Performance Analysis of Semi-supervised Intrusion Detection System

V. K. Pachghare
Assistant Professor,
Department of Comp. Engg. & IT
College of Engineering,
Pune-5

Vaibhav K Khatavkar
Assistant Professor,
Department of Comp. Engg. & IT
College of Engineering,
Pune-5

Dr. Parag Kulkarni
Adjunct Professor
Department of Comp.. Engg. & IT
College of Engineering,
Pune-5

ABSTRACT
Supervised learning algorithm for Intrusion Detection needs labeled data for training. Lots of data is available through internet, network and host. But this data is unlabeled data. The availability of labeled data needs human expertise which is costly. This is the main hurdle for developing supervised intrusion detection systems. We can intelligently use both labeled and unlabeled data for intrusion detection. Semi-supervised learning has attracted the attention of the researcher working in Intrusion Detection using machine learning. Our goal is to improve the classification accuracy of any given supervised classifier algorithm by using the limited labeled data and large unlabeled data. The key advantage of the proposed semi-supervised learning approach is to improve the performance of supervised classifier. The results show that the performance of the proposed semi-supervised algorithm is better than the state-of the- art supervised learning algorithms. We compare the performance of our DS-AdaBoost algorithm as well as 5 standard algorithms available in WEKA for supervised and semi-supervised approach.

Keywords
Intrusion Detection, supervised learning, semi-supervised learning, pattern recognition

1. INTRODUCTION
Any set of actions that attempt to compromise the integrity, confidentiality, or availability of a system is defined as intrusion [1]. The system which is used to detect an intrusion is known as intrusion detection system (IDS), which involves detecting unusual patterns or patterns of activity that are known to have some relation with intrusions. IDS can broadly classify into two main types: anomaly and misuse detection. According to the difference in monitoring objects, IDSs are divided into network-based IDSs (NIDS) and host-based IDSs (HIDS).

IDS can be implemented using: unsupervised, supervised and semi-supervised machine learning algorithms. Unsupervised learning use unlabeled data. This method can detect the intrusions that have not been previously learned. Examples of unsupervised learning for intrusion detection include K-means-based approaches and self-organizing map (SOM)-based approaches. In supervised learning for intrusion detection, the labeled data is needed for training. There are mainly neural network (NN)-based approaches, and support vector machine (SVM)-based approaches for IDS. The third method is semi-supervised learning in which both the labeled and unlabeled data is used for training.

In this paper we propose semi supervised approach for IDS. Semi-supervised learning approach can leverage unlabeled data in addition to labeled data. They have received significant attention of the researcher, and are more suiTable for intrusion detection because they require a small quantity of labeled data while still taking advantage of the large quantities of unlabeled data.

The proposed method also offers the advantage of not requiring a separate method to label the data. Instead of that this suggests that we should choose only the top few most confident data points. This 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 other semi-supervised learning approaches. It also improves the performance of other supervised classifiers.

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
Semi-supervised learning methods use both labeled and unlabeled data for training.

These algorithms can be classified as: generative, discriminative, or a combination of both. The oldest semi-supervised learning model is generative model. The joint probabilities of data and their labels are modeled in this model. The second model is Discriminative models, which restricts themselves from determining the most likely class for a given data by estimating the probability of each class given the data. They do not model the classes, so generation of new class examples is difficult.

There are many semi-supervised learning methods. The initial work in semi-supervised learning is done by H. J. Scudders in

his work on "self-learning" in 1965. An earlier work in sequential learning by Robbins and Monro can also be viewed as related to semi-supervised learning [20]. Some often-used methods include: EM with generative mixture models, self-training, co-training, transductive support vector machines, and graph-based methods. The first bootstrapping algorithm to become widely known in computational linguistics was the Yarowsky algorithm [21]. An alternative algorithm, co-training [4], has subsequently become more popular. Perhaps in part, because it has proven amenable to theoretical analysis, in contrast to the Yarowsky algorithm, which is as yet mathematically poorly understood.

3. ARCHITECTURE

It is important to distinguish the problem of semi-supervised improvement from the existing supervised classification approaches. In the semi-supervised improvement problem, we aim to build a 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.

Figure 1 shows the architecture we used for the semi-supervised approach for intrusion detection system. We use the labeled data for training the system as supervised approach. After training we test the system using unlabeled data. We will have to add the tested data to the training data so as to implement semi-supervised approach.

