International Journal of Computer Networks and Communications Security

VOL. 2, NO. 12, DECEMBER 2014, 485–490
Available online at: www.ijcncs.org
ISSN 2308-9830

A survey on Data Mining based Intrusion Detection Systems

Mala Bharti Lodhi[1], Vineet Richhariya[2] and Mahesh Parmar[3]

[1, 2, 3] LNCT, Department of Computer Science Engineering, RGPV University, India

E-mail: [1]mala141086@yahoo.co.in, [2]Vineet_rich@yahoo.com

ABSTRACT

In communication the security is an essential objective. Therefore various security systems are developed for networks, among them the IDS are an essential contribution for security. In this paper IDS technology is investigated in detail. In addition of that recent development on IDS systems is also investigated in this paper. After evaluation of previously developed methodology a new IDS system is proposed for enhancing the performance of the recently developed IDS systems.

Keywords: *Network Security, IDS, Data Mining, Classification, Review.*

1 INTRODUCTION

Now in these days the network communication is growing continuously and adopted rapidly. The network technology having a large number of applications, this is now used for banking applications, shopping and others. Thus a significant amount of sensitive and private data is traversing through these networks. During to data transmission data is travelling through untrusted network, therefore loss of security and data is an essential concern in the network technology. The presented study provides a detailed investigation of the security aspects and their flaws. Thus in this study intrusion detection systems are learned and a new concept of intrusion system design is presented. The key objectives of the presented study involved the following work.

a. Study of intrusion detection system: in this phase, intrusion detection system is studied. In addition of that their properties and standard datasets are obtained.

b. Investigation of intrusion properties: in this phase the attributes and their properties are explored for finding their importance in IDS system. Additionally various different recently developed techniques and methods are also studied for finding the optimum method for classifying the KDD CUP 99's datasets.

c. Design and development of hybrid IDS system: in this phase a new algorithm is designed and implemented using suitable technology.

d. Performance analysis of proposed system: after design and implementation of the desired technique the performance of system is evaluated using accuracy, error rate, memory consumption (space complexity) and time consumption (time complexity).

IDS systems are a kind of security filter designed using software or hardware configuration. Therefore, intrusion detection system (IDS) examines all inbound and outbound network activities such as user activities, packet transactions and identifies suspicious patterns. These patterns are analysed using any network administrator defined rules, predefined constrains for network or using machine learning algorithms. There are numerous ways to categorize IDS:

- Misuse detection vs. anomaly detection: in misuse detection, the IDS analyses the information it collects and compares it to huge databases of attack signatures. Like a virus detection system, misuse detection software is only compare with the database of attack signatures. In anomaly detection, the system administrator defines the baseline or norms such as networks traffic load, protocol, breakdown, and packet size. The anomaly detector monitors network divides to compare with normal baseline.

- Network-based vs. host-based systems: in a network-based system or NIDS, the specific packets flowing through a network analyser. The NIDS can identify malicious packets. In a host-based system, the IDS inspects at the activity on each individual computer or host.
- Passive system vs. reactive system: in a passive system, the IDS detect a potential security breach, log the information and signal an alert. In a reactive system, the IDS respond to the doubtful activity by logging off a user or by reprogramming the firewall to obstruct network traffic from the supposed malicious source.

An IDS varies from a firewall, in a firewall looks out for intrusions in order to prevent them from happening. The firewall confines the access between networks in order to thwart intrusion and does not signal an attack from inside the network. An IDS compares a supposed intrusion once it has taken place and indicates an alarm. An IDS also watches for attacks that initiate from within a system.

Therefore in this work first kind of system is modified to make more powerful and better performance IDS. Moreover it includes a new pattern detection technique for intrusion detection.

2 BACKGROUND

This section includes the study of different intrusion detection systems which are recently developed for KDD CUP's 99 dataset classification.

