Journal of Engineering and Applied Sciences 14 (23): 8699-8704, 2019
ISSN: 1816-949X

An Efficient Cluster based Outlier Detection Algorithm

M. Priya and M. Karthikeyan
Department of Computer and Information Science, Faculty of Science, Annamalai University,
608002 Annamalai Nagar, Tamil Nadu, India
mpriyaau@gmail.com

Abstract: Outlier analysis is becoming an important technique in data mining whose task is to identifying the data objects that are completely different from the majority of all objects. Outlier detection is necessary and useful with numerous applications in many fields like medical, fraud detection, fault diagnosis in machines, etc. In this study, we tend to propose a cluster based outlier detection algorithm which can be fulfilled in two stages. In the first stage we construct cluster using mutual nearest neighbor graph clustering algorithm. In the second stage we find the cluster outlier factor based on size of each cluster. The concept is to find outlier value of object and outlier clusters are extended to the formation of cluster. This algorithm is used to identify objects must confided as outlier object and outlier clusters in a database. This algorithm is based on the thought of mutual nearest neighbor graph clustering. The proposed algorithm can be used to identify the outlier value factor in the database and to detect the outliers and outlier clusters efficiently. The simulation result shows that the proposed method yields better results in outlier detection.

Key words: Data mining, outlier detection, cluster, outlier clusters, mutual nearest neighbor, fault diagnosis

INTRODUCTION

Data mining is mining knowledge from data that is the steps of knowledge discovery. The data mining is an essential step within the process of information discovery from large dataset by performing data cleaning, integration, selection, mining, pattern evaluation and knowledge presentation. The data mining involves three common tasks viz., association rule mining, clustering and classification. Outlier detection is a primary part within the fields of data mining and has attracted many researches. Outlier detection is one amongst the foremost necessary tasks in data analysis. Mining outlier has been extensively used in the fields like medical and public healthcare, master card fraud detection, fault diagnosis in machines, financial industry, quality control, weather prediction, pharmaceutical analysis, network robustness analysis, intrusion detection and etc. (Han *et al*., 2011).

Outliers are named as abnormalities, discordant, deviants, anomalies, etc. An outlier, according to Hawkins (1980) "an observation that deviates most from different observations on arouse that it had been generated by a unique mechanism". An outlier generally, contains helpful information about the abnormal characteristics of a system which can reflect the method of data generation. Outlier detection algorithms are developed for mechanically identifying valuable objects but irrelevant in the data. Once developing and applying outlier detection strategies, we have a tendency to use any one of the two methods, supervised and unsupervised. There are several outlier detection algorithms are proposed in the literature and they can be grouped into five categories viz., distribution-based methods, depth-based methods, distance-based methods, density-based methods and clustering-based methods.

Distribution based methods initiate from statistics, wherever observation is consider as an outlier, if it deviates too much from underlying distribution. Hence, distribution based methods is effective to identify the outliers, if we all know the distribution of the datasets (Barnett and Lewis, 1994). Depth based methods depends on the computation of various layers of k-d convex hulls. In this method, outliers are objects within the outer layer of those hulls (Johnson *et al*., 1998). Distance based methods suspects that outlier as an observation that is d_{min} distance faraway from p proportion of observations within the dataset. The problem is then finding suitable d_{min} and p, so that, outliers would be properly detected with a tiny number of false detections (Knorr *et al*., 2000). Outlier can be detected using the local density of observations. A low density on the observation is a sign of a possible outlier. Several density based outlier detection algorithms are proposed in the literature (Jin *et al*., 2006; Ha *et al*., 2014). Breunig *et al*. (2000)

Corresponding Author: M. Priya, Department of Computer and Information Science, Faculty of Science, Annamalai University, 608002 Annamalai Nagar, Tamil Nadu, India, mpriyaau@gmail.com

proposed Local Outlier Factor (LOF) that is one of the most commonly used approach in outlier detection. The LOF method is the concept of "reachability distance" between two objects, p and q that indicates the value of maximum distance between $d_k(p)$ and d(p, q) is applied for evaluate the density around an object. Every form of the outlier detection methods has its advantage and disadvantage. So that, it is possible to solve the problem of outlier cluster, we propose a novel outlier detection algorithm based on clustering method which can be done in two stages. First, we construct clusters using mutual nearest neighbor graph clustering algorithm. Then, we find the cluster outlier factor based on size of each cluster. This proposed algorithm can be used to identify the outlier value of a database and detect the outlier objects and outlier clusters efficiently.

