ISSN 2349-0292
Impact Factor 3.802
GLOBAL JOURNAL OF ADVANCED ENGINEERING TECHNOLOGIES AND SCIENCES

A NOVEL APPROACH FOR IMAGE IMPAINTING USING ADAPTIVE SVM

Kulbir Kaur Sandhu
* Assistant Professor, Department of Computer Science
Baba Farid College, Bathinda

DOI: 10.5281/zenodo.4568089

ABSTRACT

Data security is critical for most businesses and even home computer users. As, all the sensitive and important information is manipulated online. Therefore, it is essential to protect the data from intruders. We need to find the best ways to protect our system from different types of attacks.Intrusions refer to the network attacks against vulnerable services, data-driven attacks on applications, host-based attacks like privilege escalation, unauthorized logins and access to sensitive files, or malware like viruses, worms and Trojan horses. The problem in current IDS is the high false positive rate and low detection rate.As far as system complexity is concern clustering algorithm having higher space and time complexity.Moreover taking consider all the above mentioned limitation there is requirement of novel data mining technique which can give better accuracy, system should make automated self-configured self-optimized and having least complexity and very quick response time as we are taking into consideration the IDS system which is very important aspect of cloud computing, also it should not make system heavier, does not affect the QOS, and so on, hierarchical clustering seems to be better solution for all the problem also in future we try to add some classification technique in this system so that would make fully automated system.

KEYWORDS: Image Impainting; SVM; Accuracy; False positive Ration, False Negative ration.

INTRODUCTION

Intrusions refer to the network attacks against vulnerable services, data-driven attacks on applications, host-based attacks like privilege escalation, unauthorized logins and access to sensitive files, or malware like viruses, worms and trojan horses. It is an activity or attempt which affects the integrity, confidentiality & availability of a resource. Data integrity is data must not be changed in transit, and steps must be taken to ensure that data cannot be altered by unauthorized people. Data confidentiality is roughly equivalent to privacy. Measures undertaken to ensure confidentiality are designed to prevent sensitive information from reaching the wrong people, while making sure that the right people can in fact get it. Access must be restricted to those authorised to view the data. Data availability is guarantee of reliable access to the information by authorised people. The network should be tough to Denial of Service attacks [10].One of the main challenges in the security management of large-scale high-speed networks is the detection of suspicious anomalies in network traffic patterns due to Distributed Denial of Service (DDoS) attacks or worm propagation [1] [2] a secure network must provide the following:

- **Data confidentiality:** Data that are being transferred through the network should be accessible only to those that have been properly authorized.
- **Data integrity:** Data should maintain their integrity from the moment they are transmitted to the moment they are actually received. No corruption or data loss is accepted either from random events or malicious activity.
- **Data availability:** The network should be resilient to Denial of Service attacks.

The first threat for a computer network system was realized in 1988 when 23-year old Robert Morris launched the first worm, which overid over 6000 PCs of the ARPANET network. On February 7th, 2000 the first DoS attacks of great volume where launched, targeting the computer systems of large companies like Yahoo!, eBay, Amazon, CNN, ZDnet and Dadet. More details on these attacks can be found at [3]. These threats and others that are likely to appear in the future have lead to the design and development of Intrusion Detection Systems. According to webopedia [4] an intrusion detection system (IDS) inspects all inbound and outbound network activity and identifies suspicious patterns that may indicate a network or system attack from someone attempting to break into or compromise a system.

ISSN 2349-0292
Impact Factor 3.802

RELATED WORK

Mahbod T. And E. Bagheri et al. [20], designs an intrusion detection system which is divided into two parts. It contains two data mining techniques: association rules and clustering. The first step is association phase in which frequent item set are produced by Apriori algorithm. The second step is a clustering phase in which clusters are created by K-Means. This technique uses the standard KDD99 intrusion detection contest dataset. The authors define that this IDS technique can detect the attacks/intrusions and classifies them into different categories: U2R (User to Root), probe, R2L (Remote to Local) and Daniel of Service (DoS).The execution time is 120 ms(approx) that an approach takes to produce results. The CPU uses is 74%(approx) and memory usage is 54%(approx).

