Application research of a data stream clustering algorithm in network security defense

Canshi Zhu¹, Xiaoyang Wang², Lin Zhu*³

¹Information Engineering College, Xijing University, Xi’an 710123, China
²Xi’an Satellite Control Center, Xi’an 710043, China
³Equipment Management and Safety Engineering College, Air Force Engineering University, Xi’an 710051, China

*Corresponding author and e-mail: Lin Zhu, 3399241158@qq.com

Abstract. The traditional intrusion detection system feature model is based on static data mining. Its mining algorithm relies on too many assumptions, which makes it difficult for intrusion detection systems to adapt to dynamic and real-time system detection requirements. Using attenuated sliding window technology, data stream mining technology and fusion technology with intrusion detection system, a data flow clustering algorithm based on attenuated sliding window is designed to improve and optimize the feature pattern extraction method of intrusion detection system to solve the dynamics of intrusion detection system. Through algorithm design, algorithm application and intrusion detection system simulation verification, the feasibility and accuracy of the algorithm and the optimized intrusion detection system are proved.

1. Introduction

Intrusion detection system is an effective network security defense strategy [1]. The method is to detect data packets that want to flow into the network under test, and review security threats to the defended system at the front end of the defended system [2]. It can block the access operations of the threat system to protect the protected network system resources from attack.

Intrusion detection system is simple and convenient to implement. It does not need to cross multiple physical network segments, but only monitors on ports [3]. It can passively and silently collect the information flowing into the network at the entrance of the protected network system, analyze and detect the behavior mode of entering the network. It allows access to the network for normal and authorized user operations, and prevents and counteracts user access for abnormal and unauthorized behavior, so as to achieve the purpose of network security defense [4]. The key link of intrusion detection system is the extraction of intrusion feature patterns or rules, and the accuracy of feature patterns or rules extraction. Timeliness and blocking ability are the key technologies of intrusion detection system [5].

Traditional intrusion feature extraction is based on static data analysis and mining [6]. This method needs periodic analysis and frequent updating of feature pattern library. Therefore, the traditional intrusion detection is less effective and dynamic, and the security defense effect is limited. The traditional intrusion detection or anomaly detection schematic is shown in Figure 1.
Commonly used intrusion detection is classified according to the extraction technology rules, such as intrusion detection technology based on expert system, intrusion detection technology based on neural network, intrusion detection technology based on immune algorithm and data mining-based intrusion detection technology [7-8].

The intrusion detection technology based on the expert system encodes the intrusion or attack behavior to form an expert system rule base. Each of these rules is a code for an intrusion scenario that detects intrusion behavior by matching the audit record with the rules.

The intrusion detection technology based on neural network is mainly neural network algorithm. The algorithm allows the neural network to simulate user phase behavior and adjust to new changes. The network receives the input event data and compares it with the historical behavior of the reference to determine the similarity or degree of deviation between the two.

The intrusion detection technology based on immune algorithm is to achieve the purpose of eliminating the dissident and maintaining the balance of the system through the affinity of the antibody and the antigen. The most basic ability of the immune system is to be able to identify tissues that belong to the normal body, and those that are not part of the normal body are abnormal.

Data mining-based intrusion detection technology applies data mining technology to intrusion detection. This method can comprehensively and effectively audit data and obtain characteristic patterns of attack behavior, so as to accurately distinguish and capture actual intrusion and normal behavior patterns.

With the rapid development of network technology, network attack methods and technologies are constantly developing and improving [9]. Coupled with the increase of network bandwidth, the increase of network speed and the rapid increase of network traffic, these pose new challenges to network security defense. Traditional intrusion detection technology has been difficult to cope with this challenge. The main reason is that system feature extraction is implemented in a static data environment. The timeliness of feature patterns lags, and real-time processing capabilities cannot meet system requirements [10]. Therefore, how to dynamically extract feature patterns in massive data streams not only solves the accuracy of feature patterns, but also improves the timeliness of pattern extraction, which is the key to improving the performance of intrusion detection systems.

