Improved Data Stream Clustering Algorithm for Anomaly Detection

Chunyong Yin[1(✉)], Sun Zhang[1], and Jin Wang[2]

[1] School of Computer and Software,
Jiangsu Engineering Center of Network Monitoring,
Nanjing University of Information Science and Technology, Nanjing, China
yinchunyong@hotmail.com
[2] College of Information Engineering, Yangzhou University, Yangzhou, China

Abstract. Intrusion detection provides important protection for network security and anomaly detection as a type of intrusion detection, can recognize the pattern of normal behaviors and label the behaviors which departure from normal pattern as abnormal behaviors. We think that the traditional methods based on dataset do not satisfy the needs of dynamic network environment. The network data stream is temporal and cannot be treated as static dataset. The concept and distribution of data objects is variety in different time stamps and the changing is unpredictable. Therefore, we propose an improved data stream clustering algorithm and design the frame of anomaly detection according to the improved algorithm. It can modify the established model with the changing of data stream and detect abnormal behaviors in time.

Keywords: Intrusion detection · Anomaly detection · Data stream · Clustering

1 Introduction

The popularity of internet application brings great challenge to network security. In recent years, hacker intrusion, network paralysis and user information leakage have caused extensive damage to society and economy. Wenke Lee [1] proposed the concept of intrusion detection in 1998 and it provides important protection for network. Intrusion detection can be divided into two types: misuse detection and anomaly detection. Misuse detection analyzes the characteristics of known attack behaviors and builds rule base which is used to match with behaviors. The behavior with higher similarity will be labeled as abnormal behavior. Anomaly detection recognizes the pattern of normal behaviors and label the behaviors which departure from normal pattern as abnormal behaviors.

Early anomaly detection has high false alarm rate and the introduction of data mining makes it a great development. The application of data mining in anomaly detection can be divided into two types according to the processing objects. One is the method based on dataset and most research results is based on dataset. The other one is the method based on data stream. The dataset is static and the data model based on dataset is permanent. Conversely, data stream is temporal and it is changing by time [2]. The data model based on data stream is variable in different time stamps. Because

J.J. (Jong Hyuk) Park et al. (eds.), *Advanced Multimedia and Ubiquitous Engineering*,
Lecture Notes in Electrical Engineering 448, DOI 10.1007/978-981-10-5041-1_98

the concept and distribution of data objects is variety in different time stamps and the changing is unpredictable [3]. We think that data object is transmitted in network as the form of stream, so the method based on data stream is more appropriate.

The main methods of data mining are classification analysis, clustering analysis and regression analysis. Classification analysis trains the model by labeled datasets and recognizing unlabeled data records in testing phase. Clustering analysis belongs to unsupervised method and it can divide data points into different clusters according to their similarity. The data points in different clusters will have farther distance and the cluster has high similarity inside. The data model based on data stream need to be modified in time. The adjustment of classification model is harder than clustering model, because classification model is supervised and it needs extra label resources. Thus, we improve clustering algorithm based on data stream as the core of anomaly detection and design the anomaly detection frame.

2 Anomaly Detection Model

Recent research works of anomaly detection and clustering algorithm mainly focus on datasets and it can obtain perfect performance in simulate experiments. We summarize and compare the difference between the methods of dataset and data stream as shown in Table 1.

Table 1. Comparison of methods based on dataset and data stream

Types	Methods based on dataset	Methods based on data stream
Data model	Permanent	Temporary
Mathematical relation	Data objects set	Data objects sequence
Saved all	Yes	No
Processing	One-time	Continuous
Results	Accurate	Approximate
Time	Longer	Shorter
Memory	More	Less

We compare them from seven points: (1) The dataset is static and the data model based on dataset is permanent. The data stream is changing with time and the data model based on data stream is temporary; (2) Dataset is the set of data objects and data stream is the sequence of data objects. The definition of data stream is a sequence that constructed by continuous and ordered data points; (3) The method based on dataset usually read all data records into the memory. But for data stream, it is continuous and infinite. The memory consumption will increase with time. Therefore, the method based on data stream utilize the fixed memory to save summary statistics information; (4) The process of dataset is one-time and that of data stream is continuous. However, the process for each data object in data stream is one-time; (5) The results on dataset is always accurate, because it is calculated by accurate value of data objects. In the process of data stream, it only saves summary statics information of data objects that

will cause the approximate result. But the approximate result does not affect the detection of anomaly behaviors; (6) the time consumption on dataset is longer than that on data stream. In some related works, the data mining methods can get pretty better results, but it consumes too long time. To process the data object in time, the algorithm on data stream should have high efficiency; (7) Because of the reason in third point, the method based on dataset takes much more memory than that based-on data stream.

