Traffic Clustering Method Based on Adaptive Density Grid

Haichuan Sun\textsuperscript{1,a}, Liang Gu\textsuperscript{1,b}, Yaning Meng\textsuperscript{1,c}, Haoguang Xin\textsuperscript{2,d}, Chenni Dong\textsuperscript{1,e}, Xin Ma\textsuperscript{1,f}, Xixi Zhao\textsuperscript{1,g}, Liguo Wang\textsuperscript{1,h}, Xiaorong Duan\textsuperscript{1,i}, Xin Gong\textsuperscript{1,j} and Sheng Bi\textsuperscript{1,k}

\textsuperscript{1}Information and Communication Branch of State Grid Shanxi Electric Power Company, Taiyuan, Shanxi, 030000, China.
\textsuperscript{2}College of Information and Computer, Taiyuan University of Technology, Taiyuan, Shanxi, 030000, China.

\textsuperscript{a}email: sunhaichuan1993@gmail.com, \textsuperscript{b}email: Guliang_1998@163.com, \textsuperscript{c}email: xieminqi@hotmail.com, \textsuperscript{d}email: iambishopshawn@gmail.com, \textsuperscript{e}email: dongchenni123@163.com, \textsuperscript{f}email: max0202@163.com, \textsuperscript{g}email: zxx9307@126.com, \textsuperscript{h}email: wlg402215289@163.com, \textsuperscript{i}email: gongxin0507@163.com, \textsuperscript{j}email: blueseavic@163.com

*Corresponding author’s e-mail: duanxr0821@163.com

Abstract. At present, the scale of internal network of some enterprises is expanding with the development of enterprises. The terminal of enterprises is becoming more and more diversified in terms of the number of types and the scope of application. The form of terminal security protection is becoming more and more severe. In this paper, based on adaptive adjust the density of the grid flow clustering method, by aiming at characteristics of the grid density value difference reflect the characteristics of the data of uncertainty on the influence of density threshold, adjustment of adaptive threshold density, determine the key characteristic vector, and the flow characteristics and the agreement that certain learning training, finally realizes the flow clustering.

1. Introduction
The data clustering method in correlation technology can only cluster data sets of fixed size, and the data structure stored in memory will be released after clustering in correlation technology, and the data will be read from the data set again for clustering in the next clustering, which can't carry out accurate real-time clustering analysis for time sensitive data. It is especially not suitable for network traffic analysis and clustering with large amount of data. Therefore, streaming clustering algorithm is widely used in clustering analysis of large amount of real-time data. There are many implementation algorithms for stream clustering algorithm. Compared with other stream clustering algorithms, the grid density based algorithm has some advantages. For example, compared with the partition based algorithm, the grid density based algorithm can adapt to the evolution characteristics of data stream, and can identify clusters of arbitrary shape; in addition, for hierarchical clustering, although the algorithm can adapt to the evolution of data stream, the distance based clustering is not as good as the algorithm based on grid density; in addition, compared with the density based clustering algorithm, the grid density based
algorithm uses the method of dividing the data grid when processing the data, and then carries on the grid mapping to the new inflow data. When updating the grid, it does not update all the grid density, but updates the data mapped grid, and in the later operation, it processes the grid instead of the data. Therefore, compared with other algorithms, it has advantages in running efficiency. Therefore, the clustering algorithm based on grid density is more suitable for real-time clustering analysis of network traffic. However, the speed and efficiency of data stream clustering algorithm based on density grid have been greatly improved, but its essence is to improve the density grid method. Therefore, for the burst of network traffic and the continuous entry or exit of terminal devices in the network, the original data flow clustering algorithm based on density grid has some defects. For example: 1. It is difficult to set the density threshold parameter of the grid cell, and it is difficult to set the parameter properly for the burst traffic and the emerging new terminals in the network; 2. The boundary of cluster is difficult to be accurate, and the grid is used to compress and store the data, which will lose the distribution information of the data in the grid cell and easily lead to the deviation of clustering results; 3. With more and more factors involved in target analysis, terminal devices constantly enter or exit the network, which will lead to the continuous change of traffic characteristics, resulting in individual terminals being treated as irritable points, which affects the efficiency of clustering to a certain extent. However, high-quality, efficient and secure enterprise network and terminal environment is an important guarantee for the good development of enterprises. However, the network monitoring means in related technologies have been unable to meet the needs of some enterprises for real-time monitoring, rapid identification and timely interruption of internal terminal network behavior. Compared with home users, the value of data and other assets of enterprise terminals is higher. The internal network composed of terminals, servers and other different software and hardware brings more complex virus sources, infection and transmission channels. Therefore, enterprise users are facing more severe network behavior security challenges of end users, and have more stringent requirements on protection, management, application and other aspects.

