Classification of Intrusion Using AdaBoost Classifier with J48 Decision Tree

Gaurav Kumar Das
Computer Science & Engineering
Kautilya Institute of Technology & Engineering
Jaipur, India
Dr. Vijay Kumar
Computer Science & Engineering
Kautilya Institute of Technology & Engineering
Jaipur, India

Abstract*— Classification of KDD intrusion dataset is proposed along with noise reduction, clustering and feature selection. DBSCAN algorithm has been applied to reduce noise present in KDD dataset. After noise removal genetic search approach is utilize to pick relevant feature. K-Means clustering method is utilized to cluster the dataset and resultant dataset is tested by decision tree based Adaboost classifier. It is examined that suggested approaches gives 98.822% accuracy. A comparative analysis performed between proposed methods and KMSVM (Simple K-mean with SVM classification) and it is observed that proposed method gives better results.

Keywords*— Intrusion Detection System, Density based algorithm, K-Means Clustering, Decision Tree J48, AdaBoost Classifier.

I. Introduction

An Intrusion Detection System (IDS) is a blend of software and hardware which are used for detecting intrusion. It accumulates and analyzes the network traffic & detects the malicious patterns and finally alert to the proper authority [1].

IDS named into two classifications: Misuse detection and Anomaly detection. In misuse detection, pattern of perceived malicious movement is put away with-in the dataset and select uncertain data by methods for assess new sample with the store pattern of attacks. Anomaly identifier screens new samples. The recently arrived attacks are comparing with the baseline, if there may be any difference from base-line, data is notified as interference [2].

In IDS, the records deals from more than one source which includes network traffic or logs, device logs, software logs, alarm messages, & so forth. Owing to varied data source & format, the complexity increased in audit & analysis of data. Data Mining has vast benefit in data extraction from huge volumes of data that are noisy & dynamic, thus it's of first-rate significance in IDS [3].

Density-based methods are for the purpose of unpacking clusters of subjective shape. The chief idea is that, for every info within the given category, during a given vary of space should contain a minimum of a specific range of points. This procedure can be utilized to filter the "noise" outlier data [4]. K-Means clustering frame-work is utilize to parcel the training data into k clusters with the support of Euclidean distance correspondence. It is an iterative clustering algorithm where products are process amongst group of clusters in anticipation of the favored group is attain [3].

AdaBoost is a machine learning algorithm which executes in cycles. In every cycle, the weights of misclassified tests are extended while the weights of legitimately ordered delineations are diminished. This type of weight updating is most important since AdaBoost can later focus on the classification of difficult samples to decrease the training error [5]. J48 is the extension of C4.5 and it is the decision tree based algorithm. With this technique a tree is worked to show the arrangement procedure in decision tree the interior nodes of the tree indicates a test on a characteristic, leaf node holds a class mark, branch speak to the aftereffect of the test and the highest node is the root node. Model created by decision tree predicts new occurrences of data [6].

II. Literature Survey

Yang Jian [2009] suggested an enhanced intrusion detection model based on DBSCAN which demonstrates the idea of producing a cluster by merging small clusters and using density method. The paper portrays a additional realistic density-based clustering algorithm for intrusion detection (IIDGB), using coherent technique to calculate the space and plan the technique of parameter selection [7].

Z. Muda, et al. [2011] projected innovative work of IDS based on OneR classification and K-means clustering; they presented a strategy which relate the methods of OneR classification and K-means. The main objective of paper is to use K-means to split and group data into usual and attack instances. The algo partition the dataset into k clusters as per an initial value known as seed point into each cluster's centroids or cluster centers. The mean estimation of every data enclosed inside each bunch is named as centroids [8].

Zhengjie et al. [2011] proposed anomaly IDS in light of K-means algorithm with PSO. Particle swarm optimization (PSO) algo is an evolutionary calculating technology which is based on swarm intelligence has good universal search ability. The proposed algo has defeated falling into local minima and has sensibly great entire joining [9].

Karthick et al [2012] characterize a versatile system IDS, that uses two stage design. In the first phase a probabilistic classifier is utilize to identify prospective inconsistencies in the traffic. In the second phase a HMM based traffic display is used to slight down the planned assault IP addresses. Numerous design selections that were made to make this scheme practical and difficulties faced in integrating with present models are also defined. We illustrate that this system accomplishes good performance empirically [10].

