Ma, W., Tran, D., & Sharma, D. (2008). A Study on the Feature Selection of Network Traffic for intrusion Detection Purpose. In C. Yang, W. Chen, W. Hsu, & T. Wu (Eds.), *Proceedings of the IEEE Intelligence and Security Informatics Conference 2008 (ISI 2008)* (pp. 245-247). United States: IEEE, Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ISI.2008.4565069

A Study on the Feature Selection of Network Traffic for Intrusion Detection Purpose

Wanli Ma, *Member, IEEE*, Dat Tran, *Senior Member, IEEE*, and Dharmendra Sharma, *Senior Member, IEEE*

***Abstract*—The 3 most important issues for anomaly detection based intrusion detection systems by using data mining methods are: feature selection, data value normalization, and the choice of data mining algorithms. In this paper, we study primarily the feature selection of network traffic and its impact on the detection rates. We use KDD CUP 1999 dataset as the sample for the study. We group the features of the dataset into 4 groups: Group I contains the basic network traffic features; Group II is actually not network traffic related, but the features collected from hosts; Group III and IV are temporally aggregated features. In this paper, we demonstrate the different detection rates of choosing the different combinations of these groups. We also demonstrate the effectiveness and the ineffectiveness in finding anomalies by looking at the network data alone. In addition, we also briefly investigate the effectiveness of data normalization. To validate our findings, we conducted the same experiments with 3 different clustering algorithms - K-means clustering, fuzzy C means clustering (FCM), and fuzzy entropy clustering (FE).**

***Index Terms*—Intrusion detection, Clustering methods, Feature extraction**

I. Introduction

IN GENERAL, there are two types of intrusion detection systems (IDS): signature based IDS and anomaly detection based IDS. Signature based IDS are reactive. Intrusion patterns have to be provided beforehand, and the system always legs behind the new attacks. On the other hand, anomaly detection based IDS promises proactive detections through continuously machine learning, with little human intervening. The learning process could be unsupervised just from network data or supervised from labeled data.

Many different types of technology have been proposed as the detection engines. Due to the highly irregular distribution of the network data, which has "power-law distribution" and is "one-sided and heavy tailed" [1], using clustering method is strongly advocated by a number of research groups. The 3 most important issues for anomaly detection by using clustering methods, and indeed any machine learning algorithm, are: feature selection, data value normalization, and the choice of algorithms.

In this paper, we study the impact of the feature selection of network traffic data on anomaly network traffic detection rates. We use KDD CUP 1999 dataset as the sample. We classify the fields of the dataset vectors into 4 groups. Group I fields are the basic network traffic attributes; Group II fields are actually not network traffic related, and they are from host based monitoring sensors; Group III and IV fields are time based attributes. We study the different combinations of these groups. To validate our findings, which are not just accidental under a set of one-off experiments, we choose 3 different clustering algorithms – K means clustering, fuzzy C-means clustering (FCM), and fuzzy entropy clustering (FE) – to conduct the same experiments. We are aware of the criticisms on claiming detection rates solely based on a single dataset [2]. Without losing the generality, we do not try to fine-tune the clustering algorithms to achieve the premium detection rates for this particular dataset. Our primary focus in this paper is the comparison of different detection rates achieved by selecting different features and the consistence of the comparison results under different algorithms.

The rest of the paper is as follows: in Section II, we briefly discuss the related work. Section III provides the background information. In Section IV, we study the impact of the different combinations of the features with the discussions of our observations. We conclude the paper with future work in Section V.

II. Related Work

All the proposals listed in this section use either KDD CUP 1999 dataset [4] or DARPA 1999 dataset [7]. By no mean do we inclusively take the account of all possible proposals.

In [3], Portnoy proposed to use a simple variant of single-linkage clustering method to learn network traffic patterns on unlabelled noisy data. The author made 2 assumptions: the number of normal activities is far larger than that of abnormal activities, and the sample data reflects the distribution of day to day network operation. It is not clear from the paper which fields (attributes) are used. The approach achieves 40%-55% detection rate with 1.3%-2.3% false positive rate.

NATE [5, 6] was proposed by Taylor and Alves-Foss. The approach is similar to Portnoy's, but the authors suggested that

Manuscript received January 30, 2000.

Wanli Ma, Dat Tran, and Dharmendra Sharma are with Faculty of Information Sciences and Engineering, University of Canberra, Australia (phone: +61.2.62012838; fax: +61.2.62015231; e-mail: Wanli.Ma@ canberra.edu.au).

ISI 2008, June 17-20, 2008, Taipei, Taiwan
1-4244-2415-3/08/$20.00 ©2008 IEEE

it might not be the case that the number of normal network activities are always far larger than that of abnormal activities. From the papers, it is unclear how and also if data normalization is carried out.

Chan et al [1] also used DARPA dataset for their clustering based IDS – CLAD (Clustering for Anomaly Detection). The clustering algorithm used is k-NN, and the training process is unsupervised. The authors first converted the symbolic values into digital values, and then normalized these values based on logarithm.

Li and Ye [8] proposed to use CCAS clustering algorithm, supervised clustering and classification. Instead of using the network traffic records, they used BSM audit records. Interestingly, they only used one attribute (event type).

