PACIS 2016 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

Summer 6-27-2016

OUTLIER DETECTION VIA MINIMUM SPANNING TREE

Xin Tang
Xi'an Jiaotong University, tangxin.900616@stu.xjtu.edu.cn

Wei Huang
Xi'an Jiaotong University, waynehw21st@gmail.com

Xue Li
The University of Queensland, x.li@uq.edu.au

Shengli Li
Xi'an Jiaotong University, lishengli@mail.xjtu.edu.cn

Yuewen Liu
Xi'an Jiaotong University, liuyuewen@mail.xjtu.edu.cn

Recommended Citation

Tang, Xin; Huang, Wei; Li, Xue; Li, Shengli; and Liu, Yuewen, "OUTLIER DETECTION VIA MINIMUM SPANNING TREE" (2016). *PACIS 2016 Proceedings*. Paper 211.
http://aisel.aisnet.org/pacis2016/211

OUTLIER DETECTION VIA MINIMUM SPANNING TREE

Xin Tang, Center of Data Science and Information Quality, School of Management, Xi'an Jiaotong University, Xi'an, Shaanxi, China, tangxin.900616@stu.xjtu.edu.cn

Wei Huang, Center of Data Science and Information Quality, School of Management, Xi'an Jiaotong University, Xi'an, Shaanxi, China, waynehw21st@gmail.com

Xue Li, School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, Australia, x.li@uq.edu.au

Shengli Li, Center of Data Science and Information Quality, School of Management, Xi'an Jiaotong University, Xi'an, Shaanxi, China, lishengli@mail.xjtu.edu.cn

Yuewen Liu, Center of Data Science and Information Quality, School of Management, Xi'an Jiaotong University, Xi'an, Shaanxi, China, liuyuewen@mail.xjtu.edu.cn

Abstract

In the big data era, analysis with data sets becomes more and more important. How to obtain valuable information from the data records is all we care about. However, most of the time, there are outliers among the data records. Outliers can lead to wrong information extracted from the data sets, detecting them can help us modify these rules or get them easier. In this paper, we combine the distance-based and clustering-based outlier detection methods, use the theory of minimum spanning tree and standard normal distribution to define a new method of outlier detection. At the same time, our algorithm can find the data records which we should pay attention to in the data sets. The algorithm works with two phases. During the first phase, we build a minimum spanning tree by all data records, compute the average weight and the standard deviation of it. In the second phase, we use the distance of each data record with its K nearest neighbours to discover the outliers. Experimental results show our algorithm is more accurate and efficient.

Keywords: *Distance-based Outlier Detection, Clustering-based Outlier Detection, Minimum Spanning Tree, Standard Normal Distribution.*

1 INTRODUCTION

Outlier is a concept which is originally produced from the statistics. In statistics, an outlier is an observation that is numerically distant from the rest of the data (Grubbs 1969). In a more general setting, outliers are the data points which deviate so much from other points in the same dimensionality. It means those data points which has different properties from others. As the outlier detection is an important application, it has been widely used in practical domains, such as fraud detection for credit cards, insurance and tax (Bolton& David 2002), disease outbreaks in the medical field (Wong& Moore& Cooper& Wagner 2002), and the gene analysis (Li C, Wong W H 2001; Ying X& Olman V& Dong X 2002). In recent years, outlier detection has also been used in internet security, weather forecast, air quality test, traffic, analytical finance (Sheng B& Li Q& Mao W& et al 2007) and so on (Hodge& Austin 2004; Chandola& Banerjee& Kumar 2009).

Therefore, the outlier detection has become a popular field of research, and many different techniques have been proposed. At present, there are six main methods, such as statistical tests, depth-based, deviation-based, distance-based, density-based, clustering-based approaches (Krogel 2010). But there is no unique way to suit for all the detection requirements. The statistics tests, depth-based and deviation-based methods relay on the statistical models. They find out outliers based on statistical and probabilistic data analysis. The distance-based and density-based methods utilize the models based on spatial proximity. More specifically these kinds of methods depend on the data points to their nearest neighbours to judge the outliers from the normal ones. The clustering-based approach divides the data points by their features into clusters of similar ones and defines the outliers that are the kinds of clusters with a small number of data points. There are also many other different types of outlier detection algorithms that have been proposed. However, they are beyond the scope of study and will not be discussed here.

