the objects in $\underline{X}_B$ can do that.

In the following, we construct two kinds of sequence – sequence of attributes and sequence of attribute subsets (Jiang et al., 2009).

Definition 3.4 (*Sequence of attributes*). Let $IS = (U,A,V,f)$ be an information system, where $A = \{a_1, a_2, \ldots, a_m\}$. For any $X \subseteq U$ and $X \neq \emptyset$, we construct a *sequence* $S_X = \langle a'_1, a'_2, \ldots, a'_m \rangle$ *of attributes* in A, such that for every $1 \leqslant j < m$, $\left|IB_{\{a'_j\}}(X)\right| \leqslant \left|IB_{\{a'_{j+1}\}}(X)\right|$, where $IB_{\{a'_j\}}(X)$ is the inner boundary of X, for every singleton subset $\{a'_j\}$ of A.

Next, through decreasing the attribute set A gradually, we can determine a descending sequence of attribute subsets.

Definition 3.5 (*Descending sequence of attribute subsets*). Let $IS = (U,A,V,f)$ be an information system, where $A = \{a_1, a_2, \ldots, a_m\}$. For any $X \subseteq U$ and $X \neq \emptyset$, let $S_X = \langle a'_1, a'_2, \ldots, a'_m \rangle$ be the sequence of attributes defined above. Given a sequence $AS_X = \langle A_1, A_2, \ldots, A_m \rangle$ of attribute subsets, where $A_1, A_2, \ldots, A_m \subseteq A$. If $A_1 = A$, $A_m = \{a'_m\}$ and $A_{j+1} = A_j - \{a'_j\}$ for every $1 \leqslant j < m$, then we call AS_X a *descending sequence of attribute subsets* with respect to X in IS.

From the above definition, we can see that in $AS_X = \langle A_1, A_2, \ldots, A_m \rangle$, for every $1 \leqslant j < m$, A_{j+1} is the attribute subset transformed from A_j by removing the element a'_j from A_j, where a'_j is the jth element in sequence S_X. Therefore, given an information system $IS = (U,A,V,f)$, for every sequence S_X of attributes in A, we can uniquely determine a descending sequence AS_X of attribute subsets with respect to S_X.

Most current methods for outlier detection give a binary classification of objects (data records): is or is not an outlier. In real life, it is not so simple. For many scenarios, it is more meaningful to assign to each object a degree of being an outlier. Therefore, Breunig et al. proposed a method for identifying density-based local outliers (Breunig et al., 2000). He defined a *local outlier factor* (*LOF*) that indicates the degree of outlierness of an object using only the object's neighborhood. In this paper, similar to Breunig's method, we shall define a *hybrid outlier factor* (*HOF*), which can indicate the degree of outlierness for every object with respect to a given subset of universe in an information system. And before giving the definition of *hybrid outlier factor*, we define another preliminary concept – *deviation factor*, which indicates the degree of outlierness for every object under a given indiscernibility relation.

Definition 3.6 (*Deviation factor*). Let $IS = (U,A,V,f)$ be an information system, $X \subseteq U$ and $X \neq \emptyset$, where $A = \{a_1, a_2, \ldots, a_m\}$. For any $x \in X$ and $B \subseteq A$, the *deviation factor* $DF_X^B(x)$ of x with respect to X under relation $IND(B)$ is defined as

$$DF_X^B(x) = \frac{|\{y \in EB(X) : d(x,y) \leqslant d_1\}|}{|X|} + \frac{|\{y \in PB_B(X) : d(x,y) \geqslant d_2\}|}{|X|} + \frac{|\{y \in \underline{X}_B : d(x,y) \geqslant d_3\}|}{|X|},$$

where $EB(X)$, $PB_B(X)$ and $\underline{X}_B$ denote the exceptional boundary, principal boundary, and lower approximation of X under relation $IND(B)$, respectively; $d(x,y)$ denotes the distance between objects x and y under a given distance metric (in our experiment, the overlap metric in rough set theory is adopted (Jiang et al., 2009)); d_1, d_2 and d_3 are three given thresholds (In our experiment, we set $d_1 = |A|/3$, $d_2 = |A|/2$, and $d_3 = 0.9 \times |A|$, respectively. The reason for adopting that setting will be discussed in Section 4).

In the above definition, according to the distance-based method, we calculate the distance between x and each object in $EB(X)$, $PB_B(X)$, and $\underline{X}_B$, respectively (Knorr and Ng, 1998; Knorr et al., 2000). But differing from the traditional distance-based method, we adopt different attitudes to objects from different parts of X when determining outliers in X. That is, since objects in $EB(X)$ have the most likelihood to be an outlier, the more the objects in $EB(X)$ that are a short distance from object x (i.e. the distance between them is less than a given threshold d_1), the more the likelihood of x to be an outlier. On the other hand, since objects in $\underline{X}_B$ have the least likelihood to be an outlier, the more the objects in $\underline{X}_B$ that are a great distance from object x (i.e. the distance between them is more than a given threshold d_3), the more the likelihood of x to be an outlier. And for objects in $PB_B(X)$, the likelihood to be an outlier is moderate. Therefore, the more the objects in $PB_B(X)$ that are at a distance from object x (i.e. the distance between them is more than a given threshold d_2), the more the likelihood of x to be an outlier.

It should be noted that in the above definition, threshold d_3 must be bigger than d_2, since the objects in $\underline{X}_B$ have less likelihood to be an outlier than objects in $PB_B(X)$.

Definition 3.7 (*Hybrid outlier factor*). Let $IS = (U,A,V,f)$ be an information system, where $A = \{a_1,a_2,\ldots,a_m\}$. For any $X \subseteq U$ and $X \neq \emptyset$, let $AS_X = \langle A_1,A_2,\ldots,A_m\rangle$ be the descending sequence of attribute subsets with respect to X in IS. For any $x \in X$, the *hybrid outlier factor* $HOF_X(x)$ of x with respect to X in IS is defined as

$$HOF_X(x) = \frac{\sum_{j=1}^{m}\left(DF_X^{\{a_j\}}(x) \times W_X^{\{a_j\}}(x)\right) + \sum_{j=1}^{m}\left(DF_X^{A_j}(x) \times \frac{\sqrt{|A_j|}}{|A|}\right)}{2 \times |A|}, \tag{7}$$

where $DF_X^{A_j}(x)$ and $DF_X^{\{a_j\}}(x)$ are the deviation factors of x with respect to X, for every attribute subset $A_j \in AS_X$ and singleton subset $\{a_j\}$ of A, $1 \leqslant j \leqslant m$. And for every singleton subset $\{a_j\}$ of A, $W_X^{\{a_j\}} : X \to [0,1)$ is a weight function such that for any $x \in X$, $W_X^{\{a_j\}}(x) = 1 - \frac{\sqrt{|[x]_{\{a_j\}} \cap X|}}{|X|}$, $1 \leqslant j \leqslant m$. $[x]_B$ denotes the equivalence class of relation $IND(B)$ that contains element x.

The weight function $W_X^{\{a_j\}}$ in the above definition should express such an idea that outlier detection always concerns the minority group in the data set and objects belonging to the minority group are more likely to be outliers than those belonging to the majority group (Jiang et al., 2008, 2009). Since from the above definition, we can see that the more the weight, the more the rough outlier factor, objects belonging to the minority group should have more weight than objects belonging to the majority group. Therefore if the objects in X that are indiscernible with x are few, that is, the percentage of objects in X that are indiscernible with x is small, then we may consider x belonging to the minority group, and assign a big weight to x. Moreover, since $W_X^{\{a_j\}}$ is in fact an empirical function, the definition of $W_X^{\{a_j\}}$ is also affected by the experiment results in Section 4. We introduce a square root operation in $W_X^{\{a_j\}}$, since this operation can effectively improve the performance of our outlier detection method.

