ISSN 2277-2685

IJESR/July. 2020/ Vol-10/Issue-3/1-07

Analysis of the Effectiveness of Several Intrusion Detection Methods that Use Supervised Machine Learning

Manjula C. Belavagi

Assoc. Professor

Department of CSE

Manjula C@GMAIL.COM

Abstract

Intrusion detection system plays an important role in network security. Intrusion detection model is a predictive model used to predict the network data traffic as normal or intrusion. Machine Learning algorithms are used to build accurate models for clustering, classification and prediction. In this paper classification and predictive models for intrusion detection are built by using machine learning classification algorithms namely Logistic Regression, Gaussian Naive Bayes, Support Vector Machine and Random Forest. These algorithms are tested with NSL-KDD data set. Experimental results shows that Random Forest Classifier out performs the other methods in identifying whether the data traffic is normal or an attack.

Introduction

In the present era, it is difficult to envision world without internet. Every person has dependency on internet. It has become an important model in various applications such as education, business and others. So security of the data that is communicated through internet is necessary. Secure network is maintained by Intrusion Detection System (IDS). IDS observes the data traffic carefully and identifies it as normal or spam. Nowadays most of the applications depends on the advance network technologies namely wireless networks, wireless sensor networks and bluetooth. In case of wireless sensor networks security mechanisms such as key-management protocols, authentication techniques and security protocols cannot be used because of resource constraints. Intrusion Detection System is the ideal security mechanism for wireless sensor networks.

1.1 Intrusion Detection System

A security mechanism used to monitor the abnormal behavior of the network is an Intrusion Detection System (IDS)12. The IDS identifies and informs that whether the user activity is normal or not. The users activities are compared by the IDS with the already stored intrusion records to identify the intrusion. Accurate predictive models can be built for large data sets using supervised machine learning techniques, that is not possible by traditional methods.

As specified by Tom Mitchell3, machine learning based intrusion detection falls under two categories Anomaly and Misuse.
IDS learns the patterns by the training data, so the misuse based method is used. Misuse based detection can detect only the known attack, new attacks cannot be identified. Anomaly based IDS observes the normal behavior and if there is a change in the behavior then it considers that behavior as anomaly. So anomaly based IDS can detect new attacks that are not learned from the training model.
Till now different machine learning techniques such as Artificial neural networks7, Support Vector Machine4 and Naive Bayes5,6, based techniques are proposed for the intrusion detection. A new detection by combining different techniques, a hybrid detection technique is proposed by8. The literature on comparison of supervised machine learning techniques in intrusion detection is limited. Hence this paper aims at understanding the implications of using supervised machine learning techniques on intrusion detection.
The paper is organized as follows. Section 2 discusses some research work on Intrusion Detection System. In Section 3 various supervised machine learning methods used are discussed. Section 4 gives brief introduction about the data-set used. Section 5 discusses the methodology. Conclusion and future scope is given in section 7.

Related Work

Recently Yousef *et al.*[9] used algorithms namely Random Forest, Naive Bayes, K-means and Support Vector Machine to identify four types of attacks. They also proposed best feature selection method. Concluded that the Random Forest Classifier (RFC) outperforms the other methods. They have mentioned that hierarchical clustering method can be used to improve the performance.

Nadiammai *et al.*[10] proposed semi supervised machine learning based intrusion detection. Authors have not considered the resource consumption. Combination of different classifiers to identify the intrusion is proposed by Panda *et al.*[8]. They used supervised classification or unsupervised clustering for filtering of the data. They used NSL-KDD data-set and tested with decision tree classifier. But the proposed method works only for binary class classification.

Sangkatsanee *et al.*[11] proposed intrusion detection system using supervised machine learing techniques to identify the on line network data as normal or not. The proposed method identifies probe and Denial of Service attacks only, but the other attacks are not considered.

A framework of machine learning approach is proposed by Yu *et al.*[12] and Campos *et al.*[13]. Intrusion is identified by analyzing the local features. Levent *et al.*[14] proposed Naive Bayes based multiclass classifier to identify the intrusions. They suggested that intrusion detection is possible by Hidden Naive Bayes (HNB) model. Denial of Service attacks are identified with good accuracy compared to other attacks.

