ATLANTIS PRESS
Advances in Engineering Research (AER), volume 61
International Conference on Mechanical, Electronic, Control and Automation Engineering (MECAE 2017)

Abnormal Traffic Classification based on Feature Entropy Vector

Lulu Chen [a], Wenpu Guo [b], Hao He [c]
Xi'an High - Tech Research Institute, Xi'an 710025, China
[a]879784162@qq.com, [b]879784162@163.com, [c]460827753@qq.com

Keywords: Information Entropy, Feature Vector, Anomaly Detection, Classification.

Abstract. Existing anomaly detection technology is mainly concerned with the detection of anomalous flow, and it is not enough to study anomaly type. Therefore, a method based on information entropy and k-means clustering is proposed to construct the anomalous traffic entropy feature vector to achieve fast and accurate judgment of anomaly types. The method is simple and easy to operate. Simulation results show that the proposed method is effective in classifying and determining the common types of network attack anomalies.

1. Introduction

With the continuous development of network scale, network attacks become more and more complex, more and more types. Facing the increasingly complex network situation, it is very necessary to distinguish different types of network attacks quickly and accurately. Anomaly classification is based on abnormal traffic detection and abnormal flow identification, and classifies the abnormal traffic and finally diagnoses the root event type that causes the exception to occur. There are many types of network anomalies, such as port scanning, DDoS attacks, worms and other malicious behaviors. They can also be caused by network mismatch and fault (such as link failure, Routing problems, device buffer overflows), and even legitimate behavior such as unreasonable use of network bandwidth.

In order to speed up the classification speed and improve the detection precision, many kinds of abnormal traffic classification methods have been proposed, and the entropy-based detection method is one of the most important ones. In order to improve the accuracy of the entropy computation, and there is sufficient accuracy to ensure. Anomaly detection based on entropy is a more fine-grained detection method than traffic-based anomaly detection. It can detect more covert anomaly types, and the effect of sampling on its accuracy is not obvious. The entropy sequence of each dimension in the network is correlated under both normal and abnormal conditions. Using the correlation between the dimensions can effectively reduce the false alarm rate and increase the accuracy of the classification (Cormode G et al, 2004).

In order to adapt to the characteristics of Internet traffic data and dynamic change of characteristic attributes, the use of data mining method to deal with abnormal traffic classification has become a new research hotspot in network security field. The object of the study is network flow, that is, a group of packets with the same five-tuple (transport protocol, source IP, destination IP, source port, destination port) From the point of view of data mining, the problem of anomaly classification can be abstracted as follows: how to use data mining method to discover hidden patterns between abnormal events and their characteristic attributes in the case of network anomaly set T and set of anomalous flows S with known type, and then constructs the classification model of the abnormal flow, and uses this model to classify the abnormal flow with unknown type(Xu et al ,2012). Core work includes two parts:

(1) Selecting the appropriate abnormal flow feature attributes to construct the attribute vector;

(2) Select the appropriate data mining algorithm to construct the classification model.

In this paper, we focus on two problems, namely, inefficient classification of anomalous traffic and low classification accuracy in the field of network anomaly detection. We use multidimensional entropy classifying method to make full use of the efficient entropy calculation algorithm and the correlation between entropy values which can effectively improve the classification efficiency and

accuracy. Firstly, based on the netflow anomaly traffic data, the entropy change characteristics of different types of anomalies are analyzed, then the k-means classification is used to construct the eigenvector of abnormal traffic types. Finally, the method of feature database and feature matching is used to identify the abnormal types quickly and accurately.

2. Related technology research

In recent years, there have been some progresses in the research of automatic classification of abnormal events, but there are still some problems. The association rules between different SNMP MIB variables are used to classify the abnormal traffic(Thottan M,2003).A method based on rule reasoning is proposed to distinguish different types of anomalies in the sampling flow(Kim M S et al,2004).

The cross - entropy is used to measure the change of traffic distribution of each attribute, and the eigenvector of the attacking behavior is established. The attack type is distinguished by the similarity between the anomaly vector and each characteristic eigenvector (Yan et al, 2010). However, this method only uses cross-entropy as a measure of feature attributes, cannot effectively distinguish the abnormal type, resulting in higher false positive rate.