In this paper, we propose a simple but effective outlier detection method which is based on the normal distribution of the edges in the minimum spanning tree (MST) of a given data set. It is built upon both the distance-based and clustering-based outlier detection algorithm. In general, our outlier detection has two phases. In the first phase, with data records as the vertices and Euclidean distance between them as the weight made a Euclidean minimum spanning tree (EMST) is. Then the average weight μ and the standard deviation δ are computed from the edges of EMST. In the second phase, for each vertex, we compute the Euclidean distance to its K nearest neighbours and then calculate the mean distances (M_i) of them. Next we use the formula of Normal Distribution to Standard Normal Distribution to standardize each M_i (explain in section 4). Finally, the result with the standard deviation δ is used as the threshold to judge which point is an outlier.

Our main contributions can be summarised as follow:
(1) A new method of outlier detection is proposed which combines the advantages of statistical tests, distance-based and clustering-based approaches.
(2) Sufficient experimental evaluations on different datasets and comparison with other state of the art outlier detection algorithms demonstrate that our method is an effective way to detect outliers.
(3) With different precision adjusted, by using different threshold values, our method can allow users to detect outliers according to different demands.
(4) Most algorithms of outlier detection just care about which point is different from others and may the outlier. As a by-product, in our algorithm, we not only pay attention to the traditional outliers but also discuss about the points which have similar features.

The rest of the paper is organized as follows. In section 2, we review some related works on statistical tests, distance-based, density-based and clustering-based outlier detection algorithms. In section 3, our proposed method is presented. In section 4, the analyses of our algorithm and some comparative experiments are given. Also, we use the results to evaluate the performance and present the practical value of our algorithm. In the end, the conclusions are made and future works are discussed in section 5.However, our algorithm has some advantages, we need further research to extend the algorithm to high dimensional data sets and to adapt our method of threshold select more efficiently of data sets to suit for clustering propose to different fields.

2 RELATED WORK

There are four approaches of outlier detection related to our research, namely, statistical test, distance-based, density-based, and clustering-based outlier detection technique. Every approach will be discussed in the follow.

2.1 Statistical Tests Outlier Detection

The concept of outlier starts from statistics. Its basic assumptions are that normal data records follow a known distribution and occur in a high probability region of this model. But outliers are data records which deviate strongly from this distribution. Assuming a given data set satisfies a certain statistical distribution, the parameters of the model statistics can be computed from the data records. For example, the mean and standard deviation value of Normal Distribution. The outliers are records that have a low probability to be generated by the overall distribution. However, the number of models is limited. Using the certain kinds of statistical distribution to detect outliers is meaningful only in some areas. As a result, other kinds of outlier detection methods have been proposed recently.

2.2 Distance-Based Outlier Detection

One of the most popular approaches of outlier detection is to use the distances of a data point to its K nearest neighbours, giving rise to distance-based outlier detection technique. Proposed by Knorr and Ng, they gave a definition of outliers like this, "An object O in a dataset T is a distance-based outlier, denoted by DB(p, D)-outlier, if at least a fraction p of the objects in T lies greater than distance D from O, where the term DB(p, D)-outlier is a shorthand notation for a Distance-Based outlier (detected using parameters p and D)"(Knorr& Ng 1998).Given a distance measure on a feature space, there are many different definitions of distance-based outlier detection. Three most popular definitions are:
(1)Given a real number *d* and an integer *p*, a data item is an outlier if there are fewer than *p* other data items within distance *d* (Knorr& Ng 1998; Knorr& Ng 1999).
(2)Given two integers, *n* and *k*, outliers are the data items whose distance to their *k-th* nearest neighbor is among top *n* largest ones (Knorr& Ng& Tucakov 2000).
(3)Given two integers, *n* and *k*, outliers are the data items whose average distance to their *k* nearest neighbours is among top *n* largest ones (Ramaswamy& Rastogi& Shim 2000; Angiulli& Pizzuti 2002).

