Research Journal of Applied Sciences, Engineering and Technology 9(4): 257-261, 2015
DOI:10.19026/rjaset.9.1402
ISSN: 2040-7459; e-ISSN: 2040-7467
© 2015 Maxwell Scientific Publication Corp.
Submitted: July 18, 2014 Accepted: October 17, 2014 Published: February 05, 2015

Research Article

An Effective Pruning based Outlier Detection Method to Quantify the Outliers

[1]Kamal Malik, [2]Harsh Sadawarti and [3]G.S. Kalra
[1]MMICT and BM, MMU, Mullana, Haryana,
[2]RIMTIET (Affiliated to Punjab Technical University)
[3]Lovely Professional University, Punjab, India

Abstract: Outliers are the data objects that do not conform to the normal behaviour and usually deviates from the remaining data objects may be due to some outlying property which distinguishes them from the whole dataset. Usually, the detection of outliers is followed by the clustering of the dataset which sometimes ignores the prominency of outliers. In this study, we have tried to detect the outliers and pruned the clustering elements initially so that the outliers can be prominently highlighted. We have proposed an algorithm which effectively prunes the similar data objects from the large datasets and its experimental results compare the neighbouring points and show the better performance than the existing methods.

Keywords: Clusters, distance-based, pruning

INTRODUCTION

Outlier detection is one of the very important aspects of the data mining. Outliers are the data objects or the points that do not comply with the normal behaviour of the datasets. The important applications of the data mining include data cleaning, fraud detection stock market analysis, intrusion detection marketing, network sensors etc. To find the suspicious or erroneous pattern is to find out the outliers. There are many approaches used by the researchers for the outlier detection which may be classified as supervised that includes the exploitation of the training data set of the normal and abnormal objects, semi supervised which includes only normal examples and unsupervised which searches the unlabelled data set to detect the outliers without giving the proper reason for having their outlying property. The problem of detecting the outliers has been extensively studied in the statistics community (Barnett and Lewis, 1994; Angiulli *et al*., 2006). Distance-Based techniques are very popular for relating each pair of objects in the data set. There are so many methods for defining the outliers from the perspective of distance based outliers methods which are based on the concepts of local neighbourhood of K-Nearest Neighbours (KNN) of the data points (Angiulli and Pizzuti, 2005). The notion of the distance based outlier methods is that the user might not have the idea about the underlying assumptions of the data and usually generalizes the concepts from the distribution methods. Moreover, distance based methods are usually very easy to scale up for the higher dimensional data to detect the projected outliers. There are several metrics that are used to measure the outlierness of the data points. In this study, we have tried to find the various outliers and analyse them using the clustering and distance function. The notion behind it is first to prune the points of nearest neighbourhood of the centroid i.e., the points which are of similar properties are collected in a different dataset and are pruned and the remaining points which are outliers are taken into consideration. So, without considering the points of the cluster, only the outliers are taken into the account which reduces the computations and complexity as a whole. Moreover, the ODF i.e., outlier defining factor is used to provide the proper ranking or the measure of outlierness of an outlier.

The main aim of this study includes three very important points:

- The detection of the subset S from the input population which are homogenous in nature that is the points which are quite similar to each other.
- Detecting the outliers which crosses the threshold of their inner radius and are at much distance from the centroid.
- Thirdly, to quantify the outliers using the outlierness defining factor, so as to detect the deviations and differences from the cluster.

