Network Anomaly Detection using Hybrid Feature Selection

Eunhye Kim*, Sehun Kim*
* Department of Industrial Engineering,
Korea Advanced Institute of Science and Technology
ehkim@tmlab.kaist.ac.kr, shkim@kaist.ac.kr

Abstract

In this paper, we propose a hybrid feature extraction method in which Principal Components Analysis is combined with optimized k-Means clustering technique. Our approach hierarchically reduces the redundancy of features with high explanation in principal components analysis for choosing a good subset of features critical to improve the performance of classifiers. Based on this result, we evaluate the performance of intrusion detection by using Support Vector Machine and a nonparametric approach based on k-Nearest Neighbor over data sets with reduced features. The Experiment results with KDD Cup 1999 dataset show several advantages in terms of computational complexity and our method achieves significant detection rate which shows possibility of detecting successfully attacks.

I. INTRODUCTION

Intrusion Detection Systems (IDS) have become important and widely used tools for ensuring network security. However, as defense mechanisms to identify and prevent intrusions been purposed and deployed for IDS, also attacks have been more sophisticated and intelligent. For more reliable detection of attacks, there are two main challenges to be improved. First, the amount of data that an IDS needs to examine is very large even for a small network by using extraneous features, although some of the data features may be redundant or contribute little to the detection process. IDS must therefore reduce the amount of data to be processed for real-time detection. Second, new intrusion types are unaware and difficult to detect.

Most current network intrusion detection systems employ signature based methods and learning algorithms which rely on labeled data to train. These methods generally have difficulty in detecting new types of attack and training data is typically expensive.

Anomaly detection, on the other hand, builds models of normal data and detects any deviation from the normal model in the observed data. Given a set of normal data to train, the goal is to determine whether the test data belong to normal or to an anomalous behavior. The anomaly detection algorithms have the advantage that they can detect new types of intrusions as deviations from normal usage [1, 2].

In this paper, the purpose of our anomaly detection scheme is 1) to identify important input features in building a IDS that is

"This research was supported by the MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Assessment)" (IITA-2005-(C1090-0502-0020))

computationally efficient and effective and 2) to evaluate the performance of statistical intrusion detection techniques in order to learn normal and anomalous patterns from training data and generate classifiers used to detect attacks.

Ⅱ. RELATED WORK

Novelty detection is one of the fundamental requirements of a good classification or identification system since sometimes the test data contains information about objects that were not known at the time of training the model. In a probabilistic sense, novelty detection is equivalent to deciding whether an unknown test pattern is produced by the underlying distribution that corresponds to the training set of normal patterns [3].

The major algorithm is that they typically use only normal patterns as training examples to build a generative model of normal behavior. The novelty detection approach is particularly attractive under situations where novel or anomalous patterns are expensive or difficult to obtain for model construction.

Ⅲ. HYBRID FEATURE SELECTION

For computationally efficient and effective IDS, it is essential to identify important input features. We propose a hybrid feature selection algorithm by combining Factor Analysis based on Principal Components Analysis and k-Means clustering. Factor analysis is a statistical technique used to identify a relatively small number of factors that can be used to represent relationships among sets of many interrelated variables. The general expression for the estimate of the i^{th} factor, F_i, is

$$F_i = \sum_{j=1}^{n} C_{ij} X_j$$

where C_{ij} is factor score coefficients and n is the number of variables.

In our hybrid feature selection, first, we apply factor analysis based on Principal Components Analysis for factor extraction and then k-Means clustering algorithm on each training data set. Based on k clusters of features through k-Means clustering, we extract features of which factor score coefficients is higher than a threshold. And then we reduce redundancy of high score features if they are in same cluster. The algorithm in detail proceeds as follows.

1. Factor Analysis to the whole set of p features.
   a) Input : training data set
   b) Factor extraction algorithm : Principal Components Analysis
   c) Output : C_{ij} (factor score coefficients) for features
2. K-Means clustering to the whole set of p features.
   a) Input : training data set
   b) Output : S_k (k feature clustering set)
3. Extract features of which factor score coefficients are higher than a threshold
4. Compare k - clustered features of training set with extracted features of high score coefficients
5. Reduce redundancy of features in the same cluster by choosing the feature with the maximum factor score.

