iJETRM
International Journal of Engineering Technology Research & Management

INTRUSION DETECTION SYSTEM USING MACHINE LEARNING TECHNIQUES

Goli Sushma[1], G Shilpa[2], Kanakandla Vasudha[3]
[1,2,3]Assistant professor, Guru Nanak Institutions Technical Campus
Ibrahimpatnam, Telangana, India

Abstract:
Relevant features can considerably increase the accuracy and performance for intrusion detection engine recently with emerging of new technologies, vast amount of traffic is being generated Therefore we need an dynamic system which can collect data and analyze the incoming traffic. However not all the features may not contribute to identify the incoming traffic, hence we need a reduced subset of features that can improve the speed and accuracy of intrusion detection system In this paper feature selection method has been applied to identify relevant features which can contribute in improving accuracy and speed of intrusion detection system and selected relevant feature based on the score. The main objective is to identify relevant features that can contribute in improving speed and accuracy of intrusion detection system To achieve this in this paper the feature selection approach has been applied on NSL-KDD dataset and used scikit library. relevant features has been identified and observed accuracy rate has been increased.

Keywords:
Network traffic, Intrusion Detection system, classification, NSL-KDD 1999, scikitlearn.

I. INTRODUCTION

Modern technologies like Big Data, Cloud Computing, and Social Network have been emerged. discovering knowledge from these emerging technologies has become very challenging task for business analyst ,marketers and data scientists "With this amount of data transmitted over a network or internet, security becomes a major concern, although multiple intrusion prevention technologies have been built in the past decade to eliminate potential threats despite that, attacks still continue and increase in complexity, this is the reason there is a need of a mechanism to detect any suspicious or unwanted traffic which may cause damage on a particular network. This security mechanism can be implemented using an Intrusion Detection System (IDS) which can be described as a collection of software or hardware devices able to collect, analyze and detect any unwanted, suspicious or malicious traffic either on a particular computer host or network [1]. hence, to achieve its task, an IDS should use some statistical or mathematical method to read and interpret the information it collects and subsequently reports any malicious activity to the network administrator"[2]. "There still exists one main issue regarding the actual intrusion detection technique that is the involvement of human interaction when it comes to label the traffic between an intrusion and a normal one, another major concern is the new challenge of “Big Data” and “Cloud Computing”. These two ubiquitous technologies produce a large amount of data that must be collected and analyzed by the intrusion detection engine dynamically and often the IDS need to deal with multi-dimensional data generated by these large quantities of data. It is necessary to consider that the intrusion dataset can be huge in size, not only the number of observations grown, but the number of observed attributes can also increase significantly and may generate a considerable number of false positives results as it can contain many redundant or duplicate records". Machine learning helps to optimize performance criterion using example data or past experience using a computer program, models are defined with some parameters, and learning is the execution of the programming computer to optimize the parameters of the model using training data. The model can be predictive to make predictions in the future, or descriptive to gain knowledge from data"[4].

iJETRM
International Journal of Engineering Technology Research & Management

II. PROPOSED SYSTEM

"Intrusion Detection is the process used to identify intrusions. An intrusion detection system (IDS) is a device or software application that monitors network or system activities for malicious activities or policy violations and produces reports to a management station"[5].

The proposed model focused on banking sector to detect the intrusion or malicious activities. We collect the network logs from several sources and upload the data to the system to detect and generate the percentage of intrusion. If several times the inner person attacks the same system through the network then it will be difficult to find that some attack has been happened in the network. However we can identify inner and outer attacks using id's through the network , it will use K-means data clustering method to separate the normal activity from abnormal activity. The proposed method in this paper considerably maximize the effectiveness in identifying the attacks and achieve better accuracy as well as low false alarm rate. Mining methodologies are evaluated using KDD Cup Dataset.

