12-1-2011

Outlier detection using modified-ranks and other variants

Huaming Huang
Syracuse University

Kishan Mehrotra
Syracuse University, mehrotra@syr.edu

Chilukuri K. Mohan
Syracuse University, ckmohan@syr.edu

Department of Electrical Engineering and Computer Science

Technical Report

SYR-EECS-2011-12 December 01, 2011

Outlier Detection Using Modified-ranks and Other Variants

Huaming Huang hhuang13@syr.edu

Kishan Mehrotra kishan@ecs.syr.edu

Chilukuri K. Mohan mohan@ecs.syr.edu

ABSTRACT: Rank-based algorithms provide a promising approach for outlier detection, but currently used rank-based measures of outlier detection suffer from two deficiencies: first they take a large value from an object near a cluster whose density is high even through the object may not be an outlier and second the distance between the object and its nearest cluster plays a mild role though its rank with respect to its neighbor. To correct for these deficiencies we introduce the concept of modified-rank and propose new algorithms for outlier detection based on this concept. Our method performs better than several density-based methods, on some synthetic data sets as well as on some real data sets.

KEYWORDS: Outlier detection, ranking, neighborhood sets, clustering

Syracuse University - Department of EECS,
4-206 CST, Syracuse, NY 13244
(P) 315.443.2652 (F) 315.443.2583
http://eecs.syr.edu

Outlier Detection Using Modified-ranks and Other Variants

Technical Report Number: SYR-EECS-2011-12

Huaming Huang, Kishan Mehrotra, Chilukuri K. Mohan

Department of EECS, Syracuse University

Abstract. Rank-based algorithms provide a promising approach for outlier detection, but currently used rank-based measures of outlier detection suffer from two deficiencies: first they take a large value for an object near a cluster whose density is high even through the object may not be an outlier and second the distance between the object and its nearest cluster plays a mild role though its rank with respect to its neighbor. To correct for these deficiencies we introduce the concept of modified-rank and propose new algorithms for outlier detection based on this concept. Our method performs better than several density-based methods, on some synthetic data sets as well as on some real data sets.

Keywords: Outlier detection, ranking, neighborhood sets, clustering.

1 Introduction

Outlier detection is an important task for data mining applications. Several effective algorithms have been successfully applied in many real-world applications. Density-based algorithms such as "local outlier factor" (LOF) and connectivity-based outlier factor (COF) were proposed by [1] and [7] respectively. Jin *et al.* [5] proposed another modification, called INFLO, which is based on a symmetric neighborhood relationship. Tao and Pi [8] have proposed a density-based clustering and outlier detection (DBCOD) algorithm. Outliers detection based on clustering has been proposed in the literature, see Chandola et al. [3], where an object is declared as an outlier if it does not belong to any cluster. This in turn, requires a new clustering philosophy in which all objects of a given data set are not required to be a cluster. Tao and Pi's [8] clustering approach belongs to this category. In this paper we modify their approach towards this goal, but we differ in the outlier detection step; we use clustering to eliminate the objects that are not suspected outliers and evaluate outlierness of the remaining objects only.

Another rank based detection algorithm (RBDA) was recently proposed by Huang et al.[**?**]. It was observed that RBDA demonstrates superior performance than LOF, COF, and INFLO. However, RBDA is found to assign a large outlierness value to an object in the vicinity of a large cluster, although the object may not be an outlier. In this paper we present few approaches to rectify this deficiency of RBDA — first is a simple modification to RBDA whereas in the

second and third approaches the size of the cluster is explicitly addressed; in all cases clustering acts as a preprocessing step.

The paper is organized as follows. In Section 2, after introducing key notations and definitions, we briefly describe RBDA and DBCOD. In Section 3, first we illustrate the above described weakness of RBDA followed by suggested measures of outlier detection. These new measures are compared with RBDA and DBCOD using one synthetic and three real data sets. Brief descriptions of data sets and a summary of our findings are presented in Section 4, followed by the conclusions and future work.

2 Notation and Definitions

In following notations and concepts are used throughout the paper.

2.1 Notation

- D denotes the given dataset of all observations.
- $d(p, q)$ denotes the distance between two points $p, q \in D$. This distance measure could be any appropriate distance but for concreteness we use the Euclidean distance.
- $d_k(p)$ = the distance between p and its kth nearest neighbor, k is a positive integer.
- $\mathcal{N}_k(p) = \{q \in D - \{p\} : d(p, q) \leq d_k(p)\}$ denotes the set of k nearest neighbors of p.
- $r_q(p)$ denotes the rank of p among neighbors of $q \in \mathcal{N}_k(p)$; i.e., $r_q(p)$ is the rank of $d(q, p)$ in $\{d(p, o) : O \in D - \{q\}\}$.
- $\mathcal{RN}_k(p) = \{q : q \in D \text{ and } p \in \mathcal{N}_k(q)\}$ denotes the set of reverse k nearest neighbors of p.

2.2 Definitions

The following definitions are used in the proposed clustering algorithm; all definitions are relative with respect to a positive integer ℓ. In other words, for example, D-reachable defined below should be viewed as D-reachable given ℓ

D-reachable – An object p is directly reachable (D-reachable) from q, if $p \in \mathcal{N}_\ell(q)$.
*R*eachable – An object p is reachable from q, if there is a chain of objects $p \equiv p_1, \ldots, p_n \equiv q$, such that p_i is D-reachable from p_{i+1} for all values of i.
*C*onnected – If p is reachable from q, and q is reachable from p, then p and q are connected.
*N*eighborhood Clustering – A subset C of D is a cluster of non-outliers if the following three conditions are satisfied:

1. For any two objects p and q in C, $p \neq q$, p and q are connected.

2. For $p \in C$, p is D-reachable from at least two other objects in C.
3. $|C| \geq m^*$, where m^* is the minimum number of objects in a cluster, it is pre-defined by users (domain experts).

Condition 3 above is used to avoid treating a small number of outliers as a cluster. We denote the clustering method as $\mathcal{NC}$-clustering; more formally as $\mathcal{NC}(\ell, m*)$. For instance, $\mathcal{NC}$(6,5) means that a cluster contains connected objects for ℓ =6 and a cluster must contain at least 5 objects.

The values of ℓ and m^* are mainly decided based on domain knowledge. If ℓ is small $\mathcal{NC}$-clustering method will find small and tightly connected clusters and large value of ℓ will find large and loose clusters. If the clusters are small and tight, we expect to find more objects that don't belong to any cluster whereas in the latter case, only a few objects will be declared as outliers. In real world applications (such as credit card fraud detection) most of the transactions are normal and only 0.01% or less of the transactions are fraudulent. In this case, a small value of ℓ is more suitable than a large ℓ.

The value of m^* has a similar effect: if m^* is too small, then the cluster size may also be too small, and a small collection of outliers may be considered as a cluster, which is not what we want. In our experiments, m^* is set to a fixed value of 6.

RBDA is a rank-based outlier detection approach that identifies outliers based on mutual closeness of a data point and its neighbors. For $p, q \in D$, if $q \in \mathcal{N}_k(p)$ and $p \in \mathcal{N}_k(q)$, then p and q are "close" to each other. To capture this concept we define a measure of "outlierness" of p, as follows:

$$O_k(p) = \frac{\sum_{q \in \mathcal{N}_k(p)} r_q(p)}{|\mathcal{N}_k(p)|}.$$

If $O_k(p)$ is 'large' then p is considered an outlier.

Density-based clustering and outlier detection algorithm (DBCOD) For $p \in D$ Tao and Pi [8] define the local density, the neighborhood-based density factor, and neighborhood-based local density factor of p, respectively, as:

$$\mathrm{LD}_k(p) = \frac{\sum_{q \in \mathcal{N}_k(p)} \frac{1}{d(p,q)}}{|\mathcal{N}_k(p)|}, \ \mathrm{NDF}_k(p) = \frac{|\mathcal{RN}_k(p)|}{|\mathcal{N}_k(p)|}, \text{ and } \mathrm{NLDF}_k(p) = \mathrm{LD}_k(p) \times \mathrm{NDF}_k(p).$$

The threshold of NLDF, denoted as thNldf, is defined as:

$$\mathrm{thNldf} = \begin{cases} \min_k(\mathrm{NLDF}_k(p)) & \text{if for all objects } p \in D, \mathrm{NDF}_k(p) = 1 \\ \max_k(\mathrm{NLDF}_k(p)) & \text{otherwise} \end{cases}$$

Using the above definitions, Tao and Pi's [8] find the clusters based on the definitions in section 2.2, *except their definition of D-reachability is as follows: p and q are in each other's k-neighborhood and* $NLDF_k(q) < thNldf$. Points outside the clusters are declared as outliers.

