Hybrid System for Anomaly Intrusion Detection using Enhanced K Strange Points Clustering and Naïve Bayes Classifier

Mario Dias[1] Valen D'souza[2] Abhijeet Bhangle[3] Nishad Dangui[4] Kedar Sawant[5]
[1,2,3,4]BE Student
[1,2,3,4,5]Department of Computer Engineering
[1,2,3,4,5]Goa University India

Abstract— An intrusion detection system (IDS) monitors the system activity and tracks abnormal activity patterns thus ensuring system and file integrity. Proposed research is based on the combination of clustering and classification techniques which are used in Hybrid IDS. Clustering is a technique which groups similar data objects into a single cluster. Classification is a technique which predicts a new class for the test object. Proposed IDS works on NSL-KDD dataset. NSL-KDD dataset is a revised version of KDD99 dataset. First, clustering is performed using Enhanced K Strange Points clustering algorithm on NSL KDD consisting of Denial of Service (DoS) attacks. This output is given to the Naive bayes classifier, which classifies the dataset into 6 types of DoS attacks. The results of the proposed system are then compared with existing IDS which uses Kmeans clustering and KNN classifier. The proposed concept aims at improving the detection rates and classification rates of existing Intrusion Detection System (IDS) by using the new approach. It also focuses on reducing the false positive rates compared to the existing system.

Key words: NSL-KDD Dataset, Naïve Bayes Classifier

I. INTRODUCTION

Data mining techniques have been widely used in Analysis of Financial Data, Telecommunication Industry, Biological Data Analysis, and Other Applications. As more and more systems are being connected over the internet almost every day, the task of securing the data on the system also increases, and if the security of the system is compromised even for a small amount of time it may result in the loss and manipulation of data. Although most organisations are using firewalls having inert configurations that block attacks based on some attributes, it does not provide security in preventing and detecting attacks. As a result, an IDS is used in order to detect all possible kinds of attacks and take appropriate actions to counter them. Computers over the internet are vulnerable to snooping, hacking and other type of attacks. The shielding devices used by some organisations such as firewalls, do not guarantee complete protection against these attacks. However an IDS provides better detection of attacks as compared to firewall. Upon knowing the type of intrusion suitable action can be taken to trace the attacker or the source of attack [1]. A firewall could be thought of as a security guard at the gate whereas an intrusion detection system would be a surveillance camera which will detect all the threats that might pass through the gate. In the proposed concept we have applied Data Mining techniques in anomaly detection field of intrusion detection. The main objective of the paper is to improve the accuracy of intrusion detection system.

II. RELATED WORK

Reference [1] proposed a concept for intrusion detection system, which will enhance efficiency as compare existing intrusion detection system. The proposed concept is using data mining techniques. Most of the data mining techniques like, clustering, association rule mining and classification have been functional on intrusion detection, where pattern mining and classification is the significant technique. In order to maintain the soaring detection rate and accuracy even as at the same time to decrease the false alarm rate, the proposed technique is the combination of three learning approaches. In this proposed technique, the author used K-Means (KM) clustering, K-Nearest Neighbor (KNN) algorithm and Decision Table Majority (DTM) rule based approach. First apply the k-means algorithm to the given dataset to split the data records into normal cluster and anomalous clusters. It specifies the number of clusters as five to the k-means and clusters the records in the dataset into normal cluster and anomalous clusters. The anomalous clusters are User to Root (U2R), Remote to local (R2L), PROBE, and DoS.

III. PROPOSED CONCEPT

A. Proposed Architecture

The proposed concept of the intrusion detection system is shown in figure 1 which enhances the efficiency and performance compared to the existing intrusion detection system. The input that is given to the proposed intrusion detection system is the NSL-KDD dataset. Data preprocessing techniques are then applied on the dataset to reduce the dimensionality and to remove the redundant records present in the dataset. The processed data is given to the clustering and classification data mining techniques. The NSL dataset is an improved version of the traditional KDD CUP 99 dataset. The redundant entries contained in KDD CUP 99 are eliminated in NSL KDD thus making it more efficient for the data mining tasks [3]. Only those attributes and records that are relevant to Denial of Service attacks are selected [4]. The attributes chosen are protocol_type, service, loggen_in, dst_host_count, dst_host_srv_count. The data entries are then normalized using min-max normalization technique.