MATERIALS AND METHODS

Distance based outlier techniques were first of all introduced by Tucakov *et al*. (2000) and Knorr and Ng (1998). According to them-An object p in a data set DS

Corresponding Author: Kamal Malik, MMICT and BM, MMU, Mullana, Haryana, India

is a DB (q,dist) - an outlier if at least fraction of the objects in DS lie at the greater distance from p, it can generalize the several statistical tests. Then Ramaswamy *et al*. (2000) proposed the extension of the above method as they proposed a notion that all the outlier points are ranked based on the outlier score. Moreover, Angiulli and Pizzuti (2005) proposed the way of ranking the outliers by considering the whole neighbourhood of the objects. In this case, the points are ranked on the sum of the distances from the k-nearest neighbours, rather than considering individual distance from centroid. Then Breunig *et al*. (2000) also proposed the local outlier factor to indicate the degree of outlierness in each outlier. They were first to quantify the outliers and used the term local outlier factor because only neighbourhood of each object is taken into account. It was a density method and has the stronger capability in a sense that a data object is gathered by how many members of its neighbourhood that decides its density. More the data object is denser, lesser the probability of being its outlier. Then, Zhang *et al*. (2009), proposed the local distance based outlier detection method which is known as ldof function, which has its overall complexity as O (N^2), where N is the no. of the points in the data set. Moreover, there are many clustering algorithms like DBSCAN (Ng and Han, 1994), CURE (Guha *et al*., 1998), BIRCH (Zhang *et al*., 1996) etc, to detect the outliers, but the limitations of these clustering algorithms is that they may optimize the clustering of various data objects but they do not optimize the detection of the outliers which is our prime priority. In this study, we have used the pruning based algorithm PLDOF (Pamula *et al*., 2011), which is actually pruning out the data items which are similar and only the outliers are detected and taken into consideration, we have tried to extend this study by using ODF function in order to make pruning more effective by enhancing the inner radius which will be quantified according to their deviations.

In our study, we have used two measures LDOF (Zhang *et al*., 2009) and PLDOF in which former indicates how much a point is deviating from its neighbours and is a probable candidate of outliers and later prunes out the data items of clusters and focuses only on the outlier candidates only. The factor ldof is calculated as:

Ldof of m: The local distance based outlier factor of m is defined as:

$$\text{Ldof (m)} := \frac{dm}{Dm} \qquad (1)$$

d_m is the KNN of m. If N_m is the set of k- nearest neighbours of object m. Let dist (m, n)≥0 be the distance measure between objects m and n. The k nearest neighbour distance of the object m is:

$$d_{m:} = \frac{1}{k}\sum_{q\varepsilon Nm} dist(m,n) \qquad (2)$$

D_m is the kNN inner distance of m. Inner distance D_m is defined as:

$$D_m = \frac{1}{k(k-1)}\sum_{q\, q'\varepsilon Nm, q\neq q'} dist(q,q') \qquad (3)$$

Based upon this ldof, PLDOF i.e., pruning based local outlier detection measure was proposed (Zhang *et al*., 1996). The main idea underlying the pruning based algorithm is to first cluster the complete data set into clusters and then the points which are not the outliers are pruned out.

Effective pruning based local outlier detection method (Proposed method): In this section we will describe our proposed method which is a further improvement over ldof and pldof. The main idea of the effective pruning based outlier detection is to effectively prune the data items which are basically the part of the clusters and only consider the outliers which are deviating from the nearest neighbours. In this algorithm, the pruning is increased effectively by increasing the threshold distance due to which only the genuine outlier candidate are taken into account. Moreover, the KNN inner distance is also decreased due to which the percentage of pruned data items is enhanced. We briefly describe the steps that we need to perform by our pruning based algorithm:

- **Generation of clusters:** Initially, K-Means algorithm is used to cluster the entire data set into c clusters and then the radius of each cluster is calculated. If any of the clusters contains less no. of points than the required no. of outliers, then the radius pruning will not be done for that cluster.
- **Pruning the points with the increased threshold:** Firstly, calculate the distance of each point from the centroid of the cluster and half the distance of the radius from the centroid so as to enhance the threshold value and then effectively prune out the elements which are the part of the clusters and the outliers are taken into consideration.
- **Quantify the outliers:** All the outliers detected are then quantified using ODF function which is based upon ldof in such a way that a numerical value is associated with each outlier and hence the top n outliers are taken into consideration. Hence, the outlier points are computed in an effective way using ODF function.