Li *et al.* proposed[15] Intrusion detection technique using Support Vector Machine (SVM). They also used feature removal method to improve the efficiency. Using the proposed feature removal method they selected best nineteen features from the KDD-CUP99 data-set. In the proposed method the data set used is very small. A light weight IDS is proposed by Sivatha Sindhu *et al.*[16]. The proposed method mainly focused on pre-processing of the data so that only important attributes can be used. The first step is to remove the redundant data so that the learning algorithms give the unbiased result.

A survey on intrusion detection systems was conducted by Butan *et al.*[17] Information about IDSs such as classification, Intrusion type, computing location and infrastructure are discussed. They discussed about the Mobile Adhoc Networks (MANET) IDS. They compared MANETIDS and the Wireless Sensor Networks (WSN) IDS. Authors suggested that for mobile applications distributed and cooperative IDS schemes are suitable. For stationary applications centralized IDSs are suitable and for cluster based applications hierarchical IDSs are suitable. Farooqi *et al.*[18] proposed intrusion detection framework to detect routing attacks. Specification based approach is used to detect routing attacks. Authors claim that the proposed method has low False Positive Rate (FPR) and good intrusion detection rate. The proposed method works only for static networks. Wang *et al.*[19] developed IDS for Sink, Cluster Head (CH) and for a Sensor Node (SN) separately and combined altogether to identify the intrusion in heterogeneous Cluster Based Wireless Sensor Networks (CWSN) but the detection rate for U2R, R2L and Probe attacks is very low.

Supervised Machine Learning Techniques

Our work is to design a network intrusion detection system with the different supervised machine learning classifiers. This paper is to investigate the performance of the classifiers namely Logistic Regression, Support Vector Machine (SVM), Gaussian Naive Bayes (GNB) and Random Forest in intrusion detection. These classifiers are discussed below.

1.2 Logistic regression

To solve the classification problems Logistic Regression (LR) is used. TLR works for both binary classification and multiclass classification. Probability of occurrence of an event is predicted by fitting data to the Logistic function. The values selected by the logistic function[20] is in the range 0 and 1. If the value is 0.5 and above then it is labeled as otherwise 0.

$$h_\theta (x) = g(1/1 + e^{-\theta^T x}) \tag{1}$$

1.3 Support vector machine

Mainly for classification problems Support Vector Machine algorithm is used, but it can be used in regression problems also. N-dimensional feature space is considered to plot each data item as a point with the value of each feature as a particular coordinate. Then classification is made by finding the hyper-plane that differentiate the two classes very well. Support Vectors are the co-ordinates of specific observation that lies closest to the boarder line[21]. In case of SVM training samples are divided into different subsets called as support vectors, the decision function is specified by these support vectors. This paper is based on the liner kernel method of SVM for intrusion detection.

1.4 Gaussian naive bayes

The Gaussian Naive Bayes algorithm is the supervised learning method. Probabilities of each attribute which belongs to each class are considered for a prediction. This algorithm is assumes that the probability of each attribute belonging to a given class value is not depends on all other attributes. If the value of the attribute is known the probability of a class value is called as the conditional probabilities. Data instances provability can be found out by multiplying all attributes conditional probabilities together. Prediction can be made by calculating the each class instance probabilities and by selecting the highest probability class value[21].

$$P(M|N) = P(N|M) * P(M)/P(N) \tag{2}$$

1.5 Random forest classifier

In 2001 Breiman proposed the random forest machine learning classifier. It is a collaborative method which works based on the proximity search. It is decision tree based classifier. It makes use of standard divide and conquer approach to improve the performance. The main principle behind random forest is that strong learner group is created by a group of weak learners22. It is applicable to disjunctive hypothesis.

Comparison of above mentioned classifiers based on the Precision, Recall, F1-Score and Accuracy is discussed in section 6.

2. Intrusion Detection Data-set

The standard intrusion detection data set KDDCUP99[23] has redundant records. This may lead to unfair result of the machine learning algorithms. So the supervised machine learning algorithms are tested NSL-KDD[24] data-set which is the advanced version of the KDDCUP99 intrusion detection data-set. It has 42 features and the four simulated attacks.

