Appl Intell (2015) 42:303–322
DOI 10.1007/s10489-014-0591-4

Outlier detection based on granular computing and rough set theory

Feng Jiang · Yu-Ming Chen

Published online: 9 October 2014

Abstract In recent years, outlier detection has attracted considerable attention. The identification of outliers is important for many applications, including those related to intrusion detection, credit card fraud, criminal activity in electronic commerce, medical diagnosis and anti-terrorism. Various outlier detection methods have been proposed for solving problems in different domains. In this paper, a new outlier detection method is proposed from the perspectives of granular computing (GrC) and rough set theory. First, we give a definition of outliers called GR(GrC and rough sets)-based outliers. Second, to detect GR-based outliers, an outlier detection algorithm called ODGrCR is proposed. Third, the effectiveness of ODGrCR is evaluated by using a number of real data sets. The experimental results show that our algorithm is effective for outlier detection. In particular, our algorithm takes much less running time than other outlier detection methods.

Keywords Outlier detection · Granular computing · Rough set theory · Accuracy of approximation · Data mining

1 Introduction

Recently, interest in the detection of outliers in data mining has grown considerably [14, 26, 27]. Many researchers have begun to focus on outlier detection and attempted to apply algorithms of outlier detection to tasks such as fraud detection [6], identification of computer network intrusions [11, 28], detection of employers with poor injury histories [27]. As an important task of data mining, outlier detection aims to find small groups of objects that are exceptional when compared with the remaining objects. From a data mining standpoint, studying the extraordinary behavior of outliers can help us uncover valuable information [14].

Outliers, generated from different sources, exist extensively in the real world. With increasing awareness of outlier detection in the literature, more concrete definitions of outliers have been made for solving problems in specific domains. For instance, Hawkins gave the following definition for outliers: "An outlier is an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism" [16].

Outlier detection has a long history in statistics [4, 16]. To avoid the problems of statistical methods, Knorr and Ng have proposed a distance-based approach [26], which uses the distance between any two objects as a measure of unusualness [3, 27, 48]. Although the distance-based approach is effective for outlier detection, its computational complexity is very high. For instance, the outlier detection algorithms based on nested loops typically require $O(n^2)$ distance computations, where n denotes the number of trheb objects.

In many application domains such as banking, government censuses, and geographic information systems, n is usually too large for an algorithm with $O(n^2)$ complexity to be practical. Moreover, we must select an appropriate metric for a distance-based approach, and it is difficult to do so for many practical tasks.

F. Jiang (✉)
College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, People's Republic of China
e-mail: jiangkong@163.net

Y.-M. Chen
Department of Computer Science, Xiamen University of Technology, Xiamen 361024, People's Republic of China

In general, outliers can be classified into three categories: *global outliers;contextual (or conditional) outliers*; and *collective outliers* [14]. For any object x in a given data set, x is a global outlier if it deviates significantly from the rest of the data set. Global outliers are sometimes called point anomalies. Most outlier detection methods are aimed at finding global outliers. Given a data set (T), a subset (S) of T forms a collective outlier if the objects in S as a whole deviate significantly from the full data set T. Importantly, the individual objects in S need not be outliers. In a given data set, an object is a contextual outlier if it deviates significantly with respect to a specific context for the object. In contextual outlier detection, the context has to be specified as part of the definition of the problem [14].

In this paper, we mainly discuss the definition and detection of global outliers. Based on granular computing (GrC) and rough set theory, we give a new definition of global outliers, GR-based outliers, and propose an algorithm to detect such outliers.

GrC is a general computational theory for using granules such as classes, clusters, subsets, groups, and intervals to build an efficient computational model for complex applications [34]. In 1979, Zadeh first introduced the notion of information granulation and suggested that fuzzy set theory may find potential applications in this respect [62, 64]. The term " granular computing" originated with a suggestion from Lin in discussion of the BISC Special Interest Group on Granular Computing [30, 64]. Lin proposed a method for GrC based on neighborhood systems [30, 31]. Yao and Zhong also examined some GrC methods with neighborhood systems [57, 58]. As an emerging conceptual and computing paradigm for information processing, GrC provides a conceptual framework for studying many issues in data mining and pattern recognition [32, 39, 43, 60].

In 1982, Pawlak proposed the theory of rough sets, which provides a concrete example of GrC [40]. To an extent, rough set theory emphasizes the importance of the notion of granulation [60]. Rough set theory has become a popular mathematical framework for GrC [34]. The essence of rough set theory is that it is a very general technique for approximating a given set. Approximation is carried out in terms of two sets, the lower and upper approximations. Via the lower and upper approximations, a basic notion, accuracy of approximation, is introduced, which can be used to measure the quality of approximation of decision classes on the universe U [40, 41].

Recently, there have been numerous applications of GrC and rough sets in the field of data mining [34, 47, 61], and many outlier detection algorithms based on rough sets or GrC have also been proposed. For instance, Nguyen proposed a method for the detection and evaluation of outliers using multilevel approximate reasoning schemes [37]. Shaari et al. proposed a new method to detect outliers using the concept of Non-Reduct from rough set theory [49]. Xue and Liu proposed a rough set-based semi-supervised outlier detection method [56]. Chen et al. proposed an outlier detection algorithm based on the neighborhood rough set model [9]. Albanese et al. extended outlier detection to spatiotemporal data by using a new rough set approach [2].

Moreover, in a series of papers, we have proposed various methods to detect outliers within the framework of rough set theory or GrC [8, 21–23]. In [21], we proposed a boundary-based outlier detection method using rough set theory. For any object x in boundary regions and any object y in lower approximations, the possibility of x being an outlier is greater than that of y. In [22], we presented an RMF(rough membership function)-based outlier detection method, by virtue of the notion of rough membership function in rough sets. Further, in [23] we proposed a BD(boundary and distance)-based method for outlier detection, which combines the boundary-based and distance-based methods [21, 26]. We also proposed a GrC-based outlier detection method [8], which uses the information table-based GrC model to detect outliers [58].

Although the above methods have demonstrated the effectiveness of rough set theory and GrC for outlier detection, some problems remain. For instance, the boundary-based and RMF-based methods are not feasible for dealing with very large data sets as their time complexities are too high [21, 22]. Moreover, when using the BD-based or GrC-based method to detect outliers, we must compute the distance between any two objects or two granules with respect to a given metric [8, 23]. However, it is difficult to select appropriate distance metrics for many practical tasks, and it is also difficult to set the values of distance parameters in each of the two methods. It may involve too many trials to find suitable distance metrics and set the values of parameters.

This paper is an extension of our previous work. To solve the problems of our previous methods, we propose a novel outlier detection approach based on GrC and rough set theory. Our approach uses the information table-based GrC model and rough set theory to detect outliers [40, 58]. Given an information table $IS = (U, A, V, f)$, for any object $x \in U$ and a set P of indiscernibility relations (i.e. available information/knowledge) on U we can obtain a granule containing x with respect to each relation in P. First, for each granule g containing x, we calculate the accuracy of approximation of g from the perspective of rough set theory [41]. Second, the degree of outlierness of granule g is calculated based on the accuracy of approximation of g. Finally, the degree of outlierness of object x is calculated by using the degrees of outlierness of those granules containing x. If

the degrees of outlierness of those granules containing x are always high, then the degree of outlierness of x is also high. Moreover, if the degree of outlierness of x is greater than a given threshold, then we call x a GR-based outlier in IS.

The remainder of this paper is organized as follows. In Section 2, we introduce some relevant preliminaries. In Section 3, we give the definition of GR-based outliers, as well as an example of finding such outliers. In Section 4, we present the outlier detection algorithm ODGrCR. Experimental results are given in Section 5, and a detailed discussion is presented in Section 6. Finally, Section 7 concludes the paper.

2 Preliminaries

In this section, we introduce some basic notions in GrC and rough set theory [40, 41, 58].

GrC may be regarded as a label of theories, methodologies, techniques, and tools that make use of granules in the process of problem solving. A general framework of GrC was presented by Zadeh in the context of fuzzy set theory [63]. Many specific models of GrC have also been proposed [29, 39, 42, 45, 46, 50, 52, 54, 65]. Yao and Zhong proposed a GrC model using information tables [58]. In an information table, each object of a finite nonempty universe is described by a finite set of attributes. Based on attribute values of objects, one may decompose the universe into parts called granules. Objects in each granule share the same or similar description in terms of their attribute values. Within this model, various methods for the construction, interpretation, and representation of granules were examined [58].

Moreover, Pedrycz and Vukovich [44], and Hu et al. [18], used information granulation and fuzzy sets for feature selection. Lin and Louie presented a fast association rule algorithm based on GrC [33]. Miao et al. proposed a web mining approach based on GrC [10, 34, 35].

In the rough set data model, information is stored in a table, where each row represents facts about an object. In rough set terminology, a data table is also called an *information table* [52], which is formally defined as follows [41, 51].

Definition 1 [Information Table] An information table 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 $x \in U$, $f(x, a) \in V_a$.

Given an information table $IS = (U, A, V, f)$, for any $B \subseteq A$, the *indiscernibility relation* $IND(B)$ on U is defined as [52]:

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

The indiscernibility relation $IND(B)$ is also an equivalence relation on U [41], which partitions U into disjoint equivalence classes. Let $U/IND(B)$ denote the set of all equivalence classes with respect to $IND(B)$, each element in $U/IND(B)$ may be viewed as a granule consisting of indiscernible objects in U [58]. Objects having the same description are indiscernible and may be put into the same granule [58], from which we can obtain a specific model of GrC, i.e., the information table-based GrC model. For any object $x \in U$, let $[x]_B$ denote the granule that contains x with respect to indiscernibility relation $IND(B)$.

Definition 2 [Lower and Upper Approximations] Given an information table $IS = (U, A, V, f)$, for any $B \subseteq A$ and $X \subseteq U$, the *B-lower and B-upper approximations* of set X are respectively defined as [41, 51]:

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

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

Definition 3 [Accuracy of Approximation] Given an information table $IS = (U, A, V, f)$, for any $B \subseteq A$ and $X \subseteq U$, the accuracy of approximation of set X with respect to relation $IND(B)$ is defined as [41, 51]:

$$\alpha_B(X) = \frac{|\underline{X}_B|}{|\overline{X}_B|}, \tag{4}$$

where $\underline{X}_B$ and $\overline{X}_B$ respectively denote the B-lower and B-upper approximations of set X.

It can be seen that $0 \leq \alpha_B(X) \leq 1$. If X is a union of some equivalence classes in $U/IND(B)$, then $\underline{X}_B = \overline{X}_B = X$, and hence $\alpha_B(X) = 1$. Moreover, $\alpha_B(X) = 0$ if and only if $\underline{X}_B = \emptyset$ [41, 51]. Especially, if $X = \emptyset$ then $\alpha_B(X) = 1$.

3 GR-based outliers

In this section, we first provide definitions concerning GR-based outliers, and then give an example of finding GR-based outliers.

3.1 Definitions

Given an information table $IS = (U, A, V, f)$, for any $X \subseteq U$ the current outlier detection methods do not usually consider the unusualness of X. Instead, they consider the unusualness of every object in U. It may be meaningful to discuss the unusualness of a given subset of objects, and so we will address this below. In the information table-based GrC model, for any object $x \in U$ and a set of indiscernibility relations on U we can obtain a granule containing x with respect to each of these relations. Before calculating the degree of outlierness of x, we first calculate the degrees of outlierness of those granules containing x, and the degree of outlierness of x is then determined by the degrees of outlierness of those granules containing x.