Numerous researches have argued that Artificial Neural Networks (ANNs) can advance the performance of intrusion detection systems (IDS) when compared with traditional methods. However for ANN-based IDS, detection precision, particularly for low-frequent attacks, and detection stability are still required to be enhanced. In this paper, Gang Wang et al [2] propose a new approach, called FC-ANN, based on ANN and fuzzy clustering, to solve the trouble and help IDS achieve higher detection rate, less false positive rate and stronger stability. The universal procedure of FC-ANN is as follows: firstly fuzzy clustering technique is used to generate dissimilar training subsets. Subsequently, based on different training subsets, dissimilar ANN models are trained to create different base models. Finally, a meta-learner, fuzzy aggregation module, is utilized to aggregate these results. Experimental results on the KDD CUP 1999 dataset show that proposed noval approach, FC-ANN, outperforms BPNN and other well-known methods such as decision tree, the naïve Bayes in terms of detection precision and detection stability.

In this research, anomaly detection using neural network is introduced. This research aims to experiment with user behaviour as parameters in anomaly intrusion detection using a back propagation neural network. Here ManoranjanPradhan et al [3] wanted to see if a neural network is capable to classify normal traffic properly, and detect known and unknown attacks without using a large amount of training data. For the training and testing of the neural network, they used the DARPA Intrusion Detection Evaluation data sets. In final experiment, author has got a classification rate of 88% on known and unknown attacks. Compared with other researches our result is very promising.

Reyadh Shaker Naoum et al [4] is performed a study for the potential threats and attacks that can be caused by intrusions have been increased quickly due to the dependence on network and internet connectivity. In order to prevent such attacks, Intrusion Detection Systems were designed. Different soft computing based methods have been proposed for the development of Intrusion Detection Systems. In this paper a multilayer perceptron is trained using an enhanced resilient back propagation training algorithm for intrusion detection. In order to increase the convergence speed an optimal or ideal learning factor was added to the weight update equation. The performance and evaluations were performed using the NSLKDD anomaly intrusion detection dataset. The experiments results demonstrate that the system has promising results in terms of accuracy, storage and time. The designed system was capable to classify records with a detection rate about 94.7%.

Network activity has become an essential part of daily life of almost any modern person or company. At the same time the number of network threats and attacks of various types in private and corporate networks is constantly increasing. Therefore, the development of effective methods of intrusion detection is an urgent problem at the present day. In this paper Vladimir Bukhtoyarov et al [5] propose a new approach to intrusion detection in computer networks based on the use of neural networks ensembles. This approach can be implemented in distributed intrusion detection systems (IDS) which better meets the challenges of the present time, in contrast to the traditional use of neural networks in host based IDS. In the paper the basic steps of the neural networks ensembles designing are described and some of the methods to complete these steps are expounded. Peculiarities of using neural networks ensembles to solve classification problems are discussed. Then the basic scheme of neural networks ensemble approach to intrusion detection systems is proposed. Conditions and results of the experimental investigation of the proposed approach on a number of classification problems are presented, including the problem of classifying probe attacks. Possible development of the proposed approach and areas for future research are discussed in the end.

IDS can offer protection from external users and internal attackers, where traffic doesn't go past the firewall at all. The research on IDS attempted to use neural networks for intrusion detection has been carried on and will continue. Such systems were trained on normal or attack behaviour information and then detect intrusions or attacks. In this paper, Xiao Hang Yao [6] have described five kinds of Neural Network technologies that are used in IDS. An IDS combining with GA and BP is put forward, and functions of each module are detailed. Finally, a discussion of the future NN technologies, which promise to enhance the detection ability of IDS is provided.

3 PROPOSED WORK

Data mining is a task of data analysis where using computer driven algorithms are utilized to find the essential pattern from the data. There are various applications and decisions are made based on the data mining and their mining techniques, the use of data mining is in a large verity of applications such as business intelligence, research, medical data analysis and others.The main concept behind that is the application can change their face according to the application area and the kind of data required to be analyse. An intrusion detection system (IDS) is a device or software application that monitors network and/or system behaviour for malicious activities or policy violations and produces reports to a management station.

In this proposed work a data mining based IDS system is prepared and implemented, the proposed data model is able to accept the KDD cup dataset and produces the outcomes of classifiers. The proposed implementation of IDS uses the concept of supervised and unsupervised learning for improving the classification ability of detection.

In recent development there is various IDS design concepts are available in these systems the following issues are considered for improvement.

