International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

Performance and Computation Time Enhancement of Various Machine Learning Techniques for NSL-KDD Dataset

Pradeep K V, K Anusha, S. Nachiyappan

Abstract*: To develop an effective intrusion detection system we definitely need a standardize dataset with a huge number of correct instances without missing values. This would significantly help the system to train and test for real-time use. Previously for research purpose, KDD-CUP'99 dataset has been used, but later on, it has been seen that it is not so useful for training the model as it consists a lot of missing and abundant values. All this issue have been tackled in NSL dataset. To analyze the capabilities of the dataset for intrusion detection system we have analyzed various machine learning classification algorithm to classify the attack over any network. This paper has explored many facts about the dataset and the computation time.

***Keywords* : -KDD, Computation Time, Intrusion Detection.**

I. INTRODUCTION

Since the internet has reached a global level, data security has become an important issue. Everyday millions of network attack are going on across the globe. This could lead to personal data manipulation, cyber-crime, and financial breakdown of a company. To prevent all this previously firewall and encryption methods has been used, but intruders have become aware of these technologies and build multiple techniques to avoid firewall, other prevention technique. Now machine learning has come over to solve the problems of intruders getting into the network. Previously few machine learning approaches have been made and that has given a significant outcome.

This paper is looking forward to the computation speed and detection rate of machine learning models. We have considered SVM, MLP, Gaussian Naïve Bayes algorithm for the classification effectiveness. To evaluate and implement our proposed model, we have considered NSL-KDD dataset.

Following part of this paper has explained the II. Related work III. Proposed Technique IV. Results V. Conclusion VI. Future work VII. Reference.

Manuscript received on March 15, 2020.
Revised Manuscript received on March 24, 2020.
Manuscript published on March 30, 2020.
* Correspondence Author
Pradeep K V*, Asst. Prof, SCOPE, VIT University, Chennai.
Anusha K. Assoc. Prof. SCOPE, VIT University, Chennai.
Nachiyappan S. Asst. Prof.(Sr), SCOPE, VIT University, Chennai.

II. RELATED WORK

Previously many researchers have proposed different kind of analysis over NSL-KDD dataset to improve intrusion detection system using WEKA [1] tool. K-means clustering algorithm utilizes the NSL-KDD [2] data set to prepare and test different existing and new attacks. A relative report on the NSL-KDD data set with its antecedent KDD99 cup data set is made in [3] by utilizing the Artificial Neural Network (Self Organization Map).

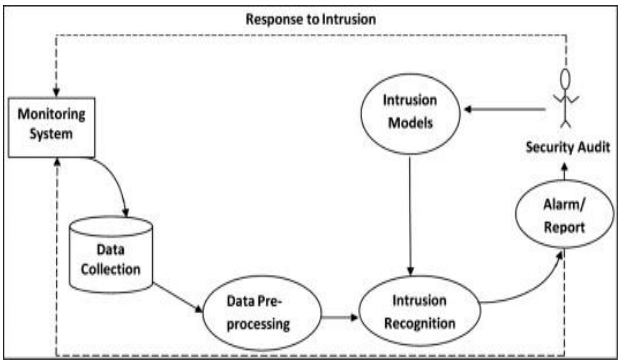
A comprehensive investigation of different informational collections like KDD99, GureKDD and NSLKDD are made in utilizing different datamining algorithms like SVM, Decision Tree, K-nearest neighbor, K-Means. The inherent problem of KDD dataset leads to new version of NSL KDD dataset that are mentioned in [6, 7]. It is very difficult to signify existing original networks, but still it can be applied as an effective benchmark data set for researchers to compare different intrusion detection methods [4].

In [7] they have conducted a statistical analysis on this data set and found two important issues which highly affect the performance of evaluated system, and results in very poor evaluation of anomaly detection approaches. To solve these issues, they proposed a new dataset, NSL-KDD, which consists of only selected records form the complete KDD dataset and does not suffer from any of the mentioned shortcomings.

