International Journal of Hybrid Information Technology
Vol.8, No. 5 (2015), pp. 129-134
http://dx.doi.org/10.14257/ijhit.2015.8.5.14

Clustering Outlier Detection Algorithm

Huangtao[1] and Tan Yanna[2]

[1]*Harbin University of science and Technology, Harbin, China*
[2]*China United Network Communications Corp Harbin branch, Harbin, China*
ht1124@163.com

Abstract

Outlier detection and clustering technologies are an important branch of data mining, such as combining the two technologies can improve the mining significance. In this paper, both clustering and outlier detection can be the starting point, proposed a DBSCAN-LOF algorithm is the core idea is to use k_ neighbors thought, DBSCAN redefine the core of the object, making the only non-core objects LOF The operation, thereby reducing the original LOF algorithm is computing the number of global objects, and makes no DBSCAN algorithm input parameters Eps. Real and simulated data sets by experimental results confirm that the algorithm to improve the operating efficiency and the LOF algorithm DBSCAN clustering effect, and while producing clustering and outlier detection results.

Keywords: *Data mining, clustering, DBSCAN, outlier, LOF algorithm*

1. Introduction

Hawkins gives the essence of outliers: outlier is an observation point, it deviates from the large extent of other observation points, and then it is suspected that the observation point generated by different mechanisms [1]. The reason that outliers are produced may be artificially measurement process error, execution error or special events occur. In some applications, from the perspective of knowledge discovery point, those things what rarely happen often will be more interesting than what often happens, then it can be more research value, for ten thousand normal recording is likely to cover only one rule, and ten outlier is likely to mean ten different rules[2]. Currently, outlier detection is becoming a hot topic in the field of machine learning researchers, databases, statistics and so on.

At present, some outliers can be found from clustering methods, such as CLARANS, DBSCAN, BIRCH, STING[3] and so on, just as a byproduct of clustering outlier identification process, after identifying outlier ,for making the impact of outlier is minimized, simply delete or set aside outliers, makes some important hidden information is lost. In outlier detection, because of the distribution of data is unclear, making analysis of the causes and sources of these outliers are not clear enough, so in many cases outliers have lost its practical value. The outlier detection algorithm combined with clustering analysis, the meaning of data mining can be proposed. In this paper, the DBSCAN and LOF algorithm are combined by k- neighbors ,then the DBSCAN-LOF algorithm is proposed, the basic idea is the introduction of DBSCAN clustering technology in the LOF, using the core meaning of the object [4], the data pruning to reduce original LOF algorithm computation times, improve the efficiency of LOF algorithm; and makes DBSCAN algorithm without input parameters Eps[5], DBSCAN algorithm reduces the sensitivity of the parameters Eps, so that the algorithm is more suitable for those who dataset unevenly distributed.

In recent years, in outlier detection algorithm， the outlier detection method which is based on clustering is appeared, the clustering technology is fused into outlier detection, in some extent, the outlier detection efficiency is improved [6],

ISSN: 1738-9968 IJHIT

and the data mining significance is improved. The algorithm can be defined as a data object is a cluster-based outlier, if the data object does not belong to any cluster. Roughly it is divided into the following two methods.

A use of clustering techniques to detect outlier's method is discarded away from other clusters of small clusters [7]. This method can be used in conjunction with any clustering technique, but requires a small distance between the cluster and the other clusters and the minimum cluster size threshold. Thus, the process can be simplified to drop smaller than the minimum size of a cluster. In this way the number of the selected cluster is highly sensitive to outliers and it is difficult to attach an object to the data point degree. As in [8], the first cluster using one-pass algorithm divides the data set to be almost the same radius of the hierarchy data, and then calculate the value of each cluster OF, and sorted, and finally through the other small clusters the threshold value of the distance between clusters is determined to obtain the outlier clusters. Because of this method stepwise clustering and outlier detection, making the results of outlier detection is constrained by the clustering effect.

Another more systematic approach is as follow, firstly all data objects is clustered, and then assess the extent of the data objects belonging to the cluster [7]. For the prototype-based clustering techniques, you can use the data object to its distance from the cluster center to measure the extent of the data objects which is belong to the cluster. In addition, the clustering technology which is based on the objective function, the size of the objective function can be used to assess the extent of the data objects that belong to any cluster. As the literature [8] proposed a computational algorithm to reduce the LOF method, it introduces the idea of micro-cluster-based local outlier detection density, finding the Top-n which is most likely to become data objects outliers, reducing the amount of computation for each object LOF calculation to some extent. But it requires the user to enter additional clustered micro maximum radius, reducing the reliability of the algorithm.

