An Improved Feature Selection Algorithm Based on MAHALANOBIS Distance for Network Intrusion Detection

Zhao Yongli [1]
[1]State Key Laboratory of Hybrid Process Industry Automation System and Equipment Technology, Automation Research and Design Institute of Metallurgical Industry
Beijing, China
zylhappy47@163.com

Zhang Yungui [1, 2]
[1]State Key Laboratory of Hybrid Process Industry Automation System and Equipment Technology, Automation Research and Design Institute of Metallurgical Industry
Beijing, China
[2]School of Electrical Engineering and Automation, Harbin Institute of Technology,
Harbin, China

Tong Weiming [2]
[2]School of Electrical Engineering and Automation, Harbin Institute of Technology,
Harbin, China

Chen Hongzhi [1]
[1]State Key Laboratory of Hybrid Process Industry Automation System and Equipment Technology, Automation Research and Design Institute of Metallurgical Industry
Beijing, China

***Abstract* — Network Intrusion Detection System (NIDS) plays an important role in providing network security. Efficient NIDS can be developed by defining a proper rule set for classifying network audit data into normal or attack patterns. Generally, each dataset is characterized by a large set of features, but not all features will be relevant or fully contribute identifying an attack. Since different attacks need different subsets to have better detection accuracy, this paper describes an improved feature selection algorithm to identify most appropriate subset of features for a certain attack. The proposed method is based on MAHALANOBIS Distance feature ranking and an improved exhaustive search to choose a better combination of features. We evaluate the approach on the KDD CUP 1999 datasets using SVM classifier and KNN classifier. The results show that classification is done with high classification rate and low misclassification rate with the reduced feature subsets.**

Keywords—Feature Selection, intrusion detection, MAHALANOBIS Distance, classification, SVM, KNN

I. INTRODUCTION

With the increased growth of networked systems and applications, the demand for network security is high. Besides firewalls, anti-virus software, VPNs, Intrusion Detection Systems as a second strobe of security find a wide application network security fields to detect attacks. Network Intrusion Detection Systems (NIDSs) are the war-horses of network security. Two different approaches are by far dominant in the research literature and commercial IDS security devices: signatures-based and anomaly detection [1, 2].

Signatures-based detection systems are highly effective to detect those attacks which are programmed to alert on. However, they cannot defend the network against new attacks. Indeed, signatures-based systems have two challenges: the diagnosis of a new attack and the construction of the new signature. On the other hand, anomaly detection uses instances of normal-operation traffic to build normal-operation pro-files, detecting anomalies as activities that deviate from this baseline. Such methods can detect new kinds of net-work attacks. Nevertheless, they often induce high false-alarm rates [2, 3]. To tackle these problems, there have been many proposed approaches, such as machine learning based approaches, data mining based approaches, and statistical data based approaches, to generate attack signatures automatically and improve classification performance. Feature selection is one of most critical processes.

Feature selection is a process that selects a subset of features from input data, such as network traffic, to reduce overheads of data processing and to improve the accuracy of attack detection; moreover, dimensionality reduction also decreases the computational load of models. These reasons lead to the development of a huge number of feature selection algorithms in the past few years. The large majority of them assume the datasets are either continuous or symbolic. However, in many contexts in IDS, data come in a mixed way [4, 5]. In this paper, we proposed an improved algorithm based on MAHALANOBIS Distance, which can resolve the mixed problem.

This algorithm first use feature ranking based on MAHALANOBIS Distance as the principle selection mechanism. Feature ranking is a filter approach and possesses several advantages over other feature selection methods: 1) it is computationally and statistically scalable to large datasets; 2) it is simple to use; 3) it shows good success for a variety of real-world applications [6]. Then it uses an improved exhaustive

search to choose optimal ranked features. The selection criterion is built as a quadratic expression with a minimum value based on the optimal classification rate and misclassification rate pair, which can realize a global optimal search.

The rest of paper is organized as follow. Section II describes the proposed algorithm. In Section III, experimental evaluation of the algorithm using SVM classifier and KNN classifier is provided. Conclusions are given in Section IV.

