Outlier Detection Based on Granular Computing

Yuming Chen, Duoqian Miao, and Ruizhi Wang

Department of Computer Science and Technology,
The Key Laboratory of Embedded System and Service Computing, Tongji University
Shanghai, 201804, P.R. China
cym0620@163.com

Abstract. As an emerging conceptual and computing paradigm of information processing, granular computing has received much attention recently. Many models and methods of granular computing have been proposed and studied. Among them was the granular computing model using information tables. In this paper, we shall demonstrate the application of this granular computing model for the study of a specific data mining problem - outlier detection. Within the granular computing model using information tables, this paper proposes a novel definition of outliers - GrC (granular computing)-based outliers. An algorithm to find such outliers is also given. And the effectiveness of GrC-based method for outlier detection is demonstrated on three publicly available databases.

Keywords: Granular computing, outlier detection, rough sets, data mining.

1 Introduction

L. A. Zadeh introduced the concept of granular computing in 1979 under the name of information granularity [2]. And the term "granular computing" came to life with a suggestion from T. Y. Lin in the discussion of BISC Special Interest Group on Granular Computing [3]. Basic ingredients of granular computing are granules such as subsets, classes, and clusters of a universe. Furthermore, Andrzej Skowron, et al. introduced the discovery of information granules and information granules in distributed environment [4-5]. D. Q. Miao, et al. proposed an approach to web mining based on granular computing [6-9]. Specially, Y. Y. Yao and N. Zhong proposed a granular computing model using information tables [1, 10]. 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. Although the model is simple, it is powerful for the study of fundamental issues in granular computing, and has many potential applications in data mining.

Data mining is an important issue in the development of data- and knowledge-base systems. Usually, the tasks of data mining can be classified into four general

C.-C. Chan et al. (Eds.): RSCTC 2008, LNAI 5306, pp. 283–292, 2008.

categories: (a) dependency detection, (b) class identification, (c) class description, and (d) outlier/exception detection [11]. In contrast to most tasks of data mining, 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, from a data mining standpoint.

Outliers exist extensively in the real world. While there is no single, generally accepted, formal definition of an outlier, Hawkins' definition captures the spirit: an outlier is an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism [11-12].

With increasing awareness on outlier detection in literatures, more concrete meanings of outliers are defined for solving problems in specific domains. But to our best knowledge, there are few works about outlier detection in granular computing community [14]. In this paper, we aim to exploit the granular computing model using information tables proposed by Yao for outlier detection. The basic idea is as follows. Given an information table $S = (U, A, V, f)$, where U is a non-empty finite set of objects, A a set of attributes, V the union of attribute domains, and $f : U \times A \to V$ a function such that for any $x \in U$ and $a \in A$, $f(x, a) \in V_a$. In S, each attribute subset $B \subseteq A$ determines an indiscernibility relation $IND(B)$ on U. $IND(B)$ induces a partition of U, which is denoted by $U/IND(B)$, where each element from $U/IND(B)$ is a granule (equivalence class), and the element containing $x \in U$ is called the granule containing x under relation $IND(B)$. For a given object $x \in U$ and a set of indiscernibility relations (available information/knowledge) on U, we can obtain a granule containing x under each of these indiscernibility relations. Then through calculating the degree of outlierness for each of these granules containing x, we can decide whether object x behaves normally according to the given knowledge at hand. That is, if the degrees of outlierness of the granules containing x under these indiscernibility relations are always very high, then we may consider object x as a GrC (granular computing)-based outlier in U wrt S. A GrC-based outlier in U wrt S is an element such that the granules containing it always have a high degree of outlierness in view of the given knowledge.

The paper is organized as follows. In the next section, we introduce some preliminaries that are relevant to this paper. In section 3, based on the granular computing model using information tables, we give the definition of GrC-based outliers. An algorithm to find GrC-based outliers is also given. In section 4 we give the experimental results. Section 5 concludes the paper.

2 Preliminaries

An information table is a quadruple $S = (U, A, V, f)$, 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 $x \in U$, $f(x, a) \in V_a$.