1. The implementation of IDS systems leads to store a significant amount of data for processing, the pattern detection and recognition of the real time data from this huge data is a time consuming issue.
2. Learning with large data is affecting the performance of classifiers in terms of accuracy. Therefore, using essential method required to keep preserve the performance during the data analysis and learning.
3. Due to weak learning situations there are less false alarm rate, therefore learning ability improvement also required.

In order to overcome the issues and challenges in the proposed study the following suggestions are made.

1. Pre-process data by which significant attributes and features are extracted
2. Improve classification technique by combining more then on classification approach
3. Implement data optimization technique
4. Search for strong classifier that provides the significant improvement on classification

The proposed methodology for designing an effective intrusion detection system is given in figure 1. In this technique for essential features extraction two different algorithms are adopted first K-mean clustering which performs the clustering over data and in unsupervised manner cluster whole the dataset into parts.

The verified data is produced as input to the genetic algorithm, which is a kind of by part learning. That is sometime used for performance enhancement of classifiers such as boosting. Genetic algorithm optimizes the solutions for finding the more appropriate patterns in learning datasets. These recognized patterns are then classified using KNN algorithm and performance of the algorithm is evaluated. Genetic algorithm mainly employed here for finding the similar patterns in input data, these patterns are preserved first for performing classification in addition of KNN just used for classifying the data set.

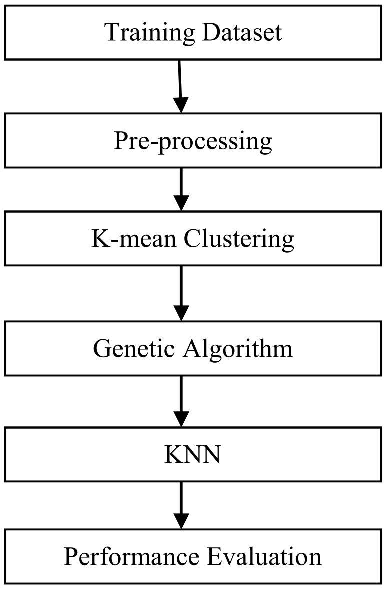

Fig. 1. proposed technique

4 ALGORITHM STUDY

This section includes the different algorithms that are used in the proposed model.

K-Means clustering

The K-Means clustering algorithm is a partition-based cluster analysis method [1]. According to the algorithm we initially choose k objects as preliminary cluster centers, then compute the distance between each object and each cluster center and allocate it to the nearest cluster, renew the averages of all clusters, replicate this process until the criterion function converged. Square error criterion for clustering

$$E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left\| x_{ij} - m_i \right\|^2$$

$x_{ij} is$ the sample j of i-class,$m_i$is the center of i-class,n_ii is the number of samples of i-class. K-means clustering algorithm is simply described as

Input: N objects to be cluster (xj, Xz…xn), number of clusters k;

Output: k clusters and the sum of variation between each object and its nearby cluster center is the small;

Process:

1. Arbitrarily select k objects as initial cluster centers$(m_1, m_2, \dots, m_k)$;
2. Compute the distance between each object Xi and each cluster center, then allocate each object to the nearby cluster, formula for calculating distance as:

$$d(x_i, m_i) = \sqrt{\sum_{j=1}^{d} \left(x_i - m_{j1}\right)^2}, i = 1 \dots N, j = 1 \dots k$$

$d(x_i, m_i)$is the distance between data i and cluster j.

3. Compute the mean of objects in each cluster as the fresh cluster centers,

$$m_i = \frac{1}{N} \sum_{j-1}^{n_i} x_{ij}, i = 1,2, \dots, K$$

$N_i$is the number of samples of current cluster i;

4. Repeat 2) 3) until the principle function E converged, return$(m_1, m_2, \dots, m_k)$Algorithm terminates.

Genetic Algorithm

The genetic algorithm uses the three main concepts for solution discovery: reproduction, natural selection and diversity of the genes.

Genetic Algorithm processes a pair of individuals these individuals are the sequence of symbols that are participating in solution space. The new generation is produced using the selection process and genetically inspired operators. The brief description of the overall search process is given as.

