Research on outlier detection for high dimensional data stream

Liping Yu [a], Yunfei Li* and Juncheng Jia

School of Soochow University, Suzhou 215000, China.

[a]fromthesmile@163.com

Abstract. The development of the Internet of things has put forward new requirements to the data processing capacity, and outlier detection has found an increasingly wide utilization in the field of data mining. The accuracy of the outlier detection algorithm based on Euclidean distance in the high dimensional data detection cannot be guaranteed, what is worse, the processing time is too long. This paper constructs the small data sets of the best set of data grid and recently data grid, in order to calculate the abnormal degree of the newest data point by measuring angle variance of the high dimensional data stream; as data stream capture, the best data grid and data grid updated incently, whose aim is to solve the concept transferring of big data flow. The experimental results show that compared with the ABOD algorithm and the classical algorithm, this algorithm is more suitable for the outlier detection of the high dimensional data stream in the Internet of things.

Keywords: High dimensional; data stream; outlier detection.

1. Introduction

Based on IOT technology, fuzzy data mining can from a large number of complex data to extract potentially valuable information, and networking data exist in the form of high dimension, with massive, heterogeneous noise characteristics, which makes the traditional data analysis algorithm to detect the outliers inefficient such as high time-complexity and has potential limitations. High dimensional data flow anomaly detection faces many challenges.

At home and abroad, a large number of scholars have studied the data mining of wireless sensor networks and proposed a series of methods to detect the anomaly of high dimensional data stream. Such as the maximum margin criterion algorithm of high dimensional data is projected to the position of the feature space using the literature [1], and then the outlier data mining using minimax probability machine, but for real-time data flow of the running time of the algorithm complexity is high; the literature [2] a set of data pretreatment, in small-scale centralized to find K nearest neighbors, according to the calculation the degree of outlier selection of outliers, the algorithm is realized by parallel to improve the computational efficiency of large-scale data sets, but does not apply to high dimension data set.

Kriegel and other proposed based anomaly detection algorithm (Outlier Detection Angle-based, ABOD) [3]. In high dimensional space, the angle is more stable than the distance, and the cosine angle variance to a certain extent can reflect the abnormal degree of the data [1]. As shown in Figure 1, the normal point in the area around the data by other data from all directions, angles, and angle variance; and abnormal data and other data points form the angle change little angle variance. ABOD algorithm considered that small variance of the point as outliers. However, the algorithm has the error of the clustering ensemble, and the operation cost of the algorithm is high with the increase of time.

Grid division is a common method for data processing in the discretization of spatial data [6]. High dimensional data stream is a continuous time series, in a certain range of normal circumstances, there is a small range of fluctuations. There is a certain deviation between the emergence of a point and the last data, or a certain time series, the data points deviate from the best data set. In view of the characteristics of high dimensional data stream, this paper proposes an improved algorithm for anomaly detection of data streams based on the variance of the angle.

2. Relevant definition

The following is a brief introduction to the relevant knowledge.

The first definition: high dimensional data stream is a kind of K-meta relation, in which the tuple arrives continuously, t is the time when the data point arrives, and K is the dimension of the dataset.

The second definition : Defines a data set on a given space as well as a sample point, randomly selected a pair of sample points from the data set, with the angle of the vectors, with all the variance values :

$$VOA_p = Var\left[\theta_{mpn}\right] = MOA_2(p) - (MOA_1(p))^2 \tag{1}$$

According to the analysis of the angle of the vectors, ABOD algorithm can effectively reduce the impact of the "Curse of dimensionality" on the detection results, but highify the time complexity, with the increase of data size and dimension, the performance of the algorithm is difficult to guarantee.

The third definition: Optimal data set grid is a small scale data pattern of data stream. In the data space, the data point of which the angle variance is within the threshold will be classify as the optimal data data grid.

3. Outline of the improved algorithm

3.1 The algorithm outline

The main idea of improvement based on the variance of the angle of high dimensional data flow anomaly detection algorithm: according to the classified data distribution of data stream, based on the literature [6] in all of the data grid to filter the normal region, but this method does not consider the concept of data stream transfer problems in the implementation process of this algorithm in real time, maintain an optimal set of data and recent data sets, according to the latest collection of high dimensional data to calculate the angle of abnormal factor variance, variance factor in computing part of the data flow in the perspective of historical data, low computational cost.

