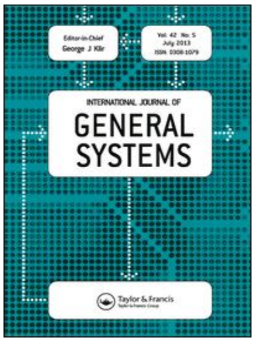


International Journal of General Systems

ISSN: 0308-1079 (Print) 1563-5104 (Online) Journal homepage: http://www.tandfonline.com/loi/ggen20

A rough set approach to outlier detection

Feng Jiang , Yuefei Sui & Cungen Cao

To cite this article: Feng Jiang , Yuefei Sui & Cungen Cao (2008) A rough set approach to outlier detection, International Journal of General Systems, 37:5, 519-536, DOI: 10.1080/03081070701251182

To link to this article: https://doi.org/10.1080/03081070701251182

Published online: 22 Aug 2008.

International Journal of General Systems
Vol. 37, No. 5, October 2008, 519–536

A rough set approach to outlier detection

Feng Jiang[ab]*, Yuefei Sui[a] and Cungen Cao[a]

[a]*Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, P.R. China;* [b]*Graduate School of Chinese Academy of Sciences, Beijing, P.R. China*

(*Received 8 May 2008; final version received 16 January 2008*)

"One person's noise is another person's signal" (Knorr and Ng 1998). 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. Detecting such outliers is important for many applications such as criminal activities in electronic commerce, computer intrusion attacks, terrorist threats, agricultural pest infestations. In this paper, we suggest to exploit the framework of rough sets for detecting outliers. We propose a novel definition of outliers—*RMF (rough membership function)-based outliers*, by virtue of the notion of rough membership function in rough set theory. An algorithm to find such outliers is also given. And the effectiveness of RMF-based method is demonstrated on two publicly available data sets.

Keywords: outlier detection; rough sets; rough membership function; KDD

1. Introduction

Knowledge discovery in databases (KDD), or data mining, is an important issue in the development of data- and knowledge-base 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 and 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 rest 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. Studying the extraordinary behaviours of outliers can help us uncover the valuable information hidden behind them. Recently researchers have begun focusing on outlier detection and attempted to design algorithms for 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).

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 outliers, 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). With increasing awareness on outlier detection in literatures, more concrete meanings of outliers are defined for solving problems in specific domains. Nonetheless, most of these definitions follow the spirit of Hawkins' (Chiu and Fu 2003).

*Corresponding author. Email: jiangkong@163.net

ISSN 0308-1079 print/ISSN 1563-5104 online

DOI: 10.1080/03081070701251182
http://www.informaworld.com

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

(1) *Distribution-based approach* is the classical method in statistics. It is based on some standard distribution model (e.g. Normal, and Poisson) and those objects which deviate from the model are recognized as outliers (Rousseeuw and Leroy 1987). Its biggest 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 follows (if there is any).

(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 *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.

(3) *Clustering approach* 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.

(4) *Distance-based approach* was originally proposed by Knorr and Ng (1998) and 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 kind of outliers are 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 *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 neighbourhood.

Rough set theory introduced by Pawlak (1982), Pawlak (1991), Pawlak and Skowron (1994) and Pawlak *et al.* (1995), as an extension of naive set theory, 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 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. Successful applications of the rough set model in a variety of problems have demonstrated its importance and versatility.

To our best knowledge, there are few works 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. 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 a set of attributes, V the union of attribute domains, and $f{:}U \times A \rightarrow V$ a function such that for any $x \in U$ and $a \in A$, $f(x,a) \in V_a$. In IS, each attribute subset $B \subseteq A$ determines an indiscernibility relation IND(B) on U. For any $X \subseteq U$ ($X \neq \emptyset$), $B \subseteq A$ and $x \in X$, the rough membership function (RMF) $\mu_X^B : X \rightarrow (0, 1]$ expresses how strongly object x belongs to set X in view of available information about x expressed by B (i.e. indiscernibility relation IND(B)). When given a set of indiscernibility relations (or available information/knowledge) on U, if the values of RMF of x wrt X under these indiscernibility relations are always small, then we may consider the object x as not behaving normally according to the given knowledge at hand. We call such objects *RMF (rough membership function)-based outliers* wrt X. Therefore, we can say that a RMF-based outlier wrt X is an element whose degrees of membership wrt X are always small in view of the given knowledge.

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 RMF-based outliers in information systems of rough set theory. An example and an algorithm to find RMF-based outliers are also given. Experimental results are given in Section 4. Finally, Section 5 concludes the paper.

2. Preliminaries

In a rough set data model, information is represented 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: $U \times A \rightarrow V$ 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 the indiscernibility relation, and is defined as follows:

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

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

Given any $B \subseteq A$, the 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

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

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

The pair $(\underline{X}_B, \overline{X}_B)$ is called the rough set with respect to X. 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 $\overline{X}_B$ may or may not belong to X.

In classical set theory, either an element belongs to a set or it does not. The corresponding membership function is the characteristic function for the set, i.e. the function takes values 1 or 0 for an element. In the case of rough sets, the notion of membership is different. A RMF is usually defined as follows (Pawlak and Skowron 1994).

