Expert Systems with Applications 37 (2010) 8745–8749

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Neighborhood outlier detection

Yumin Chen [*], Duoqian Miao, Hongyun Zhang

Department of Computer Science and Technology, Xiamen University of Technology, Xiamen 361024, China
Key Laboratory of Embedded System and Service Computing, Ministry of Education of China, Tongji University, Shanghai 201804, China

ARTICLE INFO

Keywords:
Outlier detection
Rough sets
Neighborhoods
Data mining

ABSTRACT

KNN (k nearest neighbor) is widely discussed and applied in pattern recognition and data mining, however, as a similar outlier detection method using local information for mining a new outlier, neighborhood outlier detection, few literatures are reported on. In this paper, we introduce neighborhood model as a uniform framework to understand and implement neighborhood outlier detection. Furthermore, a neighborhood-based outlier detection algorithm is also given. This algorithm integrates rough set granular technique with outlier detecting. We propose a neighborhood-based metric on outlier detection, and compare neighborhood outlier detection with DIS, KNN and RNN. The experimental results show that neighborhood-based metric is able to measure the local information for outlier detection. The detected accuracies based on neighborhood outlier detection are superior to DIS, KNN for mixed dataset, and a litter better than RNN for discrete dataset.

© 2010 Elsevier Ltd. All rights reserved.

1. Introduction

In contrast to traditional pattern recognition question that aims to construct a general pattern map to the majority of data, outlier detection targets to find the rare data whose behavior is very exceptional when compared with rest large amount of data. One of the most popular outlier detection techniques is distance-based outlier, introduced by Knorr and Ng (1998, 1999). A distance-based outlier in a dataset D is a data object with pct% of the objects in D having a distance of more than d_{min} away from it. This notion generalizes many concepts from distribution-based approach and enjoys better detected accuracy. What is more, it is extended based on the distance of a point from its kth nearest neighbor, which is called KNN method (Ramaswamy, Rastogi, & Kyuseok, 2000). It ranks the top k points by the distance to its kth nearest neighbor as the outliers. Efficient algorithms for mining top-k outliers are also studied. Furthermore, in the algorithm proposed by Angiulli and Pizzuti (2002), the outlier factor of each datum point is computed as the sum of distances from its k nearest neighbors, which obtained better result comparing with traditional KNN. However, as KNN outlier detections computing all the dimensional distances of the points from one another, it is time-consuming if the available objects are of very great size. Besides, the direct application of KNN methods to high dimensional problems often results in unexpected performance and qualitative costs due to the curse of dimensionality.

With increasing awareness on outlier detection in literatures, more concrete meanings of outliers are defined for solving problems in specific domains (Breunig, Kriegel, Ng, & Sander, 2000; Jain, Murty, & Flynn, 1999; Jiang, Sui, & Cao, 2009; Johnson, Kwok, & Ng, 1998; Kovacs, Vass, & Vidacs, 2004; Rousseeuw & Leroy, 1987). In addition to distance-based outlier approach, the other approaches to outlier detection can be classified into five categories, which are distribution-based approach, depth-based approach, clustering approach, density-based approach and RST-based approach (Kovacs et al., 2004). Distribution-based approach is the classical method in statistics. It is based on some standard distribution model (Normal, Poisson, etc.) and those objects which deviate from the model are recognized as outliers (Rousseeuw & Leroy, 1987). Its greatest disadvantage is that the distribution of the measurement data is unknown in practice. Depth-based approach is based on computational geometry and compute different layers of k–d convex hulls and flags objects in the outer layer as outliers (Johnson et al., 1998). However, it is a well-known fact that the algorithms employed suffer from the dimensionality curse and cannot cope with large k. Clustering approach classifies the input data. It detects outliers as by-products (Jain et al., 1999). Since the main objective is clustering, it is not optimized for outlier detection. Density-based approach was originally proposed by Breunig et al. (2000). A local outlier factor (LOF) is assigned to each sample based on their local neighborhood density. Samples with high LOF value are identified as outliers. The disadvantage of this solution is that it is very sensitive to parameters defining the neighborhood. Rough set theory (RST) is proposed by Pawlak