The main realization of the algorithm is:

Input: real time data points, the best data grid threshold, the weight and the recent data centralization;

Output: abnormal point set

Step 1 acquisition of real-time data, access to the latest data points in the high dimensional data stream;

Step 2 initialization of the best data set Best_list and recent data sets Latest_list;

Step 3 on the latest data points with respect to the Best_list and Latest_list and set the weight of the angle of the variance factor VOA, respectively, to determine whether it is abnormal;

Step 4 according to the latest data points VOA value and the set of the best set of threshold value to determine if the angle of variance factor is greater than the threshold value is marked as the normal data, on the contrary, the data points are labeled as abnormal data;

Step 5 if the data points are labeled as normal data, according to the first out of the way to update the latest data grid in real time, and vice versa is not updated;

In order to make the algorithm more efficient, the best data set is screened out before the variance of the variance of the angle is calculated. Because of the large amount of data, such as the calculation of the variance of each data point of view, long run time, the cost of computing. The optimal data set is a key parameter for mesh generation. When the different values are taken, the result of grid clustering is very big. When the grid is too large, the mesh size is small, the number of the grid increases, thus increasing the amount of computation. On the contrary, when the hour is over, the grid cell is very small. At this time, although the computational efficiency is improved, but the clustering effect is also significantly decreased.

The weights and threshold, grid is the main factor affecting the accuracy of the algorithm, and the weights and thresholds of different in different accuracy. In order to obtain the optimal weights and threshold algorithm requires a large number of test. In general, every application data sets corresponding to the rules and characteristics, but also the corresponding a set of optimal weights and threshold. A large number of experiments show that, as long as the optimal weights and threshold to obtain a set of data sets and the corresponding.

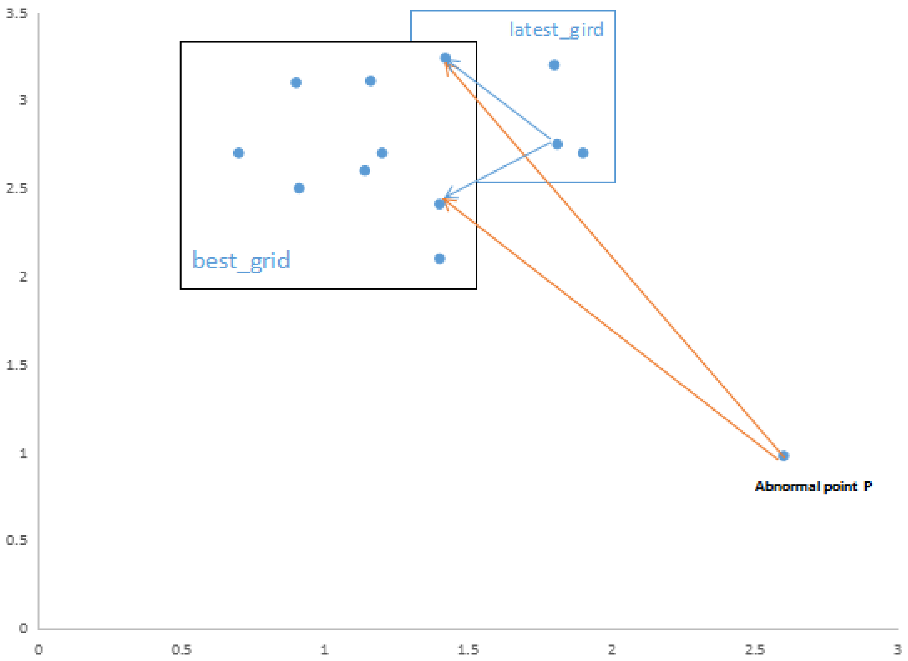

Fig. 1 The improved ABOD algorithm

3.2 The algorithm sample

```
Best_grid→∅, Latest_grid→∅
for input Xi From Data Stream
for all Di in best_list, project Di to each space,calculate the Xi VOA
for all Fi in Latest_list, project FI to each space,calculate the Xi VOA
calculate the angle-variance according to the above two and relative weight.
if VOA of  $X_i$ < $\mu$
label $X_i$  as normal and add it to Latest_grid, dequeue the first record
else label it as a outlier
end
```

4. Experiment

In this paper, the experimental environment we are accessible is the computer Core (R) i5-3470 (TM) CPU Intel @ 3.20GHz, memory 8.0GB. Development platform for PyCharm 2015, the use of Python language and pandas analysis package on the improved algorithm and ABOD algorithm and the classic LOF algorithm for comparison.

Experimental data set is intrusion detection data set KddCup99. The data set has the characteristics of high dimension, here we chose 1000 data, of which is a outlier record of 10.

The effectiveness analysis of KddCup99 algorithm used in this paper, in order to analyze the effect of different threshold on the accuracy of algorithm, and different data flow calculation set algorithm running time and accuracy, this paper carried out the following 4 groups experiment:

In Experimen 1, we compare the ABOD algorithm and the improved algorithm, we get the precision of them on data set we have chosen.