Three definitions are explaining the distance-based outlier detection from different angles. The first one sets a fixed distance *d*, and a fixed number of point *p*, within this distance. But it does not give a ranking. The other two notions both give a ranking and the number of outliers there may exist. But in practice, it is very sensitive to K and hard to say how many outliers there are in a data set at most time.

2.3 Density-Based Outlier Detection

The general idea of density-based outlier detection is to compare the density around a data point with that around its nearest neighbours. Density-based outlier detection techniques work well for detecting outliers in datasets which contain one or more clusters with similar density. However, for many real world datasets which have complex structures in the sense that different portions of a database can exhibit very different characteristics, they might not be able to find all interesting outliers (Wang& Li Wang& Wilkes 2012). A classic algorithm which called Local Outlier Factor (LOF) (Tang& Chen& Fu& Cheung 2002) is proposed by Breunig et al. in 2000.

2.4 Clustering-Based Outlier Detection

Clustering is the means to partition data records into different clusters by their features. The data records in small clusters are regarded as the outliers. However, these methods have a major problem (Duan L& Xu L& Liu Y& Lee J 2008). We often need to know how many clusters will exist in advance. Besides, it is not clear how small a cluster can be treated as an outlier scale, but it is difficult to judge which cluster is among the small ones.

3 OUR DB-MST ALGORITHM

From the work presented in the last section we can observe that outliers are closely related with their K nearest neighbours and meaningful to the whole data records. The method considers both the distances of ever record to their nearest neighbours and the distances in an MST over the whole data records. We call it distance-based with minimum spanning tree outlier detection algorithm or DB-MST.

A minimum spanning tree is an undirected graph without close cycle. The MST of a weighted graph is very suitable to use to cluster the points. Removing the longest edge of the MST can partition the points into two clusters. Given a data set, compute the Euclidean distance between each point in a data set. The distance will be the weight of an edge to construct the minimum spanning tree. Assuming the edge weights of the MST follow a normal distribution which is given by Equation (3.1) and (3.2). Then the average weight μ and the standard deviation δ of edges in the entire MST can be computed. Next, we compute the average Euclidean distances M_i, of K nearest neighbours for every point, Pi. To standardize M_i, we calculate μ_i by taking M_i into Equation (3.3) to fit the standard normal distribution.

Normal Distribution:

$$\mathrm{F(x)} = \frac{1}{\sqrt{2\pi}\delta} e^{-\frac{(x-\mu)^2}{2\delta^2}} \tag{3.1}$$

Where x is dependent variable, μ is the mean or expectation of the normal distribution, σ is its standard deviation, e is the Euler's number.

Standard Normal Distribution:

$$\mathrm{F(y)} = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \tag{3.2}$$

Where y is dependent variable, μ is the mean or expectation of the standard normal distribution, σ is its standard deviation, e is the Euler's number.

Normal Distribution Standardization:

$$\mu i = \frac{Mi-\mu}{\sigma} \tag{3.3}$$

Where M_i is dependent variable, μ is the mean or expectation and δ is standard deviation of the MST which we have built.

In a standard normal distribution, the ratio of absolute value within a standard deviation (δ=1) is about 68%; the ratio of absolute value within two times standard deviation (δ=2) is more than 95%; and the ratio of absolute value within three times standard deviation (δ=3) close to is 99%. When K=1, the μ_i is satisfying with this rule completely, and our method is almost the same as the distance-based outlier detection did.