Ⅳ. STATISTICAL INTRUSION DETECTION TECHNIQUES

Based the reduced feature set, we evaluate the performance of intrusion detection by using Support Vector Machine and a nonparametric approach based on k-Nearest Neighbor classifiers over data sets with reduced features. Our experiments use novelty detection concept.

1. Support vector machines

Support vector machines, or SVMs, are learning machines that plot the training vectors in a feature space. The main idea of the SVM is to derive a unique separating hyperplane that maximizes the separating margin between the two classes. The feature vectors that lie on the boundary defining this separating margin are called support vectors. Classifiers that exploit this property are called support vector machines [,5, 6].

In our experiment, one-class SVM is employed to define the support region of normal network traffic. A new data will be

signified as abnormal if it lies outside of this support region. In order to separate the dataset, we need to solve the following optimization problem:

$$\min_{\alpha} \quad \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j K(x_i,x_j)$$

Subject to

$$0 \le \alpha_i \le \frac{1}{vl}$$

$$\sum \alpha_i = 1$$

where α_i is a Lagrange multiplier, v is a parameter that controls the trade off between maximizing the distance of the hyperplane from the origin and the number of data points in the region created by the hyperplane, l is the number of points in the training dataset, $K(x_i,x_j)$ is the kernel function to project input vectors to a feature space and the decision function is

$$f(x) = sign(\sum_{i=1}^{l}\alpha_i K(x,x_i) - \rho)$$

Three common kernels are as follows:
Polynomial kernels:

$$K(x,y) = (x\cdot y+1)^d$$

Radial Basis Function (RBF) kernels:

$$K(x,y) = \exp(-\|x-y\|^2/2\sigma^2)$$

Sigmoid kernels (with gain k and offset θ)

$$K(x,y) = \tanh(k(x\cdot y)+\theta)$$

The objective of our SVM experiments is to separate normal and attack patterns. In our experiments, the freeware LIBSVM is used and we used the Radial Basis Function kernel [9].

2. k-Nearest Neighbor classifier

Nearest Neighbor is a predictive technique suitable for classification models. When a new case or instance is presented to the model, the algorithm looks at all the data to find a subset of cases that are most similar to it and uses them to predict the outcome. There are two principal drivers in the k-NN algorithm: one is the number of nearest cases to be used (k) and the other is a metric to measure similarity of data samples.

To classify a class-unknown data vector *x*, the k-Nearest Neighbor classifier algorithm ranks the test data's neighbors among the training sample vectors, and then the test sample is classified to the class represented by a majority of the kNN's [11]. The classes of these neighbors are weighted using the similarity of each neighbor to x, where similarity is measured by Euclidean distance between two data vectors as follows:

$$dist(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

where x and y are two records and n is the number of variables measured.

The k-Nearest Neighbor classifier is based on the assumption that the classification of an instance is most similar to the classification of other instances that are nearby in the vector space.

V. EXPERIMENTS

The data we used for testing is the KDD Cup 1999 data set [7]. It originated from the 1998 DARPA Intrusion Detection Evaluation Program managed by MIT Lincoln Labs. For our experiment we partitioned the KDD dataset into subsets and selected three subsets among them, which contain a good mix of various intrusions types. Another three subsets are consisted of only normal connections as training data. The data sets are then normalized. The training and test sets comprise of approximately 5100 and 6700 records respectively. We labeled the 41 features in order as A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, Z, AA, AB, AC, AD, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO and the class label named as AP.

To evaluate our system we used two indicators of performances : true negative rate and true positive rate. The true negative rate is defined as the number of normal instances classified by the system divided by the total number of normal instances present in the test set. The true positive rate is defined as the number of intrusion instances detected by the system divided by the total number of intrusion instances present in the test set.

Table 1 depicts the selected features of three sets containing normal and attack data by our feature selection algorithm. These sets were used for test data passing through the trained model by normal data to detect intrusions. Table 2 shows the selected features of three sets only containing normal data for training.