II. Improved Feature Selection Algorithm Based on MAHALANOBIS Distance for Network Intrusion Detection

This section is devoted to the description of the proposed feature selection algorithm. The two essential elements of the algorithm, namely, the feature ranking algorithm and the improved exhaustive search algorithm (including the global optimal selection criterion) are given in the following 2 subsections.

A. Feature Ranking Algorithm

Generally speaking, feature ranking can be performed in two different ways, namely, ranking by evaluating individual features (single-feature ranking criteria) and ranking by evaluating subsets of features (subset-based ranking criteria). Also, feature selection is grouped according to the attribute evaluation measure: depending on the type (filter or wrapper techniques). The filter model relies on general characteristics of the data to evaluate and select feature subsets without involving any mining algorithm. The wrapper model requires one predetermined mining algorithm and uses its performance as the evaluation criterion. The wrapper model often is more computationally expensive and more complex than filter model [6, 7]. In this paper, we perform feature ranking based on MAHALANOBIS Distance using single-feature ranking criteria and the filter model.

TABLE I. Single-based Feature Ranking Algorithm

Input: - $\overline{k} = [k_1, k_2, ..., k_n]$ - the original feature set

Output: - Ranking criterion function, c and ranked features

1) Initialize: $\overline{f} = []$

2) **For** each feature $k_i \in \overline{k}$

a) obtain the scalar MAHALANOBIS distance $dM_{1,k}$ (the value between the mean of a particular group A and the whole set of group B) and $dM_{2,k}$ (the value between the mean of a particular group B and the whole set of group A)

b) calculate merit score f_k using the ranked criterion: $f_k = dM_{1,k} + dM_{2,k}$

c) store f_k into $\overline{f}$

3) End of **for** loop

4) Rank feature $\overline{k}$ based on $\overline{f} = [f_1, f_2, ..., f_n]$

To use the single-feature ranking criterion for ranking individual features, one can select one feature at a time and calculate the merit score using the ranking criterion for the single feature. As a result, one would expect that for a good feature, such calculated merit index would be higher than that for a bad feature. The single-based feature ranking algorithm is shown in TABLE I [8].

This algorithm is realized on either labeled data which is classified into different classes or unlabeled data which is classified by a classifier. The advantage of MAHALANOBIS distance criterion is to eliminate the correlation between variables and units, which is important for single-based feature ranking [6, 9, and 14].

B. Improved Exhaustive Search Algorithm

Algorithms begin with the phase where attributes are individually evaluated and provide a ranking according to a filter criterion. In the next step, a feature subset evaluator is applied to a fixed number of attributes from the previous ranking following a search strategy [7]. In this paper, we proposed an improved exhaustive search algorithm.

The time complexity of different search strategies, namely exhaustive, heuristic and random search, is exponential in term of data dimensionality for exhaustive search and quadratic for heuristic search. Experiments show that in order to find best feature subset, the number of iterations required is usually at least quadratic to the number of features [7]. In order to obtain optimal feature subsets and reduce the number of iterations, this paper proposes an improved exhaustive search algorithm, which is shown in TABLE II.

TABLE II. An Improved Exhaustive Search Algorithm

Input: - $\overline{k} = [k_1, k_2, ..., k_n]$ - the ranked feature set

Output: - Evaluation criterion function, $f(c, m)$ and optimal feature subsets R.

1) Initialize: $R = []$; $CR = []$; $MR = []$.

2) Choose a reduced feature subsets which are greater than a threshold value: $F = [k_1, k_2, ..., k_m]$, m<=n.

3) Do a discriminate analysis with F using MAHALANOBIS distance in stepwise statistics. To obtain the output by the labeled data (group A and group B): C = classification rate and M = misclassification rate.

4) **For** each feature $k_i \in F$ **and** $k_i = \max(F)$

a) Select highest ranked feature F_k and obtain: $R = \{R \cup F_k\}$; $F = \{F - F_k\}$.

b) Do a discriminate analysis with R using MAHALANOBIS distance in stepwise statistics. To obtain the output by the labeled data (group A and group B):

$c_k = current\ classification\ rate$

$m_k = current\ misclassification\ rate$.

c) Store c_k and m_k into CR and MR separately.

