IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Volume-5, Issue-11 E-ISSN: 2347-2693

Addressing Challenges in Big Data Intrusion Detection System using Machine Learning Techniques

Saqr Mohammed H. Almansob[1*], Santosh Shivajirao. Lomte[2]

[1*]Department of Computer Science, Radhai Mahavidyalaya (BAM University), Aurangabad, India
[2] Principal, School of Engineering Technology, VDF, Latur, India

[*]*Corresponding Author: saqrmohammed2014@gmail.com*

Available online at: www.ijcseonline.org

Received: 10/Oct/2017, Revised: 29/Oct/2017, Accepted: 16/Nov/2017, Published: 30/Nov/2017

Abstract**—** In the last few years, the number of people around the world is increasing day by day in matching the use of the internet and social media. For this reason, a large volume of data is generated by the internet and social media from gigabytes (GB) to petabytes (PB) with high speed. In this work, it is proposed Intrusion Detection System (IDS) with large amounts of data to address challenges in various types of network attacks using machine learning techniques. On another hand, it is proposed Principal Components Analysis method to reduce high dimensionality and features of data. Therefore, in order to reduce amounts of calculations and improve an accuracy of classification of data. That is, why the use of DARBAI data set in this model and it is applied to K-nearest neighbour method for classification.

Keywords**—** *Big data; Intrusion Detection System (IDS), Principal Component Analysis (PCA), K-Nearest neighbour (KNN)*

I. Introduction

Nowadays, computer and the internet have become a part of life in all fields. So, social media and company are facing many challenges. One of these challenges is security. To protect the data and accounts of users from external attacks have become the big problem now. So Intrusion Detection System able to addressing these challenges in all sectors by monitoring and analysing all the normal and malicious activities over a network. In addition to all these, the false positive alerts about malicious activities when there are malicious activities in the system an increasing scale. For this reason, the reduction of false positive has become a very important work to protect the system from external attacks. The section II has discussed the related work. The section III will characterize of the proposed system. The section IV describes the experiment and results. The section IIV includes the conclusion and reference.

II. Related Work

Md. Al Mehedi Hasan et al [1]. have proposed support vector machine (SVM) and Random forest (RF) for Intrusion Detection System (IDS). Authors applied KDD99 data sets as training and testing phases which do not include any redundant records. The obtained results of this model which applied on KDD99 data sets indicate support vector machine (SVM) is better than Random Forest (RF) in accurate results of classification. Yuteng Guo et al [2]. have explored feature selection based on Rough Set and Genetic algorithm to improve detection accuracy and efficiency. The authors applied KDD cup99 datasets on the model. Feature selection applied to remove all the unimportant feature of KDD data set to improve the accuracy. Furthermore, this is bound to increase the detection rate and reduce the false positive rate over the network. Dong Seong et al [3] have demonstrated that use of principal components analysis (PCA) and Back-propagation Neural Network based Genetic algorithm to optimize Intrusion Detection System (IDS).The authors have addressed some problems over Intrusion Detection System (IDS) such as increased detection rate. On the hand, applied genetic algorithm based PCA and BNN enable Intrusion Detection System to increase detection rate. KDD cup99 data sets applied in their model. Z. Elkhadir et al [4] have proved Principal Components Analysis (PCA) and Kernel Principal Components for intrusion detection system. The dimension in a data set which is big challenge to improve the detection rate. The authors proposed this approach to reduce the high dimensionality of data and select appropriate features for enhancing detection and accuracy rate. Furthermore, applied K-NN algorithm as classification, the KDDcup99 data applied in the model. Gholam Reza et al [5] have proposed Category and Principal Components Analysis (PCA) for intrusion detection. The authors applied this approach to improve the detection speed and reduced high dimensionality of data se so the DARBAI data sets which are applied in the model. The obtained results of Category based PCA approach indicates that dimension can reduce the process of time and improve the accuracy in an Intrusion Detection System. Bahareh Gholipour et al [6] have identified SVM and ABC approach to select features of KDDcup99 data set to enhance network intrusion detection and improve accuracy rate. The authors applied support vector machine (SVM) as a classifier and in order to enhance classification results. The obtained results of the model indicate that their approach has improved the results in Intrusion Detection System. Gan Xu-shing [7] reports the Partial Least Square (PLS) and Vector Machine algorithms for addressing anomaly intrusion detection problem. The authors applied PLS algorithm to reduce high dimensionality and extract features of data set whereas applied CVM as fast classification algorithm. KDD99 data set used in the model so obtained results of applied PLS-CVM algorithms indicates that solves the problem of anomaly intrusion detection.

