International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Data Mining Based Intrusion Detection in Wireless Sensor Network

Asha R N, Venkatesan S

Abstract*: Intrusion detection is the one of the challenging task in wireless sensor network and prevents the system and network resources from being intrude or compromised. One of the ongoing strategies for recognizing any anomalous activities presented in a network is done by intrusion detection systems (IDS) and it becomes an essential part of defense system against attacker problems. The primary goal of our work is to study and analyze intrusion detection technique meant for improving the performance of Intrusion Detection using hybrid ANN based Clustering technique. To estimate the effectiveness of the proposed strategy, KDD CUP 99 dataset is utilized for testing and assessment. Based on the analysis, it is noticed that the proposed ANN clustering performs much better than other methods with respect to accuracy which attains an average high accuracy of 93.91%when compared with other methods.

Index Terms*: classification, *data mining, dataset, intrusion detection.

I. INTRODUCTION

Data mining algorithm uses a different data analyses tool to identify the pattern and relation among data that might be utilized to make substantial predictions. The concealed and valuable informations were taken from the accessible dataset through data mining. It involves various steps of procedure. It contains different gatherings of inter connecting steps which will assist to obtain the required information for making decision. Data mining searches the database to find unknown patterns and anticipate data to enhance the associated business. Data mining is the non trivial extraction of certain presumptuously unknown, fascinating and conceivable needed data. Classification, grouping, expectation, affiliation, rule extraction and detection are the different kinds of issues that can be illuminated through data mining. Statistics, Al and pattern acknowledgement are the methods used in data mining. It incorporates statistical methods, case based thinking, neural network, decision trees, rule acceptance, Bayesian network, fuzzy sets, harsh sets and genetic algorithm.

Revised Manuscript Received on 30 July 2019.
***** Correspondence Author
Asha R N*, Computer Science & Engineering, Global Academy of Technology, Bangalore, India.
Dr. Venkatesan, Computer Science & Engineering, DSCE, Bangalore, India.

Because of the advanced web technology and local networks, intrusion in computer systems is becoming more. Due to the increased network connections, computer systems are winding up progressively powerless against attacks. The general objective of attack is to subvert the conventional security technique on the system and accomplish operation to authorize intruders. These intruders access the protected reading or doing malignant harm to the system files. By constructing complex tools for continuous monitoring and reporting, the security system could find potentially noxious activities. Intrusion detection systems are significantly progressing in keeping network security at appropriate level. A perfect intrusion detection system has the ability to classify ordinary and anomalous activities. These include any occasion, state, substance or character that is viewed as anomalous against predefined standard. It is essential for IDSs to generate rules to recognize ordinary and irregular characteristics by viewing datasets, which are the records of exercises created by the system that are entered in sequential arrangement. Intrusion detection has influenced more interest among the analyst because of the quick improvement and familiarization of the Internet and local network. Despite lot of procedures and tools available, more research is carried out to develop good intrusion detection systems.

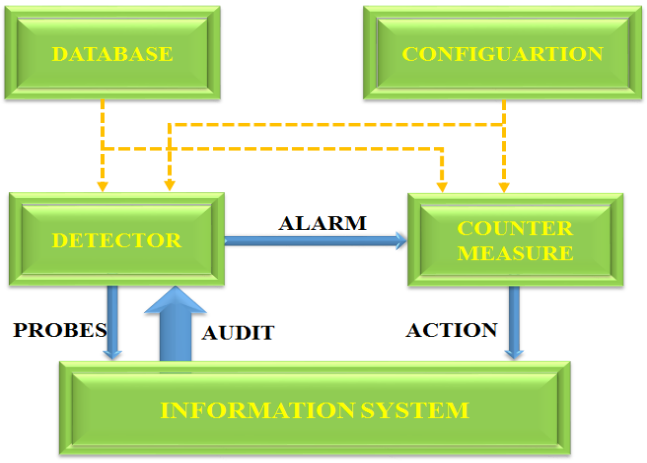

Figure 1.1 Simple Intrusion Detection Systems

The detector eliminates the unwanted data from the audit. The synthetic perspective of the security concerned issues considered during typical operation of the network and a synthetic aspect of the present security state of the network is presented. A conclusion is made by estimating the likelihood of these actions and considered as manifestation of intrusion.