The complexity of the K-Means algorithm is c*it*N where *c* is the no. of clusters to be formed, *it* is the no of iterations and N is the no. of the data points. Total computations in our method are $c*it*N+c*n_p+(x+N)^2$ where n_p represents average no. of points in each cluster and x indicates the fraction of

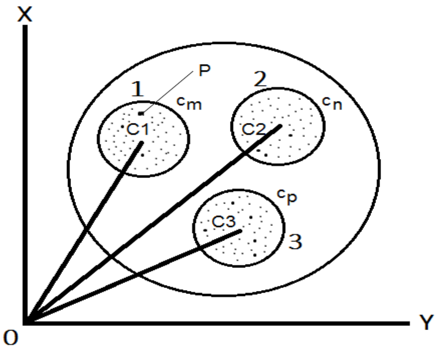

Fig. 1: Shows a dataset with three clusters with their centroids and inner radius

the data points after pruning, which depends upon the threshold value as in our case. As the outliers are very less in our case because most of the data points which are the cluster candidates are pruned out due to which the value of x is very small and hence the complexity is very much reduced even from $O(N^2)$ (Fig. 1).

Outlier metric-outlier defining functions: All the arbitrary points are divided into three discrete clusters c_i, c_j, $c_{k.}$ Then the distance of the centroids c_1, c_2, c_3 of the clusters c_i, c_j, c_k respectively is calculated from the origin using the ordinary distance formula of coordinate geometry. Then we define a metric known as ODF i.e., Outlier Defining Factor which defines and quantifies the outlierness of the points. The higher value of the ODF indicates and tells that how much a point is deviating from its neighbours and probably it can be an outlier and provides its ranking among all the points.

We have considered the unsupervised way of learning here because we want that the outliers should be quantified well enough even though we do not have the training classes with us. To consider that the point p is an outlier or not, consider the first cluster, its distance dist (c_1, c_m) where c_m, c_n, c_p are the points on the cluster circumference or the farthest point on the cluster. c_i, c_j, c_k, respectively:

$$\text{dist}\ (c_1,c_m) = \text{Radius}/\ 2 = R_{0.5} \tag{4}$$

Now, dist $(c_1, p)<R_{0.5}$, then p is an outlier, but if dist $(c_1, p) > R_{0.5}$, then p is not an outlier and ODF is not calculated for it. If the later holds, then the data item is pruned out, otherwise the point p is termed as an outlier and the ODF has to be calculated using Euclidean or simple distance formulae of co-ordinate geometry. Similarly, various points of cluster c_i are taken and their corresponding distances are calculated and then the mean of all the points are taken and is termed as D(p) where,

$$\text{ODF} = D(p) = 1/k \sum_{p\in ci} dist(c1,p) \tag{5}$$

This ODF function is based on ldof and hence it provides the information that how much a point is deviating from its neighbours but with a difference that we have calculated the inner radius $R_{0.5}$ and every point will be compared with it in order to decide whether it is an outlier or not. In case of outliers, its ODF will be calculated otherwise the data objects will be pruned out.

Algorithm 1: Effective Pruning Outlier Detection Algorithm

```
Dataset: X, iteration: no. of loops, cluster_ no, assump_outlier
Step 1. Set X← K Means(c, k, S)
    for i=0 to cluster_no_1
Do
    Cluster_ centre $_i$ =Point $P_i \in$ Dataset
Done
End for
While (iteration---------)
Do
For i=0 to Dataset – 1
Do
If (dist_mean$_j$[i] < dist_mean$_{j+1}$[i] <dist_mean$_{j+2}$[i])
  cluster_no[i] = j
    cluster $c_j$← Point $P_i$
    endif
    ∀ Cluster $c_i$
do
  cluster_centre $_i$ =$\sum_{j=1}^{n}$ Pj ∈ Ci / Ci
done
    ∀ Cluster ($c_i$)
Do
Radius$_i$ ← Radius ($c_i$)
Do
If ($c_i$.elemnt>assump_outlier)
For j= 0 to $c_i$.elemnt-1
Do
If (dist ($P_i \in c_i$, cluster_centre) < = (radius*0.8)
Prune ($P_i$)
Else
Move $P_i$ to resultant_ cluster
End if
Done
Else
∀ point Pi ∈ $C_i$
Do
Move $P_i$ to resultant _cluster
Done
End if
Done
∀ Point $P_i$, T is the resultant cluster
Do
ODF ($P_i$) // Applying the ODF function
Find n points with higher ODF values and the desired outliers. According to the Implementation, the graph is plotted for the desired outliers and cluster points.
Done
Done
```

