ANOMALY BASED INTRUSION DETECTION SYSTEM FOR NETWORKS USING K-MEANS CLUSTERING, DISCRETIZATION, AND NAÏVE BAYES CLASSIFICATION: A HYBRID APPROACH

S. SIVANANTHAM
Assistant Professor, Adhiyamaan College of Engineering, Hosur, Tamilnadu, India,
sivanantham17@gmail.com.
V. MOHANRAJ
Professor, Sona College of Technology, Salem, Tamilnadu, India, vmohanraj06@gmail.com

Abstract

The role of Intrusion Detection System (IDS) is to discover, assess and account unauthorized access, illegal activities and security issues in Network. The aim of this research is to model an Anomaly-Based Intrusion Detection System as a Semi-supervised machine learning method involving K - Means Clustering and Naïve Bayesian Classification. This proposed method focuses on reducing data loss during clustering and improving the exactness of classification. This is achieved by adding an intermediate pre-processing technique named Proportional K - Interval Discretization (PKID). The proposed combination of K - Means Clustering, PKID, and Naïve Bayesian Classification is evaluated against KDD Cup 99 Dataset. The results show that the proposed scheme gives an average Accuracy rate of 99.32%, Detection rate of 99.45% and False alarm rate of 0.128 during training and testing phase.

Keywords*: Intrusion Detection System, K-Means Clustering, Naïve Bayesian Classification and Discretization.*

1. Introduction

Security has been a mythical term in the field of Networks. Identifying authorized and unauthorized users in the Network without affecting their privileges is a major concern in Network. An illegitimate user who accesses the network resources is called an Intruder. An Intrusion Detection System (IDS) is used to identify and mitigate the intruders. Intrusion Detection System protects the network by analyzing the network streaming traffic data and finds the trace of attacks if any and checks for anomalous data. An efficient Intrusion Detection System should maintain baseline between normalcy and anomaly, keeping the false positive rate low as much as possible.

Data mining is a critical technology used to improve the efficiency of Intrusion Detection Systems significantly. Data mining includes many approaches like Clustering, Classification, Spectral Decomposition, Statistics based approaches etc [1] [2]. Though many Data mining techniques are available, Classification draws much attraction because of its suitability for streaming data [3]. In this paper, we propose a new Hybrid, Semi-supervised and efficient Data mining model for Intrusion Detection. It combines K-Means Clustering (KMC) - PKID Discretizator - Naïve Bayesian Classifier (NBC).

Discretization is often used to make learning accurate and speedy. Discretization can be with no trouble used on numeric instances on which filtering methods may not be applied. Discretization works by grouping continuous value into a number of discrete intervals. The main factors which differentiate the Discretization methods are grouping strategy of continuous data, a number of intervals and position of cut points to split. The Proportional K - Interval Discretization (PKID) uses a probability estimation to regulate the number and size of intervals to a number of instances [17]. K - Means Clustering is a method of grouping instances into k disjoint clusters based on the feature value of the data instances where k value is user defined. Naïve Bayesian classification is a supervised technique which classifies data instances using a probabilistic model representing the probabilistic independencies among the data instances which is developed during training [2] [3].

The benefit of the proposed scheme is it enjoys the advantage of both supervised and unsupervised techniques [4]. Another advantage of the scheme is the use of classifier, which will give the exact set of anomalies. To evaluate the proposed method, the scheme is applied on KDD Cup 99 Dataset. Primarily 10% of the complete dataset containing 4, 94,020 instances, is used to train the model and later the test data set is used and results are obtained. Initially Naïve Bayesian classifier (NBC) is applied over the dataset and results are recorded. Next the hybrid combo of K- Means Clustering and Naïve Bayesian classifier (KMC-NBC) is applied. Finally the proposed scheme of K- Means Clustering with PKID Discretization and Naïve Bayesian classifier (KMC-PKID-NBC) is applied to the dataset. Comparing the results, the proposed KMC-PKID-NBC scheme gives a better accuracy of 99.95% and 99.94% during the training & testing phases respectively. The entire scenario is implemented and tested in the Data mining tool, WEKA.

