ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

A Hybrid Approach for Intrusion Detection Using Data Mining

Meghana solanki [1], Vidya Dhamdhere [2]

P.G. Student, Department of Computer Engineering, G.H.R.C.E.M, Pune, Maharashtra, India[1]

Assistant Professor, Department of Computer Engineering, G.H.R.C.E.M, Pune, Maharashtra, India[2]

ABSTRACT: Intrusion detection is an essential and important technique in research field. 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 different kinds of network attack. We give attacks normally identified by intrusion detection systems. Differentiation can be done in existing intrusion detection methods and systems based on the underlying computational methods used. Intrusion detection methods started appearing in the last few years. In this paper we propose an Intrusion detection method using three different methods. These methods are K-means clustering, neuro fuzzy models and C4.5 algorithm. We propose a three level framework for Intrusion detection. In First step k-means clustering are used to create number of training subsets. In second step different neuro fuzzy models are trained depending on the training subsets. At last we perform classification using C4.5 decision tree algorithm and find whether data is anomalous or not.

KEYWORDS: k-means , Fuzzy Neural Network, Intrusion Detection System , Network Intrusion Detection System , SVM , C 4.5.

I. INTRODUCTION

Intrusion detection is the act of identifying burglar on network**.** It is hot research topic. Security of Information is one of the keystones of Information Society. Because of increasing number of Network attacks over the past few years, intrusion detection system (IDS) is increasingly becoming a censorious component to protect the network. In last years, many researchers are using data mining techniques for building IDS**.** Intrusion detection is the issue of identifying unauthorized use of computer system. It detect misuse of computer system .it diagnose abuse of computer system. This can be present at many different levels such as at the network (point of entry) firewall level, at any point in the network level or at a specific host level. This is usually made by matching signatures (or patterns) of known attacks against network traffic.

There is one system known as IDS (Intrusion Detection Sensor).it is a system that can be placed at the various points in your system to detect and possibly act on intrusion events. This usually is a piece of software that is loaded on a host Operating System but can also be on a dedicated appliance. The data and the networks are protected from attackers and malicious activity using number of different approach. These are many security methods such as password protected access according to needs firewalls are used. Many times these are not efficient. Also the systems are under the threat. Networks are also under the thread.

One technique to guard data is using Intrusion Detection Systems on critical systems. This system work for early detection of attacks. Due to this the recovery of lost and corrupted data becomes simple. In case of intrusion detection, there are two points on which researchers are working, these are efficiency and accuracy. The intrusions at network level or operating system level are detected by making use of these efforts. They are not able to detect corrupted data due to malicious transactions in databases. Intrusion detection is the means of auditing the events in computer world. It is used for evaluating them for signs of possible incidents. They are violations of violation of policies which applies in computer security. Intrusion detection system detects attacks in the network .also it detect anomalies in the network [1].

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

There are various aspects of intrusion detection. Due to which a researcher can become quickly familiar with every aspect of intrusion detection. There are large number of intrusion detection methods, techniques and systems [2].

II. LITERATURE SURVEY

Network anomaly detection is a roomy research area, which already bluster a number of surveys, review articles, as well as books. In this paper author present an efficient technique for intrusion detection by making use of k-means clustering, fuzzy neural network and radial support vector machine. System uses different techniques for intrusion detection. K-means is prototype based. Neural networks are flexible and powerful. Neuro fuzzy are universal approximators. Proposed technique has obtained a reliable peak score for all types of intrusion [3].

In this paper author present an intrusion detection system developed using Bayesian probability. The system developed is a naïve Bayesian classifier that is used to identify possible intrusions. Bayesian probability reduces false positive alert rate. This method improves accuracy of R2L attack. To improve accuracy of IDS several Bayesian filters are used in parallel [4].

In this paper author propose a new, quantitative-based approach for the detection and the prevention of intrusions. Our model is able to probabilistically predict attacks before their completion by using a quantitative Markov model built from a corpus of network traffic collected on a *honey pot.* It predicts attack before their completion [5].