The data point P_i whose μ_i is greater than the standard deviation of the EMST (μ_i>1) is regarded as the outlier. If the μ_i is a negative number and the absolute value of μ_i is greater than the standard deviation of the EMST (μ_i<-1) can we suggest that this data point is very close to its neighbours can be clustering, and this data point should be paid attention to. The threshold does not need to be the standard deviation, it can be *T* times of the standard deviation, and often depends on the actual demand. The resulting outliers and the data points that should be paid attention to are fewer when *T* is large; and the results are more if *T* is small.

```
Algorithm: DB-MST
Input: a data set S
Output: the outliers of the input data set R
Initialize: K $\leftarrow$ 4 //the number of the considered nearest neighbours of every point
        T $\leftarrow$ 1//the multiple of the standard deviation of the standard normal//distribution
Construct an EMST from S
Compute the average $\mu$ distance of all the edges in the EMST
Compute the standard deviation $\delta$ distance of all the EMST
For each Pi $\in$ S
            Find K nearest neighbours $Q_j$ of $P_i$
            $M_i \leftarrow M_i$ + the distance between P and $Q_j$
            $M_i \leftarrow M_i / K$      //average Euclidean distance $M_i$ of K nearest neighbours for
            $\mu_i \leftarrow (M_i - \mu) / \delta$   //standardized value of every point $P_i$
            Rank the $\mu_i$
            If $\mu_i > T$ Then  $P_i$ is an outlier
            Elseif $\mu_i < -T$ Then
                  $P_i$ should be paid attention
```

The advantage of our algorithm is the accuracy and efficiency. In other words, the EMST contains all data records, the outliers are judged by the average weight μ and the standard deviation δ of edges in the EMST, our algorithm considers about the global outlier detection. At the same time, it takes advantage of the importance of nearest neighbours. Most of the algorithms running time are the time used to construct the MST again and again. Our method just needs to build the MST one time and propose new threshold to find outliers.

4 EXPERIMENTAL RESULTS

We tested our algorithm on number of datasets of different distributions. We found that the performance of detection was excellent, especially for irregular distributions. In the first part of this section, the result of our outlier detection algorithm is analyzed. In the second part, we show the comparative experiments between our algorithm and other classical outlier detection algorithms such as DB (Ramaswamy& Rastogi& Shim 2000; Angiulli& Pizzuti 2002), MST, COF (Breunig& Kriegel& Ng & Sander 2000) and LOF to explain the advantages of our algorithm.

4.1 Algorithm Analysis

In our algorithm, there are two important parameters K and T which should be initialized. K means how many nearest neighbours to be considered here. T represents the criteria of outlier threshold. First, we fix T to be 1 and change K form 2 to 8, the results are presented in Figure 1.

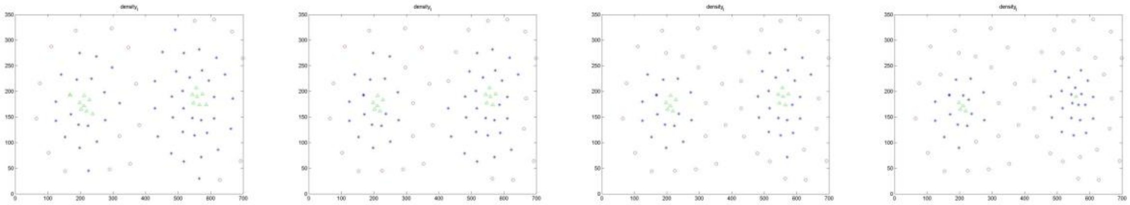

Figure 1. K=2, 4, 6, 8 T=1

With K increases while fixing T=1, more outliers show up while the number of points which should be paid attention to become less. The more nearest neighbours are considered, the larger the average distance of data points is. Fixed T=1, the greater μ_i, the less the data points whose μ_i is smaller than -T.

Second, we fix K to be 4 and change T from 0.6 to 1.2, the results are given in Figure 2. From the figure we know that with T increases, the numbers of outliers and the points that should be paid attention to both decrease when the K is unchanged. With K fixed, μ_i is fixed, the lager the T is, the smaller the detected outliers and corresponding data points that attention should be paid to.