2. Clustering Outlier Detection Algorithm DBSCAN-LOF

Definition1. k- Radius. For p∈D, data objects neighborhood radius k of p, it is defined that the object's distance from the neighborhood to distance p, the formula is as follows:

$$k\_radius(p) = \frac{\sum_{o \in N_k(p)} dist(p,o)}{|N_k(p)|}$$

Where p represents a direct distance of the object with the object o, and p represent the object of k_ neighbor neighborhood, said the object p is the number of k-nearest neighbor neighborhood object (except p itself outside) of.

Highly dependent on the parameters determining DBSCAN algorithm core objects, and when the data distribution is sparse, it is difficult to accurately determine the core object; the paper redefines the concept of core object.

Definition 2 core object (core point), right p∈D, if p is the core of the object, there $k\_radius(p) \le \frac{\sum_{o \in N_k(p)} k\_radius(o)}{|N_k(p)|}$

Intuitive, one data point is at the core objects around it k- nearest neighbor density-related, so this article will define the core object of k neighborhood radius of its core object p k- little neighborhood k objects the mean radius.

2.1. The Description of DBSCAN-LOF

Based on the idea described above, and the definitions of non-core objects, the thought of DBSCAN-LOF algorithm is described as below .The Dimensions of the data set D is

d, | D | denotes the number of data objects in the data set D, k is the positive integer parameters which is given by the user.

Algorithm DBSCAN-LOF (data set D, the number of nearest neighbors k)
Input: data set D, the number of nearest neighbors' k.
Output: outliers and clusters.
Steps:
1. **For** the each objects p of the data set D
2. {
3 .To calculate the distance t between the object p and other objects
4 .Using the sorting method, in turn find the k values which is the nearest distance from the object p
5. Made the k- neighborhood point set $N_k(p)$ of object p.
}
6. **For** the each objects p of the data set D
7. {
8. Set the two properties of the object, the initial value are NULL; // the identity object of attribute1 is a core.
The object is non-core objects; his identity object of attribute2 is the core point, the boundary points is also the outliers.
9. **If** ($k\_radius(p) \le \frac{\sum_{o \in N_k(p)} k\_radius(o)}{|N_k(p)|}$)
10. {
The attribute1of the object p is marked as the core object, and in the $N_k(p)$,the addition2 of all the objects except the core objects are marked as boundary point.
11. **If** (the attributes 2 of the object p is NULL)
12. the attributes2 of p is marked as core points;
13. **Else**
14. the attributes2 of p is marked as core points once again;
15.}
16. **Else**
17 the attributes1 of p is marked as non-core points // then p is the non-core object.
18}
19. For each core object p
20. {
21. **If** ($dist(p,o) \ge k\_radius(o)$) // $o \in N_k(p)$, and o is the core object ;
22. the attributes1 of p is marked as non-core points, the attributes 2 is marked as NULL;
23}
24. **For** the object p which is non-core objects and each attributes 2 is NULL.
25. {
26. calculate the LOF value of the object p;
27. **If** (1 / (1 +) LOF (p) (1 +))
28. the attributes2 of p is marked as boundary points;
29. **Else**
30. the attributes2 of p is marked as outliers;
31.}
32. **While** (the core object p is marked as traverse objects)
33 {
34. the core object p is marked by class number ,labeled as ClusterIDp;
35. **While** (the density of the object p is up to the object q)
36 {
37. **If** (q class label is NULL)
38 .q will be set to ClusterIDp;
39. **Else**
40. the class of all elements q of the same class label is updated for ClusterIDp;
41.}
42.}
43 .the boundary point of an object is assigned to the cluster associated with the core object;
44. Output outliers and clusters.

3. Experiment Analysis

In this section, by experiments the DBSCAN-LOF algorithm, clustering quality and LOF efficiency is analysis.

The Algorithm is written by JAVA, it is run in 4.3.00GHz, 1GB memory, 160GB hard drive, under JBulider9.0 environment Pentium (R). Test data set information in Table 1, using the synthetic data sets TDB and Zoo dataset in the UCI, in order to assess the outliers detection method based on clustering, it is difficult to know exactly what data

objects are realistic data sets produced by different mechanisms outliers. Although many outlier detection of cases have been studied, but a set of possible outlier data is often incomplete, it is difficult to assess whether all the abnormal data is detected. Therefore, this new method of Gaussian random points were generated .In order to accurately assess the impact of the number of different dimensions and different sizes of data, we generate different sets of 25 dimensions. We choose the data sets which include 1000 and 5000 data.