The incremental K-means clustering algorithm is used to classify the sample points online (Qian, 2011). However, K-means clustering algorithm needs to set the number of clusters in advance, and the number of abnormal events in the network is usually unknown, resulting in the classification accuracy rate is reduced.

It is possible to identify worms by the error messages generated during the worm scanning (Bakos G et al, 2012). The accuracy of this method is high, but this packet-by-packet analysis method cannot meet the processing requirements of Internet mass data.

Comprehensive analysis of the above methods, the focus is to determine whether the flow anomalies, and for specific types of anomalies cannot be clear or poor classification. Therefore, it is of practical significance to construct the eigenvector of anomaly type based on the data mining method.

3. An Efficient Traffic Classification Method Based on Feature Entropy Vector.

3.1 Multi-dimensional entropy vector

Information entropy is a concept used in information theory to measure the amount of information. Information entropy marks the amount of information contained, which is a description of system uncertainty. The information entropy is used to describe the characteristics of network traffic, and the traffic data preprocessing not only enhances the detection capability of network traffic anomalies, but also facilitates the classification of traffic anomalies. Taking the Netflow data collected as a discrete information source, the information entropy can be analyzed by considering each attribute in the data as a set of random events. Suppose that the set of destination IP in a certain period is denoted by X, and i means that there are i different destination IPs. n_i represents the number of times that the ith destination IP occurs. $X=\{n_i, i=1,2,3,\cdots N\}$ Then the information entropy of the destination IP (X) in this period is defined as the formula 1, this paper uses the information entropy of the indicators active / destination IP, source / destination port. The calculation method of different index entropy can refer to formula (1). Table 1 shows the impact of different types of anomalies on the entropy of each attribute.

$$H(X)=-\sum_{i=1}^{N}(\frac{n_i}{S})\log 2(\frac{n_i}{S}),\quad S=\sum_{i=1}^{N}n_i \tag{1}$$

The multidimensional entropy vector is N_i . $N_i=(H_{X_1},H_{X_2},H_{X_3},H_{X_4})$ For the numerical properties of samples, different attribute characteristics have different metrics, therefore, each feature in the distance measurement will result in the clustering results to varying degrees. In order to eliminate the impact of different measures on clustering, the measured data should be normalized.

Table 1. Common abnormal flux entropy change characteristics

Type	Description	Entropy Characteristics
DOS	Single source attack on single (multi) source points	Source IP entropy decreases, Destination IP entropy increases
DDOS	Attack on Single Source Point by Multi - Source	Source IP entropy increases Destination IP entropy decreases
Port Scanning	A single source sends a large amount of port scan information to a single source for a short time	Source IP entropy decreases, Destination IP entropy decreases Destination port entropy increases
Worms	A small number of destination ports on multiple destination hosts are detected maliciously	Source IP entropy decreases Destination IP entropy increases Destination port entropy decreases

3.2 K-means clustering method

Clustering algorithm is a process of dividing a data set into several clusters. The data in the same cluster has high similarity, but the data in different clusters do not have similarity. Similar or dissimilar as measured by the attribute values of the descriptive data, a distance-based approach is usually used. Clustering can find dense and sparse areas of data, discovering the distribution patterns of data as a whole, and the meaningful association between data attributes.

The k-means algorithm is an unsupervised clustering algorithm, which uses the sum of squares of errors as the clustering criterion and classifies the data by using a clustering center parameter. It initially defines k centroids and finds the data that is closer to the centroid. The centroid is recalculated based on the newly added data until the centroid of all the clusters is no longer transmitted, ie, the objective function is minimized.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^j - c_j \right\|^2 \tag{2}$$

In the formula (2), $\left\| x_i^j - c_j \right\|^2$ represents the square of the distance between all data points x_i^j and c_j, and c_j represents the center of n data points in the same class. It has the advantages of fast calculation speed, simple method, and flexible adaptability to complex and variable requirements. For small data sets have good processing power.

3.3 Classification of feature vectors based on entropy

The distribution of abnormal sample points in space has the clustering phenomenon. Considering the diversity and complexity of the anomaly types, this paper will use the classification method based on cluster analysis to determine the anomaly types. This method combines the k-means algorithm with the pattern matching of anomaly features. First, the clustering result set of the entropy space of the abnormal sample points is obtained by k-means clustering method. Then, the relationship between the mean and the standard deviation of the clusters in the four-dimensional entropy space is used to determine the anomaly type If the mean of the clustering result set is less than 3 times the standard deviation of the normal data set, the characteristic entropy is denoted as zero. If the average clustering result set is greater than the mean of the normal data set, and greater than 3 times the standard deviations, the characteristic entropy is denoted as 1.If the average clustering result set is less than the mean of the normal data set, and less than 3 times the standard deviation, the characteristic entropy is denoted as -1.