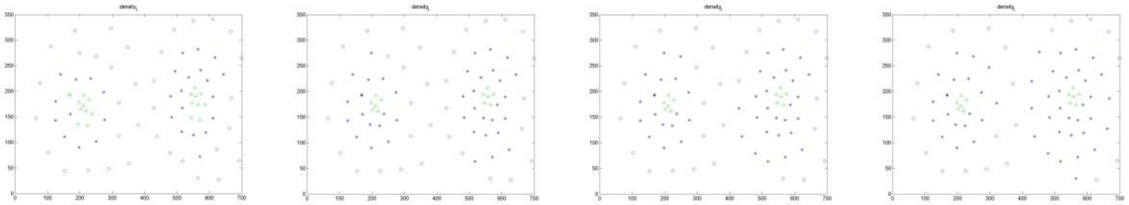

Figure 2. K=4 T=0.6, 0.8, 1.0, 1.2

From what has been discussed above, about the results, we observe that when K=4 and T=1, the results of our outlier detection algorithm are better. We next conduct do the experiments on several relatively difficult datasets.

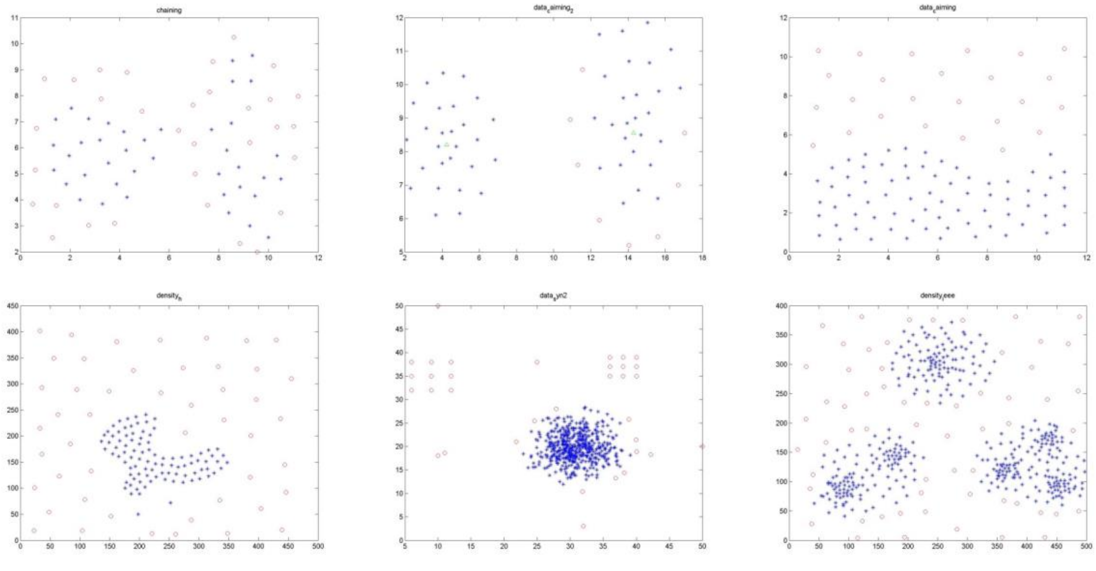

Figure 3. Three different groups of data sets

In Figure 3, we show results of three different groups of data sets. The first group consists of two clusters on the first row. Our algorithm can find the outliers around each cluster accurately, but it is hard to find the points which should be paid attention to when the two clusters are too closed. It is because the standard deviation δ of the MST of the data set on the left becomes smaller lead to a larger μ_i. For the data sets on the second row, over all, they display the two clusters in a dataset, our algorithm also can distinguish them. But there are two data points in the wrong cluster in the right picture. The reason is that these two data points are too close to the points in the other cluster than the others. The two data sets on the last row are more common in the sense that they show the clusters with regular or irregular outliers. From the figures we can see our algorithm also can detect them accurately.