In this paper, we use the hybrid outlier factor $HOF_X(x)$ to indicate the degree of outlierness of object x. From Definition 3.7, we can see that $HOF_X(x)$ is based on the definition of deviation factor $DF_X^B(x)$ given in Definition 3.6. And from Definition 3.6, we can see that $DF_X^B(x)$ can only indicate the degree of outlierness of x under a given indiscernibility relation $IND(B)$. Obviously, it is insufficient to obtain the degree of outlierness of x by calculating $DF_X^B(x)$ only under one indiscernibility relation. Since using one indiscernibility relation to calculate $DF_X^B(x)$, we can only investigate x from one point of view. To obtain a more comprehensive and reliable result for the degree of outlierness of x, it is necessary to calculate $DF_X^B(x)$ under many different indiscernibility relations.

In rough set theory, given an information system $IS = (U,A,V,f)$, each attribute subset $B \subseteq A$ determines an indiscernibility relation $IND(B)$ on U, which is also deemed as the knowledge in IS (Wang et al., 2009, 2011). Therefore, there exist $2^{|A|}$ different indiscernibility relations (or knowledge) on U, which form a knowledge base in IS. Although there exist $2^{|A|}$ different indiscernibility relations in IS, it is impracticable to use all these relations to calculate $DF_X^B(x)$. Since the time complexity of our method will be exponential with respect to the number of attributes, it cannot handle data sets with more than 30 attributes. Therefore, as described in Definition 3.7, we adopt a compromise strategy. That is, we use each singleton subset $\{a_j\}$ of A to calculate $DF_X^{\{a_j\}}(x)$, and use each subset A_j in sequence AS_X to calculate $DF_X^{A_j}(x)$. The two kinds of results are further added together by using the corresponding weights, and the sum is exactly the hybrid outlier factor $HOF_X(x)$ of x.

Definition 3.8 (*BD-based outliers*). Let $IS = (U,A,V,f)$ be an information system, $X \subseteq U$ and $X \neq \emptyset$. Let v be a given threshold value, for any $x \in X$, if $HOF_X(x) > v$ then x is called a *BD* (*boundary and distance*)*-based outlier* with respect to X in IS, where $HOF_X(x)$ is the hybrid outlier factor of x with respect to X in IS.

In general, an outlier detection method only gives a degree of outlierness for each object. Before using the method to detect outliers, the users should first input an empirical value k to denote the number of outliers they expect. Obviously, the value k is varying for different data sets. And the users should determine the value k by many trials on the given data set.

In our outlier detection method, the setting of parameter v also depends on the value k provided by the users. If we have calculated the hybrid outlier factor $HOF_X(x)$ for every object $x \in X$, and sorted these objects according to their hybrid outlier factors in descending order. Let $\langle x_1,x_2,\ldots,x_n\rangle$ denote the sequence of objects in X after sorting, where $n = |X|$, $HOF_X(x_i) \geqslant HOF_X(x_{i+1})$, $1 \leqslant i < n$. Then we can set v as follows: $HOF_X(x_{k+1}) \leqslant v < HOF_X(x_k)$. By using this setting for v, we can guarantee that our method can find the k objects $x_1,x_2,\ldots,x_k$, whose degrees of outlierness are higher than those of other objects in X. And the k objects will be returned to the users as the outliers.

3.2. An example

Example 1. Given an information system $IS = (U,A,V,f)$, where $U = \{u_1,u_2,u_3,u_4,u_5,u_6\}$, $A = \{a,b,c\}$, as shown in Table 1.

Let $X = \{u_1,u_2,u_5,u_6\}$, and threshold $v = 0.4$. We assume that the distance metric is overlap metric in rough set theory (Jiang et al., 2009), and we set thresholds $d_1 = |A|/3$, $d_2 = |A|/2$, and $d_3 = 0.9 \times |A|$, respectively.

Table 1
Information system IS.

$U\backslash A$	a	b	c
u_1	0	0	0
u_2	1	0	1
u_3	0	2	0
u_4	2	0	0
u_5	0	2	1
u_6	1	1	2

For $u_1 \in X$, $[u_1]_{\{a\}} = \{u_1, u_3, u_5\}$, $[u_1]_{\{b\}} = \{u_1, u_2, u_4\}$, $[u_1]_{\{c\}} = \{u_1, u_3, u_4\}$;
For $u_2 \in X$, $[u_2]_{\{a\}} = \{u_2, u_6\}$, $[u_2]_{\{b\}} = \{u_1, u_2, u_4\}$, $[u_2]_{\{c\}} = \{u_2, u_5\}$;
For $u_5 \in X$, $[u_5]_{\{a\}} = \{u_1, u_3, u_5\}$, $[u_5]_{\{b\}} = \{u_3, u_5\}$, $[u_5]_{\{c\}} = \{u_2, u_5\}$;
For $u_6 \in X$, $[u_6]_{\{a\}} = \{u_2, u_6\}$, $[u_6]_{\{b\}} = \{u_6\}$, $[u_6]_{\{c\}} = \{u_6\}$.

From Definition 3.1, $IB_{\{a\}}(X) = \{u_1, u_5\}$; $IB_{\{b\}}(X) = \{u_1, u_2, u_5\}$; $IB_{\{c\}}(X) = \{u_1\}$.

From Proposition 3.1, $\underline{X}_{\{a\}} = \{u_2, u_6\}$; $\underline{X}_{\{b\}} = \{u_6\}$; $\underline{X}_{\{c\}} = \{u_2, u_5, u_6\}$.

And from definition 3.2, $EB(X) = IB_{\{a\}}(X) \cap IB_{\{a\}}(X) \cap IB_{\{a\}}(X) = \{u_1\}$.

Correspondingly, $PB_{\{a\}}(X) = IB_{\{a\}}(X) - EB(X) = \{u_5\}$; $PB_{\{b\}}(X) = \{u_2, u_5\}$; $PB_{\{c\}}(X) = \emptyset$.

Next, we construct two sequences from $IB_{\{a\}}(X)$, $IB_{\{b\}}(X)$, and $IB_{\{c\}}(X)$. From Definition 3.4, the sequence of attributes is $S_X = \langle c, a, b\rangle$. Correspondingly, the descending sequence of attribute subsets is $AS_X = \langle A_1, A_2, A_3\rangle = \langle\{a, b, c\}, \{a, b\}, \{b\}\rangle$.

Analogously, we can obtain that $IB_{A_1}(X) = \emptyset$; $IB_{A_2}(X) = \{u_5\}$; $IB_{A_3}(X) = \{u_1, u_2, u_5\}$. $PB_{A_1}(X) = \emptyset$; $PB_{A_2}(X) = \{u_5\}$; $PB_{A_3}(X) = \{u_2, u_5\}$. And $\underline{X}_{A_1} = \{u_1, u_2, u_5, u_6\}$; $\underline{X}_{A_2} = \{u_1, u_2, u_6\}$; $\underline{X}_{A_3} = \{u_6\}$.