A counter measure is taken for preventive action to either stop the activities from being implemented or to change the system back to a protected state.

Intrusion Detection System is implemented with preventive security components as a second line of barrier to ensure data system, such as access for controlling and confirmation. The reasons for making intrusion identification as an essential portion of the defense system are given below. Most customary system and application were created without considering safety issue. At next, system and application were created to do a task in various environments and become vulnerable when conveyed without any intrusion detection system. Even if the preventive security technique can safeguard information system successfully, it is necessary to know the type of intrusions that have occurred or are going on so that the security threats and problems can be understood for ensuring future attacks.

The data mining field has been growing because of its vast success in scientific progress and its wide range of applications. The data mining based applications are deployed in different fields like financial analysis, customer management, hospitality, business organization, communication, duplicate detection and threat analysis etc. The consistently increasing difficulties in various fields and enhancements in technology have presented new challenges for data mining. The problems of various data groups, data from different areas, propels in computing and system administration assets, research and logical areas, regularly increasing business challenges and so on.

II. RELATED WORKS

The concept of IDS was presented initially by Anderson [1] and later reframed by Denning [2], has got noticed in the recent years. The aim of IDSs is to detect intrusion, i.e., sets of activities to bargain the integration, confident or accessibility of a system resource [3]. A Host dependent Intrusion Detection has host dependent sensor and a network dependent Intrusion detection has network based sensor [4]. Host dependent technique verifies files which are accessed and applications that are executed [5]. Network-based intrusion detection is used in solving issues related to detection of unauthorized utilization of systems connected in a network [6].

The intrusion identification system has the ability to recognize typical and irregular activities of user [7]. This would incorporate any occasion, state, substance and conduct which is viewed as anomalous by a pre characterized format [8]. The intrusion detection system for data mining is categorized based on two detection methodology. They are misuse identification and anomalous identification. Abuse identification utilizes pattern of familiar assault or weak areas of the system to detect intrusion [9]. Anomaly detection determines the deviation that is established from normally used patterns and flagged as intrusion [10]. The major challenge in intrusion detection is the identification of the hidden intrusion from a large quantity of typical communication practices [11].

Almost all the present feature extraction strategies involves in transformation of the real input pattern of high dimensional vectors into low dimensional new vectors [12].The Principal Component Analysis is the renowned technique for reducing the dimensionality. But, selecting the number of directions is the major issue with this method. It does not compute the principal components of high dimension feature spaces which are related with input feature by non-linear mapping [13]. Linear Discriminate Analysis based feature reduction is the recent method used in the detection of network attack. This technique decreases the number of input features besides increasing the accuracy of classification. In addition, the preparation and the testing time of the classifier can also be reduced by choosing the most discriminating features [14].

The manner in which the best feature set is chosen forms the major issues experienced by the analysts because all the features were not linked with the learning algorithms. In certain cases, superfluous and redundant features may create noisy data and divert the learning algorithms and corrupt the detector exactness, causes training and testing processes to take more time. Feature selection was shown to have a significant effect on the classifier performance [15].

Since research is going on to detect new attacks with the key component as data mining [16]. Data mining is the investigation of data to create relation and detect concealed pattern of data or to prevent the data from unnoticed. Numerous researchers stayed in the intrusion detection of data using data mining [17].

Various data mining strategies are proposed for intrusion detection. K-Mean Clustering [18] is an unsupervised data mining method to detect intrusion. K-Means is a well known effortlessness execution and partition based clustering algorithm and is commonly implemented in diverse applications.