H. Fatma & L. Mohamed [2013] suggested a two phase method to progress intrusion detection scheme based on data mining algorithm. They adopted a two phase method in order to improve the accuracy of sensors. The first phase purpose to produce meta-alerts through clustering and the second phase purposes to reduce the rate of false alarms using a binary classification of the produced meta-alerts. For the first phase they utilized two alternatives, self-organizing map (SOM) with K-means algorithm and neural GAS with fuzzy c-means algorithm and for the second phase they utilize three methods, SOM with K-means algorithm, decision trees and support vector machine [11].

Nutan Farah Haq et. al [2015] an ensemble IDS system is define via a sequence of machine learning classifiers and a hybrid FS approach. This build a appropriate NSL-KDD train dataset, decrease features into 12 from 41 through the describe hybrid FS method; Building up classification models utilizing training data; Classification of test examples utilizing larger part vote and prior assembled arrangement models as base classifier. The FP) rate of the define replica is 0.021 with a TP rate of 98.0%.The outcome demonstrates that the proposed ensemble model is a consistent and exceed other classifiers performance [12].

Susheel Kumar Tiwari et al [2018] proposes up a network intrusion detection method based on the fusion algorithm combining with information entropy and K-means, analyze results demonstrate that the blend of algorithm has upgraded the identification proportion and lessened the false alarm proportion related with customary K-means algorithm. Though, the execution of the combination did not consider the algorithm execution efficiency, which needs the additional study [13].

III. Proposed Methodology

To enhance the classification accuracy of KDD dataset proposed a new approach. Proposed approach consists of following four phases shown in the figure 3.1.

1. Noise Reduction
2. Feature selection
3. Clustering
4. Classification

Proposed Algorithm:

1. Apply KDD Cup99 dataset. Select one lakh records from them.
2. Choose DBSCAN algorithm to remove noise.
3. For selecting attribute choose genetic algorithm. Here out of 41 features only 13 relevant attributes are selected for further processing.
4. Selected attribute are clustered by using K-means algorithm. It produces 4 cluster.
5. Apply AdaBoost classifier with J48 decision tree.
6. Check intrusion happens or not.
7. Stop.

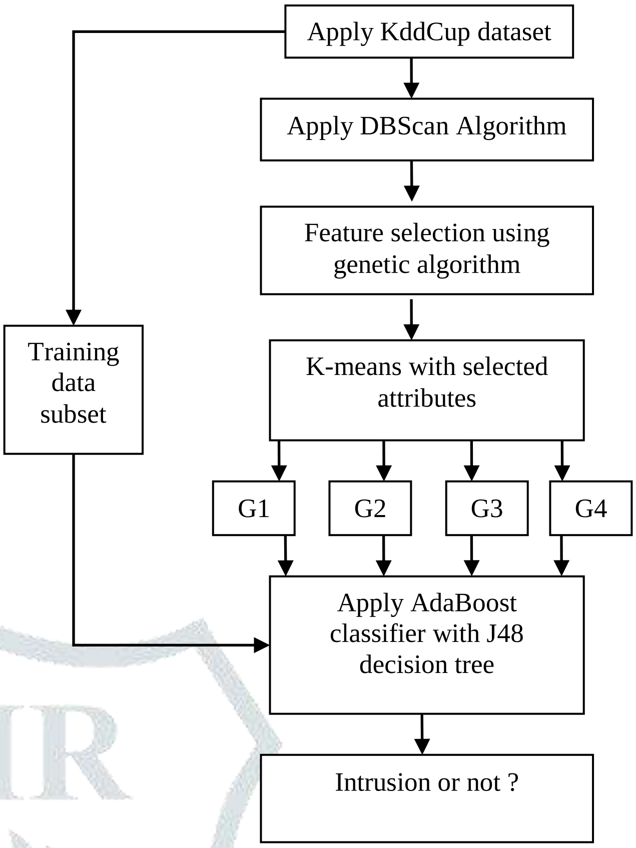
Apply KddCup dataset
Apply DBScan Algorithm
Feature selection using genetic algorithm
Training data subset
K-means with selected attributes
G1
G2
G3
G4
Apply AdaBoost classifier with J48 decision tree
Intrusion or not ?

