Expert Systems with Applications 36 (2009) 4680–4687

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Some issues about outlier detection in rough set theory

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

[a] College of Information and Science 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

Keywords:
Outlier detection
Rough sets
Distance metric
KDD

ABSTRACT

"One person's noise is another person's signal" (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)). In recent years, much attention has been given to the problem of outlier detection, whose aim is to detect outliers – objects which behave in an unexpected way or have abnormal properties. Detecting such outliers is important for many applications such as criminal activities in electronic commerce, computer intrusion attacks, terrorist threats, agricultural pest infestations, etc. And outlier detection is critically important in the information-based society. In this paper, we discuss some issues about outlier detection in rough set theory which emerged about 20 years ago, and is nowadays a rapidly developing branch of artificial intelligence and soft computing. First, we propose a novel definition of outliers in information systems of rough set theory – *sequence-based outliers*. An algorithm to find such outliers in rough set theory is also given. The effectiveness of sequence-based method for outlier detection is demonstrated on two publicly available databases. Second, we introduce traditional distance-based outlier detection to rough set theory and discuss the definitions of distance metrics for distance-based outlier detection in rough set theory.

1. Introduction

Knowledge discovery in databases (KDD), or data mining, is an important issue in the development of data- and knowledge-based systems. Usually, knowledge discovery tasks can be classified into four general categories: (a) dependency detection, (b) class identification, (c) class description, and (d) outlier/exception detection (Knorr & Ng, 1998). In contrast to most KDD tasks, such as clustering and classification, outlier detection aims to find small groups of data objects that are exceptional when compared with the remaining large amount of data, in terms of certain sets of properties. For many applications, such as fraud detection in E-commerce, 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, researchers have begun focusing on outlier detection, and attempted to apply algorithms for finding outliers to tasks such as fraud detection (Bolton & Hand, 2002), identification of computer network intrusions (Eskin, Arnold, Prerau, Portnoy, & Stolfo, 2002; Lane & Brodley, 1999), data cleaning (Rulequest Research), detection of employers with poor injury histories (Knorr, Ng, & Tucakov, 2000), and peculiarity-oriented mining (Zhong, Yao, Ohshima, & Ohsuga, 2001).

Outliers exist extensively in the real world, and are generated from different sources: a heavily tailed distribution or errors in inputting the data. 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 & Ng, 1998). With increasing awareness on outlier detection in the literatures, more concrete meanings of outliers are defined for solving problems in specific domains. Nonetheless, most of these definitions follow the spirit of Hawkins' definition (Chiu & Fu, 2003).

Roughly speaking, the current approaches to outlier detection can be classified into the following five categories (Kovács, Vass, & Vidács, 2004):

(1) *Distribution-based approach* is the classical method in statistics. It is based on some standard distribution models (Normal, Poisson, etc.), and those objects which deviate from the model are recognized as outliers (Rousseeuw & Leroy, 1987). Its greatest disadvantage is that the distribution of the measurement data is unknown in practice. Often a large number of tests are required in order to decide which distribution model the measurement data follow (if there is any).

* Corresponding author.
E-mail addresses: jiangkong@163.net (F. Jiang), yfsui@ict.ac.cn (Y. Sui), cgcao@ict.ac.cn (C. Cao).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.06.019

(2) *Depth-based approach* is based on computational geometry and computes different layers of k–d convex hulls and flags objects in the outer layer as outliers (Johnson, Kwok, & Ng, 1998). However, it is a well-known fact that the algorithms employed suffer from the dimensionality curse and cannot cope with a large k.
(3) *Clustering approach* classifies the input data. It detects outliers as by-products (Jain, Murty, & Flynn, 1999). However, since the main objective is clustering, it is not optimized for outlier detection.
(4) *Distance-based approach* was originally proposed by Knorr and Ng (Knorr & Ng, 1998; Knorr et al., 2000). An object o in a data set T is a distance-based outlier if at least a fraction p of the objects in T are further than distance D from o.This outlier definition is based on a single, global criterion determined by the parameters p and D. Problems may occur if the parameters of the data are very different from each other in different regions of the data set.
(5) *Density-based approach* was originally proposed by Breunig, Kriegel, Ng, and Sander (2000). A Local Outlier Factor (LOF) is assigned to each sample based on its 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.

Rough set theory, introduced by Zdzislaw Pawlak in the early 1980s (Pawlak, 1982, 1991, Pawlak, Grzymala-Busse, Slowinski, & Ziarko, 1995), is 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 object 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. Successful applications of the rough set model in a variety of problems have demonstrated its importance and versatility.

To the best of our knowledge, there is no existing work about outlier detection in rough set community. The aim of this work is to combine the rough set theory and outlier detection to show how outlier detection can be done in rough set theory. We suggest two different ways to achieve this aim. First, we propose sequence-based outlier detection in information systems of rough set theory. Second, we introduce traditional distance-based outlier detection to rough set theory.