[*] Corresponding author at: Department of Computer Science and Technology, Xiamen University of Technology, Xiamen 361024, China.
E-mail addresses: cym0620@163.com (Y. Chen), miaoduoqian@163.com (D. Miao), zhanghongyun583@sina.com (H. Zhang).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.06.040

(1982), which is applied in machine learning, data mining and pattern recognition successfully. RST-based approach was originally proposed by Jiang et al. (2009). A sequence-based outlier was defined based on RST in information systems. An algorithm to find such outliers was also given, which is effective for discrete data. Furthermore, an expanded distance-based approach using RST was also proposed. However, his RST-based method is suitable to discrete data rather than continues data, for which only considering equivalence class and equivalence relation.

In fact, neighborhoods and neighborhood relations are a class of important concepts in topology. Lin (1988, 1997) pointed out that neighborhood spaces are more general topological spaces than equivalence spaces and introduced neighborhood relation into rough set theory. Yao (1998) and Wu and Zhang (2002) discussed the properties of neighborhood approximation spaces. It is a powerful tool to attribute reduction, feature selection, classification and reasoning with uncertainty (Hu, Yu, & Xie, 2006a; Hu, Yu, Xie, & Liu, 2006b; Jensen & Shen, 2004; Swiniarski & Skowron, 2003). However, few applications of neighborhood rough set model were reported these years. In this paper, we will review the basic concepts on neighborhood and neighborhood rough sets. And then we will propose a method to outlier detection based on neighborhood rough set. This approach integrates its powerful granular ability of uncertainty data with outlier detection, and detects an outlier in the selected subspaces based on the majority class in the neighborhood of the data. Furthermore, as recent rough set outlier detection method only dealing with discrete data, our proposed approach is not only suitable to discrete data but also to continues data. Some experimental analysis is conducted on UCI data sets. The results show that the detected accuracies of proposed detection systems outperform the popular DIS and KNN outlier detection for mixed dataset, and a little better than RNN for discrete dataset.

The remainder of the paper is organized as follows. The basic concepts on neighborhood rough set model are shown in Section 2. The neighborhood outlier detection method is introduced in Section 3. Section 4 presents the experimental analysis. Then the conclusion is given in Section 5.

2. Neighborhood model

Formally, An information system for data mining can be written as a quadruple $IS = (U,A,V,f)$, where: U is a non-empty finite set of objects, called a universe, A is a non-empty finite set of features, V is the union of feature domains such that $V = \bigcup_{a \in A} V_a$ for V_a denotes the value domain of feature a, f: $U \times A \to V$ is an information function such that $f(x,a) \in V_a$ for every $a \in A$ and $x \in U$. We can split set A of features into two subsets: $C \subset A$ and $D = A - C$, conditional set of features and decision (or class) feature(s), respectively. The condition features represent measured features of the objects, while the decision features are a posteriori outcome of classification.

Consider a universe U and a distance function D: $f(x,y) \to R+$, where $R+$ is the set of non-negative real numbers. Given any $x \in U$, $B \subseteq C$ and $q \in R+$, the neighborhood $n_B^q(x)$ of x in the subspace B is defined as

$$n_B^q(x) = \{y | x, y \in U, D_B(x,y) \leqslant q\}$$

D is a distance function, which satisfies

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

A distance metric is a distance function on a set of points, mapping pairs of points into the non-negative real numbers. In general, there are three metric functions that are widely used. Consider that x_1 and x_2 are two objects in n-dimensional space $A = a_1, a_2, \ldots, a_n$, $f(x,a_i)$ denotes the value of sample x in the ith dimension a_i, then a general metric, named Minkowsky distance, is defined as

$$D_p(x,y) = \left(\sum_{i=1}^{n} |f(x,a_i) - f(y,a_i)|^p \right)^{1/p}$$

where (1) it is called Manhattan distance D_1 if $p = 1$; (2) Euclidean distance D_2, if $p = 2$; (3) Chebychev distance D_∞, if $p = \infty$.