Latifur Khan et al [19] has proposed a composition of SVM and DGSOT technique, which begins with an underlying preparation set and extended slowly by utilizing the clustering approach created by the DGSOT algorithm. Rasha G. Mohammed Helali [20] has surveyed data mining based Intrusion Detection System. He introduced highlights of signature based NIDS and presented the Data Mining based NIDS approaches. Many researchers have contended that Artificial Neural Networks (ANNs) can enhance the performance of detection system when contrasted with conventional approach. ANN is one of the broadly utilized methods and is effective in analyzing numerous complex real-time issues.

III. IDS DATA PROCESSING TECHNIQUES

Data Mining is one of the steps in KDD process which uses explicit algorithm for obtaining patterns from data. The term KDD alludes to the complete process of discovering valuable information from the data. The further steps in KDD process are training, selection, classification of data. Initially, data is received from different sources followed by data preprocessing step such as data cleaning and data selection. This makes information distribution center. Relevant data is taken from data warehouse task and data mining is implemented on this. In data mining, pattern evaluation is applied to extricate knowledge. Therefore in knowledge discovery process, data mining plays an important function. The KDD process alludes to the entire process of transforming low level information into high level knowledge which may be automatic or semi-automatic revelation of patterns and relations in large databases.

In a network every connection is treated as common or as attack, with precisely one explicit attack type. The four main types of assault are:

- *DOS: denial of service, (syn. flood)*
- *R2L: unapproved accession from a remote machine, (speculating password)*
- *U2R: unapproved accession to neighbor user (root), e.g., different "buffer overflow" assault;*
- *probing: reconnaissance and other probing (port filtering).*

It is essential to notice that the data to be examined should not be from the similar likelihood dispersion of the preparation stage data but it incorporates explicit assault types which are not in the training data. This makes the operation much practical. Some intrusion expert accepts that most assaults are variation of known assault and the "signature" of known assault is adequate to obtain novel variation. The KDD cup 99 datasets contain 24 training assault categories with extra 14 types of the test data only.

IV. EVALUATION OF IDS USING ANN

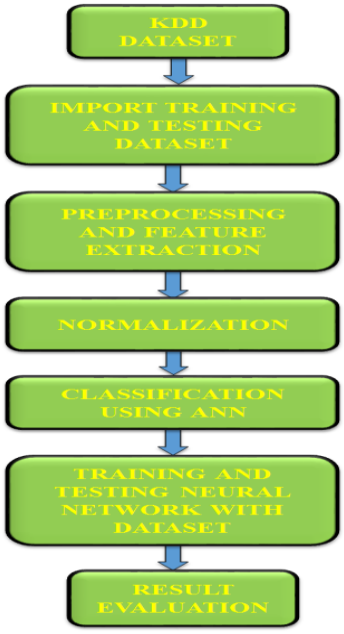

Figure 2: Flowchart of proposed ANN

A. Preprocessing and feature extraction

The input data of the ANN is in the scope of (0 1) or (-1 1). Subsequently the preprocessing and feature extraction of data is needed. The KDDCUP99 data are preprocessed. Each record in KDDCUP99 dataset has 41 features and they are in continuous, discrete and emblematic form with remarkable different ranges. The input may have different structure depending on the type of neural network and hence it requires various preprocessing techniques. Some neural network can take only binary data and some can take continuous estimated data. In Pre-processing, after KDDCUP99 features are obtained from each record, all the features are converted into numerical format from text and symbolic format. To convert the symbol as number format, an integral code is allotted to each symbol.

B. Normalization

Preprocessing changes all the symbol and text formats into numbers. The scope of estimation of the various features is not uniform. Some features that have large span of values will affect the performance more than other features that has less span of values. Consequently normalization is applied to the features to change the scope of quantities to fall between 0 to 1. Various techniques are proposed to normalize the data.

In order to normalize feature esteems, statistical investigation is made on the estimation of all features based on the KDDCUP99 data set and the optimal estimation of each feature is measured. Based on the maximum value and by utilizing the basic formulae for normalizing the feature value in the scope (0, 1), the normalization procedure is given below.