In an information table $S = (U, A, V, f)$, each subset $B \subseteq A$ of attributes determines a binary relation $IND(B)$, called indiscernibility relation, defined as $IND(B) = \{(x, y) \in U \times U : \forall a \in B(f(x, a) = f(y, a))\}$.

For any two objects $u_1, u_2 \in U$, if $(u_1, u_2) \in IND(B)$ then one cannot differentiate u_1 from u_2 based solely on their values on attributes of B. We say that u_1 and u_2 are indistinguishable. Since each indiscernibility class may be viewed as a granule consisting of indistinguishable elements, u_1 and u_2 may be put into the same granule.

Given any $B \subseteq A$, relation $IND(B)$ induces a partition of U, which is denoted by $U/IND(B)$, where an element from $U/IND(B)$ is called an equivalence class or elementary set. Each equivalence class of relation $IND(B)$ is a granule. The equivalence class of $IND(B)$ that contains object $x \in U$, written $[x]_B$, is defined by collecting all objects whose value on each attribute $a \in B$ is the same as x's value:

$$[x]_B = \{y \in U : \ \forall a \in B(f(y, a) = f(x, a))\} \tag{1}$$

For every object $x \in U$, $[x]_B$ is called the granule containing x under relation $IND(B)$. When B is a singleton subset of A, the elements in $U/IND(B)$ are called *elementary granules*, as they are the smallest granules derivable. From the elementary granules, large granules may be built by taking a union of a family of elementary granules. One can build a hierarchy of granules [1].

3 GrC-Based Outlier Detection

3.1 Definitions

Given an information table S, we first define a granular outlier factor (GOF), which can indicate the degree of outlierness for every granule in the granular computing model using S [13]. Then we define an object outlier factor (OOF) by virtue of GOF, which can indicate the degree of outlierness for every object.

Definition 1 (Distance Between Granules). *Let $S = (U, A, V, f)$ be an information table. Given any $B \subseteq A$, relation $IND(B)$ induces a partition of U, which is denoted by $G = U/IND(B)$, where each equivalence class of relation $IND(B)$ is a granule. For any two granules $g_1, g_2 \in G$, the distance between granules g_1 and g_2 in S is defined as follows:*

$$M(g_1, g_2) = \frac{\sum_{p \in g_1, q \in g_2} \delta(p, q)}{|g_1| \times |g_2|} \tag{2}$$

where $M : G \times G \to [0, \infty]$ is a distance function such that for any $g_1, g_2 \in G$, $M(g_1, g_2)$ denotes the distance between sets g_1 and g_2. And δ is a given distance metric on U for nominal attributes.

In the above definition, to calculate the distance between any two granules, we consider the average distance between the objects in the analyzed two granules, which is adopted in the average linkage algorithm of hierarchical clustering [15].

Definition 2 (Granular Outlier Factor). *Let $S = (U, A, V, f)$ be an information table. Given any $B \subseteq A$, relation $IND(B)$ induces a partition of U, which is denoted by $G = U/IND(B)$, where each equivalence class of relation $IND(B)$ is a granule. For any granule $g \in G$, the granular outlier factor of g in S is defined as follows:*

$$GOF(g) = \frac{|\{g' \in G : M(g, g') > d\}|}{|G|} \tag{3}$$

where $M(g, g')$ denotes the distance between granules g_1 and g_2, d is a given parameter, and $|K|$ denotes the cardinality of set K.

Definition 3 (Object Outlier Factor). *Let $S = (U, A, V, f)$ be an information table. For any $x \in U$, the object outlier factor of x in S is defined as*

$$OOF(x) = \frac{\sum_{a \in A} (GOF([x]_{\{a\}}) \times W_{\{a\}}(x))}{|A|} \tag{4}$$

where for every singleton subset $\{a\}$ of A, $W_{\{a\}} : U \to (0, 1]$ is a weight function such that for any $x \in U$, $W_{\{a\}}(x) = 1 - \frac{|[x]_{\{a\}}|}{|U|}$. $[x]_{\{a\}} = \{u \in U : f(u, a) = f(x, a)\}$ denotes the indiscernibility class of relation $IND(\{a\})$ that contains element x, i.e. the elementary granule containing x under relation $IND(\{a\})$. $GOF([x]_{\{a\}})$ denotes the granular outlier factor of granule $[x]_{\{a\}}$ and $|K|$ denotes the cardinality of set K.

