Computing, Performance and Communication Systems (2017) 2: 1-7
Clausius Scientific Press, Canada

An improved outlier detection algorithm K-LOF based on density

Zhang Shaomin[a*], Luo Xiangyu[b], Wang Baoyi[c]

School of Control and Computer Engineering, North China Electric Power University, Baoding, 071003, China

a.email: zhangshaomin@126.com,

b.email:1922637717@qq.com, c.email:wangbaoyiqj@126.com

Keywords: Data mining; Clustering algorithm; Outlier detection

Abstract: The local outlier factor (LOF) algorithm is one of the representative algorithms based on the density outlier detection algorithm. But the algorithm has the problem of high time complexity, not suitable for large data sets and high dimensional data set. Therefore, this paper proposes a new outlier detection algorithm, clustering the data sets determines the data center of data space through the K-means clustering algorithm, building data set primary model by setting the distance threshold of the data set object to the data center, and optimizing the screening process combined the neighbor distribution of data objects. Although the use of clustering algorithm for abnormal data set screening will increase the computational complexity of the algorithm, but the data center space once identified will no longer need to repeat the calculation, so with the increase of data, the advantages of the algorithm will become more and more obvious. After testing, the algorithm can effectively improve the detection accuracy of anomaly factors, and reduce the computational complexity of the algorithm, and can complete the local outlier detection.

1. Introduction

Outlier detection is an important branch of data mining, mainly used to detect abnormal data which deviated from the normal distribution pattern, it can extract the potential and valuable information from the large, fuzzy complex data, and is widely used in data processing[1].

When the property and behavior of a data object are clearly distinguished from its nearest neighbor, it is treated as an outlier [2].Data outlier detection is widely used in many aspects, such as network intrusion detection, credit card fraud, environmental monitoring and others, as a prerequisite for the normal operation of the system, it has received extensive attention in academic and industrial circles.

There are many anomaly detection algorithms, including anomaly detection methods based on distributed, clustering, classification, depth, distance and density. With the development and

progress of machine learning, artificial intelligence, pattern recognition and other fields, more and more new and effective anomaly detection techniques and methods have been put forward[3].Literature [4] presents a SOMRNN detection algorithm based on the k- nearest neighbor distance, although the algorithm in efficiency and accuracy has been greatly improved, but it ignores the situation that the changes of the amount of data can cause the changes of k- neighbor distance. In addition, some scholars have proposed an outlier detection algorithm based on multiple clustering [5].

The outlier detection algorithm based on density can detect more outliers, and the method can be used to analyze data sets with uneven density distribution [6].The most representative of outlier detection algorithm based on density is the local outlier factor (LOF) algorithm.

Literature [7] presents a detection algorithm based on density, the algorithm introduces local outlier factor (Local Outlier Factor, LOF), and uses LOF value of the data to determine whether the data is abnormal, this method is only applicable to the static data detection. Literature [8] proposed a new anomaly detection algorithm DLOF, the algorithm reduces the computational complexity of the algorithm, but it computes in the whole data set, with the increase of data, the running time will show significant growth. Literature [9] improved an anomaly detection method based on improved density clustering that density clustering is in each feature column, and a normal behavior profile is established to compare the difference between some behavioral data and the normal behavior contour. It can be seen from the improved algorithm that it is effective to reduce the computational complexity and shorten the running time by preprocessing the data sets and screening the initial abnormal data sets.