In this Paper author focus on improving intrusion system in wireless local area network by using Support Vector Machines (SVM). SVM performs intrusion detection based on recognized attack patterns. It recognizes anomalies and raises an alarm. It produces better results in terms of detection efficiency [6].

In this paper author propose method Feature Vitality Based Reduction Method, to identify important reduced input features. Author applies one of the efficient classifier naive bayes on reduced datasets for intrusion detection. It reduces attributes and gives better performance. Naive bayes classifier operates on strong independence assumption [7]. In this paper, author propose a clustering based framework for outlier detection in evolving data streams that assigns weights to attributes depending upon their respective relevance. Weighted attributes reduces effect of noisy attributes. It is incremental and adaptive to concept evolution. It gives higher outlier detection [8].

In this paper author propose an outlier mining method based on symmetric neighbourhood relationships. Author evaluates the method with UCI ML Repository datasets, benchmark dataset KDD Cup 1999 and real time intrusion datasets. An efficient outlier detection method is presented [9].

Most anomaly detection methods are typically implemented in batch mode, and thus cannot be easily extended to large-scale problems without sacrificing computation and memory requirements. In this paper, author proposes an online oversampling principal component analysis (osPCA) algorithm to address this problem, and we aim at detecting the presence of outliers from a large amount of data via an online updating technique. It applies to online data. It uses online updating technique [10].

III. PROPOSED IMPLEMENTATION

Intrusion Detection system classified into following types:

I) Missuse based intrusion detection system

Misuse detection is one of an approach used in detecting different kinds of an attack. In misuse detection approach, we first define abnormal system behavior. Then define other behavior as normal behavior. It stands against anomaly detection approach which utilizes the reverse approach, defining normal system behavior and defining any other behavior as abnormal. In misuse detection, the IDS analyse the data it gathers and compares it to large databases of attack signatures. Essentially, the IDS glance for a specific attack. Such a specific attack has already been documented. Like a virus detection system, misuse detection system is only as good as the database of attack signatures

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

that it uses to compare packets against. In other words anything we don't know is normal. Using attack signatures in IDSes is an example of this approach. Misuse detection has also been used to refer to all kinds of computer misuse [1].

II) Network Based Intrusion Detection System

Network based Intrusion Detection System audit the traffic as it flows to other end user. Supervising criteria for a specific host in the network can be increased or decreased with relative ease. NIDS can be able to stand against large amount of network traffic to remain effective. As network traffic increases exponentially NIDS must collect all the traffic and evaluate in a timely manner. A network-based intrusion detection system (NIDS) is used to monitor and analyse network traffic to protect a system from network-based threats. A NIDS scans all inbound packets and look for any doubtful patterns. When threats are discovered, based on its grimness, the system responses by taking action such as notifying administrators, or barring the source IP address from accessing the network [1].

III) Anomaly Based Intrusion Detection System

Anomaly based Intrusion Detection System supervise ongoing traffic, activity, transactions and behavior in order to identify intrusions by detecting anomalies. It works on the assumption that attack behavior differs enough from normal user behavior such that it can be detected by cataloguing and identifying the differences involved. The system administrator defines the baseline of normal behavior. Anomaly-based IDS systems face down due to a lot of false positives. The anomaly detection technique focuses on the concept of a baseline for network behavior. This baseline is a description of accepted network behavior, which is learned or specified by the network administrators. Events in an anomaly detection engine are caused by any behaviors that fall outside the predefined or accepted model of behavior. .Anomaly-based IDS systems can cause heavy processing overheads on the computer system. In this case we have two possibilities: (1) False positive: Anomalous activities that are not intrusive but are flagged as intrusive. (2) False Negative: Anomalous activities that are intrusive but are flagged as non intrusive [1].