To calculate the degree of outlierness of a given granule, we use the accuracy of approximation in rough sets. The accuracy of approximation is an important notion in rough sets, which is typically used to measure the quality of approximation of decision classes on U [40, 41]. Here, we use it to define the degree of outlierness of a granule. For a given granule g, we calculate the accuracies of approximation of g with respect to a set of indiscernibility relations. If the accuracies of approximation of g with respect to these relations are always low, then we may consider g as not behaving normally and thus the degree of outlierness of g will be high.

The accuracy of approximation of a granule is defined as follows.

Definition 4 [Accuracy of Approximation of Granule] Given an information table $IS = (U, A, V, f)$ and any $B \subset A$, where $|A - B| \geq 2$. Let $G = U/IND(B)$ be the partition of U induced by $IND(B)$, for any granule $g \in G$ and $E \subseteq A - B$, the accuracy of approximation of g with respect to relation $IND(E)$ is defined as:

$$\alpha_E(g) = \frac{|\ \underline{g}_E\ |}{|\ \overline{g}_E\ |}, \tag{5}$$

where $\overline{g}_E$ and $\underline{g}_E$ respectively denote the E-upper and E-lower approximations of g.

For any object $x \in U$, most of the current outlier detection methods only give a binary classification of x, i.e., x is or is not an outlier [26, 27]. However, in many cases, it is more meaningful to assign a degree of being an outlier to x. In the following, we introduce two concepts: "degree of outlierness of granule" and "GR-based outlier factor (GROF)", where the former quantifies the degree of outlierness of a given granule, and the latter the degree of outlierness of a given object.

Definition 5 [Degree of Outlierness of Granule] Given an information table $IS = (U, A, V, f)$ and any $B \subset A$, where $|A - B| \geq 2$. Let $C = A - B = \{a_1, ..., a_n\}$, and $G = U/IND(B)$ denote the partition of U induced by relation $IND(B)$, for any granule $g \in G$, the *degree of outlierness* of g with respect to relation $IND(B)$ is defined as:

$$DOG_B(g) = 1 - \frac{\alpha_C(g) + \sum_{i=1}^{n} \left(\alpha_{C-\{a_i\}}(g) + 1\right)/2}{n+1} \times \frac{|\ g\ |}{|\ U\ |}, \tag{6}$$

where $\alpha_C(g)$ and $\alpha_{C-\{a_i\}}(g)$ respectively denote the accuracies of approximation of g with respect to relations $IND(C)$ and $IND(C - \{a_i\})$, $1 \leq i \leq n$.

In Definition 5, $\frac{|g|}{|U|}$ is the weight of granule g, which is in accordance with the opinion that outlier detection always concerns the minority of objects in the data set and these objects are more likely to be outliers than others. For any granule $g \in G$, if the cardinality of g is smaller than those of other granules in G, then we consider the objects in g as belonging to the minority of objects, and assign a low weight to g so that the degree of outlierness of g will be high.

In rough set theory, the uncertainty of a rough set can be measured by its roughness [41]. Given an information table $IS = (U, A, V, f)$, for any $X \subseteq U$ and $B \subseteq A$, the uncertainty of X with respect to indiscernibility relation $IND(B)$ can be measured by the roughness of X with respect to $IND(B)$, denoted by $\rho_B(X)$. Roughness is a complementary concept to the accuracy of approximation. Let $\alpha_B(X)$ denote the accuracy of approximation of X with respect to $IND(B)$, since $\alpha_B(X) = 1 - \rho_B(X)$, we can also use $\alpha_B(X)$ to measure the uncertainty of X [41].

In Definition 5, we use $DOG_B(g)$ to measure the abnormality of granule g, and the abnormality of g is described by the uncertainty of g. Because the accuracy of approximation of g can be used to measure the uncertainty of g, we calculate $DOG_B(g)$ by using the accuracy of approximation of g. Given a set of indiscernibility relations on U, if the accuracies of approximation of g with respect to these relations are always low, then we consider g as not behaving normally and $DOG_B(g)$ will be high. To calculate the accuracies of approximation of g with respect to various indiscernibility relations, we should not attempt to check all subsets of C, as each subset of C determines an indiscernibility relation on U, and so we would have $2^{|C|}$ relations. It is impracticable to calculate the accuracies of approximation of g with respect to all these relations, because the time complexity

will be exponential with respect to $|C|$. Therefore, we just calculate the accuracies of approximation of g with respect to relations $IND(C)$ and $IND(C - \{a_i\})$, $1 \le i \le n$.

Definition 6 [GR-based Outlier Factor] Given an information table $IS = (U, A, V, f)$, where $|A| \ge 3$. For any object $x \in U$, the *GR(GrC and rough sets)-based outlier factor* of x in *IS* is defined as:

$$GROF(x) = \frac{\sum_{a \in A} \left(DOG_{\{a\}}([x]_{\{a\}}) \times W_{\{a\}}(x) \right)}{|A|}, \tag{7}$$

where for every singleton subset $\{a\}$ of A, $W_{\{a\}} : U \to [0, 1)$ is a weight function such that for any $x \in U$, $W_{\{a\}}(x) = 1 - \sqrt[3]{|[x]_{\{a\}}|/|U|}$. $[x]_{\{a\}}$ denotes the granule in $U/IND(\{a\})$ that contains object x, and $DOG_{\{a\}}([x]_{\{a\}})$ denotes the *degree of outlierness* of granule $[x]_{\{a\}}$ with respect to $IND(\{a\})$.

Each subset of A determines an indiscernibility relation on U, and we can obtain a granule containing x with respect to each of these relations, hence there exist $2^{|A|}$ granules containing x in *IS*. It is impracticable to calculate the degrees of outlierness of all granules containing x, because the time complexity is exponential with respect to $|A|$. Therefore, in Definition 6 we use only the $|A|$ granules determined by each singleton subset $\{a\}$ of A, to calculate $GROF(x)$.

The weight function $W_{\{a\}}$ of Definition 6 is also in accordance with the opinion that outlier detection always concerns the minority of objects in the data set and these objects are more likely to be outliers than others. From Definition 6, it can be seen that the higher the weight, the greater the GR-based outlier factor, hence the minority of objects should have higher weight than the majority of objects. For any $x \in U$ and $a \in A$, if the cardinality of the granule $[x]_{\{a\}}$ in $U/IND(\{a\})$ is smaller than those of other granules in $U/IND(\{a\})$, then we consider x as belonging to the minority of objects in U, and assign a high weight to x.

Equations (6) and (7) in Definitions 5 and 6 are derived from Hawkins's opinion about outliers [16], that is, given an information table $IS = (U, A, V, f)$, for any object $x \in U$, if x has some abnormal characteristics with respect to other objects in U, then we may consider x as an outlier in *IS*. In (6) and (7), the uncertainty is regarded as the abnormal characteristic. If the uncertainty of a granule g is always very high, then we may consider g as not behaving normally and the degree of outlierness of g will be high. Furthermore, from the perspective of GrC, a granule is a clump of objects drawn together by indistinguishability, similarity or proximity. For any $x \in U$, if the degrees of outlierness of the granules containing x are always high, then the degree of outlierness of x should also be high. Therefore, in (7), the degree of outlierness of x is proportional to the degrees of outlierness of the granules containing x.

Definition 7 [GR-based Outliers] Given an information table $IS = (U, A, V, f)$, where $|A| \ge 3$. Let μ be a given threshold, for any $x \in U$, if $GROF(x) > \mu$ then we call object x a *GR(GrC and rough sets)-based outlier* in IS, where $GROF(x)$ is the GR-based outlier factor of x in *IS*.

3.2 An example

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

Let threshold $\mu = 0.27$. From Table 1, we can obtain the partitions of U with respect to various subsets of A, where

$$\begin{aligned}
U/IND(A) &= \{\{u_1\}, \{u_2\}, \{u_3\}, \{u_4\}, \{u_5\}, \{u_6\}\};\\
U/IND(\{a\}) &= \{\{u_1, u_3\}, \{u_2, u_5, u_6\}, \{u_4\}\};\\
U/IND(A - \{a\}) &= \{\{u_1\}, \{u_2, u_4\}, \{u_3\}, \{u_5\}, \{u_6\}\};\\
U/IND(\{b\}) &= \{\{u_1\}, \{u_3, u_6\}, \{u_2, u_4, u_5\}\};\\
U/IND(A - \{b\}) &= \{\{u_1\}, \{u_2\}, \{u_3\}, \{u_4\}, \{u_5\}, \{u_6\}\};\\
U/IND(\{c\}) &= \{\{u_1, u_6\}, \{u_2, u_4\}, \{u_3, u_5\}\};\\
U/IND(A - \{c\}) &= \{\{u_1\}, \{u_2, u_5\}, \{u_3\}, \{u_4\}, \{u_6\}\}.
\end{aligned}$$

Let $U/IND(\{a\}) = \{g_1, g_2, g_3\} = \{\{u_1, u_3\}, \{u_2, u_5, u_6\}, \{u_4\}\}$, for each granule $g \in U/IND(\{a\})$, the accuracies of approximation of g

Table 1 Information table *IS*

$U \backslash A$	a	b	c
u_1	0	0	3
u_2	1	2	1
u_3	0	1	2
u_4	2	2	1
u_5	1	2	2
u_6	1	1	3

with respect to relations $IND(A-\{a\})$, $IND(\{c\})$ and $IND(\{b\})$ are as follows.

$$\alpha_{A-\{a\}}(g_1) = \frac{|\ \underline{g_1}_{A-\{a\}}\ |}{|\ \overline{g_1}_{A-\{a\}}\ |} = \frac{|\ \{u_1,u_3\}\ |}{|\ \{u_1,u_3\}\ |} = 1;$$

$$\alpha_{\{c\}}(g_1) = \frac{|\ \underline{g_1}_{\{c\}}\ |}{|\ \overline{g_1}_{\{c\}}\ |} = \frac{|\ \emptyset\ |}{|\ \{u_1,u_3,u_5,u_6\}\ |} = 0;$$

$$\alpha_{\{b\}}(g_1) = \frac{|\ \underline{g_1}_{\{b\}}\ |}{|\ \overline{g_1}_{\{b\}}\ |} = \frac{|\ \{u_1\}\ |}{|\ \{u_1,u_3,u_6\}\ |} = 1/3;$$

$$\alpha_{A-\{a\}}(g_2) = \frac{|\ \underline{g_2}_{A-\{a\}}\ |}{|\ \overline{g_2}_{A-\{a\}}\ |} = \frac{|\ \{u_5,u_6\}\ |}{|\ \{u_2,u_4,u_5,u_6\}\ |} = 1/2;$$

$$\alpha_{\{c\}}(g_2) = \frac{|\ \underline{g_2}_{\{c\}}\ |}{|\ \overline{g_2}_{\{c\}}\ |} = \frac{|\ \emptyset\ |}{|\ \{u_1,u_2,u_3,u_4,u_5,u_6\}\ |} = 0;$$

$$\alpha_{\{b\}}(g_2) = \frac{|\ \underline{g_2}_{\{b\}}\ |}{|\ \overline{g_2}_{\{b\}}\ |} = \frac{|\ \emptyset\ |}{|\ \{u_2,u_3,u_4,u_5,u_6\}\ |} = 0;$$

$$\alpha_{A-\{a\}}(g_3) = \frac{|\ \underline{g_3}_{A-\{a\}}\ |}{|\ \overline{g_3}_{A-\{a\}}\ |} = \frac{|\ \emptyset\ |}{|\ \{u_2,u_4\}\ |} = 0.$$

$$\alpha_{\{c\}}(g_3) = \frac{|\ \underline{g_3}_{\{c\}}\ |}{|\ \overline{g_3}_{\{c\}}\ |} = \frac{|\ \emptyset\ |}{|\ \{u_2,u_4\}\ |} = 0;$$

$$\alpha_{\{b\}}(g_3) = \frac{|\ \underline{g_3}_{\{b\}}\ |}{|\ \overline{g_3}_{\{b\}}\ |} = \frac{|\ \emptyset\ |}{|\ \{u_2,u_4,u_5\}\ |} = 0.$$