For any two objects in X, the distance between them is as follows

$d(u_1, u_1) = 0$; $d(u_1, u_2) = 2$; $d(u_1, u_5) = 2$; $d(u_1, u_6) = 3$;
$d(u_2, u_1) = 2$; $d(u_2, u_2) = 0$; $d(u_2, u_5) = 2$; $d(u_2, u_6) = 2$;
$d(u_5, u_1) = 2$; $d(u_5, u_2) = 2$; $d(u_5, u_5) = 0$; $d(u_5, u_6) = 3$;
$d(u_6, u_1) = 3$; $d(u_6, u_2) = 2$; $d(u_6, u_5) = 3$; $d(u_6, u_6) = 0$.

(1) For $u_1 \in X$, from Definition 3.6, $DF_X^{\{a\}}(u_1) = \frac{|\{u_1\}|+|\{u_5\}|+|\{u_6\}|}{|X|} = \frac{1+1+1}{4} = \frac{3}{4}$; $DF_X^{\{b\}}(u_1) = \frac{|\{u_1\}|+|\{u_2,u_5\}|+|\{u_6\}|}{|X|} = \frac{1+2+1}{4} = 1$; $DF_X^{\{c\}}(u_1) = \frac{|\{u_1\}|+|\{u_6\}|}{|X|} = \frac{1+1}{4} = \frac{1}{2}$.
Analogously, we can obtain that $DF_X^{A_1}(u_1) = \frac{|\{u_1\}|+|\{u_6\}|}{|X|} = \frac{1+1}{4} = \frac{1}{2}$; $DF_X^{A_2}(u_1) = \frac{|\{u_1\}|+|\{u_5\}|+|\{u_6\}|}{|X|} = \frac{1+1+1}{4} = \frac{3}{4}$; $DF_X^{A_3}(u_1) = \frac{|\{u_1\}|+|\{u_2,u_5\}|+|\{u_6\}|}{|X|} = \frac{1+2+1}{4} = 1$.
Hence the hybrid outlier factor of u_1 with respect to X is as follows

$$HOF_X(u_1) = \frac{\frac{3}{4}\times\left(1-\frac{\sqrt{2}}{4}\right) + 1\times\left(1-\frac{\sqrt{2}}{4}\right) + \frac{1}{2}\times\left(1-\frac{1}{4}\right)}{2\times 3} + \frac{\frac{1}{2}\times\frac{\sqrt{3}}{3} + \frac{3}{4}\times\frac{\sqrt{2}}{3} + 1\times\frac{1}{3}}{2\times 3} \approx 0.4137 > v.$$

Therefore u_1 is a BD-based outlier with respect to X in IS.

(2) For $u_2 \in X$, from Definition 3.6, $DF_X^{\{a\}}(u_2) = \frac{|\{u_5\}|}{|X|} = \frac{1}{4}$; $DF_X^{\{b\}}(u_2) = \frac{|\{u_5\}|}{|X|} = \frac{1}{4}$; $DF_X^{\{c\}}(u_2) = \frac{0}{|X|} = \frac{0}{4} = 0$. Analogously, we can obtain that $DF_X^{A_1}(u_2) = \frac{0}{|X|} = \frac{0}{4} = 0$; $DF_X^{A_2}(u_2) = \frac{|\{u_5\}|}{|X|} = \frac{1}{4}$; $DF_X^{A_3}(u_2) = \frac{|\{u_5\}|}{|X|} = \frac{1}{4}$. Hence we can obtain that $HOF_X(u_2) \approx 0.0874 < v$. Therefore u_2 is not a BD-based outlier with respect to X in IS.

(3) For $u_5 \in X$, from Definition 3.6, $DF_X^{\{a\}}(u_5) = \frac{|\{u_6\}|}{|X|} = \frac{1}{4}$; $DF_X^{\{b\}}(u_5) = \frac{|\{u_2\}|+|\{u_6\}|}{|X|} = \frac{2}{4}$; $DF_X^{\{c\}}(u_5) = \frac{|\{u_6\}|}{|X|} = \frac{1}{4}$. Analogously, we can obtain that $DF_X^{A_1}(u_5) = \frac{|\{u_6\}|}{|X|} = \frac{1}{4}$; $DF_X^{A_2}(u_5) = \frac{|\{u_6\}|}{|X|} = \frac{1}{4}$; $DF_X^{A_3}(u_5) = \frac{|\{u_2\}|+|\{u_6\}|}{|X|} = \frac{2}{4}$. Hence we can obtain that $HOF_X(u_5) \approx 0.1878 < v$. Therefore u_5 is not a BD-based outlier with respect to X in IS.

(4) For $u_6 \in X$, from Definition 3.6, $DF_X^{\{a\}}(u_6) = \frac{|\{u_5\}|}{|X|} = \frac{1}{4}$; $DF_X^{\{b\}}(u_6) = \frac{|\{u_2,u_5\}|}{|X|} = \frac{2}{4}$; $DF_X^{\{c\}}(u_6) = \frac{|\{u_5\}|}{|X|} = \frac{1}{4}$.
Analogously, we can obtain that $DF_X^{A_1}(u_6) = \frac{|\{u_1,u_5\}|}{|X|} = \frac{2}{4}$; $DF_X^{A_2}(u_6) = \frac{|\{u_5\}|+|\{u_1\}|}{|X|} = \frac{2}{4}$; $DF_X^{A_3}(u_6) = \frac{|\{u_2,u_5\}|}{|X|} = \frac{2}{4}$.
Hence we can obtain that $HOF_X(u_6) \approx 0.2359 < v$. Therefore u_6 is not a BD-based outlier with respect to X in IS. □

3.3. Algorithm for detecting BD-based outliers

Algorithm 1.

```
Input: information system $IS = (U, A, V, f)$ and $X \subseteq U$, where
    $|U| = n$, $A = \{a_1, a_2, \ldots, a_m\}$, and $|X| = n_X$; thresholds $v$, $d_1$,
    $d_2$, and $d_3$
Output: a set $E$ of BD-based outliers with respect to $X$ in $IS$
Initialization: Let $E = \emptyset$

For every $a \in A$
{
    Sort all objects from $U$ according to a given order (e.g. the lexicographical order) on domain $V_a$ of attribute $a$;
    Determine the partition $U/IND(\{a\})$;
    Calculate the inner boundary $IB_{\{a\}}(X)$ of $X$ under relation $IND(\{a\})$;
    Calculate the lower approximation $\underline{X}_{\{a\}}$ of $X$ under $IND(\{a\})$
}
Let $EB(X) = \bigcap_{i=1}^{m} IB_{\{a_i\}}(X)$ be the exceptional boundary of $X$;
Determine the sequence $S_X = \{a'_1, a'_2, \ldots, a'_m\}$ of attributes in $A$, where for each $1 \leqslant j < m$, $|IB_{\{a'_j\}}(X)| \leqslant |IB_{\{a'_{j+1}\}}(X)|$;
Determine descending sequence $AS_X = \{A_1, \ldots, A_m\}$ of attribute subsets;
For $1 \leqslant j \leqslant m$
{
    Sort all objects from $U$ according to a given order (e.g. the lexicographical order) on domain $V_{A_j}$ of attribute subset $A_j$;
    Determine the partition $U/IND(A_j)$;
    Calculate the inner boundary $IB_{A_j}(X)$ of $X$ under relation $IND(A_j)$;
    Calculate the lower approximation $\underline{X}_{A_j}$ of $X$ under $IND(A_j)$
    Calculate the principal boundary $PB_{A_j}(X)$ of $X$ under $IND(A_j)$;
    Calculate the principal boundary $PB_{\{a_j\}}(X)$ of $X$ under $IND(\{a_j\})$
}
For every $x \in X$
{
    For every $y \in X$, calculate the distance $d(x, y)$ between $x$ and $y$;
    For $1 \leqslant j \leqslant m$
    {
        Calculate $DF_X^{\{a_j\}}(x)$, the deviation factor of $x$ with respect to $X$ under relation $IND(\{a_j\})$;
        Assign a weight $W_X^{\{a_j\}}(x)$ to $x$, where $W_X^{\{a_j\}}(x) = 1 - \frac{\sqrt{|[x]_{\{a_j\}} \cap X|}}{|X|}$;
        Calculate $DF_X^{A_j}(x)$, the deviation factor of $x$ with respect to $X$ under relation $IND(A_j)$
    }
    Calculate $HOF_X(x)$, the hybrid outlier factor of $x$ with respect to $X$;
    If $HOF_X(x) > v$ then $E = E \cup \{x\}$
}
Return $E$.
```

Usually, the time complexity for calculating the partition induced by an indiscernibility relation is $O(n^2)$, where n is the cardinalities of universe U. In algorithm 1, we use a method proposed by Nguyen and Nguyen (1996) which can calculate the partition induced by an indiscernibility relation $IND(B)$ in $O(k \times n\log n)$ time, where k and n are the cardinalities of B and U respectively.