A.System Architecture

System architecture of Intrusion Detection System is shown in fig.1.System architecture consists of following methods:

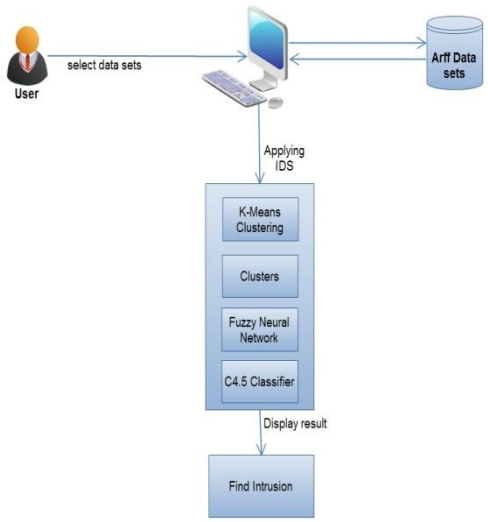

Fig. 1 Architecture of the Proposed System

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

K-means Clustering module:

Clustering is the method of combining the records into classes or clusters, so that entities within the cluster have high resemblance in compare to one another on the further hand are very dissimilar to entities in other clusters. K-means algorithm is useful for undirected knowledge discovery and is relatively simple. If the number of data is less than the number of cluster then we assign each data as the centroid of the cluster. Each centroid will have a cluster number. K-means has found wide spread usage in lot of fields, ranging from unsupervised learning of neural network, Pattern recognitions, Classification analysis, Artificial intelligence, image processing, machine vision, and many others. There are alternate approaches which are used in clustering process Partitioning methods, Hierarchical methods, etc. it is used to group unrelated data. Our input dataset consist of normal data, attack data, training data. It is grouped into 5 clusters using k-means clustering technique-means is a prototype based-means is unsupervised clustering algorithm. It solves well known problem in many field. Goal of clustering is expressed by using objective function. Through k-means clustering modules we cluster training data. It is clustered into 5 subsets. Four subsets of type called attack data set. And one subset of type called normal dataset. K-means clustering is shown in fig.2

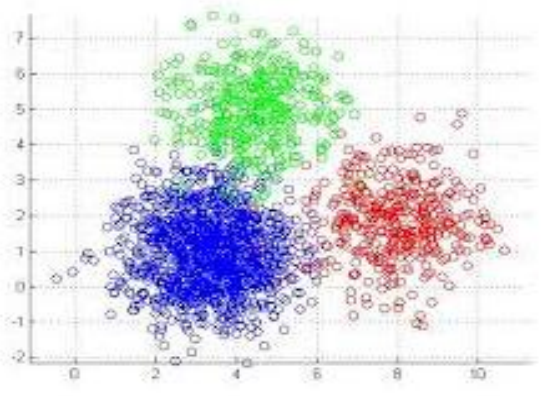

Fig.2 K-means clustering

Neuro-Fuzzy module:

Fuzzy Logic is a problem solving control Structure approach that gives itself to implementation in the systems which are ranging from multichannel PC or Workstation acquisition and control systems. Neural networks are compelling tool for classification. They are flexible. They are powerful. Impossible interpretation of the functionality is the drawback of neural network. It also faces problem in determining number of layers and it can be engaged in hardware, software, or in both. Combination of neural network and fuzzy logic known as FNN. They are universal approximators. It can handle any kind of information. It has self learning capability. It is self organizing. It makes the use of backpropagation learning .It offers a simple manner to attain on a definite decision based upon indefinite, ambiguous, inaccurate, noisy, or absent input information.FNN is shown in fig.3