Definition 2.1. Let IS $= (U, A, V, f)$ be an information system, $B \subseteq A$, $X \subseteq U$. The function $\mu_X^B : U \rightarrow [0, 1]$ such that for any $x \in U$

$$\mu_X^B(x) = \frac{|[x]_B \cap X|}{|[x]_B|} \tag{4}$$

is called a RMF, where $[x]_B$ denotes the indiscernibility class of relation IND(B) that contains element x, and $|M|$ denotes the cardinality of set M.

The following proposition collects the basic properties for the RMF of definition 2.1 (Pawlak and Skowron 1994).

Proposition 2.2. The RMF μ_X^B of definition 2.1 has the following properties

(1) $\mu_X^B(x) = 1$ iff $x \in \underline{X}_B$;
(2) $\mu_X^B(x) = 0$ iff $x \in U - \overline{X}_B$;
(3) $0 < \mu_X^B(x) < 1$ iff $x \in BN_B(X)$;
(4) if $(x, y) \in \mathrm{IND}(B)$ then $\mu_X^B(x) = \mu_X^B(y)$;
(5) $\mu_X^B(x) = 1 - \mu_{U-X}^B(x)$;
(6) $\mu_{X\cup Y}^B(x) \geq \max\{\mu_X^B(x), \mu_Y^B(x)\}$;
(7) $\mu_{X\cap Y}^B(x) \leq \min\{\mu_X^B(x), \mu_Y^B(x)\}$;
(8) for any pairwise disjoint collection P of concepts, $\mu_{\cup P}^B(x) = \sum\limits_{Y\in P} \mu_Y^B(x)$.

3. Rough membership function-based outliers

3.1 Definitions

First, in order to be used in outlier detection, a slightly different definition of RMF is given below.

Definition 3.1. Let IS $= (U, A, V, f)$ be an information system, $B \subseteq A$, $X \subseteq U$ and $X \neq \emptyset$. The function $\mu_X^B : X \longrightarrow (0, 1]$ such that for any $x \in X$

$$\mu_X^B(x) = \frac{|[x]_B \cap X|}{|[x]_B|} \tag{5}$$

is called a RMF, 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, and $|M|$ denotes the cardinality of set M.

In definition 3.1, the domain of the RMF is a subset X of U, not the universe U, as it is the case in definition 2.1 of Section 2. That is, in definition 3.1, we do not consider the degrees of membership wrt X for all elements that belong to set $U - X$. Correspondingly, the basic properties for the RMF of definition 3.1 will also differ somewhat from those in proposition 2.2 of Section 2. We use another proposition to represent them.

Proposition 3.2. The RMF μ_X^B of definition 3.1 has the following properties

(1) $\mu_X^B(x) = 1$ iff $x \in \underline{X}_B$;
(2) $\mu_X^B(x) > 0$;
(3) $\mu_X^B(x) < 1$ iff $x \in X - \underline{X}_B$;
(4) if $(x, y) \in \mathrm{IND}(B)$ then $\mu_X^B(x) = \mu_X^B(y)$;
(5) if $\mathrm{IND}(B) \cap (X \times (U - X)) = \emptyset$ then $\mu_X^B(x) = 1$;
(6) given $X, Y \subseteq U$, for any $x \in X \cap Y$, $\mu_{X\cup Y}^B(x) = \mu_X^B(x) + \mu_Y^B(x) - (|[x]_B \cap X \cap Y|)/(|[x]_B|)$
(7) given $X, Y \subseteq U$, for any $x \in X \cap Y$, $\mu_{X\cap Y}^B(x) = \mu_X^B(x) + \mu_Y^B(x) - (|[x]_B \cap (X \cup Y)|)/(|[x]_B|)$;
(8) given $X_1, X_2 \subseteq U$ and $X_1 \subseteq X_2$, for any $x \in X_1$, $\mu_{X_1}^B(x) \leq \mu_{X_2}^B(x)$.

In a rough set data model, a data set can be formally described using an information system. Therefore we should discuss the issue of outlier detection in an information system. Our definition of RMF-based outliers in an information system follows the spirit of Hawkins' one. 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 the attributes in A, we may call x an outlier wrt X in IS.

Especially, our definition of RMF-based outliers has a characteristic that is ignored by most current definitions of 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 wrt any subset X of U, where X maybe a particular subset of U which we are interested in or anything else which we are willing to separate from other elements of U.

Furthermore, 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.* (2000) proposed a method for identifying density-based local outliers. He defines a local outlier factor (LOF) that indicates the degree of being an outlier for every object using only the object's neighbourhood. Similar to Breunig's method, we shall define a rough outlier factor (ROF), which can indicate the degree of being an outlier for every object wrt a given subset of the universe in an information system.