From the above results, we can calculate the degree of outlierness of each granule in $U/IND(\{a\})$ with respect to relation $IND(\{a\})$, where

$$DOG_{\{a\}}(g_1) = 1 - \frac{\alpha_{A-\{a\}}(g_1) + \left(\alpha_{\{c\}}(g_1)+1\right)/2 + \left(\alpha_{\{b\}}(g_1)+1\right)/2}{3} \times \frac{|\ g_1\ |}{|\ U\ |} = 1 - \frac{1+\frac{1}{2}+\frac{2}{3}}{3} \times \frac{2}{6} = \frac{41}{54};$$

$$DOG_{\{a\}}(g_2) = 1 - \frac{\alpha_{A-\{a\}}(g_2) + \left(\alpha_{\{c\}}(g_2)+1\right)/2 + \left(\alpha_{\{b\}}(g_2)+1\right)/2}{3} \times \frac{|\ g_2\ |}{|\ U\ |} = 1 - \frac{\frac{1}{2}+\frac{1}{2}+\frac{1}{2}}{3} \times \frac{3}{6} = \frac{3}{4};$$

$$DOG_{\{a\}}(g_3) = 1 - \frac{\alpha_{A-\{a\}}(g_3) + \left(\alpha_{\{c\}}(g_3)+1\right)/2 + \left(\alpha_{\{b\}}(g_3)+1\right)/2}{3} \times \frac{|\ g_3\ |}{|\ U\ |} = 1 - \frac{0+\frac{1}{2}+\frac{1}{2}}{3} \times \frac{1}{6} = \frac{17}{18}.$$

In a similar method, we can calculate the degree of outlierness of each granule in $U/IND(\{b\})$ with respect to relation $IND(\{b\})$, where

$$DOG_{\{b\}}(\{u_1\}) = 1 - \frac{1+\frac{1}{2}+\frac{1}{2}}{3} \times \frac{1}{6} = \frac{8}{9};$$

$$DOG_{\{b\}}(\{u_3,u_6\}) = 1 - \frac{1+\frac{1}{2}+\frac{1}{2}}{3} \times \frac{2}{6} = \frac{7}{9};$$

$$DOG_{\{b\}}(\{u_2,u_4,u_5\}) = 1 - \frac{1+\frac{3}{4}+\frac{5}{8}}{3} \times \frac{3}{6} = \frac{29}{48}.$$

Moreover, the degree of outlierness of each granule in $U/IND(\{c\})$ with respect to relation $IND(\{c\})$ is as follows.

$$DOG_{\{c\}}(\{u_1,u_6\}) = 1 - \frac{1+\frac{2}{3}+\frac{1}{2}}{3} \times \frac{2}{6} = \frac{41}{54};$$

$$DOG_{\{c\}}(\{u_2,u_4\}) = 1 - \frac{\frac{1}{3}+\frac{1}{2}+\frac{5}{8}}{3} \times \frac{2}{6} = \frac{181}{216};$$

$$DOG_{\{c\}}(\{u_3,u_5\}) = 1 - \frac{\frac{1}{3}+\frac{1}{2}+\frac{1}{2}}{3} \times \frac{2}{6} = \frac{23}{27}.$$

From Definition 6, we can obtain that

$$GROF(u_1) = \frac{\sum_{a\in A}\left(DOG_{\{a\}}([u_1]_{\{a\}}) \times W_{\{a\}}(u_1)\right)}{|A|} = \frac{\frac{41}{54}\times\left(1-\frac{\sqrt[3]{9}}{3}\right) + \frac{8}{9}\times\left(1-\frac{\sqrt[3]{36}}{6}\right) + \frac{41}{54}\times\left(1-\frac{\sqrt[3]{9}}{3}\right)}{3} \approx 0.2885 > \mu.$$

In the same way, we can obtain that $GROF(u_2) \approx 0.1788 < \mu$; $GROF(u_3) \approx 0.2442 < \mu$; $GROF(u_4) \approx 0.2688 < \mu$; $GROF(u_5) \approx 0.1802 < \mu$; and $GROF(u_6) \approx 0.2087 < \mu$. Therefore, u_1 is a GR-based outlier in IS, and other objects are not GR-based outliers.

4 Outlier detection algorithm ODGrCR

In this section, we present the outlier detection algorithm ODGrCR. In ODGrCR, we need to calculate the accuracies of approximation of various granules, hence we first use the following algorithm to calculate the accuracies of approximation of granules.

Algorithm 1 Calculating the accuracies of approximation of granules.

Input: information table $IS = (U, A, V, f)$, $a \in A$, $a' \in C = A - \{a\}$ and the partitions $U/IND(\{a\}) = \{g_1, ..., g_m\}$, $U/IND(C-\{a'\})$, $U/IND((C-\{a'\}) \cup \{a\}) = U/IND(A-\{a'\})$.
Output: $\alpha_{C-\{a'\}}(g_i)$, $1 \le i \le m$.
Initialization: For each $1 \le j \le |U/IND(C-\{a'\})|$, $Flag[j] \leftarrow F$.

(1) For each $E \in U/IND(C-\{a'\})$, assign a number $N(E)$ to E, where $1 \le N(E) \le |U/IND(C-\{a'\})|$.
(2) Compute the cardinality of each granule in $U/IND(C-\{a'\})$ and $U/IND(A-\{a'\})$.
(3) For each granule $g_i \in U/IND(\{a\})$, $1 \le i \le m$
(4) {
(5) $count1 \leftarrow 0$.
(6) For each $x \in g_i$
(7) {
(8) If $|[x]_{(C-\{a'\})}| = |[x]_{(A-\{a'\})}|$ and $Flag[N([x]_{(C-\{a'\})})] = F$ then
(9) {
(10) $count1 \leftarrow count1 + |[x]_{(C-\{a'\})}|$.
(11) $Flag[N([x]_{(C-\{a'\})})] \leftarrow T$, where $[x]_{(C-\{a'\})} \in U/IND(C-\{a'\})$ and $1 \le N([x]_{(C-\{a'\})}) \le |U/IND(C-\{a'\})|$.
(12) }
(13) }
(14) $LA[i] \leftarrow count1$.
(15) }
(16) For each $1 \le j \le |U/IND(C-\{a'\})|$, $Flag[j] \leftarrow F$.
(17) For each $g_i \in U/IND(\{a\})$, $1 \le i \le m$
(18) {
(19) $count2 \leftarrow 0$.
(20) For each $x \in g_i$
(21) {
(22) If $Flag[N([x]_{(C-\{a'\})})] = F$ then
(23) {
(24) $count2 \leftarrow count2 + |[x]_{(C-\{a'\})}|$.
(25) $Flag[N([x]_{(C-\{a'\})})] \leftarrow T$.
(26) }
(27) }
(28) $UA[i] \leftarrow count2$.
(29) For each $x \in g_i$, if $Flag[N([x]_{(C-\{a'\})})] = T$ then $Flag[N([x]_{(C-\{a'\})})] \leftarrow F$.
(30) }
(31) For each $1 \le i \le m$, compute the accuracy of approximation of granule g_i with respect to $IND(C-\{a'\})$, i.e., $\alpha_{C-\{a'\}}(g_i) = LA[i]/UA[i]$.
(32) Return $\alpha_{C-\{a'\}}(g_i), 1 \le i \le m$.

In Algorithm 1, for each granule $g_i \in U/IND(\{a\})$, $1 \le i \le m$, to calculate the lower and upper approximations of g_i, we should examine each equivalence class $[x]_{(C-\{a'\})}$ in $U/IND(C-\{a'\})$. Besides object x, $[x]_{(C-\{a'\})}$ may contain other objects, hence the array $Flag$ is used to indicate whether $[x]_{(C-\{a'\})}$ has been examined. If $Flag[N([x]_{(C-\{a'\})})] = F$ then we examine $[x]_{(C-\{a'\})}$, else we omit it.

Moreover, we assume that the partitions $U/IND(\{a\})$, $U/IND(C-\{a'\})$ and $U/IND((C-\{a'\}) \cup \{a\})$ have already been calculated, that is, we need not to calculate them in Algorithm 1. Since $\sum_{g \in U/IND(C-\{a'\})} |g| = |U|$ and $\sum_{g \in U/IND(A-\{a'\})} |g| = |U|$, the time complexity of Step (2) is $O(|U|)$. Since $\sum_{g_i \in U/IND(\{a\})} |g_i| = |U|$, the time complexity of Steps (3)-(9) is $O(|U|)$, and the time complexity of Steps (11)-(19) is also $O(|U|)$. In the worst case, the time complexity of Algorithm 1 is $O(|U|)$, and its space complexity is $O(|U|)$.

Given an information table $IS = (U, A, V, f)$, for any $a \in A$ and each granule $g \in U/IND(\{a\})$, we can also calculate $\alpha_{A-\{a\}}(g)$ (i.e., the accuracy of approximation of g with respect to relation $IND(A-\{a\})$), by using a method similar to Algorithm 1.

Given any $B \subseteq A$, if we use the traditional method to calculate the partition $U/IND(B)$, the time complexity is $O(|U|^2)$. To reduce the time complexity for calculating $U/IND(B)$, Nguyen and Nguyen proposed an algorithm to calculate $U/IND(B)$ by sorting objects from U [36], and the time complexity is $O(|B||U| \log |U|)$.

In Algorithm 2, we adopt the counting sort based method to calculate the partition $U/IND(B)$ [55], and the time complexity is $O(|B||U| + k)$, where $k = \sum_{a \in B} k_a$ (for any $a \in B$, k_a is the range of attribute values in domain V_a).

Algorithm 2 **ODGrCR**.

Input: information table $IS = (U, A, V, f)$, threshold μ.
Output: set O of GR-based outliers.

(1) Calculate the partition $U/IND(A)$ based on the counting sort.
(2) For each $a \in A$, calculate partitions $U/IND(\{a\})$ and $U/IND(A-\{a\})$ based on the counting sort (we assume that $U/IND(\{a\}) = \{g_1, ..., g_m\}$).
(3) For each $a \in A$
(4) {
(5) For each $a' \in A - \{a\}$
(6) {
(7) Calculate $U/IND(A-\{a, a'\})$ based on the counting sort.
(8) For each $1 \le i \le m$, calculate $\alpha_{A-\{a,a'\}}(g_i)$ using Algorithm 1.
(9) }
(10) For each $1 \le i \le m$, calculate $\alpha_{A-\{a\}}(g_i)$.
(11) For each $1 \le i \le m$, calculate $DOG_{\{a\}}(g_i)$, the degree of outlierness of granule g_i with respect to relation $IND(\{a\})$.
(12) }
(13) For each $x \in U$
(14) {
(15) For each $a \in A$, assign a weight $W_{\{a\}}(x)$ to x.
(16) Calculate $GROF(x)$, the GR-based outlier factor of x in IS.
(17) If $GROF(x) > \mu$ then $O \leftarrow O \cup \{x\}$.
(18) }
(19) Return O.

