**International Journal of Information & Network Security (IJINS)**
Vol.1, No.3, August 2012, pp. 228~234
ISSN: 2089-3299 

# Pattern Based Network Security Using Semi-Supervised Learning

**V K Pachghare*, V K Khatavkar*, Dr Parag Kulkarni****
* Assistant Professor, Department of Computer Engg. & IT, College of Enineering, Pune.
** Adjunct Professor, Department of Computer Engg. & IT, College of Enineering, Pune.

**Article Info**

***Article history:***

Received Jun 30th, 2012
Revised July 10 th, 2012
Accepted July 25th, 2012

***Keyword:***

Intrusion Detection
supervised learning
semi-supervised learning
patterns matching

**ABSTRACT**

Network security is becoming increasingly important in today's internet-worked systems. With the development of internet, its use on public networks, the number and the severity of security threats has increased significantly. Intrusion Detection System can provide a layer of security to these systems. The goal of intrusion detection system is to identify entities who attempt to subvert in-place security controls. The field of machine learning is gaining increasing attention in the development of intrusion detection systems. The machine learning techniques used for solving intrusion detection problem can be broadly classified into three broad categories: Unsupervised, supervised and semi-supervised. The supervised learning method exhibits good classification accuracy for known attacks. But it requires large amount of training data. In real world the availability of labeled data is time consuming and costly. An emerging field of semi-supervised learning offers a promising direction for further research. So in this work we propose a semi-supervised approach for pattern based IDS to improve performance of supervised approach. The experimentation is performed on KDD CUP99 dataset.



***Corresponding Author:***

Mr. V. K. Pachghare
Assistant Professor,
Department of Comp. Engg. & IT College of Engineering, Pune-5
Email : vkp.comp@coep.ac.in

## 1. INTRODUCTION

Network security is becoming increasingly important in modern internet-worked systems. With the development of networking and interoperation on public networks, the number and the severity of security threats has increased significantly. Intrusion Detection System can provide a layer of security to these systems. An intrusion is defined by Heady et al. as any set of actions that attempt to compromise the integrity, confidentiality, or availability of a resource [1]. An intrusion detection system (IDS) is a system for the detection of intrusions. Intrusion detection involves detecting unusual patterns of activity or patterns of activity that are known to correlate with intrusions. We can classify IDS into two main types: anomaly and misuse detection. The anomaly detection approach establishes the profiles of normal behavior of users, systems, system resources, network traffic and/or services and detects intrusions by identifying significant deviations from the normal behavior patterns observed from profiles. The misuse detection approach defines suspicious misuse signatures based on known system vulnerabilities and a security policy. According to the difference in monitoring objects, IDSs are divided into network-based IDSs and host-based IDSs. Machine learning and pattern recognition methods have been utilized to detect intrusions. Learning algorithms can be categorized as unsupervised, supervised and semi-supervised. Unsupervised learning studies how systems can learn to represent particular input patterns in a way that reflects the statistical structure of the overall collection of input patterns. It learns from unlabelled examples. Supervised machine learning methods require labeled data for training. The objective of supervised learning is to learn about assigning correct

labels to new unseen examples of the same task. With the immense amount of network and host data available, expert labeling of the data is very expensive. The labeled data available is often from controlled environments. For instance, the 1998 and 1999 intrusion detection evaluations from DARPA/MIT Lincoln lab [2] have the ground truth information, but the data itself have been shown to be not representative of real environments. This proves to be a bottleneck in applying supervised learning methods to detect novel or unknown attacks. Hence, relying only on supervised learning methods which requires a large amount of labeled data is impractical for real network environment. This motivates a need for a new and more practical learning framework. Semi-supervised learning methods can leverage unlabeled examples in addition to labeled ones. Semi-supervised learning methods received significant attention, and are more suitable for real network environment because these methods require a small quantity of labeled data while still taking advantage of the large quantities of unlabeled data. In this paper we propose semi supervised approach for intrusion detection.