Literature review: In data mining, outlier detection is an important technique and to get more and more attention. Cluster based outlier detection algorithms can solve the problems and several cluster based outlier detection algorithms are proposed and some of the existing research as discussed. Duan *et al.* (2009) developed a Cluster-Based Outlier detection algorithm (CBOF). The concept of CBOF is to calculate outlier factor of point p, denoted as CBOF(p), this method clustering the dataset and to compute CBOF based on the result of clustering. LDBSCAN finds the Local Outlier Factor (LOF) of each point of dataset (Duan *et al.*, 2007). Brito *et al.* (1997) presented an outlier detection algorithm (CMKNNG) based on the "Connectivity of the Mutual K-Nearest-Neighbor Graph". This algorithm is to develop the relationship between the connectivity of a mutual k-nearest-neighbor graph and the presence of clustering structure and outliers in the data. Jobe and Pokojovy (2015) proposed a cluster based outlier detection approach for multivariate data. Min (2015) proposed an efficient outlier detection algorithm based on data clustering over massive data. Ramaswamy *et al.* (2000) presented a method within which n largest kNN distances are denoted as outliers. This can be seen as "sparseness estimate" of a vector, in which the n sparsest vectors are identify as outliers.

Definitions: The following are some of existing work in outlier detection algorithm and definitions in the field of data mining.

Definition 1 (density based local outlier): Suppose given a point p in a dataset D is a density based local outlier, if LOF (p)>LOFLB, the lowest acceptable bound of LOF (Breunig *et al.*, 2000). The basic idea of LOF is to compare its local density with the local densities of its neighbors.

Definition 2 (core object): An object p is a core object with respect to LOFUB, if LOF (p)• LOFUB. Where LOFUB is upper bound parameter which can be set manually. If LOF (p) is enough small value, it means that object p is not an outlier and it must belong to some clusters. Therefore, it can be considered as a core object. Then LDBSCAN algorithm continues to expand from core object to all of the objects are visited. After clustering, CBOF captures the boundary between normal and outlier (abnormal) clusters.

Definition 3 (upper bound of the cluster based outlier): Let the dataset D, discovered the set of clusters C = {c_1, c_2, c_3, ..., c_k} by LDBSCAN in the sequence $|c_1|$• $|c_2|$•, ..., • $|c_k|$. The value of UBCBO in the cluster C_i, the number of the objects is • given parameter, if ($|c_1|+|c_2|+$, ..., $+|c_{(i-1)}|$• $|D|$*• and ($|c_1|+|c_2|+$, ..., $+|c_{(i-2)}|$• $|D|$*• . This definition gives quantitative measure to UBCBO. Consider that in the dataset, most of the data objects are not outliers, therefore, clusters that have a large part of data objects should not be considered as outliers.

Definition 4 (cluster based outlier): Suppose D is the dataset and the set of clusters C = {c_1, c_2, c_3, ..., c_k} discovered by LDBSCAN. The clusters that the number of the objects is no more than UBCBO are outlier clusters.

For example, in some cases the abnormal cluster deviated from a large cluster might contain more objects than a certain small normal cluster. In fact, due to spatial and temporal locality, it would be more proper to choose the clusters which have small spatial or temporal span as cluster-based outliers than the clusters which contain few objects. The notion of cluster-based outliers depends on situations. The cluster-based outliers are discovered and then compute CBOF, the cluster-based outlier factor.