2. Analysis of Data Stream Clustering Mining Algorithm
The traditional clustering mining algorithm is based on static data, its data will be reused and updated in small amount in real time. It relies on too many assumptions, which makes it difficult for intrusion detection systems to adapt to dynamic and real-time system detection requirements.
Data streams are characterized by continuous fast, short-lived, unrecoverable, and unpredictable. Therefore, designing data stream mining algorithms should solve the following technical problems.

Firstly, the algorithm requires solving the problem of unlimited data volume mining with limited memory resources. The traditional data is limited, the storage space can be satisfied, but the data stream has an unlimited amount of data, and the memory capacity cannot be satisfied.

Secondly, the algorithm has high requirements for time complexity. Traditional data mining is carried out in static data environment, and multiple iterations can be performed using traditional data mining algorithms. However, data stream mining is performed in a dynamic data network environment. The data stream is irreversible and the data can only be processed once.

Thirdly, the data stream is a specific type, requiring the data stream mining algorithm to adapt to changes. For example, as time progresses, the knowledge or models that are mined should also change constantly. This can be called concept drift or evolution. As far as the data stream mining algorithm is concerned, it can be divided into clustering and classification. This paper is based solely on clustering algorithm analysis. The clustering algorithm generally divides the mining objects into N groups. The similarity of the objects in the group is as high as possible, and the similarity of the objects outside the group is as low as possible. Typical data stream clustering algorithms are k-means, CluStream, and DBSCAN algorithms.

The k-means algorithm first determines the k-value, and divides the data object into k groups. Each group of data objects is not empty. Each object belongs to a group if and only belongs to one group. Each group forms a cluster, and the algorithm loops to find the optimal cluster center.

The CluStream algorithm is an evolution of the data stream clustering algorithm, which mainly deals with the clustering of data streams in different time domains. There are two stages in the process of dealing with data stream which may be changing constantly. One is the online maintenance of candidate micro-clusters, the other is the offline generation of clustering results.

The DBSCAN algorithm is a density clustering algorithm, which uses data space D as a data class and D as a high-density object region. If we want to study a point p of the space D, and use p as the core, query the neighborhood of the D area, and all the points in the neighborhood and p belong to the same class. These points will be the next object of investigation. In this iterative way, the class is constantly expanded until a complete class is found.

The research content of this project comes from the part of the team of Professor Zhu Canshi of Xijing College in 2019, supported by the Shaanxi Provincial Science and Technology Department, China, "Big Data Security and Application Research in the Age of Sharing". Aiming at the shortcomings of data stream clustering algorithm, this paper proposes and designs a clustering mining algorithm based on attenuated sliding window technology, which improves the clustering quality and efficiency of traditional clustering algorithm.

3. A Data Stream Clustering Algorithm Based on Attenuated Sliding Window (ASWD Stream)

Data stream mining is a new technology for dynamically extracting feature patterns in data streams. The core technology is data stream mining algorithm, which is based on the full analysis of classical clustering algorithms.

The processing idea of the algorithm is to analyze the critical candidate micro-cluster and determine whether it will be converted into a candidate micro-cluster or an out-of-group micro-cluster. Clustering is performed in a sliding window. The sliding window is divided into N sub-windows. Firstly, clustering operations are performed on each sub-window to obtain N candidate micro-cluster sets, and then candidate micro-clusters are used instead of specific records for processing. Thereby reducing the impact of the order of data inflow on the clustering results.

The algorithm sequentially reads the network data into each sliding sub-window, and performs clustering in each sliding sub-window to obtain a plurality of candidate micro-clusters and a plurality of critical candidate micro-clusters. The candidate micro clusters are saved into the corresponding set, and the critical candidate micro clusters are saved to another corresponding set. After all the sub-windows in the sliding window have been processed, the saved critical candidate micro clusters and
candidate micro clusters are processed. After obtaining the micro-cluster in the sliding window, the algorithm merges the micro-clusters into the pyramid structure that has been established in advance by using an offline method, and saves the current cluster cluster in a certain snapshot form. It uses a clustering result based on the preservation history of the attenuation model. After the sliding sub-window processing is completed, the candidate micro clusters and the critical candidate micro clusters in the set are processed. The processing flow of the data stream algorithm is shown in Figure 2.