The network anomaly traffic sample set $T = C_1, C_2, \cdots, C_n$ in a certain time is known, where the traffic data at each sample instant can be represented by a multidimensional vector containing the m-network flow characteristics. Such as sample $C_1 = N_1, N_2, \cdots, N_t$, represents the multidimensional entropy vector $N_i = H_1, H_2, \cdots H_m$ of the anomaly type C_1 at t different instants. It can be seen from Table 1 that the change of information entropy of the same type of network anomalies is consistent, that is, the different multi-dimensional entropy vector N_i is also consistent with the

normal flow multi-dimensional entropy vector. The different multidimensional entropy vector N_i is expressed as a feature vector $S = x_1, x_2, \cdots x_m$, $x_i \in$ -1,0,1 . The specific steps are as follows: The specific steps are as follows:

Step1: Determine the source / destination IP and the source / destination port entropy of netflow format anomaly traffic set detected t times in the time period, and get the multi-dimensional entropy vector at different abnormal flow time.

Step2: The k-means clustering method is used to classify the anomaly set multidimensional entropy vector N_i into different classes (multidimensional entropy matrix).

Step3: Calculate the four-dimensional entropy mean value of each type of anomaly set and calculate the mean and standard deviation of the four-dimensional entropy of the normal data set.

Step4: Converting N_i to $S = x_1, x_2, x_3, x_4$, $x_i \in$ -1,0,1 for each class of anomalies according to the above-mentioned decision method.

Specific anomaly classification algorithm flow shown in Fig.1.

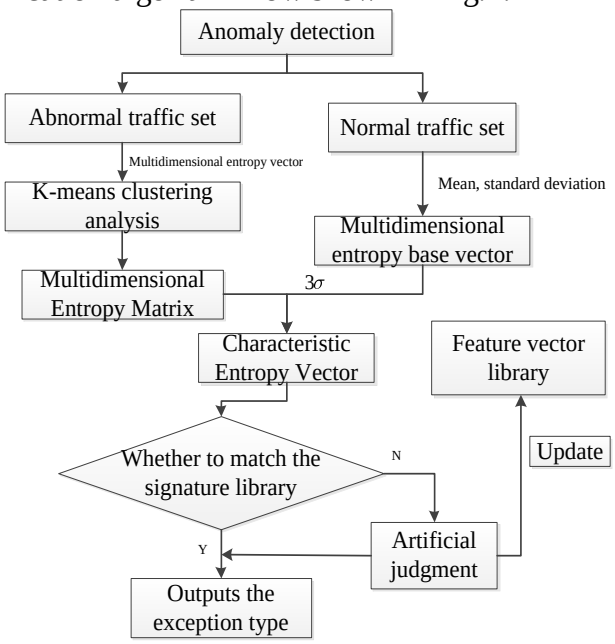

Fig.1 The flow of abnormal classification algorithm

4. Simulation experiment and analysis.

In order to verify the effectiveness of the proposed algorithm, the experimental data are obtained from the Abilene backbone network data collected from March 1, 2009 to March 7, 2009. The network is an IP backbone network in the United States. Each node corresponds to a city. It consists of 12 nodes, with 12×12 = 144 OD flows. All nodes are configured to collect NetFlow to collect the actual OD flow. The collection interval is 5 minutes, so the number of times collected in one week is 7×24×60/5=2016, and the number of OD flow is 144, so the OD flow matrix is a matrix of 144×2016.

But Abilene data set is not obvious abnormalities, so the use of artificial injection of abnormal flow method for experiment. In the experiment, the interval between sampling points is 5min, the abnormal injection process is as follows: from the 300th sampling point to the 800th sampling point, every 50 sampling points in turn injects a low-speed DDOS attack and a DDOS attack in turn, and The attack duration is 30min (continuous 6 sampling points), and the attack time is 30min (continuous 6 sampling points); from 1000th to 1500th sampling point, every 50 sampling points are injected with a port scan; From the 1700th to the 1900th sampling point, continuous injections of worm attacks, including infection to the outbreak of the complete process (continuous 200 sampling points).