Definition 5 (distance between two clusters): Let D is the dataset C_1 and C_2 are clusters in D. The distance between C_1 and C_2 is defined as Eq. 1 as follows:

$$\text{dist}(C_1, C_2) = \min\{\text{dist}(p, q) \mid p \in C_1, q \in C_2\} \quad (1)$$

Definition 6 (cluster-based outlier factor): Suppose C_1 be an outlier cluster and C_2 be the nearest non-outlier cluster of C_1. The cluster-based outlier factor of C_1 is defined as Eq. 2 as follows:

$$\text{CBOF}(C_1) = |C_1| * \text{dist}(C_1, C_2) * \sum_{p_i \in C_2} \frac{\text{lrd}(p_i)}{|C_2|} \quad (2)$$

Where:

|C| = The No. of the objects in cluster C

$\text{lrd}(p_i)$ = The local reachability density of object p_i

The outlier factor of cluster C_1 gets the degree to which we call C_1 an outlier cluster. However, one can see that the CBOF is computed by the above method and analysis based on clustering result. Hence, the clustering result is not suitable; the CBOF(p) is unrepresentative which is used to detect outliers. Then CBOF needs to improve by introductory the parameter • to discover the outlier clusters.

MATERIALS AND METHODS

In the proposed method, a cluster based outlier detection algorithm is introduced as follows:

- Classify the dataset into cluster using mutual nearest neighbor graph clustering algorithm
- Determine the clusters outlier factor based on size of each cluster

This algorithm can used to compute the outlier value of a dataset, to detect the outlier objects and outlier clusters.

Let D be a database, m and n be some objects in D and k is a positive integer to indicate the number of neighbors of each object.

Mutual Neighbor (MN): If m is the neighbor of n and n is the neighbor of m at the same time. Then, we call m is a mutual neighbor of n and similarly, n is a mutual neighbor of m that is.

Mutual Nearest Neighbor Graph (MNNG): Mutual nearest neighbor graph clustering can be constructed group by connecting each object to its mutual neighbors. The number of mutual neighbors of each object is k. However, it is possible that different objects have different number of mutual neighbors. Figure 1 shows that the simple example of mutual nearest neighbors clustering. Here the point a lies that the dense region to occupy more mutual neighbors than points that lies in the sparse region. Moreover, point b is a local outlier and does not have any mutual neighbors. In addition, from Fig. 1, we can see that the point a will be rise up, if the value of k is increased, then the numbers of mutual neighbors are also increased. The point that is greater than one and less than k value is considered as an outlier cluster, the point b is a local outlier object.

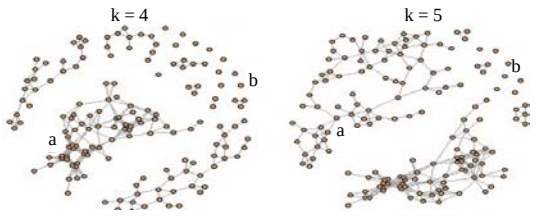


Fig. 1: The example of mutual nearest neighbors (k = 4 and k = 5)

To detect out the outlier clusters, first we want to cluster the datasets. In this study, we propose a clustering algorithm to construct a rough cluster based on mutual nearest neighbors graph clustering algorithm as shown in algorithm 1 and to detect the outlier objects and outlier clusters based on outlier detection algorithm as shown in algorithm 2.

Algorithm 1; Clustering algorithm:

Input: Cluster (D, k)//D as dataset and k as number of mutual neighbors
Output: The clustering results CL = {cl_1, cl_2, cl_3, ..., cl_n}
Step 1: Constructing the mutual nearest neighbor graph (n = 1)
Step 2: Select an object a randomly
reached(a) = true
cl_n = cl_n MNN(a)
Step 3: While exist b cl_n and reached (b)• true then
Reached (y) = true
cl_n = cl_n MNN(b)
Step 4: If exist c• D and reached (c)• true, then n = n+1
Step 5: Go to Step 2