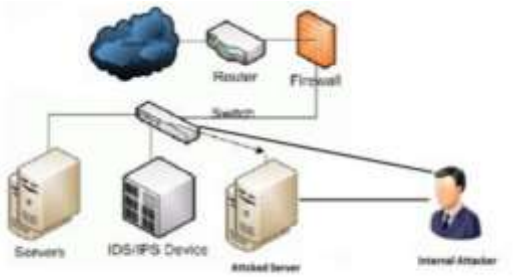


FIGURE: ARCHITRCHTURE OF IDS

III. METHODOLOGY

A. Scikit-Learn Library

In this research scikit-learn library is used, the data to be stored in a two-dimensional matrix or in the form of an array. The size of the expected matrix represented as [samples, features], where the first parameter samples define the number of samples, and each sample is an item to be processed, the second parameter is the number of features that can be used to describe each item in an quantitative manner, In general the value may be real-valued but sometimes it may be Boolean or discrete-valued in few cases.

Fig. 2 shows the data representation in scikit-learn.

$$\text{feature matrix}: \quad \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1D} \\ x_{21} & x_{22} & \cdots & x_{2D} \\ x_{31} & x_{32} & \cdots & x_{3D} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{ND} \end{bmatrix}$$

$$\text{label vector}: \quad \mathbf{y} = [y_1, y_2, y_3, \cdots y_N]$$

Fig 2: data in scikit-learn

- where, N samples and D features

B. Methodology

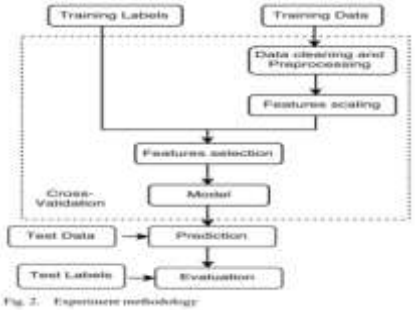
The proposed methodology in this paper, is shown in fig. 1 and describe as follow:

- **Preprocessing**

iJETRM
International Journal of Engineering Technology Research & Management

Data cleaning is the first step to be employed to preprocess the data, it eliminates redundancy, inconsistency and removes noisy in the data. NSL KDD data set, which is used for experimental analysis has already been cleaned.

"Generally, the estimator (classifier) defined in the scikit-learn works well with numerical inputs, so a one-of-K or one-hot encoding method is used to make that transformation. This technique will transform each categorical feature with m possible inputs to n binary features, with one active at the time only"[6].



- **Features scaling**

Features are adjectives, helps in identifying the class of an object. Feature scaling eliminates the feature with large values because that may affect the outcome. "For each feature, calculate the average, subtract the mean value from the feature value, and divide the result by their standard deviation. After scaling, each feature will have a zero average, with a standard deviation of one"[7].

- **Features Selection**

"Feature selection is used to eliminate the redundant and irrelevant data. It is a technique of selecting a subset of relevant features that fully represents the given problem alongside a minimum deterioration of presentation". [12],

- **Step 4: Model**

In Decision tree model , data can be partition using entropy and information gain within the dataset . "This is a very simple but yet an effective hierarchical method for supervised learning (classification or regression) whereas the local space (region) is recognized in a sequence of repetitive splits in a reduced number of steps (small). At each test, a single feature is used to split the node according to the feature's values. If after the split, for every branch, all the instances selected belong to the similar class, the split is considered complete or pure "[10].

Entropy or information gain are the possible best methods to measure a good split among the dataset . The training set entropy is represented by H and calculated in 'bits' of information , it is represented as:

$$H = -\sum_{i=1}^{n} P(c_i)\, log_2\, P(c_i)$$

- **Step 5: Evaluation and prediction**

The test data is used to predict the proposed, different factors were considered such as the accuracy, precision, recall, f-measure and a confusion matrix. A 10-fold cross-validation was also performed during the process.