Caruso and Malerba [9] tested Weka data mining tools (K-means and EM) on their firewall logs. The selected features are time stamps, protocol, destination IP, Source IP, Service port, number of packets, duration, and the country of source IP address. From the paper, it is unclear how symbolic values (e.g., protocols) are handled, and also how and if data normalization is carried out.

Wang and Megalooikonomou [10] proposed to use the Fuzzy-Connectedness Clustering (FCC) algorithm. They achieved 94% detection rate and a false alarm rate below 4% on the KDD CUP dataset. However, in the paper, they did not mention how many features were used, how symbolic values were processed, nor if there is any normalization.

III. Grouping KDD CUP 1999 Dataset

KDD CUP 1999 dataset was based on MIT Lincoln Lab intrusion detection dataset, also known as DARPA dataset. The raw network traffic records have already been converted into vector format. Each vector has 41 fields (features), Table I. We refer the readers to [4] and [11] for the meanings of the fields. In this paper, we ignore the fields with symbolic values, i.e., field 1, 2, 3, and 6. The rest of the fields are grouped into 4 groups:

- **Group I**: fields 0, 4, 5, and 7, these fields are the basic characteristics of a connection. They are the durations, the octets transferred, and wrong fragmentation flags of the connection.
- **Group II**: fields 10-19, these fields are actually not traffic features. The values cannot be obtained by looking at the traffic records alone. The help from host based logs is needed.
- **Group III**: fields 22-30, these fields are time based traffic features. They are the statistics of traffic features in the previous 2 seconds time window. The calculation is based on the source IP address.
- **Group IV**: fields 31-40: the same as Group III, except for that the calculation is destination IP address oriented.

IV. Experimental Results and Observations

The proposed method for the network intrusion detection was evaluated using the KDD CUP dataset for training and the "corrected" dataset for testing. Training sets for the 23 attacks were extracted from KDD CUP dataset and the maximum number of feature vectors for each of the training sets was set to 1000. All 311029 feature vectors in the testing set were used.

Because the feature values have different ranges, the following normalization of features is therefore used:

$$x'_{tj} = \frac{x_{tj} - \mu_j}{s_j} \tag{7}$$

where x_{tj} is the j-th feature of the t-th vector, μ_j the mean value of all T vectors for feature j, and s_j the absolute standard deviation, that is

$$s_j = \frac{1}{T}\sum_{t=1}^{T} | x_{tj} - \mu_j | \tag{8}$$

For each of the clustering algorithms, we trained 23 models for the 23 attacks using the training sets extracted from the KDD CUP dataset. We have conducted the experiments with 15 different combinations of Group I, II, III, and IV. Each individual experiment is conducted with the raw data and the normalized data.

There are a few interesting observations, take for example the run of all features (i.e., Group I, II, III, and IV) with K-means clustering algorithm. The recognition rates for the labels A (back), G (land), K (nmap), O (pod), and R (Satan) are pretty high. Labels A, G, and O are all denial of services attacks, and Labels K and R are port scanning activities. These types of activities have distinct network features.

TABLE I, THE RECOGNITION RATES (%) FOR VECTORS WITH THE "NORMAL" LABEL. DATA VALUES ARE NORMALIZED.

	group	K-means	FCM	FE
1	I	40.7	38.4	35.5
2	II III	62.1	62.2	68.1
3	I II III	68.3	61.3	63.2
4	I II	69.4	70.2	68.8
5	II	70	70	70
6	I II III IV	82.1	80	80.4
7	I III	83	86	83.6
8	I III IV	85.4	83.6	85.2
9	III IV	87.1	79.5	72.7
10	IV	88.3	91.1	87.7
11	I IV	88.8	88.7	87.6
12	II IV	88.8	61.7	89.2
13	I II IV	89	88	89.1
14	III	89.1	80.9	78.5
15	II III IV	90.4	81.3	80.3

On the other hand, the recognition rates for the labels B (buffer_overflow), C (ftp_write), D (guess_passwd), F (ipsweep), H (loadmodule), I (multihop), J (neptune), M (perl), N (phf), S (smurf), and W (warezmaster) are very low. Among them, labels B, C, D, H, M, N, and W actually cannot be detected by checking network traffic data alone. They belong to host based intrusion detection. Monitoring data at host level is needed. Due to the lack of extensive host level data, the low recognition rates are understandable. Label I represents complicated multiday activities, and more work is needed to improve the low recognition rate. Label F describes a type of port scanning activities. It is largely misclassified as label C. We don't know the reasons yet. Both label J and label S belong

to denial of services attacks. Label J is misrecognized mainly as R (40.3%) and D (29.3%). Both J and R have large number of SYN packets. It is the reason why 40.3% J vectors are misclassified as R vectors. Label S is basically recognized as U (76.6%), which is another type of denial of services attacks.

For the purpose of our experiments, from this point on, we will concentrate only on the recognition rates of the vectors with the "normal" label. All the other labels are regarded as abnormal (or anomaly). Table I lists the recognition rates, under the different combinations of features, on normalized data values, with the 3 different clustering algorithms, and Table II lists the same results, but on the raw data values, i.e., without normalization.