In the traffic classification method, Williams [1] and others compared five classification algorithms, namely Bayesian network, naive Bayesian discretization, naive Bayesian core density estimation, naive Bayesian tree and C4.5. Among the supervised methods, C4.5 is the fastest. The paper [2] in order to reduce the amount of computation in the process of anomaly detection, a method based on empirical mode decomposition (EMD) is proposed, which uses weighted self-similar parameters to realize traffic detection. Bernaille [3] and others extracted the characteristics of TCP packets, and used K-means clustering method to detect abnormal traffic. In the field of encrypted traffic classification, korczynski [4] uses Markov chain to model SSL / TLS packets in the traffic generated from server to client, and realizes the encrypted traffic classification method.

How to accurately monitor and predict the behavior of internal network users through real-time analysis and detection of internal network traffic is a serious problem faced by the enterprise internal business construction and management. Therefore, it is the key to find an efficient and accurate network traffic analysis method, which can adapt to the changes of network state and traffic. We propose an adaptive method to adjust the density parameters of the network, which reflects the influence of the uncertain characteristics of the data on the density threshold through the density difference of the feature grid at the previous survey time of the current survey time, It can effectively avoid the inaccuracy and deviation of clustering results caused by the global uniform density threshold in the grid based density clustering algorithm. By setting the key feature vector, the new terminal vector is not compressed and eliminated, so as to adapt to the behavior portrait of the terminal in the LAN or to monitor the abnormal network traffic behavior.

2. Main methods
At present, the traditional clustering methods mainly have inaccurate clustering methods. This paper proposes a network traffic clustering method which can adjust the density grid adaptively and dynamically. It can be used for network terminal behavior portrait or network traffic abnormal behavior monitoring. Firstly, the training samples composed of characteristic traffic are marked with traffic
characteristics and traffic types such as IP address, TCP / UDP port number, TCP session flag and packet length, and the key feature vectors are determined. Then the traffic characteristics and traffic type labels are input into the clustering model for training. By meshing the vector space of each feature domain and training the density distribution of traffic on the grid, the feature traffic detection model (preset model) is formed. Then, the IP address, TCP / UDP port number, TCP session flag and other features are extracted from the real-time original traffic data collected from the network, and the extracted feature vector features are input into the detection model. The model classifies the input traffic features according to the grid density distribution of each feature vector space. According to the classification result deviation, the grid adjustment interval and density threshold are adjusted in real time to further improve the detection accuracy.

This paper proposes a clustering method of network traffic with adaptive dynamic density adjustment grid, which is divided into four modules: data acquisition module, first clustering module, adjustment module and second clustering module (with graph). The data acquisition module is used to obtain the first time feature vector of the network traffic characteristics of the local area network. The network traffic characteristics include: source MAC address, destination MAC address, data frame protocol identification, source IP address, destination IP address, IP packet type, TTL, TOS, IP packet fragmentation identification, IP packet fragmentation offset value, IP extension field, source TCP port number, IP address, IP address, IP address, IP packet type, TTL and TOS Destination TCP port number, TCP connection identification, TCP receiving window value, TCP confirmation field, TCP serial number field, source UDP port, destination UDP port, data length, protocol content characteristic value and IP packet acquisition time; The first clustering module is used to cluster the feature vectors at the first time according to the density threshold of the grid to obtain the first clustering result; The adjustment module is used to adjust the grid spacing and density threshold according to the first clustering result; The second clustering module is used to cluster the feature vectors of the network traffic characteristics of the local area network at the second moment by using the grid after adjusting the spacing and density threshold to obtain the second clustering result.