The preprocessed dataset is then given to the data mining algorithms. These include the clustering and classification algorithms. Clustering is carried out using Enhanced K Strange points algorithm and Classification is carried out using Naïve Bayes classifier. Hence the accuracy is improved using a hybrid approach. The outputs are then compared with the traditional IDS which uses K means and K-Nearest Neighbours (KNN) in combination [1].

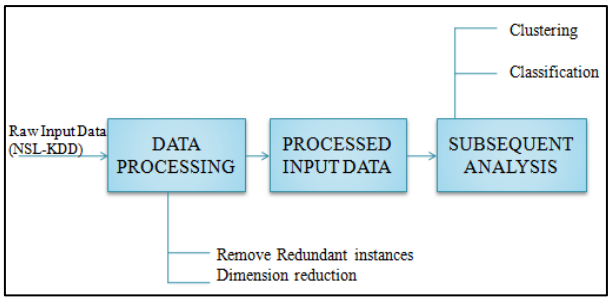

Fig. 1: Block Diagram of KDD Process

In the proposed method, we first apply Enhanced K strange points clustering to the preprocessed test set by specifying the number of clusters [5]. In this context two clusters will be obtained which are normal and abnormal. The abnormal cluster contains different types of DoS attacks. Initially the algorithm finds two points from the dataset which are at maximum distances to each other that is K_{min} and K_{max}. These become the cluster centroids for the two clusters respectively. Further, closeness of every data point from the dataset is then calculated with respect to K_{min} and K_{max} and accordingly they get assigned to these cluster. After all the points are assigned, two clusters are formed. Since the number of normal points are significantly large as compared to the anomalies, the cluster with the smaller size will be the abnormal cluster. This cluster is then given as a test input to the Naïve Bayes classifier. Naïve Bayes is a supervised learning algorithm which assumes a probabilistic model [7]. Supervised learning means that the algorithm is first trained using a training set which helps it to predict the class label of an unknown test sample. The training dataset consists of 22,683 records while the test dataset contains 12,900 records. Naïve Bayes is used to classify different types of DoS attacks in the anomalous cluster obtained as a result of clustering. It starts by calculating the prior probability P(c) for each attack (class) from the training set. This is then followed by estimating the conditional probability P(x/c) which is the likelihood probability of a given class. Using these probabilities it calculates the posterior probability P(c/x). The posterior probabilities are then compared and the class having highest posterior probability is assigned to the test sample. The output of this approach is then compared with the traditional approach to find its accuracy.

1) Clustering

Clustering is a technique which groups objects having similar properties. It is an unsupervised technique because class labels are unknown. Each cluster contains objects which are similar to each other and different from those in other clusters. Clusters are differentiated based on the distance measure. Euclidean Distance measure is most commonly used to find distance between two data points. Clustering is efficient when difference between two cluster is more. Some data mining clustering techniques are: K-Means, Bisecting Kmeans, K Strange, Hierarchical Clustering etc. In the proposed work we have applied Enhanced K Strange points clustering algorithm on NSL-KDD dataset [3].

2) Classification

Classification is a method by which class labels are assigned to the test records. Classification algorithm is first trained with a training data set. The main task of classification is to correctly predict the target class for each record in the data. Classification algorithm uses supervised learning approach. To assign a class label to a test record, the classification algorithm builds a classification model using the training data set and then predicts the class label for the given test record. In our proposed concept we have used probabilistic classification algorithm that is Naïve Bayes classifier which is applied to identify test record generated by clustering algorithm.