3 Weighted RBDA and other improvements

In general RBDA performs better than density-based algorithms such as LOF, COF and INFLO. A sample performance table is presented in Section 4. these density based measures do not assign appropriate measures of outlierness to one or two objects that are clearly far away from a cluster whereas RBDA is mostly successful. A simple example illustrates this observation. Consider the synthetic dataset in Figure 1. This dataset contains two clusters of different densities and an 'outliers' A. For $k = 5, 6, 7$ or 8, the density-based algorithms such as LOF,

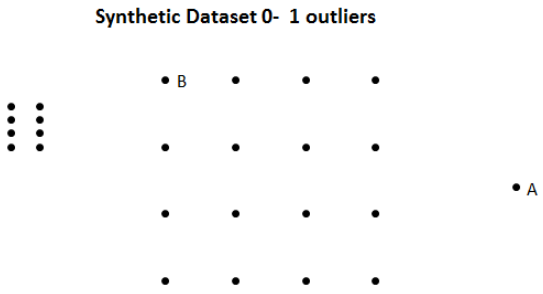

Fig. 1. Synthetic dataset-0 with one outlier, but LOF, COF and INFLO identify B as the most significant outlier.

COF and INFLO do not identify A as the most significant outlier. Instead, B is their top choice, which is wrong. Why B gets a higher outlier value? The reason is that some of B's k-neighbors are from a high density cluster while the others are from a low density cluster and due to mix density of neighborhoods density-based algorithms fail to identify object A as an outlier. RBDA identifies the object A as the most significant outlier.

However, behavior of RBDA is also inconsistent with expectation when an object is near a dense cluster, which we identify as the'cluster density effect'. Consider the data in Figure 2 where two points are of special interest; A in the neighborhood of a cluster with low density (25 objects) and B in the neighborhood of a cluster with high density (491 objects).

By visual inspection, it can be argued that the object 'A' is an outlier whereas object 'B' is a possible but not definite outlier. For k=20, $O_{20}(A)$=25 because rank of 'A' is 25 from all of its neighbors. On the other hand, the ranks of 'B' with respect to its neighbors are: 2, 8,. . . , 132, 205, 227; so that $O_{20}(B)$ is 93.1. RBDA concludes that 'B' is more likely outlier than 'A'. It is clearly an artifact due to large and dense cluster in the neighborhood of 'B', i.e., a point closer to a dense cluster is likely to be misidentified as an outlier, even though it may not be. Such behavior of RBDA, due to cluster density, is observed for some values of k.

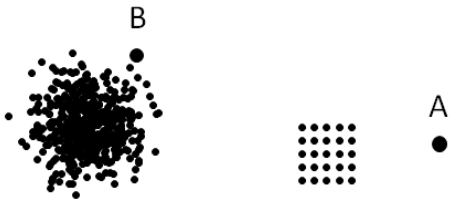

Fig. 2. An example to illustrate 'Cluster Density Effect' on RBDA; RBDA assigns larger outlierness measure to B.

By visual inspection, we generally conclude that a point is an outlier if it is 'far away' from the cluster. This implies that the distance of the object (from the cluster) plays an important role; but accounted for in RBDA only through ranks. Perhaps this deficiency in RBDA can be fixed by incorporating distance in RBDA. The distance can be measured in many ways; either collectively for objects in $\mathcal{N}_k(p)$ or by accounting for the distance of each $q \in \mathcal{N}_k(p)$ separately. These different ways of accounting for distance lead to potentially many possible measures of outlierness. We have explored some of them but in the next subsection we present only one that performed better than others.

3.1 Weighted RBDA

Rank-based approach ignores useful information contained in the distance of the object from other neighboring objects. To overcome the weakness of RBDA due to "cluster density effect", we propose to adjust the value of RBDA by the average distance of p from its $k-$neighbors. Step by step description of this rank and distance based detection algorithm (RADA) is given below:

1. Choose three positive integers k, ℓ, m^*.
2. Find the clusters in D by $\mathcal{NC}(\ell, m^*)$ method.
3. An object o is declared a potential-outlier if it is does not belong to any cluster.
4. Calculate a measure of outlierness: $W_k(p) = O_k(p) \times \frac{\sum_{q \in \mathcal{N}_k(p)} d(q,p)}{|\mathcal{N}_k(p)|}$.
5. If p is a potential-outlier and $W_k(p)$ is large, declare p is an outlier.

For the dataset in Figure 2 we observe that $W_{20}(\text{A}) = 484.82$ and $W_{20}(\text{B}) = 396.19$ implying that A is more likely outlier than B, illustrating that RADA is capable of fixing the discrepancy observed in RBDA.

3.2 Outlier detection using modified-ranks (ODMR)

In this section we propose an alternative procedure to overcome the cluster density effect. We have observed that the size of neighboring cluster plays an

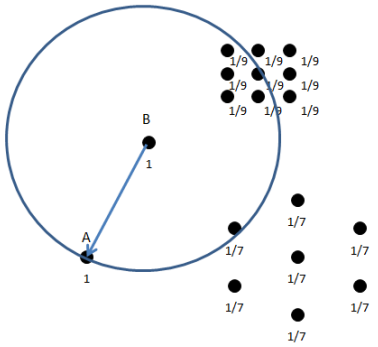

Fig. 3. Assignment of weights in different clusters and modified-rank (modified-rank of A, with respect to B, is $1 + 1 + 5 \times \frac{1}{9} + \frac{1}{7}$.)

important role when calculating the object's outlierness via RBDA. To modify this effect, all clusters of all sizes are assigned equal weights (including isolated points viewed as a cluster of size 1) and all $|C|$ observations of the cluster are assigned equal weights $= 1/|C|$.[1] The rank $r_q(p)$ of an observation p is equal to the number of points within a circle of radius $d(q,p)$ centered at q. In RBDA we sum $r_q(p)$ for all values of $q \in \mathcal{N}_k(p)$. In the proposed version, we calculate "modified-rank" of p, which is defined as the sum of weights associated with all observations within the circle of radius $d(q,p)$ centered at q; that is

$$\text{modified-rank of } p \text{ from } q = mr_q(p) = \sum_{s \in \{d(q,s) \le d(q,p)\}} \text{weight}(s),$$

and sum the "modified-ranks" in $q \in \mathcal{N}_k(p)$.

Figure 3 illustrates how modified-rank is calculated. Step by step description of the proposed method is as follows:

1. Choose three positive integers k, ℓ, m^*.
2. Find clusters in D by $\mathcal{NC}(\ell, m^*)$. All objects not belonging to any cluster are declared as potential-outliers.
3. If C is a cluster and $p \in C$, then the weight of p is $b(p) = \frac{1}{|C|}$.
4. For $p \in D$ and $q \in N_k(p)$, Q denotes the set of points within a circle of radius $d(q,p)$, i.e., $Q = \{s \in D | d(q,s) \le d(q,p)\}$. Then the modified-rank of p with respect to q, denoted as $mr_q(p)$, is computed as $mr_q(p) = \sum_{s \in Q} b(s)$.
5. For a potential outlier p, its ODMR-outlierness, denoted as $\text{ODMR}_k(p)$, is defined as: $\text{ODMR}_k(p) = \sum_{q \in \mathcal{N}_k(p)} mr_q(p)$
6. If p is a potential outlier and $\text{ODMR}_k(p)$ is large, we declare p is an outlier.

[1] We have experimented with another weight assignment to points within a cluster, equal to $1/\sqrt{|C|}$, but the results are not as good as when weights are $1/|C|$.

3.3 Outlier detection using modified-ranks with distance (ODMRD)

Influenced by the distance consideration of section 3.1, in this section we present yet another algorithm that combines ODMA and distance. $\text{ODMRD}_k(p)$ is obtained by implementing all steps as before except Step 5 of the previous algorithm is modified as follows:

(5*) For a potential outlier p, its ODMRD-outlierness, denoted as $\text{ODMRD}_k(p)$, is defined as: $\text{ODMRD}_k(p) = \sum_{q \in \mathcal{N}_k(p)} mr_q(p) \times d(q,p)$

4 Experiments

4.1 Datasets

We use one synthetic and three real datasets to compare the performance of RBDA with RADA, ODMR, ODMRD and DBCOD.