2. Related Work

In the recent days, much research has been undertaken by researchers in the development of efficient Intrusion Detection Systems (IDS) using Data mining Techniques. Let us review some noteworthy contributions towards IDS and Data mining approaches in this section,

Arif Jamal Malik et.al, proposed an IDS using binary PSO and Random forest. Binary PSO was used to enhance feature selection in classification and the Random forest is a category of the classifier [5].

Natesan et.al introduced a new method of combining ADA boost filter, Decision tree and Naïve Bayes classifiers to improve the performance of IDS and they have tested their method against KDD cup 99 Dataset [6].

Preeti Aggarwal et.al analyzed KDD Dataset using Random forest method based on the four classes basics, traffic, content, and host [7].

Mrutyun Jaya Panda et.al investigated variously Supervised and Unsupervised filtering methods used in Intelligent IDS development [8].

W. Yassin et al implemented an integrated a machine learning approach involving K-Means clustering and Naïve Bayesian classification to maximize accuracy and detection rate [9].

Anusha Jayasimhan et al proposed Anomaly-based IDS using K - Means clustering to reduce the false negative rate and to detect novel attacks [10].

Jose F Nieves et al, developed Anomaly detection system combining K - Means clustering and Particle Swarm Optimization (PSO) to solve local optima problem and to find volume anomalies [11].

Reda M et al introduced A hybrid approach which combines random forest and weighted K-Means clustering which improves detection rate but has the issue of difficulties in encoding rules and in finding new attacks [12].

Tahir Md et al combined K- Means clustering and Naïve Bayesian Classifier with Discretization and proposed a hybrid approach to improve Accuracy and False Alarm rate. Time taken for building the model reduces considerably when data is discretized and the proposed method is tested against ISCX 2012 Intrusion evaluation dataset [13].

Majority of the literature work given above focus on improvising the accuracy and detection rate while keeping the false alarm rate minimum. The review shows clearly that the Hybrid Data mining approaches provide better performance rather than individual ones. In this proposed scheme, we have chosen PKID Discretization which is considered to be most suitable one to be used with Naïve Bayesian classifier [14] and also the literature shows K - Means clustering is the appropriate method to be combined with Bayesian Classifier.

3. Experimental Setup

In the proposed method, PKID Discretization is used along with K Means Clustering and NB Classifier to improve the accuracy and detection rate. The entire experiment is setup in WEKA Tool and tested against KDD Cup 99 Dataset.

About KDD Cup 99 Dataset

Selection of appropriate Dataset is a mission-critical task in any comparative study of data mining methods. In this work, we chose KDD Cup 99 Dataset which is most important Network Intrusion Evaluation dataset in KDD Archive. The Dataset contains 41 attributes and the instances are represented as single connection vectors [15]. The vectors are well labeled as Normal and Attack. The 10% of the dataset contains around 5 Lac instances which could be used as training set. In this training set, 19.69% connections are labeled Normal and 80.31% connections are labeled Attack [16]. The attacks are categorized as four types namely,

Denial of Service Attack (DOS):

In this attack, the Computer system is compromised and it is made either too busy or too full to process legitimate requests. Sometimes, even authorized access to the system also might be blocked. DOS includes attacks like teardrop, pod, smurf, land, back, and neptune.

User to Root Attack (U2R):

The attack starts with entry to a system as a normal user and getting access to the root of the system. U2R includes attacks like rootkit, buffer_overflow, load module and perl.

Remote to Local Attack (R2L):

This attack is one in which the attacker sends data to a system in the Network and tries to obtain local access to the system. R2L includes attacks like warezmaster, imap, phf, multihop, ftp_write, guess_password, warezclient and spy.

Probing Attack (PROBE):

The attack begins when the attacker tries to collects information about the system and network in order to breach the security. PROBE includes attacks like nmap, portsweep, satan and ipsweep.

The protocols considered in KDD are Transmission Control Protocol (TCP), User Datagram Protocol (UDP) and Internet Control Message Protocol (ICMP).