5) End of **for** loop

6) To obtain the optimal feature subsets R based on evaluation criterion function $\min(f(c, m))$:

$f(c, m) = sqrt((1-c) \cdot 100 \cdot m \cdot 100)$

In [13], a simple greedy algorithm is used to select the feature subset, which can reduce the iterations to some extent; but the feature subset is not the optimal. We built the evaluation criterion function $f(c, m) = sqrt((1-c) \cdot 100 \cdot m \cdot 100)$, which is a quadratic function with the minimum value and we can get the optimal feature subset based on the evaluation

criterion function. The value range of the evaluation criterion function is between 0 and 100.

The final output of this method provides important features for identifying every attack.

III. Experimental Evaluation

To evaluate our proposed feature selection method, we carried out the algorithm by MATLAB software tool based on KDD CUP 1999 datasets and calculated the classification rate and misclassification rate. In the experiments, we used Polynomial Kernel function of the Support Vector Machine (SVM) [4].

Although there are some criticisms [9] towards KDD CUP 1999 Data sets, we used the data in our experiments based on two reasons. First, the data have been widely used for evaluating various intrusion detection methods. Second, the data provides numerous types of anomalies [10].

The raw data contain traffic in a simulated military network that consists of hundreds of hosts. We use a subset in the experiments. In the data sets, each network connection is labeled as either normal, or as an exactly one specific kind of attack. The network connection data contain 41 features. These features were divided into three groups: basic features of individual TCP connections, traffic features and content features within a connection suggested by the domain knowledge. Among these 41 features, 34 are continuous; 4 are discrete values and 3 are text features [9, 10].

In the International Knowledge Discovery and Data Mining Tools Competition, only "10% KDD" dataset is employed for the purpose of training [13]. It is a concise from of "Whole KDD". This dataset has only 22 attack types and they are mostly of denial of service category. Whereas "Corrected KDD" dataset provides a data with different statistical distributions of attacks compared to "10% KDD" or "Whole KDD". It contains 37 types of attacks. We selected "10% KDD" as the training data set and "Corrected KDD" as the test data set. Finally， we use SVM classifier to validate the algorithm. Since SVM classification uses only numerical data for testing and training, so text features are needed to be converted into numerical values [12]. To simplify the calculation, only 38 numeric features were used in the experiments.

Our experiments have two phases namely selecting optimal feature subsets for every attack and then classifying the testing data. In the first phase, important attributes from training data of "Corrected KDD" are ranked by single-based feature ranking values and then an improved exhaustive search algorithm is used based on the evaluation criterion function to select the optimal feature subset. In the second phase, the training data of "10% KDD" to train SVM classifier and the testing data of "Corrected KDD" using SVM classifier to classify the data with 38 features and selected optimal feature subset separately and find the classification rate and misclassification rate. We run our experiments on a system with 1.88 GHz RAM, Intel(R) Core (TM) i3-2310M CPU @ 2.10 GHz and 2.09 GHz running Windows XP. All the processing is done using MATLAB R2010a. TABLE III gives brief descriptions of the six datasets for feature selection and TABLE IV gives the optimal feature subsets for each attack after feature selection algorithm.

TABLE III. The Six Datasets for Feature Selection

Type		# of data
DOS	BACK	1098
	NEPTUNE	2000
PROBE	IPSWEEP	306
R2L	GUESS_PASSWD	2000
U2R	HTTPTUNEL	158
NORMAL	-	2000

TABLE IV. Optimal Feature Subset for Each Attack After Feature Selection Algorithm

Attack Names	Selected features
BACK	5, 6,32,33, 1,24
NEPTUNE	6,5,33,32,23,1,24,12,34,40,41,28, 27,29,31,36,10,35,30,26
IPSWEEP	6,5,32,33,1,23,24,12,36,37
GUESS_PASSWD	6,5,32,33,24,1,23,11,12,31,34,36, 32,10,40,27,28,41,37,29,35,30,19 ,39,26,17,14,38
HTTPTUNEL	6,5,33,32,1,24,23,12,34,28,27,41, 31,30,13