III. Results

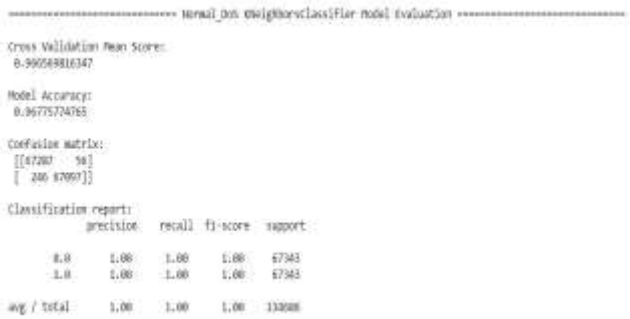
iJETRM
International Journal of Engineering Technology Research & Management

Fig 3: Training &testing data

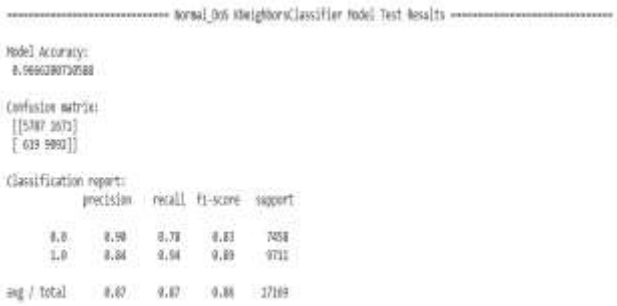
Fig 4: Normal DoS Naive Bayes Classifier Model

iJETRM
International Journal of Engineering Technology Research & Management

Fig 5. Normal DoS Decision Tree Classifier Model

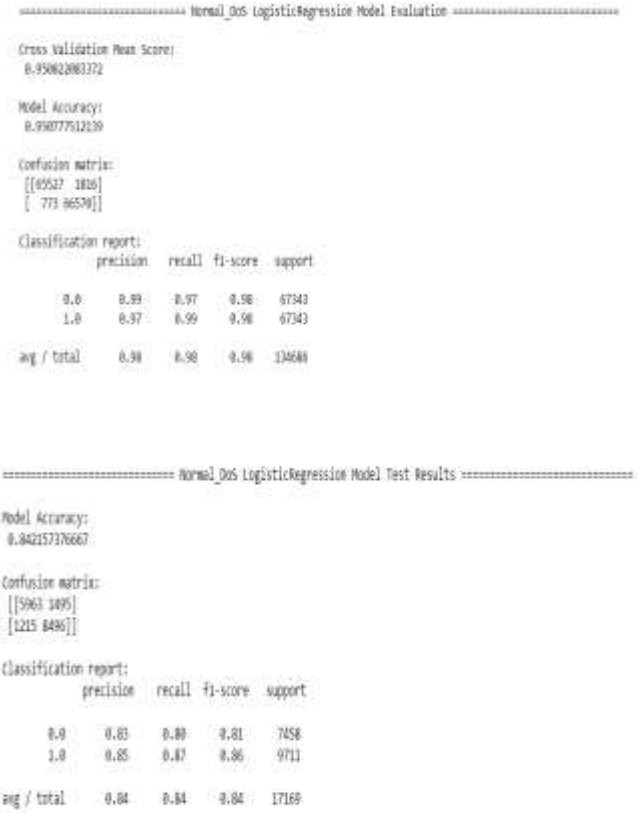


iJETRM
International Journal of Engineering Technology Research & Management



```
========== Normal_DoS KNeighborsClassifier Model Test Results ==========

Model Accuracy:
 0.9696380750588

Confusion matrix:
 [[5787 1671]
 [ 619 9093]]

Classification report:
             precision    recall  f1-score   support

        0.0       0.90      0.78      0.83      7458
        1.0       0.84      0.94      0.89      9711

avg / total       0.87      0.87      0.86     17169
```

Fig 6. K-Neighbours Classifier Model



```
========== Normal_DoS LogisticRegression Model Evaluation ==========

Cross Validation Mean Score:
 0.950822083372

Model Accuracy:
 0.950777512139

Confusion matrix:
 [[65527  1816]
 [  773 66570]]

Classification report:
             precision    recall  f1-score   support

        0.0       0.99      0.97      0.98     67343
        1.0       0.97      0.99      0.98     67343

avg / total       0.98      0.98      0.98    134686


========== Normal_DoS LogisticRegression Model Test Results ==========

Model Accuracy:
 0.842157376667

Confusion matrix:
 [[5963 1495]
 [1215 8496]]

Classification report:
             precision    recall  f1-score   support

        0.0       0.83      0.80      0.81      7458
        1.0       0.85      0.87      0.86      9711

avg / total       0.84      0.84      0.84     17169
```

Fig 7. Logistic Regression Model

CONCLUSION

In this paper it has been demonstrated that the significance of relevant attributes can adequately improves the accuracy of classification for intrusion detection system. the process continuously build a model by including considering each feature aside and repeats the process with remaining attributes available in dataset the

iJETRM
International Journal of Engineering Technology Research & Management

effectiveness of this method is evaluated using different performance metrics and has been proven that by eliminating irrelevant features can increase the accuracy of classification. The proposed method achieved high results.