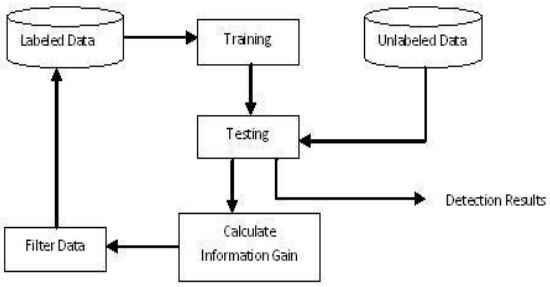

Figure 1: Architecture for proposed Semi-supervised IDS

Test data is comparatively large than train data. It is bad idea to add whole test data to train data. So we will have to select particular data from test data to add into train data. Our classifier uses its own predictions to teach itself. The predictor logic is based on entropy. The entropy can be calculated as:

Entropy is a measure of the average information content. The entropy for data D given the p_i the probability of i^{th} feature, can be calculated as shown in Equation 1.

$$E(D) = -\sum_{i=1}^{n} p_i\,(d_i)\log_2\big(p_i\,(d_i)\big) \qquad (1)$$

Now, a data set D with n points arranged in a frequency distribution with k classes. The class mark of the i^{th} class is denoted x_i; the frequency of the i^{th} class is denoted f_i and the relative frequency of the i^{th} class is denoted $p_i = \frac{f_i}{n}$. Next step is to calculate entropy for each class of data, namely, normal, DOS, R2L, U2R, and PROBE in the data set using Equation 1. Then filter the data from each class in the range is calculated by using mean, standard deviation and variance given in Equations 2, 4 and 3 respectively.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} f_i * x_i = \sum_{i=1}^{n} p_i * x_i \qquad (2)$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} f_i * (x_i - \mu)^2 = \sum_{i=1}^{n} p_i * (x_i - \mu)^2 \qquad (3)$$

$$\sigma = \sqrt{\sigma^2} \qquad (4)$$

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:

1. Train the system with supervised approach.
2. Give unlabelled data for testing.
3. Calculate entropy of test data in the dataset after testing.
4. Using statistical methods filter the data.
5. Add this filtered data to the training data

Train the system with this new data. After testing our approach we have the conclusion that the filtered data is not more than 6% of the actual unlabeled data.

The selected instances from the test data are added in the original data set.

4. EXPERIMENTATIONS AND RESULTS

Experiments are carried on two data sets, i.e. KDDCup99 [2] and NSL data set [22]. For comparison the performance of our supervised and semi-supervised DS-Adaboost algorithm, the other standard algorithms are used. We first discuss the results of semi-supervised Ds-Adaboost algorithm on the same data set. Next we run other standard algorithms such as SMO, AdaBoost-M, Bagging, J48, Naïve Bayes using WEKA [23] on the same data set and compare their results with our algorithm.

4.1 Experiments on KDDCup99 data set

Table 1 shows the number of samples in the training data set. When the semi-supervised algorithm is applied on the training data set, the results were obtained as shown in Table 2 with confusion matrix.

Table 1: Training Data set

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

Table 2: Performance of semi-supervised algorithm in training data set

	Normal	DOS	R2L	U2R	PROBE	%
Normal	108167	19	09	00	32	**99.94**
DOS	21	391423	02	04	18	**99.98**
R2L	14	00	2881	04	04	**99.24**
U2R	02	00	00	47	04	**88.67**
PROBE	40	11	09	00	6877	**99.13**
%	**99.92**	**99.99**	**99.31**	**85.45**	**99.16**	**99.96**

Table 3: Detection Results in Training Data Set

Training Set	
FPR (%)	**DR (%)**
0.055	**99.96**

We can find the performance of the semi-supervised method is better than those of supervised learning method. Semi-supervised method using self training is introduced to improve purely supervised methods for intrusion detection. Experiment is presented the comparison between these two methods.

Table 4: Comparison of Supervised and Semi-supervised Method

Algorithm	Detection Rate		False Positive Rate	
	Supervised	Semi-supervised	Supervised	Semi-supervised
Adaboost M1	97.6424	98.0134	0.092	0.0834
Naïve Bayes	92.1764	92.748	0.1	0.081
J48	99.67	99.8527	0.073	0.067
Bagging	99.6494	99.68	0.079	0.072
SMO	99.52	99.57	0.083	0.08
DS-AdaBoost	99.70	99.96	0.06	0.055

Experiments of comparing this method with other traditional supervised learning methods are presented here. Results show that the performance is much better than those of the other supervised learning methods on detecting attacks. Finally, after analyzing the results of experiments carefully and deeply, we propose a potential learning method— self training semi-supervised learning, which may be more adept in detecting attacks with the aid of accumulated training examples of known attacks.