Example 1. Given an information system $IS = (U,A,V,f)$, where $U = x_1, x_2, x_3, x_4, x_5$, $A = \{a\}$, as shown in Table 1.

Using the Manhattan distance, supposed $q = 0.1$, we can have the following neighborhoods for objects of U:

$$n_a^q(x_1) = (x_1, x_2), \quad n_a^q(x_2) = (x_1, x_2, x_3), \quad n_a^q(x_3) = (x_2, x_3),$$
$$n_a^q(x_4) = (x_4, x_5), \quad n_a^q(x_5) = (x_4, x_5)$$

Given an information system $IS = (U,A,V,f)$, the family of neighborhoods $n_B^q(x) = \{y | x, y \in U, D_B(x,y) \leqslant q, B \subseteq A\}$ forms a neighborhood system, which covers the universe. A neighborhood relation R over the universe can be written as a relation matrix $M(R) = (r_{ij})_{n \times n}$, where $r_{ij} = 1$ if $D(x_i,x_j) \leqslant q$, otherwise $r_{ij} = 0$. It is easy to show that R satisfies the following properties:

(1) Reflexivity: $r_{ij} = 1$;
(2) Symmetry: $r_{ij} = r_{ji}$.

Obviously, neighborhood relations are one class of similarity relations, which satisfy reflexivity and symmetry. Specially, $n(x)$ is an equivalent class and R is an equivalence relation if $q = 0$, this case is applicable to discrete data. Neighborhood relations draw the objects together for similarity or indistinguishability in terms of distances.

3. Neighborhood-based outlier detection

3.1. The value difference metric under the neighborhood relation

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

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

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

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

$$d_f(x_f, y_f) = (P(x_f) - P(y_f))^2$$

where $P(x_f)$ is the probability of object x on feature f and $P(x_f)$ is the probability of object y on feature f.

Table 1
An example of neighborhoods.

U	A
x_1	0.1
x_2	0.2
x_3	0.3
x_4	0.7
x_5	0.8

Since traditional rough set theory is suitable to discrete data, it deals with not only discrete data but also continue data if we introduce neighborhood relation to rough set. Next we give the revised definition of *VDM* in rough set theory under the neighborhood relation.

Definition 1. Given an information system $IS = (U,A,V,f)$, where U is a non-empty finite set of objects and A is non-empty finite set of attributes. Let $x, y \in U$ be any two objects between which we shall calculate the distance. The value difference metric in rough set theory under the neighborhood relation VDM_N: $U \times U \rightarrow [0,\infty]$ is defined as $VDM_N(x,y) = \sum_{a \in A} d_a(x_a, y_a)$ where $d_a(x_a,y_a)$ denotes the distance between two objects on attribute a, and x_a is the value of object x on attribute a. For any $a \in A$, let q_a is a neighborhood parameter, define

$$d_a(x_a, y_a) = \left(\frac{|n_a^{q_a}(x)|}{|U|} - \frac{|n_a^{q_a}(y)|}{|U|} \right)^2$$

where $n_a^{q_a}(x)$ is a neighborhood of object x on attribute a and $n_a^{q_a}(y)$ is a neighborhood of object y on attribute a.

If values on attribute a are discrete, we set the neighborhood parameter $q_a = 0$, otherwise, set $q \in (0,\infty)$. Obvious, $\frac{|n_a^{q_a}(x)|}{|U|}$ is similar to $P(x_f)$ in the above definition.

Example 2. Given an information system $IS = (U,A,V,f)$, where $U = \{x_1,x_2,x_3,x_4,x_5\}$, $A = \{a,b,c\}$, as shown in Table 2.

The second column and the fourth column are continues data. The third column is discrete data. Let $q_a = 0.1$, $q_b = 0$, $q_c = 0.1$. Using the distance metric defined by Definition 1, we can calculate the distance for every pair of objects in U. Because of the limitation of space, we just present the procedure for calculating the distance between x_1 and y_2.