Counting sort is a non-comparative sorting algorithm. It operates by counting the number of objects that have a distinct key value, and using arithmetic on those counts to determine the position of each key value in the output sequence. Since integers can represent strings of characters (e.g., names or dates) and specially formatted real numbers, counting sort is not limited to integers.

In Algorithm 2, we assume that for any $a \in B$, the domain V_a of a is a set of non-negative integers. If a is a categorical attribute, we may recode V_a with a set of non-negative integers. For example, an attribute representing color might have values such as red, green, blue, brown, black and white, which could be represented by 1-6, respectively. Moreover, if a is a continuous attribute, we may also recode V_a with a set of non-negative integers. For example, an attribute representing weight might have values such as 18.72, 18.73 and

33.27, which could be represented by 1872, 1873 and 3327, respectively. Based on the above assumption, we can calculate $U/IND(B)$ by using the counting sort based method.

In the worst case, the time complexity of Algorithm 2 is $O(|A|^2(|A||U|+k))$, where $k=\sum_{a\in A}k_a$ (for any $a\in A$, k_a is the range of attribute values in domain V_a). The space complexity of Algorithm 2 is $O(|A||U|+M)$, where $M=Max(\{k_a : a\in A\})$.

5 Experimental results

As mentioned earlier in Section 1, we have proposed four outlier detection methods in previous work, i.e., the boundary-based method [21]; the RMF-based method [22]; the BD-based method [23]; and the GrC-based method [8]. In this section, we will compare ODGrCR with the above four methods. In addition, we will compare ODGrCR with the other two methods: the traditional distance-based method [26] and the KNN-based method [48].

5.1 The experimental setup

To evaluate the performance of ODGrCR algorithm, the following four data sets were used:

(1) Lymphography data set;
(2) Breast Cancer data set;
(3) Yeast data set; and
(4) KDD Cup 1999 data set.

The above data sets were obtained from the UCI Machine Learning Repository [5]. Experiments were conducted on a 3.2GHz Pentium 4 machine with 2GB RAM, running the Windows XP operating system.

To evaluate the performance of a given outlier detection method, we adopted the evaluation metric proposed by Aggarwal and Yu [1]. The metric tests the outlier detection method on a given data set which contains several rare classes, and calculates the percentage of objects which belong to one of the rare classes. Those objects belonging to the rare classes are deemed as outliers. If the given outlier detection method works well, then we expect that such abnormal classes would be over-represented in the set of objects detected [1].

In the experiments, for the BD-based method, the overlap metric was used to calculate the distance between any two objects [23], and the three distance parameters d_1, d_2 and d_3 were respectively set as: $d_1=|A|/3$; $d_2=|A|/2$; and $d_3=0.9\times|A|$, where A is the set of attributes [23]. For the GrC-based method, we used the overlap metric to calculate the distance between any two granules [8]. For the KNN-based method, the parameter k was set to 5 [48].

The traditional distance-based outlier detection method only gives a binary classification of objects [26, 27]. To compare it with ODGrCR, we introduced a notion called "distance outlier factor (DOF)", which can indicate the degree of outlierness of a given object [22, 23].

Definition 8 [Distance Outlier Factor] Given an information table $IS=(U,A,V,f)$, for any object $x\in U$, the *distance outlier factor of* x in IS is defined as:

$$DOF(x)=\frac{|\{y\in U : dist(x,y)>d'\}|}{|U|}, \tag{8}$$

where $dist(x,y)$ denotes the distance between objects x and y, and d' is a given parameter.

5.2 Lymphography data set

The Lymphography data set contains 148 objects, which are partitioned into 4 classes:"normal find" (1.35 %), ''metastases" (54.73 %), " malign lymph" (41.22 %) and "fibrosis" (2.7 %), where"normal find" and "fibrosis" are regarded as rare classes.

In the experiments, data in the Lymphography data set was stored in an information table $IS_L=(U_L,A_L,V_L,f)$, where U_L contains 148 objects and A_L contains 19 attributes. Since the boundary-based, RMF-based and BD-based methods are designed to detect outliers with respect to some subsets of the given data set, to compare ODGrCR with these methods, we detected outliers with respect to four subsets $L_1,...,L_4$ of U_L. $L_1,...,L_4$ are respectively described as follows.

(1) $L_1=\{x\in U_L : f(x,\text{bl\_lymph\_c})=1\}$;
(2) $L_2=\{x\in U_L : f(x,\text{early\_uptake})=1 \vee f(x,\text{block\_affere})=1\}$;
(3) $L_3=\{x\in U_L : f(x,\text{special\_forms})=3 \vee f(x,\text{dislocation})=1\}$;
(4) $L_4=\{x\in U_L : f(x,\text{block\_affere})=1\}$.

For each $1\le j\le 4$, let R_{L_j} be the set of all outliers in L_j. For the parameter d in the GrC-based method and the parameter d' in the distance-based method, we set $d=|A_L|/2$ and $d'=|A_L|/2$. The results of various outlier detection methods on the Lymphography data set are given in Tables 2, 3, 4 and 5.

In Tables 2–5, "BD", "RMF", "RB", "DIS", "GrC" and "KNN" respectively denote the BD-based, RMF-based, boundary-based, distance-based, GrC-based and KNN-based outlier detection methods. For any given object x, various outlier detection methods were respectively used to calculate the degree of outlierness of x. Let M be the current outlier detection method, "Top ratio (number of objects)" denotes the percentage (number) of objects selected from

Table 2 Experimental results with respect to L_1 in IS_L

Top ratio (number of objects)	$L_1: \|L_1\| = 122, \|R_{L_1}\| = 4$						
	Number of rare classes included (coverage)						
	BD	RMF	RB	DIS	GrC	KNN	ODGrCR
2 %(2)	2(50 %)	2(50 %)	2(50 %)	2(50 %)	2(50 %)	2(50 %)	2(50 %)
3 %(4)	4(100 %)	3(75 %)	2(50 %)	4(100 %)	3(75 %)	2(50 %)	4(100 %)
4 %(5)	4(100 %)	4(100 %)	2(50 %)	4(100 %)	4(100 %)	2(50 %)	4(100 %)
6 %(7)	4(100 %)	4(100 %)	2(50 %)	4(100 %)	4(100 %)	3(75 %)	4(100 %)
6.5 %(8)	4(100 %)	4(100 %)	2(50 %)	4(100 %)	4(100 %)	4(100 %)	4(100 %)
88 %(107)	4(100 %)	4(100 %)	4(100 %)	4(100 %)	4(100 %)	4(100 %)	4(100 %)
Accuracy	100 %	89.6 %	8.2 %	100 %	89.6 %	73.3 %	100 %

Table 3 Experimental results with respect to L_2 in IS_L

Top ratio (number of objects)	$L_2: \|L_2\| = 90, \|R_{L_2}\| = 5$						
	Number of rare classes included (coverage)						
	BD	RMF	RB	DIS	GrC	KNN	ODGrCR
2 %(2)	2(40 %)	2(40 %)	2(40 %)	2(40 %)	2(40 %)	2(40 %)	2(40 %)
4 %(4)	4(80 %)	3(60 %)	3(60 %)	3(60 %)	4(80 %)	3(60 %)	4(80 %)
5 %(5)	4(80 %)	3(60 %)	3(60 %)	4(80 %)	4(80 %)	3(60 %)	5(100 %)
8 %(7)	5(100 %)	3(60 %)	3(60 %)	4(80 %)	5(100 %)	3(60 %)	5(100 %)
10 %(9)	5(100 %)	4(80 %)	3(60 %)	4(80 %)	5(100 %)	3(60 %)	5(100 %)
12 %(11)	5(100 %)	4(80 %)	3(60 %)	4(80 %)	5(100 %)	5(100 %)	5(100 %)
14 %(13)	5(100 %)	5(100 %)	3(60 %)	5(100 %)	5(100 %)	5(100 %)	5(100 %)
70 %(63)	5(100 %)	5(100 %)	5(100 %)	5(100 %)	5(100 %)	5(100 %)	5(100 %)
Accuracy	84.5 %	64.5 %	19.9 %	64.5 %	84.5 %	69.2 %	100 %

Table 4 Experimental results with respect to L_3 in IS_L

Top ratio (number of objects)	$L_3: \|L_3\| = 105, \|R_{L_3}\| = 5$						
	Number of rare classes included (coverage)						
	BD	RMF	RB	DIS	GrC	KNN	ODGrCR
2 %(2)	2(40 %)	2(40 %)	2(40 %)	2(40 %)	2(40 %)	2(40 %)	2(40 %)
4 %(4)	4(80 %)	4(80 %)	3(60 %)	4(80 %)	4(80 %)	3(60 %)	4(80 %)
6 %(6)	4(80 %)	4(80 %)	3(60 %)	4(80 %)	4(80 %)	4(80 %)	5(100 %)
7 %(7)	5(100 %)	5(100 %)	3(60 %)	4(80 %)	4(80 %)	4(80 %)	5(100 %)
8 %(8)	5(100 %)	5(100 %)	3(60 %)	5(100 %)	5(100 %)	4(80 %)	5(100 %)
11 %(12)	5(100 %)	5(100 %)	4(80 %)	5(100 %)	5(100 %)	4(80 %)	5(100 %)
12 %(13)	5(100 %)	5(100 %)	4(80 %)	5(100 %)	5(100 %)	5(100 %)	5(100 %)
24 %(25)	5(100 %)	5(100 %)	5(100 %)	5(100 %)	5(100 %)	5(100 %)	5(100 %)
Accuracy	84.7 %	84.7 %	50 %	79.8 %	79.8 %	65.2 %	91.2 %

Table 5 Experimental results with respect to L_4 in IS_L

Top ratio (number of objects)	L_4 : $\|L_4\| = 66$, $\|R_{L_4}\| = 4$						
	Number of rare classes included (coverage)						
	BD	RMF	RB	DIS	GrC	KNN	ODGrCR
2 %(1)	1(25 %)	1(25 %)	1(25 %)	1(25 %)	1(25 %)	1(25 %)	1(25 %)
4 %(3)	3(75 %)	3(75 %)	2(50 %)	3(75 %)	3(75 %)	2(50 %)	3(75 %)
6 %(4)	4(100 %)	3(75 %)	2(50 %)	4(100 %)	3(75 %)	2(50 %)	4(100 %)
9 %(6)	4(100 %)	4(100 %)	2(50 %)	4(100 %)	4(100 %)	3(75 %)	4(100 %)
10 %(7)	4(100 %)	4(100 %)	2(50 %)	4(100 %)	4(100 %)	4(100 %)	4(100 %)
65 %(43)	4(100 %)	4(100 %)	4(100 %)	4(100 %)	4(100 %)	4(100 %)	4(100 %)
Accuracy	100 %	81.7 %	23.2 %	100 %	81.7 %	76.2 %	100 %

L_j, whose degrees of outlierness calculated by M are greater than those of other objects in L_j, $1 \leq j \leq 4$. Let $S_j \subseteq L_j$ contain those objects selected from L_j,"Number of rare classes included" denotes the number of objects in S_j that are outliers, and "coverage" denotes the ratio of "Number of rare classes included" to $|R_{L_j}|$, $1 \leq j \leq 4$ [17, 22].

In Tables 2–5, "Accuracy" is used to measure the effectiveness of each outlier detection method on the current data set, which is defined as:

$$Accuracy(L_j) = \left(\frac{|R_{L_j}|}{|S_j|} + \left(1 - \frac{|S_j \cap N_{L_j}|}{|N_{L_j}|} \right) \right) / 2, \quad (9)$$

where N_{L_j} denotes the set of all normal objects in L_j, and S_j denotes the set of all objects selected from L_j by the given outlier detection method.