The attacks are grouped protocol wise as,

Attacks corresponding to TCP Protocol	**Attacks corresponding to UDP Protocol**	**Attacks corresponding to ICMP Protocol**
neptune,portsweep,guess_password,buffer_overflow,land,phf,warezmaster,warezclient,ipsweep,perl,multihop,back,ftp_write,rootkit,loadmodule, imap, spy, satan.	teardrop, satan, nmap, root kit.	Portsweep, satan, ipsweep, nmap, smurf, pod.

About WEKA

WEKA can be abbreviated as Waikato Environment for Knowledge Analysis. WEKA is a powerful tool in which Data Pre-processing, Data mining and Data Visualization can be carried out [15]. Various dataset formats namely ARFF, CSV, C4.5, XRFF, BSI, libsvm data files, Json instance files, MATLAB ASCII files etc, are supported in WEKA and this makes WEKA inevitable choice for Data mining. To name a few, the Data mining methods like Clustering, Classification, Regression, Association etc, could be employed in WEKA.

Figure 1. Shows implementation of the proposed scheme, initially, the test dataset is loaded into WEKA through the WEKA Explorer. Then the data has to undergo 2 stages of pre-processing. First, Simple K - Means Clustering technique is applied by selecting Add Cluster under Unsupervised attribute filters and second, Discretization is applied by choosing PKID Discretize under Unsupervised attribute filters. Once the pre-processing is over, from the classify tab, Naïve Bayesian Classifier is chosen and applied. To improve accuracy the cross-validation is fixed as 10 folds. The same procedure is repeated for a different number of clusters by setting the value of K= 3, 4, 5, 6&7. Even then, the number of clusters is increased in each iteration, the proposed scheme provides best results.

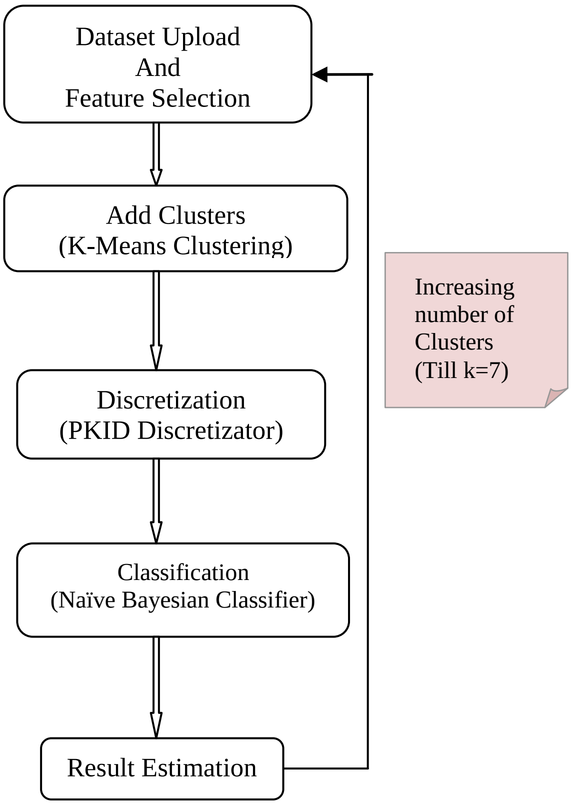

Figure 1.Steps to implement Proposed Scheme

4. Result and Discussion

The results of the proposed scheme include the calculation of Accuracy, Detection rate (DR) and False Alarm rate (FAR). The above said measures are calculated from the obtained rate of True positive (TP), True negative (TN), False positive (FP) and False negative (FN).

True positive:- Data instances rightly sampled as Attacks.

True negative:- Data instances rightly sampled as Non-Attacks.

False positive:- Right Data instances sampled as Attacks.

False negative:- Data instances wrongly sampled as Attacks.

Accuracy, Detection rate (DR) and False Alarm rate shall be calculated as,

Accuracy = Number of Data instances correctly sampled / Total number of Data instances.

Detection rate = TP / (TP+FN).

False Alarm Rate=FP/ (FP+TN).