This paper is organized as follows. In the next section, we introduce some preliminaries in rough set theory and outlier detection. In Section 3, we give some definitions concerning sequence-based outliers in information systems of rough set theory. The basic idea is as follows: Given an information system $IS = (U, A, V, f)$, where U is a non-empty finite set of objects, A is a set of attributes, V is the union of attribute domains, and f is a function such that for any $x \in U$ and $a \in A$, $f(x, a) \in V_a$. Since each attribute subset $B \subseteq A$ determines an indiscernibility (equivalence) relation $IND(B)$ on U, we can obtain the corresponding equivalence class of relation $IND(B)$ for every object $x \in U$. If we decrease attribute subset B gradually, then the granularity of partition $U/IND(B)$ will become coarser, and for every object $x \in U$ the corresponding equivalence class of x will become bigger. So when there is an object in U whose equivalence class always does not vary or only increases a little in comparison with those of other objects in U, then we may consider this object as a sequence-based outlier in U with respect to IS. An algorithm to find sequence-based outliers is also given. In Section 4, we apply traditional distance-based outlier detection to rough set theory. Since classical rough set theory is better suited to deal with *nominal* attributes, we propose the revised definitions of two traditional distance metrics for distance-based outlier detection in rough set theory – *overlap metric* and *value difference metric in rough set theory*, both of which are especially designed to deal with *nominal* attributes. Experimental results are given in Section 5, and Section 6 discusses the advantages of our sequence-based approach by comparing with other approaches to outlier detection. Section 7 concludes the paper.

2. Preliminaries

In rough set data model, information is stored in a table, where each row (tuple) represents facts about an object. All we know about an object from the table is the corresponding tuple in the table.

In rough set terminology, a data table is also called an information system. 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. $f : U \times A \to V$ is an information function such that for any $a \in A$ and $x \in U$, $f(x, a) \in V_a$.

We can split set A of attributes into two subsets: $C \subset A$ and $D = A - C$, conditional set of attributes and decision (or class) attribute(s), respectively. The condition attributes represent measured features of the objects, while the decision attributes are *a posteriori* outcome of classification.

Each subset $B \subseteq A$ of attributes determines a binary relation $IND(B)$, called indiscernibility relation, which is defined as follows:

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

It is obvious that $IND(B)$ is an equivalence relation on U and $IND(B) = \bigcap_{a \in B} IND(\{a\})$.

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. 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 $IND(B)$.

The distance-based approach is now widely used for outlier detection. An object o in a data set S is a distance-based (DB) outlier with parameters p and d, denoted by $DB(p, d)$, if at least a fraction p of the objects in S lie at a distance greater than d from o. The advantage of the distance-based approach is that no explicit distribution is needed to determine unusualness, and it can be applied to any feature or attribute space (the vector space spanned by some features is called a feature space) for which we can define a distance metric. It should be noted that metrics are measures possessing metric properties which express the degree or strength of a quality factor. And many measures that we often use are in fact not metrics. A distance metric is a distance function on a set of points, mapping pairs of points into the nonnegative real numbers. In general, any distance metric which obeys the following conditions can be used in similarity measures (Li, Chen, Li, Ma, & Vitnyi, 2003):

(1) $D(x, y) \geqslant 0$: Distances cannot be negative.
(2) $D(x, y) = 0$ if and only if $x = y$.
(3) $D(x, y) = D(y, x)$: Distance is symmetric.
(4) $D(x, y) + D(y, z) \geqslant D(x, z)$: Triangular inequality.

Usually, an attribute can be linear or nominal, and a linear attribute can be continuous or discrete. A *continuous* (or *continuously valued*) attribute uses real values, such as the mass of a planet or the velocity of an object. A *linear discrete* (or *integer*) attribute can have only a discrete set of linear values, such as *number of children*. A *nominal* (or *symbolic*) attribute is a discrete attribute whose values are not necessarily linearly ordered. For example, a variable representing color might have values such as *red, green, blue, brown, black* and *white*, which could be represented by 1–6, respectively (Wilson & Martinez, 1997).

By now, there are many distance metrics that have been proposed, and most of these metrics are only defined for *linear* attributes, including the Euclidean, Minkowsky, Mahalanobis distance metrics, the context-similarity measure, hyperrectangle distance functions and others. When the attributes are nominal, these definitions of similarity measures become less trivial. For *nominal* attributes, a simple but commonly used measure is the overlap metric (Stanfill & Waltz, 1986). Under this metric, for two possible values v_i and v_j, the distance is defined as zero when v_i and v_j are identical, and one otherwise. For binary attributes, this overlap metric reduces to the so-called Hamming distance. Moreover, a very popular real-valued metric for nominal attributes is the value difference metric (Cheng, Li, Kwok, & Li, 2004; Stanfill & Waltz, 1986).

3. Sequence-based outlier detection in rough set theory

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.

At present, in rough set community, the research concerning KDD mainly focuses on the first three categories of tasks in knowledge discovery. However, there is no concern on the fourth category of tasks – outlier/exception detection in rough set theory. Therefore in this section, we discuss issues about outlier definition and detection in information systems of rough set theory. In the following subsection, we first give some definitions concerning sequence-based outliers. Next an algorithm to find sequence-based outliers is presented.

3.1. Definitions

Our definition for sequence-based outliers in an information system follows the spirit of Hawkins' definition for outliers. That is, given an information system $IS = (U, A, V, f)$, for any $x \in U$, if x has some characteristics that differ greatly from those of other objects in U, in terms of the attributes in A, we may call x an outlier in U with respect to IS.

Our basic idea is as follows. Given an information system $IS = (U, A, V, f)$, since each attribute subset $B \subseteq A$ determines an indiscernibility (equivalence) relation $IND(B)$ on U, we can obtain the corresponding equivalence class for every object $x \in U$ under relation $IND(B)$, that is, the equivalence class that contains x. If we decrease B gradually, then the granularity of partition $U/IND(B)$ will become coarser, and for every object $x \in U$ the corresponding equivalence class of x will become bigger. So when there is an object in U whose equivalence class always does not vary or only increases a little in comparison with those of other objects in U, then we consider this object as an outlier in U with respect to IS. In a word, an outlier in U is an element whose corresponding equivalence class (the set of all indiscernibility elements) always does not vary or varies a little when we regularly coarsen the granularity of partition $U/IND(B)$ about U.