From Tables 2–5, it can be seen that for the Lymphography data set, ODGrCR algorithm has the best performance. For each of L_1, ..., L_4, the accuracy of ODGrCR is higher than or equal to those of other methods. For instance, in Table 3, when the top ratio (number of objects) is set to 5 %(5), the number of outliers detected by ODGrCR is 5, that is, the five objects selected by ODGrCR are all outliers in L_2, but for BD, RMF, RB, DIS, GrC and KNN, only 4, 3, 3, 4, 4 and 3 outliers are found, respectively. From another point of view, to find all outliers in L_2, ODGrCR needs to check 5 % of objects in L_2, but for BD, RMF, RB, DIS, GrC and KNN, they need to check 8 %, 14 %, 70 %, 14 %, 8 % and 12 % of objects in L_2, respectively. This also demonstrates the effectiveness of ODGrCR on the Lymphography data set. Especially, for L_1, L_2 and L_4, the accuracy of ODGrCR is 100 %, which means that there does not exist any misjudgment when applying ODGrCR to L_1, L_2 and L_4.

Table 6 Statistical information for the accuracies of various methods on Lymphography

Outlier detection methods	Mean	Standard deviation
BD	92.3 %	7.7 %
RMF	80.1 %	9.5 %
RB	25.3 %	15.3 %
DIS	86.1 %	14.9 %
GrC	83.9 %	3.7 %
KNN	71 %	4.2 %
ODGrCR	97.8 %	3.8 %

Table 6 gives the statistical information for the accuracies of various outlier detection methods on L_1, ..., L_4.

From Table 6, it can be seen that the average accuracy of ODGrCR is higher than those of BD, RMF, RB, DIS and KNN, and the standard deviation of the accuracies of ODGrCR is lower than those of the five methods. Although the standard deviation of the accuracies of ODGrCR is a little higher than that of GrC, the average accuracy of ODGrCR is markedly higher than that of GrC.

5.3 Breast cancer data set

The Breast Cancer data set contains 699 objects, which are partitioned into 2 classes: "benign" (65.5 %) and "malignant" (34.5 %) [5]. To form an unbalanced distribution, we removed some of the malignant objects from the data set [15, 17]. The resultant data set contains 39 (8 %) malignant objects and 444 (92 %) benign objects, where the malignant objects are regarded as outliers. Moreover, the 9 continuous attributes in the data set were respectively transformed into discrete attributes [1].

Data in the Breast Cancer data set was stored in an information table $IS_W = (U_W, A_W, V_W, f)$. We detected outliers with respect to four subsets W_1, ..., W_4 of U_W in IS_W. The four subsets W_1, ..., W_4 are respectively described as follows.

(1) $W_1 = \{x \in U_W : f(x, \text{Clump\_thickness}) = 5\}$;

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

Table 7 Experimental results with respect to W_1 in IS_W

Top ratio (number of objects)	$W_1: \lvert W_1\rvert = 87, \lvert R_{W_1}\rvert = 4$						
	Number of rare classes included (coverage)						
	BD	RMF	RB	DIS	GrC	KNN	ODGrCR
2 %(2)	2(50 %)	2(50 %)	2(50 %)	2(50 %)	2(50 %)	2(50 %)	2(50 %)
3 %(3)	3(75 %)	3(75 %)	3(75 %)	2(50 %)	3(75 %)	3(75 %)	3(75 %)
5 %(4)	4(100 %)	3(75 %)	3(75 %)	3(75 %)	3(75 %)	4(100 %)	4(100 %)
6 %(5)	4(100 %)	4(100 %)	3(75 %)	3(75 %)	3(75 %)	4(100 %)	4(100 %)
7 %(6)	4(100 %)	4(100 %)	3(75 %)	4(100 %)	4(100 %)	4(100 %)	4(100 %)
8 %(7)	4(100 %)	4(100 %)	4(100 %)	4(100 %)	4(100 %)	4(100 %)	4(100 %)
Accuracy	100 %	89.4 %	76.8 %	82.1 %	82.1 %	100 %	100 %

Table 8 Experimental results with respect to W_2 in IS_W

Top ratio (number of objects)	$W_2: \lvert W_2\rvert = 42, \lvert R_{W_2}\rvert = 5$						
	Number of rare classes included (coverage)						
	BD	RMF	RB	DIS	GrC	KNN	ODGrCR
7 %(3)	3(60 %)	3(60 %)	3(60 %)	2(40 %)	3(60 %)	3(60 %)	3(60 %)
10 %(4)	4(80 %)	4(80 %)	3(60 %)	3(60 %)	4(80 %)	4(80 %)	4(80 %)
15 %(6)	5(100 %)	4(80 %)	3(60 %)	5(100 %)	4(80 %)	5(100 %)	5(100 %)
17 %(7)	5(100 %)	4(80 %)	4(80 %)	5(100 %)	4(80 %)	5(100 %)	5(100 %)
20 %(8)	5(100 %)	5(100 %)	4(80 %)	5(100 %)	5(100 %)	5(100 %)	5(100 %)
21 %(9)	5(100 %)	5(100 %)	5(100 %)	5(100 %)	5(100 %)	5(100 %)	5(100 %)
Accuracy	90.3 %	77.2 %	72.4 %	90.3%	77.2 %	90.3 %	90.3 %

Table 9 Experimental results with respect to W_3 in IS_W

Top ratio (number of objects)	$W_3: \lvert W_3\rvert = 363, \lvert R_{W_3}\rvert = 7$						
	Number of rare classes included (coverage)						
	BD	RMF	RB	DIS	GrC	KNN	ODGrCR
1 %(4)	4(57 %)	4(57 %)	3(43 %)	3(43 %)	4(57 %)	4(57 %)	4(57 %)
2.2 %(8)	6(86 %)	6(86 %)	6(86 %)	6(86 %)	6(86 %)	6(86 %)	7(100 %)
2.5 %(9)	7(100 %)	6(86 %)	6(86 %)	7(100 %)	6(86 %)	7(100 %)	7(100 %)
2.8 %(10)	7(100 %)	7(100 %)	6(86 %)	7(100 %)	7(100 %)	7(100 %)	7(100 %)
4 %(15)	7(100 %)	7(100 %)	6(86 %)	7(100 %)	7(100 %)	7(100 %)	7(100 %)
6.2 %(23)	7(100 %)	7(100 %)	7(100 %)	7(100 %)	7(100 %)	7(100 %)	7(100 %)
Accuracy	88.6 %	84.6 %	63 %	88.6 %	84.6 %	88.6 %	93.6 %

Table 10 Experimental results with respect to W_4 in IS_W

Top ratio (number of objects)	W_4 : $\|W_4\| = 454$, $\|R_{W_4}\| = 23$						
	Number of rare classes included (coverage)						
	BD	RMF	RB	DIS	GrC	KNN	ODGrCR
1 %(5)	4(17 %)	4(17 %)	4(17 %)	4(17 %)	5(22%)	4(17 %)	4(17 %)
2 %(9)	8(35 %)	8(35 %)	7(30 %)	6(26 %)	8(35%)	7(30 %)	8(35 %)
3 %(14)	11(48 %)	12(52 %)	11(48 %)	10(43 %)	12(52 %)	10(43 %)	11(48 %)
4 %(18)	15(65 %)	15(65 %)	13(57 %)	12(52 %)	14(61 %)	12(52 %)	15(65 %)
5 %(23)	16(70 %)	18(78 %)	18(78 %)	15(65 %)	17(74 %)	16(70 %)	18(78 %)
6 %(27)	18(78 %)	20(87 %)	20(87 %)	18(78 %)	19(83 %)	19(83 %)	20(87 %)
7 %(32)	23(100 %)	22(96 %)	21(91 %)	23(100 %)	22(96 %)	23(100 %)	23(100 %)
7.2 %(33)	23(100 %)	23(100 %)	21(91 %)	23(100 %)	23(100 %)	23(100 %)	23(100 %)
10 %(45)	23(100 %)	23(100 %)	22(96 %)	23(100 %)	23(100 %)	23(100 %)	23(100 %)
12 %(54)	23(100 %)	23(100 %)	23(100 %)	23(100 %)	23(100 %)	23(100 %)	23(100 %)
Accuracy	84.9 %	83.7 %	67.7 %	84.9 %	83.7 %	84.9 %	84.9 %

(2) $W_2 = \{x \in U_W : f(x, \text{Marginal\_adhesion}) = 2\}$;
(3) $W_3 = \{x \in U_W : f(x, \text{Uniformity\_cell\_shape}) = 6 \vee f(x, \text{Single\_cell\_size}) = 2\}$;
(4) $W_4 = \{x \in U_W : f(x, \text{Mitoses}) = 1\}$.

For each $1 \leq j \leq 4$, let R_{W_j} be the set of all outliers in W_j. For the parameter d in the GrC-based method and the parameter d' in the distance-based method, we set $d = |A_W|/2$ and $d' = |A_W|/2$. The results of various outlier detection methods on the Breast Cancer data set are summarized in Tables 7, 8, 9, and 10.

From Tables 7–10, it can be seen that for the Breast Cancer data set, ODGrCR also has the best performance. For each of W_1, ..., W_4, the accuracy of ODGrCR is higher than or equal to those of other methods. For instance, in Table 9, when the top ratio (number of objects) is set to 2.2 %(8), the number of outliers detected by ODGrCR is 7, but other methods can only find 6 outliers. From another point of view, to find all outliers in W_3, ODGrCR needs to check 2.2 % of objects in W_3, but for BD, RMF, RB, DIS, GrC and KNN, they need to check 2.5 %, 2.8 %, 6.2 %, 2.5 %, 2.8 % and 2.5 % of objects, respectively.

Table 11 Statistical information for the accuracies of various methods on Breast Cancer

Outlier detection methods	Mean	Standard deviation
BD	91 %	5.6 %
RMF	83.7 %	4.3 %
RB	70 %	5.2 %
DIS	86.5 %	3.2 %
GrC	81.9 %	2.9 %
KNN	91 %	5.6 %
ODGrCR	92.2 %	5.5 %

Table 11 gives the statistical information for the accuracies of various outlier detection methods on W_1, ..., W_4.

From Table 11, it can be seen that the average accuracy of ODGrCR is higher than those of BD and KNN, and the standard deviation of the accuracies of ODGrCR is lower than those of the two methods. Although the standard deviation of the accuracies of ODGrCR is higher than those of RMF, RB, DIS and GrC, the average accuracy of ODGrCR is markedly higher than those of the four methods.

5.4 Yeast data set

The Yeast data set contains 1484 objects with eight continuous attributes [5]. All objects in the Yeast data set are partitioned into 10 classes. In the experiments, only the ERL class and the first three classes (i.e., CYT, NUC and MIT) were selected, where the ERL class is regarded as a rare class. The resultant data set contains 1136 normal objects and 5 outliers. Moreover, the 8 continuous attributes in the data set were transformed into discrete attributes by using the MDL-based discretization method proposed by Fayyad and Irani [12].

Data in the Yeast data set was stored in an information table $IS_Y = (U_Y, A_Y, V_Y, f)$. We detected outliers with respect to four subsets Y_1, ..., Y_4 of U_Y in IS_Y. The four subsets Y_1, ..., Y_4 are respectively described as follows.