In [5] they use k mean clustering technique on NSLKDD dataset to find the accuracy for intrusion detection. Shilpa et.al [8] used principal component analysis on NSL KDD dataset for feature selection and dimension reduction technique for analysis on anomaly detection. Generally, Data mining and machine learning technology has been widely applied in network intrusion detection and prevention system by discovering user behavior patterns from the network traffic data.

III. PROPOSED ARCHITECTURE

This paper has proposed 4 important stages of the evaluation process A. Dataset; B. Pre-processing; C. Dimension reduction; D. Classification; E. Evaluation.



A. DATASET :

Selection of the dataset of any machine learning model is a crucial stage, whereas for IDS there are few datasets available on the internet. After reading a few papers we have come to a point that NSL-KDD is the best dataset for research purpose.

1. Dataset does not consist of any missing value.
2. Dataset has no duplicate instances.

Dataset has 41 single vector feature and a class label for each instance. The class label has been defined as either attack or normal.

B. PREPROCESSING:

The dataset is containing an exactly 41 number of features. Which is precisely to be 3 categorical data and 38 numerical data and 1 class label for each instance of both the datasets? Since 3 of the categorical data are not binary so converting them into numerical is a huge task.

Also, we tried finding out the effectiveness of the categorical values by comparing the whole dataset with categorical data extracted dataset. Categorical data slows down the training process and makes the architecture of the classifier much complicated. Thus the categorical data has been extracted from the raw dataset to level up the performance and computational time.

C. DIMENSION REDUCTION:

Dimension reduction is data mining techniques by which we can save the vast majority of the data of a crude dataset additionally reduce the quantity of the feature present in the crude dataset and produce another dataset. A lesser number of a feature in a dataset is dependably taken less time of training and testing.

Principle Component Analysis : PCA is a measurable system of dimension reduction which incorporates covariance, eigenvectors, and eigen values. To calculate the average of a feature in a dataset we use the below formula:

$$\bar{X} = \frac{(\sum X_i)}{N}$$

Where, 'N' is the total number of observation for each X_i , after calculating the average we need to calculate mean adjusted value for each X_i . Now using the following formulas we need to calculate Covariance:

$$COV(X,Y) = \frac{\sum(X_i - \bar{x})\,(Y_i - \bar{y})}{N-1}$$

$$COV(X,X) = \frac{\sum(X_i - \bar{x})^2}{N-1}$$

$$COV(Y,Y) = \frac{\sum(Y_i - \bar{y})^2}{N-1}$$

For further calculation we need to prepare the covariance matrix:

$$A = \begin{bmatrix} COV(X,X) & COV(X,Y) \\ COV(Y,X) & COV(Y,Y) \end{bmatrix}$$

Now to calculate the Eigen values following formula to be used

$$DET|A - \lambda I| = 0$$

Once the λ_1, λ_2 values are calculated then eigenvector will be calculated using the below formula:

$$(\mathbf{A} - \boldsymbol{\lambda}\mathbf{I}) = \mathbf{0}$$

Using the Eigen vector and the mean adjusted value will produce the new dataset which would represent the raw dataset with the lesser number of feature.

$$ND_I = [V_{11} \times (X_i - \bar{x})_I] + [V_{12} \times (Y_i - \bar{y})_I]$$

In this paper, we have used Weka tool to apply PCA on our dataset

Dataset	Raw Dataset (feature count + class label)	After pre-processing (feature count + class label)	PCA COV_800 (feature count + class label)
NSL KDD	41 + 1	38 + 1	15 + 1

Table:-1 Principle Component Analysis on Data Set

D. CLASSIFICATION

The main functionality of our model to detect the attack over any network, to train the model according to that we have used a few classification techniques like SVM, MLP, GAUSSIAN NAÏVE BAYES.