Through reasoning about the changes of attribute subsets and the corresponding equivalence classes in an information system, we can find all the outliers. In fact, most of our idea above is derived from the excellent work by *A*. Skowron et al., who discussed the approximate reasoning methods based on information changes in a series of papers (Bazan et al., 2002; Skowron & Synak, 2004). They introduced basic concepts for approximate reasoning about information changes across information maps, and measured degree of changes using information granules. Any rule for reasoning about information changes specifies how changes of information granules from the rule premise influence changes of information granules from the rule conclusion. Changes in information granules can be measured, e.g., using expressions analogous to derivatives.

To implement the above idea, we construct three kinds of sequence (Arning, Agrawal, & Raghavan, 1996). The first is the sequence of attributes.

Definition 3.1 (*Sequence of attributes*). Let $IS = (U, A, V, f)$ be an information system, where $A = \{a_1, a_2, \ldots, a_m\}$. For every attribute $a_i \in A$, let $U/IND(\{a_i\}) = \{E_{i_1}, \ldots, E_{i_k}\}$ denote the partition induced by equivalence relation $IND(\{a_i\})$. And we use $Var(a_i)$ to denote the variance of set $\{|E_{i_1}|, \ldots, |E_{i_k}|\}$, that is,

$$Var(a_i) = \frac{(|E_{i_1}| - Ave)^2 + \ldots + (|E_{i_k}| - Ave)^2}{k} \tag{2}$$

where $Ave = (|E_{i_1}| + \ldots + |E_{i_k}|)/k$, $|X|$ denotes the cardinality of set X, and $1 \leqslant i \leqslant m$.

We construct a *sequence* $S = \langle a'_1, a'_2, \ldots, a'_m \rangle$ *of attributes* in A (In this paper, a sequence is denoted by a tuple), where for every $1 \leqslant i < m$, $Var(a'_i) \geqslant Var(a'_{i+1})$.

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

Definition 3.2 (*Descending sequence of attribute subsets*). Let $IS = (U, A, V, f)$ be an information system, where $A = \{a_1, a_2, \ldots, a_m\}$. Let $S = \langle a'_1, a'_2, \ldots, a'_m \rangle$ be the sequence of attributes defined above. Given a sequence $AS = \langle A_1, A_2, \ldots, A_m \rangle$ of attribute subsets, where $A_1, A_2, \ldots, A_m \subseteq A$. If $A_1 = A$, $A_m = \{a'_m\}$ and $A_{i+1} = A_i - \{a'_i\}$ for every $1 \leqslant i < m$, then we call AS a *descending sequence of attribute subsets* in IS.

By the above definition, in $AS = \langle A_1, A_2, \ldots, A_m \rangle$, for every $1 \leqslant i < m$, A_{i+1} is the attribute subset transformed from A_i by removing the element a'_i from A_i, where a'_i is the ith element in sequence S.

Given any object $x \in U$, for the descending sequence AS of attribute subsets, we can construct a corresponding ascending sequence of equivalence classes of x.

Definition 3.3 (*Ascending sequence of equivalence classes*). Let $IS = (U, A, V, f)$ be an information system, where $A = \{a_1, a_2, \ldots, a_m\}$. Let $AS = \langle A_1, A_2, \ldots, A_m \rangle$ be a descending sequence of attribute subsets in IS. For any object $x \in U$, let $[x]_{A_i}$ be the equivalence class of x under relation $IND(A_i)$ for every $1 \leqslant i \leqslant m$, then we have a sequence $ES(x) = \langle [x]_{A_1}, [x]_{A_2}, \ldots, [x]_{A_m} \rangle$ of equivalence classes, where $[x]_{A_1}, [x]_{A_2}, \ldots, [x]_{A_m} \subseteq U$. $ES(x)$ is called an *ascending sequence of equivalence classes* of object x in IS.

Most current methods for outlier detection give a binary classification of objects: is or is not an outlier, e.g. the distance-based outlier detection. However, for many scenarios, it is more meaningful to assign to each object a degree of being an outlier. Given a degree of outlierness for every object, the objects can be ranked according to this degree, giving the data mining analyst a sequence in which to analyze the outliers. Therefore, Breunig et al. introduced a novel notion of local outlier in which the degree to which an object is outlying is dependent on the density of its local neighborhood, and each object can be assigned a *Local Outlier Factor (LOF)*, which represents the likelihood of that object being an outlier (Breunig et al., 2000).

Similar to Breunig's method, we define a *sequence outlier factor (SOF)*, which indicates the degree of outlierness for every object in an information system.

Definition 3.4 (*Sequence Outlier Factor*). Let $IS = (U, A, V, f)$ be an information system, where $A = \{a_1, a_2, \ldots, a_m\}$. For any $x \in U$, let $ES(x) = \langle [x]_{A_1}, [x]_{A_2}, \ldots, [x]_{A_m} \rangle$ be the ascending sequence of equivalence classes of x in IS. The *sequence outlier factor* of x in IS is defined as follows:

$$SOF(x) = \sum_{j=2}^{m} \left(1 - \frac{|[x]_{A_j}| - |[x]_{A_{j-1}}|}{|U|} \right) \times W(x) \tag{3}$$

where $W : U \rightarrow [0, 1)$ is a weight function such that for any $x \in U$, $W(x) = \sum_{a \in A} \left(1 - \frac{|[x]_{\{a\}}|}{|U|} \right) \Big/ |A|$. $[x]_{\{a\}} = \{u \in U : f(u, a) = f(x, a)\}$ denotes the indiscernibility class of relation $IND(\{a\})$ that contains element x and $|M|$ denotes the cardinality of set M.

