Intrusion Awareness Based on Data Fusion and SVM Classification

Ramnaresh Sharma[1], Manish Shrivastava[2]
PG Scholar LNCT, Bhopal (M.P)[1]
Head of Information Technology Bhopal (M.P.) Department, LNCT, Bhopal (M.P.)[2]

Abstract

Network intrusion awareness is important factor for risk analysis of network security. In the current decade various method and framework are available for intrusion detection and security awareness. Some method based on knowledge discovery process and some framework based on neural network. These entire model take rule based decision for the generation of security alerts. In this paper we proposed a novel method for intrusion awareness using data fusion and SVM classification. Data fusion work on the biases of features gathering of event. Support vector machine is super classifier of data. Here we used SVM for the detection of closed item of ruled based technique. Our proposed method simulate on KDD1999 DARPA data set and get better empirical evaluation result in comparison of rule based technique and neural network model.

Keywords

Intrusion awareness, data fusion, SVM and KDDCUP1999.

I. Introduction

Cyber and network security is a complex task. The current networked computing environment is based on insecure operating systems, complex software from multiple vendors, quick and dirty system deployment and ubiquitous high-bandwidth network connections [4]. It is threatened by blended attacks, monetization of hacking and exploitation, and the easy access to exploitation tools. Needless to say, an organization's network security mission is difficult to staff, develop and maintain. Tools must be developed from both the data/algorithm (bottom up) and the human analyst (top down) perspectives to achieve better situational awareness. To gain better situational awareness, the individual alerts from multiple IDSs must be aggregated, fused or correlated in some fashion. Methods of correlation can bring better situational awareness if the method provides some additional meaning to the data [6]. For instance, alerts that identify a common source might provide information about the identity of the intruder, or at a minimum, provide the analyst with information crucial to the defense of the network, such as blocking all traffic from the source address [8]. Alerts that have a common destination may provide information about the vulnerabilities on a given host in the defended network. Alerts with a common attack signature may provide information about the types of attacks in use and suggest methods for defense, or most certainly suggest where the analyst should pay attention. The process of alert correlation is non-trivial. It is not as simple to aggregate data based on these simple factors in a large scale network and even more difficult to do so in real-time. Data fusion techniques combine data from different sources together. The main objective of employing fusion is to produce a fused result that provides the most detailed and reliable Information possible. Fusing multiple information sources together also produces a more efficient representation of the data [10]. In this paper we proposed fused Support Vector Machine (fSVM) algorithm for detection and classification of security attack dataset. As we know that the performance of support vector machine is greatly depend on the kernel function used by SVM. Therefore, we modified the Gaussian kernel function in data dependent way in order to improve the efficiency of the classifiers. The relative results of the both the classifiers are also obtained to ascertain the theoretical aspects. The analysis is also taken up to show that FSVM performs better than rule based [12]. The classification accuracy of FSVM remarkably improve (accuracy for Normal class as well as DOS class is almost 100%) and comparable to false alarm rate and training, testing times. The remainder of the paper is organized as follows. In Section II, we present KDDCUP'99 dataset. The Preliminary work of security attack detection and classification is formulated in Section III. In section IV FSVM is proposed. In section V Experimental and result analysis. In section V conclusion and future work.

II. Kddcup99 Dataset

To check performance of the proposed algorithm for distributed cyber-attack detection and classification, we can evaluate it practically using KDD'99 intrusion detection datasets [6]. In KDD99 dataset these four attack classes (DoS, U2R, R2L, and probe) are divided into 22 different attack classes that tabulated in Table I. The 1999 KDD datasets are divided into two parts: the training dataset and the testing dataset[14]. The testing dataset contains not only known attacks from the training data but also unknown attacks. Since 1999, KDD'99 has been the most wildly used data set for the evaluation of anomaly detection methods. This data set is prepared by Stolfo et al. [11] and is built based on the data captured in DARPA'98 IDS evaluation program [12]. DARPA'98 is about 4 gigabytes of compressed raw (binary) tcpdump data of 7 weeks of network traffic, which can be processed into about 5 million connection records, each with about 100 bytes. For each TCP/IP connection, 41 various quantitative (continuous data type) and qualitative (discrete data type) features were extracted among the 41 features, 34 features (numeric) and 7 features (symbolic). To analysis the different results, there are standard metrics that have been developed for evaluating network intrusion detections. Detection Rate (DR) and false alarm rate are the two most famous metrics that have already been used [16].