The specific implementation is shown in Figure 1, obtain the feature vector of network traffic features at the first moment, and then standardize the range of the feature vector at the first moment to get the processed feature vector: cluster the processed feature vector according to the density threshold of the grid, and standardize the range of the feature vector at the first moment to get the processed feature vector. According to the density threshold of the grid, the feature vectors at the first time are clustered. Using the preset model, the density threshold of the grid is adjusted according to the first clustering result, and the density threshold of the grid is adjusted. The second clustering result is obtained by clustering the feature vectors of the network traffic of the enterprise LAN at the second moment (that is, the next moment of the first moment) using the grid after adjusting the spacing and density threshold.
3. Experimental Evaluation
The training process of the experimental model is shown in Figure 2, which is divided into six steps. First, the network traffic is extracted to confirm the key vectors. The features mainly come from the network layer and the transport layer. Then these data features are quantified.

3.1 experimental data
In order to evaluate the traffic clustering method proposed in this paper, the data set selected in the experiment is the open source data set ISCX2012 [5], which is the intrusion detection data set. Compared with KDD99, the attack traffic data contained in this data set tends to the real environment, and the attack types are more comprehensive. The data set contains five types of seven days traffic data, the traffic types are normal, infiltriting, httpdos, DDoS, sshbruteforce, and the data type is pcap package.
3.2 Experimental steps

3.2.1 Data input
The input network traffic is extracted according to the data packet. The key feature vector is determined and the key feature vector is established. The specific extracted features include: source MAC address, destination MAC address, data frame protocol identification, source IP address, destination IP address, IP datagram type, TTL, TOS, IP packet fragmentation identification, IP packet fragmentation offset value, IP extension field, source TCP port number, destination TCP port number, TCP connection identification, TCP receiving window value, TCP confirmation field, TCP sequence number field, source UDP port Destination UDP port, data length, protocol content characteristic value (ARP, ICMP, DNS, HTTP, etc.), packet acquisition time.

3.2.2 Data feature quantification
After feature extraction, the data is added to the range standardization processing, that is, the average value of the data is subtracted from the corresponding value of each dimension, and then divided by the range to calculate. After the range standardization transformation, the corresponding mean value of the sample is zero, so as to reduce the error of data processing. The IP address is quantized according to 32-bit integer value, and the non numeric text type is quantized according to the text classification enumeration number. The specific range standardization formula is as follows:

\[ X_{ij}' = \left( \frac{X_{ij} - \bar{X}_j}{\max X_{ij} - \min X_{ij}} \right) \]

Among them, \( X_{ij} \) represents the j-th eigenvector in the i-th eigenvector before processing, \( \bar{X}_j \) represents the mean value of the i-th eigenvector before processing, \( \max X_{ij} \) represents the maximum value in the i-th eigenvector before processing, \( \min X_{ij} \) represents the minimum value in the i-th eigenvector before processing, and \( X_{ij}' \) represents the j-th eigenvector in the i-th eigenvector after processing.

3.2.3 Network mapping
The grid mapping of eigenvector is processed by time window, that is, every gap time interval, the data in the time window is meshed and mapped. When new data arrives, it is stored in the time window first, and then processed when the next gap arrives. The specific treatment process is as follows:
1. A density space is established for each eigenvector.
2. The parameter \( K (K \geq \) the number of flow types) is determined, and the density space is divided into \( k \) grids.
3. The input eigenvalue is mapped to the grid.
4. The gap is calculated according to the formula.

\[ \text{gap} = \log_{\frac{N}{M}} \left( \frac{C_L}{C_M} \cdot \frac{N - C_L}{N - C_M} \right) \]

\( N \) is the amount of input data, \( M \) is the number of features, \( CL \) is the sparse grid density threshold, \( CM \) is the dense grid density threshold.
5. Judge whether the gap time interval is reached at this time. If the sparse grid of the key feature vector space has been detected, if a new manic point appears in the sparse grid of the key feature vector space, and the manic point does not appear in the key feature vector list, then keep the sparse grid and add the manic point to the key feature vector list. Otherwise, the sparse mesh is deleted and the mesh density is adjusted.