B. *Proposed Algorithms*

1) Enhanced K Strange Points Clustering Algorithm

– Input: Pre-processed NSL-KDD dataset with n objects $D=\{D_1,D_2,\ldots D_n\}$
– Output: Set of K=2 clusters

1) Step 1: Find K_{min}, i.e. the Minimum of dataset based on the Euclidean Distance Measure.
2) Step2: Find a Point K_{max}, which is at maximum distance from K_{min}.
3) Step3: Locate a third point S which is farthest from K_{min} and K_{max} and also equidistance from each other
4) Step 4**:** if $(D(K_{min},S) == D(K_{max},S))$

$$K_{str} = S$$

else if$(D(K_{min},S) < D(K_{max},S))$

$$Kstr = Kstr_{prv} + X_m \quad [\frac{|Kmax - Kstrprv|}{K-1}]$$

else if$(D(K_{min},S) > D(K_{max},S))$

$$Kstr = K_{min} + X_m \left[\frac{|Kstrprv - Kmax|}{K-1}\right]$$

Where K is the number of clusters

$$X_m = X_1, X_2, \ldots, Xk_{-2}$$

$Kstr_{prv}$ = Uncorrected values of S

Kstr = Corrected values of S

5) Step 5: Repeat the above procedure until we locate K strange points.
6) Step 6: Assign the remaining Points in the dataset into clusters formed by these non-collinear K Strange points.
7) Step7: Output K clusters.

2) Naive Bayes Classification Algorithm Steps

– Convert the data set into a frequency table.
– Create Likelihood table by finding the probabilities of the attributes.
– Use Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.

There are two types of probabilities

– Posterior Probability [P(c/x)]
– Prior Probability [P(c)]

Where x is data tuple and c is some hypothesis.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \qquad (1)$$

$$P(c|x) = P(x_1|c)*P(x_2|c)*\ldots P(x_n|c)*P(c) \qquad (2)$$

P(c|x) is the posterior probability of class (c,target) given predictor (x,attribute) (3.1).
P(c) is the prior probability of class.
P(x|c) is likelihood probability of predictor class.
P(x) is the prior probability of predictor.

IV. Result Analysis

It is observed that the new hybrid approach (Enhanced K Strange Points Clustering + Naïve Bayes Classifier) is more accurate and faster in terms of both clustering and classification when compared with the old approach (K means Clustering + K nearest neighbor Classifier). The run time, detection rates and classification rates are significantly

improved in the new approach. The false positive rate of the traditional technique is more as compared to the new technique indicating that the new technique has a better classification rate over the old one.

A. K means Clustering + K nearest neighbor Classifier

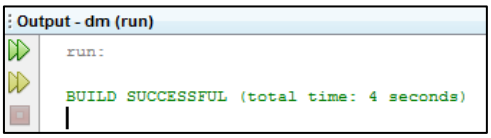

Fig. 2: Run time for K means

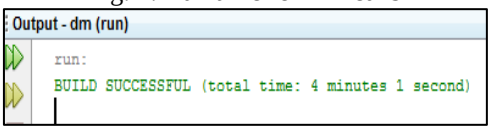

Fig. 3: Run time for K nearest neighbor

B. Enhanced K Strange Points Clustering + Naïve Bayes Classifier

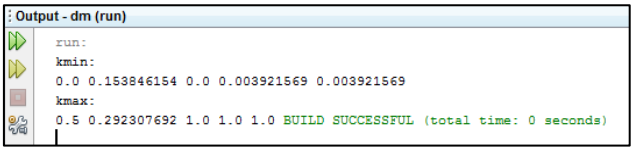

Fig 4. Run time for Enhanced K Strange

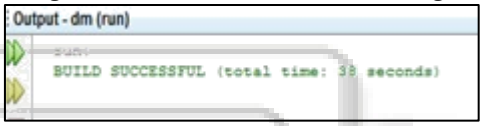

Fig 5. Run time for Naïve Bayes

Hence from the above results it is concluded that the overall run time of new approach is better and is thus faster in terms of both clustering and classification.

C. Quantitative Performance Analysis

The detection rates DR (3), false positive rates FPR (4) and the classification rates CR (5) are calculated as follows:

$$\text{Detection rate} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{False Positive rate} = \frac{FP}{FP+FN} \quad (4)$$

$$\text{Classification rate} = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

Where,

TP = True Positives and is the number of anomalous records that are correctly classified as intrusion.