Table1: different Types of Attacks in KDD99 Dataset

4 Main Attack Classes	22 Attack Classes
Denial of Service (DoS)	back, land, neptune, pod, smurt, teardrop
Remote to User (R2L)	ftp_write, guess_passwd, imap, multihop, phf,spy, warezclient, warezmaster
User to Root (U2R)	buffer_overflow, perl, loadmodule, rootkit
Probing(Information Gathering)	ipsweep, nmap, portsweep, satan

DR is computed as the ratio between the number of correctly detected attacks and the total number of attacks, while false alarm (false positive) rate is computed as the ratio between the number of normal connections that is incorrectly misclassified as attacks and the total number of normal connections. In the KDD Cup 99, the criteria used for evaluation of the participant entries is the Cost Per Test (CPT) computed using the confusion matrix and a given cost matrix .A Confusion Matrix (CM) is a square matrix in which each column corresponds to the predicted class, while rows correspond to the actual classes. An entry at row i and column j, CM (i, j), represents the number of misclassified instances that originally belong to class i, although incorrectly identified as a member of class j. The entries of the primary diagonal, CM (i, i), stand for the number of properly detected instances. Cost matrix is similarly defined, as well, and entry C (i, j) represents the cost penalty for misclassifying an instance belonging to class i into class j. Cost matrix values employed for the KDD Cup 99 classifier learning contest are shown in Table 2. A Cost per Test (CPT) is calculated by using the following formula: [17]

$$\mathrm{PT} = 1/N \sum_{i=1}^{m} \quad \sum_{j=1}^{m} \mathrm{CM}(\mathrm{i},\mathrm{j}) * \mathrm{C}(\mathrm{i},\mathrm{j})(1)$$

Where CM and C is confusion matrix and cost matrix, respectively, and N represents the total number of test instances, m is the number of the classes in classification. The accuracy is based on the Percentage of Successful Prediction (PSP) on the test data set.

$$PSP = \frac{number\ of\ successful\ instance\ classification}{number\ of\ instance\ in\ the\ test\ set} \ (2)$$

III. Proposed Method

In this paper we proposed a fused cascaded SVM classifier for intrusion awareness alert generation and data generation. Data fusion is collective collection of data event frequency for expected event. In SVM we perform the cascading process with data fusion. . All of the features are ranked based on their KullbackLeibler (K-L) distances, which is an alternative way to measure the importance of a feature in discriminating two classes. The features discriminating based on the equiliden distance formula for finding a similarity of features based on attack category. After calculation of discriminate we apply parallel support vector machine [18]. SVM which was developed by Vapnikis one of the methods that is receiving increasing attention with remarkable results. SVM implements the principle of Structural Risk Minimization by constructing an optimal separating hyper plane in the hidden feature space, using quadratic programming to find a unique solution. Originally SVM was developed for pattern recognition problems. Recently, a regression version of SVM has emerged as an alternative and powerful technique to solve regression problems by introducing an alternative loss function[13].

Although SVM has been successfully applied in many fields, there is a conspicuous problem appeared in the practical application of SVM. In parallel SVM machine first we reduced non-classified features data by distance matrix of binary pattern. From this concept, the cascade structure is developed by initializing the problem with a number of independent smaller optimizations and the partial results are combined in later stages in a hierarchical way, as shown in figure 1, supposing the training data subsets and are independent among each other.