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

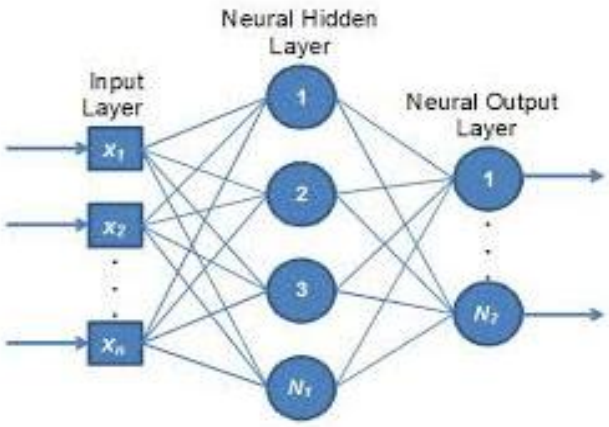

Fig.3 Fuzzy neural network

SVM module:

A Support Vector Machine (SVM) performs classification by finding the hyper plane that maximizes the margin between the two classes. The vectors (cases) that define the hyper plane are the support vectors. Support vector machines (SVM) are learning machines that plot the training vectors in high dimensional feature space, labelling each vector by its class. SVMs classify data by determining a set of support vectors, which are members of the set of training inputs that outline a hyper plane in feature space. Computing the hyper plane to separate the data points leads to a quadratic optimization problem. There are two main reasons that we used SVMs for intrusion detection. The first reason is that its performance is in terms of execution speed, and the second reason is scalability. SVMs are relatively insensitive to the number of data points, and the classification complexity does not depend on the dimensionality of the feature space.

C4.5 module:

Classification algorithms have attracted considerable significance both in the machine learning and in the data mining research areas. Among classification algorithms, the C4.5 system gains a special mention for several reasons. On the other side, it shows the result of research in machine learning that traces back to the ID3 system. It has always been taken as the point of reference for the improvement and analysis of innovative proposals. On the other hand, the results show that the C4.5 tree-induction algorithm gives good classification accuracy and is the fastest among the compared main-memory algorithms for machine learning and data mining. The algorithm constructs a decision tree starting from a training set T S, which is a set of cases, or tuples in the database terminology. Each case specifies values for a collection of attributes and for a class. Each attribute may have either discrete or continuous values. Moreover, the special value unknown is allowed, to denote unspecified values. The class may have only discrete values. We denote with C1..... CN Class the values of the class.

B.Proposed System Algorithm

1. K-Means Algorithm:

Step-I Define number of cluster 'K'

Step-II Initialize the K cluster centroids.

This is achieved by partitioning all

objects into k clusters. Next their centroids are computed. Then it is verify that whether all centroids are initialized to k arbitrarily chosen, different objects.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

Step-III Iterate over all data points in the data set and compute the distances to the centroids of all clusters. Assign each data points to the cluster with the nearest centroids.
Step-IV Re calculate 'K' new centroids for each cluster which is the outcome of the preceding step.
Step-V Repeat step 3 until the centroids do not change any more.

2. Fuzzy Neural Network (FNN):
Implement the fuzzy IF-THEN rules by trainable neural network architecture.
Step-I Ri : if x is Ai, then y is Bi
Where, Ai and Bi are fuzzy numbers, i=1….n.
Fuzzy rule is based on three rules.
1. If x is small then y is negative.
2. If x is medium then y is about zero.
3. If x is big then y is positive.

3. C4.5 Algorithm:
Step-I Apply the test instance over the c4.5 decision tree of the computed closest cluster.

FormTree (T)
1) ComputeClassFrequency(T)
2) If OneClass or FewCases
   Return a leaf;
   Create a decision node N;
3) ForEach Attribute A
   ComputeGain (A);
4) N.test = AttributeWithBestGain;
5) if N.test is continuous
   Find Threshold;
6) ForEach T' in the splitting of T
7) if T' is Empty
   Child of N is a leaf
   else
8) Child of N = FormTree(T')
9) ComputeErrors of N;
   Return N

Step-II Classify the test instance as normal or anomaly and include it in the cluster.
Step-III Update the centre of the cluster.