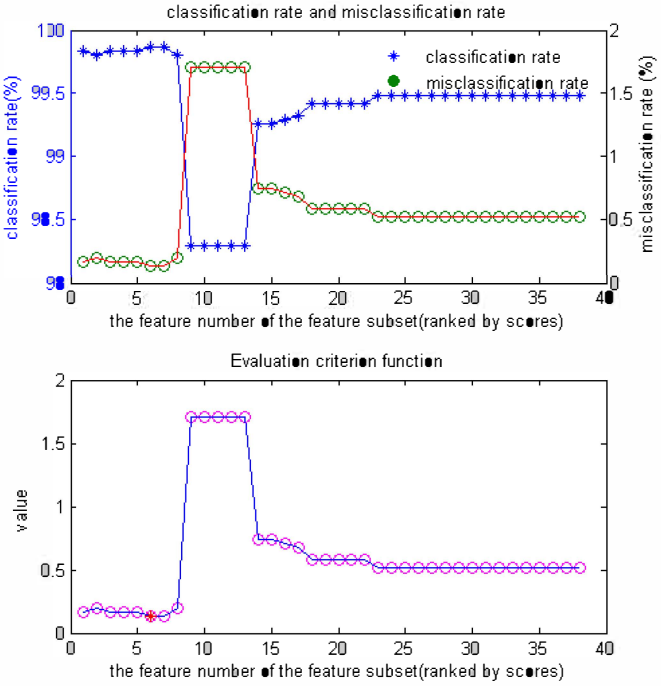

Fig. 1 The classification rate, misclassification rate and evaluation criterion function curves of the BACK attack

Fig. 1 and Fig. 2 describe the feature distributions of the two typical examples of attacks. Classification rate and misclassification rate are defined as the following:

$$\text{CR (Classification Rate)} = \frac{\text{(the number of the sample data classified correctedly)}}{\text{(the total number of the sample data)}}$$

$$\text{MR (Misclassification Rate)} = \frac{\text{(the number of the sample data classified uncorrectedly)}}{\text{(the total number of the sample data)}}$$

In Fig. 1, we choose NORMAL data set and BACK data set to select optimal feature subset using our proposed algorithm based on MAHALANOBIS distance discriminate criterion. In Fig. 2, we choose NORMAL data set and NEPTUNE data set to select optimal feature subset using our proposed algorithm based on MAHALANOBIS distance discriminate criterion.

Seen form them, the evaluation criterion function gives the optimal solution with maximum classification rate and minimum misclassification rate.

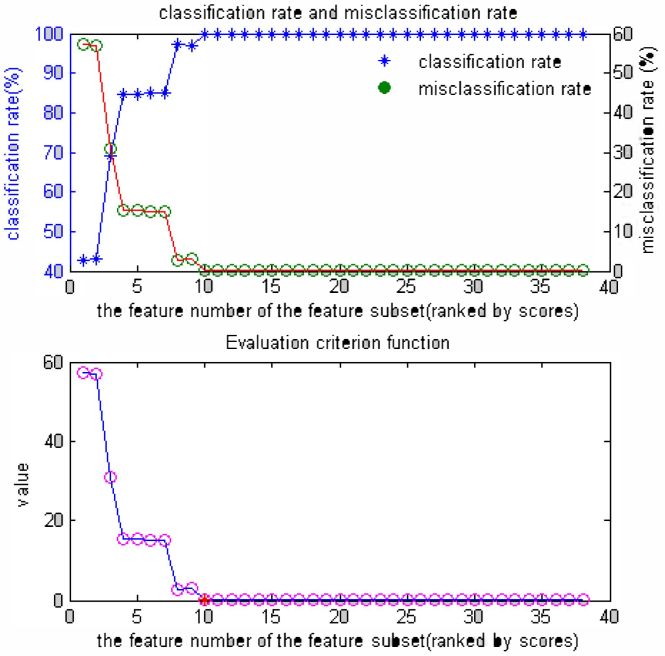

Fig. 2 The classification rate, misclassification rate and evaluation criterion function curves of the IPSWEEP attack

We select two datasets of DOS (BACK and NEPTUNE) attack using SVM classifier and KNN classifier to prove the effectiveness of our algorithm.