The performance of the algorithm is evaluated by three comprehensive indexes.

(1) False positive rate: the total number of misclassified samples the proportion of the total number of samples. (2) Accuracy: the total number of samples correctly classified as the proportion of the total number of samples divided into categories. (3) The total accuracy rate: normal classification of the total number of samples the proportion of the total number of samples.

4.1 k-means clustering results

Fig.2. shows that black represents DoS attack and DDoS attack, red represents worm intrusion and blue represents port scan. DoS performance is that SrcIP entropy is small, DstIP entropy is small, and it is a single source point to a single point of attack match; DDoS performance is SrcIP entropy, DstIP entropy is small, with multiple source point to single source attack; The performance of the worm is the entropy value SrcIP small, DstIP entropy is large, with a small number of the sender through a port on a large number of targets sniffer match. Similarly, the network traffic anomalies in the three-dimensional entropy space also showed a significant aggregation phenomenon, although the abnormal flow in the three-dimensional entropy space distribution is very irregular, but with obvious aggregation phenomenon. Considering the diversity and complexity of anomaly types, we will use the four-dimensional space method to judge the network anomaly types. Such as port scanning attacks and worm intrusion, are only difficult to distinguish from the entropy values of the two stream signatures, since they both exhibit large DstIP entropy values and small DstPort entropy values, and therefore need to be added at the SrcIP The entropy size can be distinguished. Because the source / destination IP address and the source / destination port-based network flow feature describe the characteristics of the terminal host of the network flow, the anomaly entropy spatial distribution is affected by the four flow characteristics, so the anomaly classification method based on cluster analysis The four-dimensional space judgment method of network anomaly is used to discriminate the anomaly types.

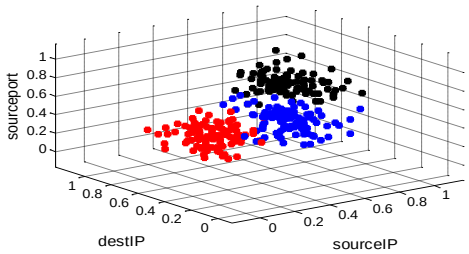

Fig 2 The k-means clustering effect in three-dimensional entropy space

4.2 Abnormal Classification Based on Feature Entropy Vector

Table 2. Abnormal features entropy vector library

ID	SrcIP	Srcport	DstIP	Dstport	Flag
1	0	-1	0	0	DOS
2	-1	1	0	0	DDOS
3	-1	-1	1	1	Port scanning
4	0	0	-1	1	Port scanning
5	-1	0	1	-1	worm
6	-1	0	0	-1	worm
7	0	0	0	0	unknown

Based on the anomaly detection of network traffic, according to the classification algorithm of network traffic abnormity designed in section 3.3, the abnormal classification results are obtained step by step. Firstly, according to the anomaly classification algorithm, the model library of initial information anomaly information entropy space is constructed (Wei et al, 2012)., as shown in Table 2. The four-dimensional network anomaly judgment method is mainly described by the mean and standard deviation of each measure in the clustering result set and the number of clustering samples. The patterns of the four measures in each cluster are compared with the patterns in the feature pattern library. The network anomaly classification is based on the anomaly detection, and the network traffic classification classification model is adopted. Table 3 shows the center position and standard deviation of the sample points in the four-dimensional residual entropy space at each latitude.

According to the clustering results, the cluster 1 matches the characteristic pattern 5 (-1, 0.1.-1), which is consistent with the decrease of the source IP entropy and the decrease of the destination port entropy when the destination IP entropy increases. Then the characteristic patterns of other clusters are obtained in the same order. Specific classification results are shown in Table 4. From the classification results can be seen to have achieved good results.