4.2 Metrics for Measurement

To evaluate the performance of the algorithms, three metrics were selected called Precision and Recall. Assuming that in a given outlier detection algorithm, we identify m most suspicious data points in data set which contains dt true outliers and let mt be the number of true outliers among m data points.
Then, precision, which measures the proportion of true outliers in the top m suspicious instances, is Precision = mt/m
And recall, which measures the accuracy of an algorithm, is Recall = |mt|/|dt|

4.3 Comparative Experiments

In our algorithm, we consider both the distance of every data record's nearest neighbours and the whole data records in the data set. The performance of it is better than other classical distance-based, density-based and clustering-based outlier detection algorithm.

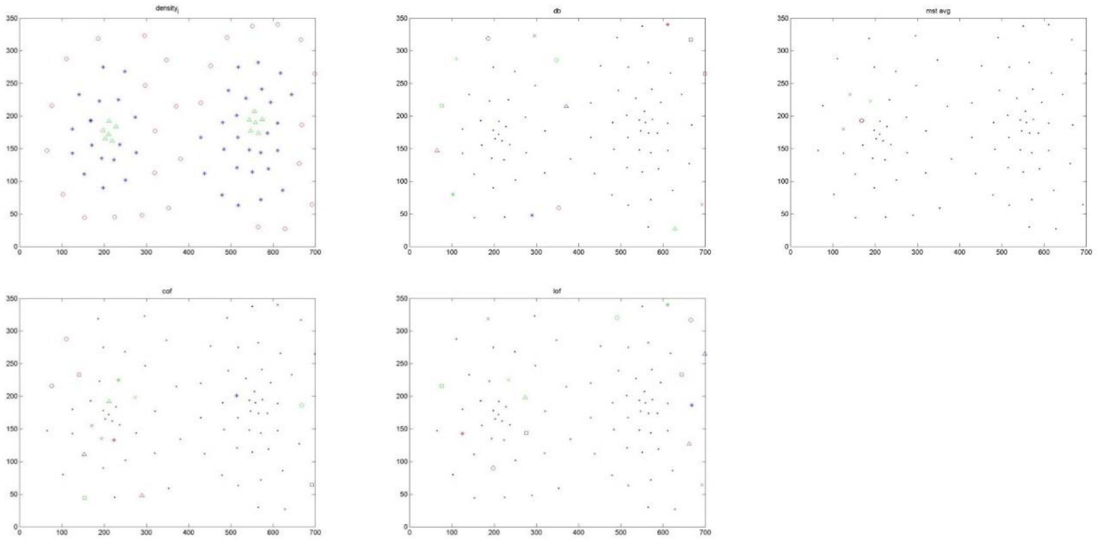

Figure 4. Detection results of our algorithm

In Figure 4, we show the detection results of our algorithm with respect to four other state-of-the-art outlier detection methods. But MST, COF and LOF are not useful here. The other four algorithms require the number of outliers to be specified. From the above figure, it can be seen that our algorithm is the most effective algorithm.

In table 1, we summarize the performances of the algorithms, where N represents the number of data points in the data set, Pre represents Precision and Rec represents Recall. From the table 1, it infers that the accuracy of our algorithm is better than others.

	N	m	mt	dt	Pre	Rec
MST-DB	84	28	23	25	0.82	0.92
DB	84	25	22	25	0.88	0.88
MST	84	25	2	25	0.08	0.08
COF	84	25	9	25	0.36	0.36
LOF	84	25	13	25	0.52	0.52

Table 1. The performances of different algorithms

From the table 1, we identify m most suspicious data points in data set which contains dt true outliers and let mt be the number of true outliers among m data points. Then, precision, which measures the proportion of true outliers in the top m suspicious instances, is Precision = mt/m, and recall rate which measures the accuracy of an algorithm, is Recall = |mt|/|dt|, we find that our algorithm MAT-DB is more efficient and accurate, compared with other four algorithms, MST-DB and DB algorithm are the most precise algorithm of all, but the recall rate is better than DB algorithm.