Support Vector Machine (SVM): SVM works on different types of pattern for classifying. This paper has used the linear pattern technique to classifying the dataset. To create the optimal hyperplane SVM uses below formula

$$g(\vec{x}) = \vec{w}^{\,T}\vec{x} + w_0$$

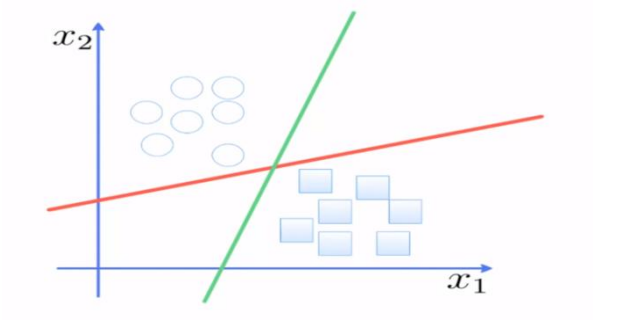


Fig-1 Hyperplane Diagram

Depending on the classes it creates multiple hyperplanes. Then it creates a margin for each hyperplane. These margins are the closest object of each class. Among all the hyperplane whichever has the highest margin distance that will be considered as classifier line

$$g(\vec{x}) \geq 1, \forall \vec{x} \in class1$$

$$g(\vec{x}) \leq -1, \forall \vec{x} \in class2$$

To minimize $\vec{w}$ is nonlinear optimization task, which can be solved using Lagrange multiplier λ_i

$$\vec{w} = \sum_{i=0}^{N} \lambda_i y_i \vec{x}_i$$

$$\sum_{i=0}^{N} \lambda_i y_i = 0$$

Now using with the weight vector we can classify the dataset.

Multilayer Perceptron (MLP): It is based on artificial neural network. Where a perceptron consists of weights of inputs, bias and activation function. Theta is the threshold value by which, it is classified.

$$w_1x_1 + w_2x_2 + w_3x_3 + .. + w_nx_n > \theta \rightarrow 1$$

$$w_1x_1 + w_2x_2 + w_3x_3 + .. + w_nx_n \leq \theta \rightarrow 0$$

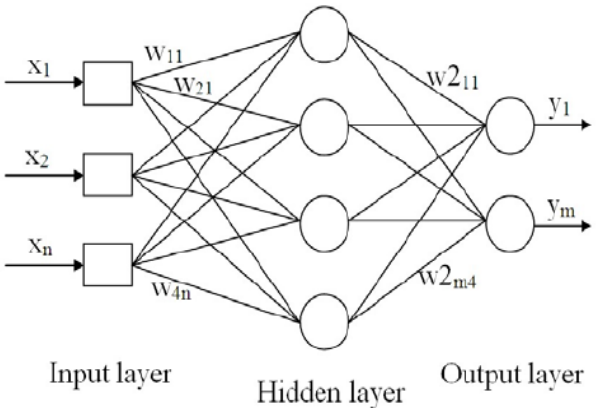


Fig-2: Architecture of Multilayer Perceptron

We also use hidden layers in between input and activation function to make the not linear data frame to classify. Multiple activations can be used for the technique.

Gaussian Naïve Bayes: This classifier works on the principle of the Bayes theorem, where we try to calculate the probability of an event depending on an event which has happened before.

$$P(A|B) = \frac{P(A|B)P(A)}{P(B)}$$

Now depending on the total probability of an event happened independently. We calculate all probability of a given input, whichever is higher will be the class label.

$$P(y|x_1, \ldots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1)P(x_2)\ldots P(x_n)}$$

Gaussian Naïve Bayes classifier assumes that the entire feature is in a continuous form associated with the normal distribution of the probability. So the conditional probability is denoted by:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