In the above definition, we can see that in the granular computing model using information table S, only those elementary granules are used to calculate the object outlier factor. We do not consider other granules in the model.

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

Definition 4 (Granular Computing-based Outliers). *Let $S = (U, A, V, f)$ be an information table. Let μ be a given threshold value, for any $x \in U$, if $OOF(x) > \mu$ then x is called a granular computing (GrC)-based outlier in S, where $OOF(x)$ is the object outlier factor of x in S.*

3.2 Algorithm

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

Algorithm 3.1

Input: information table $S = (U, A, V, f)$, where $|U| = n$ and $|A| = m$; threshold value μ, d
Output: a set O of GrC-based outliers

(1) For any two objects $u_1, u_2 \in U$, calculate the distance between them under a given distance metric on U, that is, $\delta(u_1, u_2)$;
(2) For every $a \in A$
(3) {
(4) Sort all objects from U according to a given order (e.g. the lexicographical order) on domain V_a of attribute a [16];
(5) Determine the partition $U/IND(\{a\})$;
(6) For any $g_1, g_2 \in U/IND(\{a\})$, calculate the distance $M(g_1, g_2) = \frac{\sum_{p \in g_1, q \in g_2} \delta(p,q)}{|g_1| \times |g_2|}$
(7) }
(8) For every $x \in U$
(9) {
(10) For every $a \in A$
(11) {
(12) Calculate the granular outlier factor of $[x]_{\{a\}}$ in S, i.e. $GOF([x]_{\{a\}}) = \frac{|\{g' \in U/IND(\{a\}):\ M([x]_{\{a\}}, g') > d\}|}{|U/IND(\{a\})|}$;
(13) Assign weight $W_{\{a\}}(x) = 1 - \frac{|[x]_{\{a\}}|}{|U|}$ to x
(14) }
(15) Calculate $OOF(x)$, the object outlier factor of object x in S;
(16) If $OOF(x) > \mu$ then $O = O \cup \{x\}$
(17) }
(18) Return O.

4 Experimental Results

4.1 Experiment Design

In this section, following the experimental setup in [17], we use three real life data sets (*lymphography*, *annealing* and *cancer*) to demonstrate the performance of our algorithm against traditional distance-based method [11], FindCBLOF algorithm [18] and KNN algorithm [19]. In addition, on the *cancer* data set, we

add the results of RNN-based outlier detection method for comparison, these results can be found in the work of Harkins et al. [20, 21].

For algorithm 3.1, in order to calculate the distance between any two granules, we should first calculate the distances between objects contained in these granules under a given distance metric on U. In our experiment, we adopt the *overlap metric in rough set theory*, which is defined as follows:

Definition 5. *Given an information table $S = (U, A, V, f)$, let $x, y \in U$ be any two objects between which we shall calculate the distance. The overlap metric in rough set theory is defined as*

$$\Delta(x, y) = |\{a \in A : a(x) \neq a(y)\}| \tag{5}$$

where $\Delta : U \times U \to [0, \infty]$ is a function from $U \times U$ to the non-negative real number, and $|M|$ denotes the cardinality of set M.

And in algorithm 3.1, in order to calculate the granular outlier factor for a given granule, we should specify a value for parameter d, we set $d = |A| /2$ in our experiment, where $|A|$ denotes the cardinality of attribute set A.

Furthermore, in our experiment, the two parameters needed by FindCBLOF algorithm are set to 90% and 5 separately as done in [18]. And for the KNN algorithm, the results were obtained by using the 4^{th} nearest neighbor [19].

4.2 Lymphography Data

The first is the Lymphography data set, which can be found in the UCI machine learning repository [22]. It contains 148 instances with 19 attributes (including the class attribute). The 148 instances are partitioned into 4 classes: "normal find" (1.35%), "metastases" (54.73%), "malign lymph" (41.22%) and "fibrosis" (2.7%). Classes 1 and 4 are regarded as rare classes.