Table 1: Effective pruning ratio for WDBC data set

	Percentage of data pruned		Precision		
K	PLDOF	EPLDOF	LDOF	PLDOF	EPLDOF
10	57.18	60.08	0.4	0.48	0.482
20	55.52	60.02	0.7	0.72	0.78
30	55.52	58.18	0.8	0.80	0.79
40	53.59	57.23	0.8	0.80	0.82
50	54.14	57.02	0.8	0.80	0.80
60	54.14	57.73	0.8	0.80	0.80

RESULTS AND DISCUSSION

In this section, we have compared the outlier detection performance of our Effective Pruning Based Outlier Detection method with the PLDOF and LDOF methods.

WDBC (Medical Diagnosis Data): To validate our experiment, we have used the medical data set WDBC, (diagnosis) from UCI repository which has already been further used by the nuclear feature extraction for the Breast Cancer Diagnosis. This data set contains 569 medical diagnosis records, each with 32 real valued input features. This diagnosis is Binary i.e., cancer data can be Benign or Malignant. We assume that the objects labelled as benign are normal data whereas the malignant are considered as abnormal one or an outlying data. In our experiment, initially, we use all the 360 Benign Diagnosis records as normal objects and added five malignant records in them as outliers. This process is repeated a no. of times by varying the values of the neighbourhood objects. Every time, the value of the neighbourhood size i.e., k is varied (may be increased or decreased). The three measures that are highly affected by varying the neighbourhood size k are:

- The n- top potential outliers
- The detection of the precision
- The percentage of the data pruned

In order to verify this effect, we will repeat this experiment 10 times by adding the random no. of outliers (Malignant cancer data) every time. With the help of the various independent runs, say from 10 to 60, the average detection precision is calculated and both the top n- outliers and the percentage of the pruned data are varied.

In EPLDOF method as shown in Table 1, the percentage of the pruned data is more as compared to the PLDOF. This is basically because of the ODF function that we have used in our EPLDOF algorithm. Due to the increase in the threshold value for the inner radius of the clusters, more data is pruned out and when the data gets much more pruned, the time and the space complexity are much more suppressed and the outliers will be more prominently highlighted. Moreover, when the precision is compared with LDOF and PLDOF, EPLDOF reaches at par at k = 30 even though the 60% of the data is pruned initially. Hence the time complexity and computation time are further decreased. This comparison is shown in the Table 1.

CONCLUSION

In this study, we have proposed an effective and an efficient outlier detection algorithm which is based on already existing methods LDOF and PLDOF but with