(1) $Y_1 = \{x \in U_Y : f(x, \text{gvh}) = 1\}$;
(2) $Y_2 = \{x \in U_Y : f(x, \text{nuc}) = 0\}$;
(3) $Y_3 = \{x \in U_Y : f(x, \text{mit}) = 0 \vee f(x, \text{mit}) = 1\}$;

Table 12 Experimental results with respect to Y_1 in IS_Y

Top ratio (number of objects)	$Y_1: \lvert Y_1\rvert = 402, \lvert R_{Y_1}\rvert = 5$						
	Number of rare classes included (coverage)						
	BD	RMF	RB	DIS	GrC	KNN	ODGrCR
0.8 %(3)	3(60 %)	3(60 %)	3(60 %)	3(60 %)	2(40 %)	2(40 %)	3(60 %)
1 %(4)	3(60 %)	3(60 %)	3(60 %)	3(60 %)	3(60 %)	2(40 %)	3(60 %)
1.5 %(6)	4(80 %)	3(60 %)	3(60 %)	3(60 %)	4(80 %)	3(60 %)	4(80 %)
1.7 %(7)	5(100 %)	3(60 %)	3(60 %)	3(60 %)	4(80 %)	4(80 %)	5(100 %)
2 %(8)	5(100 %)	3(60 %)	4(80 %)	4(80 %)	5(100 %)	4(80 %)	5(100 %)
2.2 %(9)	5(100 %)	3(60 %)	5(100 %)	5(100 %)	5(100 %)	4(80 %)	5(100 %)
2.5 %(10)	5(100 %)	4(80 %)	5(100 %)	5(100 %)	5(100 %)	4(80 %)	5(100 %)
2.7 %(11)	5(100 %)	5(100 %)	5(100 %)	5(100 %)	5(100 %)	5(100 %)	5(100 %)
Accuracy	85.5 %	72 %	77.2 %	77.2 %	80.9 %	72 %	85.5 %

Table 13 Experimental results with respect to Y_2 in IS_Y

Top ratio (number of objects)	$Y_2: \lvert Y_2\rvert = 774, \lvert R_{Y_2}\rvert = 4$						
	Number of rare classes included (coverage)						
	BD	RMF	RB	DIS	GrC	KNN	ODGrCR
0.4 %(3)	2(50 %)	2(50 %)	2(50 %)	2(50 %)	3(75 %)	0(0 %)	3(75 %)
0.5 %(4)	3(75 %)	3(75 %)	3(75 %)	3(75 %)	3(75 %)	0(0 %)	4(100 %)
0.6 %(5)	4(100 %)	4(100 %)	4(100 %)	4(100 %)	3(75 %)	1(25 %)	4(100 %)
0.8 %(6)	4(100 %)	4(100 %)	4(100 %)	4(100 %)	3(75 %)	2(50 %)	4(100 %)
0.9 %(7)	4(100 %)	4(100 %)	4(100 %)	4(100 %)	4(100 %)	2(50 %)	4(100 %)
1.4 %(11)	4(100 %)	4(100 %)	4(100 %)	4(100 %)	4(100 %)	3(75 %)	4(100 %)
2.4 %(19)	4(100 %)	4(100 %)	4(100 %)	4(100 %)	4(100 %)	4(100 %)	4(100 %)
Accuracy	89.9 %	89.9 %	89.9 %	89.9 %	78.4 %	59.6 %	100 %

Table 14 Experimental results with respect to Y_3 in IS_Y

Top ratio (number of objects)	$Y_3: \lvert Y_3\rvert = 948, \lvert R_{Y_3}\rvert = 5$						
	Number of rare classes included (coverage)						
	BD	RMF	RB	DIS	GrC	KNN	ODGrCR
0.3 %(3)	3(60 %)	3(60 %)	3(60 %)	3(60 %)	2(40 %)	2(40 %)	3(60 %)
0.5 %(5)	3(60 %)	3(60 %)	4(80 %)	3(60 %)	3(60 %)	3(60 %)	4(80 %)
0.6 %(6)	3(60 %)	3(60 %)	5(100 %)	3(60 %)	4(80 %)	4(80 %)	5(100 %)
0.7 %(7)	4(80 %)	4(80 %)	5(100 %)	4(80 %)	4(80 %)	4(80 %)	5(100 %)
0.8 %(8)	5(100 %)	5(100 %)	5(100 %)	5(100 %)	4(80 %)	5(100 %)	5(100 %)
1 %(9)	5(100 %)	5(100 %)	5(100 %)	5(100 %)	5(100 %)	5(100 %)	5(100 %)
Accuracy	81.1 %	81.1 %	91.6 %	81.1 %	77.6 %	81.1 %	91.6 %

Table 15 Experimental results with respect to Y_4 in IS_Y

Top ratio (number of objects)	Y_4 : $\|Y_4\| = 1134$, $\|R_{Y_4}\| = 5$						
	Number of rare classes included (coverage)						
	BD	RMF	RB	DIS	GrC	KNN	ODGrCR
0.3 %(3)	3(60 %)	3(60 %)	3(60 %)	3(60 %)	3(60 %)	3(60 %)	3(60 %)
0.4 %(5)	3(60 %)	5(100 %)	3(60 %)	3(60 %)	5(100 %)	5(100 %)	5(100 %)
0.5 %(6)	4(80 %)	5(100 %)	3(60 %)	4(80 %)	5(100 %)	5(100 %)	5(100 %)
0.6 %(7)	5(100 %)	5(100 %)	3(60 %)	5(100 %)	5(100 %)	5(100 %)	5(100 %)
0.8 %(9)	5(100 %)	5(100 %)	3(60 %)	5(100 %)	5(100 %)	5(100 %)	5(100 %)
10.4 %(118)	5(100 %)	5(100 %)	5(100 %)	5(100 %)	5(100 %)	5(100 %)	5(100 %)
Accuracy	85.6 %	100 %	47.1 %	85.6 %	100 %	100 %	100 %

(4) $Y_4 = \{x \in U_Y : f(x, \text{mcg}) = 2 \ \vee \ f(x, \text{erl}) = 0\}$.

For each $1 \leq j \leq 4$, let R_{Y_j} be the set of all outliers in Y_j. For the parameter d in the GrC-based method and the parameter d' in the distance-based method, we set $d = |A_Y|/3$ and $d' = |A_Y|/3$. The results of various outlier detection methods on the Yeast data set are summarized in Tables 12, 13, 14 and15.

From Tables 12–15, it can be seen that for the Yeast data set, ODGrCR also has the best performance. For each of $Y_1, ..., Y_4$, the accuracy of ODGrCR is higher than or equal to those of other methods. For instance, in Table 13, when the top ratio (number of objects) is set to 0.5 %(4), the number of outliers detected by ODGrCR is 4, but other methods can only find 3 or fewer outliers. From another point of view, to find all outliers in Y_2, ODGrCR needs to check 0.5 % of objects in Y_2, but for BD, RMF, RB, DIS, GrC and KNN, they need to check 0.6 %, 0.6 %, 0.6 %, 0.6 %, 0.9 % and 2.4 % of objects, respectively.

Table 16 gives the statistical information for the accuracies of various outlier detection methods on $Y_1, ..., Y_4$.

From Table 16, it can be seen that the average accuracy of ODGrCR is higher than those of RMF, RB, GrC and KNN, and the standard deviation of the accuracies of ODGrCR is lower than those of the four methods. Although the standard deviation of the accuracies of ODGrCR is higher than those of BD and DIS, the average accuracy of ODGrCR is markedly higher than those of the two methods.

Table 16 Statistical information for the accuracies of various methods on Yeast

Outlier detection methods	Mean	Standard deviation
BD	85.5 %	3.1 %
RMF	85.8 %	10.4 %
RB	76.5 %	17.8 %
DIS	83.5 %	4.8 %
GrC	84.2 %	9.2 %
KNN	78.2 %	14.7 %
ODGrCR	94.3 %	6.1 %

5.5 KDD cup 1999 data set

The KDD Cup 1999 data set includes 41 attributes and a class label which specifies the status of each connection record as either normal or specific attack type [5]. Because the KDD Cup 1999 data set is too large for our purpose, we generated a concise subset of it by using the *random sampling* (*without replacement*) technology. The subset is called KDD-Subset, which contains 20000 normal objects and 1000 attack objects. Moreover, the continuous attributes in KDD-Subset were also transformed into discrete attributes by using the MDL-based discretization method [12].

Data in KDD-Subset was stored in an information table $IS_K = (U_K, A_K, V_K, f)$. We detected outliers with respect to four subsets $K_1, ..., K_4$ of U_K in IS_K. The four subsets $K_1, ..., K_4$ are respectively described as follows.

(1) $K_1 = \{x \in U_K : f(x, \text{srv\_count}) = 2\}$;
(2) $K_2 = \{x \in U_K : f(x, \text{count}) = 1\}$;
(3) $K_3 = \{x \in U_K : f(x, \text{protocol\_type}) = 0 \ \wedge \ f(x, \text{duration}) = 0\}$;
(4) $K_4 = \{x \in U_K : f(x, \text{dst\_host\_count}) = 1 \ \vee \ f(x, \text{dst\_host\_count}) = 2\}$.

For each $1 \leq j \leq 4$, let R_{K_j} be the set of all outliers in K_j. For the parameter d in the GrC-based method and the parameter d' in the distance-based method, we set $d = |A_K|/3$ and $d' = |A_K|/3$. The results of various methods on KDD-Subset are summarized in Tables 17, 18, 19 and 20.

Table 17 Experimental results with respect to K_1 in IS_K

Top ratio (number of objects)	K_1 : $\|K_1\| = 4641$, $\|R_{K_1}\| = 458$						
	Number of rare classes included						
	BD	RMF	RB	DIS	GrC	KNN	ODGrCR
1 %(46)	0	46	46	46	46	44	46
2 %(93)	37	93	93	93	93	91	93
5 %(232)	176	232	232	232	232	230	232
9.93 %(461)	401	458	457	457	453	401	458
10.02 %(465)	405	458	458	457	453	405	458
10.4 %(483)	423	458	458	458	454	423	458
11.16 %(518)	458	458	458	458	455	446	458
12.54 %(582)	458	458	458	458	458	446	458
79.8 %(3703)	458	458	458	458	458	458	458
Accuracy	93.5 %	99.6 %	99.2 %	97.1 %	87.9 %	17.4 %	99.6 %

From Tables 17–20, it can be seen that for KDD-Subset, ODGrCR also has the best performance. For each of K_1, ..., K_4, the accuracy of ODGrCR is higher than or equal to those of other methods.

Table 21 gives the statistical information for the accuracies of various outlier detection methods on K_1, ..., K_4.

From Table 21, it can be seen that the average accuracy of ODGrCR is higher than those of RMF, RB, and DIS, and the standard deviation of the accuracies of ODGrCR is lower than those of the three methods. Although the standard deviation of the accuracies of ODGrCR is higher than those of BD, GrC and KNN, the average accuracy of ODGrCR is markedly higher than those of the three methods.

5.6 Average running time comparison

So far, we have compared the performance of ODGrCR with those of other outlier detection methods. It is still interesting to know the average running time of each outlier detection method.

