Discriminative Multinomial Naïve Bayes for Network Intrusion Detection

Mrutyunjaya Panda
Department of AE&IE
Gandhi Institute of Engg. and Tech.
Gunupur-765022, India
e-mail: mrutyunjaya74@gmail.com

Ajith Abraham
Machine Intelligence Research
Labs (MIR Labs), USA
e-mail: ajith.abraham@ieee.org

Manas Ranjan Patra
Department of Computer Science
Berhampur University, India
e-mail: mrpatra12@gmail.com

Abstract— **This paper applies discriminative multinomial Naïve Bayes with various filtering analysis in order to build a network intrusion detection system. For our experimental analysis, we used the new NSL-KDD dataset, which is considered as a modified dataset for KDDCup 1999 intrusion detection benchmark dataset. We perform 2 class classifications with 10-fold cross validation for building our proposed model. The experimental results show that the proposed approach is very accurate with low false positive rate and takes less time in comparison to other existing approaches while building an efficient network intrusion detection system.**

Keywords- **Intrusion detection, Discriminative parameter learning, DMNB, filtered classifier, NSL-KDD dataset, Accuracy.**

I. INTRODUCTION

Intrusion detection systems (IDS) are becoming an important part of today's network security architectures, where it analyzes the network traffic and looks for potential threats. Traditionally, intrusion detection techniques fall into two categories: Signature detection and anomaly detection. Signature or misuse detection searches for well known patterns of attacks, called attack signatures while anomaly detection is based on establishing a normal activity profile for a system.

Intrusion detection is broadly considered to be a classification problem. The main issue in standard classification problems lies in minimizing the probability of error while making the classification decision. Hence, the key point is to how to choose an effective classification approach to build accurate intrusion detection systems in terms of high detection rate while keeping a low false alarm rate. A number of intelligent paradigms have been proposed for building a network intrusion detection model that includes Support Vector Machines (SVM) [1], ensemble voting system [2], Junction tree algorithm [3], and many others [4, 5, 6, 24, 25] .

The purpose of this paper is to address some of the issues in most commonly used KDDCup 1999 dataset. We used NSL-KDD dataset [7], a modified KDDCup 1999 intrusion detection benchmark dataset for our experimentation. The proposed work that combines discriminative parameter learning using Naïve Bayes (DMNB) classifier with principal component analysis (PCA) as a filtering approach; produces better classification accuracy with other existing approaches. We have performed 2 class (attack or normal) classifications for our proposed research work.

The outline of the paper is as follows: A review of the currently available literature in the filed of intrusion detection system is provided in Section 2. Section 3 introduces technical analysis of the various machine learning approaches, followed by the proposed methodology in Section 4. Section 5 addresses some issues that are encountered in KDDCup 1999 dataset and provides an insight into the NSL-KDD dataset. Experimental results and discussions are provided in Section 6. Finally, we conclude the paper in Section 7.

II. RELATED RESEARCH

Mukkamala et al. [8] demonstrated the use of genetic programming approach for building an efficient network intrusion detection system. In [9], the authors propose various feature reduction techniques in order to build a network intrusion detection model in terms of detection accuracy and computation time. Network intrusion detection using Naïve Bayes classifiers is proposed in [10]. In [11], the authors use Bayesian belief network with genetic local search for intrusion detection. An evolutionary support vector machine for intrusion detection is proposed in [12]. In this, the authors have combined evolutionary programming into support vector machines. They concluded that their model is able to detect new attacks as well as experienced attacks. A hybrid statistical approach which includes data mining and decision tree classification is used in [13]. Authors used decision tree and rule based classifiers for the performance comparisons in terms of accuracy and false alarm rate. A hybrid DTNB approach is used in [14] by combining Decision Table (DT) with Naïve Bayes (NB)

to design an efficient intrusion detection model. The authors used Neuro-Fuzzy techniques (NEFCLASS) and JRip classifier to reduce false alerts in [15]. They take SNORT alerts as input and learning from training in order to achieve their goal. All the above papers used the KDDCup 1999 dataset for their experimentation. So, keeping in view the drawbacks of this dataset, as mentioned in [7], we propose to use NSTL-KDD dataset to build our intrusion detection system using novel discriminative parameter learning with Naïve Bayes.