If (MaxF $\leq$ f) then Nf=1;
else Nf = (f / MaxF)

F: Feature, f: Feature esteem, MaxF: optimum value of F, Nf: Normalized value of F

C. Classification

Classification techniques surmise a model from the database. The database contains numerous qualities that infer the tupple class and these are known as anticipated traits while the remaining traits are called foreseeing attribute. The combined value for the anticipated trait indicates a class. The system need to find the rules for predicting the class while learning classification rules from the anticipated traits, hence in the initially stage the user has to state the constraints for each class and then the data mining system should construct the portrayals for each class. Essentially, the system should be given with a case or tupple of certain known trait values so that it can be able to anticipate the type of class that each case belongs to.

There are different data mining classification methods available which are Decision Tree based Methods, Rule based techniques, Naive Bays and Bayesian Belief Network, Nearest Neighborhood Method Neural Network, Support Vector Machines, Ensemble Method for classification and expectation.

To find the relation between input and output vector, neural network utilize their learning algorithm and standardized them to get new relation between input and output. The main objective of neural network is to study the characteristics of vectors in the intrusion detection system (e.g., clients, domains). It is come to know that statistical analysis is comparable with neural network. The benefit of neural network over statistics analysis is its simple way to express non-linear relation among variables and in learning about their relation.

D. ANN

ANN is an organically roused form of distributed algorithm. It is made up of basic processing units or hubs, and link between them. The link between any two nodes has some weight, which is utilized to decide the amount of one unit to influence the other. The subset of unit functions as input node and some subset work as output node, which perform summation and thresholding. An early stopping system is utilized to surpass the over-fitting issue. The early halting strategy monitors the performance of network by utilizing separate approval set. Typically, as the network fits the data, the errors in the validation set will gets decreased and after that error gets increased as the network fits with peculiarities in the training data.

ANN Algorithm

Start

Assume some small arbitrary value for each w_{ij}
While
For each $<\vec{x}, t>$ in training dataset
Input the unit $\vec{x}$ and calculate the output
For each weight w_{ij}, process the above and update
$\Delta w_g\ (t+1) = \Delta w\ t + \eta\delta\ x_t$
while hub j is output hub,
$\delta_j = o_j(1 - o_j)(\ t_j - o_j)$
while hub j is shrouded hub,
$\delta_j = x'_j\ (1- x'_j)\sum_k \delta_k\ w_{jk}$
Until converge
End

For IDS based on hybrid ANN, more than one ANN technique is executed in a steady progression. A set of the input and anticipated yield are given as hybrid ANN is a supervised learning method. ANN can have any number of concealed layers. Maximum number of shrouded layers for a given number of input and output is chosen by the experimental strategy. In case of detecting and categorizing the attack problems, one hidden layer is not adequate. Henceforth more number of hidden layers is executed at the cost of increasing system complexity and reduced intermingling rate. The complexity of the system may increase or decrease with the number of units in each layer. Excess number of concealed layers will diminish the system performance, while less number of hidden layers will lessen the detection rate.

V. EXPERIMENTAL VERIFICATION

We have executed the Hybrid ANN by utilizing KDD CUP 1999 dataset. We have considered 10% of the dataset for training and 90 % dataset as the testing. The database consists of 24 sorts of preparation assaults, with extra 14 categories as the test data. Basically the assaults are categorized as DOS, R2L, U2R and Probe attacks. In our experiments, we have considered 41 inputs, 1 output with constant learning rate 0.9 and initial weights were assumed to be any arbitrary value. Two concealed layers with 41 shrouded units in each layers is taken. Our demonstration shows that the time taken by the model with one concealed layer is increased as the number of concealed layer increases. Also the convergent rate for the model with one shrouded layer is high. Detection performance of hybrid ANN is good for both the known and obscure assault. In order to train the hybrid ANN, number of the epochs required was very high which consumes more time for training. The performance is decreased when network is over-trained and it is required to define the early halting condition. It is conceivable to classify the attack category of KDD CUP 99 dataset using the various output layer of hybrid ANN.