Table 3. Four-dimensional entropy spatial clustering results

Clustering	SrcIP	Srcport	DstIP	Dstport	Number of Sample Points	Mode ID
1	0.49±0.15	0.40±0.1	0.28±0.16	0.28±0.17	137	5
2	0.42±0.13	0.58±0.17	0.39±0.14	0.23±0.11	32	2
3	0.59±0.17	0.58±0.19	0.74±0.24	0.43±0.11	25	1
4	0.33±0.12	0.48±0.15	0.53±0.19	0.28±0.09	47	3
5	0.24±0.13	0.35±0.23	0.48±0.14	0.18±0.05	28	4
6	0.13±0.07	0.59±0.19	0.28±0.12	0.51±0.15	43	6
7	0.22±0.10	0.32±0.14	0.13±0.05	0.57±0.10	8	7

Table 4. Algorithm classification performance

Type	False Positive Rate	Accuracy	Overall Accuracy
DOS	0%	83.3%	90.6%
DDOS	6.4%	94.2%	
Port Scanning	11.2%	89.3%	
Worm	4.3%	90.5%	

4.3 Performance comparison with other algorithms

Fig.3. shows the comparison of the algorithm of this paper with the cross entropy algorithm and the k-means clustering algorithm. Algorithm 1 is the method of this paper, algorithm 2 is cross entropy method, and algorithm 3 is only k-means method.

It can be seen that the classification accuracy of this algorithm is better than K-means and cross-entropy at the same false alarm rate. This is because K-means clustering algorithm needs to set the number of clusters in advance, but the number of anomalies in the network is usually unknown. But only use cross-entropy as a feature of the metric cannot effectively distinguish between abnormal, resulting in a high false positive rate. In this paper, in the case of finite class anomaly, the algorithm does not depend on the k-choice, and the entropy eigenvector library is constructed to obtain good results.

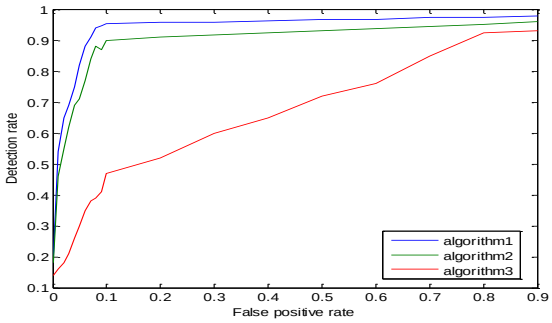

Fig. 3 Comparison of different algorithms

5. Summary

In this paper, we propose an eigenvector based on abnormal traffic entropy and k-means classification to construct anomalous traffic patterns and classify anomalies by feature matching. The simulation results show that more than 90% of abnormal events can be correctly classified. The disadvantage of this paper is that other entropy features are not fully utilized for abnormal traffic. The next step is to improve the feature vector library, on the one hand to increase the entropy

characteristics, on the other hand is a specific type of abnormal events for accurate classification, improve the feature library.

References

[1] Cormode G, Korn F, Muthukrishnan S, et al. Diamond in the Rough:Finding Hierarchical Heavy Hitters in Multi-dimensional Data[C]//Proceedings of SIGMOD'04. Paris, France: [s. n.], 2004:155-166
[2] Xu Qian, Chen Dong-nian. Network Traffic Anomaly Classification Algorithm Based on Hierarchical Clustering [J]. Computer Engineering, 2010, 38(23): 131-135.
[3] Thottan M, Ji C. Anomaly Detection in IP Networks [J]. IEEE Trans. on Signal Processing, 2003, 51(8): 2191-2204.
[4] Kim M S, Kang H J, Hung S C, et al. A Flow-based Method for Abnormal Network Traffic Detection[C]//Proceedings of IEEE/IFIP Network Operations and Management Symposium. [S. 1]: IEEE Press, 2004.
[5] Yan Ru-yu, Zheng Qing-hua.Application of cross-entropy detection and classification network abnormal flow [J] Journal of Xi'an Jiaotong University (Natural Science Edition) ,2010, 44 (61): 10-15.
[6] Qian Ye-kui, Chen Ming, Hao Qiang, et al.On-line detection and classification of the entire network (1): 111-120 [J].Journal of Communications, 2011, 32 (1): 111-120.
[7] Bakos G Berk V Early Detection of Internet Worm Activity byMetering ICMP Destination Unreachable Activity[C]//Proceedings of SPIE Conference on Sensors, and Command, Control, Communications and Intelligence. Orlando, USA: [s. n.], 2012: 33-42
[8] Wei Xiang-lin, Chen Ming, Zhang Guo-min,et al. NMF-NAD: detecting network-wide traffic anomaly based on NMF[J]. Journal on Communications, 2012, 33(4):54-61.