Fig. 3 shows the attack classification rate for Polynomial Kernel of SVM classifier. TABLE V displays the classification rate and running time for KNN classifier. In the experiments, we used two DOS attacks in KDD datasets: the BACK attack and the NEPTUNE attack. As you can see, the features selected with our algorithm show higher classification rate and lower misclassification rate than that of all 38 features in Fig. 3. Although the classification rates basically stay the same between that of the optimal feature subsets and all 38 features, the running time reduces greatly. In the experiments with polynomial kernel of SVM classifier and KNN classifier, we used various training data, but it shows that approximately same classification rate with the optimal feature subsets. Even though the size of the training data is small, the proposed algorithm shows approximately same classification rate with a trained machine with large training data, which indicates that it can reduces the storage and computational resources to some extent using small training data with our selected feature subsets.

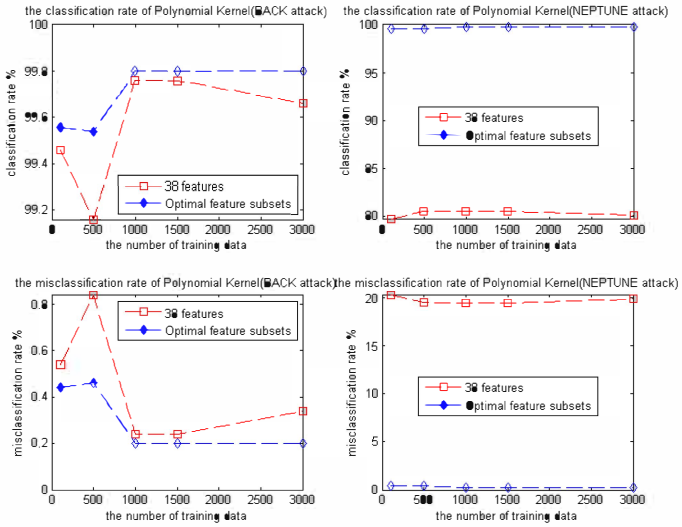

Fig. 3 The classification rate and misclassification rate of SVM (Polynomial Kernel)

TABLE V. THE CLASSIFICATION RATE AND RUNNING TIME OF KNN CLASSIFIER

Num	BACK				NEPTUNE			
	Classification rate (%)		*Running Time(S)*		*Classification rate (%)*		*Running Time(S)*	
	38	*opt*	*38*	*opt*	*38*	*opt*	*38*	*opt*
100	99.7	99.8	3.07	2.75	79.9	79.9	0.37	0.23
500	99.7	99.8	3.60	2.98	79.9	79.9	0.77	0.58
1000	99.7	99.8	5.59	3.36	79.9	79.9	2.50	0.82
1500	99.7	99.8	6.68	3.76	79.9	79.9	3.22	2.25
3000	99.7	99.8	12.0	5.00	79.9	79.9	7.72	3.49

TABLE V shows that the classification rate of NEPTUNE attack is lower than that of BACK attack. Because the KNN classifier use the Euclidean distance criterion to classify the data and the data attribute types and unit dimensions affect the classification results greatly, which reduced the classification rate to some extent. Also the selected feature subset of NEPTUNE attack includes more symbol attributes than that of BACK attack. We could improve the classification rate of KNN classifier by Attribute Normalization [10].

IV. CONCLUSIONS

In this paper, we proposed an improved feature selection algorithm for network intrusion detection that performs data reduction by selecting important subset of attributes. The performance of our proposed approach on the KDD datasets achieved Stability and robustness for DOS attack class. It also reduced the cost of resources and the misclassification rate. The experimental results to manifest that significant attribute selection can improve the performance of network intrusion

detection. The attacks of KDD dataset detected with more than 99% classification rate using our proposed approach.