The weight function W in the above definition expresses such an idea that outlier detection always concerns the minority of objects in the data set and the minority of objects are more likely to be outliers than the majority of objects. Since from the above definition, we can see that the more the weight, the more the sequence outlier factor, the minority of objects should have more weight than the majority of objects. Therefore for every $a \in A$, if the objects in U that are indiscernible with x under relation $IND(\{a\})$ are always few, that is, the percentage of objects in U that are indiscernible with x is small, then we consider x belonging to the minority of objects in U, and assign a high weight to x.

Furthermore, the above definition accords with our basic idea about sequence-based outliers. That is, when we regularly coarsen the granularity of the partitions on domain U of an information system IS, if the corresponding equivalence class of an object $x \in U$ always does not vary or varies a little, then the likelihood of x being an outlier, i.e. the sequence outlier factor of x, is high.

Definition 3.5 (*Sequence-based Outliers*). Let $IS = (U, A, V, f)$ be an information system, where $A = \{a_1, a_2, \ldots, a_m\}$. Let μ be a given threshold value, for any $x \in U$, if $SOF(x) > \mu$ then x is called a *sequence-based outlier* in U with respect to IS, where $SOF(x)$ is the sequence outlier factor of x in IS.

3.2. Algorithm

Algorithm 3.1. Input: information system $IS = (U, A, V, f)$, where $|U| = n$ and $|A| = m$; threshold value μ. Output: a set O of sequence-based outliers

```
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 $a$ (Nguyen, 1996);
  Determine the partition $U/IND(\{a\}) = \{E_{i_1}, \ldots, E_{i_k}\}$;
  Calculate $Var(a)$, the variance of set $\{|E_{i_1}|, \ldots, |E_{i_k}|\}$
}
Determine the sequence $S = \{a'_1, a'_2, \ldots, a'_m\}$ of attributes in $A$, where
for every $1 \leqslant i < m$, $Var(a'_i) \geqslant Var(a'_{i+1})$;
Construct a descending sequence $AS = \{A_1, \ldots, A_m\}$ of attribute
subsets, in terms of sequence $S$;
For $1 \leqslant i \leqslant m$
{
  Sort all objects from $U$ according to a given order (e.g. the)
  (lexicographical order) on domain $V_{A_i}$ of attribute subset $A_i$;
  Determine the partition $U/IND(A_i)$
}
For every $x \in U$
{
  Construct an ascending sequence $ES(x)$ of equivalence classes,
  where $ES(x) = \langle [x]_{A_1}, [x]_{A_2}, \ldots, [x]_{A_m} \rangle$, and $[x]_{A_i} \in U/IND(A_i)$,
  $1 \leqslant i \leqslant m$;
  Assign a weight $W(x)$ to $x$;
  Calculate $SOF(x)$, the sequence outlier factor of $x$ in $IS$;
  If $SOF(x) > \mu$ then $O = O \cup \{x\}$
}
Return $O$.
```

In Algorithm 3.1, we use a method proposed by Nguyen (1996) to calculate the partition induced by an indiscernibility relation in an information system.

In the worst case, the time complexity of Algorithm 3.1 is $O(m^2 \times nlogn)$, and its space complexity is $O(m \times n)$, where m and n are the cardinalities of A and U, respectively.

4. Distance-based outlier detection in rough set theory

Knorr and Ng (1998) and Knorr et al. (2000) proposed a non-parametric approach to outlier detection based on the distance of an object in a given data set to its nearest neighbors. In this approach, one looks at the local neighborhood of points for an object typically defined by the k-nearest objects (also known as neighbors). If the neighboring points are relatively close, then the object is considered to be normal; if the neighboring points are far away, then the object is considered to be unusual (Bay & Schwabacher, 2003).

By virtue of the ideas of distance-based outlier detection, objects in an information system are considered to be outliers if they do not have enough neighbors. Although rough sets can also be used for non-nominal attributes, they are better suited to nominal data sets. In order to calculate the distance between any two objects in an information system, we should find a suitable distance metric for nominal attributes in rough set theory. Next, we shall give the revised definitions of the traditional overlap metric and value difference metric – *overlap metric* and *value difference metric in rough set theory*, respectively.

4.1. Overlap metric in rough set theory

Definition 4.1. Given an information system $IS = (U, A, V, f)$, let $x, y \in U$ be any two objects between which we shall calculate the distance. The *overlap metric in rough set theory* is defined as

$$\Delta(x, y) = |\{a \in A : a(x) \neq a(y)\}| \tag{4}$$

where $\Delta : U \times U \rightarrow [0, \infty]$ is a function from $U \times U$ to the non-negative real number (Stanfill & Waltz, 1986).

It should be noticed that the above definition is closely related to the definition of discernibility matrices (Skowron & Rauszer, 1992). In fact, if we have calculated the discernibility matrix of a given information system, we can easily observe the distance between any two objects in the information system from the discernibility matrix.