In Experiment 2, the optimal threshold value is 0.3, the best data centralization is 0.62, and the latest grid is 0.38;

Based on Experiment 3, the threshold value of the optimal data set is set to 0.5, and the weight and fixed value of the optimal data set is 2;

In Experiment 4, the threshold value of the optimal data set is fixed, and the weights are set to 0.8 and 0.2 respectively based on the experiment 2;

Based on the test data set, the above experimental test accuracy as shown in Figure 2 and 3, can be seen through a lot of experiments to obtain the optimal parameters of this algorithm can ensure the ability to identify abnormal state, such as in Experiment 1; with the increase of the best data set threshold, detection precision decrease abnormal points, such as experiment 1 and 2 fixed threshold; that means the weight change data points by the best data set and the impact of the recent data sets to different degrees, and the data set relied more on the best data set.

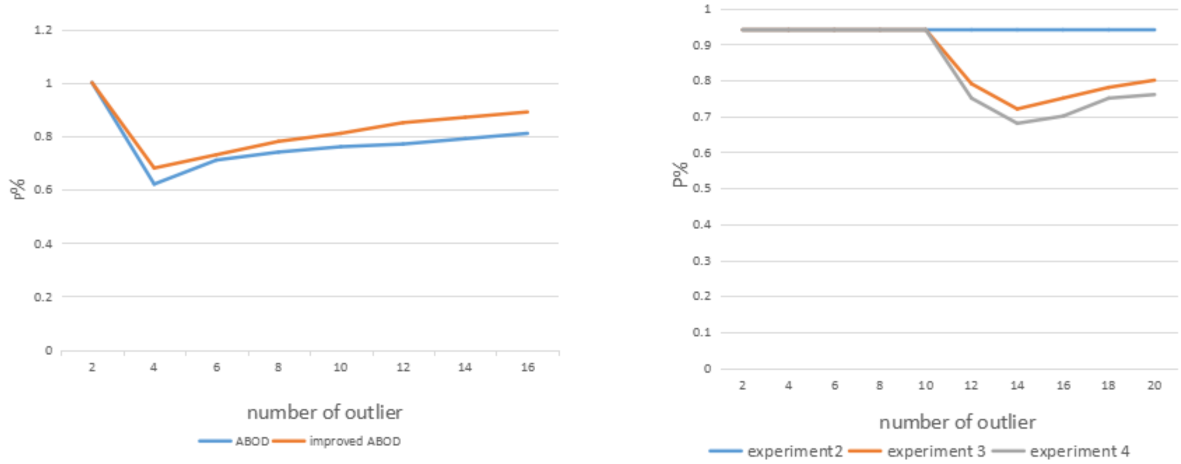

Fig. 2 The experiment of precision contrast

5. Summary

The high dimensional data point distribution based anomaly detection method, proposed an improved anomaly detection algorithm, the detection process of grid partition of data set in the data flow, maintain the best data set grid and recent data set grid, reduce the computational cost of angle variance, and by setting the weight value to solve the problem of the transfer of data stream concept them. Experimental results show that the proposed algorithm can effectively identify outliers in high dimensional data streams, and the time complexity of the algorithm is lower than that of the ABOD algorithm. How to adapt the weights according to the characteristics of the data stream is the content of the next step.

References

[1] Ding J, Wang L, Shen D, et al. An anomaly detection system on big data [J]. Natural Science Journal of Hainan University, 2015.

[2] Kriegel H P, Hubert M S, Zimek A. Angle-based outlier detection in high-dimensional data [C]. Proc of KDD. 2008: 444-452.

[3] Kriegel H P, Schubert M, Zimek A. Angle-based outlier detection in high dimensional data [C]. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2008: 444-452.

[4] Pham N, Pagh R. A near-linear time approximation for angle-based outlier detection in high dimensional data [C]. Proceeding of the 18th ACM SIGKDD International Conference on Knowledge. Discovery and Data Mining. New York, USA: ACM, 2012:877-885.

[5] Zhou D H, Wei M H, Xiao Sheng, et al. A survey on anomaly detection, life prediction and maintenance decision for industrial processes [J]. Acta Automatica Sinica, 2013, 39: 711-722

[6] Feldman D, Schmidt M, Sohler C. Turning big data into tiny data: constant-size coresets for K -means, PCA and projective clustering [C] Proc of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms.2013: 1434-1453.

[7] Zhang Y, Hamm N, Meratnia N. Statistics-based outlier detection for wireless sensor networks [J]. International Journal of Geo graphical Information Science, 2012, 26(8):1373-1392.

[8] Han J, Kamber M. Data mining: concepts and techniques [M]. 3rd ed.2011:1-18.