III. Technical Analysis of Existing Data Mining Approaches

Data Mining is considered to be a new approach for detecting network intrusions. In this, the raw data is first converted into ASCII network packet information which in turn is converted to connection level information like duration, service, flag, etc. Data Mining approaches are then applied to the intrusion data to create efficient network intrusion detection model. Naïve Bayes is also called as Idiot's Bayes, Simple Bayes and independent Bayes, which is considered to be important for its simplicity, elegance and robustness. Naïve Bayes is an efficient and effective classification algorithm which assumes that all attributes are independent given the class (conditional independence assumption). But the attribute conditional assumption of Naïve Bayes rarely holds in real world applications. So, it needs to relax the assumption effectively to improve its classification performance. Decision trees construct easily interpretable models, which is useful for a system administrator to inspect and edit. It can also be used effectively in large data set which makes them useful in real time applications. The ability to detect the unseen or new attacks is made possible due to the generalization property of the decision trees. Complex decision trees can be difficult to understand, for instance because of information about one class is usually distributed throughout the tree. The rules are of the form "if A and B and C and …. Then class Z", where rules for each class are grouped together. A case is classified by finding the first rule whose conditions are satisfied by the case; if no rule is satisfied then the case is assigned to a default class. Support vector machines (SVM) are learning machines which is based on intuitive geometric principles that linearly separates the training data so that minimum expected risk is achieved. More details about this can be found in [16].

Ensemble of classifiers can often better than any individual classifier, which makes them an efficient approach in detecting network intrusions. The AdaBoost algorithm proposed by Y. Freund and R. Schapire is one of the most important ensemble methods, since it has solid theoretical foundation, very accurate prediction, great simplicity and very wide and successful applications. More details about the ensemble approach can be found in [17].

IV. Proposed Methodology

We used a Meta filtered classifier approach, where class for running an arbitrary classifier on data that has been passed through an arbitrary filter. Like the classifier, the structure of the filter is based exclusively on the training data and test instances will be processed by the filter without changing their structure. This approach is shown in Figure 1. Here, we use various supervised and unsupervised filter like Principal component analysis (PCA), Random Projection (RP) and Nominal to Binary (N2B) for attribute selection along with some base classifier in order to build an intrusion detection system.

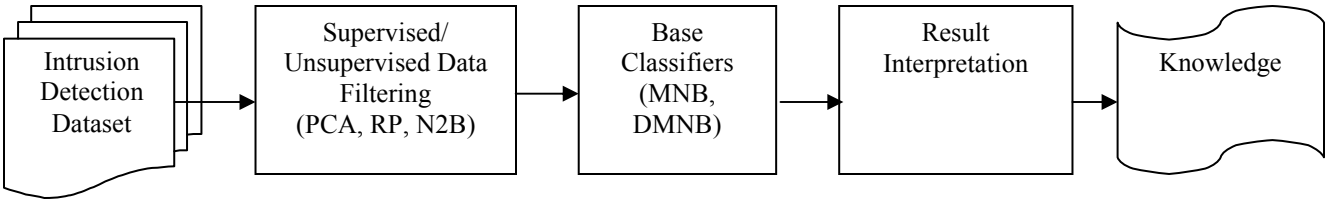

Figure 1. Proposed Meta Filtered Classifier using Discriminative Parameter Learning

In this paper, we propose a discriminative multinomial Naïve Bayes classifiers, which uses a simple, efficient, and effective discriminative parameter learning method. The discriminative parameter learning method learns parameters by discriminatively computing frequencies from intrusion data. Empirical studies show that the discriminative parameter learning, sometimes referred to as Discriminative Frequency Estimate (DFE) integrates the advantages of both generative and discriminative learning techniques. Bayesian networks are very often considered as stable classifiers in detecting network intrusions. While classifying data, Naïve Bayesian learning is performed by using frequency estimate (FE) that determines parameters by computing the appropriate

frequencies from data. The major advantage of FE is its frequency i.e. it only needs to count each training instance