In the worst case, the time complexity of algorithm 1 is $O((m \times n_X^2) + (m^2 \times n\log n))$, and its space complexity is $O(m \times (n + m))$, where m, n, n_X are the cardinalities of A, U and X respectively.

4. Experimental results

4.1. Experiment design

To evaluate BD-based method for outlier detection, we ran our algorithm on two real life data sets (*lymphography* and *cancer*) obtained from the UCI Machine Learning Repository (Bay, 1999). Since in our previous papers (Jiang et al., 2005, 2008, 2009), we have proposed three different methods for outlier detection in rough set theory. In this section, we compare the performance of BD-based method with the three methods on identifying true outliers. In addition, we compare the performance of BD-based method with the other two methods proposed by other researchers: GrC (granular computing)-based and KNN-based outlier detection methods (Ramaswamy et al., 2000; Chen et al., 2008).

First we give a brief description for these methods.

In (Jiang et al., 2008), we presented a RMF (rough membership function)-based method for outlier detection, in which we introduced a rough outlier factor (ROF) to indicate the degree of outlierness for every object in an information system. And the rough outlier factor is defined by virtue of the notion of rough membership function in rough sets. The objects whose degrees of membership with respect to a given subset X of universe U are small have more likelihood of being an outlier, i.e. the rough outlier factors of them are high.

And in (Jiang et al., 2009), we introduced distance-based outlier detection to rough set theory and proposed the definitions of distance metrics for distance-based outlier detection in rough set theory. Since in traditional distance-based outlier detection, being an outlier is regarded as a binary property, we only know that an object is an outlier or not. In paper (Jiang et al., 2008), in order to compare distance-based method with other methods, we revised the definitions of distance-based outlier detection by introducing a *distance outlier factor (DOF)*, which can be defined as follows.

Definition 4.1 (*Distance outlier factor*). Given an information system $IS = (U,A,V,f)$ and $X \subseteq U$. For any object $x \in X$, the percentage of the objects in X lie greater than d_0 from x is called the *distance outlier factor of* x with respect to X in IS, denoted by

$$DOF_X(x) = \frac{|\{y \in X : d(x,y) > d_0\}|}{|X|}, \tag{8}$$

where $d(x,y)$ denotes the distance between objects x and y under a given distance metric in rough set theory, d_0 is a given threshold (In our experiment, the overlap metric in rough set theory is adopted (Jiang et al., 2009), and we set $d_0 = |A|/2$, where $|A|$ denotes the cardinality of set A).

Furthermore, in (Jiang et al., 2005), we proposed a boundary-based method for outlier detection. Based on the notions of boundary region and lower approximation in rough set theory, we first defined the notions of *inner boundary* and *boundary degree*. Then by virtue of boundary degree, we defined the notion of *exceptional degree* for every object in X. Similar to HOF and DOF, the *exceptional degree* of an object indicates the degree of outlierness for that object. And in paper (Jiang et al., 2008), in order to compare boundary-based method with other methods and obtain a finer effect for boundary-based method, we revised the definitions for boundary degree and exceptional degree in (Jiang et al., 2005), which can be defined as follows.

Definition 4.2 (*Boundary degree*). Given an information system $IS = (U,A,V,f)$ and $X \subseteq U$ $(X \neq \emptyset)$, where $A = \{a_1,\ldots,a_m\}$. Let $IB = \{IB_1,IB_2,\ldots,IB_m\}$ be the set of all inner boundaries of X under each equivalence relation $IND(\{a_j\})$, $1 \leqslant j \leqslant m$. For every object $x \in X$, the *boundary degree* of x with respect to X in IS is defined as:

$$BD_X(x) = \sum_{j=1}^{m} \left(f(x,IB_j) \times W_X^{\{a_j\}}\right), \tag{9}$$

where f is a characteristic function for set IB_j (that is, if $x \in IB_j$ then $f(x,IB_j) = 1$ else $f(x,IB_j) = 0$). $W_X^{\{a_j\}} : X \to [0,1)$ is a weight function such that for any $x \in X$, $W_X^{\{a_j\}}(x) = 1 - (|[x]_{\{a_j\}} \cap X|/|X|)$, $1 \leqslant j \leqslant m$. $[x]_{\{a_j\}} = \{u \in U : f(u,a_j) = f(x,a_j)\}$ denotes the equivalence class of relation $IND(\{a_j\})$ that contains element x and $|M|$ denotes the cardinality of set M.

Definition 4.3 (*Exceptional degree*). Given an information system $IS = (U,A,V,f)$ and $X \subseteq U$ $(X \neq \emptyset)$, where $A = \{a_1,\ldots,a_m\}$. Let $IB = \{IB_1,IB_2,\ldots,IB_m\}$ be the set of all inner boundaries of X under each equivalence relation $IND(\{a_j\})$, $1 \leqslant j \leqslant m$. For any object $x \in X$, the cardinality of set IB divided by the boundary degree of x with respect to X is called the *exceptional degree of* x with respect to X in IS, denoted by

$$ED_X(x) = \frac{BD_X(x)}{|IB|}. \tag{10}$$

In (Chen et al., 2008), Chen et al. presented a GrC (granular computing)-based method for outlier definition and outlier detection, which exploited the granular computing model using information tables to detect outliers. The main idea of the method is that an object has more likelihood of being an outlier if the granules containing it have a high degree of outlierness. Moreover, in (Ramaswamy et al., 2000), Ramaswamy et al. proposed a KNN (the kth nearest neighbor)-based method for outlier detection, which is based on the distance of a point from its kth nearest neighbor. They ranked each point on the basis of its distance to its kth nearest neighbor and declared the top n points in this ranking to be outliers. In our experiment, for the KNN-based method, the results were obtained by using the 5th nearest neighbor and the overlap metric in rough set theory (Jiang et al., 2008).

4.2. Lymphography data

The first is the lymphography data set, which can be found in the UCI Machine Learning Repository (Bay, 1999). It contains 148 instances (or objects) with 19 attributes (including the class attribute). The 148 instances are partitioned into 4 classes: "normal find" (2 or 1.35%), "metastases" (81 or 54.73%), "malign lymph" (61 or 41.22%) and "fibrosis" (4 or 2.7%).

In (Aggarwal and Yu, 2001), Aggarwal and Yu proposed a practicable way to test the effectiveness of an outlier detection method. That is, we can run the outlier detection method on a given data set and test the percentage of points (instances) which belonged to one of the rare classes. Aggarwal considered those kinds of class labels which occurred in less than 5% of the data set as rare labels. And those points belonged to the rare class are considered as outliers. If the outlier detection method works well, we expect that such abnormal classes would be over-represented in the set of points found.