4. Simulation Verification of Data Stream Clustering Algorithm (ASWD Stream)

The validation data set is based on the data provided by the 3rd international knowledge discovery and data mining competition (KDD CUP 1999). The KDD CUP 1999 data is highly authoritative in assessing the performance of intrusion detection systems.

When verifying, the following parameter settings are used. The initial data points are N=1000, the data flow velocity is v = 1000, the attenuation factor is \( \lambda = 0.25 \), \( \varepsilon_1 = \varepsilon_2 = 13 \), \( \mu = 10 \), and the outlier threshold is \( \beta = 0.2 \).

The data stream clustering algorithm (ASWDStream) and the CluStream algorithm use the same parameters as above. The clustering process of the data stream clustering algorithm (ASWDStream) after MATLAB simulation is shown in Figure 3.
Figure 3. Micro-cluster formation process of ASWD Stream algorithm

The two algorithms are compared in terms of time complexity. The simulation results of the two algorithms are compared as shown in Figure 4.

Figure 4. Running time comparison of the two algorithms

Through the above verification analysis, it can be seen that the ASWDStream algorithm is superior to the similar CluStream algorithm in terms of processing speed and clustering quality.

5. Application of ASWD Stream Clustering Algorithm in Intrusion Detection System

The intrusion detection system based on ASWDStream Clustering Algorithm is much more complicated than the traditional intrusion detection system. This is because it implements real-time mining, analysis of intrusion feature patterns, and real-time dynamic detection. After the inflow of external network information, the data stream clustering mining algorithm is first used to mine and analyze. For the suspicious feature pattern mined by the algorithm, the intrusion detection system will block the attack. Thereby realizing the security of the intranet system.

The design principle and structure of intrusion detection system based on data stream clustering mining algorithm are shown in Figure 5.
Figure 5. Intrusion detection system model based on data stream clustering mining

Firstly, the system performs sample mining on the normal data stream using the designed clustering algorithm (ASWDStream), and the obtained feature pattern is used as the normal model. Then, the intrusion detection system is laid out, and the real-time data flow into the network system is mined and analyzed by using the clustering algorithm (ASWDStream). The mining result is passed through the system and matched in real time with the normal mode. If the matching detection result is normal, the data stream can enter the network system, otherwise the data stream is blocked from entering the system.

Of course, the intrusion detection system must analyze the suspicious mode and continuously update and optimize the normal feature pattern library.

6. Intrusion Detection System Simulation Based on Data Stream Clustering Mining Algorithms

There are 7000 training samples, including 7 types of invasion. They are neptune, pod, ftp_write, spy, perl, buff_overflow, ipsweep, respectively.

There are 3000 test samples, including 9 types of invasion. They are neptune, pod, ftp_write, spy, perl, buff_overflow, eject, ipsweep, nmap, respectively.

The Distribution of test data types is shown in Table 1.

Table 1. Distribution of test data types

| Data type | Number of training samples | Percentage of training samples | Number of test samples | Percentage of test samples |
|-----------|---------------------------|-------------------------------|------------------------|---------------------------|
| Normal    | 4743                       | 67.76%                        | 1867                   | 62.23%                    |
| DoS       | 786                        | 11.23%                        | 345                    | 11.5%                     |
| R2L       | 653                        | 9.32%                         | 289                    | 9.63%                     |
| U2R       | 456                        | 6.51%                         | 168                    | 5.60%                     |
| probing   | 362                        | 5.17%                         | 331                    | 11.03%                    |
| Total     | 7000                       | 100%                          | 3000                   | 100%                      |

The Test results and Statistical results are shown in Tables 2 and 3.
Table 2. Test results

| Detected sample set | Training sample set | Test sample set |
|---------------------|---------------------|-----------------|
| Accuracy            | 99.87%              | 98.23%          |
| Misinformation rate | 0                   | 0.57%           |
| Misreporting rate   | 0.186%              | 1.13%           |