It should be noted that in paper (Jiang *et al.* 2006b), we gave a definition of rough outlier factor (ROF). But in definition 3.1 of Jiang *et al.* (2006b), for any $x \in X \subseteq U$ in an information system IS $= (U, A, V, f)$, we only considered the value of RMF of x wrt X under every indiscernibility relation IND($\{a\}$) determined by the singleton subset $\{a\}$ of A. For those indiscernibility relations that are jointly determined by more than one attribute of A, we omitted their influences on the rough outlier factor of x. Therefore, the connections among different attributes of A are missed. This may lead to inaccurate result since we only used a special kind of information from all the information/knowledge provided by the information system IS $= (U, A, V, f)$. It seems more reasonable to consider the values of RMF of x wrt X under indiscernibility relations determined by all possible subsets of attributes. However this may suffer from high computational complexity when applied to high-dimensional data. Therefore, in this paper, we propose a new definition for rough outlier factor (ROF), in which we not only consider indiscernibility relations determined by singleton subsets of A, but also indiscernibility relations determined by other subsets of A.

In the following, we first construct two kinds of sequences—sequence of attributes and sequence of attribute subsets.

Definition 3.3 [Sequence of Attributes]. Let IS $= (U, A, V, f)$ be an information system, $A = \{a_1, a_2, \ldots, a_m\}$, $X \subseteq U$ and $X \neq \emptyset$. For every $x \in X$, we construct a *sequence* $S_x = \langle a'_1, a'_2, \ldots, a'_m \rangle$ *of attributes* in A, such that for every $1 \leq j < m$, $\mu_X^{\{a'_j\}}(x) \leq \mu_X^{\{a'_{j+1}\}}(x)$, where $\mu_X^{\{a'_j\}} : X \rightarrow (0, 1]$ is a RMF defined in definition 3.1, 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 3.4 [Descending Sequence of Attribute Subsets]. Let IS $= (U, A, V, f)$ be an information system, $A = \{a_1, a_2, \ldots, a_m\}$, $X \subseteq U$ and $X \neq \emptyset$. For every $x \in X$, let $S_x = \langle a'_1, a'_2, \ldots, a'_m \rangle$ be the sequence of attributes with respect to x 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 $1 \leq j < m$, then we call AS_x a *descending sequence of attribute subsets* with respect to object x in IS.

By the above definition, in $AS_x = \langle A_1, A_2, \ldots, A_m\rangle$, for every $1 \leq j < m$, 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 $\mathrm{IS} = (U, A, V, f)$, and a subset $X \subseteq U$, for every object $x \in X$, we can uniquely determine a sequence S_x of attributes and a descending sequence AS_x of attribute subsets with respect to x.

Definition 3.5 [Rough Outlier Factor]. Let $\mathrm{IS} = (U, A, V, f)$ be an information system, $A = \{a_1, a_2, \ldots, a_m\}$, $X \subseteq U$ and $X \neq \varnothing$. For any $x \in X$, let $AS_x = \langle A_1, A_2, \ldots, A_m\rangle$ be the descending sequence of attribute subsets with respect to x, the *rough outlier factor of* x wrt X in IS is defined as

$$\mathrm{ROF}_X(x) = 1 - \frac{\sum_{j=1}^{m}\left(\mu_X^{A_j}(x)\times|A_j|\right)+\sum_{j=1}^{m}\left(\mu_X^{\{a_j\}}(x)\times W_X^{\{a_j\}}(x)\right)}{2\times|A|^2} \tag{6}$$

where $\mu_X^{A_j}$ and $\mu_X^{\{a_j\}}$ are RMFs defined in definition 3.1, for every attribute subset $A_j \subseteq A$ and singleton subset $\{a_j\}$ of A, $1 \leq j \leq m$. For every singleton subset $\{a_j\}$, $W_X^{\{a_j\}} : X \to (0, 1]$ is a weight function such that for any $x \in X$, $W_X^{\{a_j\}}(x) = \sqrt{(|[x]_{\{a_j\}}|)/(|U|)}$. $[x]_{\{a_j\}} = \{u \in U : f(u, a_j) = f(x, a_j)\}$ denotes the indiscernibility class of relation $\mathrm{IND}(\{a_j\})$ that contains element x.

The weight function $W_X^{\{a_j\}}$ 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 smaller the weights, the bigger the rough outlier factor, the minority of objects should have smaller weights than the majority of objects. Therefore if the objects in U that are indiscernible with x are few, that is, the percentage of objects in U that are indiscernible with x is small, then we may consider that x belongs to the minority of objects, and assign a small weight to x.

Definition 3.6 [Rough Membership Function-Based Outliers]. Let $\mathrm{IS} = (U, A, V, f)$ be an information system, $X \subseteq U$ and $X \neq \varnothing$. Let ν be a given threshold value, for any $x \in X$, if $\mathrm{ROF}_X(x) > \nu$ then x is called a *rough membership function*(*RMF*)*-based outlier* wrt X in IS, where $\mathrm{ROF}_X(x)$ is the rough outlier factor of x wrt X in IS.

The above definitions accord with our basic idea about RMF-based outliers mentioned in Section 1. As shown in definition 3.5, the smaller the value of RMF, the bigger the rough outlier factor (i.e. the degree of being an outlier).

