Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 30 (2012) 1 – 9

Procedia Engineering

www.elsevier.com/locate/procedia

International Conference on Communication Technology and System Design 2011

A Hybrid Intelligent Approach for Network Intrusion Detection

Mrutyunjaya Panda[a], Ajith Abraham[b], Manas Ranjan Patra[c], a*

[a]*Department of ECE, Gandhi Institute for Technological Advancement, Bhubaneswar-54, India*
[b]*Machine Intellegence Reseacrh Labs (MIR Labs), WA, USA*
[c] *Department of Computer Science, Berhampur University, India*

Abstract

Intrusion detection is an emerging area of research in the computer security and networks with the growing usage of internet in everyday life. Most intrusion detection systems (IDSs) mostly use a single classifier algorithm to classify the network traffic data as normal behaviour or anomalous. However, these single classifier systems fail to provide the best possible attack detection rate with low false alarm rate. In this paper, we propose to use a hybrid intelligent approach using combination of classifiers in order to make the decision intelligently, so that the overall performance of the resultant model is enhanced. The general procedure in this is to follow the supervised or un-supervised data filtering with classifier or clusterer first on the whole training dataset and then the output is applied to another classifier to classify the data. We use 2-class classification strategy along with 10-fold cross validation method to produce the final classification results in terms of normal or intrusion. Experimental results on NSL-KDD dataset, an improved version of KDDCup 1999 dataset show that our proposed approach is efficient with high detection rate and low false alarm rate.

Key words: Intrusion detection system; hybrid approach; sPegasis; Radial Basis function; PCA; Decision tree;

1. Main text

With the development of the internet and its wide application in all domains of everybody's life, intrusion detection is becoming a critical process in computer network security. Intrusion detection systems (IDS) attempts to recognize and notify the users' activity as either normal or anomaly (or

* Mrutyunjaya Panda. Tel.: +91-9437338270;
E-mail address: mrutyunjaya@ieee.org.

1877-7058 © 2011 Published by Elsevier Ltd.
doi:10.1016/j.proeng.2012.01.827

intrusion) by comparing the network connection records to the known intrusion patterns obtained from the human experts. As traditional methods can not detect the unknown intrusion patterns efficiently because of the problems faced by a human analyst during analyzing a faster and complex network, we concentrate on data mining based intelligent decision technology to make effective decisions to this effect. As demonstrated in [1], [2], Data Mining based intrusion detection falls into two categories: (1) misuse detection and (2) anomaly detection. IDS that employ misuse detection build the intrusion patterns by learning from the labeled data with a drawback of not bale to identify the novel attacks that are not present in the training data. In contrast, anomaly detection are capable of identifying the new or unseen intrusions, that are not available in the training data and learn the model from the behaviour of the normal activities.

Until now, a great deal of time and resources have been invested in IDS and various machine learning approaches such as: Support vector machines (Chen et al., 2005 [3], Mukkamala et al., 2003 [4]), Bayesian belief networking (Panda and Patra, 2009 [5]), Artificial neural network (Zhang et al., 2003 [6]), data mining methods (Wu and Yen, 2009 [7] and hybrid intelligent system (Peddabachigari et al., 2007 [8]) are investigated to model the IDS. However, it seems that none of them is able to detect all kind of intrusion attempts efficiently in terms of detection rate and false alarm rate. Hence, the need is to combine different classifiers as a hybrid data mining strategy to enhance the detection accuracy of the model built in order to make efficient intelligent decisions in identifying the intrusions. The rest of the paper is organized as follows. In Section 2, we discuss briefly .about the existing literature in this area of research. Section 3 introduces the various intelligent decision technologies used in this paper, followed by an analysis to the NSL-KDD, a benchmark intrusion detection dataset in Section 4. The proposed methodology adopted in this paper is explained in Section 5 with the simulation results and their analysis in Section 6. Finally, we conclude the paper with conclusion and future direction of research in Section 7.