[Table 1] Reduced features of test data

Data set	Selected features
Test Set 1	A, B, E, F, Q, V, X, Y, AA, AC, AD, AI
Test Set 2	A, B, C, E, P, Q, V, Y, AB, AC, AD, AF
Test Set 3	A, B, E, R, I, V, X, Y, AB, AD, AK

[Table 2] Reduced features of training data

Data set	Selected features
Set 1	A, B, E, L, Q, V, W, Y, AA, AC, AD, AM
Set 2	A, B, C, E, L, R, V, W, Y, AA, AC, AD
Set 3	B, D, E, L, R, V, W, Y, AC, AD, AF, AO

Through our hybrid feature selection algorithm, the 41 features are reduced to the 11 or 12 most significant features. This gives an advantage in terms of training time.

[Table 3] Detection results of Statistical Techniques

Training / Test set		SVM	k-NN
Set 1	Set 1	98.10	99.26
	Set 2	98.07	97.52
	Set 3	97.85	98.41
Set 2	Set 1	98.50	97.68
	Set 2	99.02	98.51
	Set 3	97.46	96.29
Set 3	Set 1	97.68	97.13
	Set 2	97.02	97.44
	Set 3	98.01	98.67

Table 3 shows true positive rates, the performances of intrusion detection model based on SVM and k-Nearest Neighbor classifier by using the results of features selection. The performances of the model were evaluated over each three test set by using cross validation testing concept.

The results showed that performances depend on which training set and feature set were used and revealed that there were some types of attacks that the classifier was not able to detect because of some attacks using the same feature of normal data. This affects lower detection rate. However, it can be said that our method achieves significant detection rates which shows possibility of detecting successfully attacks and our test reveals the importance of feature selection which meets well the requirement of explaining various normal and intrusion types.

VI. CONCLUSION

In this paper, we have briefly introduced a hybrid feature selection method by combining Factor Analysis based on Principal Components Analysis and k-Means clustering and described two statistical anomaly detection approaches based on Support Vector Machine and k-Nearest Neighbor classifier. In the experiments, we performed data reduction by using our proposed method for feature selection and evaluated their performance on KDD data.

The 11 or 12 most significant features give an advantage in terms of training time and our method achieves significant detection rate which shows possibility of detecting successfully attacks.

The two statistical techniques achieved similar detection rate, but the main advantage of SVM, in comparison with k-Nearest Neighbor classifier, is the low computational complexity to classify new elements. The major limitation of nonparametric approach is its relatively high computational cost of testing, although it requires essentially no training time and can be used with arbitrary distributions.

Our future research will involve modifications of our method to achieve better performance, automation and low computational complexity to classify test data.

REFERENCES

[1]. Lee W., Stolfo S. and Mok K., "A Data Mining Framework for Building Intrusion Detection Models," In Proceedings of *the IEEE Symposium on Security and Privacy*, 1999.

[2]. Shyu, M.L., Chen, S.C., Sarinnapakorn, K., Chang, L.W., "A Novel Anomaly Detection Scheme Based on Principal Component Classifier," In Proceedings of *ICDM Foundation and New Direction of Data Mining workshop*, pp. 172-179, 2003.

[3]. Markou, M., and Singh, S., "Novelty Detection: A Review Part1: Statistical Approaches"

[4]. Liu, Y., Cukic, B., Fuller, E., Gururajan, S., Yerramalla, S., "Novelty Detection for a Neural Network-Based Online Adaptive System," in Proc. of *29th Annual International Computer Software and Applications Conference (COMPSAC'05)* Volume 2, pp. 117-122

[5]. Hu, W., Liao, Y., and Vemuri, V.R., "Robust Support Vector Machines for Anomaly Detection in Computer Security," In Proceedings *International Conference on Machine Learning and Applications*, pp. 168-174, 2003.

[6]. Manevitz, L.M., Yousef, M., " One-Class SVMs for Document Classification," *Journal of Machine Learning Research 2*, pp. 139-154, 2001

[7]. KDD cup 99 Intrusion detection data set, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[8]. MIT Lincoln Laboratory. http://www.ll.mit.edu/IST/ideval

[9]. Chang, C.C., Lin, C.J., LIBSVM: a Library for Support Vector Machines http://www.csie.ntu.edu.tw/~cjlin/libsvm

[10]. Fukunaga, K., Introduction to Statistical Pattern Recognition, 2nd edition, Academic Press, New York, 1990

[11]. Duda, R.O., Hart, P.E., Stork, D.G., Pattern Classification, 2nd edition, Wiley, 2001