Table 2
Experimental results with respect to X_1 in IS_L.

Top ratio (number of objects)	Number of rare classes included (coverage)					
	BD	RMF	RB	DIS	GrC	KNN
$X_1 : \lvert X_1\rvert = 122,\ \lvert R_{X_1}\rvert = 4$						
2%(2)	2(50%)	2(50%)	2(50%)	2(50%)	2(50%)	2(50%)
3%(4)	4(100%)	3(75%)	2(50%)	4(100%)	3(75%)	2(50%)
4%(5)	4(100%)	4(100%)	2(50%)	4(100%)	4(100%)	2(50%)
6%(7)	4(100%)	4(100%)	2(50%)	4(100%)	4(100%)	3(75%)
6.5%(8)	4(100%)	4(100%)	2(50%)	4(100%)	4(100%)	4(100%)
84%(102)	4(100%)	4(100%)	3(75%)	4(100%)	4(100%)	4(100%)
88%(107)	4(100%)	4(100%)	4(100%)	4(100%)	4(100%)	4(100%)
Accuracy	100%	80%	3.7%	100%	80%	50%

In the lymphography data set, classes 1 and 4 (“normal find” and “fibrosis”) should be regarded as rare class labels since they occur in less than 5% of the data set. In our experiment, data in the lymphography data set is input into an information system $IS_L = (U,A,V,f)$, where U contains all the 148 instances of lymphography data set and A contains 18 attributes of lymphography data set (not including the class attribute). We consider detecting outliers (rare classes) with respect to four subsets $X_1,\ldots,X_4$ of U, respectively, where

(1) $X_1 = \{x \in U : f(x, \text{bl\_lymph\_c}) = 1\}$;
(2) $X_2 = \{x \in U : f(x, \text{early\_uptake}) = 1 \vee f(x, \text{bl\_affere}) = 1\}$;
(3) $X_3 = \{x \in U : f(x, \text{spec\_froms}) = 3 \vee f(x, \text{dislocation}) = 1\}$;
(4) $X_4 = \{x \in U : f(x, \text{bl\_affere}) = 1\}$.

X_1 contains those objects of U whose values on attribute “bl_lymph_c” equal to 1; X_2 contains those objects of U whose values on attribute “early_uptake” equal 1 and those objects of U whose values on attribute “bl_affere” equal 1; …Moreover, we use R_{X_j} to denote the set of all objects in X_j that belong to one of the rare classes (class 1 or 4), $1 \leqslant j \leqslant 4$.

The results from the six different outlier detection methods on the lymphography data set are summarized in Tables 2–5.

In Tables 2–5, “$|X_j|$” denotes the number of objects in X_j, “$|R_{X_j}|$” denotes the number of outliers in X_j, $1 \leqslant j \leqslant 4$. And “BD”, “RMF”, “RB”, “DIS”, “GrC”, “KNN” denote BD-based, RMF-based, boundary-based, distance-based, GrC-based and KNN-based outlier detection methods, respectively. For every objects in X_j, the degree of outlierness with respect to X_j is calculated by using the six outlier detection methods, respectively. For each outlier detection method, the “Top Ratio (Number of Objects)” denotes the percentage (number) of the objects selected from X_j whose degrees of outlierness with respect to X_j calculated by the method are higher than those of other objects in X_j. And if we use a subset $Y_j \subseteq X_j$ to contain all those objects selected from X_j, then the “ Number of Rare Classes Included” is the number of objects in Y_j that are outliers. The “Coverage” is the ratio of the “Number of Rare Classes Included” to the number of outliers in X_j (i.e., $|R_{X_j}|$), $1 \leqslant j \leqslant 4$ (Jiang et al., 2008; He et al., 2005).

Moreover, the “Accuracy” is a measure of the effectiveness of each outlier detection method on the current data set, which can be defined as:

$$Accuracy = \frac{|R_{X_j}|}{|R_{X_j}| + |F_{X_j}|}, \tag{11}$$

where $|F_{X_j}|$ denotes the number of objects in X_j, which are not outliers, but their degrees of outlierness calculated by the given outlier detection method are higher than that of some outlier in X_j, $1 \leqslant j \leqslant 4$.

From Tables 2–5, we can see that for the lymphography data set, BD-based method has the best performance. Since for subsets $X_1,\ldots$, and X_4, the accuracies of BD-based method are always the highest among all outlier detection methods. Especially for X_1 and X_4, the accuracies of BD-based method are 100%, which means that there does not exist any misjudgment when applying our method to the two subsets. For instance, in Table 5, when the Top Ratio (Number of Objects) equals 6%(4), the “ Number of Rare Classes Included” for our method is 4, that is, the 4 objects selected by our method which have the highest degree of outlierness are all outliers. However, for RMF, RB, GrC and KNN, only 3, 2, 3 and 2 outliers are found, respectively. There are misjudgments more or less for these methods. From another point of view, to find all outliers in X_4, our method only need to check 6% of objects in X_4. And for RMF, RB, GrC and KNN, to achieve that aim, they need to check 9%, 65%, 9% and 10% of objects in X_4, respectively. This also demonstrates the effectiveness of our method for outlier detection on the lymphography data set.

Table 6 gives the information about the mean and standard deviation of the accuracies of different outlier detection method on subsets $X_1,\ldots$, and X_4.

From Table 6, we can see that although the standard deviation of the accuracies of our method is higher than those of boundary-based, GrC-based and KNN-based methods, the average accuracy of our method is markedly higher than those of them. Therefore, we can further conclude that BD-based method performs better than all other methods on the lymphography data set. And the performances of RMF-based, distance-based and GrC-based methods are close, they perform markedly better than boundary-based and KNN-based methods, where the worst is the boundary-based method.

Table 3
Experimental results with respect to X_2 in IS_L.

Top ratio (number of objects)	Number of rare classes included (coverage)					
	BD	RMF	RB	DIS	GrC	KNN
$X_2 : \lvert X_2\rvert = 90,\ \lvert R_{X_2}\rvert = 5$						
2%(2)	2(40%)	2(40%)	2(40%)	2(40%)	2(40%)	2(40%)
3%(3)	3(60%)	3(60%)	3(60%)	3(60%)	3(60%)	3(60%)
4%(4)	4(80%)	3(60%)	3(60%)	3(60%)	4(80%)	3(60%)
5%(5)	4(80%)	3(60%)	3(60%)	4(80%)	4(80%)	3(60%)
8%(7)	5(100%)	3(60%)	3(60%)	4(80%)	5(100%)	3(60%)
10%(9)	5(100%)	4(80%)	3(60%)	4(80%)	5(100%)	3(60%)
11%(10)	5(100%)	4(80%)	3(60%)	4(80%)	5(100%)	4(80%)
12%(11)	5(100%)	4(80%)	3(60%)	4(80%)	5(100%)	5(100%)
14%(13)	5(100%)	5(100%)	3(60%)	5(100%)	5(100%)	5(100%)
66%(59)	5(100%)	5(100%)	4(80%)	5(100%)	5(100%)	5(100%)
70%(63)	5(100%)	5(100%)	5(100%)	5(100%)	5(100%)	5(100%)
Accuracy	71.4%	38.5%	7.9%	38.5%	71.4%	45.5%

Table 4
Experimental results with respect to X_3 in IS_L.