Boosting algorithms are very useful for improve the performance of intrusion detection system. Boosting algorithms are greedy methods for forming linear combinations of base hypotheses. In the most common scenario the algorithm is given a fixed set of labeled training data and in each iteration updates a distribution on these data. It is important to simultaneously exploit existing knowledge of attacks, to exploit the copious amounts of known normal data, and to be capable of detecting attacks unrelated to known attacks. We demonstrate a semi-supervised approach to intrusion detection that supports features of intrusion detection system and allows flexible training and adaptation. This proposed method also offers the advantage of not requiring a separate method to label the data. Instead of that we use the labeled data of testing and filtered data from the testing data is uses to refine the existing dataset and the new labeled data automatically trained the system. While when labeled data becomes available the learner incorporates it into the algorithm for training. The data we used in our experiments is KDDcup99 and is considered a benchmark for intrusion detection evaluations. Our algorithm gives better performance than supervised learning approach.

The rest of the paper is organized as follows. Section II describes the literature survey about semi-supervised methods for intrusion detection system. Section III describes our proposed approach for semi supervised learning method for intrusion detection followed by experiments and results in Section IV, followed by a conclusion in the last Section.

## 2. LITERATURE SURVEY

Semi-supervised learning methods use unlabeled data to either modify or reprioritize hypotheses obtained from labeled data alone. Recently, learning with labeled and unlabeled data, also known as semi-supervised learning has attracted much attention [13, 14]. It aims to achieve good classification performance with the help of unlabelled data in the presence of the small sample problem, and some promising results have been reported. Enlightened by this, instead of training the model with more labeled data, we incorporate the unlabelled data before active learning starts.

Many existing semi-supervised learning methods use a generative model for the classifier and employ Expectation-Maximization (EM) to estimate the label or model parameters. Other semi-supervised learning methods include self training, co-training, transductive support vector machine and graph-based methods [15]. An appropriate semi-supervised learning method whose assumptions fit the application at hand should be considered [12]. Existing semi-supervised classification algorithms may be classified into two categories based on their underlying assumptions. An algorithm is said to satisfy the manifold assumption if it utilizes the fact that the data lie on a low dimensional manifold in the input space. Usually, the underlying geometry of the data is captured by representing the data as a graph, with samples as the vertices, and the pair-wise similarities between the samples as edge-weights. Several graph based algorithms such as Label propagation, Markov random walks, Graph cut algorithms, and Spectral graph transducer and Low density separation [16, 17] are based on this assumption.

Graph-based approaches represent both the labeled and the unlabeled examples by a connected graph, in which each example is represented by a vertex, and pairs of vertices are connected by an edge if the corresponding examples have large similarity. The well known approaches in this category include Harmonic Function based approach, Spectral Graph Transducer (SGT), Gaussian process based approach, Manifold Regularization and Label Propagation approach [12]. The optimal class labels for the unlabeled examples are found by minimizing their inconsistency with both the supervised class labels and the graph structure. Semi-supervised clustering methods are mainly three types: Constraint-based, distance-based, and constraint and distance based semi-supervised clustering methods [11]. Ching-Hao Mao et al used Co-training and Active Learning based Approach for Multi-view intrusion detection for semi-supervised approach [17]. Chien-Yi Chiu et al proposed Semi-supervised Learning for False Alarm Reduction. They use Feature selection using information gain and gain ratio and Over-sampling positive points before base learner training the classifier

[18]. Jimin Li et.al proposes a novel Semi-supervised SVM Based on Tri-training for Intrusion Detection. They use three different SVMs as the classification algorithm. They use UCI data sets and application to the intrusion anomaly detection show that tri-training can improve the classification accuracy of SVM and its improved algorithms [19].

## 3. SEMISUPERVISED APPROACH

It is important to distinguish the problem of semi-supervised improvement from the existing semi-supervised classification approaches. In the semi-supervised improvement problem, we aim to build an classifier which utilizes the unlabeled samples from the output of testing stage of our supervised algorithm Supervised intrusion detection approaches use only labeled data for training. To label the data however are often difficult, expensive, or time consuming as they require the efforts of experienced human annotators. Meanwhile unlabeled data may be relatively easy to collect, but there has been few ways to use them. Semi-supervised learning addresses this problem by using large amount of unlabeled data, together with the labeled data, to build better classifiers. Because semi-supervised learning requires less human effort and gives higher accuracy, it is of great interest both in theory and in practice. Some often-used semi-supervised methods include: EM with generative mixture models, self-training, co-training, transductive support vector machines, and graph-based methods.