Initialization:

$$VDM_N(x_1,x_2) = d_a(0.1,0.2) + d_b(1,0) + d_c(0.3,0.4)$$

Step 1: Calculate $d_a(0.1,0.2)$:

$$n_a^{0.1}(x_1) = \{x_1,x_2\}, \quad n_a^{0.1}(x_2) = \{x_1,x_2,x_3\}$$

$$\begin{aligned} d_a(0.1,0.2) &= \left(\frac{|n_a^{0.1}(x_1)|}{|U|} - \frac{|n_a^{0.1}(x_2)|}{|U|} \right)^2 \\ &= \left(\frac{|\{x_1,x_2\}|}{|\{x_1,x_2,x_3,x_4,x_5\}|} - \frac{|\{x_1,x_2,x_3\}|}{|\{x_1,x_2,x_3,x_4,x_5\}|} \right)^2 \\ &= \left(\frac{2}{5} - \frac{3}{5} \right)^2 = \frac{1}{25} \end{aligned}$$

Step 2: Calculate $d_b(1,0)$:

$$n_b^0(x_1) = \{x_1\}, \quad n_b^0(x_2) = \{x_2,x_4,x_5\}$$

$$\begin{aligned} d_b(1,0) &= \left(\frac{|n_b^0(x_1)|}{|U|} - \frac{|n_b^0(x_2)|}{|U|} \right)^2 \\ &= \left(\frac{|\{x_1\}|}{|\{x_1,x_2,x_3,x_4,x_5\}|} - \frac{|\{x_2,x_4,x_5\}|}{|\{x_1,x_2,x_3,x_4,x_5\}|} \right)^2 \\ &= \left(\frac{1}{5} - \frac{3}{5} \right)^2 = \frac{4}{25} \end{aligned}$$

Table 2
An example of value difference metric.

U	a	b	c
x_1	0.1	1	0.3
x_2	0.2	0	0.4
x_3	0.3	2	0.6
x_4	0.7	0	0.7
x_5	0.8	0	0.5

Step 3: Calculate $d_c(0.3,0.4)$:

$$n_c^{0.1}(x_1) = \{x_1,x_2\}, \quad n_c^{0.1}(x_2) = \{x_1,x_2,x_5\}$$

$$\begin{aligned} d_c(0.3,0.4) &= \left(\frac{|n_c^{0.1}(x_1)|}{|U|} - \frac{|n_c^{0.1}(x_2)|}{|U|} \right)^2 \\ &= \left(\frac{|\{x_1,x_2\}|}{|\{x_1,x_2,x_3,x_4,x_5\}|} - \frac{|\{x_1,x_2,x_5\}|}{|\{x_1,x_2,x_3,x_4,x_5\}|} \right)^2 \\ &= \left(\frac{2}{5} - \frac{3}{5} \right)^2 = \frac{1}{25} \end{aligned}$$

Step 4:

$$VDM_N(x_1,x_2) = \frac{1}{25} + \frac{4}{25} + \frac{1}{25} = 0.24$$

Repeating the above calculation, we can finally obtain distances for all the other pairs of objects in U. By then, we define a neighborhood-based object outlier factor (*NOOF*), which indicates the degree of outlier for every object in an information system.

Definition 2 (*Neighborhood-based Object Outlier Factor*). Let $IS = (U,A,V,f)$ be an information system, where $A = \{a_1,a_2,\ldots,a_m\}$ and $U = \{x_1,x_2,\ldots,x_n\}$. For any $x_i \in U$, let neighborhood parameter $q = \{q_{a_1}, q_{a_2}, \ldots, q_{a_m}\}$. The neighborhood-based object outlier factor of x_i in IS is defined as follows:

$$NOOF(x_i) = \sum_{j=1, j \neq i}^{n} VDM_N(x_i, x_j)$$

Definition 3 (*Neighborhood-based Outliers*). Let $IS = (U,A,V,f)$ be an information system, where $A = \{a_1,a_2,\ldots,a_m\}$, $U = \{x_1,x_2,\ldots,x_n\}$ and neighborhood parameter $q = \{q_{a_1}, q_{a_2}, \ldots, q_{a_m}\}$. Let μ be a given threshold value, for any $x \in U$, if $NOOF(x) > \mu$ then x is called a neighborhood-based outlier in U with respect to IS, where $NOOF(x)$ is the neighborhood-based object outlier factor of x in IS.