From the above definition, we can see that the *overlap metric in rough set theory* mainly counts the number of attributes in A on which objects x and y have different attribute values. This is a reasonable choice if no information about the weight of the different attributes is available. But if we do have information about attribute weight, then we can add linguistical bias to weight the different attributes. Therefore, a number of variants for overlap metric, each using a different weighting factor, have also been proposed (Gower & Legendre, 1986). However, this overlap metric and its variants assume that all attribute values are of equal distance from each other, and thus cannot represent value pairs with differing degrees of similarities. As an example, for attribute *taste*, it may be more desirable to have the value *sour* closer to *sweet* than to *bitter*. Hence, a real-valued distance metric is often preferred over a Boolean one (Cheng et al., 2004).

4.2. Value difference metric in rough set theory

The value difference metric (VDM) was introduced by Stanfill and Waltz (1986) to provide an appropriate distance function for nominal attributes. A simplified version (without the weighting schemes) of the VDM is defined as follows:

$$VDM(x,y) = \sum_{f \in F} d_f(x_f, y_f) \tag{5}$$

where F is the set of all features in the problem domain, x and y are any two objects between which we shall calculate the distance. And $d_f(x_f, y_f)$ denotes the distance between two values x_f and y_f of feature f, where x_f is the value of object x on feature f.

For any feature $f \in F$, $d_f(x_f, y_f)$ is defined as follows:

$$d_f(x_f, y_f) = \sum_{c \in OC} (P(c|x_f) - P(c|y_f))^2 \tag{6}$$

where OC is the set of all output classes in the problem domain, $P(c|v)$ is the conditional probability that the output class is c given that feature f has the value v, and $P(c|v) = \frac{\text{number of objects with value } v \text{ and belonging to class } c}{\text{number of objects with value } v}$.

Next we give the revised definition of VDM in rough set theory. In rough set domain, given a decision table $DT = (U, C \cup D, V, f)$, we use set C of condition attributes to replace the set F of features in the above definition. And the set $U/IND(D)$ of decision categories is used to replace the set OC of output classes in the above definition.

Definition 4.2. Given a decision table $DT = (U, C \cup D, V, f)$, where C is the set of condition attributes and D is the set of decision attributes. Let $x, y \in U$ be any two objects between which we shall calculate the distance. The *value difference metric in rough set theory* $VDM_R : U \times U \to [0, \infty]$ is defined as

$$VDM_R(x,y) = \sum_{a \in C} d_a(x_a, y_a) \tag{7}$$

where $d_a(x_a, y_a)$ denotes the distance between two values x_a and y_a of condition attribute a, and x_a is the value of object x on attribute a.

For any $a \in C$, define

$$d_a(x_a, y_a) = \sum_{E \in U/IND(D)} \left(\frac{|[x]_{\{a\}} \cap E|}{|[x]_{\{a\}}|} - \frac{|[y]_{\{a\}} \cap E|}{|[y]_{\{a\}}|} \right)^2 \tag{8}$$

where $[x]_{\{a\}}$ is an equivalence class of $IND(\{a\})$ that contains object x.

For any $E \in U/IND(D)$, all elements in E have the same decision attribute values, that is, E denotes a decision category. $|[x]_{\{a\}}|$ is the number of all objects in U that have the value x_a on attribute a. And $[x]_{\{a\}} \cap E$ is the set of all objects in U that have the value x_a on attribute a and belong to decision category E. Therefore $\frac{|[x]_{\{a\}} \cap E|}{|[x]_{\{a\}}|}$ is the conditional probability that the decision category is E given that attribute a has the value x_a. This corresponds to $P(c|v)$ in the above definition.

Table 1
Decision table *DT*

$U \setminus A$	a	b	c	d	e
u_1	1	0	2	1	0
u_2	0	0	1	2	1
u_3	2	0	2	2	0
u_4	0	0	2	1	2
u_5	1	1	2	1	0

Example 4.1. Given a decision table $DT = (U, C \cup D, V, f)$, where $U = \{u_1, u_2, u_3, u_4, u_5\}$, $C = \{a, b, c, d\}$, $D = \{e\}$, as shown in Table 1.

Using the distance metric defined by Definition 4.2, we can calculate the distance for every pair of objects in U. Because of the limitation of space, we just present the procedure for calculating the distance between u_1 and u_2.

Initialization:
$U/IND(D) = U/IND(\{e\}) = \{\{u_1, u_3, u_5\}, \{u_2\}, \{u_4\}\}$,
Let $E_1 = \{u_1, u_3, u_5\}$, $E_2 = \{u_2\}$, $E_3 = \{u_4\}$.
Then $VDM_R(u_1, u_2) = d_a(1,0) + d_b(0,0) + d_c(2,1) + d_d(1,2)$.
Step 1: Evaluate $d_a(1,0)$**:**

$$[u_1]_{\{a\}} = \{u_1, u_5\}, \quad [u_2]_{\{a\}} = \{u_2, u_4\};$$

$$\begin{aligned} d_a(1,0) &= \sum_{E \in U/IND(D)} \left(\frac{|[u_1]_{\{a\}} \cap E|}{|[u_1]_{\{a\}}|} - \frac{|[u_2]_{\{a\}} \cap E|}{|[u_2]_{\{a\}}|} \right)^2 \\ &= \left(\frac{|\{u_1, u_5\}|}{|\{u_1, u_5\}|} - \frac{|\emptyset|}{|\{u_2, u_4\}|} \right)^2 + \left(\frac{|\emptyset|}{|\{u_1, u_5\}|} - \frac{|\{u_2\}|}{|\{u_2, u_4\}|} \right)^2 \\ &\quad + \left(\frac{|\emptyset|}{|\{u_1, u_5\}|} - \frac{|\{u_4\}|}{|\{u_2, u_4\}|} \right)^2 \\ &= 1 + 1/4 + 1/4 = 3/2. \end{aligned}$$