Real Datasets Real datasets consist of iris, ionosphere, and Wisconsin breast cancer datasets obtained from UCI repository. The real datasets were used in two different ways, following the criterion used in [4],[7], and [2]:

1. By making a rare set out of one the class. (1) In the Iris dataset, which is a three-class problem and contains 150 observations equally divided in three classes, 45 observations were removed randomly from the iris-setosa class. (2) In the ionosphere dataset, which is a two-class problem, out of 126 'bad' instances, 116 were randomly removed, leaving 10 'outliers'. (3) Finally, in the Wisconsin dataset, which is also a two-class problem and consists of 236 observations of benign and 236 observations of malignant cancer, after removing duplicates and observations with missing features, 226 malignant observations were removed, leaving 10 'outliers'.
2. By planting new observations in the existing datasets. These planted observations are such that one or more features are assigned the extreme values. (1) In the Iris dataset three observations were planted, (2) in the ionosphere dataset three outliers were planted and (3) in the Wisconsin dataset two outliers were planted.

Synthetic datasets The synthetic datasets are two dimensional so that it is easy to see and interpret the results. Synthetic dataset consists of 515 instances including six planted outliers; has one large normally-distributed cluster and two small uniform clusters. This datasets is intended to test the algorithms' ability to overcome the problem of "cluster density effect". This dataset and clusters obtained by an application of $\mathcal{NC}(6,6)$, are depicted in Figure 6.

4.2 Performance Measures

To measure the performance, three metrics are selected - m_t, recall and RankPower, [**?**]; briefly defined below. We list m most suspicious objects in the dataset D, by a given outlier detection algorithm, which contains exactly d_t true outliers. Let

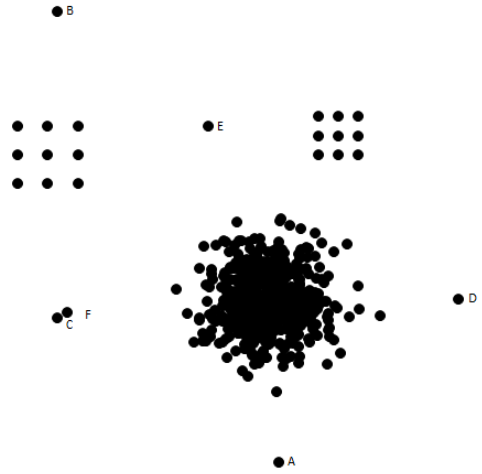

Fig. 4. Synthetic dataset

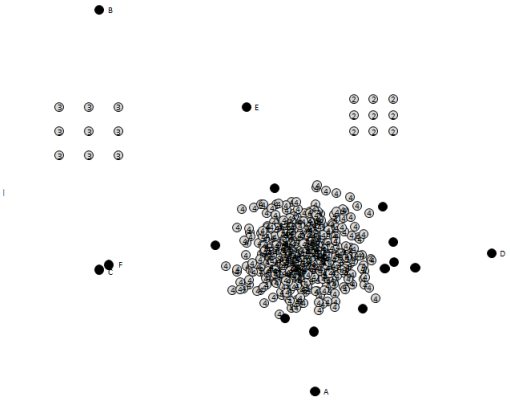

Fig. 5. Synthetic dataset with clusters found by $\mathcal{NC}(6,6)$; black object represents the outliers.

the algorithm produces m_t (true) outliers out of m. Suppose that the algorithm assigns the rank R_i to the ith outlier among m, where $R_i = 1$ represents most suspicious outlier and a larger value of R_i means that the algorithm considers that the ith outlier is less suspicions. Based on these values the performance measures we consider are:

$$\text{Recall} = \frac{|m_t|}{|d_t|}, \quad \text{RankPower} = \frac{n(n+1)}{2\sum_{i=1}^{n} R_i}.$$

RankPower summarizes the overall performance of an algorithm but an object by object assignment of ranks is naturally more illuminating.

4.3 Results

In this section we present a sample of results, extensive tables for all datasets for various values of m and k are available in the Appendix of the technical report [**?**]. Table1 compares RBDA with density based outlier detection methods LOF, COF, INFLO. Rank table of planted outliers in the synthetic dataset is presented in Table 2. In Table 3 we compare RBDA, ODMR , ODMRD, RADA , and DBCOD using RankPower for ionosphere dataset with rare class and in Table4 we summarize of RankPower for all datasets.

5 Conclusion

We observe that rank based approach is highly influenced by the density of neighboring cluster. Furthermore, by definition, ranks use the relative distances and

Table 1. Comparison of LOF, COF, INFLO and RBDA for k = 11, 15, 20 and 23 respectively for the Ionosphere dataset. Maximum values are marked as bold.

m	LOF				COF				INFLO				RBDA			
	Nrc	Pr	Re	RP	Nrc	Pr	Re	RP	Nrc	Pr	Re	RP	Nrc	Pr	Re	RP
5	**5**	**1.00**	**0.50**	**1.000**	**5**	**1.00**	**0.50**	**1.000**	**5**	**1.00**	**0.50**	**1.000**	**5**	**1.00**	**0.50**	**1.000**
15	6	0.40	0.60	**1.000**	6	0.40	0.60	0.778	6	0.40	0.60	**1.000**	**8**	**0.53**	**0.80**	0.818
30	7	0.23	0.70	0.667	7	0.23	0.70	0.560	7	0.23	0.70	0.651	**9**	**0.30**	**0.90**	**0.703**
60	8	0.13	0.80	0.409	8	0.13	0.80	0.409	8	0.13	0.80	0.419	**9**	**0.15**	**0.90**	**0.703**
85	9	0.11	0.90	0.294	9	0.11	0.90	0.290	9	0.11	0.90	0.300	**10**	**0.12**	**1.00**	**0.372**

Table 2. Outliers detected by RBDA, ODMR, ODMRD, RADA, and DBCOD in the synthetic dataset, for $k = 25$. Recall in this dataset the set of planted six outliers is S = { A, B,C, D, E,F}.

m	RBDA	ODMR	ODMRD	RADA	DBCOD
6	{A,B,C,D,E,F}	{A,B,C,D,E,F}	{A,B,C,D,E,F}	{A,B,C,D,E,F}	{ A,B,D}

Table 3. Performance measures of RBDA, ODMR , ODMRD, RADA , and DBCOD for ionosphere dataset

m	RBDA			ODMR			ODMRD			RADA			DBCOD		
	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP
5	5	0.5	1	5	0.5	1	5	0.5	1	5	0.5	1	0	0	0
15	8	0.8	0.783	8	0.8	0.783	8	0.8	0.818	8	0.8	0.818	0	0	0
30	9	0.9	0.703	9	0.9	0.682	9	0.9	0.726	9	0.9	0.726	0	0	0
60	9	0.9	0.703	9	0.9	0.682	9	0.9	0.726	9	0.9	0.726	9	0.9	0.091
85	10	1	0.369	10	1	0.364	10	1	0.390	10	1	0.387	10	1	0.098

Table 4. Summary of RBDA, ODMR, ODMRD, RADA and DBCOD for all experiments.

Dataset	RBDA	ODMR	ODMRD	RADA	DBCOD
Synthetic	3.00	1.00	1.00	1.00	5.00
Iris with rare class	2.67	2.00	2.00	2.33	5.00
Ionosphere with rare class	3.80	3.20	1.20	1.80	5.00
Wisconsin with rare class	3.33	3.00	3.67	1.67	2.67
Iris with outliers	1.00	1.00	1.00	1.00	1.00
Ionosphere with outliers	3.00	3.00	1.50	1.00	5.00
Wisconsin with outliers	1.00	1.00	1.00	1.00	5.00
Summary	2.54	2.03	1.62	**1.40**	4.10

Numbers in the table represent the average performance of the algorithms; a small value implies better performance.

ignore the 'true' distances between the observations. An outlier detection algorithm benefits from 'true' distance as well. Thus we introduce distance in RBDA and observe that the overall performance of RADA is much better than the original RBDA. That the 'true' distance plays an important role is further confirmed by the performance of the alternative algorithms ODMR and ODMRD; it is observed that in general ODMRD performs better than ODMR. We plan to further investigate the proposed algorithms for robustness and consistency.