REFERENCES

[1] M. Alazab et al. “Using feature selection for intrusion detection system,” 2012 Int. Symp. Commun. Inf. Technol., pp. 296–301, 2012.

[2] A. D. Gawande, et al. “Intrusion Detection System for Cloud Computing,” Int. J. Sci. Technol. Res., vol. 1, no. 4, pp. 67–71, 2012.

[3] T. Panchagnula, et al. “Relevance feature selection with data cleaning for intrusion detection system,” 2012 Proc. IEEE Southeastcon, pp. 1–6, 2012.

[4] K. Vinnakota, et al. “An approach for automatic selection of relevance features in intrusion detection systems,” in Proc. of the 2011 International Conference on Security and Management (SAM 11), pp. 215-219, July 18-21, 2011, Las Vegas, Nevada, USA.

[5] L. Han, et al. "Using a Dynamic K-means Algorithm to Detect Anomaly Activities," 2011, pp. 1049-1052.

[6] R. Kohavi, et al., "KDD-Cup 2000 organizers report: peeling the onion," ACM SIGKDD Explorations Newsletter, vol. 2, pp. 86-93, 2000.

[7] I. Levin, et al. "KDD-99 Classifier Learning Contest: LLSoft s Results Overview," SIGKDD explorations, vol. 1, pp. 67-75, 2000.

[8] J. McHugh, et al. “Testing intrusion detection systems: a critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory,” ACM Transactions on Information and System Security, vol. 3, no. 4, pp. 262–294, 2000.

[9] E. Bagheri, W. Lu, and A. Ghorbani, et al. “A Detailed Analysis of the KDD CUP 99 Data Set,” Submitted to Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), 2009.

[12] D. Metia, et al. “An Efficient Hybrid Multilevel Intrusion Detection System in Cloud Environment,” IOSR J. Comput. Eng., vol. 16, no. 4, pp. 16–26, 2014.

[13] Dr. S.P. Shantharajah, et al. "A Study on NSL_KDD Daaset for Intrusion Detection System Based on Classification Algorithms," International Journal of Advanced Research in Computer and Communication Engineering, vol. 4, issue 6, pp. 446-452, June 2015

[14] C. F. Tsai, et al., "Intrusion detection by machine learning: A review," Expert Systems with Applications, vol. 36, pp. 11994-12000, 2009.

[15] V. Bolón-Canedo, et al., "Feature selection and classification in multiple class datasets: An application to KDD Cup 99 dataset," Expert Systems with Applications, vol. 38, pp. 5947-5957, 2011.

[16] F. Amiri, et al., "Improved feature selection for intrusion detection system," Journal of Network and Computer Applications, 2011.

[17] Dasen Ren, et al. “An intrusion detection algorithm based on decision tree technology,” In the Proc. of IEEE Asia-Pacific Conference on Information Processing, 2009.

[18] Mohammad Zahidur Rahman, et al. "Combining Nave Bayes and Decision Tree for Adaptive Intrusion Detection," International Journal of Network Security & Its Applications, Vol. 2, No. 2, April 2010, pp. 12-25.

[19] Sidi F, Affendey L., et al. "Intrusion detection using data mining techniques," 2010 International Conference on Information Retrieval & Knowledge Management(CAMP). 2010.doi:10.1109/infrkm.2010.5466919.

[20] SivaselviK, et al. "Discriminant Analysisbased Feature Selection in KDD Intrusion Dataset," , International Journal of Computer Application VoI.31,No.ll, 2011