Aggarwal et. al. proposed a practicable way to test the effectiveness of an outlier detection method [17, 23]. That is, we can run the outlier detection method on a given data set and test the percentage of points 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 [23]). Points belonged to the rare class are considered as outliers. If the method works well, we expect that such abnormal classes would be over-represented in the set of points found.

The experimental results are summarized in table 1.

Table 1. Experimental Results in Lymphography Data Set

Top Ratio (Number of Objects)	Number of Rare Class Included (Coverage)			
	GrC	DIS	FindCBLOF	KNN
5%(7)	6(100%)	5(83%)	4(67%)	5(83%)
6%(9)	6(100%)	6(100%)	4(67%)	5(83%)
8%(12)	6(100%)	6(100%)	4(67%)	6(100%)
20%(30)	6(100%)	6(100%)	6(100%)	6(100%)

In table 1, "GrC", "DIS", "FindCBLOF", "KNN" denote GrC-based, traditional distance-based, FindCBLOF and KNN-based outlier detection methods, respectively. For every objects in U, the degree of outlierness is calculated by using the four outlier detection methods, respectively. For each outlier detection method, the "Top Ratio (Number of Objects)" denotes the percentage (number) of the objects selected from U whose degrees of outlierness calculated by the method are higher than those of other objects in U. And if we use $X \subseteq U$ to contain all those objects selected from U, then the "Number of Rare Class Included" is the number of objects in X that belong to one of the rare classes. The "Coverage" is the ratio of the "Number of Rare Class Included" to the number of objects in U that belong to one of the rare classes [17].

From table 1, we can see that for the lymphography data set, GrC-based method performs best, since it can find all outliers in U when the *Top Ratio* reaches 5%. The next one is distance-based method, which can find all outliers in U when the *Top Ratio* reaches 6%. And the worst is FindCBLOF method, since it can not achieve that goal until the *Top Ratio* reaches 20%.

Furthermore, for the lymphography data set, the *false alarm rates* (i.e., the percentage of objects in set X that are actually not outliers, where X is the set of the top-n objects with highest degrees of outlierness calculated by the given method, n is the number of outliers in U) of GrC-based, distance-based, FindCBLOF and KNN-based method are 17%, 17%, 33% and 33%, respectively.

4.3 Annealing Data

The Annealing data set is found in the UCI machine learning repository [22]. The data set contains 798 instances with 38 attributes. The data set contains a total of 5 (non-empty) classes : class 1, 2, 3, 5 and U, where class 3 has 608 instances, and the remained classes have 190 instance. Classes 1, 2, 5 and U are regarded as rare classes since they are small in size. Since Annealing data set contains 6 continuous attributes, we respectively transform these continuous attributes into categorical attributes by using the automatic discretization functionality provided by the CBA software [24].

The experimental results are summarized in table 2.

Table 2. Experimental Results in Annealing Data Set

Top Ratio (Number of Objects)	Number of Rare Class Included (Coverage)			
	GrC	DIS	FindCBLOF	KNN
10%(80)	75(39%)	73(38%)	45(24%)	21(11%)
15%(105)	96(51%)	92(48%)	55(29%)	30(16%)
20%(140)	128(67%)	121(64%)	82(43%)	41(22%)
25%(175)	161(85%)	153(81%)	105(55%)	58(31%)
30%(209)	190(100%)	178(94%)	105(55%)	62(33%)

Table 2 is similar to table 1. From table 2, we can see that for the Annealing data set, GrC-based method performs the best among the four outlier detection methods. In fact, the performances of GrC-based and distance-based methods are very close, and they perform markedly better than the other two methods — FindCBLOF and KNN-based methods.

Furthermore, for the Annealing data set, the false alarm rates of GrC-based, distance-based, FindCBLOF and KNN-based method are 6%, 12%, 45% and 68%, respectively.

4.4 Wisconsin Breast Cancer Data

The Wisconsin breast cancer data set is found in the UCI machine learning repository [22]. The data set contains 699 instances with 9 continuous attributes. Here we follow the experimental technique of Harkins et al. by removing some of the *malignant* instances to form a very unbalanced distribution [17, 20-21]. The resultant data set had 39 (8%) *malignant* instances and 444 (92%) *benign* instances. Moreover, the 9 continuous attributes in the data set are transformed into categorical attributes, respectively [1] [17].