Fig. 3.1 Flow Diagram of Proposed Approach

IV. Result Analysis

In the result analysis, the experiment of proposed work performed by using MATLAB and WEKA tool. KDD Cup99 dataset used for the implementation.

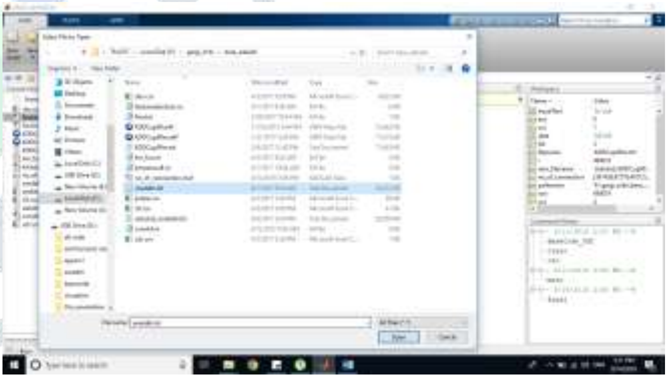

Fig. 4.1 Select 1 lakh records from KDDCup99 dataset

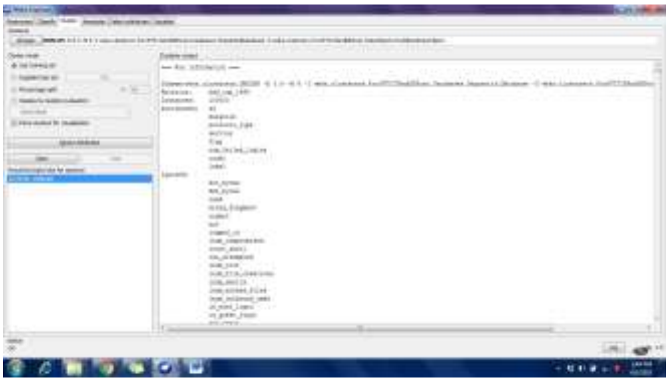

Fig. 4.2 Apply DBSCAN for removal of noise

Fig. 4.3 Final Selected Attributes through genetic method

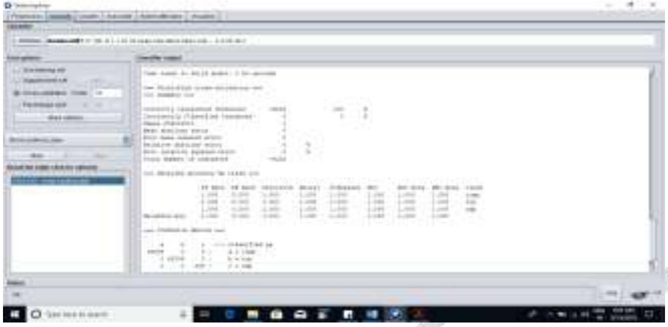

Fig. 4.4 Classified Illustrations using AdaBoost classifier with J48 decision tree

Table 4.1: AdaBoost with J48 decision tree outcomes on clustered data

Data	C1	C2	C3	C4	Average
Correctly Classified Instances	100 %	99.638%	95.652%	100%	98.822%
Incorrectly Classified Instances	0%	0.361%	4.347%	0%	1.177%
Kappa statistic	1	0.992	0.928	1	0.98
Mean absolute error	0	0.002	0.012	0	0.003
Root mean squared error	0	0.049	0.108	0	0.039
Relative absolute error	0%	0.973%	7.322%	0%	2.073%
Root relative squared error	0%	12.83%	36.938	0%	12.442%
Total Number of Instances	79232	830	230	8	20075