Top ratio (number of objects)	Number of rare classes included (coverage)					
	BD	RMF	RB	DIS	GrC	KNN
$X_3 : \|X_3\| = 105,\ \|R_{X_3}\| = 5$						
2%(2)	2(40%)	2(40%)	2(40%)	2(40%)	2(40%)	2(40%)
3%(3)	3(60%)	3(60%)	3(60%)	3(60%)	3(60%)	3(60%)
4%(4)	4(80%)	4(80%)	3(60%)	4(80%)	4(80%)	3(60%)
5%(5)	4(80%)	4(80%)	3(60%)	4(80%)	4(80%)	4(80%)
7%(7)	5(100%)	5(100%)	3(60%)	4(80%)	4(80%)	4(80%)
8%(8)	5(100%)	5(100%)	3(60%)	5(100%)	5(100%)	4(80%)
11%(12)	5(100%)	5(100%)	4(80%)	5(100%)	5(100%)	4(80%)
12%(13)	5(100%)	5(100%)	4(80%)	5(100%)	5(100%)	5(100%)
24%(25)	5(100%)	5(100%)	5(100%)	5(100%)	5(100%)	5(100%)
Accuracy	71.4%	71.4%	20%	62.5%	62.5%	38.5%

Table 5
Experimental results with respect to X_4 in IS_L.

Top ratio (number of objects)	Number of rare classes included (coverage)					
	BD	RMF	RB	DIS	GrC	KNN
$X_4 : \|X_4\| = 66,\ \|R_{X_4}\| = 4$						
2%(1)	1(25%)	1(25%)	1(25%)	1(25%)	1(25%)	1(25%)
4%(3)	3(75%)	3(75%)	2(50%)	3(75%)	3(75%)	2(50%)
6%(4)	4(100%)	3(75%)	2(50%)	4(100%)	3(75%)	2(50%)
9%(6)	4(100%)	4(100%)	2(50%)	4(100%)	4(100%)	3(75%)
10%(7)	4(100%)	4(100%)	2(50%)	4(100%)	4(100%)	4(100%)
20%(13)	4(100%)	4(100%)	3(75%)	4(100%)	4(100%)	4(100%)
65%(43)	4(100%)	4(100%)	4(100%)	4(100%)	4(100%)	4(100%)
Accuracy	100%	66.7%	9.3%	100%	66.7%	57.1%

Table 6
Statistical information of the accuracies of different methods on lymphography data set.

Outlier detection methods	Mean	Standard deviation
BD	85.7%	14.3%
RMF	64.2%	15.6%
RB	10.2%	6%
DIS	75.3%	26.2%
GrC	70.2%	6.5%
KNN	47.8	6.8

Since in BD-based method, to calculate the degree of outlierness for each object x, we should first calculate the deviation factor of x. And we need to determine the values of three parameters d_1, d_2 and d_3 when calculating the deviation factor of x. In the experiments, we first assign an original empirical value to each parameter such that $d_1 < d_2 < d_3$, and test the performances of our method on the lymphography and the cancer data sets, respectively. Then we repeatedly adjust the value of each parameter to obtain better performances for our method on the two data sets. Finally, we adopt a compromise proposal for the setting of each parameter such that $d_1 = |A|/3$, $d_2 = |A|/2$ and $d_3 = 0.9 \times |A|$ (where A is the set of attributes), which can guarantee relatively good performances for our method on both of the two data sets.

4.3. Wisconsin breast cancer data

The Wisconsin breast cancer data set (or called cancer data set) is found in the UCI Machine Learning Repository (Bay, 1999). The data set contains 699 instances with 9 continuous attributes (not including the class attribute). Each instances is labeled as *benign* (458 or 65.5%) or *malignant* (241 or 34.5%). Here we follow the experimental technique of Harkins et al. by removing some of the *malignant* instances to form a very unbalanced distribution (Harkins et al., 2002; He et al., 2005). The resultant data set had 39 (8%) *malignant* instances and 444 (92%) *benign* instances. Moreover, the 9 continuous attributes in the data set are transformed into categorical attributes, respectively.[1] Here *malignant* instances are deemed as outliers (He et al., 2005).

Similar to the treatment for the lymphography data set, data in the cancer data set is also input into an information system $IS_W = (U', A', V', f')$, where U' contains all the 483 instances of the data set and A' contains 9 categorical attributes of the data set (not including the class attribute). We consider detecting outliers (*malignant* instances) with respect to four subsets $X'_1, \ldots, X'_4$ of U', respectively, where

(1) $X'_1 = \{x \in U' : f'(x, \text{Clump\_thickness}) = 5\}$;
(2) $X'_2 = \{x \in U' : f'(x, \text{Marginal\_Adhesion}) = 2\}$;
(3) $X'_3 = \{x \in U' : f'(x, \text{Bland\_Chromatine}) = 3\}$;
(4) $X'_4 = \{x \in U' : f'(x, \text{Mitoses}) = 1\}$.

X'_1 contains those objects of U' whose values on attribute "Clump_thickness" equal to 5; ...Moreover, we use $R_{X'_j}$ to denote the set of all objects in X'_j that are *malignant*, $1 \leqslant j \leqslant 4$.

The results from the six different outlier detection methods on the cancer data set are summarized in Tables 7–10.

Tables 7–10 are similar to Tables 2–5, except that the "Number of Malignant Instances Included" is the number of objects in Y'_j that are outliers, where $Y'_j \subseteq X'_j$ contains those objects selected from X'_j that are specified as top-k ($k = |Y'_j|$) outliers by one of the six outlier detection methods. And the "Coverage" is the ratio of the "Number of Malignant Instances Included" to the number of outliers in X'_j (i.e., $|R_{X'_j}|$), $1 \leqslant j \leqslant 4$ (Jiang et al., 2008; He et al., 2005).

From Tables 7–10, we can see that for the cancer data set, BD-based method also has the best performance. Since for sub-

[1] The resultant data set is public available at: http://research.cmis.csiro.au/rohanb/outliers/breast-cancer/.

Table 7
Experimental results with respect to X'_1 in IS_W.

Top ratio (number of objects)	Number of malignant instances included (coverage)					
	BD	RMF	RB	DIS	GrC	KNN
$X'_1 : \|X'_1\| = 87,\ \|R_{X'_1}\| = 4$						
2%(2)	2(50%)	2(50%)	2(50%)	2(50%)	2(50%)	2(50%)
3%(3)	3(75%)	3(75%)	3(75%)	2(50%)	3(75%)	3(75%)
5%(4)	4(100%)	3(75%)	3(75%)	3(75%)	3(75%)	4(100%)
6%(5)	4(100%)	4(100%)	3(75%)	3(75%)	3(75%)	4(100%)
7%(6)	4(100%)	4(100%)	3(75%)	4(100%)	4(100%)	4(100%)
8%(7)	4(100%)	4(100%)	4(100%)	4(100%)	4(100%)	4(100%)
Accuracy	100%	80%	57.1%	66.7%	66.7%	100%

Table 8
Experimental results with respect to X'_2 in IS_W.

Top ratio (number of objects)	Number of malignant instances included (coverage)					
	BD	RMF	RB	DIS	GrC	KNN
$X'_2 : \|X'_2\| = 42,\ \|R_{X'_2}\| = 5$						
7%(3)	3(60%)	3(60%)	3(60%)	2(40%)	3(60%)	3(60%)
10%(4)	4(80%)	4(80%)	3(60%)	3(60%)	4(80%)	4(80%)
15%(6)	5(100%)	4(80%)	3(60%)	5(100%)	4(80%)	5(100%)
17%(7)	5(100%)	4(80%)	4(80%)	5(100%)	4(80%)	5(100%)
20%(8)	5(100%)	5(100%)	4(80%)	5(100%)	5(100%)	5(100%)
21%(9)	5(100%)	5(100%)	5(100%)	5(100%)	5(100%)	5(100%)
Accuracy	83.3%	62.5%	55.6%	83.3%	62.5%	83.3%

Table 9
Experimental results with respect to X'_3 in IS_W.