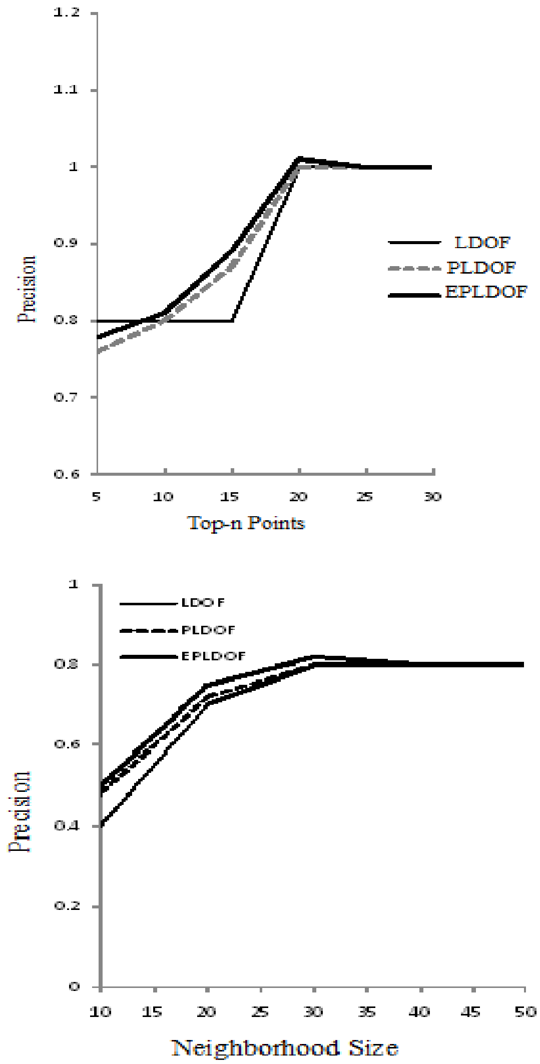

Fig. 2: Graphical comparison of LDOF, PLDOF and EPLDOF

the major difference that the pruning has been done very effectively using ODF function and hence the accuracy of EPLDOF is more as compared to PLDOF and LDOF as shown in Fig. 2 which is a graphical comparison of these three. All the points which are not the probable candidates of the outliers are pruned out and the remaining points which are termed as outliers are taken into consideration and are further quantified for their outlierness. The time and the space complexities are drastically reduced and hence the computation cost is also suppressed than the previously existing methods and the accuracy is quite high even though our pruned data is comparatively higher.

REFERENCES

Angiulli, F. and C. Pizzuti, 2005. Outlier mining in large high-dimensional data sets. IEEE T. Knowl. Data En., 17: 203-215.

Angiulli, F., S. Basta and C. Pizzuti, 2006. Distance-based detection and prediction of outliers. IEEE T. Knowl. Data En., 18(2): 145-160.

Barnett, V. and T. Lewis, 1994. Outliers in Statistical Data. John Wiley and Sons, New York.

Breunig, M.M., H.P. Kriegel, R.T. Ng and J. Sander, 2000. LOF: Identifying density-based local outliers. SIGMOD Rec., 29(2): 93-104.

Guha, S., R. Rastogi and K. Shim, 1998. CURE*:* An efficient clustering algorithm for large databases. SIGMOD Rec., 27(2): 73-84.

Knorr, E.M. and R.T. Ng, 1998. Algorithms for mining distance-based outliers in large datasets. Proceeding of 24th International Conference on Very Large Data Bases (VLDB, 1998), pp: 392-403.

Ng, R.T. and J. Han, 1994. Efficient and effective clustering methods for spatial data mining. Proceeding of the 20th International Conference on Very Large Data Bases (VLDB, 1994). Santiago, Chile, pp: 144-155.

Pamula, R., J.K. Deka and S. Nandi, 2011. An outlier detection method based on clustering. Proceeding of 2nd International Conference on Emerging Applications of Information Technology, pp: 253-256.

Ramaswamy, S., R. Rastogi and K. Shim, 2000. Efficient algorithms for mining outliers from large data sets. Proceeding of the 2000 ACM SIGMOD International Conference on Management of Data (SIGMOD '00), pp: 427-438.

Tucakov, V., E.M. Knorr and R.T. Ng, 2000. Distance-based outliers: algorithms and applications. VLDB J., 8(3-4): 237-253.

Zhang, K., M. Hutter and H. Jin, 2009. A new local distance-based outlier detection approach for scattered real-world data. Proceeding of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD ’09), pp: 813-822.

Zhang, T., R. Ramakrishnan and M. Livny, 1996. Birch: An efficient data clustering method for very large databases. SIGMOD Rec., 25(2): 103-114.