Here, we only compared the running time of each outlier detection method over the KDD Cup 1999 data set,

Table 18 Experimental results with respect to K_2 in IS_K

Top ratio (number of objects)	K_2 : $\|K_2\| = 10995$, $\|R_{K_2}\| = 45$						
	Number of rare classes included (coverage)						
	BD	RMF	RB	DIS	GrC	KNN	ODGrCR
1 %(110)	26	36	36	35	40	27	36
2 %(220)	35	36	36	35	42	29	36
3 %(330)	39	40	41	44	42	29	41
3.32 %(365)	43	41	41	44	42	29	45
3.4 %(374)	45	41	41	44	42	29	45
3.73 %(410)	45	41	45	44	42	34	45
3.87 %(426)	45	45	45	44	42	34	45
8.69 %(955)	45	45	45	44	45	34	45
9.86 %(1084)	45	45	45	45	45	34	45
58.91 %(6477)	45	45	45	45	45	45	45
Accuracy	54.5 %	53.5 %	53.8 %	47.3 %	48.2 %	21 %	54.7 %

Table 19 Experimental results with respect to K_3 in IS_K

Top ratio (number of objects)	$K_3: \lvert K_3\rvert = 16818, \lvert R_{K_3}\rvert = 943$						
	Number of rare classes included (coverage)						
	BD	RMF	RB	DIS	GrC	KNN	ODGrCR
1 %(168)	168	166	168	168	168	165	168
5 %(841)	838	839	814	841	841	838	841
10 %(1682)	931	932	923	925	923	906	934
15 %(2523)	937	938	927	931	924	906	936
20 %(3364)	940	942	942	931	934	906	942
21.85 % (3674)	941	943	943	931	937	906	943
22.37 %(3762)	943	943	943	931	937	906	943
28.63 %(4815)	943	943	943	933	943	930	943
94.94 % (15967)	943	943	943	943	943	942	943
96.19 % (16178)	943	943	943	943	943	943	943
Accuracy	53.7 %	54.2 %	54.2 %	5.6 %	47.6 %	4.9 %	54.2 %

because other data sets were too small. Each method was run 100 times and the average running time was calculated. Figures 1, 2 and 3 show the average running time of each method over K_1, ..., K_4, where the numbers of objects in K_1, ..., K_4 are 4641, 10995, 16818 and 19494, respectively.

From Figs. 1–3, it is easy to discover that ODGrCR takes much less running time than other outlier detection methods, and the running time of RMF is the biggest.

6 Discussion

Generally speaking, the current outlier detection methods can be classified into the following five categories.

(1) *Distribution-based method.* Distribution-based method was proposed in statistics. Those objects which deviate from the given model are deemed as outliers [4]. The main problem of such method is that

Table 20 Experimental results with respect to K_4 in IS_K

Top ratio (number of objects)	$K_4: \lvert K_4\rvert = 19494, \lvert R_{K_4}\rvert = 981$						
	Number of rare classes included (coverage)						
	BD	RMF	RB	DIS	GrC	KNN	ODGrCR
1 %(195)	195	194	64	193	195	192	194
5 %(975)	728	941	411	940	935	745	948
10 %(1949)	889	965	956	964	946	922	972
15 %(2924)	938	976	970	969	964	922	976
20 %(3899)	938	977	977	969	971	922	979
33.35 %(6501)	941	980	980	976	980	942	981
35.13 %(6848)	941	981	980	977	980	942	981
38.08 %(7424)	941	981	981	977	980	942	981
46.47 %(9059)	981	981	981	977	981	952	981
76.7 %(14950)	981	981	981	978	981	981	981
93.48 %(18223)	981	981	981	981	981	981	981
Accuracy	33.6 %	41.3 %	39.2 %	6.1 %	33.6 %	15.6 %	42.6 %

Table 21 Statistical information for the accuracies of various methods on KDD-Subset

Outlier detection methods	Mean	Standard deviation
BD	58.8 %	21.7 %
RMF	62.2 %	22.2 %
RB	61.6 %	22.5 %
DIS	39.0 %	37.6 %
GrC	54.3 %	20.2 %
KNN	14.7 %	6 %
ODGrCR	62.8 %	21.8 %

the distribution of the measurement data is usually unknown in practice.

Our method does not require any knowledge about the distribution of the give data, which can solve the problem of distribution-based method. Moreover, the time complexity of ODGrCR is less than that of distribution-based method.

(2) *Depth-based method.* Depth-based method is based on the computational geometry, which computes different layers of *k-d* convex hulls and flags objects in the outer layer as outliers [24]. The depth-based method suffers from the curse of dimensionality and can not deal with large data sets. Compared with the depth-based method, the time complexity of ODGrCR is relatively low.

(3) *Clustering-based method.* Clustering technology is also used to detect outliers [20]. However, outliers are only the by-products of clustering, and the quality of outlier detection is greatly impacted by that of clustering results. Our method is designed especially for outlier detection, which is different from the clustering-based method.

(4) *Distance-based method.* Distance-based method was proposed by Knorr and Ng [26, 27], which uses the distance between any two objects as a measure of unusualness. The distance-based method has the following problems:

(I) To calculate the distance between any two objects, we must select an appropriate distance metric and set the value of distance parameter d. In practice, there are no clear advantages of one particular distance metric over another. Therefore, it is difficult to select appropriate distance metrics for many practical tasks. Moreover, it is also difficult to set the value of parameter d. It may involve too many trials to find suitable distance metrics and set the value of d.

(II) The distance-based method is not feasible for dealing with very large data sets as its time complexity is too high. For instance, the outlier detection algorithms based on nested loops typically require $O(n^2)$ distance computations (where n denotes the number of objects).

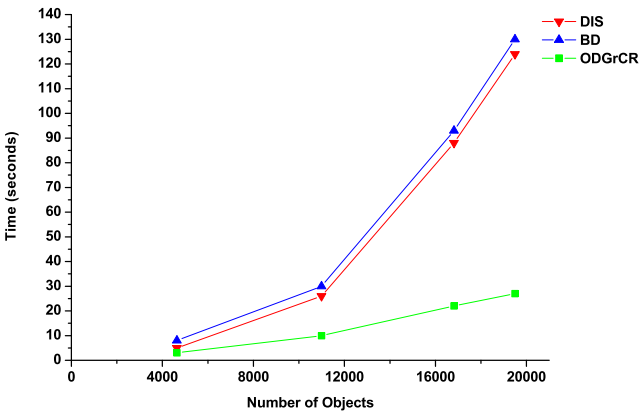

Fig. 1 Average running times of DIS, BD and ODGrCR over $K_1, ..., K_4$

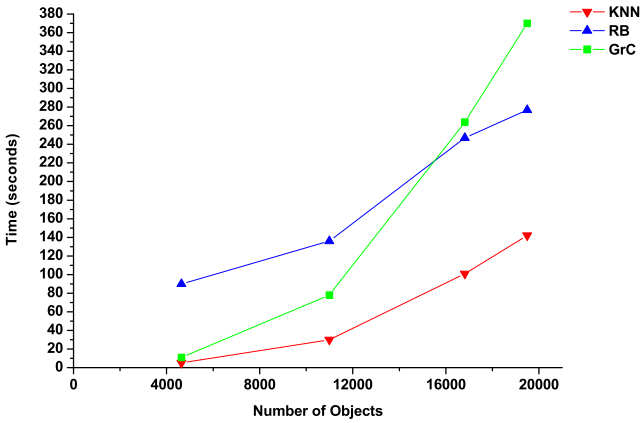

Fig. 2 Average running times of KNN, RB and GrC over $K_1, ..., K_4$

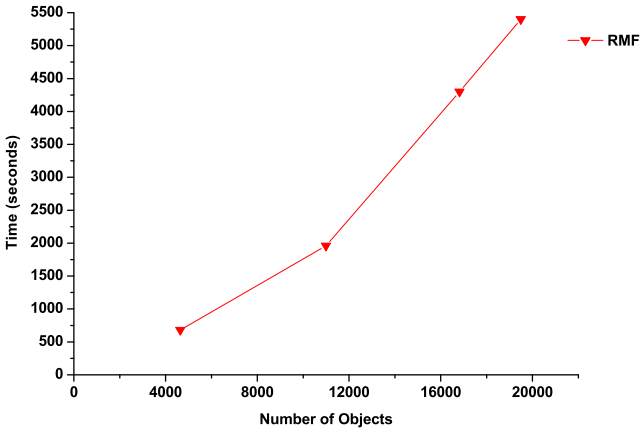

Fig. 3 Average running time of RMF over $K_1, ..., K_4$

ODGrCR algorithm can solve the above two problems, because in ODGrCR we need not to calculate the distance between any two objects. Moreover, compared with the distance-based method, the time complexity of ODGrCR is relatively low. From the experimental results in Section 5, it can be seen that the performance of ODGrCR is better than that of distance-based method, and the running time of ODGrCR is less than that of distance-based method.

(5) *Density-based method.* Density-based method was first proposed by Breunig et al. [7]. When using the density-based method to detect outliers, we need to compute the k-nearest neighbors of a given object p. Therefore, the time complexity of that method is $O(n^2)$ (where n denotes the number of objects). Moreover, the density-based method is very sensitive to the parameters defining the neighborhood. ODGrCR does not require such parameters, and the time complexity of ODGrCR is less than that of density-based method.

As mentioned earlier in Section 1, we have proposed four outlier detection methods in previous work, that is, the boundary-based method [21]; the RMF-based method [22]; the BD-based method [23]; and the GrC-based method [8]. Although these methods have demonstrated the effectiveness of rough set theory and GrC for outlier detection, some problems remain. For instance, when using the BD-based or GrC-based method to detect outliers, we must compute the distance between any two objects or two granules. Therefore, the two methods suffer from the same problems as the distance-based method, which have been discussed above. Moreover, the boundary-based and RMF-based methods are not feasible for dealing with large data sets as their time complexities are too high.

ODGrCR can solve the problems of the above four methods. In ODGrCR, we need not to calculate the distance between any two objects or two granules, which can avoid the problems of the BD-based and GrC-based methods. Moreover, from the experimental results in Section 5, it can be seen that ODGrCR takes much less running time than the boundary-based and RMF-based methods, and the performance of ODGrCR is better than those of the two methods.

In the following we will discuss the relation of our method with formal concept analysis (FCA) [13, 53]. Rough set theory and FCA provide two related methods for data analysis [38, 59]. The central notions of rough sets include the indiscernibility of objects with respect to a set of attributes and the induced approximation operators. The central notions of FCA include formal concepts and concept lattice. We can introduce the notion of concept lattice into rough sets. On the other hand, we can also introduce the notion of approximation operators into FCA [25]. For instance, based on the notion of approximation operators in rough set theory, various concept lattices have been proposed, e.g., the object oriented concept lattice, the property oriented concept lattice [59]. Since the upper and lower approximations are strongly linked to the notion of upper and lower bounds in FCA, ODGrCR algorithm may also be extended to FCA, that is, we may propose a corresponding outlier detection algorithm from the perspective of FCA.

7 Conclusions

In this paper, we have presented an outlier detection algorithm based on GrC and rough set theory. Given an information table $IS = (U, A, V, f)$, for any object $x \in U$ for which we wish to calculate the degree of outlierness, our algorithm first calculates the degree of outlierness of each granule containing x. If the degrees of outlierness of those granules containing x are always high, then the degree of outlierness of object x is also high. Moreover, we use the accuracy of approximation in rough sets to calculate the degree of outlierness of a granule. The experimental results have shown that our algorithm is effective for outlier detection.

ODGrCR was proposed based on Pawlak's classical rough set model [40, 41]. To deal with continuous attributes, the classical rough set model should replace all continuous attributes with discretized attributes by the process of discretization. However, discretization may cause the loss of information. In future work, we may extend our algorithm to the neighborhood rough set model proposed by Hu et al., which can deal with continuous attributes without discretization [19]. Moreover, we may apply ODGrCR to intrusion detection. Intrusion activities are treated as outliers, for which we shall design an unsupervised intrusion detection approach.