3.2 An example

Example 3.7. Given an information system $\mathrm{IS} = (U, A, V, f)$, where $U = \{u_1, u_2, u_3, u_4, u_5, u_6\}$, $A = \{a, b, c\}$, as shown in Table 1 below.

Let $X = \{u_1, u_2, u_5, u_6\}$ and threshold value $\nu = 0.6$.

(1) For object $u_1 \in X$, $[u_1]_{\{a\}} = \{u_1, u_3, u_5\}, [u_1]_{\{b\}} = \{u_1, u_2, u_4\}, [u_1]_{\{c\}} = \{u_1, u_3, u_4\}$. From definition 3.1, $\mu_X^{\{a\}}(u_1) = 2/3$; $\mu_X^{\{b\}}(u_1) = 2/3$; $\mu_X^{\{c\}}(u_1) = 1/3$. Therefore, from definition 3.3, the sequence of attributes with respect to u_1 is $S_{u_1} = \langle c, a, b\rangle$. Correspondingly, the descending sequence of attribute subsets with respect to u_1 is $AS_{u_1} = \langle\{a, b, c\}, \{a, b\}, \{b\}\rangle$.

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	1	1
u_6	1	1	2

From definition 3.5, the rough outlier factor of u_1 wrt X is as follows

$$\begin{aligned}\mathrm{ROF}_X(u_1) &= 1 - \frac{\sum_{j=1}^{m}\left(\mu_X^{A_j}(u_1)\times|A_j|\right)+\sum_{j=1}^{m}\left(\mu_X^{\{a_j\}}(u_1)\times W_X^{\{a_j\}}(u_1)\right)}{2\times|A|^2}\\ &= 1-\frac{(1\times 3)+(1\times 2)+\left(\frac{2}{3}\times 1\right)+\left(\frac{2}{3}\times\sqrt{\frac{3}{6}}\right)+\left(\frac{2}{3}\times\sqrt{\frac{3}{6}}\right)+\left(\frac{1}{3}\times\sqrt{\frac{3}{6}}\right)}{2\times 3^2}\\ &\approx 0.6197 > \nu.\end{aligned}$$

Therefore u_1 is a RMF-based outlier wrt X in IS.

(2) For object $u_2 \in X$, $[u_2]_{\{a\}} = \{u_2, u_6\}, [u_2]_{\{b\}} = \{u_1, u_2, u_4\}, [u_2]_{\{c\}} = \{u_2, u_5\}$. The sequence of attributes with respect to u_2 is $S_{u_2} = \langle b, a, c\rangle$. The descending sequence of attribute subsets with respect to u_2 is $AS_{u_2} = \langle\{a, b, c\}, \{a, c\}, \{c\}\rangle$.
Hence the rough outlier factor of u_2 wrt X is as follows

$$\begin{aligned}\mathrm{ROF}_X(u_2) &= 1 - \frac{\sum_{j=1}^{m}\left(\mu_X^{A_j}(u_2)\times|A_j|\right)+\sum_{j=1}^{m}\left(\mu_X^{\{a_j\}}(u_2)\times W_X^{\{a_j\}}(u_2)\right)}{2\times|A|^2}\\ &= 1-\frac{(1\times 3)+(1\times 2)+\left(\frac{2}{2}\times 1\right)+\left(\frac{2}{2}\times\sqrt{\frac{2}{6}}\right)+\left(\frac{2}{3}\times\sqrt{\frac{3}{6}}\right)+\left(\frac{2}{2}\times\sqrt{\frac{2}{6}}\right)}{2\times 3^2}\\ &\approx 0.5763 < \nu.\end{aligned}$$

Therefore u_2 is not a RMF-based outlier wrt X in IS.

(3) For object $u_5 \in X$, $[u_5]_{\{a\}} = \{u_1, u_3, u_5\}, [u_5]_{\{b\}} = \{u_5, u_6\}, [u_5]_{\{c\}} = \{u_2, u_5\}$. The sequence of attributes with respect to u_5 is $S_{u_5} = \langle a, b, c\rangle$. And the descending sequence of attribute subsets with respect to u_5 is $AS_{u_5} = \langle\{a, b, c\}, \{b, c\}, \{c\}\rangle$.
Hence the rough outlier factor of u_5 wrt X is as follows

$$\begin{aligned}\mathrm{ROF}_X(u_5) &= 1 - \frac{\sum_{j=1}^{m}\left(\mu_X^{A_j}(u_5)\times|A_j|\right)+\sum_{j=1}^{m}\left(\mu_X^{\{a_j\}}(u_5)\times W_X^{\{a_j\}}(u_5)\right)}{2\times|A|^2}\\ &= 1-\frac{(1\times 3)+(1\times 2)+\left(\frac{2}{2}\times 1\right)+\left(\frac{2}{3}\times\sqrt{\frac{3}{6}}\right)+\left(\frac{2}{2}\times\sqrt{\frac{2}{6}}\right)+\left(\frac{2}{2}\times\sqrt{\frac{2}{6}}\right)}{2\times 3^2}\\ &\approx 0.5763 < \nu.\end{aligned}$$

Therefore u_5 is not a RMF-based outlier wrt X in IS.

