IJNRD.ORG ISSN : 2456-4184

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG

An International Open Access, Peer-reviewed, Refereed Journal

A Study of Intrusion Detection System (IDS) through Machine Learning Algorithm

Preeti Gupta

Assistant Professor, Department of Computer Science and Engineering
School of Computing Science and Engineering
Galgotias University, Uttar Pradesh

Abstract

Due to the advancement of internet, the number of attacks over internet has also increased. To ensure the security of a network a good Intrusion Detection System (IDS) is required. The aim of IDS is to monitor the processes prevailing in a network and to analyse them for signs of any possible deviations. Some studies have been done in this field but a deep and exhaustive work has still not been done. This paper proposes an IDS using machine leaning for network with a good union of feature selection technique and classifier by studying the combinations of most of the popular feature selection techniques and classifiers. A set of significant features is selected from the original set of features using feature selection techniques and then the set of significant features is used to train different types of classifiers to make the IDS. Intrusion detection is performed on the dataset which contain Test data and Training data. It is finally observed that K-NN classifier produces better performance than others and, among the feature selection methods, information gain ratio based feature selection method is better.

Key Words: ID, machine learning, NSL-KDD dataset, feature selection.

1 Introduction

It is one of the challenges, to make the internet stable and secure due to the large amount of data and information stored at internet. Although security can be ensured by updating firewall and software, dynamic mechanisms can also be exploited. Intrusion detection system is one of dynamic mechanisms along with network analysers and others. Intrusion detection determines specific goal of detecting attacks. Intrusion detection system is a system that monitors and analysis data to detect any intrusion in the system or network. The large volume, variety of data generated in network made data analysis to detect attacks very difficult. There are two methods of intrusion detection: misuse and anomaly. Misuse aims to determine attack signatures in the monitored resource. Anomaly depends on knowledge of normal behaviour and any deviation from this. Anomaly detection has gained popularity as it became effective against new attacks. Machine Learning algorithms can also be used for anomaly detection. Machine learning algorithms are trained and then be applied on unseen input for the actual detection process. In machine learning, there are many classification algorithms hat can be trained and used to detect attacks on a network. To further improve the performance of these classifiers and reduce the recognition time, feature reduction algorithms can be used. Various algorithms have been developed to detect different types of network intrusions. Fig. 1.1 shows the proposed IDS model.

The feature selection method selects the significant features for classification. A dataset containing only the significant features increases the acceptance of the model with better accuracy. The set of significant features is then used for training the classifiers. After training the classifier, the test dataset is tested with the same features to determine if each instance is normal data or attack data. Sannasi Ganapathy et al. [2] presented a survey on intelligent

techniques for Intrusion Detection (ID) by feature selection and classification techniques, which includes many statistical and machine learning algorithms that are used as classifiers or feature selection techniques.

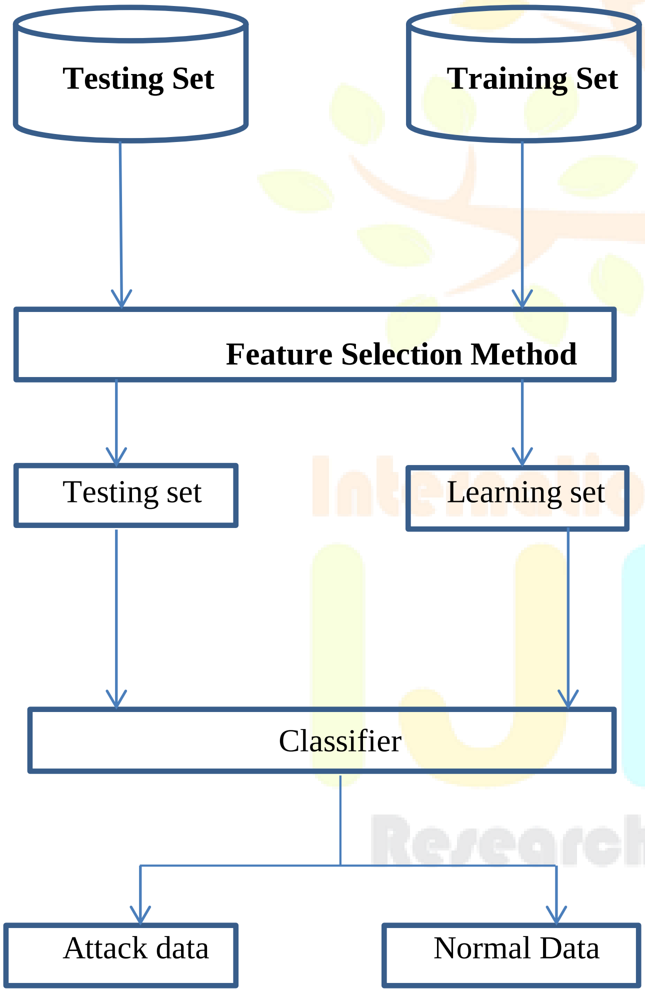


Fig. 1.1 Intrusion Detection System[1]