Table 1: Detection rate for various attacks

TYPE OF ATTACK	TOTAL ATTACK	DETECTED ATTACKS	DETECTION RATE (%)
Probe	39	38	98%
DoS	62	60	97%
R2L	52	48	95%
U2R/Data	41	38	96%
Total	194	184	97%

The performance of detection system is measured by: a) Accuracy b) Detection Rate (DR), c) Failure Analysis Rate (FAR). The precision of the system is expressed by the below articulation.

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

The Detection Rate (DR) is resolved depends on the given articulation.

$$Detection\ Rate\ (DR) = \frac{TP}{TP + FP}$$

Detection rate gives the likelihood of irregular data. With high detection rate, the algorithm can effectively predict the input anomaly.

$$Failure\ analysis\ rate\ (FAR) = \frac{FP}{TN + FP}$$

Failure analysis rate indicate the precision of intrusion detection. With low FAR, the detection accuracy becomes high.

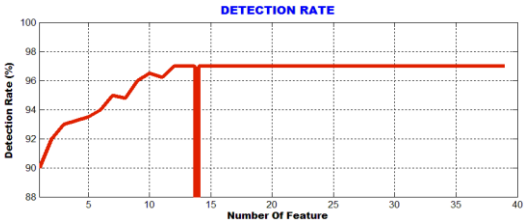

Figure 3: Number of features Vs Detection rate

From figure 3 it is found that the detection rate of hybrid ANN is excellent when compared with other ANN methods. It should note that detection rate of the any model relies on the number of variables and other conditions.

VI. CONCLUSION

We have various techniques of Artificial Neural Network to realize intrusion detection system. Each method is applicable for some particular circumstances. Hybrid ANN is more precise, adaptable and tolerant to various types of attacks, versatile to network changes, scheduled and works in real time. Hybrid ANN based model is preferably used for high detection rate requirement. A simple or hybrid ANN based approach can be used for identifying assault. The required number of the epochs and the preparation time can be reduced by combining different ANN techniques. The proposed method has attained the maximum detection rate of 97% with KDD CUP 99 dataset. Hybrid ANN is supervised learning based neural network model that are easy to implement. Number of the epochs needed to prepare the network is high when compared with other ANN techniques. Even with high detection rate, hybrid ANN suffers from the local minima and slow coverage.