2. LOF Anomaly Detection Algorithm Brief Introduction

The definition of local outlier factor algorithm LOF: For data points $p(x_i, y_i)$ in DS, the local exception factor is defined as:

$$LOF_k(p) = \frac{\sum_{o \in N_k - dist(p)} \frac{lrd(o)}{lrd(p)}}{|N_{k-dist}(p)|} \tag{1}$$

Among them: (1) using Euclidean computing method, the distance between two data points d(p,o) in DS is defined as $d(p,o) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$;

(2) For the data points $p(x_i, y_i)$ in DS, the k- distance (k-dist (P)) represents that the number of the nearest distance to P is k in the DS, and the maximum distance is denoted as k-dist (P), and the k points constitute the k- distance neighborhood of the point P, record as $N_{k-dist(p)} = \{q \in DS \setminus \{p\} | d(p,q) \le k - dist(p)\}$;

(3) For the object $p(x_i, y_i)$, $o(x_j, y_j) \in DS$, the reachable distance between the object p and the

object o is denoted as $reach-dist_k(p,o)=\max\{k-dist(o),d(p,o)\}$;

(4) The local reachable density of the object P is denoted as

$$lrd(p)=1/\left(\frac{\sum_{o\in N_k-dist(p)} reach-dist_k(p,o)}{|N_{k-dist}(p)|}\right) \tag{2}$$

Among them, $|N_{k-dist}(p)|$ is the number of data points p's K distance neighborhood; The degree of P local anomaly: the average of the local reachable density of P and its k- neighbors.

Defined by the algorithm: for the data set DS, if the distance between P and other data points in the neighborhood is smaller, P reachable distance and other data points' reachable distance is smaller. Local reachability density of P and its neighbors is closer, and value of $LOF_k(p)$ is approximately equal to 1, that represents the degree of abnormality of data points P [10].

3. K-LOF Algorithm Improvement Principle

3.1 Primary Model Based on K-means Abnormal Data Set

Clustering analysis can cluster according to the data similarity principle without knowing the pre designated categories. Typically, the normal data is relatively centralized, surrounded by a data center, and the abnormal data is deviation significantly from the data center. According to the above situation, you can use the normal data to determine a data center, and then determine the primary set of abnormal data through the deviation of the data to be measured to the data center.

Using k-means clustering algorithm to determine the data center of the normal data set, can make various parameters simplified, and also can deal with large data sets with obvious deviations between clusters and clusters, and the efficiency is higher [11].

Set data collection D containing n d dimensional data objects, remembered to D=$\{x_i|\mathrm{i}=1,2,...n\}$.

The data set D is divided into k subsets, called k-clustering, each subset is a cluster w_j, and the cluster center is obtained by formula $c_j=\frac{1}{n_j}\sum_{x\in w_j}x$, and the clustering quality is measured by

$$J=\sum_{j=1}^{k}\sum_{i=1}^{n}d_{ij}(x_j,c_j) \tag{3}$$

$d_{ij}(x_i,c_j)$ represents the distance between the data object x_i and the cluster center o_j; The objective function J represents the sum of the distances between each data object and its cluster center. The smaller the value of J, the closer the data object in the cluster to the cluster center, the

closer the data distribution. By constantly optimizing the value of J, until the minimum value of J is obtained, the optimal clustering center is the corresponding clustering center[12].

3.2 K-means Abnormal Data Primary Election Model Optimization

By setting the threshold radius r to determine the clustering center of the circle, the points on the circle are treated as normal, and the points out of the circle are added into primary abnormal data set. Because the distribution of threshold is often set conservative, it is possible to have a normal point in the small distance outside the circle (due to the normal distribution of threshold value is often too small). Therefore, this paper has improved it and expanded the query scope appropriately according to the specific data distribution, as shown in Figure 1.The steps of the optimized neighborhood query is as follows:

(1) After clustering the normal data, o is the data center, taking o as the center of the circle, r as the radius of the distribution threshold, than searching the point in the r- neighborhood, and determining the neighborhood point d (x, o) as the normal data.

(2) Select the object a, which is located outside the circle 1 and d(a, o) is smallest , with a as the center, r1=d (a, o) -r as the radius, searching in the r1 neighborhood of the object a, whether is there a point b, which satisfies d(b, a) <r1 and d(b, o) >r1. If exit, take r=d (b, o) as the query scope to continue to reduce the primary object of the outlier data set, and than move to step (2); if not exit, add the object a into the exception data set.