A Experiments Results

In this section, we will cover the detail about the datasets, and how we conduct the experiments.

A.1 Synthetic Dataset

Synthetic dataset consists of 515 instances including planted six outliers; has one large normally-distributed cluster and two small uniform clusters. It is the synthetic dataset that can be used to test the algorithms' ability to overcome the problem of "cluster density effect". Four different values of k, 25, 35 and 50 are selected and m = 6, 10, and 16.

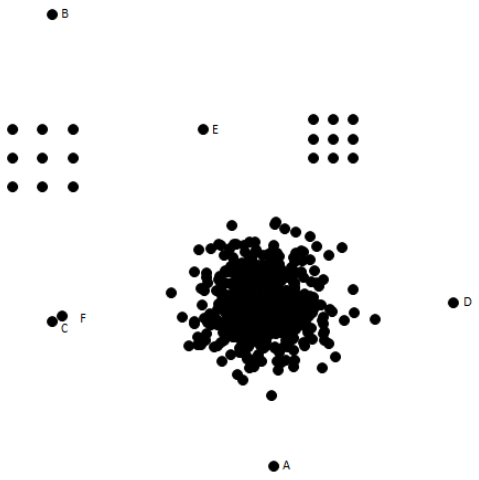

Fig. 6. Synthetic Dataset

The following figure shows the clusters and outliers found by $\mathcal{NC}(6,6)$. Black object represents the outlier object.

For k is 25, 35 and 50, ODMR, ODMRD and RADA are the best algorithms since they all rank the six real outliers in their top 6 outputs. And DBCOD

Table 5. Rank table of comparison of RBDA,ODMR,ODMRD,RADA and DBCOD for 25, 35 and 50 respectively for synthetic dataset. Number in table represents the rank in the output by descending order.

m	RBDA	ODMR	ODMRD	RADA	DBCOD
6	A,B,C D,E,F	A,B,C D,E,F	A,B,C D,E,F	A,B,C D,E,F	A,B,D

m	RBDA	ODMR	ODMRD	RADA	DBCOD
6	A,B,C D,F	A,B,C D,E,F	A,B,C D,E,F	A,B,C D,E,F	A,B,D

m	RBDA	ODMR	ODMRD	RADA	DBCOD
6	A,B,C D,F	A,B,C D,E,F	A,B,C D,E,F	A,B,C D,E,F	A,B,D

Table 6. Comparison of RBDA, ODMR, ODMRD, RADA and DBCOD for k = 25, 35 and 50 respectively for synthetic dataset. Maximum values are marked as red.

m	RBDA			ODMR			ODMRD			RADA			DBCOD		
	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP
6	6	1	1	6	1	1	6	1	1	6	1	1	3	0.5	0.857
10	6	1	1	6	1	1	6	1	1	6	1	1	4	0.667	0.667
15	6	1	1	6	1	1	6	1	1	6	1	1	5	0.833	0.577

m	RBDA			ODMR			ODMRD			RADA			DBCOD		
	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP
6	5	0.833	1	6	1	1	6	1	1	6	1	1	3	0.5	0.857
10	5	0.833	1	6	1	1	6	1	1	6	1	1	5	0.833	0.577
15	6	1	0.808	6	1	1	6	1	1	6	1	1	5	0.833	0.577

m	RBDA			ODMR			ODMRD			RADA			DBCOD		
	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP
6	5	0.833	1	6	1	1	6	1	1	6	1	1	3	0.5	0.857
10	5	0.833	1	6	1	1	6	1	1	6	1	1	3	0.5	0.857
15	5	0.833	1	6	1	1	6	1	1	6	1	1	5	0.833	0.484

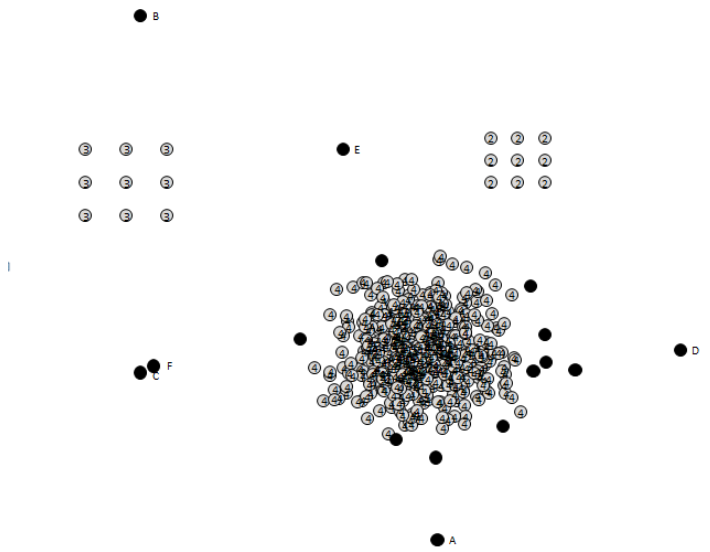

Fig. 7. Synthetic dataset with clusters found by $\mathcal{NC}(6,6)$. Black object represents the outlier object.

gets the worst performance for all values of k. In general, RBDA works better than DBCOD but worse than the others. And the ODMR, ODMRD and RADA algorithms show the very good performance and excellent ability to overcome the problem of cluster density effect since all of them get the maximum RankPower and maximum recall for all values of k.

A.2 Real Datasets:

We have used three well known datasets, namely the Iris, Ionosphere, and Wisconsin breast cancer datasets. We use two ways to evaluate the effectiveness and accuracy of outlier detection algorithms; (i) detect rare classes within the datasets (which has also been used by other researchers such as Feng *et al.* and Tang *et al.* [4, 6]) and (ii) plant outliers into the real datasets (according to datasets' domain knowledge) and expect outlier detection algorithms to identify them.

A.3 Real Datasets with Rare Classes

In this sub-section, we compare the algorithms in detecting rare classes. A class is made 'rare' by removing most of its observations. In all cases, the value of k is chosen between 1% to 10% percentage of the size of the dataset. Because the attributes are dependent, Mahalanobis distance is used to measure the distance between two points.

Iris Dataset The dataset is about iris plant and contains three classes: iris setosa, iris versicolour, iris virginica with 50 instances each. The iris setosa class is linearly separable from the other two classes, but the other two classes are not linearly separable from each other. We randomly remove 45 instances from iris-setosa class to make it 'rare'; remaining 105 instances are used in the final dataset. Three selected values of k are 5, 7, 10. Tables summarize our findings.

Table 7. Comparison of RBDA, ODMR, ODMRD, RADA and DBCOD for k = 5, 7 and 10 respectively for the Iris dataset with rare class. Maximum values are marked as red.

m	RBDA			ODMR			ODMRD			RADA			DBCOD		
	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP
5	1	0.2	0.2	3	0.6	0.75	3	0.6	0.75	1	0.2	0.2	0	0	0
10	4	0.8	0.37	5	1	0.652	5	1	0.714	5	1	0.429	1	0.2	0.125
15	5	1	0.385	5	1	0.652	5	1	0.714	5	1	0.429	4	0.8	0.2
20	5	1	0.385	5	1	0.652	5	1	0.714	5	1	0.429	4	0.8	0.2

m	RBDA			ODMR			ODMRD			RADA			DBCOD		
	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP
5	3	0.6	0.75	4	0.8	1	4	1	0.833	3	0.6	0.857	0	0	0
10	5	1	0.714	5	1	0.882	5	1	0.882	5	1	0.714	1	0.2	0.125
15	5	1	0.714	5	1	0.882	5	1	0.882	5	1	0.714	4	0.8	0.2
20	5	1	0.714	5	1	0.882	5	1	0.882	5	1	0.714	4	0.8	0.2

m	RBDA			ODMR			ODMRD			RADA			DBCOD		
	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP
5	5	1	1	4	0.8	1	4	0.8	1	5	1	1	0	0	0
10	5	1	1	5	1	0.938	5	1	0.882	5	1	1	3	0.6	0.231
15	5	1	1	5	1	0.938	5	1	0.882	5	1	1	5	1	0.288
20	5	1	1	5	1	0.938	5	1	0.882	5	1	1	5	1	0.288

For k is 5, ODMRD is the best outlier detection algorithm, and ODMR is the second best. For k is 7, ODMRD and ODMR both achieve the best performance. DBCOD has the worst RandPower and recall which means that it has the worst performance. For k is 10, RBDA and RADA work better than all others, and DBCOD is still the worst algorithm in this experiment.