The results obtained, when using training data set is weighted against the existing schemes and tabulated as

Training Dataset						
Approach Used	**TP + TN**	**FP + FN**	**Accuracy**	**Detection rate**	**False Alarm Rate**	**Time to build model (Sec)**
NBC - KMC When k=7	488329	5691	98.84	99.22	0.7728	7.7
KMC - PKID - NBC When k=7	493505	515	99.95	99.96	0.0208	0.59

Table. 1Performance Comparison of NBC - KMC & KMC - PKID - NBC using Training set

The results of Table 1 show that the proposed scheme KMC - PKID - NBC outperforms existing Data mining approach KMC - NBC. On comparing the results, when it is set k=7 it is clear that the proposed method provides a better Accuracy, Detection rate and False alarm rate than existing combo method of KMC - NBC. Apart from this, the proposed scheme is more time efficient as well.

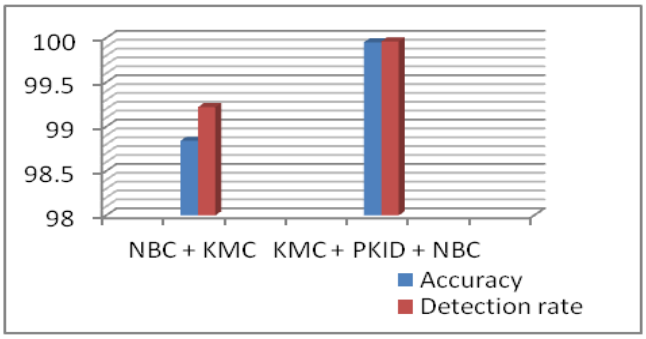

Figure 2. Performance of Existing and Proposed scheme in Training.

In the next stage of the experiment, the number of clusters is gradually increased from k=3, 4, 5, 6&7.

Training Dataset				
Approach Used	Accuracy	Detection rate	False Alarm Rate	Time to build model (Sec)
KMC - PKID - NBC When k=3	99.80	99.71	0.2060	0.75
KMC - PKID - NBC When k=4	99.81	99.76	0.1136	0.75
KMC - PKID - NBC When k=5	99.92	99.83	0.0226	0.70
KMC - PKID - NBC When k=6	99.94	99.87	0.0214	0.65
KMC - PKID - NBC When k=7	99.95	99.96	0.020	0.59

Table 1 Performance Evaluation of Proposed Scheme by increasing Number of Clusters

At each stage when the number of clusters is increased, we could visualize that the Accuracy and Detection rate increases optimally in Table 2. The False alarm rate and time taken to build model decreases as the number of cluster increases.

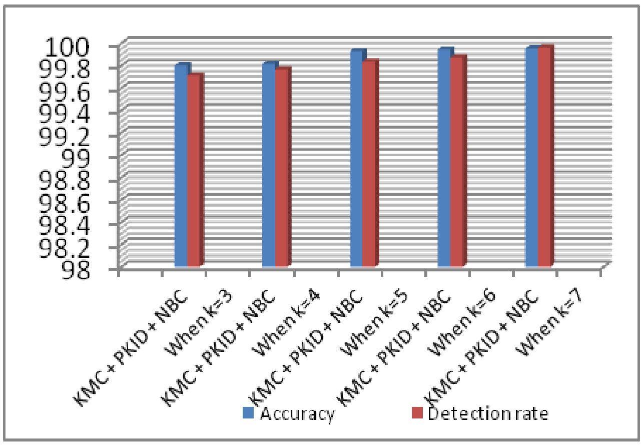

Figure 3. Performance of Proposed scheme when No. of Clusters increases in Training set

Test Dataset is prepared by reducing the number of instances in the training dataset in WEKA. 494020 instances of training dataset are reduced to 247010 instances in the test set. Then initially the hybrid method KMC - NBC is applied over the test dataset and the results are recorded. Now the proposed scheme KMC - PKID - NBC is applied over the test set, the results are tabulated in Table 3 and compared with the results of the existing method.