(3) All outliers that have not been processed are added to the primary exception data set.

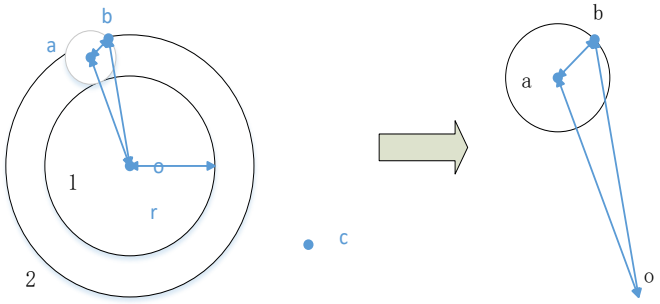

Figure 1 Schematic diagram of algorithm improvement

3.3 Algorithm Description

The process of K-LOF algorithm is as follows:

(1) Obtain the primary abnormal data set by using the K-means anomaly detection model to preprocess the data set.

(2) In the primary exception data set, select a point p that has not been processed, and calculate the weighted distance d(p,q,w) and K weighted distance k-wdistance. Use formula (4) to determine the reachable distance of each object to data center.

$$reach - dist_k(p,q) = \max\{k - wdis\tan ce(q), d(p,q,w)\} \tag{4}$$

(3) Using the formula (2) to calculate the local reachable density of each object in the primary

concentration.

(4) Using the formula (1) to calculate the local outlier factor LOF .

4. Experimental Results

This paper proves the superiority of K-LOF algorithm by comparing the detection accuracy and algorithm execution time of LOF, DLOF and K-LOF algorithms. The data source used in the experiment is the data collected from the wide area measurement system. Each object has a dimension of 5 dimensions, including voltage, current and power, etc.. In order to ensure the accuracy of the experimental results, artificial changes in the normal data, error data accounted for 5% of the data set.

(1) Accuracy comparison of algorithms

Detection accuracy = the number of outliers detected / outliers. From the experimental results shown in Figure 2, the proposed K-LOF algorithm is more accurate than the LOF and DLOF algorithms.

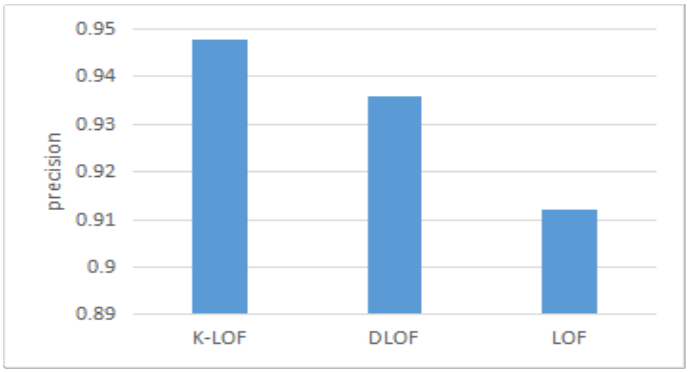

Figure 2 Comparison of the accuracy of the algorithm.

(2) Execution time comparison of algorithms

Although it takes some time to determine the clustering center, once the center is determined, no duplication is needed. The time complexity of the algorithm is related to the specific data distribution, so it is difficult to determine the time complexity theoretically. This part of the experiment by recording different detection algorithms on the same data source to detect the running time to prove that the proposed K-LOF algorithm reduces the computational complexity of outlier detection, the test results as shown in Figure 3.