Algorithm: DMNB

1. Initialize each word frequency entry F_{ic} to 0.
2. For t from 1 to M Do
   - Randomly draw a training document d^t from the training data set T.
   - Estimate the probabilities parameters using $\hat{P}(w_i \mid c) = \frac{f_{ic}}{f_c}$, where f_{ic} is the number of occurrences of w_i in documents of class c, and f_c is the total number of word occurrences in documents of class c.
   - Estimate the current frequencies $f_{ic}^{\ t}$.
   - Compute the posterior probability $\hat{P}(c \mid d^t)$.
   - Compute the loss $L(d^t)$ using $L(d) = P(c \mid d) - \hat{P}(c \mid d)$
   - For each non-zero word w_i in the document d^t
     - Let $f_{ic}^{\ t}$ is the frequency of the word w_i in the t^{th} document d^t
     - Let $f_{ic}^{\ t+1} = f_{ic}^{\ t} + L(d^t) * f_i^t$

Figure 2. Pseudo Code for Discriminative Multinomial Naïve Bayes

once. Further, It is well known that FE maximizes likelihood and thus is referred to as a typical generative learning method. Another method, called ELR [18] outperforms the generative learning method FE. But, the application of ELR is limited because of its high computational cost [19]. However, discriminative parameter learning which combines both generative and discriminative learning maximizes generalization accuracy may be an obvious choice for building an efficient network intrusion detection system. The pseudo code for the proposed DMNB is shown below in Figure 2.The more details about this can be obtained from [20].

V. DATA SET DESCRIPTION

During last decade, KDDCup 1999 intrusion detection benchmark dataset [21] is used by many researchers in order to build an efficient network intrusion detection system. However, recent study shows that there are some inherent problems present in KDDCup 1999 dataset [7].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. This new dataset, NSL-KDD dataset provided in [7] is used for our experimentation and is now publicly available for research in intrusion detection. It is also stated that though the NSL-KDD dataset still suffers from some of the problems discussed in [22] and may not be a perfect representative of existing real networks, it can be applied an effective benchmark dataset to detect network intrusions. More details about the inherent problems found in KDDCup dataset can be obtained from [7]. 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%

VI. EXPERIMENTAL RESULTS

In this Section, we provide the results obtained from our experimentation along with the discussion thereto. Here, we perform 2-class classification since most of the anomaly detection systems work with binary levels, i.e. anomalous or normal, rather than identifying the detailed information of the attacks. In this, we use NSL-KDD full training dataset that contain 25192 instances with 42 attributes which consists of 41 attributes or features with one target value or labeled class either normal or attack in order to build a network intrusion detection system. Then, we performed 10 fold cross validation to test the efficacy of the model built during the training phase. We perform all the experiments in a Pentium-IV CPU, with 512 MB RAM.

The results along with the comparison to other existing methods using NSL-KDD Dataset are shown in Table 3.

Table 3: Detection Accuracy Comparison of Machine learning algorithms using NSL-KDD Dataset

Classifier	Detection Accuracy (%)	Time Taken to Build the Model in seconds	False Alarm Rate in %
Decision Trees (J48) [7]	81.05	**	**
Naïve Bayes[7]	76.56	**	**
Random forest[7]	80.67	**	**
SVM[7]	69.52	**	**
AdaBoost [23]	90.31	**	3.38
Multinomial Naïve Bayes + N2B (ours)	38.89	0.72	27.8
Multinomial Naïve Bayes updateable + N2B (ours)	38.94	1.2	27.9
Discriminative Multinomial Naïve Bayes +PCA (*Proposed method*)	94.84	118.36	4.4
Discriminative Multinomial Naïve Bayes +RP (*Proposed method*)	81.47	2.27	12.85
Discriminative Multinomial Naïve Bayes +N2B (*Proposed method*)	**96.5**	**1.11**	**3.0**

** indicates data not provided by the author in their paper.

Inspired from the results obtained above for DMNB using Nominal to binary supervised filtering approach, we further perform some other measures like, number of iterations and time taken to build the model to prove the effectiveness of the proposed classifier in detecting network intrusions, which are shown in Figure 3 and Figure 4 respectively.