C.Mathematical module

Formal model of IDS is represented by function F.IDS is represented as an 5-tuple items, in which first item is for input , middle three are for algorithms and last item is for output.
F= (I, K, N, C, O);
Where
F is function of intrusion detection system;
I is input to system which is Dataset. Dataset is KDD 1999 dataset. Dataset consist of number of records. In records there are number of documents, which may be clean or contain some malicious data.
O is output.
O= (Normal, Anomalous);
When IDS receives input it apply k-mean, FNN & c 4.5 these techniques and determine whether records in document is normal or anomalous.
K is K-mean clustering algorithm;
N is Fuzzy neural network algorithm;
C is c 4.5 algorithm;

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

Cluster formation-

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \| X_i^{(j)} - C_j \|^2$$

Where "J" is objective function, "k" is number of cluster, "n" is number of cases, "X" is case of i & "C" is centroids for cluster j.

IV. EXPERIMENTAL RESULTS

We used 10% KDD 99 dataset for our system. We implement our system on a windows PC with i3 processor 2.30 GHz CPU and 4GB RAM. We used KDD 99 dataset for experiment. It is publically available dataset. It includes actual attacks. We used Java language for our system implementation. We used weka tool for SVM classification .We are using this dataset to design and evaluate system.KDD 99 dataset contain training data and test data. There are 41 features for each instance in dataset. There are 38 types of attack in dataset. It falls into 4 categories: PROBE, DOS, R2L, and U2R. DOS and Probe are greater frequency attack. It is difficult to achieve detection accuracy for U2R & R2L attack.KDD 99 is very huge dataset. It is tough. It requires high end machine to perform experiment. Due to this we use 10% KDD for experiment. True positive, True negative, False positive, False negative these are measurements which are used for evaluation of system. We used evaluation metrics like sensitivity, specificity and accuracy.

Sensitivity= TP/(TP+FN) Specificity= TN/(TN+FP) Accuracy= (TN+TP)/TN+TP+FN+FP

TABLE I. DATA POINTS TAKEN FOR IDS

Dataset	DOS	PROBE	U2R	R2L
KDD25000 (D1)	9267	2294	11	211
KDD50000 (D2)	11626	343	5	61
KDD75000 (D3)	33862	830	5	70
KDD100000 (D4)	41845	1808	9	102
KDD125000 (D5)	62361	1809	9	102

For Experimental result we consider 5 data records, whose size is 25000, 50000, 75000, 100000, and 125000 respectively. Data points for these records are given in table1.In table we take data points for four types of attack such as DOS, PROBE, U2R, R2L. After applying SVM & C 4.5, we get the results which are shown in fig.4.Results show that C 4.5 performs better that SVM.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