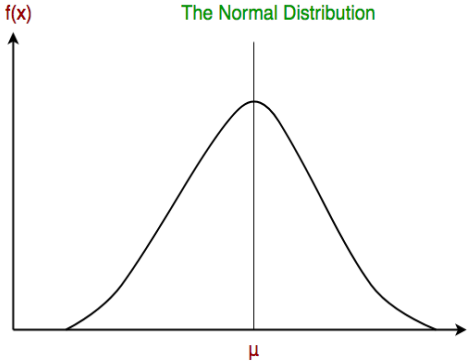
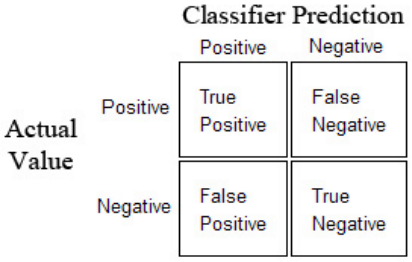


Fig-3 : Normal Probability Distribution Graph

E. EVALUTION

Confusion Matrix: The standard way of evaluating a machine learning model is to determine from the confusion matrix.



Accuracy: This value is generated from the confusion matrix by calculating how many correct predictions has been done by the model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Recall: This value is calculated by the below formula where it is also known as hit rate.

$$Recall = \frac{TP}{TP + FN}$$

Precision: Precision is calculated by the below formula which is nothing but how much positive class has been correctly classified.

$$Precision = \frac{TP}{TP + FP}$$

IV. RESULTS

We have able to achieve significant results of the dataset, which will be elaborated in this portion.

RAW DATASET :

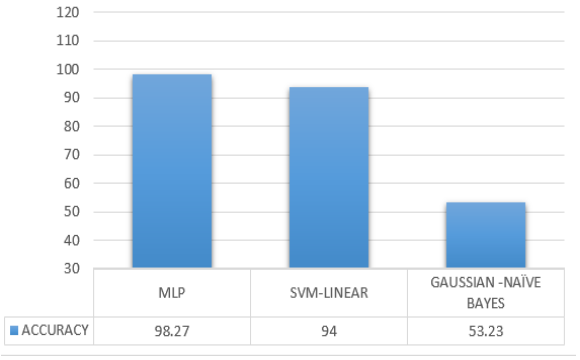


Fig-4: Accuracy Graph

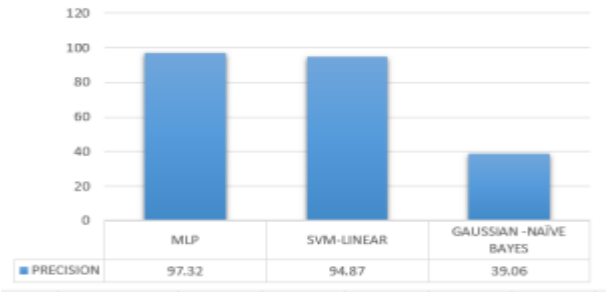


Fig-5: Precision Graph

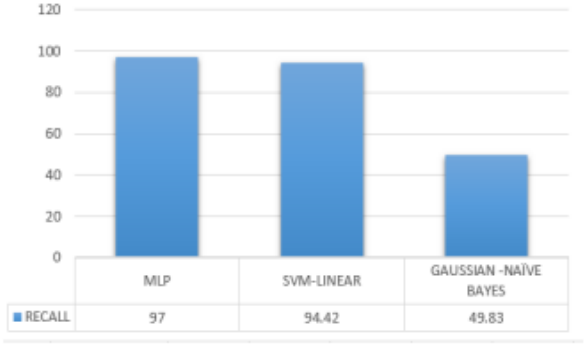


Fig-6: Recall Graph

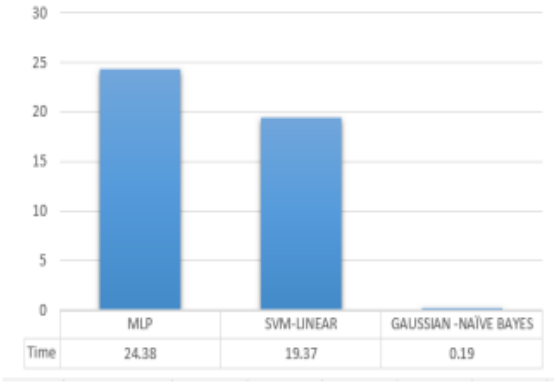


Fig-7: Time Evaluation Graph

Considering all the parameter of the results we can say that the SVM giving us optimal results in case of the raw dataset. We have further evaluated with the same algorithm after prepossessing the dataset which has given us a significant difference in all the parameter.