The mutual nearest neighbor graph based clustering algorithm has been constructed. A large complex cluster may be divided into two or more clusters using algorithm 1. By setting the parameter k is small. Once a cluster is normal, even it comes from a large cluster based on k value. The size of this cluster must be less than the value of k as formed as an outlier cluster. Figure 2 shows that normal and outlier clusters based on the output of the clustering algorithm 1. There are five clusters named c_1, c_2, c_3, c_4 and c_5 where the cluster c_3, c_4 and c_5 are normal cluster, c_1 and c_2 may be outlier cluster and remaining are outlier objects. The cluster c_1 and c_2 should be outlier cluster because the size of cluster must be lesser than other clusters as well as k value. In this study, after the clustering the datasets, we propose another algorithm, called outlier detection algorithm which is based on the size of outlier clusters as normally much lesser than the

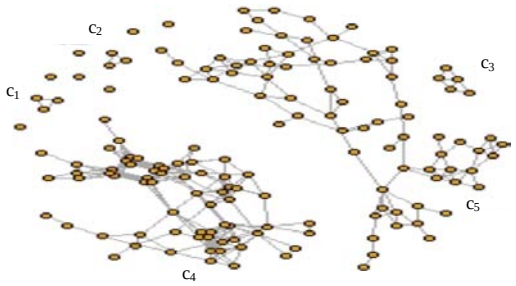


Fig. 2: Normal clusters and outlier clusters

normal clusters. Assume that the sequence set of clusters $C = \{c_1, c_2, c_3, ..., c_k\}$ such that $|cl_1| \bullet |cl_2| \bullet, ..., \bullet |cl_n|$. The changed state level of cl_k is defined as the ratio of size of cl_k and size of cl_{k+1} where k = 1, 2, ..., n-1. Hence, we define the Nearest Cluster Outlier Factor (NCOF) of cluster cl_k denoted as $NCOF(cl_k)$ is defined as Eq. 3 as follows:

$$NCF(cl_k) = \frac{e^{\frac{|cl_{k+1}|}{|cl_k|^2}} - 1}{e^{\frac{|cl_{k+1}|}{|cl_k|^2}}} = 1 - e^{\frac{-|cl_{k+1}|}{|cl_k|^2}},\ k = 1, 2, ..., n-1 \quad (3)$$

The $NCOF(cl_k)$ value ranged as (0, 1). Largest value of $NCOF(cl_k)$ indicates that cl_1, cl_2, cl_3, ..., cl_k are outlier clusters. If $NCOF(cl_k)$ is related to cl_{k+1}, then we find the correct outlier clusters. Let, cl_1, cl_2, cl_3, ..., cl_n are normal clusters of the database D identified by algorithm 1. If $NCOF(cl_p) = \max\{NCOF(cl_k)\}$ and $NCOF(cl_p) > 0.1$ then cl_1, cl_2, cl_3, ..., cl_p are outlier clusters. That is the scope of value p is [0, n-1]. In algorithm 1, we find out the clusters, if all the clusters are normal clusters then the p-value = 0 and the NCOF is small. Now, we can prove that if $NCOF(cl_p) < 0.1$ or $|cl_{p+1}|/|cl_p|^2 < 0.1$ which implies that the size of outlier cluster. Then the outlier cluster size can be small changes from cl_p to cl_{p+1}. Therefore, the value $NCOF(cl_p) < 0.1$, if, there are no outlier clusters.

Let D is a dataset, cl_1, cl_2, cl_3, ..., cl_n are clusters and cl_1, cl_2, cl_3, ..., cl_k are outlier clusters then, we compute outlier factor of dataset by using Eq. 4:

$$\text{Outlier Factor}(OF) = \frac{\sum_{k=p+1}^{n} |cl_k|}{|D|} \quad (4)$$

The outlier value is the outlier percentage and outlier clusters in a database. The significance of outlier value is used in any outlier detection algorithm. The proposed outlier detection algorithm is shown below.