(4) For object $u_6 \in X$, $[u_6]_{\{a\}} = \{u_2, u_6\}, [u_6]_{\{b\}} = \{u_5, u_6\}, [u_6]_{\{c\}} = \{u_6\}$. The sequence of attributes with respect to u_6 is $S_{u_6} = \langle a, b, c\rangle$. The descending sequence of attribute subsets with respect to u_6 is $AS_{u_6} = \langle\{a, b, c\}, \{b, c\}, \{c\}\rangle$.

Hence the rough outlier factor of u_6 wrt X is as follows

$$\begin{aligned}\mathrm{ROF}_X(u_6) &= 1 - \frac{\sum_{j=1}^{m}\left(\mu_X^{A_j}(u_6)\times|A_j|\right)+\sum_{j=1}^{m}\left(\mu_X^{\{a_j\}}(u_6)\times W_X^{\{a_j\}}(u_6)\right)}{2\times|A|^2}\\ &= 1-\frac{(1\times 3)+(1\times 2)+(1\times 1)+\left(\frac{2}{2}\times\sqrt{\frac{2}{6}}\right)+\left(\frac{2}{2}\times\sqrt{\frac{2}{6}}\right)+\left(1\times\sqrt{\frac{1}{6}}\right)}{2\times 3^2}\\ &\approx 0.5798 < \nu.\end{aligned}$$

Therefore u_6 is not a RMF-based outlier wrt X in IS.