TABLE II, RECOGNITION RATES (%) FOR VECTORS WITH THE "NORMAL" LABEL. DATA VALUES ARE NOT NORMALIZED.

	group	K-means	FCM	FE
1	I II	23.2	32.4	34.9
2	I II III	23.2	44.2	34.1
3	III	23.4	27.6	30.6
4	I III	25.1	43.4	48.4
5	I	25.12	38.7	40.1
6	I II IV	25.8	42.4	35.9
7	I II III IV	25.8	22.4	38.6
8	I IV	28.6	45.1	49.5
9	I III IV	28.6	35.4	56.5
10	II III	32.2	30.5	36.5
11	II IV	33.5	34.7	33.5
12	IV	40.5	37.5	40.1
13	III IV	52.4	52.9	52.6
14	II III IV	52.5	52.7	52.7
15	II	70	70	70

From Table II, we can see that Group III or IV alone contributes most to the recognition rate (Row 14 and 10), and the other fields actually more or less contribute negatively. Clustering on either Group III or Group IV only (Row 14 and 10, Table II) yields almost the best results in our studies. Either has almost the same result as the other. However, combining both groups together does not significantly increase the recognition rates (Row 9, Table II). More generally, adding any extra information, by adding features from the other groups, to either group does not significantly increase the recognition rates. The differences from Row 10 to Row 15, Table II, are so marginal and can be safely disregarded.

V. CONCLUSION AND FUTURE WORK

In this paper, we studied the impact of feature selection and data normalization on detecting anomaly network traffic. We use KDD CUP 1999 dataset as the sample for the study, and the detection algorithms used are K-means clustering, fuzzy C means clustering, and fuzzy entropy clustering. We have run the clustering experiments with 15 different combinations of the fields (features) from the dataset vectors. Each experiment is conducted with the raw data and also the normalised data. We found out that:

- Time based traffic features, which are temporally amalgamated values of traffic features in the last 2 seconds time window, contribute most to the recognition rates.
- Time based traffic features can be calculated based on either the sources or the destinations of the network connections. Either calculation provides almost the same results. Combining the 2 together does not increase recognition rates.
- The features which are host related (not network related) yield irregular results due to the fact that the values for these fields are exactly the same for about 70% of the vectors.
- Normalization is important.

In the near feature, we will conduct more experiments. We'd like to test the time based features on different sizes of time windows, instead of just 2 seconds as on KDD CUP 1999 dataset. We will also study the impact of different normalization methods and the impact of weighted features.

REFERENCES

[1] Chan, P.K., M.V. Mahoney, and M.H. Arshad, A Machine Learning Approach to Anomaly Detection. 2003, Florida Institute of Technology: Melbourne, FL, USA.

[2] McHugh, J. The 1998 Lincoln Laboratory IDS Evaluation (A Critique). in Recent Advances in Intrusion Detection 2000 (RAID 2000). 2000. Toulouse, France: Springer.

[3] Portnoy, L., E. Eskin, and S. Stolfo. Intrusion detection with unlabeled data using clustering. in Proceedings of ACM CSS Workshop on Data Mining Applied to Security (DMSA-2001). 2001. Philadelphia, PA, USA: ACM Press.

[4] ACM. KDD CUP 1999 data. [cited 12 January 2007]; Available from: http://kdd.ics.uci.edu//databases/kddcup99/kddcup99.html.

[5] Taylor, C. and J. Alves-Foss. An Empirical Analysis of NATE: Network Analysis of Anomalous Traffic Events. in 10th New Security Paradigms Workshop. 2002. Virginia Beach, Virginia, USA: ACM Press.

[6] Taylor, C. and J. Alves-Foss. NATE: Network Analysis of Anomalous Traffic Events, a low-cost approach. in Proceedings of New Security Paradigms Workshop. 2001. Cloudcroft, New Mexico, USA.

[7] DARPA. DARPA Intrusion Detection Evaluation Data Sets. 1999 [cited 2006 15 October 2006]; Available from: http://www.ll.mit.edu/IST/ideval/data/data_index.html.

[8] Li, X. and N. Ye. Mining Normal and Intrusive Activity Patterns for Computer Intrusion Detection. in Intelligence and Security Informatics: Second Symposium on Intelligence and Security Informatics. 2004. Tucson, AZ, USA: Springer-Verlag.

[9] Caruso, C. and D. Malerba. Clustering as an add-on for firewalls. in Fifth International Conference on Data Mining, Text Mining and Their Business Applications (DATA MINING 2004). 2004. Malaga, Spain: WIT Press, Southampton, UK.

[10] Wang, Q. and V. Megalooikonomou. A clustering algorithm for intrusion detection. in SPIE Conference on Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security. 2005. Orlando, Florida, USA.

[11] Stolfo, S.J., W. Fan, et al. Cost-based Modeling and Evaluation for Data Mining With Application to Fraud and Intrusion Detection: Results from the JAM Project. in Proceedings of 2000 DARPA Information Survivability Conference and Exposition. 2000.