REFERENCES

1. James P. Anderson. Computer security threat monitoring and surveillance. Technical report
2. Dorothy E. Denning. An intrusion detection model. IEEE Transactions on Software Engineering, SE-13(2):222–232, 1987.
3. Richard Heady, George Luger, Arthur Maccabe, and Mark Servilla. The architecture of a network level intrusion detection system. Technical report, University of New Mexico, 1990.
4. AbhijitSarmah, “Intrusion Detection Systems: Definition, Need and Challenges”, White Paper from SANS Institute, 2001.
5. Harley Kozushko, “Intrusion Detection: Host-Based and Network-Based Intrusion Detection Systems”, White Paper from Independent Study, September 11, 2003.
6. Dewan Md. Farid and Mohammad Zahidur Rahman, “Anomaly Network Intrusion Detection Based on Improved Self Adaptive Bayesian Algorithm”, Journal of Computers, Vol.5, No.1, January, 2010.
7. AnazidaZainal, MohdAizainiMaarof and Siti Maryam Shamsudin , “Research Issues in Adaptive Intrusion Detection”, In Proceedings of the 2nd Postgraduate Annual Research Seminar (PARS'06), Faculty of Computer Science & Information Systems, UniversitiTeknologi Malaysia, 24 – 25 May, 2006.
8. Fengmin Gong, “Deciphering Detection Techniques: Part II Anomaly-Based Intrusion Detection”, White Paper from McAfee Network Security Technologies Group, 2003.
9. Wenke Lee and Salvatore J. Stolfo, “Data Mining Approaches for Intrusion Detection”, Proceedings of the 7th USENIX Security Symposium, San Antonio, Texas, January 26-29, 1998.
10. Marcos M. Campos, Boriana L. Milenova, “Creation and Deployment of Data Mining-Based Intrusion Detection Systems in Oracle Database 10g”, In Proceedings of the Fourth International Conference on Machine Learning and Applications, 2005.
11. Jian Pei , Jiawei Han , Laks V. S. Lakshmanan, “Pushing Convertible Constraints In Frequent Itemset Mining”, Data Mining And Knowledge Discovery, Vol. 8, No.3, pp.227-252, May 2004.
12. Rupali Datti and Bhupen draverma, Feature Reduction for Intrusion Detection Using Linear Discriminant Analysis, International Journal on Computer Science and Engineering, Vol. 02, No. 04,pp.1072-1078, 2010.
13. Shailendra Singh, Sanjay Silakari and Ravindra Patel, —An efficient feature reduction technique for intrusion detection system, International Conference on Machine Learning and Computing, vol.3, 2011.
14. Sri latha Chebrolu, Ajith Abraham and Johnson P Thomas, —Hybrid Feature Selection for Modelling Intrusion Detection Systems, Lecture Notes in Computer Science,Vol.3316, pp 1020-1025, 2004.
15. Mohanabharathi R, T.Kalaikumaran and S.Karthi, —Feature Selection for Wireless Intrusion Detection System Using Filter and Wrapper Model,International Journal of Modern Engineering Research (IJMER), Vol.2, No.4, pp-1552-1556, 2012.
16. Gowrisona G, K. Ramarb, K. Muneeswaranc, T. Revathic, " Minimal complexity attack classification intrusion detection system", Applied Soft Computing, Vol 13, pp: 921–927, 2013.
17. Shingo Mabu, Nannan Lu, Kaoru Shimada, Kotaro Hirasawa, " An Intrusion-Detection Model Based on Fuzzy Class-Association-Rule Mining Using Genetic Network Programming", IEEE Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews, VOL. 41, NO. 1, PP: 130-139 , 2011.
18. Bahrololum M, E. Salahi and M. Khaleghi “Anomaly intrusion detection design using hybrid of unsupervised and supervised neural networks”, International Journal of Computer Networks & Communications, Vol.1, No.2, 2009.
19. Latifur Khan, MamounAwad, BhavaniThuraisingham, “A new intrusion detection system using support vector machines and hierarchical clustering”, The International Journal on Very Large Data Bases, Vol. 16, no. 4, October 2007.
20. Rasha g. Mohammed Helali, "data mining based network intrusion detection system: a survey", novel algorithms and techniques in telecommunications and networking, pp. 501-505, 2010.

AUTHORS PROFILE

Asha R N is a Computer Science & Engineering graduate. She graduated in the year 2005 and later obtained MTech degree in Computer Science & Engineering in the year 2012. Currently she is pursuing PhD from VTU in the area of Wireless Sensor Network. Currently she is working as an Assistant Professor in the department of CSE, Global Academy of Technology, Bengaluru, Karnataka, India. She has 14 years of teaching Experience. She has guided several UG and PG projects. She is also Life time member in ISTE and CSI society. **harshaaru@gmail.com**

Dr. Venkatesan S Completed B.E., Computer Science and Engineering, from University of Madras, Completed M.E., Computer Science and Engineering from Anna University and Awarded Ph.d. from Anna University Area of expertise Networks, Image Processing, Soft Computing, Optimization Techniques. 17 Years of Academic Experience, currently working as Professor in the Department of Computer Science & Engineering, Dayananda Sagar College of Engineering, Bangalore, Karnataka, India .Life Member in Indian Society for Technical Education. **selvamvenkatesan@gmail.com**