Johns Hopkins University Ionosphere Dataset The Johns Hopkins University Ionosphere dataset contains 351 instances with 34 attributes; all attributes are normalized in the range of 0 and 1. There are two classes labeled as good and bad with 225 and 126 instances respectively. There is no duplicate instances in the dataset. To form the rare class, 116 instances from the bad class are randomly removed. Final dataset has only 235 instances with 225 good and

10 bad instances. Four values of k =11, 15, 20 and 23 are used and for different value of k the m values also vary.

For k is 7, 11, 15, and 23, ODMRD works the best, and RADA performs the second best and only has a little gap of performance from ODMRD. When k is 20, RADA is even better than ODMRD for m is 15, 30 and 60. RBDA works better than DBCOD algorithms, but it performs worse than the others. In general, ODMRD is the best outlier detection algorithm in this experiment.

Wisconsin Diagnostic Breast Cancer Dataset Wisconsin diagnostic breast cancer dataset contains 699 instances with 9 attributes. There are many duplicate instances and instances with missing attribute values. After removing all duplicate and instances with missing attribute values, 236 instances labeled as benign class and 236 instances as malignant were left. Total 226 malignant instances are randomly removed following the method proposed by Cao. The final dataset consists of 213 benign instances and 10 malignant instances in our experiments.

For k is 7, RBDA and RADA both work the best. The DBCOD algorithm gets the best RankPower when m is 25, but it only detects 9 of 10 outliers and has the worst precision and recall. For other values of m, it gets the worst RankPower.
For k is 11, RADA achieves the best performance. DBCOD performs the second best. ODMRd works only better than RBDA.
For k is 22, DBCOD achieves the best performance for all values of m. And ODMR shows the second-best performance in five algorithms.

RADA shows the best performance and DBCOD gets the second best in this experiment.

A.4 Real datasets with planted outliers

Detecting rare class instances may not be adequate to measure performance of an algorithm designed to detect outliers; because it may not be appropriate to declare them as outliers. In experiments described in this subsection we plant some outliers into the real datasets according to datasets' domain knowledge.

Iris plant dataset with Outliers Three outliers are inserted into IRIS dataset, that is, there are three classes with 50 instances each and 3 planted outliers. The first outlier has maximum attribute values, second outlier has minimum attribute values, and the third has two attributes with maximum values and the other two with minimum values. Three values of k, 7, 10, and 15 are selected for this experiment.

It can be seen that all algorithms perform well in this experiment, and all get the best performance.

Table 8. Comparison of RBDA, ODMR, ODMRD, RADA and DBCOD for 7, 11, 15, 20 and 23 respectively for the Ionosphere dataset with rare class. Maximum values are marked as red.

m	RBDA			ODMR			ODMRD			RADA			DBCOD		
	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP
5	5	0.5	1	5	0.5	1	5	0.5	1	5	0.5	1	0	0	0
15	8	0.8	0.783	8	0.8	0.783	8	0.8	0.818	8	0.8	0.818	0	0	0
30	9	0.9	0.703	9	0.9	0.682	9	0.9	0.726	9	0.9	0.726	0	0	0
60	9	0.9	0.703	9	0.9	0.682	9	0.9	0.726	9	0.9	0.726	9	0.9	0.091
85	10	1	0.369	10	1	0.364	10	1	0.390	10	1	0.387	10	1	0.098

m	RBDA			ODMR			ODMRD			RADA			DBCOD		
	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP
5	5	0.5	1	5	0.5	1	5	0.5	1	5	0.5	1	0	0	0
15	8	0.8	0.818	8	0.8	0.818	8	0.8	0.818	8	0.8	0.818	0	0	0
30	9	0.9	0.703	9	0.9	0.703	9	0.9	0.726	9	0.9	0.726	0	0	0
60	9	0.9	0.703	9	0.9	0.703	9	0.9	0.726	9	0.9	0.726	9	0.9	0.098
85	10	1	0.372	10	1	0.374	10	1	0.393	10	1	0.387	10	1	0.105

m	RBDA			ODMR			ODMRD			RADA			DBCOD		
	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP
5	5	0.5	1	5	0.5	1	5	0.5	1	5	0.5	1	0	0	0
15	8	0.8	0.818	8	0.8	0.818	8	0.8	0.837	8	0.8	0.837	0	0	0
30	9	0.9	0.714	9	0.9	0.703	9	0.9	0.738	9	0.9	0.738	0	0	0
60	9	0.9	0.714	9	0.9	0.703	9	0.9	0.738	9	0.9	0.738	9	0.9	0.104
85	10	1	0.377	10	1	0.382	10	1	0.407	10	1	0.401	10	1	0.111

m	RBDA			ODMR			ODMRD			RADA			DBCOD		
	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP
5	5	0.5	1	5	0.5	1	5	0.5	1	5	0.5	1	0	0	0
15	8	0.8	0.837	8	0.8	0.818	8	0.8	0.837	8	0.8	0.857	0	0	0
30	9	0.9	0.738	9	0.9	0.726	9	0.9	0.738	9	0.9	0.75	0	0	0
60	9	0.9	0.738	9	0.9	0.726	9	0.9	0.738	9	0.9	0.75	9	0.9	0.106
85	10	1	0.387	10	1	0.39	10	1	0.414	10	1	0.414	10	1	0.114

m	RBDA			ODMR			ODMRD			RADA			DBCOD		
	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP
5	5	0.5	1	5	0.5	1	5	0.5	1	5	0.5	1	0	0	0
15	8	0.8	0.837	8	0.8	0.837	8	0.8	0.857	8	0.8	0.857	0	0	0
30	9	0.9	0.738	9	0.9	0.738	9	0.9	0.75	9	0.9	0.75	0	0	0
60	9	0.9	0.738	9	0.9	0.738	9	0.9	0.75	9	0.9	0.75	9	0.9	0.114
85	10	1	0.393	10	1	0.399	10	1	0.426	10	1	0.417	10	1	0.119

Table 9. Comparison of RBDA, ODMR, ODMRD, RADA and DBCOD for k= 7, 11, and 22 respectively for the Wisconsin dataset with rare class. Maximum values are marked as red.

m	RBDA			ODMR			ODMRD			RADA			DBCOD		
	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP
15	9	0.9	0.714	7	0.7	0.8	8	0.8	0.8	8	0.8	0.8	9	0.9	0.662
25	10	1	0.64	10	1	0.611	10	1	0.618	10	1	0.64	9	0.9	0.662
40	10	1	0.64	10	1	0.611	10	1	0.618	10	1	0.64	10	1	0.545

m	RBDA			ODMR			ODMRD			RADA			DBCOD		
	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP
15	7	0.7	0.651	8	0.8	0.655	8	0.8	0.735	8	0.8	0.783	9	0.9	0.763
25	10	1	0.573	10	1	0.579	10	1	0.573	10	1	0.604	9	0.9	0.763
40	10	1	0.573	10	1	0.579	10	1	0.573	10	1	0.604	10	1	0.598

m	RBDA			ODMR			ODMRD			RADA			DBCOD		
	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP
15	8	0.8	0.667	8	0.8	0.735	8	0.8	0.750	8	0.8	0.750	9	0.9	0.804
25	9	0.9	0.634	10	1	0.585	10	1	0.573	10	1	0.579	9	0.9	0.804
40	10	1	0.567	10	1	0.585	10	1	0.573	10	1	0.579	10	1	0.625

Table 10. Comparison of RBDA, ODMR, ODMRD, RADA and DBCOD for all selected k values(7, 10, 15) for the iris dataset with outliers. Maximum values are marked as red.

m	RBDA			ODMR			ODMRD			RADA			DBCOD		
	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP
10	3	1	1	3	1	1	3	1	1	3	1	1	3	1	1
15	3	1	1	3	1	1	3	1	1	3	1	1	3	1	1

Johns Hopkins University Ionosphere Dataset with Outliers For ionosphere dataset, two classes labeled as good and bad with 225 and 126 instances respectively are kept in resulting dataset. Three outliers are inserted into the dataset; first two outliers have maximum or minimum value in every attribute, and the third has 9 attributes with unexpected values and 25 attributes with maximum or minimum values. Unexpected value here is the value that is valid between minimum and maximum number but is never observed in real datasets[2]. Four values of k, 7, 18, 25, and 35 are chosen for this experiment.