From the comparisons above, we conclude that the method based on data stream need to establish an efficient data stream analysis model, that is, the algorithm has a smaller time and space complexity. The limited computer storage capacity cannot save infinite data objects in data stream, which requires that the memory consumption of the algorithm dose not increasing with time and it can be some fixed value [4].

We design the improved anomaly detection model according to the characteristics of data stream clustering algorithm and it is shown in Fig. 1.

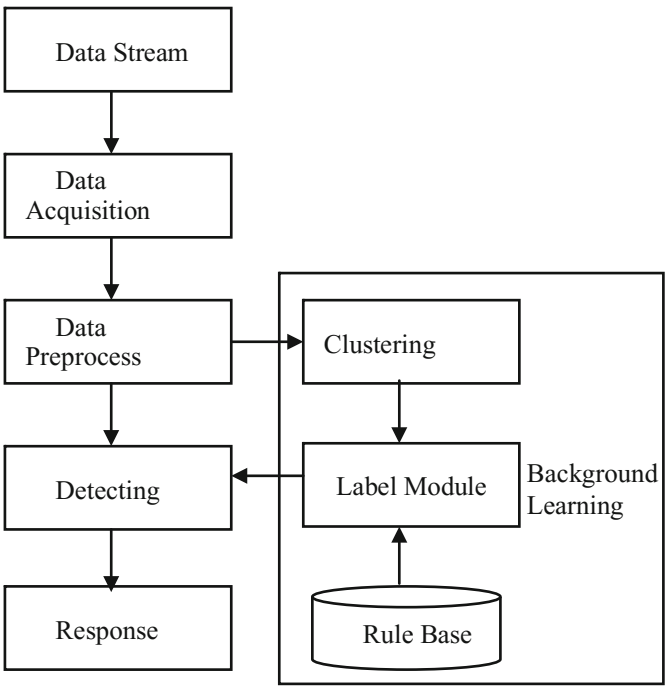

Fig. 1. Anomaly detection model. This shows the anomaly detection model which consists of several modules. Clustering module is the core of anomaly detection model.

The improved anomaly detection model consists of four modules. Data acquisition module collects data objects from data stream and data preprocess module preprocess data objects, including data cleaning and feature selection. The detecting module is composed of three parts. The clustering module is the core of model and it assign new data object into certain cluster. Suspected cluster will be sent to label module and it is

matched with the rules from rule base. Clustering algorithm is unsupervised and it cannot label these data objects. So, label module will label data objects by rule base.

3 Improved Data Stream Clustering Algorithm

In this paper, we propose the improved data stream clustering algorithm for anomaly detection. Figure 2 shows the diagram of clustering algorithm application.

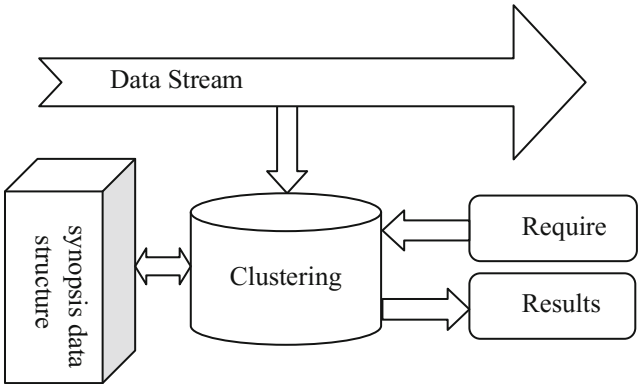

Fig. 2. Diagram of clustering algorithm application. This shows the basic application diagram of data stream clustering algorithm. *Synopsis data structure* is applied to store summary statics information.

Data stream S is a sequence of data objects o, which can be denoted as $S = \{o_1, o_2, o_3, \cdots, o_h\}$, and each data object has n features. The synopsis data structure is the important part of data stream clustering algorithm which is utilized to save the summary statistics information of clusters. In our improved data stream clustering algorithm, we use two types of synopsis data structures to store the summary statistics information of normal clusters and suspected clusters [5].