Our proposed approach focuses on two-class classification problem based on labeled data. The future work refers to multi-class classification problem, large amount of unlabeled data for training and incremental learning problems.

ACKNOWLEDGEMENTS

The research was supported by the reconstruction project of a National Engineering Technology Research Center (2011FU125Z21): Industry Control Network Security Defending Device. This work is also supported by State Key Laboratory of Hybrid Process Industry Automation System and Equipment Technology, Automation Research and Design Institute of Metallurgical Industry under grant. We thank KDD CUP 1999 for providing this data set. Thanks are also due to the anonymous reviewers for their insightful comments and suggestions.

REFERENCES

[1] Zhou Ying, Sun Mingsong, "The Network Intrusion Detection System Model Based on Clustering". Journal Harbin Univ. SCI. & Tech. vol. 12, Issue. 1, pp. 39-42, 2007.

[2] S.A.Joshi, Varsha S.Pimprale, "Network Intrusion Detection System (NIDS) based on Data Mining", International Journal of Engineering Science and Innovative Technology (IJESIT), Vol. 2, Issue 1, pp. 95-98, January 2013.

[3] Pedro Casas, Johan Mazel, Philippe Owezarski, "Unsupervised Network Intrusion Detection Systems: Detecting the Unknown without Knowledge", Computer Communications, Vol. 35, Issue 7, pp. 772-783, 2012.

[4] Jungtaek Seo, Jungtae Kim, Jongsunb Moon, Boo Jung Kang, Eul Gyu Im. Clustering-based Feature Selection for Internet Attack Defense. International journal of Future Generation Communication and Networking vol. 1, no. 1: 91-97, 2008. Science & Engineering Research Support Center, Republic of Korea.

[5] Gauthier Doquire, Michel Verleysen, "Mutual information based feature selection for mixed data", ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, April 2011.

[6] Weizhong Yan, "Fusion in Multi-Criterion Feature Ranking, Information Fusion", 10th International Conference on Digital Object Identifier, pp. 1-6, 2007.

[7] Roberto Ruiz, Jos´e C. Riquelme, Jes´us S. Aguilar-Ruiz, "Heuristic Search over a Ranking for Feature Selection", Computational Intelligence and Bioinspired Systems Lecture Notes in Computer Science, vol. 3512, pp. 742-749, 2005.

[8] Loris Nanni, "Cluster-based pattern discrimination: A novel technique for feature selection", Pattern Recognition Letters, vol. 27, pp. 682-687, 2006.

[9] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, Ali A. Ghorbani. "A Detailed Analysis of the KDD CUP 99 Data Set", CISDA'09 Proceedings of the Second IEEE international conference on Computational intelligence for security and defense applications, pp. 53-58, 2009.

[10] Wei Wang, Xiangliang Zhang, Sylvain Gombault, Svein J. Knapskog, "Attribute Normalization in Network Intrusion Detection," ispan, pp.448-453, 2009. 10th International Symposium on Pervasive Systems, Algorithms, and Networks.

[11] S. Wu, P. A. Flach, "Feature selection with labelled and unlabelled data", ECML/PKDD'02 workshop on Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning. M. Bohanec, B. Kasek, N. Lavrac, D. Mladenic, (eds.), pp. 156–167. August 2002.

[12] Shailendra Kumar Shrivastava, Preeti Jain, "Effectiove anomaly based intrusion detection using rough set theory and support vector machine", International Journal of Computer Applications, vol. 18, Issue. 3, pp. 35-41, March 2011.

[13] Dr.S.Siva Sathya, Dr. R.Geetha Ramani, K.Sivaselvi, "Discriminant Analysis based Feature Selection in KDD Intrusion Dataset", International Journal of Computer Applications, vol. 31, Issue. 11, pp.1-7, October 2011. Published by Foundation of Computer Science, New York, USA.

[14] Rupali Datti, Bhupendra verma, "Feature Reduction for Intrusion Detection Using Linear Discriminant Analysis", (IJCSE) International Journal on Computer Science and Engineering. Vol. 2, Issue 4, pp. 1072-1078, 2010.