Table 11. Comparison of RBDA, ODMR, ODMRD, RADA and DBCOD for k=7, 18, 25 and 35 respectively for the Ionosphere Dataset with Outliers. Maximum values are marked as red.

m	RBDA			ODMR			ODMRD			RADA			DBCOD		
	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	1	0.333	0.059	1	0.333	0.059	1	0.333	0.083	1	0.333	0.091	0	0	0
30	2	0.667	0.068	2	0.667	0.068	1	0.33	0.083	1	0.333	0.091	0	0	0
40	3	1	0.072	3	1	0.072	3	1	0.076	3	1	0.077	0	0	0

m	RBDA			ODMR			ODMRD			RADA			DBCOD		
	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	1	0.333	0.063	1	0.333	0.063	1	0.333	0.083	1	0.333	0.083	0	0	0
30	1	0.333	0.063	1	0.333	0.063	1	0.333	0.083	1	0.333	0.083	0	0	0
40	3	1	0.073	3	1	0.073	3	1	0.077	3	1	0.077	0	0	0

m	RBDA			ODMR			ODMRD			RADA			DBCOD		
	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	1	0.333	0.077	1	0.333	0.077	1	0.333	0.083	1	0.333	0.091	0	0	0
30	1	0.333	0.077	1	0.333	0.077	1	0.333	0.083	1	0.333	0.091	0	0	0
40	3	1	0.075	3	1	0.075	3	1	0.077	3	1	0.078	0	0	0

m	RBDA			ODMR			ODMRD			RADA			DBCOD		
	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	1	0.333	0.083	1	0.333	0.083	1	0.333	0.091	1	0.333	0.091	0	0	0
30	1	0.333	0.083	1	0.333	0.083	2	0.667	0.073	2	0.667	0.073	0	0	0
40	3	1	0.078	3	1	0.078	3	1	0.08	3	1	0.08	0	0	0

For k is 7, and 25 RADA have the best performance for all values of m. For k is 18 and 35, ODMRD and RADA have the same best performance. RBDA

[2] For example, one attribute may have a range from 0 to 100, but value of 12 never appears in real dataset.

and ODMR perform exactly same for all values of k and m. DBCOD gets the worst performance with all zeros of recall and RankPower.
In general, RADA is the best and ODMRD is the second best in this experiment.

Wisconsin Diagnostic Breast Cancer with Outliers Two outliers are planted into the dataset which has only 449 instances with 213 instances labeled as benign and 236 as malignant. There are no duplicated instances or instances with missing attribute values in the final dataset. Four values of k, 7, 22, 35 and 45 are chosen.

Table 12. Comparison of RBDA, ODMR, ODMRD, RADA and DBCOD for k=7 for the Wisconsin Dataset with Outliers. Maximum values are marked as red.

m	RBDA			ODMR			ODMRD			RADA			DBCOD		
	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP
10	2	1	1	2	1	1	2	1	1	2	1	1	2	1	0.188

Table 13. Comparison of RBDA, ODMR, ODMRD, RADA and DBCOD for k=22, 35 and 45 for the Wisconsin Dataset with Outliers. Maximum values are marked as red.

m	RBDA			ODMR			ODMRD			RADA			DBCOD		
	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP	m_t	Re	RP
10	2	1	1	2	1	1	2	1	1	2	1	1	2	1	0.75

The results table clearly shows that RBDA, RADA, ODMR and ODMRD all achieve the best performance for all values of k. DBCOD has the worst performance.

B Advantages and Disadvantages of RBDA

Compared with other outlier detection algorithms, especially with density-based algorithms, RBDA has many advantages:

- It can detect outliers which have mix density of neighborhoods effectively.
- Its computation is simple and straight forward.
- The concept of rank can be used not only for distance measurements, but also for other type of measurements.

Even RBDA shows the superior performance than LOF, COF and INFLO, it still has some weaknesses:

- Cluster density effect.
- Border effect. It may declare an object in the border of a certain dataset as an outlier even it might not be. This is common flawness for most of k-nearest neighbor based outlier detection algorithms.

As mentioned in previous section, the 'cluster density effect' only happens in certain datasets for certain values of k. In fact, if the size of small cluster is known, then simply increasing the value of k to a number that is larger than the size of small cluster can solve this weakness easily according to our observations and experiments. Unfortunately the size of small cluster near an outlier in practical applications is unknown or hard to be determined. In this case, our proposed method - ODMR, ODMRD and RADA in this paper can solve the problem.

C Distribution of RBDA

To study on the distribution of the RBDA algorithm and observe RBDA's behaviors under different distributions, we create two different distribution synthetic datasets without any outliers: uniform and Gaussian.
To get the more reliable statistical results, different data size and k are also explored. We tried dataset sizes for 100, 200, 300,. . .,2000 objects, and k for 3, 4, 5,. . .,15. For each combination of data size, k and distribution type, we generate 50 different synthetic datasets, and apply our RBDA algorithm, then analyze the results based on average of 50 statistical information such as average, standard deviation, minimum, maximum and 95

C.1 Uniform Distributed Datasets

For the uniform distributed datasets, results of RBDA are close to lognormal distribution in many cases. Some examples of fitness of distributions are shown in figures.

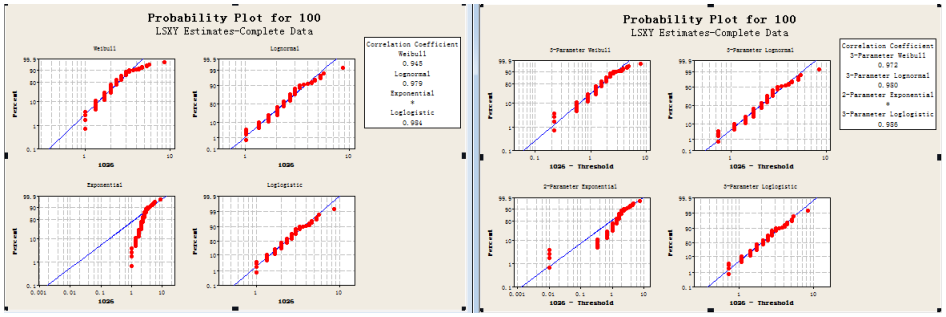

Fig. 8. RBDA distribution fitness graph of one dataset with 100 objects

According to the experiment results of uniform distributed datasets, the standard deviation (STD) of RBDA values of all objects are related with value of k. The STD varies from 1 to about 3 when k increases from 3 to 15. We observed