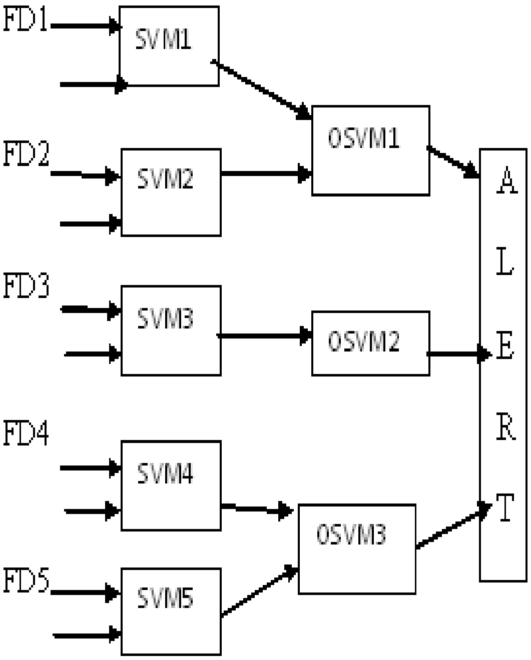

Figure 1: shows that block diagram of fused cascaded SVM

This figure shows that cascaded support vector machine, in this machine we passed five stage of features discernment and all these passes to optimized support vector machine for the processing of classification.

1. Step for data preprocessing.

- Transform data to the format of an SVM
- Conduct scaling on the data
- Consider the RBF kernel K(x; y)
- Use cross-validation to 2nd the best parameter C and
- Use the best parameter C and to train the whole training set
- Generate formatted data.

2. Step of cyber data classification.

Table 2: shows the comparative result of FSVM and rule based method

Metric		**Accuracy (%)**	**Precision (%)**	**Recall (%)**
Data-Set 1	RULEBASED	92.14	87.24	84.43
	FSVM	97.14	96.11	94.10
Data-Set 2	RULEBASED	89.90	84.32	83.23
	FSVM	95.23	92.14	91.21
Data-Set 3	RULEBASED	91.34	86.14	85.11
	FSVM	95.12	93.21	91.13
Data-Set 4	RULEBASED	92,22	88.21	87.66
	FSVM	97.13	94.52	93.67

```
    1. Read preprocessing data
    2. For all the classes are represented
   BEGIN
   Find class with no attribute
Find class at Max cross product rate
Find the class at half cross product

Repeat
Pointer= False
 Find the intervals of hyper plane
If the end condition is met
Pointer = True
If the first interval has better results we should
Use this, otherwise the other
Find the class evaluation after cross product class
Instances middle times
UNTIL pointer= False
END
3. Multiply all the classes with the best factor obtained
4. Data classified
```

IV. Experimental Result

All the experiments were performed on an Intel ® Core ™ i3 with a 2.27GHz CPU and 4 GB of RAM. We used MATLAB version 2009 software. To evaluate the performance of our proposed cyber-attack detection system, we used the KDDCUP1999 dataset. Our experiment is split into three main steps. In the first steps, we prepare different dataset for training and testing. Second, we apply data fused subset selection decision fusion algorithm (DFSDF) to the dataset. The original KDDCUP1999 dataset to select most discriminate features for intrusion attack detection. Third, we classify the intrusion attacks by using fused SVM (FSVM) as classifier. For the performance evaluation we used four different data set of KDDCUP99

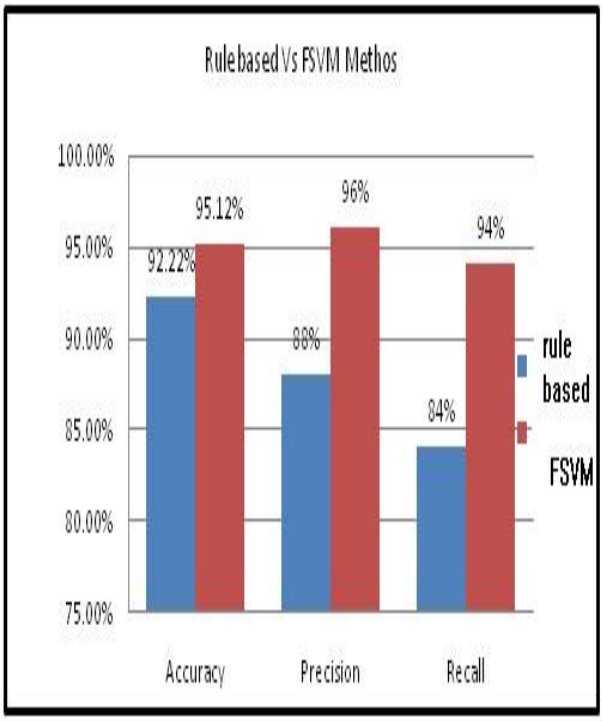

Figure 2: shows that comparative result of rule based method and DSVM

V. Conclusion

In this paper we proposed a new method for security alert generation for intrusion detection. Such method based on data fused support vector machine. In this method we used cascaded SVM with data fusion. Our empirical result shows the better performance in compression of rule based technique of security alert generation. This approach can discover new alert relations and does not depend on background knowledge. At last, we tested our methods on DARPA 2000 Dataset. The simulations showed that with the proposed methods FSVM system can efficiently analyze large amount alerts and save administrators' time and energy. In future we used auto correlation for better prediction of precision and recall. And also reduced the time complexity of data fusion process.