K. S. Desale, Chandrakant N. K. et. al , [21], creates a cloud intrusion detection system to find the masquerader attacks. Masquerader is an attacker who masquerades a lawful user in the wake of abusing the user account. In that case, the firewalls or validation protocols are useless because, subsequent to logging as a legitimate user, an attacker can misuse any user benefit. The authors of this paper used association rule mining technique of data mining. Apriori-TID algorithm has been used because it is suitable as data mining is used for the transactional database. CIDD dataset has been used to evaluate the results. Also, the comparison has been done between different datasets: DARPA(1998), KDD(1999), SSENet(2011), CIDD(2012).

Solane Duque, Dr.Mohd. Nizam bin Omar [22], uses the classification techniques of data mining to detect intrusions. The authors compare the four classification algorithms and apply it to NSL-KDD datasets and compare their results. The proposed technique has been applied to the streaming data(which changes with time & update its value). The proposed system is only intended to improve intrusion detection efficiency, not to prevent intruders. The comparison of four algorithms has been done. These are Naïve Bayes, Hoeffding Tree, Accuracy weighted ensemble, Accuracy ensemble. The experimental results show that accuracy weighted ensembles gives higher accuracy than other classifiers but takes little bit more time whereas hoeffding tree gives less accuracy in less time.

Patel and Santosh Sammarvar et al.,[23], compares the Statistical and data mining based decision tree techniques to check the efficiency of IDS classifier. Models are verified with many error measures with different partitions of data. Results reveal that decision tree based technique C5.0 is performing better than statistical technique, these techniques applied on NSL-KDD data set. A comparative study shows that C5.0 outperformance SVM in terms of accuracy, sensitivity and specificity error measures.

Liang Hu and Taihui Li et al [24], adopted clustering algorithms, the K-means algorithm and the Fuzzy C Mean (FCM) algorithm, to identify false alerts, to reduce invalid alerts and to purify alerts for a better analysis. Furthermore, they introduced an intrusion detection framework, and tested the validity and feasibility of false positive elimination in intrusion detection. Also, they defined three evaluation indexes: the elimination rate, the false elimination rate and the miss elimination rate. Set up the false positive dataset based on snort and DARPA2000 LLDOS 1.0, and then compared the effect of those two algorithms. Lastly, experiments showed the validity and feasibility of using the clustering algorithms to eliminate false positives in IDS.

Zhengjie Li and Yongzhong Li et al [25], defines the PSO-KM algorithm combines particle swarm optimization algorithm with the traditional K-means clustering algorithm. Simulation experiment on data sets KDD CUP 99 shows that PSO-KM algorithm is effective method when dealing with large data sets. Experimental results show that detection rate of PSO-KM is improved for detecting known attacks and unknown attacks. At the same time, false detection rate greatly reduces. It improves application value of K-means clustering algorithm in the field of intrusion detection.

METHODOLOGY

This section discusses about the proposed work and implementation of the proposed hybrid approach. Here the discussion is about the architecture of the proposed approach and different components used during the implementation.

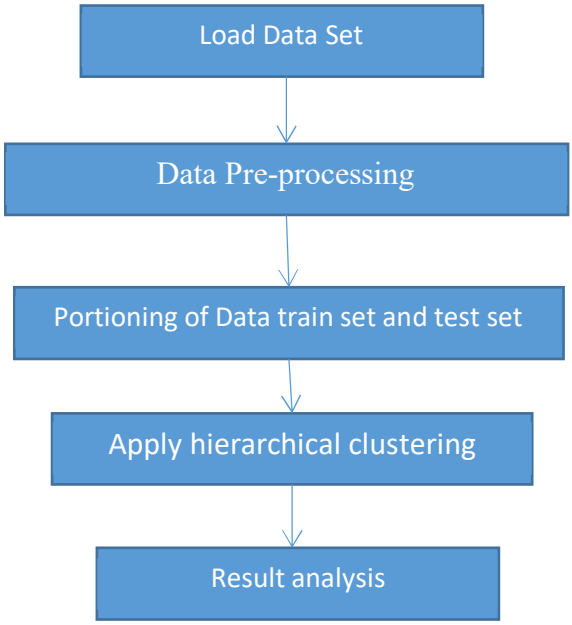

Fig 1.1 Flow chart of proposed method

A. KDD CUP 1999 Dataset

It is prepared and managed by MIT, Lincoln Labs by DARPA Intrusion Detection Evaluation Program [35], after capturing nine weeks of raw TCP dump data for LAN simulating a typical U.S. Air Force LAB. They operated the LAN as if it were a true Air Force environment, but peppered it with multiple attacks. The raw training data contains four gigabytes of compressed binary TCP dump data from seven weeks of network traffic by processed into about five million connection records. Similarly, the test data yielded around two million connections records which are captured last two weeks of the experiment. This is the data set used for The Third International Knowledge Discovery and Data Mining Tools Competition, which was held in conjunction with KDD-99 The Fifth International Conference on Knowledge Discovery and Data Mining. The competition task was to build a network intrusion detector to detect attack categories i.e. DOS, PROBE, R2L and U2R [24]. Attacks fall into four main categories [36]:

1. **Denial of Service Attack (DoS):**is an attack in which the attacker makes some computing or memory resource too busy or too full to handle legitimate requests, or denies legitimate users access to a machine.
2. **User to Root Attack (U2R):** is a class of exploit in which the attacker starts out with access to a normal user account on the system (perhaps gained by sniffing passwords, a dictionary attack, or social engineering) and is able to exploit some vulnerability to gain root access to the system.
3. **Remote to Local Attack (R2L):** occurs when an attacker who has the ability to send packets to a machine over a network but who does not have an account on that machine exploits some vulnerability to gain local access as a user of that machine.
4. **Probing Attack:** is an attempt to gather information about a network of computers for the apparent purpose of circumventing its security controls.

KDD (Knowledge Data Discovery) dataset consists of approximately 4,900,000 single connection vectors each of which contains 41 features and is labeled as either normal or as an attack, with exactly one specific attack type. It is important to note that the test data is not from the same probability distribution as the training data, and it includes specific attack types not in the training data which make the task more realistic. Some intrusion experts believe that most novel attacks are variants of known attacks and the signature of known attacks can be sufficient to catch novel variants. The datasets contain a total number of 24 training attack types, with an additional 14 types in the test data only.

KDD'99 features can be classified into three groups [38]:

1. **Basic features:** this category encapsulates all the attributes that can be extracted from a TCP/IP connection. Most of these features leading to an implicit delay in detection.

ISSN 2349-0292
Impact Factor 3.802

2. **Traffic features:** this category includes features that are computed with respect to a window interval and is divided into two groups:
   - **"same host" features:** examine only the connections in the past 2 seconds that have the same destination host as the current connection, and calculate statistics related to protocol behavior, service, etc.
   - **"same service" features:** examine only the connections in the past 2 seconds that have the same service as the current connection.
3. **Content features:** unlike most of the DoS and Probing attacks, the R2L and U2R attacks don't have any intrusion frequent sequential patterns. This is because the DoS and Probing attacks involve many connections to some host(s) in a very short period of time; however the R2L and U2R attacks are embedded in the data portions of the packets, and normally involves only a single connection. To detect these kinds of attacks, we need some features to be able to look for suspicious behavior in the data portion, e.g., number of failed login attempts. These features are called content features.

B. K-Mean Clustering

K-Mean Clustering is a technique which groups the similar data based on the behavior. K- Mean is an unsupervised task, i.e. data doesn't specify what we are trying to learn. Many researchers use K-Mean clustering in the hybrid approaches to detect the anomalous data. In proposed system, K-Mean clustering works as a pre-classification phase which groups objects based on the feature value into number of disjoint clusters.
Algorithmic steps are:

Step 1 : Choose the number of centroids objects from dataset as the initial centroids.
Step 2 : Then, calculate the Euclidean distance between each data point and the centroids.
Step 3 : If the data point is closest to the centroid, then leave it and do not make any change in its position. But if the data point is not closest to the centroid, then move it to its closest one.
Step 4 : Recalculate the centroid of both modified clusters.
Step 5 : Repeat step 3 until we get the steady centroids.

In other words, its objective is to find:

$$M = \sum_{a=1}^{k} \sum_{b=1}^{n} \quad d_{ab}\ (x_b, y_a)$$

Where, $d_{ab}\ (x_b,\ y_b)$ is an eculidean distance between the data point x_b and the centroid y_a.
Euclidean distance is:

$$d(x_b, y_a) = \parallel\ x_b\ - y_a\ \parallel$$

C. Adaptive SVM

Adaptive SVM (Support Vector Machine) aims to adapt two or more classifiers of any kind to new datasets [41]. The problem is how to select the best classifier for adaptation. The solution to this problem is to select the classifier with best parameters after estimating the performance of each classifier on the sparsely labeled dataset. The general problem of binary classification task is considered on original dataset D^o, which made up of majority of unlabeled instances D_u^o and limited number of labeled instances D_l^o ,therefore, the original dataset is:

$$D^o = D_l^o\ \cup D_u^o$$

There are one or more subordinate datasets D_1^s……,D_M^s which is different from the original dataset. The subordinate classifier f_k is used to train each of the subordinate datasets D_k^s. We have,

$$D_l^o = \{(x_i, y_i)\}_{i=1}^{N}$$

Where, x_i is the i_{th} data vector and $y_i \in \{-1, +1\}$is its binary label. Data vector x always include a constant 1 as its first element, such that, $x_i \in R^{d+1}$, where d is the number of features. There exist multiple subordinate datasets as D_1^s, $D_M^s$with $D_k^s = \{(x_i^k, y_i^k)\}_{i=1}^{N_k}$, where $x_i^k \in R^{d+1}$ and $y_i^k \in \{-1, +1\}$. The subordinate dataset description is different from the original dataset. The subordinate classifier $f_k^s(x)$ has been trained from each subordinate dataset D_k^s , which gives us the result of prediction of data label through the sign of its decision function, i.e. $\hat{y} = f^x(x)$.

The traditional SVM trains the $f(x)$ from the labeled dataset D_l^o . Adaptive SVM is used to adapt a combination of multiple existing classifiers $f_1^s(x)$ … … , $f_M^s(x)$ to the new classifier. The traditional SVM trains the $f(x)$ from

the labeled dataset D_l^o. The decision boundary is determined by the kernel function$(x, x') = \langle \Phi(x), \Phi(x') \rangle$, where $\Phi(x)$ is a feature vector. The kernel function is the inner product of two projected feature vectors. Delta function is used in Adaptive SVM in the form of $\Delta f(x) = w^T \Phi(x)$ on the basis of $f^s(x)$:

$$f(x) = f^s(x) + \Delta f(x) = f^s(x) + w^T \Phi(x) \qquad (4.1)$$

Where, w are the parameters predicted from the labeled dataD_l^o. As defined earlier, the objective is to make a group of subordinate classifiers and adapt this group to new classifier$f(x)$. By using Eq.(3), the adapted classifier's form is:

$$f(x) = \sum_{k=1}^{M} t_k \, f_k^s(x) + \Delta f(x) = \sum_{k=1}^{M} t_k \, f_k^s(x) + w^T \Phi(x) (4.2)$$

Where, $t_k \in (0,1)$ is the weight of each subordinate classifier $f_k^s(x)$, which sums to one as $\sum_{k=1}^{M} t_k = 1$.
Objective Function: To learn the parameter w of delta function Δf, the function is:

$$\min_w \frac{1}{2} || w ||^2 + C \sum_{i=1}^{N} \xi_i \qquad (4.3)$$

Such that, $\xi_i \geq 0$

$$y_i f^s(x_i) + y_i \, w^T \Phi(x_i) \geq 1 - \xi_i, \forall (x_i, y_i) \in \mathrm{D_l^o}$$

where, $\sum_{i=1}^{N} \xi_i$ measures total classification error of adapted classifier $f(x)$ and $|| w ||^2$is a regulization term that is inversely related to margin between training examples of two classes. The cost factor C in A-SVMs balances the contribution between the subordinate classifier (through the regularizer) and the training examples. The larger C is, the smaller the influence of the auxiliary classifier is.

The objective function of Adaptive SVM model which is able to leverage multiple subordinate classifiers $(f_1^s(x) \ldots \ldots \ldots f_M^s(x))$, is achieved by replacing $f_1^s(x)$ with $\sum_{k=1}^{M} t_k f_k^s(x_i)$ in Eq.(4.3):

$$\min_w \frac{1}{2} || w ||^2 + C \sum_{i=1}^{N} \xi_i$$

Such that, $\xi_i \geq 0$,

$$y_i \sum_{k=1}^{M} t_k f_k^s(x^i) + y_i \, w^T \Phi(x_i) \geq 1 - \xi_i, \forall (x_i, y_i) \in \mathrm{D_l^o}$$

D. Hybrid Approach

The best possible high detection rate and low false alarm rate can be achieved by using hybrid approaches for IDS. As later discussed that K-Mean algorithm has some disadvantages, but by combine it with other techniques, played an important role in this field. In proposed system, K-Mean is hybrid with Adaptive SVM. K-Mean is worked as pre classification phase in proposed approach. It divided the train data into meaningful clusters so that the members from the same cluster have similar properties, and the members from different clusters are different from each other. After that, the clustered data is divided into further two datasets. The main advantage of Adaptive SVM is it automatically gets the best parameters. According to these parameters, the testing is performed on whole dataset by applying Adaptive SVM on whole dataset again, therefore, the detection rate automatically increases. Following are the steps of hybrid approach:

Step 1: Load the KDD dataset containing normal data and attack data.
Step 2: Pre classification is performed by K-Mean algorithm based on training dataset.
Step 3: K- Mean divides the training data into two clusters, one is the cluster of normal data and other is the cluster of anomaly data.
Step 4: Split the clustered data into Train and Test data
Step 5: Now Split the Train data in Train_train and Train_test set
Step 6: Perform the Adaptive SVM Classification with random parameter over Train_train and Train_test set.
Step 7: Get the best parameters.
Step 8: Again perform Adaptive SVM Classification with full dataset.
Step 9: Get the output parameters i.e Efficiency rate, False Positive rate, False Negative rate.

ISSN 2349-0292
Impact Factor 3.802

RESULT AND DISCUSSION

To evaluate the hybrid approach it is compare with the K- Mean and Adaptive SVM as an individual approaches. The output parameters are accuracy, False Positive rate, False Negative rate.

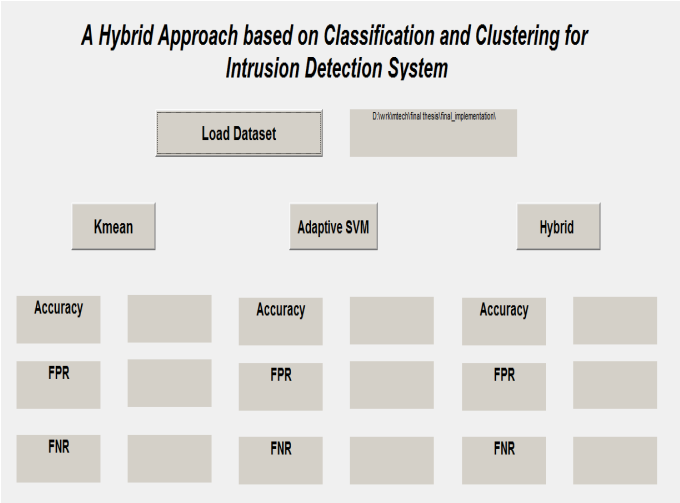

Figure 1.2: Adding a path of Train dataset

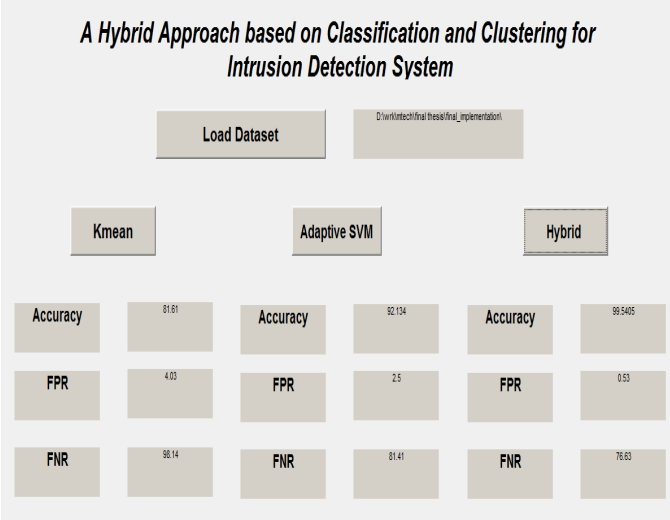

Figure 1.3: Values of Output Parameters of different algorithms after implementation

A. Accuracy

This refers to the ability of the classifiers to correctly measure the intrusions from the training dataset. This is defined as the ratio of correctly classified data to the total classified data.