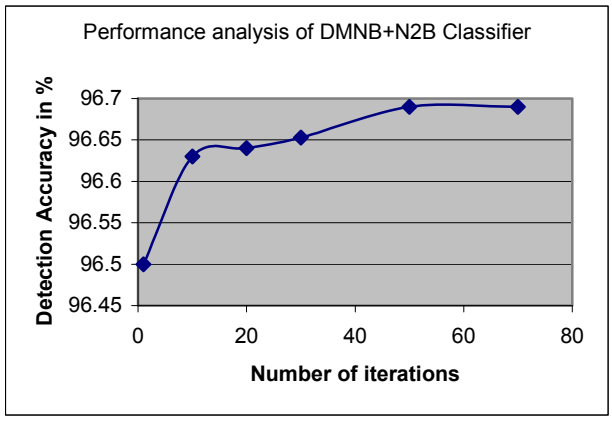

Figure 3. Accuracy Vs Number of iteration for the proposed classifier

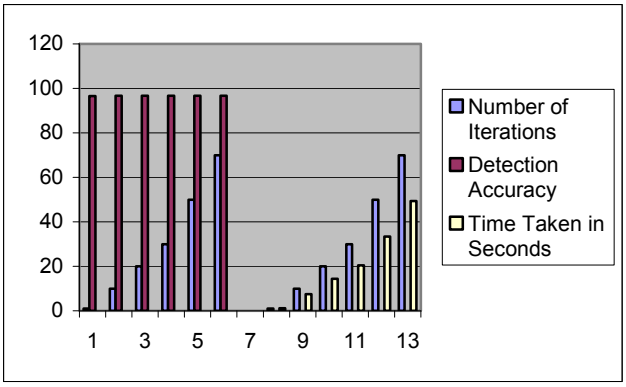

Figure 4. Accuracy comparison with number of iteration and time taken to build a classifier model

From results provided in Table 3, it can be observed that our proposed Meta classifier using DMNB as base classifier with Nominal to binary supervised filtering approach performs better than all other methods like Naïve Bayes, Decision Trees, SVM, Random Forest and AdaBoost, in terms of False alarm rate and accuracy. It is also faster as it takes only 1.11 seconds to build the proposed model to detect network intrusions. Further, it is also observed that as the number of iteration increases, the detection accuracy increases till number of iterations are 70, after which no improvement in accuracy is observed.

VII. CONCLUSIONS

In this, we proposed a novel filtered Meta classifier approach with DMNB as base classifier and Nominal to binary supervised filtering in order to build an efficient network intrusion detection system. We use a new-dataset NSL-KDD for our experiments with some highlights on the limitation of the most commonly used KDDCup 1999dataset. The results obtained are encouraging as the proposed approach is faster and accurate in comparison to other existing approaches.