3.2. Algorithm

```
Neighborhood outlier detection (NED)

Input: an information system $IS = (U,A,V,f)$, where $|U| = n$ and
   $|A| = m$; neighborhood parameter $q = q_{a_1}, q_{a_2}, \ldots, q_{a_m}$,
   threshold value $\mu$.
Output: a set $O$ of neighborhood-based outliers.

For every $a \in A$
{
   For every $x \in U$
   {
      Calculate $|n_a^{q_a}(x)|$;
   }
}
For every $x \in U$
{
   For every $y \in U$
   {
      For every $a \in A$
      {
         Calculate $d_a(x_a, y_a) = \left( \frac{|n_a^{q_a}(x)|}{|U|} - \frac{|n_a^{q_a}(y)|}{|U|} \right)^2$;
      }
      Calculate $VDM_N(x,y)$
   }
   Calculate $NOOF(x)$
   If $NOOF(x) > \mu$, then $O = O \cup x$
}
Return $O$.
```

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

4. Experimental analysis

In this section, following the experimental setup in He, Deng, and Xu (2005), we shall use two real life data sets (Annealing and Cancer) to demonstrate the performance of neighborhood-based outlier detection algorithm (NED) against traditional distance-based method (Knorr & Ng, 1998, Knorr, Ng, & Tucakov, 2000) and KNN algorithm (Ramaswamy et al., 2000). In addition, on the cancer data set, we add the results of RNN-based outlier detection method for comparison. These results can be found in the work of Harkins, He, Willams, and Baxter (2002), Willams, Baxter, He, Harkins, and Gu (2002).

For distance-based method and KNN algorithm, in order to calculate the distance between any two objects, we adopt the overlap metric in rough set theory, which is defined as follows:

Definition 4. Given an information system $IS = (U;A;V;f)$, let x, $y \in U$ be any two objects between which we shall calculate the distance. The overlap metric in rough set theory is defined as

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

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

Furthermore, in our experiment, for the KNN algorithm, the results were obtained by using the fourth nearest neighbor (Ramaswamy et al., 2000) and the overlap metric in rough set theory defined above.

4.1. Annealing data

The first is the Annealing data set, which can be found in the UCI machine learning repository (Bay, 1999). It contains 798 instances (or objects) with 38 attributes (including the class attribute). The 798 instances are partitioned into five classes. Class 3 has the largest number of instances. The remained classes are regarded as rare classes for they are small in size.

Aggarwal and Yu (2001) proposed a practicable way to test the effectiveness of an outlier detection method (Angiulli & Pizzuti, 2002; He et al., 2005). That is, we can run the outlier detection method on a given data set and test the percentage of points which 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 (Angiulli & Pizzuti, 2002)). Points belonged to the rare class are considered as outliers. If the method works well, we expect that such abnormal classes would be over-represented in the set of points found.

In our experiment, data in the Annealing data set is input into an information table $SL = (U;A;V;f)$, where U contains all the 798 instances of Annealing data set and A contains 37 attributes of Annealing data set (not including the class attribute). Since the neighborhood parameters are needed by NED, the corresponding parameters are illustrated in Table 3. These parameters were determined based on a small number of preliminary runs. The experimental results are summarized in Table 4.

Table 4 shows the results produced by the NED algorithm against the KNN algorithm and DIS algorithm. Here, the top ratio is ratio of the number of objects specified as top-k outliers to that of the objects in the dataset. The coverage is ratio of the number of detected rare classes to that of the rare classes in the dataset. For example, we let NED algorithm find the top 80 outliers with the top ratio of 10%. By examining these 80 points, we found that 51 of them belonged to the rare classes. In contrast, when we ran the KNN algorithm on this dataset, we found that only 21 of 80 top outliers belonged to rare classes.