2. Review of Related Research

In early days, Ilgun et al. [9] used rule based methods as expert systems to design IDS, where the knowledge of human experts is encoded into a set of rules. Lee and Stolfo [10] use the data mining approach to derive association rules and frequent episodes from sample data, rather than from human experts so that a predictive model can be constructed. The drawback of such frameworks is to produce the large number of association rules and thereby, increase the complexity of the system. In order to consider this limitation, association rule mining with multiple minimum support along with various interestingness measures in identifying the activities that match the defined characteristics of normal or intrusion is proposed by Panda and Patra [11]. In [12], Tavallaee et al. investigated and reviewed the current state of experimental practice along with their common pitfalls in the area of anomaly based intrusion detection. Chou et al. [13] presented an information theoretic feature selection algorithm on both high and low dimensional feature spaces with correlation analysis; thus verifying the performance of the IDS using a combination of k-nearest neighbor, fuzzy clustering and Damper-Shafer theory. A rough set based parallel genetic algorithm hybrid model is considered to address the important features in building an IDS is considered by Mahmud et al. in [14]. Panda and Patra [15] presented the effectiveness of hybrid clustering approach using COBWEB and FFT clustering in detecting novel attacks. Wang et al. [16] implemented a new Meta learner fuzzy aggregation approach to intrusion detection using artificial neural networks and fuzzy clustering and clamed that their approach provides better detection rate and stability in comparison to the back propagation neural network, Decision Tree and Naïve Bayes. Zhang and Feng [17] used support vectors and ant colony to build an intrusion detection model. PCA-ICA ensembled

intrusion detection system by pareto-optimal optimization to obtain the optimal weight for the ensemble system by Gu et al. [18] and concluded of outperforming the standard SVM, PCA SVM and ICA SVM. Panda and Patra used a hybrid NBDT by combining Naïve Bayes with Decision tree along with AdaBoost to produce the best detection rate with false positive rate in designing IDS. A hybrid intelligent approach for automated alert clustering and filtering in intrusion alert analysis is presented by Siraj et al. [19] in terms of classification accuracy and processing time.

3. Intelligent Decision Technologies

A novel aspect of our work is the use of intelligent decision techniques using data filtering through a supervised or unsupervised feature selection algorithm to select significant features followed by a classifier to model the network intrusion detection system. In particular, we investigate the combination of Decision trees, principal component analysis, SPegasos (Stochastic variant of Piramol estimated sub-gradient solver in SVM), END, Random Forest and Grading for this purpose, which are briefly discussed below.

3.1. Decision Trees

In this, the target concept is represented in the form a tree, where the tree is built by using the principle of recursive partitioning. In this, attributes are selected as a partitioning attribute or as a node based on the information gain criteria and then the process continues repeatedly for every child node until all attributes are considered and a decision tree is constructed. Some pruning techniques may further be considered so that the size of the tree is reduced and the overfitting is thereby avoided [20].

3.2. Principal Component Analysis(PCA)

PCA is an unsupervised feature selection based on multivariate statistics and its basic idea is to seek a projection that represents the data in a best possible way in a least-square sense to provide dimensionality reduction. Many researchers pointed out that PCA, which is also known as Karhunen-Loeve transformation in pattern recognition is not found suitable in feature extraction in classification process for the non inclusion of discriminatory information in calculating the optimal rotation of the feature axes.. However, Skurichina and Duin [21] advocate for the usage of PCA for better accuracy, we use PCA to determine its feasibility and to observe its suitability to increment the classification accuracy and diversity in detecting network intrusions.

3.3. Stochastic variant of Piramol estimated sub-gradient solver in SVM (SPegasos)

SPegasos implements the stochastic variant of the Pegasos (Primal Estimated sub-GrAdient SOlver for SVM) method of Shalev-Shwartz et al. [22]. This implementation globally replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes, so the coefficients in the output are based on the normalized data. In this, hinge loss (SVM) is minimized for optimizing the performance of the proposed intrusion detection system.

3.4. END

END (Ensembles of Balanced Nested Dichotomies for Multi-class Problems) is a Meta classifier for handling multi-class datasets with 2-class classifiers by building an ensemble of nested dichotomies. More details about this can be obtained from Dang et. al [23].

3.5. Grading

In this type of Meta classifier, the base classifiers are graded to enhance the performance of IDS. We use "graded" predictions (i.e., predictions that have been marked as correct or incorrect) as meta-level classes. For each base classifier, one Meta classifier is learned whose task is to predict when the base classifier results in error. Hence, the way stacking viewed as a generalization of voting, grading may be viewed as a generalization of selection by cross-validation and therefore fills a conceptual gap in the space of meta-classification schemes. More details about this can be found from Seewald and Fuernkranz [24].