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

where, TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

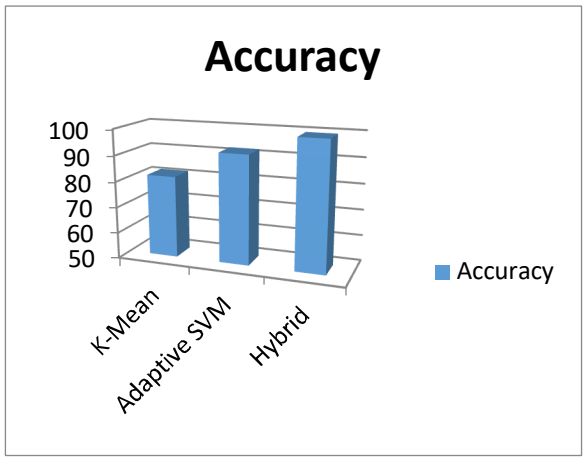

Figure 1.4: Accuracy Analysis of three Approaches

In above experimentation, the result shows the average performance of 10-Cross validation. These models are compared on the basis of each individual fold or rounds. Experimentation result shows that the hybrid approach is more accurate as compare to other approaches. The accuracy of a proposed model is near about 99.54% as shown in the Figure 5.3. This proposed method performs better than individual performances of K-Mean and Adaptive SVM. To measure the robustness and effectiveness of any model, comparison of parameters like False Positive Rate and False Negative Rate is computed and the performance of different models on the above parameters is evaluated.

B. False Positive Ratio

This is one of the main parameters to find out the effectiveness of various models and also the major concern while network setup. A normal data is considered as abnormal or attack type data. It is defined as:

$$\text{FPR} = \frac{FP}{FP+TN}$$

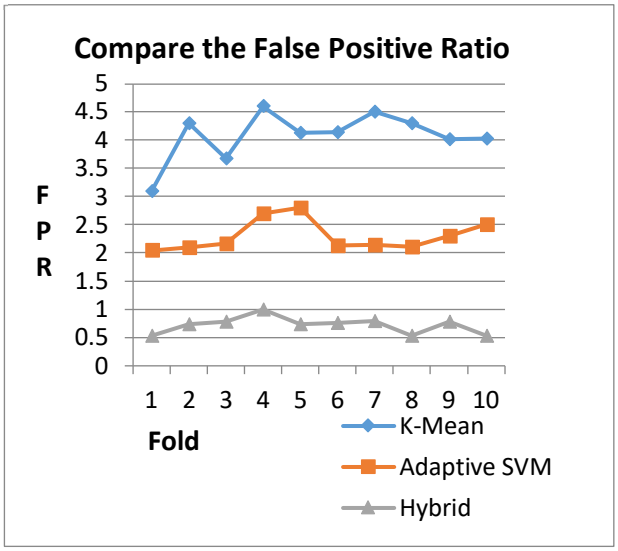

Figure 1.5: False Positive Ratio analyses of three approaches under 10 folds

C. False Negative Ratio

This is one of the main parameters used to describe a network intrusion device's inability to detect true security events under certain circumstances. An abnormal data is not detected and considered as normal data. It is defined as:

$$\text{FNR} = \frac{FP}{FN+TN}$$

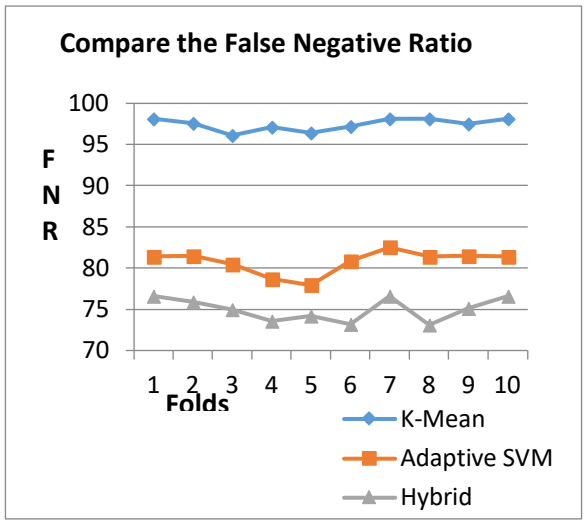

Figure 1.6: False Negative Ratio analyses of three approaches under 10 folds

From all the above experimentation results, it is shown that after applying the evaluation parameters, hybrid approach found to be the best approach in all scenarios. By applying the hybrid approach of data mining models on the dataset, the accuracy is improved for anomaly detection. So the main objective to improve the accuracy to detect the anomalous data has been met.

CONCLUSION

It is important to secure the data and system from intruders. Therefore, IDS plays an important role to detect the intrusions. In this thesis a hybrid approach using data mining is implemented which is an amalgam of two different techniques namely K-Mean and Adaptive SVM. The results of the proposed hybrid approach are compared with the results of K-Mean and Adaptive SVM individually and it outperforms them. The new approach is effective during detection of attacks and accuracy of the proposed algorithm is better than other techniques.

