ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 8, August 2019

Supervised Learning Techniques for Intrusion Detection System

Kalidindi Venkateswara Rao[1], Dr. G. Sharmila Sujatha[2]

M. Tech Student, Department of Computer Science and Systems Engineering, Andhra University, Visakhapatnam, Andhra Pradesh, India[1]

Assistant Professor, Department of Computer Science and Systems Engineering, Andhra University, Visakhapatnam, Andhra Pradesh, India[2]

ABSTRACT*:* In the previous decades, the fast development of Intrusion Detection and Prevention systems played a crucial role in computers network and security. Intrusion detection system (IDS) is one of the implemented solutions against harmful attacks. Furthermore, attackers always keep changing their tools and techniques. However, implementing an accepted IDS system is also a challenging task. In this paper, several experiments have been performed and evaluated to assess various supervised learning classifiers based on KDD intrusion dataset. It succeeded to compute several performance metrics in order to evaluate the selected classifiers. The focus was on false negative and false positive performance metrics in order to enhance the detection rate of the intrusion detection system. The implemented experiments demonstrated that the decision table classifier achieved the lowest value of false negative while the random forest classifier has achieved the highest average accuracy rate.

KEYWORDS*:* Intrusion detection, KDD dataset, Supervised learning classifiers, Performance metrics.

I. INTRODUCTION

There are many types of attacks threatening the availability, integrity and confidentiality of computer networks. The Denial of service attack (DOS) considered as one of the most common harmful attacks. A network can be protected against such attacks using an intrusion detection system. An IDS system can detect intrusions and intrusion de-generates an alert when it detects an intrusion. This intrusion detection system in a network analyses all traffic. For large datacentre's this is a difficult task. There's an enormous amount of data through the network of a datacentre. Standard intrusion systems cannot then all traffic completely.

Intrusion detection system (IDS) have been introduced as a tool designed to enhance the security of systems [1]. Various IDS approaches have been proposed in the literature since inceptions, but two of them are proposed by Steniford at al., and Denning are most relevant in this context. Denning's proposal for an intrusion detection system focused on how to develop effective and accurate methods for intrusion detection.

In general, there are two types of IDS (anomaly base or misuse base). Anomaly intrusion detection system implemented to detect attacks based on recorded normal behaviour. Therefore, it compares the current real time traffics with previous recorded normal real time traffics, this type of intrusion detection system is widely used because it has the ability to detect the new type of intrusions. But from another perspective, it registers the largest values of false positive alarm, which means there is a large number of normal packets considered as attacks packets. However, misuse intrusion detection system is implemented to detect attacks based on repository of attacks signatures. It has no false alarm but at the same time, the new type of attack (new signature) can succeed to pass-through it.

Attacks detection considered as classification problem because the target is to clarify whether the packet either normal or attack packet. Therefore, the model of accepted intrusion detection system can be implemented based on

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 8, August 2019

significant machine learning algorithms. In this paper, the following implemented the machine learning algorithms have been Implemented (Random Forest, Decision Tree, Linear Discriminant Analysis, Logistic Regression, and K-Nearest Neighbour) to evaluate and accurate the model of intrusion detection system based on a bench market dataset Knowledge Discovery in Databases (KDD) which includes the following types of attacks (DOS, R2L, U2R, and PROBE).

II. RELATED WORK

In general, this paper classifies IDSs on the basis of detection methods they employ into two categories, like (i) misuse detection and (ii) abnormality detection. By matching observed data, misuse detection identifies intrusions with pre-defined descriptions of intrusive behaviour. So prominent intrusions can be detected in an efficient manner utilizing a low false positive rate. Therefore, this technique is widely adopted in the majority of commercial systems. However, the types of new intrusions have evolved every moment and continuously, therefore. Misuse the previous techniques for intrusion detection will fail to detect new unknown intrusions. The only way to get rid over this issue is to learn from all intrusions and get update the data knowledge at every moment. This updating process can either be manual or automatic, the manual process might be very time consuming and also the human intervention is required at every moment of time. This process can work automatically using supervised machine learning techniques. Unfortunately, the preparation of datasets for training the supervised learning algorithms is very difficult and expensive, as this require collection and labelling of each event as normal or an intrusion type.