3.6. Random Forest

Random forest as suggested by Breiman [25] is an ensemble of unpruned classification or regression trees, induced from bootstrap samples of the training data, using random feature selection in the tree induction process. Prediction is made by aggregating the predictions of the ensemble by majority voting for classification. It yields generalization error rate and is more robust to noise. However, similar to most classifiers, RF can also suffer from the curse of learning from an extremely imbalanced training data set. As it is constructed to minimize the overall error rate, it will tend to focus more on the prediction accuracy of the majority class, which often results in poor accuracy for the minority class.

4. Intrusion Detection Dataset

We use NSL-KDD dataset, developed by Tavallaee et al. [26], an enhanced version of KDDCup 1999 benchmark intrusion detection dataset because of the inherent problems. The first important limitation in the KDDCup 1999 dataset is the huge number of redundant records in the sense that almost 78% training and 75% testing records are duplicated, as shown in Table 1 and Table 2; which cause the learning algorithm to be biased towards the most frequent records, thus prevent it from recognizing rare attack records that fall under U2R and R2L categories. At the same time, it causes the evaluation results to be biased by the methods which have better detection rates on the frequent records. It is also stated that though the NSL-KDD dataset still suffers from some of the problems discussed and may not be a perfect representative of existing real networks, it can be applied an effective benchmark dataset to detect network intrusions. In this NSL-KDD dataset, the simulated attacks can fall in any one of the following four categories.

- Probing Attack: this is a type of attack which collect information of target system prior to initiating an attack. Some of the examples are Satan, ipsweep, nmap attacks.

- DoS Attack: Denial of Service (DoS) attack results by preventing legitimate requests to a network resource by consuming the bandwidth or by overloading computational resources. Examples of this are Smurf, Neptune, Teardrop attacks.
- User to Root (U2R) Attack: In this case, an attacker starts out with access to a normal user account on the system and is able to exploit the system vulnerabilities to gain root access to the system. Examples are eject, load module and Perl attacks.
- Root to Local (R2L) Attack: In this, an attacker who doesn't have an account on a remote machine sends packet to that machine over a network and exploits some vulnerabilities to gain local access as a user of that machine. Some examples are ftp_write, guess password and imap attacks.

Table 1: Redundant Records in KDD 1999 Training Dataset

	Original Records	Distinct Records	Reduction Rate
Normal	972,781	812,814	16.44%
Anomaly	3,925,650	262,178	93.32%
Total	4,898,431	1,074,992	78.05%

Table 2: Redundant Records in KDD 1999 Testing Dataset

	Original Records	Distinct Records	Reduction Rate
Normal	60,591	47,911	20.92%
Anomaly	250,436	29,378	88.26%
Total	311,027	77,289	75.15%

As there are still some critiques of attack taxonomies and performance measures, we concentrate on anomaly based intrusion detection systems with 2-class classifications, i.e., anomalous and normal, rather than identifying the detailed information of the attacks.

5. Proposed Methodology

The framework for the proposed approach is shown in Fig 1 below. In this, we propose to use combining classifier strategy in order to make intelligent decisions. In this, the data filtering is done after adding supervised classification or unsupervised clustering to the training dataset. Then the filtered data is applied to the final classifier methods to obtain the final decision; which is then verified by using 10-fold cross validation method to understand the suitability of the proposed approach.

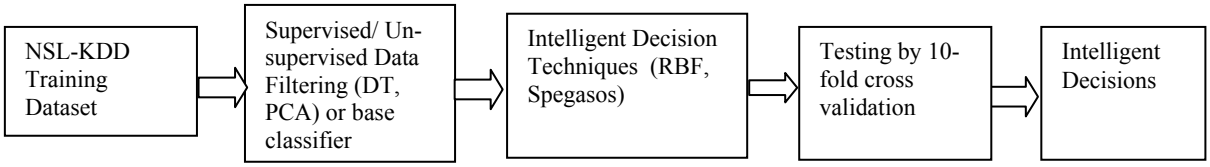

Fig 1. The proposed Hybrid Intelligent IDS

The advantage of applying such a methodology is as follows. At first, the combination strategy produce a similarity score or probability produced by each classifier , the distribution of which reveals how confident it is in making an intelligent decision. Secondly, the approach will learn from synthesized feature vectors of several classifiers and can make a better decision than a any individual strategy. Further, we use some Meta learning strategies using base classifier to enhance the performance of the proposed system using Grading and END.

5.1. Performance Evaluation

The following are some of the performance measures considered to evaluate the efficacy of the proposed IDS.