Table 3. Statistical results

| Intrusion type | Number of training samples | Number of test samples | Detection rate |
|----------------|-----------------------------|------------------------|-----------------|
| Dos            | 754                         | 231                    | 30.64%          |
| neptune        | 32                          | 114                    |                 |
| R2L ftp_write  | 432                         | 143                    | 33.10%          |
| spy            | 221                         | 146                    |                 |
| U2R perl       | 320                         | 65                     | 20.31%          |
| buff_overflow  | 136                         | 32                     |                 |
| probing        | 362                         | 178                    | 49.17%          |
| nmap           | 0                           | 153                    |                 |

From the simulation analysis, after applying ASWDStream algorithm to the intrusion detection system based on data stream mining, the intrusion detection system shows good detection performance, such as high detection rate, low misinformation rate and misreporting rate. It has better recognition ability for specific intrusion types, especially for R2L and probing intrusion types, which have high detection rate. In addition, it has better detection ability for new intrusion types.

7. Conclusion

In this paper, in view of the fact that the feature extraction of traditional intrusion detection system cannot adapt to the dynamic and real-time system detection requirements, based on the analysis of traditional intrusion detection system and data stream mining technology, a data stream clustering algorithm based on attenuated sliding window (ASWD Stream) is designed. The algorithm is integrated with intrusion detection system to design an intrusion detection system based on data stream mining, and the system is simulated and verified. The simulation results show the feasibility of data stream clustering algorithm based on attenuated sliding window. The intrusion detection system based on data stream mining constructed by this algorithm solves the problem that the feature pattern extraction of traditional intrusion detection system cannot adapt to dynamic and real-time system detection.

Acknowledgment

This work is supported by the Key Project of Soft Science Research Program of Department of science and technology of Shaanxi Province, China, No. 2019KRZ014.

References

[1] Zhu Lin, Zhu Canshi. Data stream sliding window clustering algorithm applied in IDS [J]. Computer Engineering and Applications, China. 2014, 50(1):87-90.
[2] Sun Ruxia,Zhang Sun,Yin Chunyong,Wang Jin,Min Seungwook.Strategies for data stream mining method applied in anomaly detection [J]. Cluster Computing. 2018.
[3] Phan H, Le B. A Novel Parallel Algorithm for Frequent Itemsets Mining in Large Transactional Databases [M]// Advances in Data Mining. Applications and Theoretical Aspects. Springer, Cham, 2018.
[4] Zymbler M. Parallel Algorithm for Frequent Itemset Mining on Intel Many-core
Systems [J]. 2018.

[5] V. R. B. Rohini, Dr. G. P. Saradhi Varma. FIDOOP-DP: Implementation of Data Partitioning in Frequent Itemset on Bigdata using Hadoop Pseudo Distributed Environment [J]. International Journal of Scientific Research in Science, Engineering and Technology. 2017, 8(3): 663-668.

[6] Liu Y, Ding W. A KNNS based anomaly detection method applied for UAV flight data stream[C]// 2015 Prognostics and System Health Management Conference (PHM). IEEE, 2015.

[7] Wang B, Liu D, Wang W, et al. A hybrid approach for UAV flight data estimation and prediction based on flight mode recognition [J]. Microelectronics Reliability, 2018, 84:253-262.

[8] Shabtai A. Anomaly Detection Using the Knowledge-based Temporal Abstraction Method [J]. 2016.

[9] Bangalore P, Letzgus S, Karlsson D, et al. An artificial neural network-based condition monitoring method for wind turbines, with application to the monitoring of the gearbox [J]. Wind Energy, 2017.

[10] Ahmed R, Srinivasa Pai P, Sriram N S, et al. Comparison of wavelet based denoising schemes for gear condition monitoring: An Artificial Neural Network based Approach [J]. IOP Conference Series: Materials Science and Engineering, 2018, 310:012010.