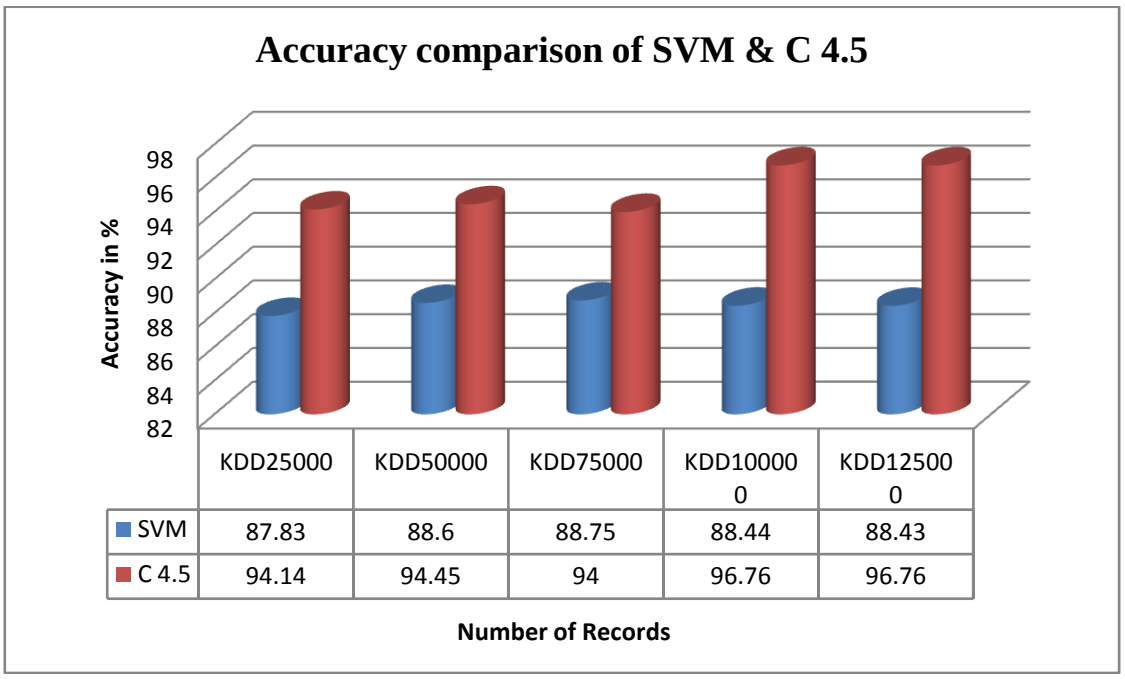

Fig. 4 Accuracy Graph

V. CONCLUSION

In this paper we present an efficient technique for intrusion detection by making use of k-means clustering, fuzzy neural network and C 4.5. We took the help of k-means clustering technique to make large, heterogeneous training data set into a number of homogeneous subsets. As a result complexity of each subset is reduced and consequently the detection performance is increased. After initial clustering in the proposed technique, training will be given to fuzzy neural network and latter classification using C 4.5 decision tree is performed. We have received the best results for C 4.5 as compare to SVM.

REFERENCES

[1] Miss Meghana Solanki, Mrs. Vidya Dhamdhere, "Intrusion Detection System by using K-Means clustering, C 4.5, FNN, SVM classifier", Volume 3, Issue 6, November-December 2014, ISSN 2278-6856.I.S.

[2] Miss Meghana Solanki, Mrs. Vidya Dhamdhere, "Intrusion Detection Technique using Data Mining Approach: Survey", Vol. 2, Issue 11, November 2014, ISSN(Online): 2320-9801.

[3] A.M. Chandrasekhar, "Intrusion Detection Technique By Using K-Means, Fuzzy Neural And Svm Classifier ", 2013 International Conference on Computer Communication and Informatics (ICCCI - 2013), Jan 04-06, 2013 Coimbatore, India.

[4] Hesham Altwaijry, "Bayesian Based Intrusion Detection System ", Journal of King Saud University – Computer and Information Sciences (2012) 24,1–6.

[5] Ammar Boulaiche, "A Quantitative Approach For Intrusions Detection And Prevention Based On Statistical N-Gram Models ", Procedia Computer Science 10 (2012) 450 – 457.

[6] Muamer N.Mohammed, "Intrusion Detection System Based On Svm For Wlan ",Procedia Technology 1 (2012) 313 – 317.

[7] Saurabh Mukherjee, "Intrusion Detection Using Naive Bayes Classifier With Feature Reduction", Procedia Technology 4 (2012) 119 – 128.

[8] Yogita,Durga Toshniwal, "A Framework for outlier Detection in Evolving Data Streams by wieghting Attributes in Clustering", Procedia Technology 6(2012) 214-222.

[9] Prasanta Gogoi, B Borah, D K Bhattachharya, J K Kalita, "OutlierIdentificationusingSymmetric Neighborhoods",Procedia Technology 6(2012) 239-246.

[10] Yuh-Jye Lee,"Anomaly Detection via Online Oversampling Principal Component Analysis",IEEE VOL. 25, NO. 7, JULY 2013.