This section presents the related works relevant to using KDD dataset for implementing machine learning algorithms. It also provides a brief overview of the different machine learning algorithms and shows how the KDD dataset is very useful for evaluating and testing various types of machine learning algorithms. As per model proposed authors imported the KDD dataset and implemented the pre-process phases e.g. normalization of the attributes range to [-1, 1] and converting symbolic attributes. Neural network feed forward was implemented in two experiments. The authors have concluded that neural network is not suitable enough for R2L and U2R attacks but on the other hand, it was recorded acceptable accuracy rate for DOS and PROBE attacks. As it relates to implement neural network against KDD intrusions, the effort of [8] the authors succeeded to implement the following four algorithms: Fuzzy ARTMAP, Radial-based Function, Back propagation (BP) and Perceptron-back propagation-hybrid (PBH). The four algorithms evaluated and tested for intrusions detection the BP and PBH algorithms recorded highest accuracy rate.

From another study, some of the authors focus on attributes selection algorithms in order to reduce the cost of computation time. In [9] the authors are focused on selecting the most significant attributes to design IDS that have a high accuracy rate with low computation time. 10% of KDD was used for training and testing. They implemented detection system based on extended classifier system and neural network to reduce false positive alarm as much as possible. On the other hand in the information gain algorithm was implemented as one of effective attributes selection. They implemented multivariate method as linear machine method to detect the denial of service intrusions.

III. SUPERVISED LEARNING TECHNIQUES

Machine learning algorithms are categorised as Supervised, Unsupervised, Semi-supervised learning and Reinforcement learning. The main aspects of supervised learning is prediction, classification and regression. Supervised can apply what has been learned in the past to new data using labelled examples to predict future events starting from the analysis of known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The classification refers to predict the discrete valued output i.e., 0 or 1. Whereas regression predicts continuous valued outputs. In classification we are trying to map input variables into discrete categories and in regression we try to map input variables to some continuous function.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 8, August 2019

Supervised learning categories and techniques

- Linear classifier (numerical functions)
  - Logistic regression, SVM, MLP etc.,
- Parametric (probabilistic functions)
  - Naïve Bayes, Gaussian Discriminant analysis etc.,
- Non-parametric (Instance based functions)
  - K-nearest neighbour, kernel regression etc.,
- Non-metric (symbolic functions)
  - Classification and regression tree, decision tree etc.,
- Aggregation
  - Ada boost, Random forest etc.,

In this paper we select an algorithm from each above category and build a model for the prediction and accuracy, precision and other metrics. The selected algorithms are as follows.

Random Tree Classifier: is one of tree classifiers using this classifying the number of trees should be fixed before implementing. Each individual tree represents a single decision tree. Each individual tree has randomly selected attributes from dataset. Therefore, the random tree classifier could be considered as a finite group of decision trees. The procedure of predicting the entire dataset is to migrate several decision trees outputs and choose the winner expected class based on total numbers of votes.

K - Nearest Neighbor (kNN) is a very simple, easy to understand, versatile and one of the topmost machine learning algorithms. KNN used in the variety of applications such as finance, healthcare, political science, handwriting detection, image recognition and video recognition. In Credit ratings, financial institutes will predict the credit rating of customers. In loan disbursement, banking institutes will predict whether the loan is safe or risky. In political science, classifying potential voters in two classes will vote or won't vote. kNN algorithm used for both classification and regression problems. kNN algorithm based on feature similarity approach.

Decision Trees algorithm is so easy compared with other classification algorithms. The decision tree algorithm tries to solve the problem, by using tree representation. Each **internal node** of the tree corresponds to an attribute, and each **leaf node** corresponds to a class label.

Linear Discriminant Analysis or LDA is a statistical technique for binary and multiclass classification. It too assumes a Gaussian distribution for the numerical input variables. You can construct an LDA model using the Linear Discriminant Analysis class2 Linear Discriminant Analysis or LDA is a statistical technique for binary and multiclass classification. It too assumes a Gaussian distribution for the numerical input variables. You can construct an LDA model using the Linear Discriminant Analysis class.

Logistic regression assumes a Gaussian distribution for the numeric input variables and can model binary classification problems. You can construct a logistic regression model using the Logistic Regression class.

