ABHIYANTRIKI
An International Journal of Engineering & Technology
(A Peer Reviewed & Indexed Journal)
Vol. 3, No. 5 (May, 2016) http://www.aijet.in/ eISSN: 2394-627X

Intrusion Detection System using Effective Feature Selection and Multi Classifier Learning

Veeresh S Pattar*
Deptt. of Computer Science & Engineering
RNS Institute of Technology, Bangalore, India

Shashidhar H R
Deptt. of Computer Science & Engineering
RNS Institute of Technology, Bangalore, India

Abstract

Anomaly based Intrusion Detection Systems (IDS) learn normal and anomalous behavior by analyzing network traffic in various benchmark datasets. Common challenges for IDSs are large amounts of data to process, low detection rates and high rates of false alarms. For performance evaluation of proposed technique the standard NSL-KDD 2009 (Network Security Laboratory-Knowledge Discovery and Data Mining) dataset is used. In existing systems feature selection method is not good, irrelevant and redundant features are present. The classifiers are not accurate for limited training data set. In this project, proposed the solution for both the problems using Entropy based Feature Selection and classifiers are Baysian classifier and Multivariate linear regression. The proposed technique outperforms other published techniques in terms of accuracy, false positive rate and detection time. Based on the experimental results achieved in the project, it shows that the proposed technique is an efficient method for network intrusion detection.

Keywords*: Feature selection, Baysian classifier, MLR classifier, Ensembler.*

****Author for correspondence*** veereshpttr@gmail.com

1. Introduction

Whenever an intrusion occurs, the security and value of a computer system is compromised. Network-based attacks make it difficult for legitimate users to access various network services by purposely occupying or sabotaging network resources and services. This can be done by sending large amounts of network traffic, exploiting well-known faults in networking services, and by overloading network hosts, use Support Vector Machines (SVM) [2] for classification. Clustering analysis helps find the boundary points, which are the most qualified data points to train SVM, between two classes. The SVM is one of the most successful classification algorithms in the data mining area, but it's long training time limits its use. Many applications, such as Data Mining and Bio-Informatics require the processing of huge data sets. The training time of SVM is a serious obstacle in the processing of such data sets. In the proposed system a new formula for feature selection based on Entropy of the features is designed. The Entropy is better indicator for the relation between the input variables and output classification result (intrusion or no intrusion). So applying this method, this approach can get the best features which are relevant and not redundant. For improving the accuracy of the classifier for limited

data set, proposed a method called Baysian Classifier [5]. Learn and use MLR (multivariate Linear Regression)[6] based on input data set and then use the MLR model learnt to generate further dataset and trained the classifier. By this way classifier can be trained well and accuracy is improved.

2. Related Work

IDS can be classified by signature, anomaly or hybrid technique. There are many soft computing techniques which have been used to detect intrusions. These techniques along with various issues are discussed here. Anomalies are detected by analyzing sparse region of this feature space. Genetic clustering based intrusion detection automatically creates normal and abnormal clusters and effectively detects intrusions. This unsupervised technique operates in two phases. In network, data is grouped by taking nearest neighbour. Second it obtains near optimal detection rate by genetic optimization [2]. Artificial Neural Network (ANN) and Support Vector Machine (SVM) are also used in IDS. Two encoding methods simple frequency-based and term frequency*inverse document frequency (tf×idf) scheme are used to detect possible intrusions [3]. Patterns of intrusions are built by IDS using random forest of training instances. After learning these patterns, intrusions are detected by the outlier detection algorithm. This hybrid approach improves the detection rate of IDS [4]. AdaBoost can be used to increase the efficiency of IDS. In this technique decision stump is used as a weak classifier. The decision rules are defined for both categorical and continuous features. Both types of features are combined to form a strong classifier. Concept drifting data streams is used in adaptive ensemble approach for classification. This model is updated automatically by traditional mining classification. Three classification algorithms namely Expectation–Maximization (EM), C4.5 and K-Nearest Neighbour (KNN) are employed to test performance of the system [5]. Hybrid Intelligent Intrusion Detection and Prevention System (IIDPS) are used to detect intrusions in early stage. Known attacks are identified by the signature based approach but others are detected using anomaly based approach. The support vector machine (SVM) with three types of kernel (Linear, polynomial and RBF) is used to detect unknown attacks in [6]. IDS faces problem to process huge network traffic dataset. Due to this learning is slow and detection time of intrusions increases. The dataset is also highly imbalanced. The difficulty to clearly separate normal and anomalous behaviour decreases accuracy and increase false positive rate [7].

3. Methodology

System architecture is the conceptual design that defines the structure and behaviour of a system. Here the training data sets are taken from KDD. This project mainly contains 3 modules namely,

I. Feature Selection
II. Multi Stage Classifier Modelling: It has 2 sub modules including:
    i. Naive Bayes Modeling
    ii. MLR (Multivariate Linear Regression) modeling
III. Intrusion Detection

Feature selection: The Data sets are taken from the KNN-KDD 1999 where it has many features which are nothing but attributes of data packets and value associated with them. So this module extracts features that are required to implement this project. This feature extraction process is done by entropy based selection where it extracts features based on the entropy algorithm. Entropy is a measure of the amount of uncertainty associated with a set of probabilities.