The self learning approach has been applied in various arenas of computer technology. This work is an attempt to implement self-learning approach in the field of intrusion detection.

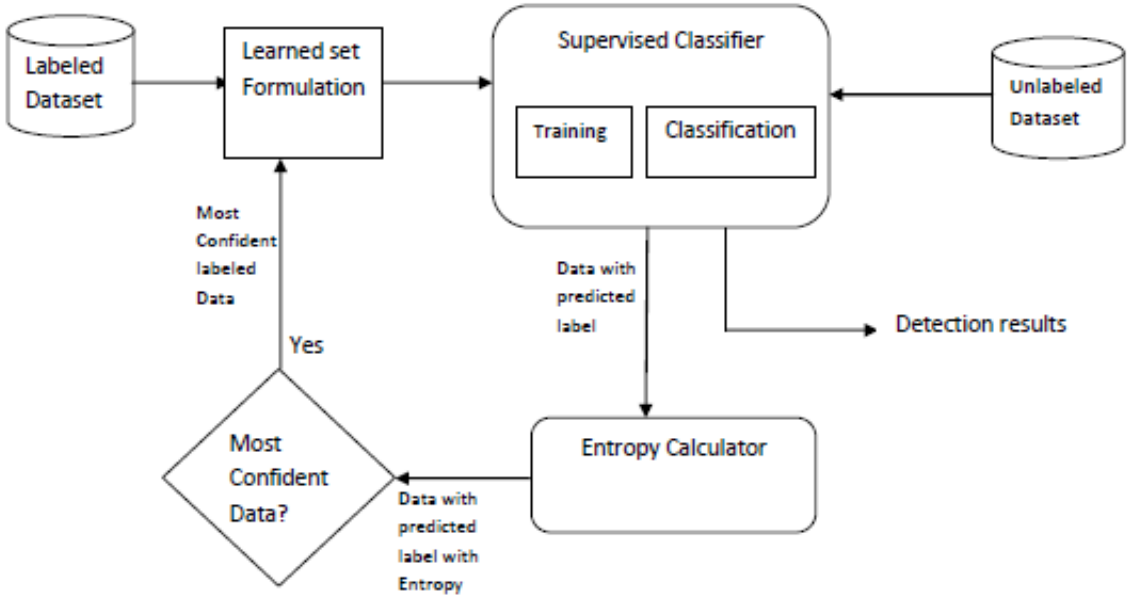


**Figure 1: Architecture for semi-supervised IDS using self learning algorithm (SLA)**

The architecture for proposed semi-supervised IDS using self learning algorithm (SLA) is as shown in Figure 1. Here the labeled data is used for training and unlabeled data is used for testing. Then the most confident data with predicted labels from the output of testing phase is selected and added in the labeled data. The learned set formulation helps to remove the data redundancy in the labeled data and controlled the size of the labeled data.

To select particular data from test data we use entropy. The entropy can be calculated as:

$$E(D) = \sum_{i=1}^{n} -pi \log_2 pi \qquad (8)$$

Where, D is data and $p_i$ is the probability of $i^{th}$ feature.

Entropy for each record is calculated and mean, variance and standard deviation for each type of label is calculated. Using this information we filter the data from test data and add to training dataset. We use statistical approach for filtering the data.

The algorithm for Semi-supervised approach can be summarized as:

**Self Learning Algorithm (SLA)**
1. Train f from labeled data
2. Predict on x ϵ unlabeled data
3. Add (x; f (x)) to labeled data
4. Repeat

The variations in self training are:
1. Add a few most confident (x, f (x)) to labeled data
2. Add all (x, f (x)) to labeled data
3. Add all (x, f (x)) to labeled data, weight each by confidence.

Train the system with this new data. After testing our approach we have the conclusion that the filtered data is not more than 10% of the actual unlabeled data.