Algorithm 2; NCOF algorithm:

Input: $\{cl_1, cl_2, cl_3, ..., cl_n\}$ and k value
Output: Outlier Cluster $OC = \{cl_1, cl_2, cl_3, ..., cl_p\} \backslash\backslash p = 1, 2, ..., n-1$
Step 1: for all $cl_k \bullet cl$
if size $(cl_k) < p$ then cl_i is outlier cluster and detected from cl
Step 2. for k = 1 to size $(cl_k)-1$
Compute $NCOF(cl_k)$
Step 3. Find p-value that $NCOF(cl_p) = \max\{NCOF(cl_k)\}$
Step 4. If $NCOF(cl_p) < 0.1$ then p = 0
Step 5. Calculate outlier factor
Step 6. Finally we get outlier factor and outlier clusters $OC = \{cl_1, cl_2, cl_3, ..., cl_p\}$

Performance evaluation: In order to show the effectiveness of the proposed method, performance evaluation based on real life datasets were studied. Powers (2007) proposed an evaluation measures for evaluating results of machine learning experiments. For performance evaluation of the algorithms, we use metrics, namely recall, precision and F-measure to evaluate the detection results. Then the metrics are defined as follows.

Recall: It can be determined by the completeness that was properly known to the total number of positive cases.

$$\text{Recall}(R) = \frac{TP}{TP+FN} \quad (5)$$

Where:
TP (True Positive) = The No. of objects that are correctly classified as outliers by an algorithm
FN (False Negative) = The No. of wrongly classified as negative classes

Precision: Precision is used to determine exactness. It is the ratio of the predicted positive cases that were accurate to the total number of predicted positive cases:

$$\text{Precision}(P) = \frac{TP}{TP+FN} \quad (6)$$

Where:
TP (True Positive) = The No. of objects that are correctly classified as outliers
FP (False Positive) = The No. of objects that are wrongly classified as outliers

The precision and recall, maximum value is 1 and the minimum value is 0. The maximum value of precision and recall is the better the results of outlier detection.

F-measure: A measure that combines precision and recall is the harmonic mean of precision and recall. It can be computed by using Eq. 7:

$$F_{Measure} = 2 * \left[\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}\right] \quad (7)$$

RESULTS AND DISCUSSION

To illustrate the practicability of this method, an real life data set is considered and identified the outliers. The evaluation of performance of the proposed method for the real world datasets which contains different characteristics. The datasets have normal cluster patterns and also sparse outliers.

Diabetic dataset is shown in Fig. 3. The data set which has been used for the records of diabetic diagnosis. The dataset contains 370 sample of data medical records

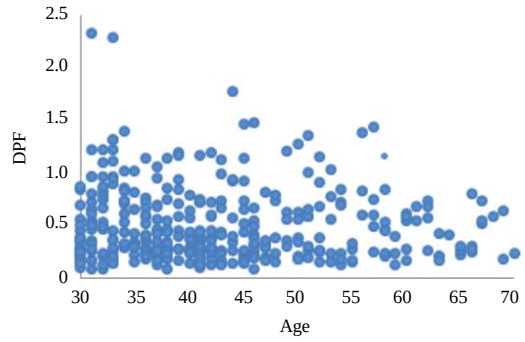


Fig. 3: Diabetic dataset

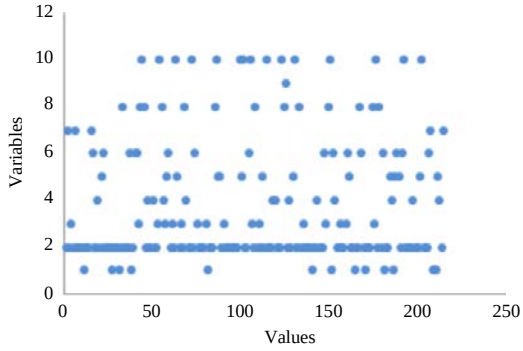
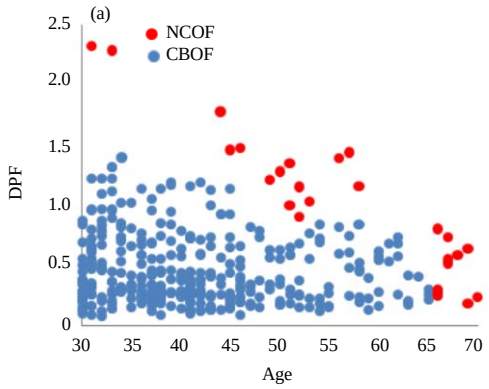