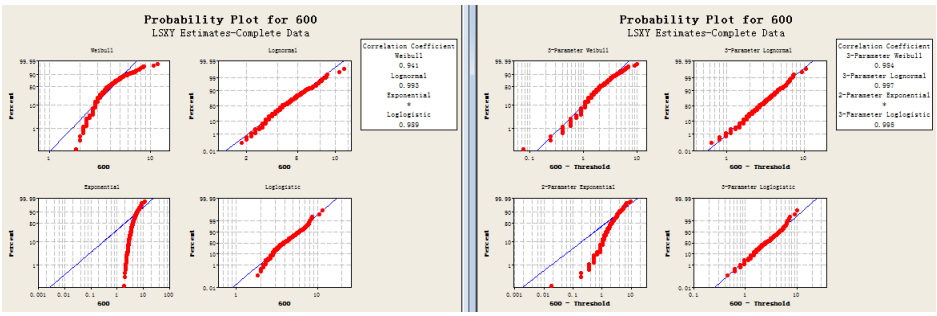

Fig. 9. RBDA distribution fitness graph of one dataset with 600 objects

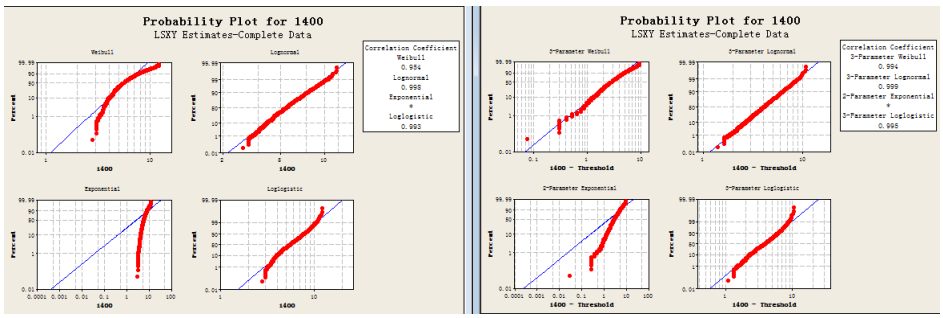

Fig. 10. RBDA distribution fitness graph of one dataset with 1400 objects

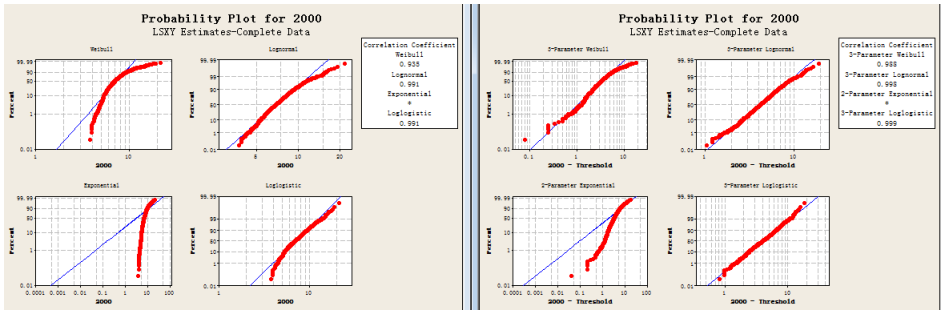

Fig. 11. RBDA distribution fitness graph of one dataset with 2000 objects

that with even with different size of dataset, the STD values are almost same with respect to same k values.

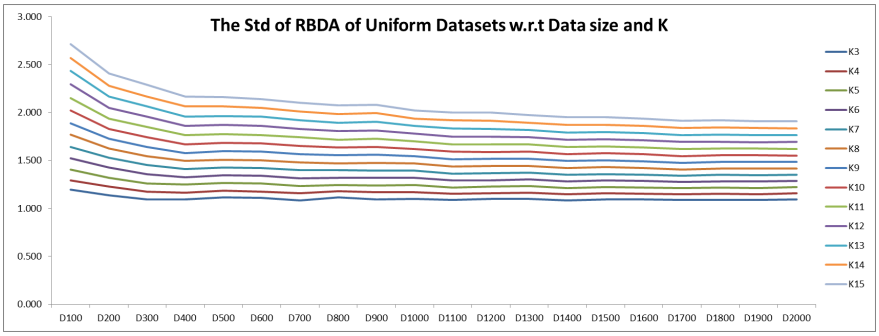

Fig. 12. The average standard deviations of values of RBDA of the datasets

The average values of RBDA of the datasets increase when k increases. For the same value of k, the average value of RBDA is almost a straight line when size of dataset is larger than 300.

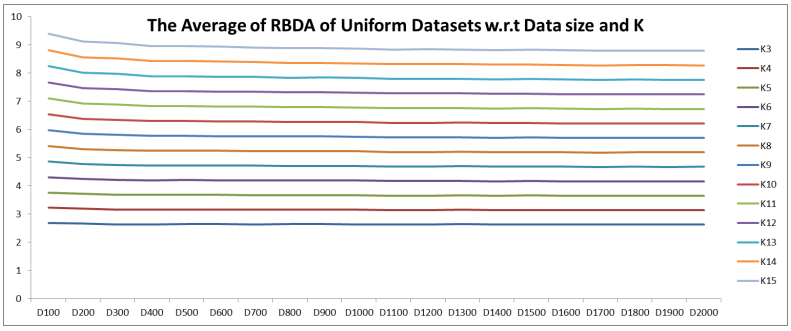

Fig. 13. The average of values of RBDA of the datasets

The regression analysis shows that the maximum value of RBDA is highly correlated with value of k, average value of RBDA and minimum value of RBDA. R square of this analysis is 0.99.

C.2 Gaussion Distributed Datasets

For Gaussian distributed datasets, the RBDA results show that distribution of RBDA is close to 3-parameter lognormal distribution. But we observe that RBDA deviates from lognormal distribution more and more from 95 % (accumulated percentage).

It can be seen that the maximum values of RBDA are higher than lognormal distributed values and the minimum values are lower than lognormal distributed

Max SUMMARY OUTPUT

Regression	
Multiple	0.997251
R Square	0.99451
Adjusted	0.994424
Standard	0.272536
Observati	260

ANOVA

	df	SS	MS	F	gnificance F
Regressic	4	3430.935	857.7338	11547.95	8E-287
Residual	255	18.94034	0.074276		
Total	259	3449.876			

	Coefficien	SE	t Stat	P-value	Lower 95%	Upper 95%	Lower 95%	Upper 95%
Intercept	7.882061	0.497998	15.82749	9.07E-40	6.901349	8.862774	6.901349	8.862774
Count	0.001057	4.33E-05	24.41873	1.06E-68	0.000972	0.001143	0.000972	0.001143
K	1.772908	0.166255	10.66376	3.39E-22	1.445499	2.100317	1.445499	2.100317
Avg	-2.48584	0.416104	-5.97408	7.74E-09	-3.30528	-1.6664	-3.30528	-1.6664
Min	1.3523	0.247757	5.45818	1.14E-07	0.864391	1.84021	0.864391	1.84021

Fig. 14. The results of regression analysis for the maximum value of RBDA of the dataset

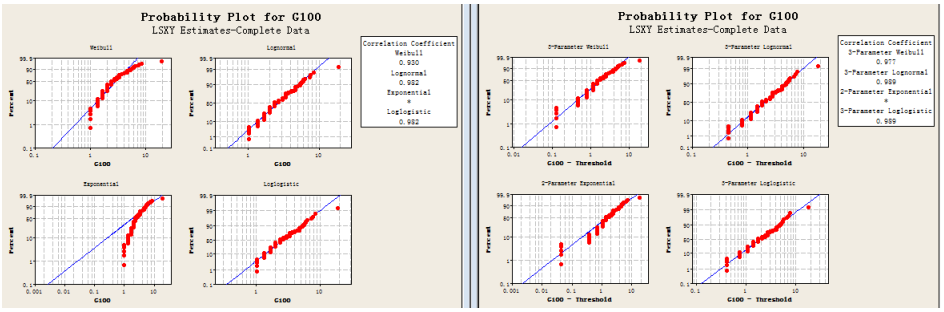

Fig. 15. RBDA distribution of fitness graph of one Gaussian dataset with 100 objects

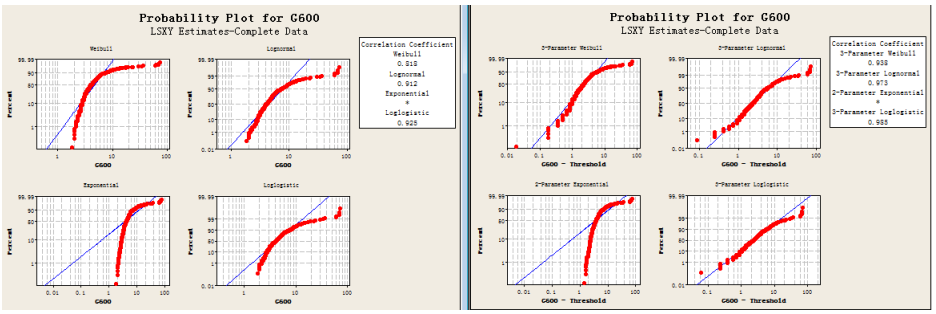

Fig. 16. RBDA distribution of fitness graph of one Gaussian dataset with 600 objects