3.3 Algorithm for detecting RMF-based outliers

Algorithm 1

```
Input: information system IS = $(U, A, V, f)$ and a subset $X$ of $U$, where $|U| = n$, $|X| = n_X$ and $|A| = m$; threshold value $\nu$
Output: a set $E$ of RMF-based outliers wrt $X$ in IS
Initialization: Let $E = \varnothing$
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 and Nguyen 1996);
   Determine the partition $U/\mathrm{IND}(\{a\})$
}
For every $x \in X$
{
   For every $a \in A$
     {
        Calculate $\mu_X^{\{a\}}(x)$, the RMF of $x$ with respect
        to $X$ under relation IND($\{a\}$);
        Assign a weight $W_X^{\{a\}}(x)$ to $x$, where $W_X^{\{a\}}(x) = \sqrt{\frac{|[x]_{\{a\}}|}{|U|}}$
  }
     Determine the sequence $S_x = \{a'_1, a'_2, \ldots, a'_m\}$ of attributes in $A$ with
     respect to $x$, where for each $1 \le j < m$, $\mu_X^{\{a'_j\}}(x) \le \mu_X^{\{a'_{j+1}\}}(x)$;
     Construct a descending sequence $AS_x = \{A_1, \ldots, A_m\}$ of attribute subsets,
     in terms of sequence $S_x$;
     For $1 \le j \le m$
     {
        Determine set $[x]_{A_j}$, i.e. the set of all objects in $U$ that are
        indiscernible with $x$ under relation IND($A_j$);
        Calculate $\mu_X^{A_j}(x)$, the RMF of $x$ with respect
        to $X$ under relation IND($A_j$)
  }
  Calculate $\mathrm{ROF}_X(x)$, the rough outlier factor of $x$ wrt $X$, where
  $\mathrm{ROF}_X(x) = 1 - \frac{\sum_{j=1}^{m}\left(\mu_X^{A_j}(x)\times|A_j|\right)+\sum_{j=1}^{m}\left(\mu_X^{\{a_j\}}(x)\times W_X^{\{a_j\}}(x)\right)}{2\times|A|^2}$;
     If $\mathrm{ROF}_X(x) > \nu$ 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 $n = |U|$. 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 $k = |B|$ and $n = |U|$.

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

4. Experimental results

4.1 Experiment design

To evaluate RMF-based method for outlier detection, we ran our algorithm on real life data sets obtained from the UCI Machine Learning Repository (Bay 1999). In our previous papers (Jiang *et al.* 2005, 2006a), we have proposed two different methods for outlier detection in rough set theory. In this section, we compare the performance of RMF-based method with these two methods on identifying true outliers. First we give a brief description for the two methods.

In Jiang *et al.* (2006a), 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. In 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 order to compare distance-based method with RMF-based method, we revise the definitions of distance-based outlier detection by introducing a *distance outlier factor (DOF)* to indicate the degree of being an outlier for every object wrt a given subset of universe in an information system.

Definition 4.1 [Distance Outlier Factor]. Given an information system IS = (U, A, V, f) and $X \subseteq U$. Let d be a parameter. For any object $x \in X$, the percentage of the objects in X whose distance from x is greater than d is called the *distance outlier factor of* x wrt X in IS, formally denoted by

$$\mathrm{DOF}_X(x) = \frac{|\{y \in X : \mathrm{dist}(x, y) > d\}|}{|X|} \tag{7}$$

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 (Jiang *et al.* 2006a), and we set $d = |A|/2$.

Furthermore, in Jiang *et al.* (2005), to detect outliers in rough set theory, we first defined the notions of *inner boundary* and *boundary degree*, where the inner boundary of a given set $X \subseteq U$ under indiscernibility relation IND(B) is defined as $X - \underline{X}_B$. Then by virtue of the *boundary degree*, we defined the notion of *exceptional degree* for every object in X. Similar to ROF and DOF, the *exceptional degree* of an object indicates the degree of being an outlier for that object. Here we call the method in Jiang *et al.* (2005) *boundary-based outlier detection*. In order to compare boundary-based method with RMF-based method and obtain a finer effect for boundary-based method, we revise the definitions for boundary degree and exceptional degree in Jiang *et al.* (2005).

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

$$BD_X(x) = \sum_{j=1}^{m} \left(f(x, IB_j) \times W_X^{\{a_j\}} \right) \tag{8}$$

where f is a characteristic function for set IB_j (that is, if $x \in IB_j$ then $f(x, IB_j) = 1$ else $f(x, IB_j) = 0$). $W_X^{\{a_j\}} : X \longrightarrow [0, 1)$ is a weight function such that for any $x \in X$, $W_X^{\{a_j\}}(x) = 1 - (|[x]_{\{a_j\}} \cap X|/|X|)$, $1 \leq j \leq m$. $[x]_{\{a_j\}} = \{u \in U : f(u, a_j) = f(x, a_j)\}$ denotes the indiscernibility 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 $= (U, A, V, f)$, where $A = \{a_1, \ldots, a_m\}$. Let $X \subseteq U$ $(X \neq \emptyset)$. Let $IB = \{IB_1, IB_2, \ldots, IB_m\}$ be the set of all inner boundaries of X under each equivalence relation IND($\{a_j\}$), $1 \leq j \leq m$. For any object $x \in X$, the cardinality of set IB divided by the boundary degree of x wrt X is called the *exceptional degree of* x wrt X in IS, denoted by

$$ED_X(x) = \frac{BD_X(x)}{|IB|}. \tag{9}$$

4.2 Lymphography data

Next we demonstrate the effectiveness of RMF-based method against distance-based and boundary-based methods on two data sets. 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%).

Aggarwal and Yu proposed a practicable way to test the effectiveness of an outlier detection method (Aggarwal and 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 (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. 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.

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) wrt 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) $X_2 = \{x \in U : f(x, \text{changes\_node}) = 2 \vee f(x, \text{no\_nodes}) = 1\}$;
(3) $X_3 = \{x \in U : f(x, \text{spec\_froms}) = 3 \vee f(x, \text{dislocation}) = 1\}$;