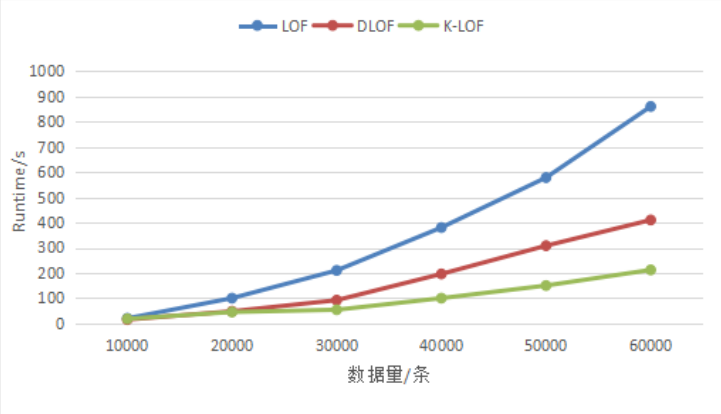

Figure 3 Comparison of the execution time of the algorithm.

5. Conclusion

Experimental results show that K-LOF algorithm is effective and feasible for outlier detection in data set. K-LOF algorithm uses k-means algorithm to cluster data sets to determine the center of data space. Through setting the distance threshold from object to data center, we establish data set primary model and optimize the selection process. Although the data center will increase computational complexity, it will greatly reduce the computational complexity of the algorithm in the case of large amounts of data. However, there are still some shortcomings in the algorithm, such as the threshold of data objects to data centers needs to be set artificially, and it will affect the speed of the algorithm.

References

[1] *ZUO Jin, CHEN Zemao. Anomaly Detection Algorithm Based on Improved K-means Clustering[J]. Computer Science, 2016, 43(8):258-261.*

[2] *CAO Ke-Yan, LUAN Fang-Jun,SUN Huan-Liang,DING Guo-Hui. Density-based Local Outlier Detection on Uncertain Data[J]. Chinese Journal of Computer.2016,(39):1-15.*

[3] *Fu PG, Hu XH. SLDOF: Anomaly Detection Algorithm Based on the Local Distance of Density-based Sampling Data[J]. Ruan Jian Xue Bao/Journal of Software (in Chinese). 2016 :1-16.*

[4] *ZHANG Zhong-ping, LIANG Yong-xin. Stream Data Outlier Mining Algorithm Based on Reverse k Nearest Neighbors[J].Computer Engineering. 2009, 35(12):11-13.*

[5] *GU Ping, LIU Hai-bo,LUO Zhi-heng. Multi-clustering based outlier detect algorithm[J]. Application Research of Computers.2013, 30(3):751-753.*

[6] *Angiulli F, Fassetti F. DOLPHIN:An efficient algorithm for mining distance-based outliers in very large datasets[J]. Acm Transactions on Knowledge Discovery from Data, 2009, 3(1):1-57.*

[7] *WANG Qian, LIU Shu-zhi. Improvement of local outliers mining based on density[J].Application Research of Computers. 2014, 31(6):1693-1696.*

[8] *HU Caiping and QIN Xiaolin.A Density-Based Local Outlier Detecting Algorithm[J].Journal of Computer Research and Development.2010, 47(12):2110-2116.*

[9] *HU Liang, REN Wei-wu,REN Fei,LIU Xiao-bo,JIN Gang. Anomaly Detection Algorithm Based on Improved*

Density Clustering[J].Journal of Jilin University(Science Edition).2009, 47(5):954-960.

[10] *HU Wei, LI Yong, CAO Yijia, ZHANG Zhipeng, ZHAO Qingzhou, DUAN Yilong.Fault identification based on LOF and SVM for smart distribution network.Electric Power Automation Equipment[J].2016, 36(6):7-12.*

[11] *Tan P N, Steinbach M, Kumar V. Introduction to data mining[M]. FAN Ming, FAN Hongjian, translated.Beijing,China: The People's Posts and Telecommunications Press[J]. 2010: 328-330.*

[12] *YAN Yingjie,SHENG Gehao,LIU Yadong,DU Xiuming,WANG Hui,JIANG Xiuchen. Anomalous State Detection of PowerTransformer Based on Algorithm Sliding Windows and Clustering. High Voltage Engineering.2016, 42(12):4020-4025.*