Top ratio (number of objects)	Number of malignant instances included (coverage)					
	BD	RMF	RB	DIS	GrC	KNN
$X'_3 : \|X'_3\| = 133,\ \|R_{X'_3}\| = 8$						
1.5%(2)	2(25%)	2(25%)	2(25%)	1(12.5%)	2(25%)	2(25%)
3%(4)	4(50%)	3(37.5%)	3(37.5%)	3(37.5%)	3(37.5%)	4(50%)
4%(5)	4(50%)	4(50%)	4(50%)	4(50%)	4(50%)	4(50%)
5%(7)	6(75%)	6(75%)	5(62.5%)	5(62.5%)	6(75%)	5(62.5%)
6%(8)	6(75%)	7(87.5%)	5(62.5%)	6(75%)	6(75%)	5(62.5%)
9%(12)	8(100%)	7(87.5%)	6(75%)	8(100%)	7(87.5%)	7(87.5%)
10%(13)	8(100%)	8(100%)	7(87.5%)	8(100%)	7(87.5%)	8(100%)
10.5%(14)	8(100%)	8(100%)	7(87.5%)	8(100%)	8(100%)	8(100%)
11%(15)	8(100%)	8(100%)	8(100%)	8(100%)	8(100%)	8(100%)
Accuracy	66.7%	61.5%	53.3%	66.7%	57.1%	61.5%

Table 10
Experimental results with respect to X'_4 in IS_W.

Top ratio (number of objects)	Number of malignant instances included (coverage)					
	BD	RMF	RB	DIS	GrC	KNN
$X'_4 : \|X'_4\| = 454,\ \|R_{X'_4}\| = 23$						
1%(5)	4(17%)	4(17%)	4(17%)	4(17%)	5(22%)	4(17%)
2%(9)	8(35%)	8(35%)	7(30%)	6(26%)	8(35%)	7(30%)
3%(14)	11(48%)	12(52%)	11(48%)	10(43%)	12(52%)	10(43%)
4%(18)	15(65%)	15(65%)	13(57%)	12(52%)	14(61%)	12(52%)
5%(23)	16(70%)	18(78%)	18(78%)	15(65%)	17(74%)	16(70%)
6%(27)	18(78%)	20(87%)	20(87%)	18(78%)	19(83%)	19(83%)
7%(32)	23(100%)	22(96%)	21(91%)	23(100%)	22(96%)	23(100%)
7.2%(33)	23(100%)	23(100%)	21(91%)	23(100%)	23(100%)	23(100%)
10%(45)	23(100%)	23(100%)	22(96%)	23(100%)	23(100%)	23(100%)
12%(54)	23(100%)	23(100%)	23(100%)	23(100%)	23(100%)	23(100%)
Accuracy	71.9%	69.7%	42.6%	71.9%	69.7%	71.9%

sets $X'_1, \ldots,$ and X'_4, the accuracies of BD-based method are always the highest among all outlier detection methods. Especially for X'_1, the accuracy of BD-based method is 100%, which means that there does not exist any misjudgment when applying our method to the subset. In Table 7, when the Top Ratio (Number of Objects) equals 5%(4), the " Number of Malignant Instances Included" for our method is 4, that is, the 4 objects selected by our method which have the highest degree of outlierness are all outliers. However, for RMF, RB, DIS and GrC, only 3 outliers are found. From another point of view, to find all outliers in X'_1, our method only need to check 5% of objects in X'_1. And for RMF, RB, DIS and GrC, to achieve that aim, they need to check

Table 11
Statistical information of the accuracies of different methods on cancer data set.

Outlier detection methods	Mean	Standard deviation
BD	80.5%	12.8%
RMF	68.4%	7.4%
RB	52.2%	5.7%
DIS	72.2%	6.8%
GrC	64%	4.7%
KNN	79.2	14.3

6%, 8%, 7% and 7% of objects in X'_1, respectively. This also demonstrates the effectiveness of our method for outlier detection on the cancer data set.

Table 11 gives the information about the mean and standard deviation of the accuracies of different outlier detection methods on subsets $X'_1, \ldots,$ and X'_4.

From Table 11, we can see that although the standard deviation of the accuracies of our method is higher than those of RMF-based, boundary-based, distance-based and GrC-based methods, the average accuracy of our method is markedly higher than those of them. And the average accuracy of KNN-based method is close to that of our method, but the standard deviation of the accuracies of it is higher than that of our method. Therefore, we can further conclude that BD-based method performs better than all other methods on the cancer data set. The next one is the KNN-based method. And the performances of RMF-based and distance-based methods are close, they perform better than boundary-based and GrC-based methods, where the worst is still the boundary-based method.

5. Discussion

Roughly speaking, the current methods to outlier detection can be classified into the following five categories (Kovács et al., 2004).

(1) *Distribution-based method* is the classical method in statistics. It is based on some standard distribution model (Normal, Poisson, etc.) and those objects which deviate from the model are recognized as outliers (Barnett and Lewis, 1994). Its greatest disadvantage is that the distribution of the measurement data is unknown in practice.
Since our method does not require knowledge about the distribution of the measurement data, it can counter the main limitations of the distribution-based method. In comparison with distribution-based method, our method avoids the excessive computation associated with fitting an observed distribution into some standard distribution and in selecting discordancy tests.

(2) *Depth-based method* is based on computational geometry and compute different layers of k-d convex hulls and flags objects in the outer layer as outliers (Johnson et al., 1998). However, it is a well-known fact that the algorithms employed suffer from the dimensionality curse and cannot cope with large k. Comparing with the depth-based method, the time complexity of our method is relatively low, which is more suitable to deal with large data sets.

(3) *Clustering-based method* classifies the input data. It detects outliers as by-products (Jain et al., 1999). However, since the main objective is clustering, it is not optimized for outlier detection. Differing from the clustering-based method, our method is designed specially for outlier detection.

(4) *Distance-based method* was originally proposed by Knorr and Ng (1998), Knorr et al. (2000). In traditional distance-based outlier detection, being an outlier is regarded as a binary property, we only know that an object is an outlier or not. In our method, we introduce a notion called hybrid outlier factor, which can indicate the degree of outlierness for each object. Moreover, although our method also calculates the distances between objects, we adopt different attitudes to objects from different parts of the given data set when detecting outliers based on distance. The experimental results in Section 4 show that this strategy can improve the effectiveness of distance-based method for outlier detection.

(5) *Density-based method* was originally proposed by Breunig et al. (2000). A local outlier factor (LOF) is assigned to each sample based on their local neighborhood density. Samples with high LOF value are identified as outliers. The disadvantage of this solution is that it is very sensitive to parameters defining the neighborhood. Unlike density-based outlier detection, our method does not require such parameters. And our method can also find the local outliers as the density-based method does, since the definition for outliers in our method has a characteristic that is ignored by most current definitions for outliers. That is, for a given data set, we consider detecting outliers with respect to any subset of the given data set, which makes it possible to find the local outliers.