III. Proposed Work

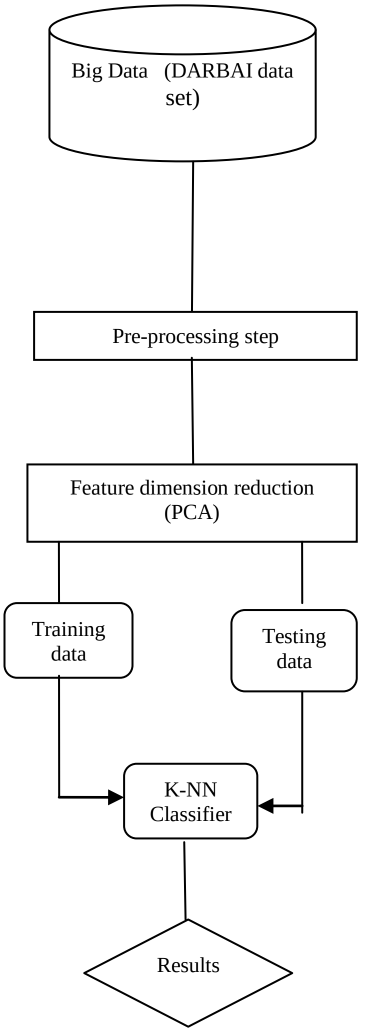

Fig.1

A. DARPA Data Set

In DARBA data set, there are several steps including pre-processing of data, reduction of features of data, classification, and evaluation. So we applied principal components analysis method has been applied to reduce the dimensionality of data while preserving information of original data.

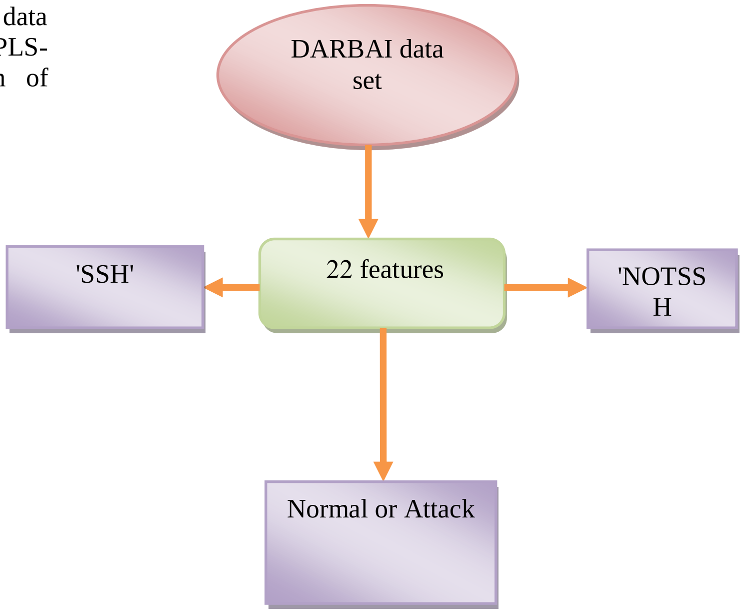

Fig1. The classified of KDD dataset.

B. Pre-processing Step

Pre-processing data is a very important method to understand data sets in a suitable and comprehensible format for improving the quality of data. Pre-process applied in this work as follows. Read training or testing data sets and pre-process to numerical, data whose all elements are 1 or 0, pre-process the input data into numerical data using some of the calculations and normalization of data finally to be numeric data whose members is a number between 0 and 1 to organizes data for more efficient access.

C. Principal Components Analysis (PCA)

The goal of applied Principal Components Analysis (PCA) approach is to reduce high dimensionality and feature extraction of DARPA Data Set by transforming a number of correlated variables into uncorrelated variables. Furthermore, computing a linear transformation from high dimensionality to lower dimensionality space so PCA is very effective algorithm for reducing high dimensionality in intrusion detection(8)

$$X = [X1, X2, X3, X4, X5 \ldots\ldots, Xn] \quad (1)$$

Each column of x is vector and defined by expected value in equation (2)

$$x = \frac{1}{M} \sum_{1=1}^{m} xi$$

Principal Component Analysis
= a1X1 + a2X2 + a3X3 · · · + adXd (2)

PCi = Principal Component, '*i*'; *Xj*— Orginal feature and '*j*'; *aj*— numerical coefficient for *Xj*.

3.4 The K-nearest neighbour algorithm (K-NN)

The K-NN algorithm is very suitable to deal with large data set which is implemented in this model for the classification. Furthermore, a supervisor, machine learning technique which applied as classification purpose (9) So, the K-nearest neighbour algorithm computes the distance for each training and testing sample in DARBAI data sets to find an exact nearest neighbour and to measure the similarity between two points. (10)

VI. Results and Discussion

Firstly, we have reduced features of data using PCA (Principal Components Analysis) because the data has many observations and features. To reduce features of data is a good method to reduce the amount of calculation and improve the accuracy of classification of data. Secondly, using reduced data and K-NN classifier is quite useful to classify the testing data and evaluate its accuracy. The training and testing data: are 349884, 150117 respectively.