IV. ABOUT THE DATASET

NSL – KDD intrusion detection dataset is used by many researchers to build an effective network intrusion detection system during past years. But the recent study illustrates there are some limitations are present in the KDD CUP 1999 dataset. It shows that the dataset has 78% training and 75% testing records having redundant records, it inhibits the IDS from detecting rare attacks such as U2R and R2L. The new dataset NSL-KDD dataset is the advanced version of KDD CUP 1999 dataset. Now it is commonly used by the researchers to apply their experiments for analyse the intrusions. The Duplicate records are eliminated to produce an un-biased result. Anming (GEP). The DARPA 1998

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 8, August 2019

dataset was used and 24 attacks can be classified into four types. Adequate numbers of records are available in the train and test datasets to execute the experiments completely. In each record there are 41 attributes are available and a label is present to categorize the data either as normal or an attack type. The attack classes grouped into four types as its predecessor.

- Statistics of redundant records in the KDD train set (Original records | Distinct records | Reduction rate)
  - Attacks: 3,925,650 | 262,178 | 93.32%
  - Normal: 972,781 | 812,814 | 16.44%
  - **Total:** 4,898,431 | 1,074,992 | 78.05%
- Statistics of redundant records in the KDD test set (**Original records | Distinct records | Reduction rate)**
  - Attacks: 250,436 | 29,378 | 88.26%
  - **Normal:** 60,591 | 47,911 | 20.92%
  - **Total:** 311,027 | 77,289 | 75.15%

V. EXPERIMENT RESULTS

Performance metrics:

These metrics are used calculating and observing the performance of the IDS and for comparing the results obtained from the dataset. The performance of the intrusion detection system (IDS) is evaluated by calculating four metric values, Accuracy, Precision, Recall and F-score, out of which accuracy plays a major role and the performance evaluation of the IDS is mainly dependent on accuracy metric.

1. Accuracy: This metric is calculated by finding the total number of instances that are correctly predicted as positive cases to the total number of data that is present, the instances are classified into positive or negative cases by calculating the data that are divided into True Positive(TP), True Negative(TN), False Positive(FP), False Negative(FN).True Positives(TP) are the data which are correctly classified as true instances, True Negatives(TN) are the data which are correctly classified as false instances, False Positives (FP) refers to the data that are negative instances but are predicted as positive and False Negative(FN) refers to the data that are positive instances which are predicted as negative. The accuracy rate at the maximum times can be taken as high though there are less number of negative instances which does not play a major role in decreasing the accuracy rate, it is calculated as: Accuracy =TP+TN/TP+TN+FP+FN

2. Precision: Precision refers to the total data which are correctly predicted to be positive over the total number of data that are predicted to be positive, by observing the false positive and true positive instances, precision can be calculated as: Precision = TP / TP+FP

3. Recall: Recall also known as a True Positive Rate (TPR), sensitivity (SN) or detection rate indicates the total number of instances that are correctly predicted as positive over the total number of actually positive instances present. While detecting the overall positive data in the dataset the recall serves as the main evaluation metric or the best performance indicator of positive data, it is calculated as follows: Recall = TP / FN+TP Precision and Recall are equally important for calculating the performance of the IDS, each individual is not sufficient for the evaluation of the performance of IDS.

4. F-score: F-score is calculated by considering both the metrics of precision and recall equally, the f1 and f2 scores are calculated, in case of f1 both the metrics are treated equally and the value is obtained by substituting 1 in the place of f-beta, in the case of f2 score the recall is considered two times more important than precision.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 8, August 2019

Table. Supervised learning technique and its accuracy

Supervised learning technique	Accuracy
Logistic regression	83.58%
Linear Discriminant Analysis	79.10%
K-Nearest neighbour	83.76%
Decision tree	74.81%
Random forest	84.55%

From the table it was clear that the random forest algorithm had more accuracy when compared to other algorithms where the decision tree algorithm had the least accuracy.

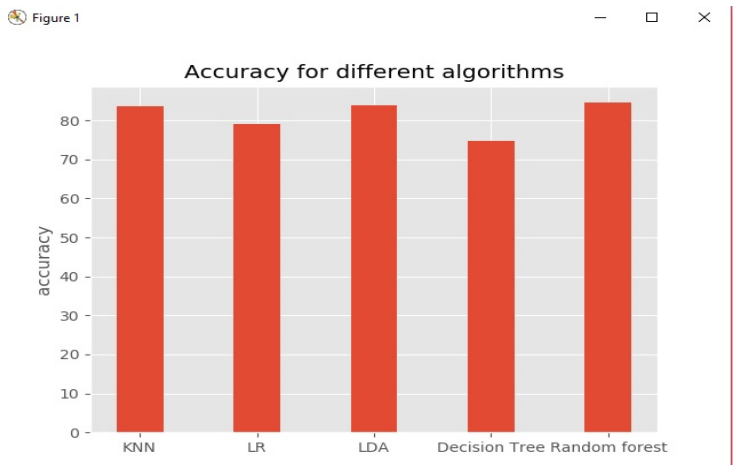

Fig.1 Graph of algorithms and its accuracy.