Denial of Service (DoS) attack: Over usage of the bandwidth or non availability of the system resources leads to the DoS attacks. Examples: Neptune, Teardrop and Smurf.

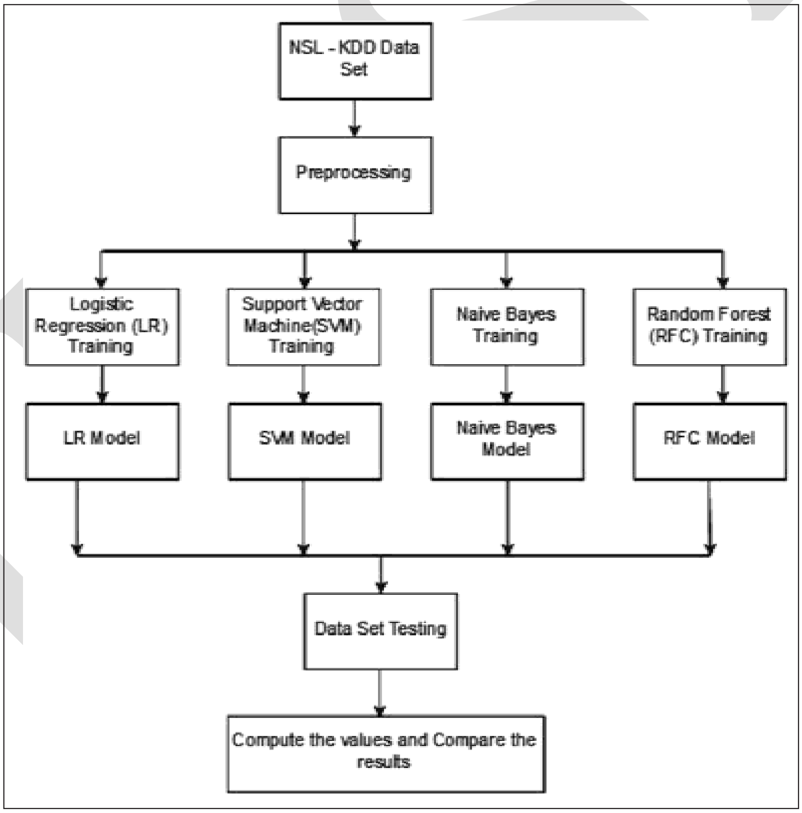
NSL - KDD Data Set
Preprocessing
Logistic Regression (LR) Training
Support Vector Machine(SVM) Training
Naive Bayes Training
Random Forest (RFC) Training
LR Model
SVM Model
Naive Bayes Model
RFC Model
Data Set Testing
Compute the values and Compare the results

Fig. 1. Methodology.

- User to Root (U2R) Attack: Initially attacker access normal user account, later gain access to the root by exploiting the vulnerabilities of the system. Examples: Perl, Load Module and Eject attacks.
- Probe Attack: Have an access to entire network information before introducing an attack. Examples: ipsweep, nmap attacks.
- Root to Local (R2L) Attack: By exploiting some of the vulnerabilities of the network attacker gains local access

by sending packets on a remote machine. Examples: imap, guess password and ftp-write attacks.

3. Methodology

The methodology used is shown in the Fig 1. In pre-processing step all the categorical data which are in textual form are converted to numerical form. Pre-processed data is divided as testing data and training data. The models are built using Logistic Regression, Gaussian Naive Bayes, Support Vector Machine and Random Forest classifiers. These models are used for predicting the labels of the test data. Actual labels and predicted labels are compared. Accuracy, True Positive Rate (TPR) and False Positive Rate (FPR) are computed. Based on these parameters performance of the models are compared.

Following steps are used to build the models.

Pre-process the data set.