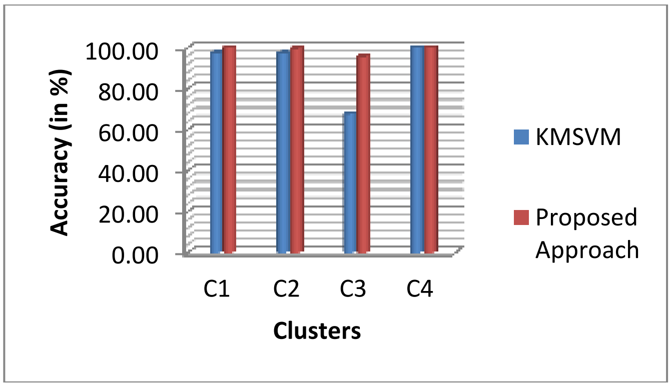

Fig. 4.5 Comparison graph of Accuracy

V. Conclusion

Security is the most concerning issue in the modern time. In this report a novel approach is proposed which is a hybrid model of classification and clustering.KDD99Cup dataset suffered from noise which is removed by using DBSCAN algorithm than feature selection technique applied to select relevant attribute. K means algorithm purposeful to bunch dataset into G1, G2, G3 & G4. At last AdaBoost classification algorithm applied on clustered data with J48 decision tree to know intrusion happened or not. The overall approach improves the accuracy. It is observed that obtained accuracy is 98.822% which is better than KMSVM. In the future it can be done by another clustering method or classification algorithm. In place of AdaBoost classifier can be used machine learning techniques.

References

[1] Amanpreet Chauhan, Gaurav Mishra, Gulshan Kumar “Survey on Data Mining Techniques in Intrusion Detection” International Journal of Scientific & Engineering Research Volume 2, Issue 7, July-2011 1 ISSN 2229-5518.

[2] D. Shona, A.Shobana “A Survey on Intrusion Detection using Data Mining Technique” International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 12, December 2015.

[3] Chandrakant Jain, Aumreesh Kumar Saxena “General Study of Mobile Agent Based Intrusion Detection System (IDS)” 5 February 2016; accepted 10 April 2016; published 13 April 2016 Copyright © 2016.

[4] M.Suresh Babu, Dr. N.Geethanjali, Prof B.Satyanarayana, Clustering Approach to Stock Market Prediction, Int. J. Advanced Networking and Applications Volume: 03, Issue: 04, Pages:1281-1291 (2012).

[5] Shuqiong Wu and Hiroshi Nagahashi, “Parameterized AdaBoost: Introducing a Parameter toSpeed Up the Training of Real AdaBoost”,IEEE Signal Processing Letters, Vol. 21, No. 6, June 2014

[6] Manisha Kansra and Pankaj Dev Chadha, “Cluster Based detection of Attack IDS using Data Mining”, International Journal of Engineering Development and Research (IJEDR), Volume 4, Issue 3, 2016, pp. 1082-1087.

[7] Yang Jian, "An Improved Intrusion Detection Algorithm Based on DBSCAN", Micro Computer Information, 25, 1008-0570(2009)01- 3- 0058-03, 58-60, 2009.

[8] Z. Muda, W. Yassin, M.N. Sulaiman and N.I.Udzir, "Intrusion Detection based on K-Means Clustering and OneR Classification" In Information Assurance and Security (IAS), 7th International conference, 2011.

[9] Zhengjie Li, Yongzhong Li, Lei Xu, "Anomaly intrusion detection method based on K-means clustering algorithm with particle swarm optimization", In ICM, 2011.

[10] R Rangadurai

[11] Karthick et.al, “Adaptive Network Intrusion Detection System using a Hybrid Approach”, IEEE, 2012, pp. 1-7.

[12] H. Fatma, L. Mohamed, "A two-stage technique to improve intrusion detection systems based on data mining algorithms", In ICMSAO, 2013

[13] Suvendra Kumar Jayasingh, Jibendu Kumar Mantri, P. Gahan , “Comparison between J48 Decision Tree, SVMand MLP in Weather Forecasting”, SSRG International Journal of Computer Science and Engineering (SSRG-IJCSE) – volume 3 Issue 11–November 2016.

[14] Susheel Kumar Tiwari, Dr. Manish Shrivastava” Implementation of Improved K-Mean Algorithm for Intrusion Detection System to Improve the Detection Rate”, 2018 IJSRCSEIT, Volume 3, Issue 1,| ISSN: 2456-3307.