PREPROCESSED DATASET:

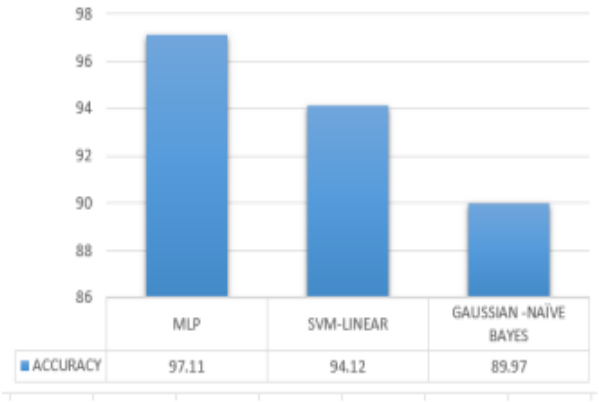


Fig-8: Accuracy Graph

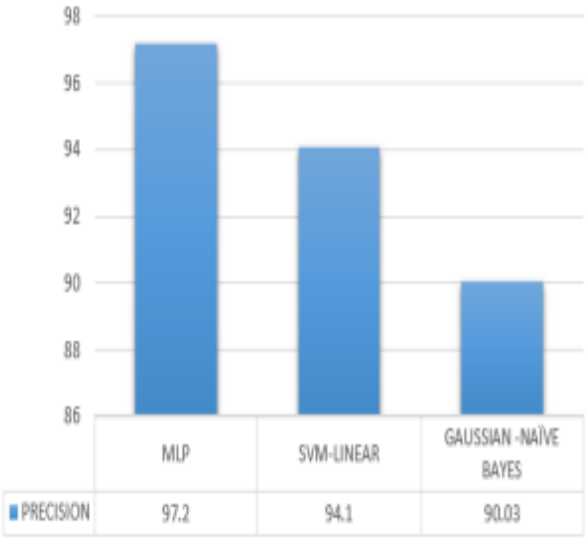


Fig-9: Precision Graph

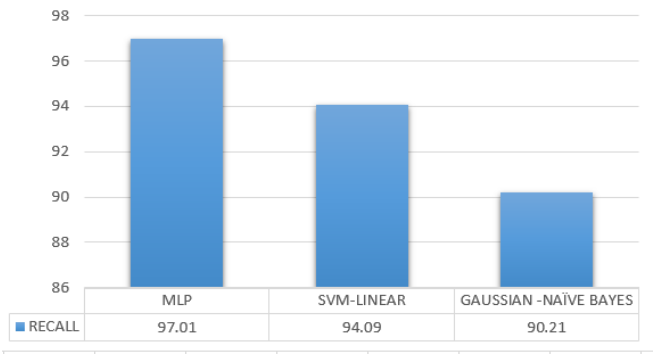


Fig-10: Recall Graph

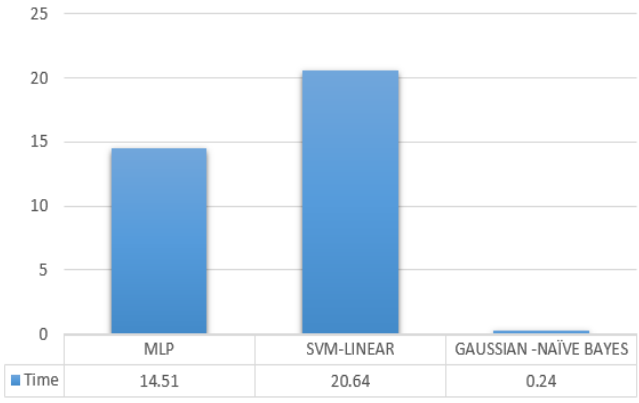


Fig-11: Evaluation Time Graph

After analysis of pre-processed data, we came to a point where we can say that MLP is giving the best outcome regarding all the parameter.