The data set is divided as training data and testing data

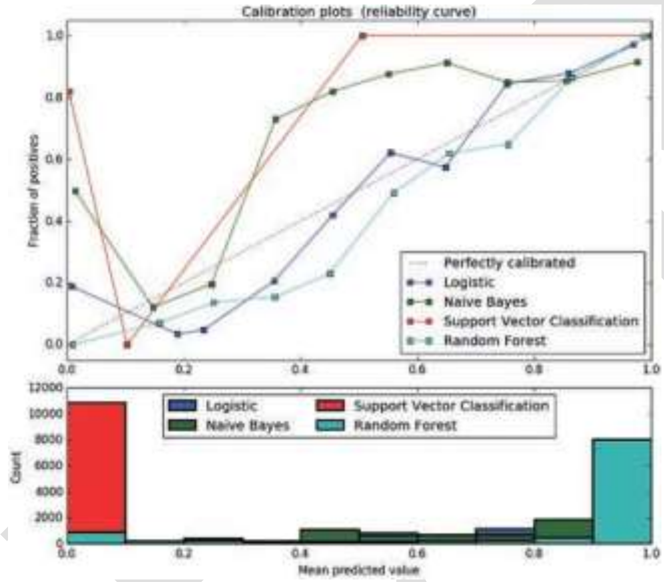

Fig. 2. Calibration Plots.

- Support Vector Machine
- Gaussian Naive Bayes
- Random Forest

2. Read the test data
3. Test the classifier models on training data
4. Compute and compare TPR, FPR, Precision, Recall, F1-Score and Accuracy for all the models.

4. Experimental Results and Discussions

Supervised machine learning algorithms namely Logistic Regression, Gaussian Naive Bayes, Support Vector Machine and Random Forest are tested on NSL-KDD dataset, the new standard intrusion detection data-set. These algorithms are tested on Intel Core (TM) i5-3230M CPU @2.60 GHZ, 4 GB RAM and coding is done by Python[20]. The result of the experiment is represented as a Reliability curve. In Reliability curve estimated probabilities are plotted against the true empirical probabilities. Figure 2 shows the Reliability Curve for the above mentioned supervised machine learning classifiers. Reliability curve for the ideal classifier falls near the diagonal because the estimated probabilities and empirical probabilities are nearly equal.

X-axis probability space is divided into ten bins as shown in Fig. 2. Estimated probabilities values ranging from 0 to 0.1, 0.1 to 0.2 and so on. The values 0 to 0.1 belongs to I bin, 0.1 to 0.2 belongs to II bin and similarly the other ranges. From the graph shown in Fig. 2, it can be concluded that the Random Forest classifier out performs the other methods in identifying the network traffic as normal or an attack. Where as the SVM identifies the intrusion with the lowest probability estimate.

Quality of the classification models is identified by plotting the Receiver Operating Characteristics (ROC) curve. In ROC curve shows FPR verses TPR. ROC curve for the above mentioned classifiers is shown in the Fig. 3. Random Forest has highest TPR. Hence, the ROC curve for Random forest is plotted separately. By observing the graphs, it can be concluded that the Random forest classifier has lowest FPR and highest TPR in identifying attacks. It outperforms the other techniques. Where as Support Vector Machine has highest FPR (39%) and minimal TPR (75%) for intrusion detection. This is due to the fact that too many features from the data set is considered[15] and SVM's linear kernal function is used.

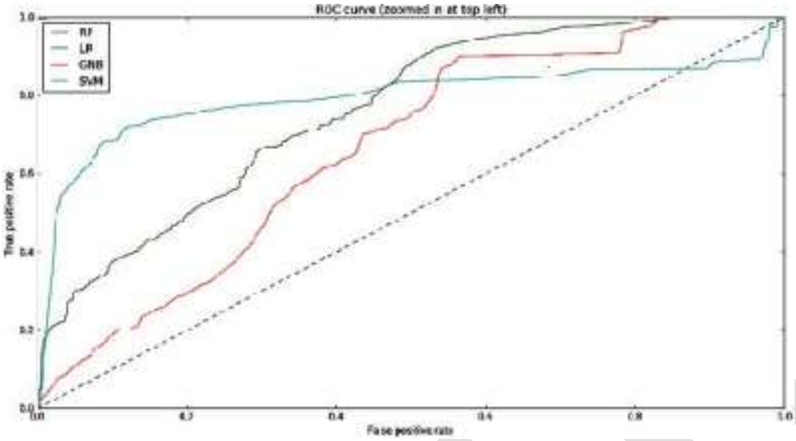

Fig. 3a. Receiver Operating Characteristics (ROC) curve.