TN = True Negatives and is the number of legitimate records that are not classified as intrusion.

FP = False Positives and is the number of records that are incorrectly classified as intrusion.

FN = False Negatives and is the number of records that are incorrectly classified as legitimate activities.

Based on the above metrics the detection rates, false positive rates and classification rates are calculated and shown in the table 1.

Comparison Criteria	Enhanced K Strange +Naïve Bayes Neighbor	K Means + K Nearest
Detection	82.48% Rate	67.10%
False Positive	5.20% Rate	11.08%
Classification	92.03% Rate	84.01%

Table 1: Detection Rates, False Positive Rates and Classification Rates

The detection rate of new method is 84.48% which is much more than the old method with only 67.1%. The new method has an excellent classification rate (92.03%) compared to the old one. Furthermore the false positive rate of the new method also is reduced significantly to 5.2% with new approach.

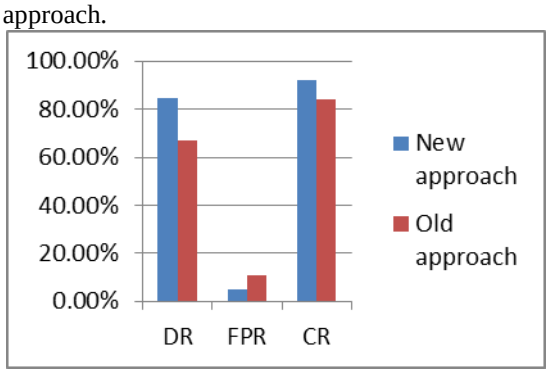

Fig. 2: Graphical representation of DR, FPR and CR

The above graph shows comparison of Detection rate, False positive rate and classification rate between existing IDS and Proposed IDS.

V. Conclusion

A hybrid approach can find more accurate probability of normal and abnormal packets. When compared with traditional system the hybrid system, yields more faster and efficient results. The detection rates and classification rates are significantly improved over the old approach. The use of Enhanced K Strange Clustering and Naïve Bayes Classifier also reduces the false positive rates thus classifying the samples more accurately. One of the limitations is that this model works only with limited number of attributes to yield better results. Further work can be done to improve the limitations. Decision table majority rule based approach could be applied to make classification more efficient.

References

[1] Vasim Iqbal Memon and Gajendra Singh Chandel, "A Design and Implementation of New Hybrid Sytem for Anomaly Intrusion Detection System to Improve Efficiency", Int. Journal of Engineering Research and Applications, ISSN: 2248-9622, Vol. 4, Issue 5 (Version 1), May 2014, pp.01-07.

[2] Sundus Juma, Zaiton Muda and Warusia Yassin, "Reducing False Alarm Using Hybrid Intrusion Detection Based on X-Means Clustering And Random Forest Classification", Journal of Theoretical and Applied Information Technology, ISSN: 1992-8645, Vol. 68 No. 2, pp.01-06.

[3] L.Dhanabal and Dr. S.P. Shantharajah, "A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms", International Journal of Advanced Research in Computer and Communication Engineering , ISSN 2319-5940, Vol 4, Issue 6, June 2015, pp.01-07.

[4] Wei Wang, Sylvain Gombault and Thomas Guyet, " Towards fast Detecting intrusion using key attributes of network traffic", Dream Team, IRISA, France.

[5] Terence Johnson and Dr. Santosh Kumar Singh, "Enhanced K strange Points Clustering Algorithm" 2015

International Conference on Emerging Information Technology and Engineering Solutions.

[6] Chai, K.; H. T. Hn, H. L. Chieu; "Bayesian Online Classifiers for Text Classification and Filtering", Proceedings of the 25th annual international ACM SIGIR conference on Research and Development in Information Retrieval, August 2002, pp. 97-104.

[7] Sayali D. Jadhav and H. P. Channe, "Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques", International Journal of Science and Research, ISSN (Online): 2319-7064, 2014, pp.01-04