(4) $X_4 = \{x \in U : f(x, \text{ changes\_lym}) = 2 \vee f(x, \text{ exclusion}) = 2\}$.

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 “changes_node“ equal 2 and those objects of U whose values on attribute “no_nodes“ 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), $1 \leq j \leq 4$.

The results from the three different outlier detection methods on the lymphography data set are summarized in Tables 2 and 3.

In Tables 2 and 3, $|X_j|$ denotes the number of objects in X_j, $|R_{X_j}|$ denotes the number of objects in X_j that belong to one of the rare classes, $1 \leq j \leq 4$. And "RMF", "RBD", "DIS" denote RMF-based, boundary-based and distance-based outlier detection methods, respectively. For every objects in X_j, the degree of being an outlier wrt X_j 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 X_j whose degrees of being an outlier wrt X_j calculated by the method are higher than 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 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 X_j that belong to one of the rare classes (i.e. $|R_{X_j}|$), $1 \leq j \leq 4$ (He *et al.* 2005).

From Tables 2 and 3, we can see that for the lymphography data set, RMF-based and distance-based methods perform markedly better than boundary-based method. And the performances of RMF-based and distance-based methods are very close.

4.3 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 nine 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, Williams *et al.* 2002, He *et al.* 2005). The resultant data set had 39 (8%) *malignant* instances and 444 (92%) *benign* instances. Moreover, the nine 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 Wisconsin breast cancer data set is also input into an information system $\mathrm{IS}_W = (U', A', V', f')$, where U' contains all the 483 instances of the data set and A' contains nine categorical attributes of the data set (not including the class attribute). We consider detecting outliers (*malignant* instances) wrt four subsets $X'_1, \ldots, X'_4$ of U', respectively, where

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

The results from the three different outlier detection methods on the Wisconsin breast cancer data set are summarized in Tables 4 and 5.

Tables 4 and 5 are similar to Tables 2 and 3, except that the "Number of *Malignant* Instances Included" is the number of objects in Y'_j that are *malignant*, 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 three outlier detection methods. The "Coverage" is the ratio of the "Number of *Malignant* Instances Included" to the number of objects in X'_j that are *malignant* (i.e. $|R_{X'_j}|$), $1 \leq j \leq 4$ (He *et al.* 2005).

From Tables 4 and 5, we can obviously see that for the Wisconsin breast cancer data set, RMF-based method performs better than the distance-based method and boundary-based

Table 2. Experimental results wrt X_1, X_2 in IS_L.

$X_1 : \lvert X_1 \rvert = 122, \lvert R_{X_1} \rvert = 4$				$X_2 : \lvert X_2 \rvert = 85, \lvert R_{X_2} \rvert = 4$			
	Number of rare classes included (coverage)				Number of rare classes included (coverage)		
Top ratio (number of objects)	RMF	RBD	DIS	Top ratio (number of objects)	RMF	RBD	DIS
2%(2)	2(50%)	2(50%)	2(50%)	2%(2)	2(50%)	2(50%)	2(50%)
3%(4)	3(75%)	2(50%)	4(100%)	4%(3)	3(75%)	3(75%)	3(75%)
4%(5)	4(100%)	2(50%)	4(100%)	5%(4)	3(75%)	3(75%)	4(100%)
20%(24)	4(100%)	2(50%)	4(100%)	7%(6)	4(100%)	3(75%)	4(100%)
84%(102)	4(100%)	3(75%)	4(100%)	8%(7)	4(100%)	3(75%)	4(100%)
88%(107)	4(100%)	4(100%)	4(100%)	9%(8)	4(100%)	4(100%)	4(100%)

Table 3. Experimental results wrt X_3, X_4 in IS_L.

$X_3 : \lvert X_3 \rvert = 105, \lvert \mathrm{R}_{X_3} \rvert = 5$				$X_4 : \lvert X_4 \rvert = 132, \lvert \mathrm{R}_{X_4} \rvert = 4$			
	Number of rare classes included (coverage)				Number of rare classes included (coverage)		
Top ratio (number of objects)	RMF	RBD	DIS	Top ratio (number of objects)	RMF	RBD	DIS
2%(2)	2(40%)	2(40%)	2(40%)	1%(1)	1(25%)	1(25%)	1(25%)
3%(3)	3(60%)	3(60%)	3(60%)	2%(3)	3(75%)	2(50%)	3(75%)
4%(4)	4(80%)	3(60%)	4(80%)	3%(4)	3(75%)	2(50%)	3(75%)
5%(5)	4(80%)	3(60%)	4(80%)	4%(5)	3(75%)	3(75%)	3(75%)
7%(7)	5(100%)	3(60%)	4(80%)	4.5%(6)	4(100%)	4(100%)	3(75%)
8%(8)	5(100%)	3(60%)	5(100%)	5%(7)	4(100%)	4(100%)	4(100%)
11%(12)	5(100%)	4(80%)	5(100%)				
24%(25)	5(100%)	5(100%)	5(100%)				

Table 4. Experimental results wrt X'_1, X'_2 in IS_W.

$X'_1 : \lvert X'_1 \rvert = 87, \lvert \mathrm{R}_{X'_1} \rvert = 4$				$X'_2 : \lvert X'_2 \rvert = 119, \lvert \mathrm{R}_{X'_2} \rvert = 9$			
	Number of Malignant instances included (coverage)				Number of Malignant instances included (coverage)		
Top ratio (number of objects)	RMF	RBD	DIS	Top ratio (number of objects)	RMF	RBD	DIS
2%(2)	2(50%)	2(50%)	2(50%)	3%(4)	4(44%)	4(44%)	4(44%)
3%(3)	3(75%)	3(75%)	2(50%)	5%(6)	6(67%)	5(56%)	4(44%)
5%(4)	3(75%)	3(75%)	3(75%)	6%(7)	6(67%)	5(56%)	5(56%)
6%(5)	4(100%)	3(75%)	3(75%)	7%(8)	7(78%)	5(56%)	6(67%)
7%(6)	4(100%)	3(75%)	4(100%)	8%(10)	8(89%)	5(56%)	7(78%)
8%(7)	4(100%)	4(100%)	4(100%)	9%(11)	9(100%)	6(67%)	8(89%)
				10%(12)	9(100%)	7(78%)	8(89%)
				11%(13)	9(100%)	8(89%)	9(100%)
				22%(26)	9(100%)	9(100%)	9(100%)

Table 5. Experimental results wrt X'_3, X'_4 in IS$_W$.

$X'_3 : \|X'_3\| = 194$, $\|R_{X'_3}\| = 12$				$X'_4 : \|X'_4\| = 454$, $\|R_{X'_4}\| = 23$			
	Number of Malignant instances included (coverage)				Number of Malignant instances included (coverage)		
Top ratio (number of objects)	RMF	RBD	DIS	Top ratio (number of objects)	RMF	RBD	DIS
1.5%(3)	3(25%)	3(25%)	3(25%)	1%(5)	4(17%)	4(17%)	4(17%)
3%(6)	5(42%)	5(42%)	5(42%)	2%(9)	8(35%)	7(30%)	6(26%)
4%(8)	7(58%)	7(58%)	7(58%)	3%(14)	12(52%)	11(48%)	10(43%)
5%(10)	8(67%)	8(67%)	8(67%)	4%(18)	15(65%)	13(57%)	12(52%)
6%(12)	10(83%)	9(75%)	10(83%)	5%(23)	18(78%)	18(78%)	15(65%)