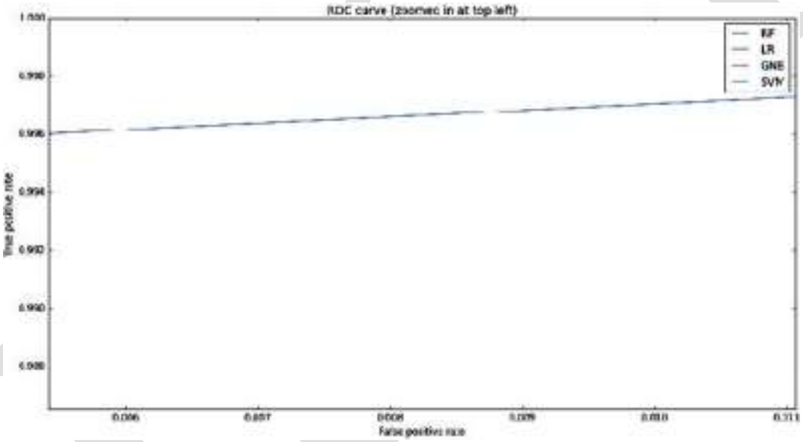

Fig. 3b. ROC for Random Forest Classifier.

Table 1. Performance Measures.

–	Precision	Recall	F1-Score	Accuracy
LR	0.83	0.85	0.82	0.84
GNB	0.79	0.81	0.78	0.79
SVM	0.76	0.79	0.77	0.75
RFC	0.99	0.99	0.99	0.99

Table 1 shows Precision, Recall, F1-Score and Accuracy of the supervised machine learning classifiers in identifying the intrusion. Based on the results shown in the Table 1 it can be identified that Random Forest classifier with the highest accuracy, outperforms the other methods. Whereas SVM has the lowest accuracy, Logistic Regression algorithm has the good accuracy than Gaussian Naive Bayes and SVM.

5. Conclusions and Future Scope

An attempt has been made to check the performance of the supervised machine learning classifiers namely Support Vector Machine, Random Forest, Logistic Regression and Gaussian Naive Bayes are compared for an intrusion detection. These algorithms are tested with the NSL-KDD data-set. Effective classifier is identified by comparing the performances based on the precision, recall, F1-Score and accuracy. From the observed results it can be concluded that the Random forest classifier outperforms other classifiers for the considered data-set and parameters. It has the accuracy of 99%. The work can be extended by considering the classifiers for multiclass classification and considering only the important attributes for the intrusion detection.

References

M. Kemiche and R. Beghdad, Intelligent Systems in Science and Information 2014: Extended and Selected Results from the Science and Information Conference 2014, Cham: Springer International Publishing, ch. Towards Using Games Theory to Detect New U2R Attacks,
pp. 351–367, (2015). [Online]. Available: http://dx.doi.org/10.1007/978-3-319-14654-6-22

S. Patil, D. V. K. B. P, S. Singha and R. Jamil, A Survey on Authentication Techniques for Wireless Sensor Networks, *International Journal of Applied Engineering Research*, vol. 7, (2012).

T. M. Mitchell, Machine Learning, 1st ed, New York, NY, USA: McGraw-Hill, Inc., (1997).

D. S. Kim and J. S. Park, Network-Based Intrusion Detection with Support Vector Machines, Information Networking: International Conference, ICOIN 2003, Cheju Island, Korea, February 12–14, (2003). Revised Selected Papers, Berlin, Heidelberg: Springer Berlin Heidelberg, ch. pp. 747–756, (2003). [Online]. Available: http://dx.doi.org/10.1007/978-3-540-45235-5-73

H. Altwaijry and S. Algarny, Bayesian Based Intrusion Detection System, *Journal of King Saud University – Computer and Information Sciences*, vol. 24, no. 1, pp. 1–6, (2012). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1319157811000292

M. Panda and M. R. Patra, Semi-naive Bayesian Method for Network Intrusion Detection System, In *Neural Information Processing, 16th International Conference, ICONIP 2009*, Bangkok, Thailand, December 1–5, 2009, Proceedings, Part I, pp. 614–621, (2009). [Online].
Available: http://dx.doi.org/10.1007/978-3-642-10677-4-70

G. Poojitha, K. N. Kumar and P. J. Reddy, Intrusion Detection Using Artificial Neural Network, In *2010 International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pp. 1–7, July (2010).