The hybrid approach effectively classifies the data either as normal or attack. The accuracy of the proposed hybrid approach is 99.54%, which is better than the techniques performed individually. It can be concluded that this approach is simple and efficient in terms of reducing the false positive ratio and increases the false negative ratio.

REFERENCES

[1] CERT/CC Statistics 1988-2003, http://www.cert.org/stats/cert_stats.html
[2] Jones, Anita K., Sielken, Robert S., "Computer system intrusion detection: A survey", Technical Report, Computer Science Dept., University of Virginia, 1999.
[3] Koziol, Jack, "Intrusion detection with Snort", Sams Publishing, 2003.
[4] Bace, Rebecca Gurley, Intrusion detection", Macmillan Technical Publishing, 2000.
[5] Crothers, Tim, "Implementing intrusion detection systems", Wiley Publishing, Inc., 2003.
[6] Mohammadian, M., "Intelligent Agents for Data Mining and Information Retrieval," Hershey, PA Idea Group Publishing, 2004
[7] Wang, J., "Data mining: Opportunities and challenges," Idea Group Publishing, September, 2003.
[8] McCulloch, W.S., and Pitts, W., "A logical calculus of the ideas immanent in nervous activity," Bulletin of Mathematical Biophysics, vol. 5, pp. 115-133, 1943.
[9] Haykin, Simon, "Neural networks, A comprehensive foundation", Macmillan College Publishing Company, Inc.
[10] Christodoulou, C., Georgiopoulos, M., "Applications of Neural Networks in Electromagnetics", Artech House.
[11] whatis.techtarget.com/definition/confidentiality-integrity-and-availability-CIA
[12] Subaira.A.S and Mrs.Anitha.P, " Efficient Classification Mechanism for Network Intrusion Detection System Based on Data Mining Techniques:a Survey", IEEE 8th Proceedings International Conference on Intelligent Systems and Control (ISCO)(2014).
[13] Wang, "Anomalous Payload-Based Network Intrusion Detection (PDF) " ,Proceedings of Springer: Recent Advances in Intrusion Detection.

[14] Azmi R, and Khansari M. et al. "Host-based web anomaly intrusion detection system, an artificial immune system approach".
[15] Dae-Ki Kang and Doug Fuller et al. , " Learning Classifiers for Misuse and Anomaly Detection Using a Bag of System Calls Representation", IEEE Workshop on Information Assurance and Security United States Military Academy (2005).
[16] K. Shivshankar E., "Combination of Data Mining Techniques for Intrusion Detection System", IEEE International Conference on Computer, Communication and Control (IC4-2015).
[17] M. Junedul H. and Khalid.W. Magld et al., "An Intelligent Approach for Intrusion Detection Based on Data Mining Techniques", proceedings of IEEE,2012.
[18] "Nsl-kdd data set for network-based intrusion detection systems." Available on: http://nsl.cs.unb.ca/KDD/NSLKDD.html, March 2009.
[19] Brijesh Sharma and Huma Gupta, A Design and Implementation of Intrusion Detection System by using Data Mining, IEEE Fourth International Conference on Communication Systems and Network Technologies (2014).
[20] Mahbod T. And E. Bagheri et al. " A detailed Analysis of the KDD CUP 99 Data set", IEEE Symposium on Computational Intelligence in Security and Defence Applications (2009).
[21] K. S. Desale, Chandrakant N. K. et. al , "Efficient Intrusion Detection System using Stream Data Mining Classification Technique", IEEE International Conference on Computing Communication Control and Automation (2015).
[22] Solane Duque, Dr.Mohd. Nizam bin Omar, "Using Data Mining Algorithms for Developing a Model for Intrusion Detection System (IDS)" , proceedings of Sciencedirect, Conference Organized by Missouri University of Science and Technology 2015-San Jose, CA. pp. 46-51.
[23] Patel and Santosh Sammarvar et al., "Data Mining Vs Statistical Techniques for Classification of NSL-KDD Intrusion Data", International Journal of Computer Science and Information Technologies, Vol. 5 (4) ,(2014).
[24] Liang Hu and Taihui Li et al., "False Positive Elimination in Intrusion Detection Based on Clustering", IEEE 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD) (2015).
[25] Zhengjie Li and Yongzhong Li et al., "Anomaly Intrusion Detection Method Based on K-means Clustering Algorithm with Particle Swarm Optimization", International Conference of Information Technology, Computer Engineering and Management Sciences (2011).