## 4. RESULTS AND ANALYSIS

We utilize the KDD CUP 1999 data set for our experiments. It was originated from MIT's Lincoln Lab and developed for IDS evaluations by DARPA. Despite of several drawbacks mentioned, it has served as a reliable benchmark data set for many researches on network based intrusion detection algorithms. In this data set, each TCP/IP connection has been labeled, and 41 features had been extracted, some of which are continuous and others are categorical. So we don't have to do the task of "Feature extraction" and "Data labeling". Hence we can focus on the effectiveness and accuracy of our algorithms of pattern based network security for semi-supervised learning. There is a high imbalance in the data when we do a one vs rest classification. While a knowledge of priors may be used to incorporate this imbalance into semi-supervised learning to achieve high performance, we assume that nothing is known about the data other than the similarity information and a few training examples.

The attacks are categorized into four general categaries: DOS (denial of service), U2R (user to root), R2L (remote to local) and PROBE. In each of the four, there are many low level types of attacks.

**TABLE 1: Training Data set for semi-supervised approach**

| **Normal** | **Attack** | | | | **Total** |
|---|---|---|---|---|---|
| | ***DOS*** | ***R2L*** | ***U2R*** | ***PROBE*** | |
| | *391468* | *2903* | *53* | *6937* | |
| 108227 | 401361 | | | | 509588 |

The number of samples of various types in the training data set used for semi-supervised approach is listed in Table 1.

When the semi-supervised algorithm is applied on the training data set, the results were obtained as shown in figure 2.
After filter the data we get 15567 labeled data which is just 6% of the unlabeled data we used for testing. After adding this labeled data from the output of testing our supervised approach to our labeled data, we train the system. The accuracy of training of our semi-supervised approach is graphically shown below.

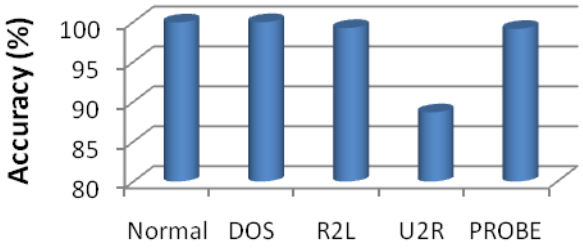

**Fig: 2 Accuracy of proposed semi-supervised algorithm**

Table 2 shows the false-alarm rate and detection rate for training data set for proposed semi-supervised approach.

**TABLE 2: Detection Results in Training Data Set for proposed semi-supervised Algorithm**

| FPR (%) | DR (%) |
|---|---|
| 0.055 | 99.96 |

The detection rate and the false positive rate of the proposed semi-supervised algorithm are better than our supervised algorithm. So, using this approach we improve the performance of supervised approach. The performance of the proposed algorithm is good as compare to other semi-supervised approaches. Figure 3, shows the comparative performance of various approaches of the semi-supervised learning. From figure 3, we observe that the false positive rate of our proposed algorithm is better than the other approaches.

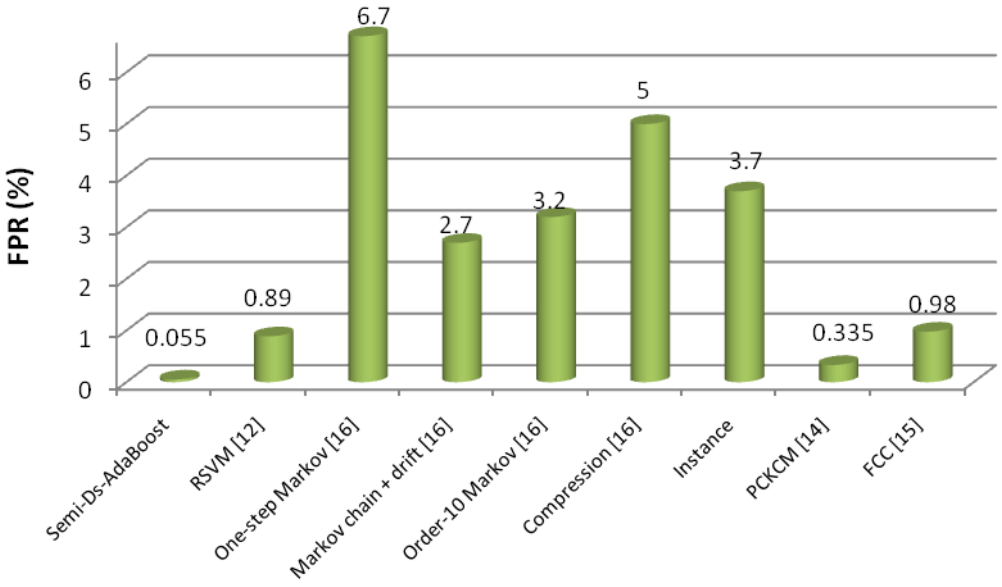


**Fig: 3 Comparison of FPR of Proposed algorithm with other algorithms**

Table 3 shows the Detection rate comparison of proposed semi-supervised algorithm with state of art algorithms.