From Table 4, the performance of the NED algorithm outperformed that of the KNN algorithm and the DIS algorithm in all the five cases, especially, when the top ratio is relative small, the NED algorithm worked much better. Anneal dataset has not only discrete data, but also continue data. The experiment shows that the NED algorithm is suitable to mixed data.

Table 3
Neighborhood parameters of annealing data set.

Attribute label	Parameter	Parameter value
4	q_4	5
5	q_5	10
9	q_9	200
33	q_{33}	0.2
34	q_{34}	300
35	q_{35}	800
37	q_{37}	50
Other labels	q_a	0

Table 4
Experimental result in annealing dataset.

Top ratio (number of objects)	Number of rare class included (coverage)		
	NED	KNN	DIS
10% (80)	51 (27%)	21 (11%)	33 (17%)
15% (105)	67 (35%)	30 (16%)	44 (23%)
20% (140)	81 (43%)	41 (22%)	61 (32%)
25% (175)	84 (44%)	58 (31%)	77 (41%)
30% (209)	92 (48%)	62 (33%)	84 (44%)

Table 5
Experimental result in Wisconsin breast cancer dataset.

Top ratio (number of objects)	Number of rare class included (coverage)			
	NED	DIS	KNN	RNN
1% (4)	4 (10%)	4 (10%)	4 (10%)	3 (8%)
2% (8)	7 (18%)	7 (18%)	7 (18%)	6 (15%)
4% (16)	14 (36%)	14 (36%)	13 (33%)	11 (28%)
6% (24)	19 (49%)	21 (54%)	20 (51%)	18 (46%)
8% (32)	26 (67%)	28 (72%)	27 (69%)	25 (64%)
10% (40)	31 (79%)	32 (82%)	32 (82%)	30 (77%)
12% (48)	36 (92%)	36 (92%)	38 (97%)	35 (90%)
14% (56)	38 (97%)	39 (100%)	39 (100%)	36 (92%)
16% (64)	39 (100%)	39 (100%)	39 (100%)	36 (92%)
18% (72)	39 (100%)	39 (100%)	39 (100%)	38 (97%)
20% (80)	39 (100%)	39 (100%)	39 (100%)	38 (97%)
28% (112)	39 (100%)	39 (100%)	39 (100%)	39 (100%)

4.2. Wisconsin breast cancer data

The Wisconsin breast cancer dataset is found in the UCI machine learning repository (Bay, 1999). The data set contains 699 instances with 9 attributes. Here we follow the experimental technique of Harkins et al. by removing some of the malignant instances to form a very unbalanced distribution (Angiulli & Pizzuti, 2002). The resultant dataset had 39 (8%) malignant instances and 444 (92%) benign instances.

Data in the Wisconsin breast cancer data set is also input into an information table $SW = (U; A; V; f)$, where U contains all the 483 instances of the data set and A contains nine attributes of the data set (not including the class attribute). We consider detecting outliers (malignant instances) in SW. Since the dataset is discrete, the parameters for NED are set to 0. The experimental results are summarized in Table 5.

Table 5 is similar to Table 4. From Table 5, we can see that for the Wisconsin breast cancer dataset, the NED performs better than RNN method, and a litter weaker than KNN and DIS. In fact, the performance of NED is more suitable to continue dataset than discrete dataset.

5. Conclusion and future work

Outlier detection is becoming critically important in many areas. In order to deal with not only discrete data but also continue data set, we proposed a new definition of the traditional distance metrics by considering neighborhood information. A measure for identifying the significance of an outlier is also presented. Furthermore, we give the neighborhood-based algorithm for discovering outliers. The experimental results show that our approach outperformed existing methods on identifying meaningful and interesting outliers for mixed dataset.

In the future work, for the neighborhood-based outlier detection algorithm, we shall consider using rough set feature select method to reduce the features while preserving the performance of it. For the performance of the computation of our method, we will sort all objects according to a given order on values of feature to improve the computational complexity.