As in the graph accuracy is taken on one axis and algorithms are on other axis and compared where random forest is shown as the best.

VI. CONCLUSION

The authors have analysed the class specific detection with the KDD dataset, using the supervised machine learning algorithm Random Forest for IDS and the test data and the training data is constructed for evaluating the performance to detect different types of attacks. The training time of the model is observed with the respective increase in the size of the dataset. The increase in the values of evaluation metrics (Accuracy, Precision, Recall, F-score) by increasing the size of the dataset in steps is observed. From the experiment conducted the authors obtained the results with an accuracy of 84%. The criteria included accuracy, complexity, time for training a model, time for classifying a unknown data and understating final solution. It is difficult to identify a better approach of ML method based on only one factor like accuracy.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 8, August 2019

REFERENCES

[1] G. C. Kessler, "Defenses against distributed denial of service attacks," SANS Institute, vol. 2002, 2000.View publication stats.
[2] H. A. Nguyen and D. Choi, "Application of data mining to network intrusion detection: classifier selection model," in Asia-Pacific Network Operations and Management Symposium. Springer, 2008, pp. 399–408.
[3] S. Paliwal and R. Gupta, "Denial-of-service, probing & remote to user (r2l) attack detection using genetic algorithm," International Journal of Computer Applications, vol. 60, no. 19, pp. 57–62, 2012.
[4] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the kdd cup 99 data set," in Computational Intelligence for Security and Defense Applications, 2009. CISDA 2009. IEEE Symposium on. IEEE, 2009, pp. 1–6.
[5] P. Amudha, S. Karthik, and S. Sivakumari, "Classification techniques for intrusion detection-an overview," International Journal of Computer Applications, vol. 76, no. 16, 2013.
[6] M. K. Lahre, M. T. Dhar, D. Suresh, K. Kashyap, and P. Agrawal, "Analyze different approaches for ids using kdd 99 data set," International Journal on Recent and Innovation Trends in Computing and Communication, vol. 1, no. 8, pp. 645–651, 2013.
[7] F. Haddadi, S. Khanchi, M. Shetabi, and V. Derhami, "Intrusion detection and attack classification using feed-forward neural network," in Computer and Network Technology (ICCNT), 2010 Second International Conference on. IEEE, 2010, pp. 262–266.
[8] Z. Zhang, J. Li, C. Manikopoulos, J. Jorgenson, and J. Ucles, "Hide: a hierarchical network intrusion detection system using statistical preprocessing and neural network classification," in Proc. IEEE Workshop on Information Assurance and Security, 2001, pp. 85–90.
[9] W. Alsharafat, "Applying artificial neural network and extended classifier system for network intrusion detection." International Arab Journal of Information Technology (IAJIT), vol. 10, no. 3, 2013.
[10] N. Bhargava, G. Sharma, R. Bhargava, and M. Mathuria, "Decision tree analysis on j48 algorithm for data mining," Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, no. 6, 2013.
[11] C. Fleizach and S. Fukushima, "A naive bayes classifier on 1998 kdd cup," 1998.
[12] M. Alkasassbeh, G. Al-Naymat, A. B. Hassanat, and M. Almseidin, "Detecting distributed denial of service attacks using data mining techniques," International Journal of Advanced Computer Science & Applications, vol. 1, no. 7, pp. 436–445.
[13] S. D. Bay, "The uci kdd archive [http://kdd. ics. uci. edu]. irvine, ca: University of california," Department of Information and Computer Science, vol. 404, p. 405, 1999.
[14] M. Al-Kasassbeh, "Network intrusion detection with wiener filter-based agent," World Appl. Sci. J, vol. 13, no. 11, pp. 2372–2384, 2011.
[15] S. K. Pal and S. Mitra, "Multilayer perceptron, fuzzy sets, and classification," IEEE Transactions on neural networks, vol. 3, no. 5, pp. 683–697, 1992.
[16] A. Cutler and G. Zhao, "Pert-perfect random tree ensembles," Computing Science and Statistics, vol. 33, pp. 490–497, 2001.
[17] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5–32, 2001.
[18] J. R. Quinlan, C4. 5: programs for machine learning. Elsevier, 2014.