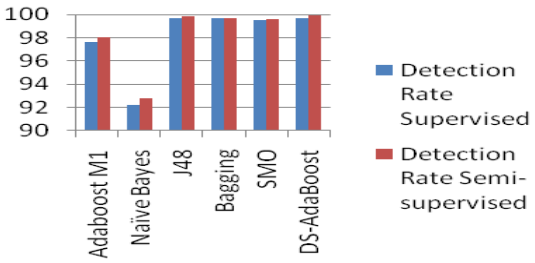

Figure 2: Detection Rate Comparison

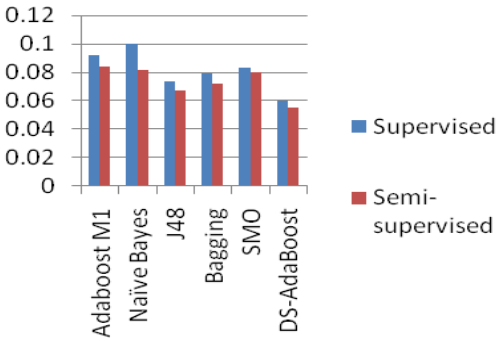

Figure 3: False Positive Rate Comparison

4.2 Experiments on NSL data set

We perform the same experiments on NSL data set and we have the following results for our algorithm with supervised approach and other standard algorithms as shown in Table 5.

Table 5: Comparison of FPR and DR of Various algorithms

Algorithm	False Positive Rate	Detection Rate
AdaboostM1	0.097	94.6838
Naïve Bayes	0.499	90.5933
J48	0.15	99.7245
Bagging	0.27	99.7899
SMO	0.29	97.4574
DS-AdaBoost	0.089	99.8467

Above results are shown in Figure 4 below. The detection rate of DS-Adaboost algorithm is better than other standard algorithms. The graph in Figure 5 shows the comparison of false positive rate of DS-Adaboost with other standard algorithms. The graph shows that DS-Adaboost give better false positive rate as compare to other standard algorithms.

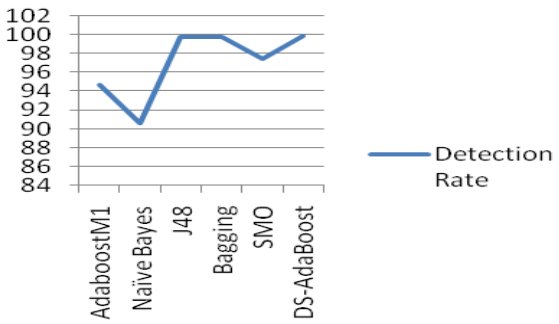

Figure 4: Detection Rate Comparison

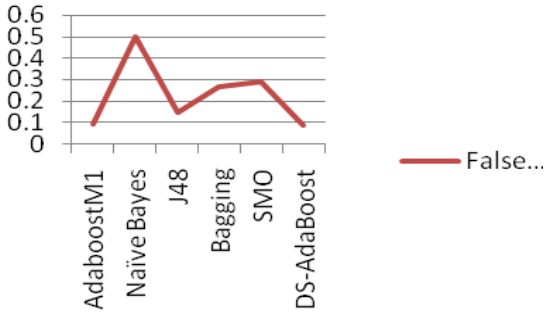

Figure 5: False Positive Rate Comparison

Comparison of Performance for both Data sets

The results of DS-Adaboost with NSL and KDD data sets for supervised and semi-supervised for DS-Adaboost algorithm is shown below in Table9.

Table 9: Performances of NSL and KDD data set

Training	NSL		KDD	
	Supervised Method	Semi-supervised Method	Supervised Method	Semi-supervised Method
Total Packets	125973	141235	494020	509588
Normal	67360	78200	97218	108244
DOS	45912	45922	391413	391453
U2R	54	56	45	55
R2L	993	2673	1102	2901
PROBE	11654	14384	4047	6935
DR	99.8467	99.8633	99.7	99.96
FPR	0.089	0.07674	0.06	0.055

The graphical comparison of detection rate of two methods for both the data sets is shown in Figure 6. It shows that for both the data set the result of semi-supervised method is better than supervised approach.