Acknowledgments This work is supported by the National Natural Science Foundation of China (grant nos. 60802042, 61273180), the Natural Science Foundation of Shandong Province, China (grant no. ZR2011FQ005), and the Project of Shandong Province Higher Educational Science and Technology Program (grant no. J11LG05).

References

1. Aggarwal CC, Yu PS (2001) Outlier detection for high dimensional data. In: Proceedings of the 2001 ACM SIGMOD International Conference on Managment of Data, California, pp 37–46
2. Albanese A, Pal SK, Petrosino A (2014) Rough Sets, Kernel Set, and Spatiotemporal Outlier Detection. IEEE Trans Knowl Data Eng 26(1):194–207
3. Angiulli F, Pizzuti C (2002) Fast outlier detection in high dimensional spaces. In: Proceedings of the Sixth European Conference

on the Principles of Data Mining and Knowledge Discovery, pp 15–26
4. Barnett V, Lewis T (1994) Outliers in Statistical Data. John Wiley & Sons, New York
5. Bay SD (1999) The UCI KDD repository. Available online at: http://kdd.ics.uci.edu
6. Bolton RJ, Hand DJ (2002) Statistical fraud detection: A review (with discussion). Statist Sci 17(3):235–255
7. Breunig MM, Kriegel H-P, Ng RT, Sander J (2000) LOF: Identifying density-based local outliers. In: Proceedings of the ACM SIGMOD international conference on management of data, Dallas, pp 93–104
8. Chen YM, Miao DQ, Wang RZ (2008) Outlier detection based on granular computing. In: Proceedings of the 6th international conference on rough sets and current trends in computing, Akron, pp 283–292
9. Chen YM, Miao DQ, Zhang HY (2010) Neighborhood outlier detection. Expert Syst Appl 37(12):8745–8749
10. Duan QG, Miao DQ, Wang RZ, Chen M (2007) An approach to web page classification based on granules. In: Proceedings of IEEE/WIC/ACM international conference on web intelligence, Silicon Valley, pp 279–282
11. Eskin E, Arnold A, Prerau M, Portnoy L, Stolfo S (2002) A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. In: Barbar D, et al. (eds) Data Mining for Security Applications. Kluwer Academic Publishers, Boston, pp 1–20
12. Fayyad UM, Irani KB (1993) Multi-interval discretization of continuous-valued attributes for classification learning. In: Proceedings of the 13th International Conference on Artificial Intelligence, pp 1022–1027
13. Ganter B, Wille R (1999) Formal Concept Analysis: mathematical foundations. Springer-Verlag, Berlin
14. Han JW, Kamber M, Pei J (2011) Data mining: concepts and techniques. Morgan Kaufmann Publishers, San Francisco
15. Harkins S, He HX, Willams GJ, Baxter RA (2002) Outlier detection using replicator neural networks. In: Proceedings of the 4th international conference on data warehousing and knowledge discovery, France, pp 170–180
16. Hawkins D (1980) Identifications of Outliers. Chapman and Hall, London
17. He ZY, Deng SC, Xu XF (2005) An optimization model for outlier detection in categorical data. In: Proceedings of the international conference on intelligent computing (ICIC(1)), Hefei, pp 400–409
18. Hu QH, Xie ZX, Yu DR (2007) Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation. Pattern Recogn 40(12):3509–3521
19. Hu QH, Yu DR, Liu JF, Wu CX (2008) Neighborhood rough set based heterogeneous feature subset selection. Inf Sci 178(18):3577–3594
20. Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv 31(3):264–323
21. Jiang F, Sui YF, Cao CG (2005) Outlier detection using rough set theory. In: Proceedings of the 10th international conference on rough sets, fuzzy sets, data mining, and granular computing (RSFDGrC (2)). LNAI 3642, Regina, pp 79–87
22. Jiang F, Sui YF, Cao CG (2008) A rough set approach to outlier detection. Int J Gen Syst 37(5):519–536
23. Jiang F, Sui YF, Cao CG (2011) A hybrid approach to outlier detection based on boundary region. Pattern Recogn Lett 32(14):1860–1870
24. Johnson T, Kwok I, Ng RT (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
25. Kent RE (1996) Rough concept analysis: a synthesis of rough sets and formal concept analysis. Fundamenta Informaticae 27(2):169–181
26. Knorr EM, Ng RT (1998) Algorithms for mining distance-based outliers in large datasets. In: Proceedings of the 24th VLDB Conference, New York, pp 392–403
27. Knorr EM, Ng RT, Tucakov V (2000) Distance-based outliers: algorithms and applications. VLDB Journal: Very Large Data bases 8(3-4):237–253
28. Lane T, Brodley CE (1999) Temporal sequence learning and data reduction for anomaly detection. ACM Trans. Inform Syst Security 2(3):295–331
29. Liang JY, Wang JH, Qian YH (2009) A new measure of uncertainty based on knowledge granulation for rough sets. Inf Sci 179(4):458–470
30. Lin TY (1997) Granular computing. Announcement of the BISC special interest group on granular computing
31. Lin TY (1998) Granular computing on binary relations I: data mining and neighborhood systems, II: rough set representations and belief functions. In: Skowron A, Polkowski L (eds) Rough sets in knowledge discovery. Physica-Verlag, Heidelberg, pp 107–140
32. Lin TY (2000) Data Mining and Machine Oriented Modeling: A Granular Computing Approach. Appl Intell 13(2):113–124
33. Lin TY, Louie E (2002) Finding association rules by granular computing: fast algorithms for finding association rules. In: Proceedings of the 12th international conference on data mining, rough sets and granular computing, Berlin, pp 23–42
34. Miao DQ, Wang GY, Liu Q, Lin TY, Yao YY (2007) Granular computing: past, present and future prospect. Science Press, Beijing
35. Miao DQ, Chen M, Wei ZH, Duan QG (2007) A reasonable rough approximation of clustering web users. In: Proceedings of the WICI international workshop on web intelligence meets brain informatics, LNCS 4845, pp 428–442
36. Nguyen SH, Nguyen HS (1996) Some efficient algorithms for rough set methods. In: IPMU'96, Granada, pp 1451–1456
37. Nguyen TT (2007) Outlier Detection: An Approximate Reasoning Approach. In: Proceedings of the International Conference on Rough Sets and Intelligent Systems Paradigms, pp 495–504
38. Pagliani P (1993) From concept lattices to approximation spaces: algebraic structures of some spaces of partial objects. Fundamenta Informaticae 18:1–25
39. Pal SK, Meher SK, Dutta S (2012) Class-dependent rough-fuzzy granular space, dispersion index and classification. Pattern Recogn 45(7):2690–2707
40. Pawlak Z (1982) Rough sets. Internat J Comput Inform Sci 11:341–356
41. Pawlak Z (1991) Rough Sets: Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht
42. Pawlak Z (1998) Granularity of knowledge, indiscernibility and rough sets. Proceedings of IEEE international conference on fuzzy systems, Anchorage, pp 106–110
43. Pedrycz W, Vukovich G (2001) Granular neural networks. Neurocomputing 36(1-4):205–224
44. Pedrycz W, Vukovich G (2002) Feature analysis through information granulation and fuzzy sets. Pattern Recognit 35(4):825–834
45. Polkowski L, Skowron A (1998) Towards adaptive calculus of granules. In: Proceedings of IEEE international conference on fuzzy systems, Anchorage, pp 111–116
46. Qian YH, Liang JY, Dang CY (2010) Incomplete multigranulation rough set. IEEE Trans. Syst. Man Cybern. Part A 40(2):420–431

47. Qian YH, Liang JY, Pedrycz W, Dang CY (2010) Positive approximation: An accelerator for attribute reduction in rough set theory. Artif Intell 174(9-10):597–618
48. Ramaswamy S, Rastogi R, Shim K (2000) Efficient algorithms for mining outliers from large datasets. In: Proceedings of the ACM SIGMOD conference on management of data, Dallas, pp 427–438
49. Shaari F, Bakar AA, Hamdan AR (2009) Outlier detection based on rough sets theory. Intell Data Anal 13(2):191–206
50. Skowron A, Stepaniuk J (1999) Towards discovery of information granules. In: Proceedings of the 3rd European conference on principles and practice of knowledge discovery in databases, LNAI 1704. Springer-Verlag, Berlin Heidelberg New York, pp 542–547
51. Wang GY (2001) Rough set theory and knowledge acquisition. Xian Jiaotong University Press, Xian
52. Wang CZ, Chen DG, Wu C, Hu QH (2011) Data compression with homomorphism in covering information systems. Internat J Approx Reason 52(4):519–525
53. Wille R (1982) Restructuring Lattice theory: An Approach Based on Hierarchies of Concepts. Ordered Sets, Reidtel, D, Dordrecht, pp 445–470
54. Wu WZ, Leung Y, Mi JS (2009) Granular computing and knowledge reduction in formal contexts. IEEE Trans Knowl Data Eng 21(10):1461–1474
55. Xu ZY, Liu ZP, Yang BR, Song W (2006) A quick attribute reduction algorithm with complexity of max($O(|C||U|)$, $O(|C|^2|U/C|)$). Chin J Comput 29(3):391–399
56. Xue ZX, Liu SY (2009) Rough-Based Semi-supervised Outlier Detection. In: Proceedings of the 6th internatonal conference on fuzzy systems and knowledge discovery, vol 1, pp 520–523
57. Yao YY (1999) Granular computing using neighborhood systems. In: Roy R, Furuhashi T, Chawdhry PK (eds) Advances in Soft Computing: Engineering Design and Manufacturing. Springer-Verlag, London, pp 539–553
58. Yao YY, Zhong N (2002) Granular computing using information tables. In: Lin TY, Yao YY, Zadeh LA (eds) Data Mining, Rough Sets and Granular Computing. Physica-Verlag, Berlin Heidelberg New York, pp 102–124
59. Yao YY (2004) A Comparative Study of Formal Concept Analysis and Rough Set Theory in Data Analysis. In: Proceedings of the 4th international conference on rough sets and current trends in computing, LNAI 3066. Springer, Berlin Heidelberg New York, pp 59–68
60. Yao YY (2006) Granular computing for data mining. In: Dasarathy B. V (ed) Proceedings of SPIE conference on data mining, intrusion detection, information assurance, and data networks security, pp 1–12
61. Ye MQ, Wu XD, Hu XG, Hu DH (2013) Multi-level rough set reduction for decision rule mining. Appl Intell 39(3):642–658
62. Zadeh LA (1979) Fuzzy sets and information granularity. In: Gupta N, Ragade R, Yager R (eds) Advances in fuzzy set theory and applications, North-Holland, pp 3–18
63. Zadeh LA (1997) Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets Syst 90(2):111–127
64. Zadeh LA (1998) Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information/intelligent systems. Soft Comput 2(1):23–25
65. Zhang B, Zhang L (1992) Theory and Applications of Problem Solving. Elsevier Science Publishers B V, North-Holland

Feng Jiang received the Ph.D. degree in computer software from the Institute of Computing Technology, Chinese Academy of Sciences, China, in 2007. He is currently an associate professor of computer science in the College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao, China. He has been a program committee member of the IEEE International Conference on Granular Computing. His research interests include data mining, rough set theory and granular computing.

Yu-Ming Chen received the Ph.D. degree in Pattern Recognition and Intelligent System from Tongji University, China, in 2010. He is currently an associate professor of computer science in the Department of Computer Science, Xiamen University of Technology, Xiamen, China. His research interests include data mining, rough set theory and granular computing.