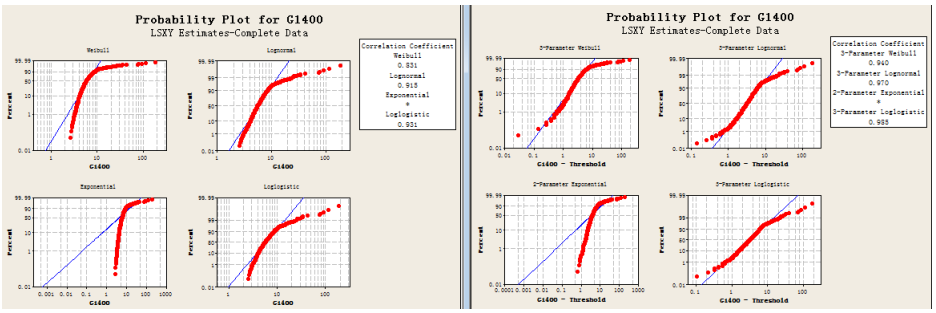

Fig. 17. RBDA distribution of fitness graph of one Gaussian dataset with 1400 objects

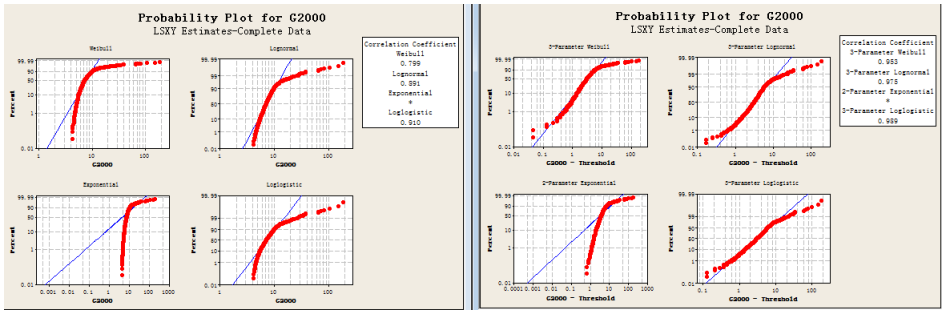

Fig. 18. RBDA distribution of fitness graph of one Gaussian dataset with 2000 objects

values. Since it is the natural fact of the RBDA, the Gaussian dataset will get the lower minimum RBDA and the higher maximum value than uniform distributed dataset.
The average and standard deviation of RBDA values of Gaussian distributed datasets show the different characteristic as those of uniform distributed datasets. The average and STD are increasing as k is increasing. The difference is that they also can be affected by the size of dataset. The average RBDA value with respect to k of 15 is changing from 12.9 in the dataset with 100 objects to 10.1 in the dataset with 2000 objects. In general, the average values of RBDA decrease when data sizes increase. The STD values here vary a lot and do not have a consistent trend compared with uniform dataset.

The regression analysis of outputs shows that the maximum value of RBDA is highly correlated with value of k, average value of RBDA and minimum value of RBDA. R square of this analysis is 0.92.

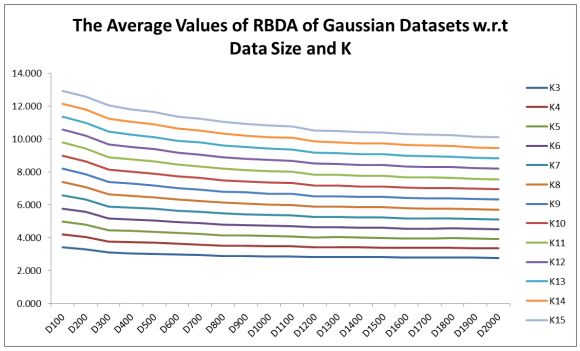

Fig. 19. The average values of RBDA of Gaussian datasets

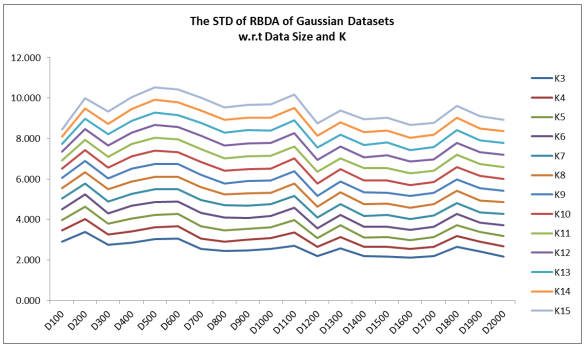

Fig. 20. The average standard deviation values of RBDA of Gaussian datasets

C.3 Regression Analysis for All Datasets

Combining the resutls of two different distribution datasets, we do regression analysis based on the size of dataset, the value of k, the average value of RBDA and the minimum value of RBDA. Our target variable is the maximum value of RBDA.

The results of regression show that the value of R square is only 0.63, which means it is not good enough.

C.4 Conclusion

The distribution of RBDA of a dataset is very close to lognormal in general, but its parts of large value and small value might deviate far from lognormal distribution according to the distribution of datasets. The regression analysis shows that we cannot predict the maximum value of RBDA precisely only based on the size of dataset , the value of k, the average value of RBDA and the

Max SUMMARY OUTPUT

Regression	
Multiple	0.962286
R Square	0.925994
Adjusted	0.924833
Standard	13.79047
Observati	260

ANOVA

	df	SS	MS	F	gnificance F
Regressic	4	606793.4	151698.4	797.6691	8.1E-143
Residual	255	48495.15	190.177		
Total	259	655288.6			

	Coefficien	SE	t Stat	P-value	Lower 95%	Upper 95%	Lower 95%	Upper 95%
Intercept	41.63601	8.565305	4.861007	2.04E-06	24.76826	58.50376	24.76826	58.50376
Count	0.015001	0.002811	5.336388	2.09E-07	0.009465	0.020536	0.009465	0.020536
K	41.42575	3.047448	13.59359	5.07E-32	35.42438	47.42712	35.42438	47.42712
Avg	-43.1279	4.427129	-9.74173	2.87E-19	-51.8463	-34.4095	-51.8463	-34.4095
Min	-7.35288	12.25878	-0.59981	0.549168	-31.4942	16.78846	-31.4942	16.78846

Fig. 21. The results of regression analysis for RBDA of Gaussian datasets

Max SUMMARY OUTPUT

回归统计	
Multiple	0.794727
R Square	0.631591
Adjusted	0.62873
标准误差	36.66557
观测值	520

方差分析

	df	SS	MS	F	gnificance F
回归分析	4	1186943	296735.7	220.7257	3.5E-110
残差	515	692347.3	1344.364		
总计	519	1879290			

	Coefficien	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	-110.797	6.687539	-16.5676	9.95E-50	-123.935	-97.6585	-123.935	-97.6585
Count	0.032964	0.003998	8.244331	1.39E-15	0.025109	0.040819	0.025109	0.040819
K	8.743572	5.697237	1.534704	0.125471	-2.44911	19.93626	-2.44911	19.93626
Avg	63.81506	2.54629	25.06198	3.23E-91	58.81266	68.81745	58.81266	68.81745
Min	-115.509	17.05542	-6.77259	3.46E-11	-149.016	-82.0026	-149.016	-82.0026

Fig. 22. The results of regression analysis for RBDA of all datasets

minimum value of RBDA. With current research results, it shows that we cannot use distribution approach to predict the threshold between outliers and non-outliers. More work need to be done in this area.

References

1. Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jorg Sander. Lof: Identifying density-based local outliers. *In Proceedings of the ACM SIGMOD International Conference on Management of Data. ACM Press*, pages 93–104, 2000.
2. Hui Cao, Gangquan Si, Yanbin Zhang, and Lixin Jia. Enhancing effectiveness of density-based outlier mining scheme with density-similarity-neighbor-based outlier factor. *Expert Systems with Applications: An International Journal*, 37(12), December 2010.
3. V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):ARTICLE 15, July 2009.
4. J. Feng, Y. Sui, and C. Cao. Some issues about outlier detection in rough set theory. *Expert Systems with Applications*, 36(3):4680–4687, 2009.
5. Wen Jin, Anthony K. H. Tung, Jiawei Han, and Wei Wang. Ranking outliers using symmetric neighborhood relationship. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 577–593, 2006.
6. J. Tang, Z. Chen, A. W Fu, and D. W. Cheung. Capabilities of outlier detection schemes in large datasets, framework and methodologies. *Knowledge and Information Systems*, 11(1):45–84, 2006.
7. Jian Tang, Zhixiang Chen, Ada Wai chee Fu, and David W. Cheung. Enhancing effectiveness of outlier detections for low density patterns. *In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 535–548, 2002.
8. Yunxin. Tao and Dechang Pi. Unifying density-based clustering and outlier detection. *2009 Second International Workshop on Knowledge Discovery and Data Mining, Paris, France*, pages 644–647, 2009.