V. CONCLUSION

Comparative analysis of this paper has come to a conclusion that in raw dataset SVM is the best performing but as our goal to minimize the computation time we can say that MLP has excel in all the categories. For future work, we are looking forward to exploring the dataset more atomic way and implement the same in deep learning models.

REFERENCES

1. C. E. Landwehr, A. R. Bull, J. P. McDermott, and W. S. Choi, "A taxonomy of computer program security flaws," ACM Comput. Surv., vol. 26, no. 3, pp. 211–254, 1994.
2. M. Shyu, S. Chen, K. Sarinnapakorn, and L. Chang, "A novel anomaly detection scheme based on principal component classifier," Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop, in conjunction with the Third IEEE International Conference on Data Mining (ICDM03), pp. 172– 179, 2003.
3. KDD Cup 1999. Available on: http://kdd.ics.uci.edu/databases/kddc up 99/kddcup99.html, October 2007. 1852 International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 12, December - 2013 IJERT ISSN: 2278-0181 IJERTV2IS120804
4. J. McHugh, "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.
5. Vipin Kumar, Himadri Chauhan, Dheeraj Panwar, "K-Means Clustering Approach to Analyze NSL-KDD Intrusion Detection Dataset", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume3, Issue-4, September 2013.
6. "Nsl-kdd data set for network-based intrusion detection systems." Available on: http://nsl.cs.unb.ca/KDD/NSLKDD.html, March 2009.
7. Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set", In the Proc. Of the IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA 2009), pp. 1-6, 2009.
8. Shilpa lakhina, Sini Joseph and Bhupendra verma, "Feature Reduction using Principal Component Analysis for Effective Anomaly–Based Intrusion Detection on NSL-KDD", International Journal of Engineering Science and Technology, Vol. 2(6), 2010, 1790- 1799.
9. Lei Li, De-Zhang Yang, Fang-Cheng Shen, "A Novel Rule-based Intrusion detection System Using Data Mining", In the Proc. Of 3rd IEEE
10. S. Revathi, Dr. A. Malathi, "A Detailed Analysis on NSL-KDDDataset Using Various Machine Learning Techniques for IntrusionDetection", International Journal of Engineering Research &Technology (IJERT), ISSN: 2278-0181, Vol. 2 Issue 12, December – 2013
11. Santosh Kumar SahuSauravranjan Sarangi Sanjaya Kumar Jena, "A Detail Analysis on Intrusion Detection Datasets", 2014 IEEEInternational Advance Computing Conference (IACC)

AUTHOR PROFILE

Pradeep K V, **Asst. Prof., SCOPE, VITCC, Chennai** and has more than 11 years of teaching experience. His area of interest in research is Cloud Computing, Image processing, Security in Cloud, Data Analytics and currently pursuing Ph.D. in cloud security. Life-Time Membership in association with Computer Society of India.

K. Anusha, Assoc. Prof. SCOPE, VITCC, has received Ph.D from VIT, Vellore. Her research interests are Network Security, Wireless N/W's, IS and Mobile Adhoc N/W's. She has various Publications and acted as editorial board members of various National, International Journals & Conferences. She has chaired many International conferences' and delivered invited, technical lectures along with keynote addresses.

Prof. S. Nachiyappan, AP(sr.), SCOPE, VITCC, Chennai, Completed his PG in Anna university in 2004 and his area of research is Software Engineering And Big Data. He is having 5 years of Industry Experience and 10 + Years of teaching experience. He is a member of ACM professional and Life-Time membership in ISTE.