Step 2: Evaluate $d_b(0,0)$:

$$[u_1]_{\{b\}} = \{u_1, u_2, u_3, u_4\}, \quad [u_2]_{\{b\}} = \{u_1, u_2, u_3, u_4\}.$$

Since $[u_1]_{\{b\}} = [u_2]_{\{b\}}$, object u_1 and u_2 belong to the same equivalence class of equivalence relation $IND(\{b\})$, and the distributions of decision categories with respect to the two objects are always the same. Therefore, $d_b(0,0)$ certainly equals 0.

Step 3: Evaluate $d_c(2,1)$:

$$[u_1]_{\{c\}} = \{u_1, u_3, u_4, u_5\}, \quad [u_2]_{\{c\}} = \{u_2\};$$

$$\begin{aligned} d_c(2,1) &= \sum_{E \in U/IND(D)} \left(\frac{|[u_1]_{\{c\}} \cap E|}{|[u_1]_{\{c\}}|} - \frac{|[u_2]_{\{c\}} \cap E|}{|[u_2]_{\{c\}}|} \right)^2 \\ &= \left(\frac{|\{u_1, u_3, u_5\}|}{|\{u_1, u_3, u_4, u_5\}|} - \frac{|\emptyset|}{|\{u_2\}|} \right)^2 + \left(\frac{|\emptyset|}{|\{u_1, u_3, u_4, u_5\}|} - \frac{|\{u_2\}|}{|\{u_2\}|} \right)^2 \\ &\quad + \left(\frac{|\{u_4\}|}{|\{u_1, u_3, u_4, u_5\}|} - \frac{|\emptyset|}{|\{u_2\}|} \right)^2 \\ &= 9/16 + 1 + 1/16 = 13/8. \end{aligned}$$

Step 4: Evaluate $d_d(1,2)$:

$$[u_1]_{\{d\}} = \{u_1, u_4, u_5\}, \quad [u_2]_{\{d\}} = \{u_2, u_3\};$$

$$\begin{aligned} d_d(1,2) &= \sum_{E \in U/IND(D)} \left(\frac{|[u_1]_{\{d\}} \cap E|}{|[u_1]_{\{d\}}|} - \frac{|[u_2]_{\{d\}} \cap E|}{|[u_2]_{\{d\}}|} \right)^2 \\ &= \left(\frac{|\{u_1, u_5\}|}{|\{u_1, u_4, u_5\}|} - \frac{|\{u_3\}|}{|\{u_2, u_3\}|} \right)^2 + \left(\frac{|\emptyset|}{|\{u_1, u_4, u_5\}|} - \frac{|\{u_2\}|}{|\{u_2, u_3\}|} \right)^2 \\ &\quad + \left(\frac{|\{u_4\}|}{|\{u_1, u_4, u_5\}|} - \frac{|\emptyset|}{|\{u_2, u_3\}|} \right)^2 \\ &= 1/36 + 1/4 + 1/9 = 7/18. \end{aligned}$$

Step 5: $VDM_R(u_1, u_2) = 3/2 + 0 + 13/8 + 7/18 \approx 3.51$.

Repeating the above calculation, we can finally obtain distances for all the other pairs of objects in U. By then, we can find out all outliers in U by using the nested loops algorithm or other algorithms for distance-based outlier detection (Knorr & Ng, 1998; Knorr et al., 2000; Ramaswamy, Rastogi, & Shim, 2000). Since many fast algorithms for finding distance-based outliers have been proposed, such as Bay and Schwabacher (2003) proposed a simple algorithm based on nested loops, which can give near linear time performance when the data is in random order and a simple pruning rule is used. And they demonstrated that their algorithm scales to real data sets with millions of examples and many features. Therefore combining the distance metrics we defined above with these fast algorithms, we can efficiently find distance-based outliers in rough set theory.

5. Experimental results

In Section 3, we proposed a sequence-based outlier detection algorithm. And in Section 4, we introduced traditional distance-based outlier detection to rough sets. In this section, following the experimental setup in He, Deng, and Xu (2005), we shall use two real life data sets (*lymphography* and *cancer*) to demonstrate the performance of sequence-based outlier detection algorithm against traditional distance-based method and KNN algorithm (Ramaswamy et al., 2000). In addition, on the *cancer* data set, we add the results of RNN-based outlier detection method for comparison, these results can be found in the work of Harkins et al. Harkins, He, Willams, and Baxter (2002) Willams, Baxter, He, Harkins, and Gu (2002).

In our experiment, for the KNN algorithm, the results were obtained by using the 5th nearest neighbor. Since in traditional distance-based outlier detection, being an outlier is regarded as a binary property, we revise the definition of distance-based outlier detection by introducing a *distance outlier factor* (DOF) to indicate the degree of outlierness for every object in an information system.

Definition 4.1 (*Distance Outlier Factor*). Given an information system $IS = (U, A, V, f)$. For any object $x \in U$, the percentage of the objects in U which lies greater than d from x is called the *distance outlier factor of* x in IS, denoted by

$$DOF(x) = \frac{|\{y \in U : dist(x, y) > d\}|}{|U|} \tag{9}$$

where $dist(x, y)$ denotes the distance between object x and y under a given distance metric in rough set theory (In our experiment, the overlap metric in rough set theory is adopted), d is a given parameter (In our experiment, we set $d = |A|/2$), and $|U|$ denotes the cardinality of set U.