2 Feature Selection and classifiers

2.1 Feature selection

Feature selection selects a representative feature set from the original feature sets. This representative set contains only the relevant elements. An important feature is that the learning algorithm takes less time to learn and creates a more general classifier as it removes unnecessary elements. A strange property of the original set. Feature selection also benefits data visualization and understanding. How to choose popular features briefly used in this article. They are presented below.

a) **Correlation based Feature Selection method**: CFS works with hypothesis that is ”Good feature subsets contain features highly correlated with the class, yet uncorrelated to each other”.
Algorithm: [3]
1) Select the dataset for pre-processing.
2) Calculate feature and feature-class correlations.
3) Search through the feature subspace and calculate feature subset based on merit.

b) **Principal Component Analysis:** Principal component analysis (PCA) determines uncorrelated attributes called principal components. [4]
Algorithm:
1. Whole d-dimensional dataset is taken ignoring the class labels.
2. The d-dimensional mean vector is calculated.
3. Covariance matrix is found for the whole data set.

c) **Minimum Redundancy Maximum Relevance**: This method tries to penalize a feature’s relevance based on its redundancy. The relevance of a feature set S for the class c is defined by the average value of all mutual information values between the individual feature fi and the class c.[5]

2.2 Classifiers

A classifier is an algorithm, or the set of rules that computers use to categorise data. On the other hand, the outcome of your classifier's machine learning is a classification model. The classifier is used to train the model, which eventually uses the classifier to classify your data. Some of the popular classification methods used in this research article is defined below:

a) **Nave Bayes:** A group of supervised learning algorithms known as naive Bayes methods utilise Bayes' theorem with the "naive" assumption that each pair of features is conditionally independent given the value of the class variable. It is based on Bayes's theory of probability.

b) **Support Vector Machine:** Its foundation is the concept of structural risk minimization, which benefits from speed and scalability. Finding a decision boundary in multidimensional space that divides hidden patterns into distinct groups is the fundamental goal of the Support Vector Machine (SVM).

c) **Decision Tree:** The non-parametric supervised learning approach used for classification and regression applications is the decision tree. It is organised hierarchically and has a root node, branches, internal nodes, and leaf nodes.

d) **Neural Network:** Deep learning techniques are based on neural networks, sometimes referred to as artificial neural networks (ANNs) or simulated neural networks (SNNs), which are a subset of machine learning. They take their cues from the way biological neurons communicate with one another, emulating the name and structure of the human brain.

e) **K-nearest neighbor (KNN) algorithm:** The k-nearest Neighbours' algorithm: what is it? The k-nearest neighbours' algorithm, sometimes referred to as KNN or k-NN, is a supervised learning classifier that employs proximity to produce classifications or predictions about the grouping of a single data point.

3 Experimental Results

Kaggle data set is used for the experiments. The size of kaggle data set is very large and nearly 22000 rows and 41 columns. It is not easy to work with large datasets so the dataset is reduced to meet the requirement. There are no null values in our dataset. There are two classes in our dataset, normal and anomaly, in which 13447 data are normal and 11743 are anomaly. Three feature selection techniques and three classifier are used in this article for intrusion detection. All the experiments are performed on python and power of all the classifiers is classifying the dataset. The performance of different classifiers in detection of Train score and Test Score dataset are shown in table.

Table 1.1 Train and Test score of three Classifiers

Model	Train Score	Test Score
KNN	0.981626	0.981477
Logistic Regression	0.928774	0.923128
Decision Tree	1	0.995766

We also calculate the Model Validation based on its precision and Mean recall.

KNeighborsClassifier Model Validation

Mean precision:
98.45 % +- 0.48

Mean recall:
98.24 % +- 0.54

Logistic Regression Model Validation

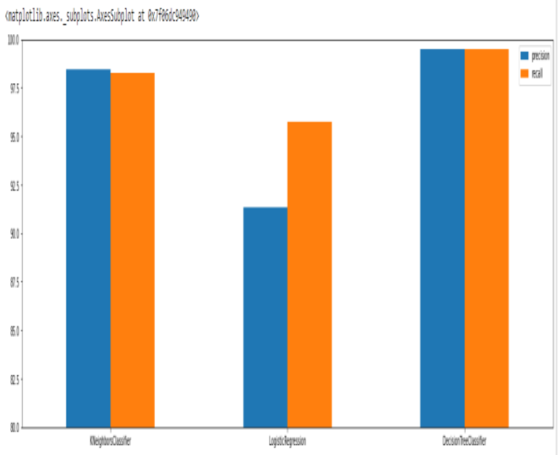
Mean precision:
91.35 % +- 0.57

Mean recall:
95.72 % +- 0.67

Decision Tree Classifier Model Validation

Mean precision:
99.51 % +- 0.19

Mean recall:
99.49 % +- 0.24



```
<matplotlib.axes._subplots.AxesSubplot at 0x7f06dc949490>
```