References

[1] J.R. Goodall, W.G. Lutters, and K. Anita, "The work of intrusion detection: rethinking the role of security analysts," Proceeding of the Tenth Americas Conf. on Information System, New York, August 2004, pp. 1421-1427.

[2] M.R. Endsley, "Design and Evaluation for Situation Awareness Enhancement," Proceeding of the human factors society 32nd annual meeting, Santa Monica, CA, 1988, pp. 97-101.

[3] T. Bass, "Multi-sensor Data Fusion for Next Generation Distributed Intrusion Detection Systems," Proceeding of the IRIS national symposium on sensor and data fusion, June, 1999, pp. 99-105.

[4] W. Yurcik, "Visualizing NetFlows for Security at Line Speed: The SIFT Tool Suit." Proceedings of 19th Usenix Large Installation System Administration Conference (LISA), San Diego, CA, USA, Dec. 2005, pp. 169-176.

[5] Carnegie Mellon's SEI. "System for Internet Level Knowledge (SILK)," Http://silktools.source forge.net, 2005.

[6] A.N. Steinburg, C.L. Bowman, and F.E. White, "Revisions to the JDL Data Fusion Model," Joint NATO/IRIS Conference, Quebec, October, 1998.

[7] D.L. Hall, Mathematical Techniques in Multisensor data Fusion. Bosston: Artech House, 2004.

[8] R.Y. Cui, and B.R. Hong, "On Constructing Hidden Layer for Three-Layered Feedforward Neural Networks," Journal of Computer Research and Development, Apr. 2004, Vol. 41, No. 4, pp. 524-530.

[9] X.D. Zhou, and W. Deng, "An Object-Oriented Programming Framework for Designing Multilayer Feedforward Neural Network," Journal of Soochow University, Soochow, China, Feb. 2006, pp. 57-61.

[10] M. Moradi, and M. Zulkernine. "A Neural Network Based System for Intrusion Detection and Classification of Attacks," Proeeding of 2004 IEEE International Conference on Advances in Intelligent Systems, Luxembourg, 2004.

[11] J. Chen, Multisensor management and information fusion. Northwest Industry University, Xian, 2002.

[12] Lincoln Laboratory, Massachusetts Institute of Technology, Darpa Intrusion Detection Evaluation, 2001, Software, Available: http://www.ll.mit.edu 358.

[13] M. Zhang, and J.T. Yao, "A Rough Sets Based Approach to Feature Selection," Proceeding of the 23rd International Conference of NAFIPS, Banff, 2004, pp. 434-439.

[14] R.P. Lippmann, and R.K. Cunningham, "Improving Intrusion Detection Performance Using Keyword Selection and Neural Networks," Computer Networks, 2000, pp. 597-603.

[15] C. Siaterlis, and B. Maglaris, "Towards multisensor data fusion for DoS detection," Proeeding of the 2004 ACM Symp. on Applied Computing, New York, 2004, pp. 439-446.

[16] J.W. Zhuge, D.W. Wang, Y. Chen, Z.Y. Ye, and W. Zou, "A Netwrok Anomaly Detection Based on the D-S Evidence Theory," Journal of Softwoare, March 2006, pp. 463-471.

[17] X.W. Liu, H.Q. Wang, Y. Liang, and J.B. Lai, "Heterogeneous Multisensor Data Fusion with Neural Network: Creating Network Security Situation Awareness," Proceeding of ICAIA'07, Hong Kong, March 2007, pp. 42-4.

[18] J. Kong, "Anonymous and untraceable communications in mobile wireless networks," Ph.D, dissertation, 2004, chair-Gerla, Mario.