Test Dataset						
Approach Used	TP + TN	FP + FN	Accuracy	Detection rate	False Alarm Rate	Time to build model (Sec)
NBC - KMC When k=7	236589	10421	98.77	90.01	0.6745	4.01
KMC - PKID - NBC When k=7	246495	515	99.94	99.03	0.0318	0.31

Table 2 Performance Comparison of NBC - KMC & KMC - PKID - NBC using test set

The comparison apparently shows that proposed scheme gives optimal results for Accuracy and Detection rate. False alarm rate and time taken to build model are minimal in the proposed scheme.

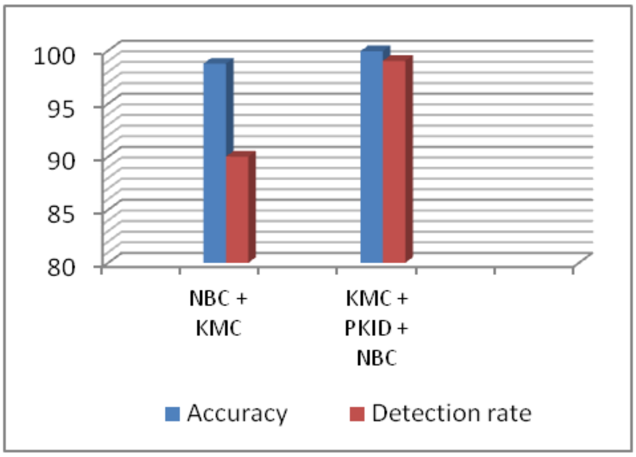

Figure 4. Performance of Existing and Proposed scheme in Testing.

Test Dataset				
Approach Used	Accuracy	Detection rate	False Alarm Rate	Time to build model (Sec)
KMC - PKID - NBC When k=3	99.57	98.68	0.2073	0.52
KMC - PKID - NBC When k=4	99.71	98.86	0.1820	0.50
KMC - PKID - NBC When k=5	99.77	98.93	0.1293	0.44
KMC - PKID - NBC When k=6	99.82	98.77	0.0996	0.33
KMC - PKID - NBC When k=7	99.94	99.03	0.0318	0.31

Table 3 Performance Evaluation of Proposed Scheme by increasing Number of Clusters.

The Proposed scheme is experimented by increasing the number clusters while using the test dataset. The Performance of the system improves as the number of clusters increased from k =3, 4, 5, 6 & 7 as shown in Table 4.

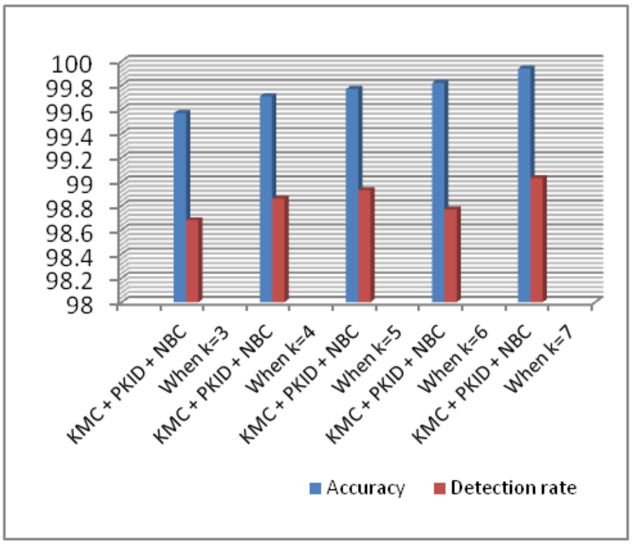

Figure 5. Performance of Proposed scheme when No. of Clusters increases in Training set

Finally, it is proved proposed scheme provides optimal results compared to existing hybrid approach. In terms of Average, the Accuracy is 99.86%, the Detection rate is 99.82% and False alarm rate is 0.0767, while tested between a number of clusters k= 3 to 7 during the training phase. The Average of Accuracy, Detection rate, and False alarm rate are 99.76%, 98.85% and 0.130 respectively during the testing phase.