**TABLE 3 Comparison of Detection Rate**

| Semi-supervised Approach | Proposed semi-supervised Algorithm | RSVM [12] | PCKCM [14] | FCC [15] |
|---|---|---|---|---|
| DR (%) | 99.96 | 90.91 | 88.50 | 98.10 |

The detection rate of our algorithm is quite higher than other algorithms. The detection rate of our algorithm is 99.96% which is much better than other algorithms. Our experiments demonstrate that the performance of the supervised learning method significantly improves using our semi-supervised learning approach. Our findings suggest that the problem of availability of the large amount of labeled data for training can be solved using semi-supervised learning. This algorithm is a multiclass algorithm whereas almost all other semi-supervised classification algorithms are currently two class algorithms. From all above discussion we conclude that the performance of proposed algorithm is better than other traditional algorithms.

## 5. CONCLUSION

We have proposed an algorithm for semi-supervised learning using a boosting framework. The strength of our proposed algorithm lies in its ability to improve the performance of any given base classifier in the presence of unlabeled samples. We have presented an experimental framework in which supervised and semi-supervised learning methods can be evaluated in an intrusion detection system. Our experiments demonstrate that the performance of the supervised learning method significantly improves using our semi-supervised learning approach. The performance of this algorithm is comparable to the state-of-the-art semi-supervised learning algorithms. The observed stability of proposed semi supervised algorithm suggests that it can be quite useful in practice. Our findings suggest that the problem of availability of the large amount of labeled data for training can be solved using semi-supervised learning. This algorithm is a multiclass algorithm whereas almost all other semi-supervised classification algorithms are currently two class algorithms.

## REFERENCES


[1] R. Heady, G. Luger, A. Maccabe, M. Servilla, "The Architecture of a Network Level Intrusion Detection System", Technical report, Department of Computer Science, University of New Mexico, 1990.

[2] S Stolfo et al, "The third international knowledge discovery and data mining tools competition" [online]. Available:http://kdd.ics.uci.eduidatabases/kddCup99/kddCup99.html, 2002.

[3] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods", 33rd Annual Meeting of the Association for Computational Linguistics, pp. 189 –196, 1995.

[4] A. Blum, T. Mitchell, "Combining labeled and unlabeled data with co-training", COLT: Workshop on Computational Learning Theory, 1998.

[5] V. Vapnik, "Statistical learning theory", Wiley-Interscience, 1998

[6] N. D. Lawrence, M. I. Jordan, "Semi-supervised learning via Gaussian processes", L. K. Saul, Y. Weiss and L. Bottou (Eds.), Advances in neural information processing systems 17. Cambridge, MA: MIT Press, 2005

[7] Xiaojin Zhu, "Semi-Supervised Learning Literature Survey", Computer Sciences Technical Report 1530, University of Wisconsin – Madison

[8] X. Zhu, Z. Ghahramani, "Towards semi-supervised classification with Markov random fields", Technical Report CMU-CALD-02-106, Carnegie Mellon University, 2002

[9] X. Zhu, Z. Ghahramani, J. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions", 20th International Conference on Machine Learning (ICML), 2003

[10] C. Kemp, T. Griffiths, S. Stromsten, J. Tenenbaum, "Semi-supervised learning with trees" Advances in Neural Information Processing System, 2003

[11] O. Nasraoui, E. Leon, "Anomaly Detection Based on Unsupervised Niche Clustering with Application to Network Intrusion Detection", Congress on Evolutionary Computation(CEC2004), IEEE 2004, pp. 502-508