Fig. 4: Breast cancer dataset



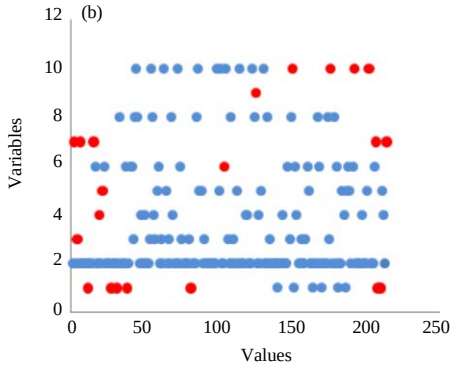


Fig. 5: Detection results for NCOF: a) Diabetic dataset and b) Breast cancer dataset

Table 1: Results corresponding the proposed method

Datasets	K	Outlier factors
Diabetic	6	0.22
Breast cancer	5	0.44

(objects). Each record contains 9 attributes which are considered as the diagnosis factors of diabetic. The output class variable labeled as 0 or 1 that is class value 1 is for diabetes and 0 is for non-diabetes. The total of 370 objects in diabetic dataset, of which 28 objects are outlier clusters or outlier objects.

Breast cancer dataset are shown in Fig. 4. This dataset has been used for recording the measurements of breast cancer cases. The data set contains the dimensions of 214 data sample medical records (objects). Each record contains 10 attributes which are considered as risk factors for the occurrence of cancer. There are two classes labeled as 0 and 1, to diagnosis of cancerous and non-cancerous. The total of 214 objects in this dataset, of which 20 objects are outlier clusters or outlier objects.

The experimental results for the proposed method are shown in Table 1. As the experiments on real datasets, NCOF finds the largest value of NCOF(cl_p) that p is the boundary between outlier clusters and normal clusters. Here, the proposed NCOF algorithm successfully detect

Table 2: Measurement metrics results

Datasets/methods	Recall (R)	Precision (P)	F-measures
Diabetic			
CBOF	0.95	0.95	0.95
NCOF	0.96	0.97	0.96
Breast cancer			
CBOF	0.95	0.94	0.95
NCOF	0.98	0.97	0.98

out outlier clusters in both datasets. Moreover, the NCOF algorithm produce the value of Outlier Factor (OF) in diabetic dataset is 0.22 and breast cancer dataset is 0.44.

Figure 5 shows the detection results for the NCOF algorithm. The outlier clusters and outlier objects detected by the proposed method are coloured in red in the experiments.

After detection of the outliers, we tend to evaluate the performance metric measurements for the datasets. The performance evaluation measurements are shown in Table 2. From the results of Table 2, we see that in the diabetic dataset, R and P as 0.95 of CBOF, then R = 0.96 and P = 0.97 of NCOF. In the breast cancer dataset, R = 0.95 and P = 0.94 of CBOF, then R = 0.98 and P = 0.97 of NCOF. Moreover, the F-measure value of NCOF for diabetic dataset as 0.96 and the breast cancer dataset as

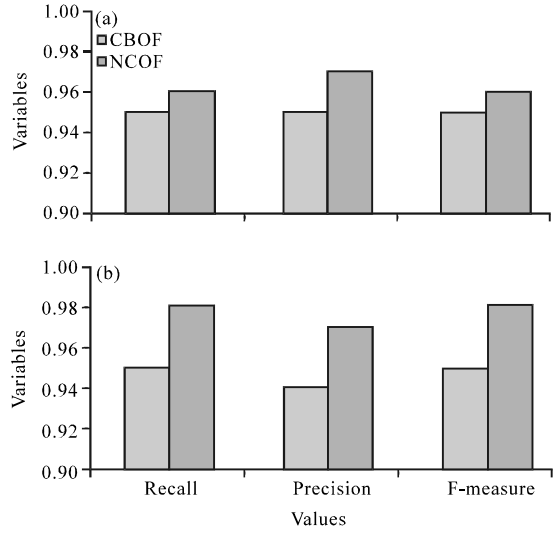


Fig. 6: Measurement metrics: a) Diabetic dataset and b) Breast cancer dataset

0.98 are the maximum value in this experiment. Figure 6 shows the graphical represents of measurement metrics for the datasets.