M. Panda, A. Abraham and M. R. Patra, A Hybrid Intelligent Approach for Network Intrusion Detection, *Procedia Engineering*, vol. 30, pp. 1–9, (2012), International Conference on Communication Technology and System Design 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877705812008375

A. T. Yousef El Mourabit, Anouar Bouirden and N. E. Moussaidr, Intrusion Detection Techniques in Wireless Sensor Network Using Data Mining Algorithms: Comparative Evaluation Based on Attacks Detection, *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 9, pp. 164–172, (2015).

G. Nadiammai and M. Hemalatha, Effective Approach Toward Intrusion Detection System Using Data Mining Techniques, *Egyptian Informatics Journal*, vol. 15, no. 1, pp. 37–50, (2014). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1110866513000418

P. Sangkatsanee, N. Wattanapongsakorn and C. Charnsripinyo, Practical Real-Time Intrusion Detection Using Machine Learning Approaches, *Computer Communications*, vol. 34, no. 18, pp. 2227–2235, (2011). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S014036641100209X

J. J. T. Zhenwei Yu, A Framework of Machine Learning Based Intrusion Detection for Wireless Sensor Networks, *IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing*, vol. 6, no. 9, pp. 272–279, (2008).

L. M. L. Campos, R. C. L. de Oliveira and M. Roisenberg, Network Intrusion Detection System Using Data Mining, Engineering Applications of Neural Networks: 13th International Conference, EANN 2012, London, UK, September 20–23, 2012. Proceedings, Berlin, Heidelberg:
Springer Berlin Heidelberg, ch. pp. 104–113, (2012). [Online]. Available: http://dx.doi.org/10.1007/978-3-642-32909-8-11

L. Koc, T. A. Mazzuchi and S. Sarkani, A Network Intrusion Detection System Based on a Hidden Naive Bayes Multiclass Classifier, *Expert Systems with Applications*, vol. 39, no. 18, pp. 13 492–13 500, (2012). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417412008640

Y. Li, J. Xia, S. Zhang, J. Yan, X. Ai and K. Dai, An Efficient Intrusion Detection System Based on Support Vector Machines and Gradually Feature Removal Method, *Expert Systems with Applications*, vol. 39, no. 1, pp. 424–430, (2012). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417411009948

S. S. S. Sindhu, S. Geetha and A. Kannan, Decision Tree Based Light Weight Intrusion Detection Using a Wrapper Approach, *Expert Systems with Applications*, vol. 39, no. 1, pp. 129–141, (2012). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417411009080

I. Butun, S. D. Morgera and R. Sankar, A Survey of Intrusion Detection Systems in Wireless Sensor Networks, *Communications Surveys and Tutorials IEEE*, vol. 16, pp. 266–282, (2013).

A. H. Farooqi, F. A. Khan, J. Wang and S. Lee, A Novel Intrusion Detection Framework for Wireless Sensor Networks, *Personal and Ubiquitous Computing*, vol. 17, no. 5, pp. 907–919, (2013).

S.-S. Wang, K.-Q. Yan, S.-C. Wang and C.-W. Liu, An Integrated Intrusion Detection System for Cluster-Based Wireless Sensor Networks,
Expert Syst. Appl., vol. 38, no. 12, pp. 15 234–15 243, (2011).

W. Richert, Building Machine Learning Systems with Python, Packt Publishing, (2013). [Online]. Available: https://books.google.co.in/books?id=C-yglCEcK0sC

K. P. Murphy, Machine Learning: A Probabilistic Perspective, The MIT Press, (2012).

L. Breiman, Random Forests, Machine Learning, vol. 45, no. 1, pp. 5–32, (2001). [Online]. Available: http://www.cs.colorado.edu/ grudic/teaching/CSCI5622_2004/RandomForests-ML-Journal.pdf

Kdd cup, (1999). [Online: Accessed December, 2015].

Nsl-kdd dataset. [Online: Accessed December, 2015].