[12] Yi Chien Chiu, Yuh-Jye Lee, Chien-Chung, Chang, Wen-Yang Luo, Hsiu-Chuan Huang, "Semi-supervised Learning for False Alarm Reduction", P. Perner (Ed.): ICDM 2010, LNAI 6171, Springer-Verlag Berlin Heidelberg 2010, pp. 595–605

[13] M. Schonlau, W. DuMouchel, W. H. Ju, A. F. Karr, M. Theus, Y. Vardi, "Computer intrusion: Detecting masquerades", Statistical Science pp. 58–74, 2001

[14] Gao Xiang, Wang Min, "Applying Semi-supervised cluster algorithm for anomaly detection", Third International Symposium on Information Processing, 978-0-7695-4261-4/10, IEEE, 2010

[15] Qiang Wang, Vasileios Megalooikonomou, "A clustering algorithm for intrusion detection", conference on Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security, vol. 5812, pp. 31-38, 2005

[16] D. Song, M. I. Heywood, and A. N. Zincir-Heywwd, "Training genetic program-ming on half a million patterns: An example from anomaly detection", IEEE Trans. Evol. Comput., vol. 9, no. 3, pp. 225-239, 2005.

[17] Ching-Hao Mao, Hahn-Ming Lee, Devi Parikh, Tsuhan Chen, Si-Yu Huang, "Semi-Supervised Co-training and Active Learning based Approach for Multi-view Intrusion Detection", 24th Annual ACM Symposium on Applied Computing, Honolulu, Hawaii, pp. 2042-2048, 2009.

[18] Chien-Yi Chiu, Yuh-Jye Lee, Chien-Chung Chang,Wen-Yang Luo, Hsiu-Chuan Huang, "Semi-supervised Learning for False Alarm Reduction", P. Perner (Ed.): ICDM 2010, LNAI 6171, Springer-Verlag Berlin Heidelberg, pp. 595–605, 2010.

[19] *Jimin Li, Wei Zhang, Kunlun Li, "*A Novel Semi-supervised SVM Based on Tri-training for Intrusion Detection", Journal of Computers, Vol 5, No 4 pp. 638-645, 2010.

## BIOGRAPHY OF AUTHORS

**Mr. V. K. Pachghare** has an experience of 22 years in the teaching field. Presently, he is working as Assistant Professor in Dept. of Computer Engg. & Information Technology, College of Engineering, Pune, India (An Autonomous Institute of Government Of Maharashtra). He worked as a member of Board of Studies, Computer Engineering, Pune University. Presently, he is a member in the Board of Computer Engineering, College of Engineering, Pune. He is author of two books namely "Computer Graphics" & "Cryptography and Information Security". He has 17 research publications in various international journals and conferences.

**Mr. V. K. Khatavkar** is working in the Department of Comp. Engg. and IT, College of Engineering, Pune, India (An Autonomous Institute of Government Of Maharashtra), since June 2010. His fields of research are network security and machine learning. He has published 5 research papers in various international journals.

An alumnus of IIT and IIM, **Dr. Parag Kulkarni** completed his Ph.D. in Computer Engineering from IIT Kharagpur. He has been working in IT industry for last 17 years. He has worked as Research head, operations head, GM, Director and was instrumental in building worldclass software product companies.He is founder Director and Chief Scientist at EKLaT research. His name and profile is selected for listing in "Marquis Who's Who in the world" (Science and Engineering) –2009. He has written many business articles. He has more than 60 International publications and two patents pending in US PTO. He is member of IASTED technical committee, WSEAS working committee, board of studies of two institutes and is guiding 7 Ph.D. students. Parag has conducted more than 25 tutorials on research and business topics at various international conferences He is visiting faulty at IIM Indore. He is pioneer of new management program "Deliverance from Success" for Executives and author of books "Deliverance from Success" and "IT strategy". His areas of research and product development include M-maps, intelligent systems, text mining, image processing, Decision systems, Semi-constrained influence diagrams, forecasting, quantitative analysis, knowledge management, IT strategy, classification, distributed computing, AI and machine learning. He is recipient of several awards including Oriental Foundation Scholarship, Professional Contribution awards. Dr. Parag authored/co-authored more than 100 research papers, he presented more than 30 tutorials across the globe and delivered more than two-dozen keynote and plenary addresses in the area of new paradigms of management, Knowledge Management and Machine Learning.