3.2.4 Flow detection
After mapping, the input traffic is classified by detection formula, and the classification result is returned. The specific detection formula is as follows:
\[ D_m = \frac{\sum_{i=1}^{n} A_i \ast \sum_{i=1}^{n} B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \ast \sqrt{\sum_{i=1}^{n} (B_i)^2}} > 0.5 \quad (3) \]

A is the grid density centroid value of the input traffic feature, B is the grid density centroid value of the traffic feature of the training set, and M is the number of features.

3.2.5. Threshold parameter adjustment
After a gap interval data detection, the sparse grid density threshold is adjusted according to the detection deviation.

\[ D_m(t) = \frac{D_m(G, t)}{k(1 - 2^{-\beta(t-m+1)})} \quad (4) \]

\( t_o \) is the gap time nearest to the current observation time \( t \), \( D_m(t) \) is the dense grid threshold value of grid cell \( g \) at the current time \( t \), \( D_{en}(G, tu) \) is the grid density value of grid cell \( G \) at the previous time, and \( K \) is the number of grid cells, \( \beta \) is the attenuation coefficient.

3.2.6. Detection model generation
When the training set is used, the grid density distribution of each feature vector is classified according to the traffic type, and the classified grid density distribution is saved in the database for detection.

3.3 experimental result
The method proposed in this paper is used for cluster recognition of iscx2012 data set, and the recognition results are shown in the following figure: through the confusion matrix, the accuracy and recall rate of each group of experiments are calculated and statistically analyzed, and the calculation method is as follows, in which TP is true positive, TN is true negative, FP is false positive, FN is false negative.

\[ \text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \]
\[ \text{Recall} = \frac{TP}{TP + FN} \times 100\% \quad (5) \]

(a) Accuracy
Recall

Figure 3 Recognition results

It can be seen from the figure that the recognition accuracy of this method for five types of traffic is more than 90%, especially for normal type traffic, the recognition accuracy is more than 98%. The average accuracy of final recognition is 97.8%. Therefore, the proposed method can be well applied to traffic clustering identification.

4. Conclusion

This paper proposes a clustering method of network traffic, which can adjust the density grid adaptively and dynamically. It can be used for network terminal behavior portrait or network traffic abnormal behavior monitoring. By meshing the vector space of each feature domain and training the density distribution of traffic on the grid, the feature traffic detection model is formed. Then, the real-time original traffic data collected from the network is extracted, and the extracted feature vector features are input into the detection model. According to the grid density distribution of the input traffic features in each feature vector space, the model classifies them. According to the classification result deviation, the grid adjustment interval and density threshold are adjusted to further improve the detection accuracy. This method realizes the clustering recognition of malicious traffic, but the recognition granularity can only reach the type of traffic, and can not identify the specific attack method. In the future work, we will achieve more fine-grained traffic clustering recognition.

REFERENCE

[1] WILLIAMS N, ZANDER S, ARMITAGE G. A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification [J]. ACM SIGCOMM Computer Communication Review, 2006, 36(5):5-16

[2] Han Jieying, Zhang J Z. Network Traffic Anomaly Detection Using Weighted Self-similarity Based on EMD[c]//Proceedings of 4th IEEE International Conference on Software Engineering and Service Science. Washington D. C. USA: IEEE Press. 2013: 23-25.

[3] Laurent B, Renata T, Ismael A, et al. Traffic Classification on the Fly[J]. SIGCOMM Comput. Commun. Rev. 2006, 36(02):23-26.

[4] KORCZYNSKI M, DUDA A. Markov Chain fingerprinting to classify encrypted traffic[c]//IEEE INFOCOM 2014 IEEE Conference on Computer Communications. IEEE, 2014.

[5] Ali Shiravi, Hadi Shiravi, Mahbod Tavallae, Ali A. Ghorbani, Toward developing a systematic approach to generate benchmark datasets for intrusion detection, Computers & Security, Volume 31, Issue 3, May 2012, Pages 357-374, ISSN 0167-4048, 10.1016/j.cose.2011.12.012.