5.1. 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" (1.35%), "metastases" (54.73%), "malign lymph" (41.22%) and "fibrosis" (2.7%). Classes 1 and 4 ("normal find" and "fibrosis") are regarded as rare classes.

Aggarwal et al., proposed a practicable way to test the effectiveness of an outlier detection method (Aggarwal & Yu, 2001; He et al., 2005). That is, we can run the outlier detection method on a given data set and test the percentage of points which belonging 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 (Aggarwal & Yu, 2001). Points belonged to the rare class are considered as outliers. If the method works well, we expect that such abnormal classes would be over-represented in the set of points found.

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) in IS_L. The experimental results are summarized in Table 2.

In Table 2, "SEQ", "DIS", "KNN" denote sequence-based, traditional distance-based and KNN outlier detection methods, respectively. For every object in U, the degree of outlierness is calculated by using the three outlier detection methods, respectively. For each outlier detection method, the "Top Ratio (Number of Objects)" denotes the percentage (number) of the objects selected from U whose degrees of outlierness calculated by the method are higher than those of other objects in U. And if we use $X \subseteq U$ to contain all those objects selected from U, then the "Number of Rare Classes Included" is the number of objects in X that belong to one of the rare classes. The "Coverage" is the ratio of the "Number of Rare Classes Included" to the number of objects in U that belong to one of the rare classes (He et al., 2005).

From Table 2, we can see that for the lymphography data set, sequence-based and distance-based methods perform markedly better than KNN method. And the performances of sequence-based and distance-based methods are very close, though the former performs a little worse than the latter.

5.2. Wisconsin breast cancer data

The Wisconsin breast cancer data set is found in the UCI machine learning repository (Bay, 1999). The data set contains 699 instances with 9 continuous attributes. 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; Willams et al., 2002). 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] (He et al., 2005).

Data in the Wisconsin breast 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) in IS_W. The experimental results are summarized in Table 3.

Table 3 is similar to Table 2. From Table 3, we can see that for the Wisconsin breast cancer data set, their performances of sequence-based, traditional distance-based and RNN-based methods are very close, the performances of are all markedly better than KNN algorithm. And among the former three methods, RNN-based method performs best, the next one is the sequence-based method, and the worst is the distance-based method.

6. Discussion

There are two broad approaches to outlier detection: parametric (e.g., distribution-based) and non-parametric (e.g., distance-based). The parametric or statistical approach to outlier detection assumes a distribution or probability model for data set and then identifies outliers with respect to the model using a discordancy

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

Table 2
Experimental results in IS_L

Top ratio (number of objects)	Number of rare classes included (coverage)		
	SEQ	DIS	KNN
5%(7)	5(83%)	5(83%)	4(67%)
6%(9)	5(83%)	6(100%)	4(67%)
8%(12)	6(100%)	6(100%)	5(83%)
10%(15)	6(100%)	6(100%)	6(100%)

Table 3
Experimental results in IS_W

Top ratio (number of objects)	Number of rare classes included (coverage)			
	SEQ	DIS	KNN	RNN
1%(4)	3(8%)	4(10%)	3(8%)	4(10%)
2%(8)	7(18%)	5(13%)	6(15%)	8(21%)
4%(16)	14(36%)	11(28%)	11(28%)	16(41%)
6%(24)	21(54%)	18(46%)	18(46%)	20(51%)
8%(32)	28(72%)	24(62%)	25(64%)	27(69%)
10%(40)	32(82%)	29(74%)	30(77%)	32(82%)
12%(48)	35(90%)	36(92%)	35(90%)	37(95%)
14%(56)	39(100%)	39(100%)	36(92%)	39(100%)
16%(64)	39(100%)	39(100%)	36(92%)	39(100%)
18%(72)	39(100%)	39(100%)	38(97%)	39(100%)
20%(80)	39(100%)	39(100%)	38(97%)	39(100%)
28%(112)	39(100%)	39(100%)	39(100%)	39(100%)

test. A major drawback of the parametric approach is that most tests are for single attributes, yet many problems on data mining require finding outliers in multidimensional spaces. Moreover, the parametric approach requires knowledge about parameters of the data set, such as the data distribution. However, in many cases, the data distribution may not be known. The representative non-parametric approach is distance-based outlier detection, which was introduced to counter the main limitations imposed by the parametric approach. In comparison with statistical-based approach, distance-based outlier detection avoids the excessive computation associated with fitting an observed distribution into some standard distribution and in selecting discordancy tests (Han & Kamber, 2000).

Distance-based outlier detection relies on a distance metric (function) to measure the similarity between any two objects. As we have noted before, most of the distance metrics that have been proposed are only defined for *linear* attributes. However, almost all real-world data sets contain a mixture of *nominal* and *linear* attributes. The *nominal* attributes are typically ignored or incorrectly modeled by existing approaches. In addition, it is well-known that, in practice, there are no clear advantages of one particular metric over another. Therefore, it is often very difficult to determine an appropriate distance metric for many practical distance-based outlier detection tasks. Moreover, it is possible in some situations that there is no appropriate distance metric to be defined for a particular data set. Then distance-based outlier detection obviously cannot be applied to those situations. Unlike distance-based outlier detection, our sequence-based outlier detection does not require a metric distance function. On the other hand, similar to distance-based outlier detection, sequence-based outlier detection is also a non-parametric approach to outlier detection, which can cope with the main limitations imposed by the parametric approach. So our approach can efficiently avoid the disadvantages of most current approaches to outlier detection.