CONCLUSION

In this study, cluster based outlier detection algorithm called NCOF has been proposed and observed that its performance method. The proposed cluster based outlier detection algorithm and the NCOF detect the outlier object and outlier clusters in a better way than other outlier detection algorithms. The algorithm NCOF has to take k value which is the number of nearest neighbors. Moreover, NCOF can compute the outlier factor value of a database and detect the outlier objects and outlier clusters efficiently. Therefore, the proposed algorithm perform better to detect the outlier objects and outlier clusters and to specify the number of outliers in a dataset than the earlier methods. Experimental results show that the proposed method yields better results in outlier detection. The algorithm discussed is study may be any set of data the applicability of outlier detection techniques in various data mining applications, like medical and public healthcare outlier detection, credit card fraud detection, fault diagnosis in machines, financial industry, quality control, weather prediction, pharmaceutical research, network robustness analysis etc.

REFERENCES

Barnett, V. and T. Lewis, 1994. Outliers in Statistical Data. 3rd Edn., Wiley, New York, USA., ISBN-10: 0471930946, pp: 604.

Breunig, M.M., H.P. Kriegel, R.T. Ng and J. Sander, 2000. LOF: Identifying density-based local outliers. Proceedings of the International Conference on Management of Data, May 15-18, 2000, Dallas, TX., USA., pp: 93-104.

Brito, M., E. Chaves, A. Quiroz and J. Yukich, 1997. Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection. Stat. Probab. Lett., 35: 33-42.

Duan, L., L. Xu, Y. Liu and J. Lee, 2009. Cluster-based outlier detection. Annal. Oper. Res., 168: 151-168.

Duan, L.,L. Xu and F. Guo, 2007. A local-density based spatial clustering algorithm with noise. Inform. Syst., 32: 978-986.

Ha, J., S. Seok and J.S. Lee, 2014. Robust outlier detection using the instability factor. Knowl. Based Syst., 63: 15-23.

Han, J., M. Kamber and J. Pei, 2011. Data Mining: Concepts and Techniques. 3rd Edn., Morgan Kaufmann Publishers, USA., ISBN-13: 9780123814791, Pages: 744.

Hawkins, D.M., 1980. Identification of Outliers. 1st Edn., Chapman and Hall, London, New York, ISBN-13: 9780412219009.

Jin, W., A.K.H. Tung, J. Han and W. Wang, 2006. Ranking Outliers Using Symmetric Neighborhood Relationship. In: Advances in Knowledge Discovery and Data Mining, Ng, W.K., M. Kitsuregawa, J. Li and K. Chang (Eds.). Springer, Berlin, Germany, ISBN-13: 9783540332060, pp: 577-593.

Jobe, J.M. and M. Pokojovy, 2015. A cluster-based outlier detection scheme for multivariate data. J. Am. Stat. Assoc., 110: 1543-1551.

Johnson, T., I. Kwok and R.T. Ng, 1998. Fast computation of 2-dimensional depth contours. J. Computation Depth, 1: 224-228.

Knorr, E.M., R.T. Ng and V. Tucakov, 2000. Distance-based outliers: Algorithms and applications. VLDB. J. Intl. J. Very Large Data Bases, 8: 237-253.

Min, J.K., 2015. An efficient outlier detection algorithms based on data clustering over massive data. Database Res., 31: 59-71.

Powers, D.M.W., 2007. Evaluation: From precision, recall and F-factor to ROC, informedness, markedness & correlation. MSc Thesis, School of Informatics and Engineering, Flinders University, Adelaide, Australia.

Ramaswamy, S., R. Rastogi and K. Shim, 2000. Efficient algorithms for mining outliers from large data sets. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data Vol. 29, May 15-18, 2000, ACM, Dallas, Texas, ISBN:1-58113-217-4, pp: 427-438.