REFERENCES

1. S. Mukkamala, A. H. sung, and Ajith Abraham. Intrusion detection using an ensemble of intelligent paradigms. Journal of Network and Computer Applications. Vol.28, pp. 167-182, 2005.
2. M. Panda and M.R.Patra . Ensemble voting system for anomaly based network intrusion detection. International journal of recent trends in engineering, Vol.2, No. 5, pp. 8-13 , 2009.
3. E. Nikolva and V. Jecheva. Anomaly based intrusion detection based on Junction tree algorithm. Journal of Information assurance and security. Vol.2, pp. 184-188, 2007.
4. S. Peddabachigari, A. Abraham, C. Grosan, and J. Thomas. Modeling intrusion detection system using hybrid intelligent systems. Journal of network and computer applications. Vol. 30, pp. 114-132, 2007.
5. M. Panda and M.R.Patra. Ensemble rule based classifiers for network intrusion detection. In Proceedings of International conference on advances in recent technologies in communication and computing (ARTCom), India, pp. 19-22, IEEE Computer Society Press, USA.
6. S. Chebrolu, A. Abraham, and J.P.Thomas. Feature deduction and ensemble design of intrusion detection system. International journal of computers and security. Vol.24, No.4, pp.295-307, 2005.
7. M. Tavallaee, E. Bagheri, W. Lu, and Ali. A. Ghorbani. A detailed analysis of the KDDCup 1999 dataset. In: Proc. Of 2009 IEEE International symposium on computational intelligence in security and defense applications (CISDA-2009). IEEE Press, USA.
8. S. Mukkamala, A. H. sung and Ajith Abraham, Modeling intrusion detection system using linear genetic programming approach, LNCS, Vol. 3029, 2004, pp. 633-642, Springer.
9. V. Venkatechalam and S. selvan, Performance comparison of intrusion detection system classification using various feature reduction techniques. International journal of simulation. Vol. 9, No. 1, pp.30-39, 2008.
10. M. Panda and M.R.Patra, Network intrusion detection using Naïve Bayes . International journal of computer science and network security, Vol. 7, No. 12, pp. 258-263, 2007.
11. M.Panda and M.R. Patra, Bayesian belief network using genetic local search for detecting network intrusions. International journal of secure digital information age (IJSDIA), Vol. 1, No. 1, pp. 34-44, 2009.
12. Sung –Hae Jun and Kyung –whan oh, An evolutionary support vector machine for intrusion detection. Asian journal of information technology, Vol. 5, No. 7, pp. 778-783, 2006.
13. N. B. Annur, H.Sallehudin, A. Gani and O.zakari. Identifying false alarm for network intrusion detection system using hybrid data mining decision tree. Malaysian journal of computer science, Vol.21, No. 2, pp.101-115, 2008.
14. M. Panda and M.R. Patra. A semi-Naïve Bayesian method for detecting network intrusions. LNCS, Vol. 5863, pp. 614-621, 2009.
15. P.Gaonjur, N. Z. Tarapore and S.G.Pokale. using neuro-fuzzy techniques to reduce false alerts in intrusion detection. In: Proceedings of International conference on Computer Networks and Security, India, pp. 1-6, 2008. IEEE Press.
16. X. Wu, V. Kumar, J.R.Quinlan et al., Top 10 algorithms in data mining. Knowledge Information system. Vol.14, pp. 1-37, 2008. Springer.
17. F.Y.Schafire. A decision theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. Vol.55, No.1, pp.119-139, 1997.

18. R. Reiner and W. Zhou. Structural extension to logistic regression:discriminative parameter learning of belief net classifiers. AAAI/IAAI, pp. 167-173, 2002.
19. D. Grossman and P. Domingos, Learning Bayesian network classifiers by maximizing conditional likelihood. ICML-2004. p. 46, ACM press,USA.
20. J. Su, H. Zhang, C.X. Ling and S. matwin. Discriminative parameter learning for Bayesian networks. In: proc. Of 25th International conference on machine learning (ICML), Finland, 2008.
21. KDDCup 1999 Dataset. Available at: http:// kdd.ics.uci.edu/databases/kddcup1999.html.
22. J. McHugh. Testing Intrusion detection system: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory, ACM Transaction on Information and system security, Vol. 3, No. 4, pp.262-294, 2000.
23. V.P. Kshirsagar and Dharmaraj R. Patil: Application of Variant of AdaBoost based Machine Learning Algorithm in Network Intrusion Detection. International Journal of Computer Science and Security (IJCSS), Vol. 4, Issue.2, pp. 1-6, 2010.
24. A. Abraham, C. Grosan and C. Martin-Vide, Evolutionary Design of Intrusion Detection Programs, International Journal of Network Security, Vol.4, No.3, pp. 328-339, 2007.
25. A. Abraham, R. Jain, J. Thomas and S.Y. Han, D-SCIDS: Distributed Soft Computing Intrusion Detection Systems, Journal of Network and Computer Applications, Elsevier Science, Volume 30, Issue 1, pp. 81-98, 2007.