- The true positive rate (TPR) is the proportion of examples which were classified as class x, among all examples which truly have class x, i.e. how much part of class were captured. It is equivalent to detection rate or sensitivity. In case of information retrieval, it is called as recall.
- The false positive rate (FPR) , which is also known as false alarm rate is the proportion of examples which were classified as class x, but belong to a different class, among all examples which are not of class x.
- Precision is another information retrieval term, which is often paired with recall. It is defined as the proportion of examples which truly have class x among all those which were classified as class x.
- F-value combines the TPR and precision into a single utility function after obtaining their harmonic mean.
- Building time is the time taken by the classifier to build the model in seconds.

6. Experimental Results and Discussions

We conduct all our experiments in an IBM PC of 2.66GHz CPU with 40GB HDD and 512 MB RAM in Java Environment. In order to make intelligent decisions with various intelligent decision technologies, NSL-KDD Dataset; a new intrusion detection benchmark dataset having 25192 training instances with 10-fold cross validation for testing is used for building an efficient Network intrusion detection system.

From Table 3, it is evident that hybridization of Random forest with nested dichotomies and END, the intrusion detection rate is 99.5% with extremely low false alarm rate of 0.1%, which is quite encouraging in comparison to all other categories. It is also faster than all others taking only 18.13 seconds to build the model, provides low root mean squared error of 0.045. Further, high F-value, high precision rate and recall rate with 99.7%, 99.9% and 99.9% respectively shows the combination strategy a good choice for making intelligent decisions. In comparison to all, the combination of data filtering by decision tree (J48) with classification by radial basis function neural network (RBF) presents low performance with only 90.7% intrusion detection rate, high false alarm rate with 5.6% and more importantly more error prone with root man squared error of 0.25.

Experiment Number/ Performance Matrices		1 Filter-Decision Tree(J48) Classifier-RBF	2 Filter-Decision Tree(J48) Classifier-SPegasos	3 Filter-PCA Classifier-SPegasos	4 SPegasos	5 Grading+ SPegasos +J48	6 END+ Nested Dichotomies +Random Forest
Detection Rate (%)	Normal	94.4	98.5	98.3	98.5	99.6	99.9
	Attack	90.7	96.3	95.7	96.3	99.5	99.5
False positive rate (%)	Normal	9.3	3.7	4.3	3.7	0.5	0.5
	Attack	5.6	1.5	1.7	1.5	0.4	0.1
F-Value (%)	Normal	93.2	97.6	97.3	97.6	99.6	99.8
	Attack	92	97.2	96.8	97.2	99.5	99.7
Time taken to build the model in Seconds		52.61	122.27	177.63	90.5	156.42	18.13
Precision (%)	Normal	92.1	96.8	96.3	96.8	99.6	99.6
	Attack	93.4	98.2	98	98.2	99.5	99.9
Recall (%)	Normal	94.4	98.5	98.3	98.5	99.6	99.9
	Attack	90.7	96.3	95.7	96.3	99.5	99.9
Root Mean Squared Error		0.25	0.16	0.17	0.16	0.068	0.045

Table 3. Comparison of Result

7. Conclusions and Future Scope

In this paper, we investigated some novel hybrid intelligent decision technologies using data filtering by adding supervised or un-supervised methods along with a classifier to make intelligent decisions in order to detect network intrusions. We use a variant of KDDCup 1999 dataset, NSL-KDD to build our proposed IDS. The performance comparison amongst different hybrid and combination of classifiers were made in order to understand their effectiveness in terms of various performance measures. Finally, we conclude that our proposed hybridization of END with nested dichotomies and random forest of 10 trees with out of bag error of 0.06 results almost 100% intrusion detection rate with 0% false alarm rate, which makes the approach as most efficient. In future, we will concentrate on 5-class classification with cost based strategy with more decision technology methods to make intelligent decisions while detecting network intrusions.

References

[1] Lee W and Stolfo S., "Data Mining techniques for intrusion detection", In: *Proc. of the 7th USENIX security symposium*, San Antonio, TX, 1998.

[2] Dokas P, Ertoz L, Kumar V, Lazarevie A, Srivastava J, and Tan P., "Data Mining for intrusion detection", In: *Proc. of NSF workshop on next generation data mining*, 2002.

[3] Chen RC, Chen J, Chen TS, Hsieh C, Chen TY and Wu KK, "Building an Intrusion detection system based on SVM and genetic algorithm", In: *Advances in NN-ISNN* 2005, pt 3, proceedings, 2005; 3498:409-14.