A. Evaluation criteria.

Standard performance measures can be written as follow:

Detection Rate = number of attack correctly identified attacks / total number of attacks in dataset

DR = TP/ (TP + FN) (3)

FPR = incorrectly identified normal events classified as attack / total number of normal events in dataset

False-Positive Rate = FP / FP + TN (4)

Accuracy Rate = number of correctly classified instances / Total number of instances in the dataset

Accuracy = TP/ (TP+TN) (5)

B. Results

The suggested PCA-K-NN techniques is implemented after applying MATLAB R2015a-64 bit installed on windows 7 Ultimate with the core i5 processor and 12 GB RAM . Table 1 shows the results of the PCA-K-NN techniques in terms of detection rate (DR), false positive rate (FPR), and accuracy rate (AR). The proposed model gives a higher performance appraisal of detection rate and accuracy. As well as, it has given a lower false positive rate.

Table1. The result of classification of DARPA99Week1-1

Name	*Value*	*Description of value*
Data	500001	The total number of data
Teidx	150117	Total number of testing data.
Tridx	349884	Total number of training data
Accuracy rate	97.17	Classification accuracy
DR	97.84	Detection Rate
FPR	0.87	False Positive Rate

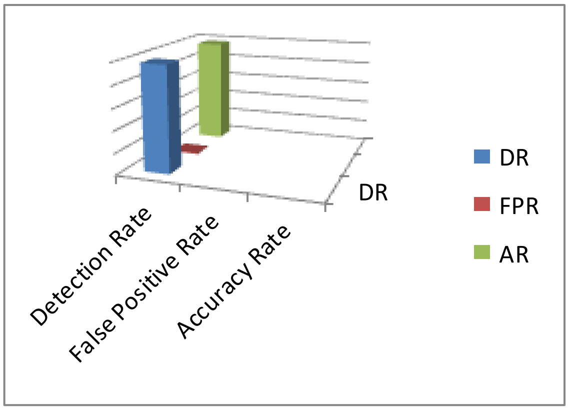

Figure2: Results of DR, AR, and FAR

V. Conclusion

In this paper the high dimensionality and false positive rate have become big challenges which facing Big Data Intrusion Detection System (IDS). The present paper has addressed two challenges which are faced by big data Intrusion Detection System. So, two algorithms have been proposed to address all the challenges. Principal Components Analysis (PCA) method is proposed to reduce high dimensionality and features selection. Therefore, amounts of calculations are reduced and an accuracy of classification is improved of data. That is, used DARBAI data set in this model and has been to applied K-Nearest Neighbour (K-NN) method for classification. The proposed model gives a higher performance appraisal of detection rate and accuracy. As well as, it has given a lower false positive rate.

References

[1]. Md. Al Mehedi Hasan et al: *Support Vector Machine and Random Forest Modeling for Intrusion Detection System (IDS)* Journal of Intelligent Learning Systems and Applications, 45-52, 2014.

[2]. Yuteng Guo et al*: Feature Selection Based on Rough Set and Modified Genetic Algorithm for Intrusion Detection*. The 5th International Conference on Computer Science & Education Hefei, China. August 24–27, 2010, 97 8-1-4244-6005-2/10/$26.00 ©2010 IEEE.

[3]. Dong Seong et al: *An Optimized Intrusion Detection System Using PCA and BNN.*

[4]. Elkhadir et al: *Intrusion Detection System Using PCA and Kernel PCA Method*. IAENG International Journal of Computer Science, 43:1, IJCS_43_1_09(Advance online publication: 29 February 2016)

[5]. Gholam Reza et al: *Category-Based Intrusion Detection Using PCA*. Journal of Information Security, 2012, 3, 259-271 http://dx.doi.org/10.4236/jis.2012.34033 Published Online October 2012

[6]. R. Wankhede, V. Chole, "*Intrusion Detection System Using Hybrid Classification Technique*", International Journal of Computer Sciences and Engineering, Vol.4, Issue.11, pp.30-33, 2016.

[7]. Gan Xu-shing et al: *Anomaly intrusion detection based on PLS feature extraction and core vector machine*.0950-7051/$-see front matter 2012 Elsevier B.V. All rights reserved.

[8]. Zyad Elkhadir et al: *Intrusion Detection System Using PCA and Kernel PCA Methods*.16 April 2016

[9]. S.Venkata Lakshmi et al*: Application of k-Nearest Neighbour Classification Method for Intrusion Detection in Network Data*. International Journal of Computer Applications (0975 – 8887) Volume 97– No.7, July 2014.

[10]. Yihua Liao et al: *Use of K-Nearest Neighbor classifier for intrusion detection*. Computers & Security Vol 21, No 5, pp 439-448, 2002 Copyright ©2002 Elsevier Science Ltd.