7%(14)	11(92%)	10(83%)	11(92%)	6%(27)	20(87%)	20(87%)	18(78%)
8%(16)	12(100%)	11(92%)	11(92%)	7%(32)	22(96%)	21(91%)	23(100%)
9%(17)	12(100%)	11(92%)	12(100%)	7.2%(33)	23(100%)	21(91%)	23(100%)
10%(19)	12(100%)	12(100%)	12(100%)	8%(36)	23(100%)	21(91%)	23(100%)
				10%(45)	23(100%)	22(96%)	23(100%)
				12%(54)	23(100%)	23(100%)	23(100%)

method. Hence, this experiment also demonstrates the effectiveness of our RMF-based method for outlier detection.

5. Conclusion

Outlier detection is becoming critically important in many areas. In this paper, we presented a RMF-based method for outlier definition and outlier detection, in which we introduced a rough outlier factor (ROF) to indicate the degree of being an outlier for every object in an information system. The rough outlier factor is defined by virtue of the notion of RMF in rough sets. The main idea is that objects whose degrees of membership wrt a given subset of universe are small have more likelihood of being an outlier. In a given information system $\mathrm{IS} = (U, A, V, f)$, every attribute subset of *A* may determine an indiscernibility relation on *U*. We constructed a sequence of attributes and a sequence of attribute subsets for every object in *U*. When we calculated the rough outlier factor for an object, in order to obtain a more accurate result, we not only considered the degrees of membership under indiscernibility relations determined by all singleton subsets of *A*, but also the degrees of membership under indiscernibility relations determined by all attribute subsets in the sequence of attribute subsets. Experimental results on real data sets demonstrated the effectiveness of our method for outlier detection. The performance of our method is not worse than that of the traditional distance-based method.

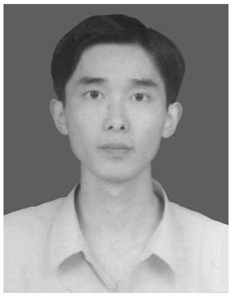
Acknowledgements

This work is supported by the Natural Science Foundation (grants no. 60273019, 60496326, 60573063 and 60573064), and the National 973 Programme (grants no. 2003CB317008 and G1999032701).

Note

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

Notes on contributors



Feng Jiang is currently a Ph.D. candidate of the Institute of Computing Technology, Chinese Academy of Sciences, China. His research interests include data mining, knowledge acquisition and the logical foundation of artificial intelligence.

Cungen Cao received the Ph.D. degree in Computer Software from the Institute of Mathematics, Chinese Academy of Sciences, China, in 1993. He is currently a professor of the Institute of Computing Technology, Chinese Academy of Sciences. His research interests are knowledge engineering and knowledge theory.

Yuefei Sui received the Ph.D. degree in Mathematics from the Institute of Software, Chinese Academy of Sciences, China, in 1988. He is currently a professor of the Institute of Computing Technology, Chinese Academy of Sciences. His research interests include the logical foundation of artificial intelligence and ontological engineering.

References

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

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

Bolton, R.J. and 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. and Sander, J., 2000. LOF: identifying density-based local outliers. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, Dallas, USA. 93–104.

Chiu, A.L. and Fu, A.W., 2003. Enhancements on local outlier detection. *Proceedings of the 7th International Database Engineering and Applications Symposium (IDEAS 03)*, Hong Kong. 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, ed. *Data mining for security applications*. Boston: Kluwer Academic Publishers.

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

Hawkins, D., 1980. *Identifications of outliers*. London: Chapman and Hall.

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

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

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

Jiang, F., Sui, Y.F. and Cao, C.G., 2006a. Some issues about outlier detection in rough set theory. Submitted to Special Issues on Rough Sets in China in LNCS Transactions on Rough Sets.

Jiang, F., Sui, Y.F. and Cao, C.G., 2006b. Outlier detection based on rough membership function. *Proceedings of the 5th International Conference on Rough Set and Knowledge Technology (RSCTC 2006). LNAI 4259*, Kobe, Japan. 388–397.

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

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

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

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

Lane, T. and Brodley, C.E., 1999. Temporal sequence learning and data reduction for anomaly detection. *ACM Transactions on Information and System Security*, 2 (3), 295–331.

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

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

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

Pawlak, Z., Skowron, A., 1994. Rough membership functions. In: R. Yager, ed. *Advances in the Dempster-Shafer theory of evidence*. New York: John Wiley & Sons, 251–271.

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

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

Williams, G.J., Baxter, R.A., He, H.X., Harkins, S. and Gu, L.F., 2002. A comparative study of RNN for outlier detection in data mining. *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002)*, Japan. 709–712.

Zhong, N., Yao, Y.Y., and Ohshima, M., 2003. Peculiarity oriented multi-database mining. *IEEE Transactions on Knowlegde and Data Engineering*, 15 (4), 952–960.

Zhong, N., Yao, Y.Y., Ohshima, M. and Ohsuga, S., 2001. Interestingness, Peculiarity, and Multi-Database Mining. *Proceedings 2001 IEEE International Conference on Data Mining (IEEE ICDM 01)*, IEEE Computer Society Press. 566–573.