Acknowledgments

The research is supported by the National Natural Science Foundation of China under Grant Nos. 60903203, 60775036, 60475019, and the Natural Science Foundation of Fujian Province of China under Grant No. 2008J0318.

References

Aggarwal, C.C., & Yu, P.S. (2001). Outlier detection for high dimensional data. In *Proceedings of the ACM international conference on management of data* (pp. 37–46).

Angiulli, F., & Pizzuti, C. (2002). Fast outlier detection in high dimensional spaces, In *Proceedings of PKDD02*.

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

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers, In *Proceedings of the 2000 ACM SIGMOD international conference on management of data, Dallas* (pp. 93–104).

Harkins, S., He, H. X., Willams, G. J., & Baxter, R. A. (2002). Outlier detection using replicator neural networks. In *Proceedings of the fourth international conference on data warehousing and knowledge discovery, France* (pp. 170–180).

He, Z. Y., Deng, S. C., & Xu, X. F. (2005). An optimization model for outlier detection in categorical data. In *International conference on intelligent computing (ICIC(1) 2005), Hefei, China* (pp. 400–409).

Hu, Q. H., Yu, D. R., & Xie, Z. X. (2006a). Information-preserving hybrid data reduction based on fuzzy-rough techniques. *Pattern Recognition Letters, 27*, 414–423.

Hu, Q. H., Yu, D. R., Xie, Z. X., & Liu, J. F. (2006b). Fuzzy probabilistic approximation spaces and their information measures. *IEEE Transactions on Fuzzy Systems, 14*, 191–201.

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

Jensen, R., & Shen, Q. (2004). Semantics-preserving dimensionality reduction: Rough and fuzzy-rough-based approaches. *IEEE Transactions of Knowledge and Data Engineering, 16*, 1457–1471.

Jiang, F., Sui, Y. F., & Cao, C. G. (2009). Some issues about outlier detection in rough set theory. *Expert Systems with Applications, 36*, 4680–4687.

Johnson, T., Kwok, I., & Ng, R. T. (1998). Fast computation of 2-dimensional depth contours. In *Proceedings of the fourth international conference on knowledge discovery and data mining, New York* (pp. 224–228).

Knorr, E. M., & Ng, R. T. (1998). Algorithms for mining distance-based outliers in large datasets. In *Proceedings of the 24th international conference on very large data bases, New York, NY* (pp. 392–403).

Knorr, E. M., & Ng, R. T. (1999). Finding intensional knowledge of distance-based outliers. In *Proceedings of the 25th international conference on very large data bases, Edinburgh, Scotland* (pp. 211–222).

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

Kovacs, L., Vass, D., & Vidacs, A. (2004). Improving quality of service parameter prediction with preliminary outlier detection and elimination. In *Proceedings of the second international workshop on inter-domain performance and simulation (IPS 2004), Budapest* (pp. 194–199).

Lin, T. Y. (1988). Neighborhood systems and relational database. In *Proceedings of 1988 ACM 16th annual computer science conference, February* (pp. 23–25).

Lin, T. Y. (1997). Neighborhood systems-application to qualitative fuzzy and rough sets. In P. P. Wang (Ed.), *Advances in machine intelligence and soft-computing* (pp. 132–155). Durham, North Carolina, USA: Department of Electrical Engineering, Duke University.

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

Ramaswamy, S., Rastogi, R., & Kyuseok, S. (2000). Efficient algorithms for mining outliers from large data sets. In *Proceedings of the ACM SIDMOD international conference on management of data*.

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

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

Swiniarski, R. W., & Skowron, A. (2003). Rough set methods in feature selection and recognition. *Pattern Recognition Letters, 24*, 833–849.

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

Wu, W. Z., & Zhang, W. X. (2002). Neighborhood operator systems and approximations. *Information Sciences, 144*, 201–217.

Yao, Y. Y. (1998). Relational interpretations of neighborhood operators and rough set approximation operators. *Information Sciences, 111*, 239–259.