Moreover, in (Jiang et al., 2005), based on the notions in rough sets, we proposed a boundary-based method for outlier detection. From the experiment results in Section 4, we can see that the performance of boundary-based method is markedly worse than those of other methods. By combining the opinions from distance-based and boundary-based methods for outlier detection, our method can effectively improve the performance of boundary-based method for outlier detection. On the other hand, our method can also solve the main problems of distance-based method.

6. Conclusion

Outlier detection is becoming critically important in many areas. In this paper, we presented a new method for outlier definition and outlier detection, which combines the opinions from boundary-based and distance-based methods for outlier detection. Given an information system, for any subset X of the domain, to detect outliers in X, we first divided X into three different parts, by virtue of the notion of boundary region in rough sets. Then we utilized the distance-based method to find outliers in X. However, differing from the traditional distance-based method, we adopted different attitudes to objects from different parts of X when detecting outliers in X. Moreover, similar to Breunig's method, we defined a hybrid outlier factor (HOF), to indicate the degree of outlierness for every object in X. Experimental results on real data sets demonstrated the effectiveness of our method for outlier detection. The performance of our method are better than those of boundary-based and distance-based methods. This indicates that if we combine the boundary-based and distance-based methods, then we can obtain a more effective method for outlier detection.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant Nos. 60802042 and 61035004), the Natural Science Foundation of Shandong Province, China (Grant Nos. ZR2010FQ027 and ZR2009GQ013), and the National High-Tech Research & Development Program of China (Grant No. 2007AA01Z325).

References

Aggarwal, C.C., Yu, P.S., 2001. Outlier detection for high dimensional data. In: Proceedings of the 2001 ACM SIGMOD International Conference on Managment of Data, California, USA, pp. 37–46.

Angiulli, F., Pizzuti, C., 2002. Fast outlier detection in high dimensional spaces. In: Proceedings of the Sixth European Conference on the Principles of Data Mining and Knowledge Discovery, pp. 15–26.

Barnett, V., Lewis, T., 1994. Outliers in Statistical Data. John Wiley & Sons.

Bay, S.D., 1999. The UCI KDD repository. Available online at: <http://kdd.ics.uci.edu>.

Bolton, R.J., Hand, D.J., 2002. Statistical fraud detection: A review (with discussion). Statist. Sci. 17 (3), 235–255.

Breunig, M.M., Kriegel, H.-P., 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, Dallas, USA, pp. 93–104.

Chen, Y.M., Miao, D.Q., Wang, R.Z., 2008. Outlier detection based on granular computing. In: Proceedings of the 6th International Conference on Rough Sets and Current Trends in Computing, Akron, USA, pp. 283–292.

Chen, Y.M., Miao, D.Q., Zhang, H.Y., 2010. Neighborhood outlier detection. Expert Syst. Appl. 37 (12), 8745–8749.

Eskin, E., Arnold, A., Prerau, M., Portnoy, L., Stolfo, S., 2002. A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. In: Barbar, D. et al. (Eds.), Data Mining for Security Applications. Kluwer Academic Publishers, Boston.

Harkins, S., He, H.X., Willams, G.J., Baxter, R.A., 2002. Outlier detection using replicator neural networks. In: Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery, France, pp. 170–180.

Hawkins, D., 1980. Identifications of Outliers. Chapman and Hall, London.

He, Z.Y., Deng, S.C., Xu, X.F., 2005. An optimization model for outlier detection in categorical data. In: International Conference on Intelligent Computing (ICIC (1) 2005), Hefei, China, pp. 400–409.

Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: A review. ACM Computing Surveys 31 (3), 264–323.

Jiang, F., Sui, Y.F., Cao, C.G., 2005. Outlier detection using rough set theory. In: Proceedings of the 10th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing (RSFDGrC (2) 2005). LNAI 3642, Regina, Canada, pp. 79–87.

Jiang, F., Sui, Y.F., Cao, C.G., 2008. A rough set approach to outlier detection. Internat. J. General Syst. 37 (5), 519–536.

Jiang, F., Sui, Y.F., Cao, C.G., 2009. Some issues about outlier detection in rough set theory. Expert Syst. Appl. 36 (3), 4680–4687.

Johnson, T., Kwok, I., Ng, R.T., 1998. Fast computation of 2-dimensional depth contours. In: Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, New York, pp. 224–228.

Knorr, E., Ng, R., 1998. Algorithms for mining distance-based outliers in large datasets. In: Proceedings of the 24th VLDB Conference, New York, pp. 392–403.

Knorr, E., Ng, R., Tucakov, V., 2000. Distance-based outliers: Algorithms and applications. VLDB Journal: Very Large Databases 8 (3–4), 237–253.

Kovács, L., Vass, D., Vidács, 2004. A: Improving quality of service parameter prediction with preliminary outlier detection and elimination. In: Proceedings of the 2nd International Workshop on Inter-Domain Performance and Simulation (IPS 2004), Budapest, pp. 194–199.

Lane, T., Brodley, C.E., 1999. Temporal sequence learning and data reduction for anomaly detection. ACM Trans. Inform. Syst. Security 2 (3), 295–331.

Miao, D.Q., Zhao, Y., Yao, Y.Y., Li, H.X., Xu, F.F., 2009. Relative reducts in consistent and inconsistent decision tables of the Pawlak rough set model. Inform. Sci. 179 (24), 4140–4150.

Nguyen, S.H., Nguyen,. H.S., 1996. Some efficient algorithms for rough set methods. In: IPMU'96, Granada, Spain, pp. 1451–1456.

Pawlak, Z., 1982. Rough sets. Internat. J. Comput. Inform. Sci. 11, 341–356.

Pawlak, Z., 1991. Rough Sets: Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht.

Pawlak, Z., Grzymala-Busse, J.W., Slowinski, R., Ziarko, W., 1995. Rough sets. Commun. ACM 38 (11), 89–95.

Qian, Y.H., Liang, J.Y., Dang, C.Y., 2008. Converse approximation and rule extracting from decision tables in rough set theory. Comput. Math. Appl. 55 (8), 1754–1765.

Ramaswamy, S., Rastogi, R., Shim, K., 2000. Efficient algorithms for mining outliers from large datasets. In: Proceedings of the ACM SIGMOD Conference on Management of Data, Dallas, pp. 427–438.

Skowron, A., Rauszer, C., 1992. The discernibility matrices and functions in information systems. Handbook of Applications and Advances of Rough Set Theory, 11. Kluwer Academic Publishers, Dordrecht, pp. 331–362.

Wang, C.Z., Wu, C.X., Chen, D.G., 2008. A systematic study on attribute reduction with rough sets based on general binary relations. Inform. Sci. 178 (9), 2237–2261.

Wang, C.Z., Chen, D.G., Zhu, L.K., 2009. Homomorphisms between fuzzy information systems. Appl. Math. Lett. 22 (7), 1045–1050.

Wang, C.Z., Chen, D.G., Wu, C., Hu, Q.H., 2011. Data compression with homomorphism in covering information systems. Internat. J. Approx. Reason. 52 (4), 519–525.

Yin, Y.F., Gong, G.H., Han, L., 2009. Control approach to rough set reduction. Comput. Math. Appl. 57 (1), 117–126.

Zhong, N., Yao, Y.Y., Ohshima, M., Ohsuga, S., 2001. Interestingness, peculiarity, and multi-database mining. In: Proceedigs 2001 IEEE International Conference on Data Mining (IEEE ICDM'01), IEEE Computer Society Press, pp. 566–573.

Zhong, N., Yao, Y.Y., Ohshima, M., 2003. Peculiarity oriented multi-database mining. IEEE Trans. Knowledge Data Eng. 15 (4), 952–960.