[4] Mukkamala S and Sung AH, "Feature selection for intrusion detection with neural network and Support vector machines", *Transportation security infrastructure prot.,* 2003; 1822:33-9.

[5] Panda M and Patra MR. Bayesian, "Belief network with genetic local search for detecting network intrusions", *International journal of secure digital information age* 2009; 1(1):34-44.

[6] Zhang CL, Jiang J and Karnel M., "Comparison of back propagation learning and radial basis function network in Intrusion detection system", In: *Rough set, fuzzy set, data mining and granular computing,* 2003; 2639:466-70.

[7] Wu S and Yen E. , "Data Mining based intrusion detectors", *Expert system with applications journal* 2009; 36(3):5605-12.

[8] Peddabachigari S, Abraham A, Grosan C and Thomas J, "Modelling intrusion detection system using hybrid intelligent" systems, *Journal of computer and network applications journal* 2007;30(1):114-32.

[9] Ligun K, Kemmerer R and Porras P., "State transition analysis: a rule based intrusion detection approach" , *IEEE Transaction on software engineering*,1995, 181-99.

[10] Lee W, Stolfo S and Mok K, "A data mining framework for building intrusion detection model", In: *Proc. of IEEE symposium on security and privacy* , 1999, 120-32.

[11] Panda M and Patra MR. Mining association rules for constructing a network intrusion detection model, *International journal of applied engineering research,* 2009,4(3):381-98.

[12] Tavallaee M, Stakhanova N and Ghorbani AA., "Towards credible evaluation of anomaly based intrusion detection methods", *IEEE Transaction on System, Man and Cybernetics, Part-c, Applications and Reviews, 2010;* 40(5):516-24.

[13] Chan TS, Yen KK and Luo J., "Network intrusion detection design using feature selection of soft computing paradigms", *International journal of computational intelligence* ,2008, 4(3):196-208.

[14] Mahmud WM, Agiza HN and Radwan E., " Intrusion detection using rough sets based parallel genetic algorithm hybrid model", In: *Proc. of the world congress on Engineering and computer Science (WCECS-2009)*, USA.

[15] Panda M and Patra MR., "A Hybrid clustering approach for network intrusion detection using cobweb and FFT", *Journal of Intelligent systems*,2009,18(3):229-45.

[16] Wang G, Hao J, Ma J and Huang L, "A new approach to intrusion detection using ANN and fuzzy clustering", Expert *systems with application journal*, 2010, Elsevier.

[17] Zhang Q and Feng, " W. Network intrusion detection by support vectors and ant colony", In: *Proc. of 2009 Intl. workshop on information security and applications (IWISA 2009),* China, p.639-42, Academy Publisher.

[18] Panda M and Patra MR., "Semi Naïve Bayesian method for anomaly based network intrusion detection", In: Proc. of ICONIP 2009, Thailand, *Lecture Notes in Computer Science,*2009*;* 5863:614-21.

[19] Siraj MMd, Maarof MdAand Hashim SZ Md., " A hybrid intelligent approach for automated alert clustering and filtering in intrusion alert analysis", *Intl. Journal of computer theory and engineering,* 2009; 1(5):539-45.

[20] Mitchell TM, "*Machine Learning"*, McGraw Hill, 1997.

[21] Skurichina M and Duin RPW , "Combining feature subsets in feature selection", In: *Proc. of 6th Intl. workshop on multiple classifier systems (MCS 2005)*, p. 165-175.

[22] Shalev-Shwarz S, Singer Y and Srebro N, "Pegasos: Piramal estimated sub-gradient solver for SVM", In: *Proc. of 24th Intl. conf. on machine learning* , 2007, p. 807-14.

[23] Dong L, Frank E and Kramer S., "Ensembles of balanced nested dichotomies for multiclass problems", In: *Proc. of PKDD* , 2005, p. 84-95.

[24] Seewald AK and Fuernkranz J, "An evaluation of grading classifier", In: *Proc. of advances in intelligent data analysis* 200, p.114-24.

[25] Breiman L. Random Forests, "Machine Learning", 2001, 45(1):5-32.

[26] Tavallaee M, Bagheri E, Lu W and Ghorbani AA., "A detailed analysis of the KDD Cup datasets", In: *Proc. of 2009 IEEE Symposium on computational intelligence in security and defence applications (CISDA-2009)*.