The experimental results are summarized in table 3.

Table 3. Experimental Results in Wisconsin Breast Cancer Data Set

Top Ratio (Number of Objects)	Number of Rare Class Included (Coverage)				
	GrC	DIS	FindCBLOF	RNN	KNN
1%(4)	4(10%)	4(10%)	4(10%)	3(8%)	4(10%)
2%(8)	7(18%)	5(13%)	7(18%)	6(15%)	7(18%)
4%(16)	14(36%)	11(28%)	14(36%)	11(28%)	13(33%)
6%(24)	21(54%)	18(46%)	21(54%)	18(46%)	20(51%)
8%(32)	28(72%)	24(62%)	27(69%)	25(64%)	27(69%)
10%(40)	32(82%)	29(74%)	32(82%)	30(77%)	32(82%)
12%(48)	37(95%)	36(92%)	35(90%)	35(90%)	38(97%)
14%(56)	39(100%)	39(100%)	38(97%)	36(92%)	39(100%)
16%(64)	39(100%)	39(100%)	39(100%)	36(92%)	39(100%)
18%(72)	39(100%)	39(100%)	39(100%)	38(97%)	39(100%)
20%(80)	39(100%)	39(100%)	39(100%)	38(97%)	39(100%)
28%(112)	39(100%)	39(100%)	39(100%)	39(100%)	39(100%)

From table 3, we can see that for the breast cancer data set, GrC-based method performs the best among the five outlier detection methods, except in the case when *Top Ratio* is 12%. In fact, the performances of GrC-based, FindCBLOF and KNN-based methods are very close, and they perform markedly better than the other two methods — RNN-based and distance-based methods.

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

Furthermore, for the Wisconsin breast cancer data set, the false alarm rates of GrC-based, distance-based, FindCBLOF, RNN-based and KNN-based method are 18%, 26%, 21%, 23% and 18%, respectively.

5 Conclusion

Finding outliers is an important task for many data mining applications. In this paper, we present a new method for outlier definition and outlier detection, which exploits the granular computing model using information tables proposed by Yao [1]. The main idea is that an object has more likelihood of being an outlier if the granules containing it have a high degree of outlierness. Experimental results on real data sets demonstrate the effectiveness of our method for outlier detection. In the next work, we may consider to further reduce the time complexity of our algorithm for finding GrC-based outliers.

Acknowledgements. This work is supported by the Natural Science Foundation (Grant Nos. 60475019 and 60775036), and the Specialized Research Fund for the Doctoral Program of Higher Education of China (Grant No. 20060247039)

References

1. Yao, Y.Y., Zhong, N.: Granular computing using information tables. In: Lin, T.Y., Yao, Y.Y., Zadeh, L.A. (eds.) Data Mining, Rough Sets and Granular Computing, pp. 102–124. Physica-Verlag (2002)
2. Zadeh, L.A.: Fuzzy sets and information granularity. In: Gupta, N., Ragade, R., Yager, R. (eds.) Advances in Fuzzy Set Theory and Applications, pp. 3–18. North-Holland, Amsterdam (1979)
3. Zadeh, L.A.: Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information/intelligent systems. Soft Computing 2(1), 23–25 (1998)
4. Skowron, A., Stepaniuk, J.: Towards discovery of information granules. In: Żytkow, J.M., Rauch, J. (eds.) PKDD 1999. LNCS (LNAI), vol. 1704, pp. 542–547. Springer, Heidelberg (1999)
5. Skowron, A., Stepaniuk, J.: Information Granules in Distributed Environment. In: Zhong, N., Skowron, A., Ohsuga, S. (eds.) RSFDGrC 1999. LNCS (LNAI), vol. 1711, pp. 357–366. Springer, Heidelberg (1999)
6. Miao, D.Q., Wang, G.Y., Liu, Q., et al.: Granular Computing Past, Present and Future Prospect. Science Press, Beijing (2007) (in Chinese)
7. Duan, Q.G., Miao, D.Q., Zhang, H.Y., Zheng, J.: Personalized Web Retrieval based on Rough-Fuzzy Method. Journal of Computational Information Systems 3(3), 1067–1074 (2007)
8. Duan, Q.G., Miao, D.Q., Wang, R.Z., Chen, M.: An Approach to Web Page Classification based on Granules. In: Proc. of 2007 IEEE/WIC/ACM Int. Conf. on Web Intelligence (WI 2007), Silicon Valley, USA, vol. 2-5, pp. 279–282 (2007)
9. Miao, D.Q., Chen, M., Wei, Z.H., Duan, Q.G.: A Reasonable Rough Approximation of Clustering Web Users. In: Zhong, N., Liu, J., Yao, Y., Wu, J., Lu, S., Li, K. (eds.) Web Intelligence Meets Brain Informatics. LNCS (LNAI), vol. 4845, pp. 428–442. Springer, Heidelberg (2007)