5. Conclusion

In ABIDS, increasing Accuracy, Detection rate and reducing False alarm is an attention-grabbing task. In the proposed scheme, better results arrived for Accuracy, Detection rate and False alarm rate compared to existing methods. This performance improvement is achieved by adding PKID to the existing KMC - NBC Scheme.

6. References

1. G V Nadiammai. Effective approach towards intrusion detection sytem using data mining techniques, Egyptian Informatics Journal, 2014, Vol.15, p. 37- 50.

2. Agrawal Shikha and Jitendra Agrawal,. Survey on Anomaly Detection using Data Mining Techniques, Proceeding of International Conference on Knowledge Based and Intelligent Information and Engineering Systems, 2015, Vol. 60, P. 708-713.

3. Albert Bifet , Geoff Holmes, Richard Kirkby , Bernhard P fahringer. MOA: Massive Online Analysis, Journal of Machine Learning Research, 2010, Vol.11 (2010) P. 1601-1604.

4. Chandore, P.R. and Dr. P.N. Chatur. Hybrid Approach for Outlier Detection over Wireless Sensor Network Real Time Data, International Journal of Computer Science and Applications, 2013 Vol. 6(2) P. 89-95.

5. Arif Jamal Malik, Waseem Shahzad, Farrukh Aslam Khan. Network intrusion Detection using hybrid binary PSO and RF algorithm, Proceeding of Conference on Security and Communication Network, 2012.

6. P.Natesan, P.Balasubramanie. Multi Stage Filter Using Enhanced Adaboost for Network Intrusion Detection, International Journal of Network Security & Its Applications, 2012, Vol .3.

7. Preeti Aggarwal, Sudhir Kumar Sharma. Analysis of KDD Dataset Attributes - Class wise For Intrusion Detection. Proceeding of 3rd International Conference on Recent Trends in Computing, 2015.

8. Mrutyunjaya panda et.al. A hybrid intelligent approach for network intrusion detection, Procedia Engineering, 2012, Vol. 45, p 1-9.

9. Z Muda, W Yassin, MN Sulaiman, and NI Udzir. Intrusion detection based on k-means clustering and na¨ıve bayes classification, Proceeding of International IEEE Conference on Information Technology in Asia (CITA 11), 2011, p. 1–6.

10. Anusha Jayasimhan and Jayant Gadge. Anomaly detection using a clustering technique, International Journal of Applied Information Systems (IJAIS), 2012, p. 2249–0868.

11. Jose F Nieves and Yu Cathy Jiao. Data clustering for anomaly detection in network intrusion detection, Journal on Research Alliance in Math and Science, 2009, p.1–12.

12. Reda M Elbasiony, Elsayed A Sallam, Tarek E Eltobely, and Mahmoud M Fahmy, A hybrid network intrusion detection framework based on random forests and weighted k-means, Ain Shams Engineering Journal, 2013, vol. 4(4), p. 753–762.

13. Hatim Mohammed Tahir, Abas Md said and Hayani osman. Improving K - Means Clustering using Discretization Techniques in IDS, Proceeding of IEEE 3rd International Conference on Computer And Information Sciences (ICCOINS), 2016.

14. V. Bolon-Canedo, N, Sanchez-Marofio and A. Alonso-Betanzos. A Combination of Discretization and Filter Methods for Improving Classification Performance in KDD Cup 99 Dataset, Proceedings of International Joint Conference on Neural Networks, 2009.

15. Mohammad Khubeb Siddiqui and Shams Naahid,. Analysis of KDD CUP 99 Dataset using Clustering based Data Mining, International Journal of Database Theory and Application, 2013 Vol. 6(5) , p.23-34.

16. Nasir Majeed Mir, Sarfraz Khan, Muheet Ahmed Butt and Majid Zaman. An Experimental Evaluation

of Bayesian Classifiers Applied to Intrusion Detection, Indian Journal of Science and Technology, 2016, Vol 9(12).

17. Y. Yang and G.I. Webb, Proportional k-Interval Discretization for Naive-Bayes Classifiers Proceedings of the 12th European Conference on Machine Learning, Springer-Verlag, 2001, p. 564-575.