5 CONCLUSIONS

In this paper, we propose a simple but effective outlier detection method which is based on the normal distribution of the edges in the minimum spanning tree (MST) of a given data set. It is built upon both the distance-based and clustering-based outlier detection algorithm. we explored a new method of outlier detection which based on MST and standard normal distribution. Our approach has advantages including:
(1)DB-MST can find the outliers accurately in the two-dimensional space which other algorithm cannot find;
(2)DB-MST do not need a predefined number of outliers, it use a newly proposed threshold value to identify the outlying data records;
(3)Our algorithm can find the data records which should be paid attention to. These data records are the center of the cluster and they have more value to mine farther.

Our future work will address:
(1)How to extend our algorithm to high dimensional data sets;
(2)How to adapt our method of threshold select more efficiently of data sets to suit for clustering propose to different fields;
(3)The research of data records which should be paid attention will be continued, we considered that these data records can make the analysis of the data set simpler and more accurate.

References

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. Techno metrics 11, 1–21.

Bolton, R. J., & David, J. H. (2001). Unsupervised profiling methods for fraud detection. Proc Credit Scoring & Credit Control VII, 5--7.

Wong, W. K., Moore, A., Cooper, G., & Wagner, M. (2004). Rule-based anomaly pattern detection for detecting disease outbreaks. Proceedings of National Conference on Artificial Intelligence, 2002, 217--223.

Li, C., & Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proceedings of the National Academy of Sciences, 98(1), 31-36.

Ying, X., Olman, V., & Dong, X. (2002). Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees. Bioinformatics, 18(4), 536-45.

Krogel, P. (2010). Outlier detection techniques. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining . ACM.

Knorr, E. M., & Ng, R. T. (1998). Algorithms for Mining Distance-Based Outliers in Large Datasets. Proceedings of the 24rd International Conference on Very Large Data Bases (pp.392--403). Morgan Kaufmann Publishers Inc..

Knorr, E. M., & Ng, R. T. (1999). Finding intensional knowledge of distance-based outliers. Vldb, 211-222..

Knorr, E. M., Ng, R. T., & Tucakov, V. (2000). Distance-based outliers: algorithms and applications. Vldb Journal — the International Journal on Very Large Data Bases, 8(3-4), 237-253.

Ramaswamy, S., Rastogi, R., & Shim, K. (2000). Efficient algorithms for mining outliers from large data sets.. Acm Sigmod Record, 29(2), 427-438.

Angiulli, F., & Pizzuti, C. (2002). Fast Outlier Detection in High Dimensional Spaces. Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery (Vol.2431, pp.15-26). Springer-Verlag..

Duan, L., Xu, L., Liu, Y., & Lee, J. (2008). Cluster-based outlier detection. Annals of Operations Research, 168(1), 151-168.

Sheng, B., Li, Q., Mao, W., & Jin, W. (2007). Outlier detection in sensor networks. Mobihoc, 306 - 309.

Hodge, V.J., Austin, J. (2004). A Survey of Outlier Detection Methodologies. Artificial Intelligence Review 22, 85–126.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: a survey.. Acm Computing Surveys, 41(3), 75-79.

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J&#. (2000). Lof: identifying density-based local outliers.. Acm Sigmod Record, 29(2), 93-104.

Tang, J., Chen, Z., Fu, W. C., & Cheung, D. W. (2002). Enhancing Effectiveness of Outlier Detections for Low Density Patterns. Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (Vol.2336, pp.535-548). Springer-Verlag.

Wang, X., Li Wang, X., & Wilkes, D. M. (2012). A minimum spanning tree-inspired clustering-based outlier detection technique. Advances in Data Mining Applications & Theoretical Aspects, 7377, 209-223.