Multistage classifier: The features extracted are sent to the multistage classifier where it has two classifiers called Naive Bayes Modelling, MLR Modelling. Which are used to calculate the correlation between the attributes and if the correlation between the attributes or feature and the required value are same then that particular feature belongs to specific class of attack? This way the features are classified.

Naive Bayes modelling: Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 10cm in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the colour, roundness and diameter features.

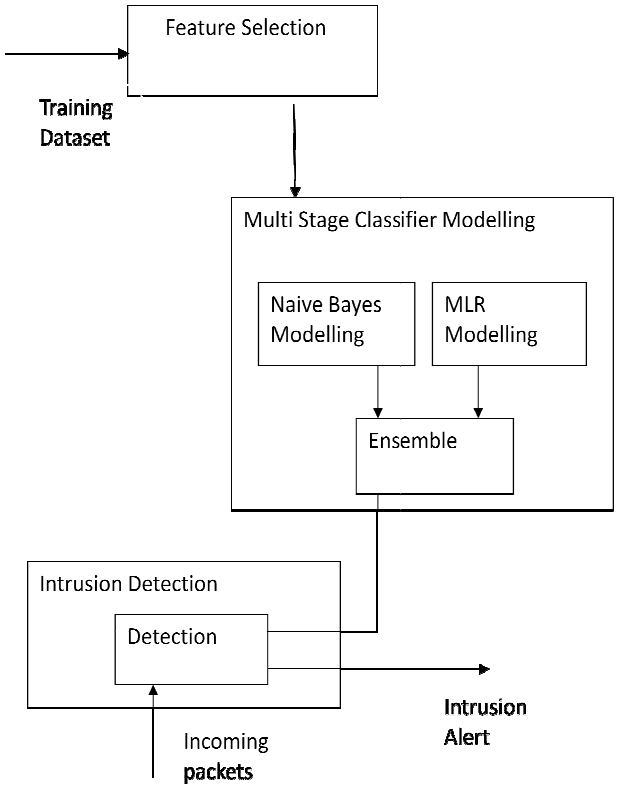

Fig.: System Architecture

Multivariate linear regression: A multivariable model can be thought of as a model in which multiple variables are found on the right side of the model equation. This type of statistical model can be used to attempt to assess the relationship between numbers of variables; one can assess independent relationships while adjusting for potential confounders. A simple linear regression model has a continuous outcome and one predictor, whereas a multiple or multivariable linear regression model has a continuous outcome and multiple predictors (continuous or categorical). A simple linear regression model would have the form

$$Y = \alpha + X*\beta + €\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

By contrast, a multivariable or multiple linear regression models would take the form

$$Y = \alpha + X1*\beta1 + X2*\beta2 + \ldots.. + Xk + \beta k + €\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Where y is a continuous dependent variable, x is a single predictor in the simple regression model, and x1, x2, …, xk are the predictors in the multivariable model.

Intrusion Detection: This module detects the intrusion. Features of the newly arrived packet are extracted and then compared with the features already existing in the system. So if the features

matches with high probability then it will be considered as the one of the attack and alert message is sent.

4. Conclusion

In the proposed system a new formula is used for feature selection based on Entropy of the features. The Entropy is better indicator for the relation between the input variables and output classification result (intrusion or no intrusion). So applying this method, we can get the best features which are relevant and not redundant. For improving the accuracy of the classifier for limited data set, a method called Baysian Classifier and Multivariate Linear Regression. Use the MLR model learnt to generate further dataset and train the classifier. By this way classifier can be trained well and accuracy is improved.

References

[1] Jacek Biesiada, W_odzis_aw Duch, Adam Kachel, Krystian M¸aczka, Sebastian Pa_ucha. International Conference on Research in Electrotechnology and Applied Informatics. August 31, September 3, 2005, Katowice, Poland

[2] S. Revathi. International Journal of Engineering Research & Technology. 2(12). 2013.

[3] Chai, K. HT Hn, & HL Chieu. Bayesian Online Classifiers for Text Classification and Filtering. Proceedings of the 25th annual international ACM SIGIR conference on Research and Development in Information Retrieval, August 2002, pp 97-104.

[4] Swati Paliwal. International Journal of Computer Applications. 60(19). 2012.

[5] Iofciu, P. Fankhauser, F. Abel, and K. Bischoff. Securing Cloud Computing Environment against DDoS Attacks. Proc. IEEE Int'l Conf. Computer Comm. and Informatics (ICCCI '12), Jan. 2012.

[6] Yang JM, Chen YF, Shen TW, Kristal BS, Hsu DF. Consensus scoring criteria for improving enrichment in virtual screening. *J Chem Inf Model*. 45(4), pp. 1134-46. 2005.

[7] International Journal of Computer Applications. 3(3), June, 2010.