7. Conclusion and future work

Outlier detection is becoming critically important in many areas. In this paper, we extended outlier detection to Pawlak's rough set theory, which has become a popular method for KDD, much due to its ability to handle uncertain and/or incomplete data. First, we proposed a new sequence-based outlier detection method in information systems of rough set theory. Experimental results on real data sets demonstrate the effectiveness of our method for outlier detection. Next, we introduced Knorr and Ng's distance-based outlier detection to rough set theory. In order to calculate the distance between any two objects in an information system, we gave two revised definitions of the traditional distance metrics for nominal attributes in rough set theory.

In the future work, for the sequence-based outlier detection algorithm, we shall consider using reducts to have smaller number of attributes while preserving the performance of it. For the distance-based outlier detection, since the proposed measures are all global ones, in the future work, we shall consider using more local versions to get better performance for distance-based method.

Acknowledgements

This work is supported by the Natural Science Foundation (Grant Nos. 60641010, 60496326, 60573063 and 60573064), the National 863 Programme (Grant No. 2007AA01Z325), and the National 973 Programme (Grant Nos. 2003CB317008 and G1999032701).

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* (pp. 37–46).

Arning, A., Agrawal, R., Raghavan, P. (1996). A linear method for deviation detection in large databases. In *Proceedings of the 2nd international conference on KDD, Oregon* (pp. 164–169).

Bay, S. D. (1999). The UCI KDD repository. <http://kdd.ics.uci.edu>.

Bay, S. D., Schwabacher, M. (2003). Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In *Proceedings of the 9th ACM SIGKDD international conference on KDD, Washington* (pp. 29–38).

Bazan, J., Osmolski, A., Skowron, A., Slezak, D., Szczuka, M., Wroblewski, J. (2002). Rough set approach to survival analysis. In *Proceedings of the 3rd international conference on rough sets and current trends in computing (RSCTC2002), Malvern, PA, USA* (pp. 522–529).

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review (with discussion). *Statistical Science, 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* (pp. 93–104).

Cheng, V., Li, C. H., Kwok, J., & Li, C. K. (2004). Dissimilarity learning for nominal data. *Pattern Recognition, 37*(7), 1471–1477.

Chiu, A. L., Fu, A. W. (2003). Enhancements on local outlier detection. In *Proceedings of the 7th international database engineering and applications symposium (IDEAS'03), Hong Kong* (pp. 298–307).

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

Gower, J. C., & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification, 3*, 5–48.

Han, J. W., & Kamber, M. (2000). *Data mining: Concepts and techniques*. California: Morgan Kaufmann Publishers.

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*. London: Chapman and Hall.

He, Z. Y., Deng, S. C., Xu, X. F. (2005). An optimization model for outlier detection in categorical data. In *Advances in intelligent computing, 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.

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, A. (2004). 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 Transactions on Information and System Security, 2*(3), 295–331.

Li, M., Chen, X., Li, X., Ma, B., Vitnyi, M. B. (2003). The similarity metric. In *Proceedings of the 14th annual ACM-SIAM symposium on discrete algorithms, Maryland* (pp. 863–872).

Nguyen, S. H., Nguyen, H. S. (1996). Some efficient algorithms for rough set methods. In *Proceedings of the 6th international conference on information processing and management of uncertainty (IPMU'96), Granada, Spain* (pp. 1451–1456).

Pawlak, Z. (1982). Rough sets. *International Journal of Computer and Information Sciences, 11*, 341–356.

Pawlak, Z. (1991). *Rough sets: Theoretical aspects of reasoning about data*. Dordrecht: Kluwer Academic Publishers.

Pawlak, Z., Grzymala-Busse, J. W., Slowinski, R., & Ziarko, W. (1995). Rough sets. *Communications of the ACM, 38*(11), 89–95.

Ramaswamy, S., Rastogi, R., Shim, K. (2000). Efficient algorithms for mining outliers from large datasets. In *Proceedings of the 2000 ACM SIGMOD international conference on management of data, Dallas* (pp. 427–438).

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York: John Wiley and Sons.

Rulequest Research. GritBot. see <http://www.rulequest.com>.

Skowron, A., & Rauszer, C. (1992). The discernibility matrices and functions in information systems. In R. Slowinski (Ed.), *Intelligent decision support – Handbook of applications and advances of the rough sets theory* (pp. 331–362). Dordrecht: Kluwer.

Skowron, A., & Synak, P. (2004). Reasoning in information maps. *Fundamenta Informaticae, 59*, 241–259.

Stanfill, C., & Waltz, D. (1986). Towards memory-based reasoning. *Communications of the ACM, 29*(12), 1213–1228.

Willams, G. J., Baxter, R. A., He, H. X., Harkins, S., Gu, L. F. (2002). A comparative study of RNN for outlier detection in data mining. In *ICDM, Japan* (pp. 709–712).

Wilson, D. R., & Martinez, T. R. (1997). Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research, 6*, 1–34.

Zhong, N., Yao, Y. Y., Ohshima, M., & Ohsuga, S. (2001). Interestingness, peculiarity, and multi-database mining. In *Proceedings of the 2001 IEEE international conference on data mining (IEEE ICDM'01)* (pp. 566–573). IEEE Computer Society Press.