KNeighborsClassifier Model Testing

[3435   63]

[65 3995]

	precision	recall	f1-score	support
normal	0.98	0.98	0.98	3498
Anamoly	0.98	0.98	0.98	4060
Accuracy			0.98	7558
macro avg	0.98	0.98	0.98	7558
weighted avg	0.98	0.98	0.98	7558

Logistic Regression Model Testing

[3127   371]

[210 3850]

	precision	recall	f1-score	support
normal	0.94	0.89	0.91	3498
Anamoly	0.91	0.95	0.93	4060
Accuracy			0.92	7558
macro avg	0.92	0.92	0.92	7558
weighted avg	0.92	0.92	0.92	7558

The following observations are made from the results.

1. In Decision tree, both precision and recall is equal. Thus decision tree predicting more true value.
2. In logistic regression, precision is slightly less than both decision tree and KNN.
3. KNN is performing good, but not as good as Decision tree.
4. For KNN, 3435 true values and true negative is 63, i.e. some 63 values are not predicting false. 65 values are predicting wrong class in KNN.
5. For logistic regression, true positive values are 3127 and true negative values are 371.

4 Conclusions

This paper proposes an IDS model which compares performances of different combinations. A subset of classifier is used to test the normal and anomaly data. Decision tree, Logistic regression and KNN classifier. In addition experimental results indicate that machine learning can be used in intrusion detection since all combination gives significant accuracy.

References

[1] Amor, N. B., Benferhat, S., and Elouedi, Z.: NB vs DTs in Intrusion Detection Systems. proceedings of ACM Symposium on Applied Computing, pp. 420-424,(2004).

[2] Ganapathy, S., Kulothungan, K., Muthurajkumar, S., Vijayalakshmi, M., Yogesh, P., and Kannan, A.:Intelligent feature selection and classification techniques for intrusion detection in networks: a survey. Journal on Wireless Communications and Networking, pp. 1-16, (2013).

[3] Jalill, K. A., Kamarudin, M. H., and Masrek, M.N.: Comparison of Machine Learning Algorithms Performance in Detecting Network Intrusion. International Conference on Networking and Information Technology, pp. 221-226, (2010).

[4] Karegowda, A., Manjunath, A. S., and Jayaram, M. A.:Comparative study of attribute selection using gain ratio and correlation based feature selection. International Journal of Information Technology and Knowledge Management, vol. 2, issue 2, pp. 271-277,(2010).

[5] Peng, H. C., Long, F., and Ding, C.: Feature selection based on mutual information: criteria of max-dependency, maxrelevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27,issue 8, pp. 1226- 1238, (2005).

[6] Reddy, R.R., Kavya, B., and Ramadevi, Y.:A Survey on SVM Classifiers for Intrusion Detection. International Journal of Computer

Applications (0975 8887) ,vol. 98, issue 19, (July 2014).

[7] Sommer, R., and Paxson, V.:Outside the Closed World: On Using Machine Learning For Network Intrusion Detection. IEEE Symposium on Security and Privacy, pp. 305-316, (2010).

[8] Wagh, S. K., Pachghare, V. K., and Kolhe, S. R.: Survey on Intrusion Detection System using Machine Learning Techniques. International Journal of Computer Applications (0975 8887), vol. 78, issue 16, (September 2013).

[9] Vinchurkar, D. P., and Reshamwala, A.: A Review of Intrusion Detection System Using NN and Machine Learning Technique. International Journal of Engineering Science and Innovative Technology, vol. 1, issue 2,(November 2012).

[10] Scarfone, K., and Mell, P.:Guide to Intrusion Detection and Prevention Systems (IDPS). National Institute Of Standards and Technology. Special Publication February-2007.

[11] Balakrishnan, S., and Kannan, V.K.:Intrusion Detection System Using Feature Selection and Classification Technique. International Journal of Computer Science and Application (IJCSA), vol. 3, issue 4, (2014).

[12] Gnes Kayack, H., Nur Zincir-Heywood, A., and Heywood, M. I.: Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets. Third Annual Conference on Privacy, Security and Trust, (2005).

[13] Mukkamala, S., Janoski, G., and Sung, A.:Intrusion detection using NNs and SVMs. IJCNN, vol. 2, pp. 1702 1707, (2002).

[14] Osareh, A., Shadgar, B.:Intrusion Detection in Computer Networks based on Machine Learning Algorithms. International Journal of Computer Science and Network Security, vol. 8,(November 2008).