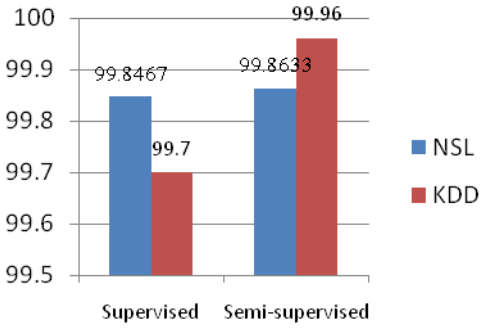

Figure 6: Detection Rate Comparison for NSL and KDD data set

The graphical comparison of false positive rate of two methods for both the data sets i.e. NSL and KDDCup 99 is shown in graph. It shows that for both the data set the false positive rate of semi-supervised method is low as compare to supervised approach.

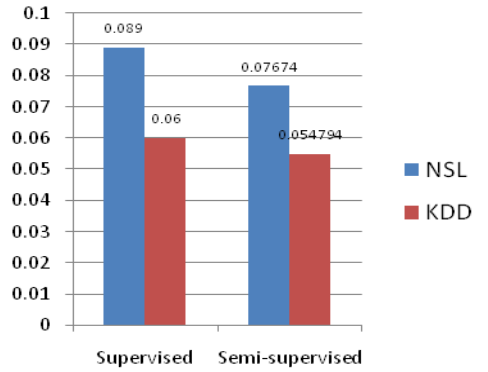

Figure 7: False Positive Rate Comparison for NSL and KDD data set

From the above results we conclude that DS-Adaboost algorithm gives better performance than standard algorithms tested for the same data set. The second conclusion is that self-training semi-supervised approach is better choice for pattern based network security than supervised approach.

5. CONCLUSION

We use supervised classifier as a black box for our semi-supervised algorithm. The results show that the performance of our proposed semi-supervised approach is better than our supervised algorithm DS-AdaBoost algorithm. Then we compare the standard algorithms available in WEKA and compare the supervised and semisupervised results. The experiments show that the performance of semi-supervised algorithm is better than supervised algorithm. Second conclusion is that our DS-Adaboost gives better detection rate and low false positive rate than SMO, AdaBoost-M, J48, Bagging and Naive Bayes algorithms.

6. REFERENCES

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

[2] Zissman, M. 1998/99 DARPA Intrusion Detection Evaluation datasets. MIT Lincoln Laboratory, URL: http://www.ll.mit.edu/IST/ideval/data/data_index.html.

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

[4] Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. *COLT: Proceedings of the Workshop on Computational Learning Theory*.

[5] Vapnik, V. (1998). *Statistical learning theory*. Wiley-Interscience.

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

[7] Xiaojin Zhu. Semi-Supervised Learning Literature Survey

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

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

[10] [Kemp, C., Griffiths, T., Stromsten, S., & Tenenbaum, J. (2003). Semi-supervised learning with trees. *Advances in Neural Information Processing System 16*.

[11] Nasraoui O. and Leon E., "Anomaly Detection Based on Unsupervised Niche Clustering with Application to Network Intrusion Detection" Proceedings of the 2004 Congress on Evolutionary Computation(CEC2004), IEEE press, Jun.2004, pp. 502-508. doi:10.1109/CEC.2004.1330898

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

[13] Schonlau, M., DuMouchel, W., Ju, W.H., Karr, A.F., Theus, M., Vardi, Y.: Computer intrusion: Detecting masquerades. Statistical Science 16 (2001) 58–74

[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 $26.00 © 2010 IEEE

[15] Qiang Wang Vasileios Megalooikonomou, "A Clustering Algorithm for Intrusion Detection",

[16] Andrew H. Sung & Srinivas Mukkamala, "Feature Selection for Intrusion Detection using Neural Networks and Support Vector Machines", TRB 2003 Annual Meeting CD-ROM

[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. Proceedings of the 2009 ACM symposium on Applied Computing SAC 09 (2009) : 2042-2048

[18] Mallapragada, P. K., Jin, R., Jain, A. K., and Liu, Y. (2007). Semiboost: Boosting for semi-supervised learning. Technical report, Department of Comp. Science and Engineering, Michigan State University

[19] Yusufovna, S.F. Integrating Intrusion Detection System and Data Mining, International Symposium on Ubiquitous Multimedia Computing , 2008

[20] Pavan Kumar Mallapragada et al., "SemiBoost: Boosting for Semi-supervised Learning", IEEE Transactions on Pattern Analysis and Machine Intelligence

[21] Yarowsky, David. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, pages 189–196.

[22] NSL data set for IDS www.iscx.ca/**NSL**-KDD/

[23] Weka a Data mining tool www.cs.waikato.ac.nz/ml/**weka**