Table 1. The Test Data Set TDB

data set	number of data points	dimension
TDB	1000, 5000	25
Zoo	101	17

More intuitive, the experimental results of DBSCAN_LOF algorithm and the experimental results of DBSCAN are shown in Table 2.

Table 2. DBSCAN_LOF Algorithm Results with the Results of the Control Algorithm DBSCAN

K value choice / accuracy / algorithm	3(*Eps* 为 2.5)	5(*Eps* 为 2.5)	12(*Eps* 为 2.5)
DBSCAN	82.37%	85.34%	71.22%
DBSCAN_LOF	92.31%	95.60%	90.63%

The experimental results can be seen that DBSCAN_LOF clustering algorithm is better than DBSCAN algorithm, through the automatic value k- neighborhood radius, it can effectively avoid the impact of the global radius Eps of clustering results.

In Figure 1, using the 25-dimensional TDB 1000 data points data set, comparing the run time of DBSCAN, DBSCAN_LOF, LOF and DBSCAN + LOF, (where k = 20, Eps = 3), it can be found that, the algorithm running time is less than DBSCAN_LOF LOF algorithm because LOF algorithm requires all the core objects outlier degree (OF) operations, while DBSCAN_LOF only for non-core subjects of non-boundary object OF arithmetic, but the run time of DBSCAN_LOF is greater than DBSCAN, because we use k- neighbors thought calculating the local density determine the core object, so the time should be slightly higher than DBSCAN. The purposes of our algorithm are not only clustering but also detect outliers. If DBSCAN and LOF algorithm is run, the run time will be higher than the run time of DBSCAN_L.

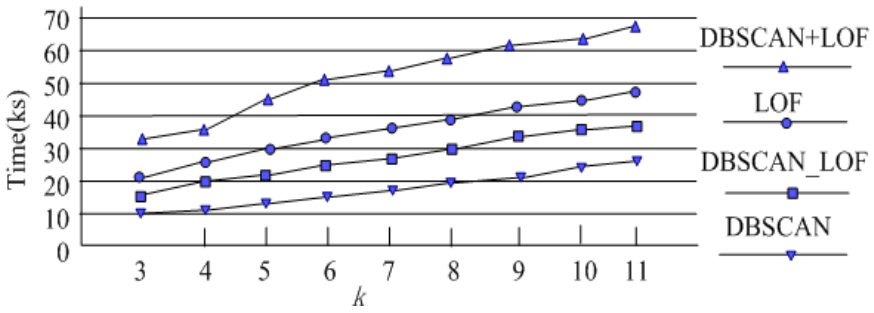

Figure 1. The Running Time of TDB (1000, 25) Data Set Algorithm

In the TDB dataset, in order to find the clear definition which is not the obvious exception of the data, we need to do the following thing: First, we randomly generate a Gaussian mixture model. Hybrid model describes the generic data points or no abnormal data points. We use a normal distribution of data space to create 10 outliers, and put them in 1000 and 5000 the data set. In this set of experiments, we compare the new method proposed in this paper and Top-n clusters outlier detection method based on sorting quality. In the experiments, we compare the difference between the set DBSCAN-LOF and Top-n two algorithms using a variety of data of different sizes and different dimensions. For Top-n, we set MinPts parameter value from 10 to 25.

According to the results we can find the 25-dimensional data sets of different sizes in Figure 1. It shows that the effect of these two methods of data collection is about the same. The reasons of this phenomenon may be that outliers inherently are difficult to be detected.