Normal cluster is denoted as $n - cluster$ which can save the necessary information. As Formula 1 shows, it has four attributes. δ is the number of data objects in the cluster; μ is the center of cluster and it is the average value of data objects; SS is the quadratic sum of data objects in the cluster. We add the attribute *flag* to identify the type of cluster.

$$n - cluster : (\delta, \mu, SS, flag). \tag{1}$$

Suspected cluster is denoted as $s - cluster$ and it is suspected to be abnormal behaviors. Formula 2 shows the five attributes of suspected cluster.

$$s-cluster:(\delta,\mu,SS,flag,list). \tag{2}$$

The attribute *list* stores the identifiers of data objects in suspected cluster and the detailed information of data objects are saved in the disk instead of memory. This attribute will increase the consumption of memory, but $s-cluster$ can be converted to $n-cluster$ and this attribute will be deleted in time. Besides, the amount of abnormal behaviors is far less than that of normal behaviors. This attribute will not occupy the memory for a long time. The main idea of improved data stream clustering algorithm is as follows and the flow diagram is shown in Fig. 3.

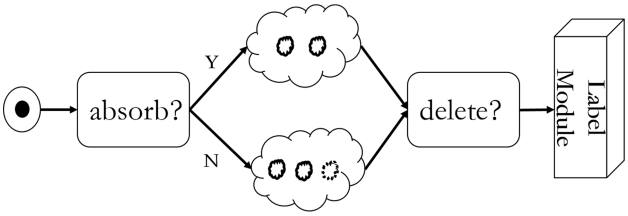

Fig. 3. Flow diagram of data stream clustering algorithm algorithm. This shows the processing of new arrived data object. The clustering algorithm decides it is absorbed by existed clusters or become new cluster. The redundant cluster should be deleted in time and sent to label module.

Step 1: At the beginning of algorithm, it waits to receive a certain number of data objects and the clustering algorithm based on dataset will generated some clusters. These clusters will be labeled as $n-cluster$ or $s-cluster$ according to the number of data objects in the cluster;

Step 2: When the new data object o arrives, it calculates the distance between data object and existed clusters. According to the distance, existed clusters decide whether to absorb this data object;

Step 3: If the distance is less than threshold, data object chooses the nearest cluster to be integrated with. The information of selected cluster should be updated as Formula 3;

$$(\delta,\mu,SS)\rightarrow(\delta+1,\frac{\mu\times\delta+o}{\delta+1},SS+o^2). \tag{3}$$

Step 4: If data object cannot be absorbed by existed clusters, it will be added into the memory as the center of new cluster and it is labeled as $s-cluster$. When the number of data objects in is more than the threshold, $s-cluster$ is transformed into $n-cluster$ and deletes the attribute from the memory;

Step 5: In order to limit the number of clusters in the memory, it should delete redundant clusters. Secure deletion method is to select $s-cluster$ which does not update for a long time;

Step 6: The center of clusters to be deleted is sent into label module. Because of the high similarity in clusters, data objects will be labeled according to the center.

4 Conclusion

In this paper, we propose the improved data stream clustering algorithm and design corresponding anomaly detection model. The clustering algorithm based on data stream is more appropriate than that based-on dataset. The clusters can be updated in time according to new arrived data object. The improved clustering algorithm requires less memory and time cost. The next works of us are to consider the situation of multiple concurrent and the improvement of label module. An appropriate method of feature selection will further improve efficiency.

Acknowledgments. This work was funded by the National Natural Science Foundation of China (61373134). It was also supported by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD), Jiangsu Key Laboratory of Meteorological Observation and Information Processing (KDXS1105) and Jiangsu Collaborative Innovation Center on Atmospheric Environment and Equipment Technology (CICAEET). We declare that we do not have any conflicts of interest to this work.

References

1. Lee, W., Stolfo, S.J., Mok, K.W.: Mining audit data to build intrusion detection models. In: International Conferenc e on Knowledge Discovery and Data Mining, pp. 66–72 (1998)
2. Silva, J.A., Faria, E.R., Barros, R.C.: Data stream clustering: a survey. ACM Comput. Surv. **46**, 125–134 (2013)
3. Aggarwal, C.C, Han, J., Wang, J., Yu, P.S.: A framework for clustering evolving data streams. In: International Conference on Very Large Data Bases, VLDB Endowment, pp. 81–92 (2003)
4. Guha, S., Meyerson, A., Mishra, N.: Clustering data streams: theory and practice. IEEE Trans. Knowl. Data Eng. **15**, 515–528 (2003)
5. Chen, Y., Tu, L.: Density-based clustering for real-time stream data. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, DBLP, San Jose, California, USA, pp. 133–142, August 2007