10. Yao, Y.Y.: A partition model of granular computing. LNCS Transactions on Rough Sets, vol. 1, pp. 232–253 (2004)
11. Knorr, E., Ng, R.: Algorithms for Mining Distance-based Outliers in Large Datasets. In: Proc. of the 24th VLDB Conf., New York, pp. 392–403 (1998)
12. Hawkins, D.: Identifications of Outliers. Chapman and Hall, London (1980)
13. Breunig, M.M., Kriegel, H.-P., Ng, R.T., Sander, J.: LOF: identifying density-based local outliers. In: Proc. of the 2000 ACM SIGMOD Int. Conf. on Management of Data, Dallas, pp. 93–104 (2000)
14. Jiang, F., Sui, Y.F., Cao, C.G.: Outlier Detection Using Rough Set Theory. In: Ślézak, D., Yao, J., Peters, J.F., Ziarko, W., Hu, X. (eds.) RSFDGrC 2005. LNCS (LNAI), vol. 3642, pp. 79–87. Springer, Heidelberg (2005)
15. Johnson, S.C.: Hierarchical Clustering Schemes. Psychometrika 2, 241–254 (1967)
16. Nguyen, S.H., Nguyen, H.S.: Some efficient algorithms for rough set methods. In: Proc. of the 6th Int. Conf. on Information Processing and Management of Uncertainty (IPMU 1996), Granada, Spain, pp. 1451–1456 (1996)
17. He, Z.Y., Deng, S.C., Xu, X.F.: An Optimization Model for Outlier Detection in Categorical Data. In: Int. Conf. on Intelligent Computing (ICIC(1) 2005), Hefei, China, pp. 400–409 (2005)
18. He, Z.Y., Deng, S.C., Xu, X.F.: Discovering Cluster Based Local Outliers. Pattern Recognition Letters 24(9-10), 1651–1660 (2003)
19. Ramaswamy, S., Rastogi, R., Shim, K.: Efficient algorithms for mining outliers from large datasets. In: Proc. of the 2000 ACM SIGMOD Int. Conf. on Management of Data, Dallas, pp. 427–438 (2000)
20. Harkins, S., He, H.X., Willams, G.J., Baxter, R.A.: Outlier detection using replicator neural networks. In: Proc. of the 4th Int. Conf. on Data Warehousing and Knowledge Discovery, France, pp. 170–180 (2002)
21. Willams, G.J., Baxter, R.A., He, H.X., Harkins, S., Gu, L.F.: A Comparative Study of RNN for Outlier Detection in Data Mining. In: Proc. of the 2002 IEEE Int. Conf. on Data Mining (ICDM 2002), Japan, pp. 709–712 (2002)
22. Bay, S.D.: The UCI KDD repository (1999), `http://kdd.ics.uci.edu`
23. Aggarwal, C.C., Yu, P.S.: Outlier detection for high dimensional data. In: Proc. of ACM SIGMOD Int. Conf. on Managment of Data, California, pp. 37–46 (2001)
24. Liu, B., Hsu, W., Ma, Y.: Integrating Classification and Association Rule Mining. In: Proc.of the 4th Int. Conf. on Knowledge Discovery and Data Mining (KDD 1998), New York, pp. 80–86 (1998)