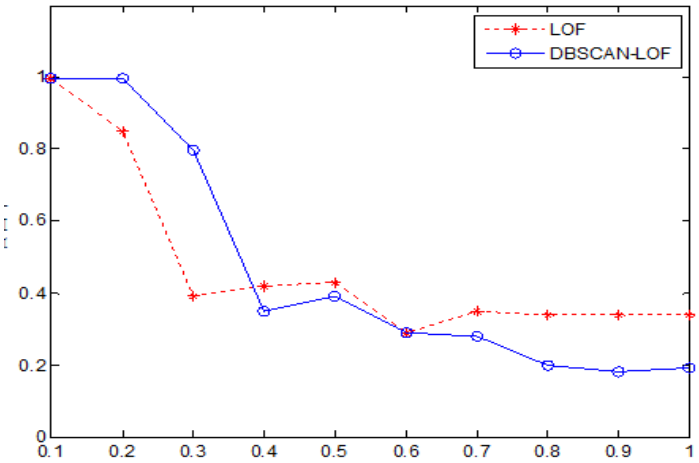

Figure 2 (a). 25d 1000 Data Point

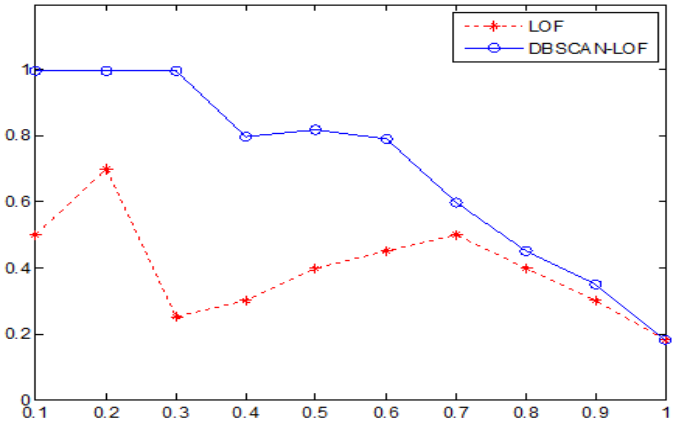

Figure 2 (b). 25d 5000 Data Point

In this paper, the clustering technology and outlier mining technology are organic combination, it presents a DBSCAN-LOF algorithm, making LOF algorithm does not need an object for each operation OF values, thereby the amount of computation of LOF is reduced ,at the same time, DBSCAN algorithm does not need to input parameters Eps, it can be adjusted according to the local density of the regions in which the parameters of the data object clustering, this method reduce the influence of the uneven distribution to data DBSCAN clustering effect, and finally cluster and outliers. can be simultaneously

detected Theoretical analysis and experimental results show that the algorithm is effective and feasible. Next, it will reduce the impact of the selected parameter points to the clustering quality, and the algorithm in the application of high-dimensional spatial data sets is further studied.

4. Conclusion

In this paper, the clustering technology and outlier mining technology are organic combination, it presents a DBSCAN-LOF algorithm, making LOF algorithm does not need an object for each operation OF values, thereby the amount of computation of LOF is reduced ,at the same time, DBSCAN algorithm does not need to input parameters Eps, it can be adjusted according to the local density of the regions in which the parameters of the data object clustering, this method reduce the influence of the uneven distribution to data DBSCAN clustering effect, and finally cluster and outliers. can be simultaneously detected Theoretical analysis and experimental results show that the algorithm is effective and feasible. Next, it will reduce the impact of the selected parameter points to the clustering quality, and the algorithm in the application of high-dimensional spatial data sets is further studied.

References

[1] D. Hawkms, "Identfication of Outllers. London Chapman and Hall, **(1980)**.
[2] M. M. Breunig, H. P. Kriegel and R. T. Ng, "LOF: identifying density-based local outliers", Proceedings of 2000 ACM SIG-MOD International Conference on Management of Data, **(2000)**, pp. 93-104.
[3] H. V. Jansen, N. R. Tas and J. W. Berenschot, "in Encyclopedia of Nanoscience and Sheng-yiz Jiang, Qing-bo", An Clustering-Based Outlier Detection Method, Proceedings of the 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery – vol. 2, **(2008)**, p 429-433.
[4] W. Jin, K. H. Anthony, T. J. Han, "Mining Top-n Local Outliers in Large Databases", Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, **(2001)**, p. 293-298.
[5] M. Halkidi, M. Vazirgia, "lni, clustering validity assessment finding the optimal partitioning of a data set", IEEE Int'l Conf. Data Mining, California, USA, **(200l)**.
[6] H. P. Kriegel, M. Schubert and A. Zimek, "Angle-Based Outlier Detection in High-dimensional Data", Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, August 2008, pp. 444-452.
[7] E. M. Knorr and R. T. Ng, "Algorithms for Mining Distance-Based Outliers in Large Datasets", Proc. 24th Int. Conf. on Very Large Data Bases, New York, NY, **(1998)**, pp. 392-403.