Generate initial population– initially the genetic algorithms are initiated with the randomly generated sequences, with the allowed alphabets for the genes. For simplifying the computational process all the generated population sequences have the same number of symbols in each sequence.

Check for termination of the algorithm–for stop the genetic algorithm a stopping criteria is required to fix for finding the optimum solution. It is possible to stop the genetic optimization process by using

1. Value of the fitness function,
2. Maximal number of iterations
3. And fixing the number of generation

Selection –that is a process of selecting the optimum symbols among all individuals, in this situation for deciding the new population two operators are used namely crossover and mutation. In this state the scaling of sequences is performed and using these best n individuals is transferred to the new generation. The elitism guarantees, that the value of the optimization function cannot produces the worst results.

Crossover –the crossover is basically the process of recombination the individuals are chosen by selection and recombined with each other. Using this new sequence is obtained. The aim is to get new population individuals, which inherit the best possible characteristics (genes) of their parent's individuals.

Mutation –the random change on some of the genes guarantees that even if none of the individuals contain the required solution genes, it is still possible to generate them using the mutation process by randomizing the search.

New generation – the selected individuals from the selection process combined with those genes that are processed with the crossover and mutation for next generation development.

K-nearest-neighbour algorithm

The K-nearest-neighbour algorithm measures the distance between a query scenario and a set of scenarios in the data base. The distance between these two scenarios is estimated using a distance function d(x,y), where x, y are scenarios developed through features, like

$$X = \{x_1, x_2, x_3, \dots\}$$

$$Y = \{y_1, y_2, y_3, \dots\}$$

The frequently used distance functions are absolute distance measuring using:

$$\mathbf{d_A(x, y)} = \sum_{\mathbf{i=1}}^{\mathbf{N}} |\mathbf{x_i} - \mathbf{y_i}|$$

And second is Euclidean distance measuring with:

$$\mathbf{d_A(x, y)} = \sum_{\mathbf{i=1}}^{\mathbf{N}} \sqrt{\mathbf{x_i^2} - \mathbf{y_i^2}}$$

The overall KNN algorithm is running in the following steps:

1. Store the output values of the M nearest neighbours to query scenario Q in vector r = {r$_1$,…...,r$_m$} by repeating the following loop M times:

   a. Go to the next scenario S$_i$ in the data set, where I is the current iteration within the domain {1……P}
   b. If Q is not set or q < d (q, S$_i$): q ←d (q, S$_i$), t ←O$_i$
   c. Loop until we reach the end of the data set.
   d. Store q into vector c and t into vector r.

2. Calculate the arithmetic mean output across r as follows:

$$\bar{r} = \frac{1}{M}\sum_{i=i}^{M} r_i$$

Return r as the output value for the query scenario q

5 CONCLUSION

In this paper a brief review on the existing IDS system design is presented. In addition of those using different classifiers a new hybrid approach is developed for finding optimum solution of IDS design. In addition of that utilized algorithm and their brief description is also reported with the proposed model. In near future the proposed IDS is implemented with the MATLAB tool and their performance study is reported.

6 REFERENCES

[1] Hari Om, AritraKundu, "A Hybrid System for Reducing the False Alarm Rate of Anomaly Intrusion Detection System", 1st Int'l Conf. on Recent Advances in Information Technology RAIT-2012, 978-1-4577-0697-4/12/$26.00 ©2012 IEEE

[2] Gang Wang, Jinxing Hao, Jian Ma, Lihua Huang, "A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering", 0957-4174, 2010 Elsevier Ltd. doi:10.1016/j.eswa.2010.02.102

[3] ManoranjanPradhan, Sateesh Kumar Pradhan, Sudhir Kumar Sahu, "Anomaly Detection using ArtificialNeural Network", International Journal of Engineering Sciences & Emerging Technologies, April 2012, ISSN: 2231 – 6604 Volume 2, Issue 1, pp: 29-36 ©IJESET

[4] Reyadh Shaker Naoum, NamhAbdulaAbid and ZainabNamh Al-Sultani, "An Enhanced Resilient Back propagation Artificial Neural Network for Intrusion Detection System", IJCSNS International Journal of Computer Science and Network Security, VOL.12 No.3, March 2012, 11

[5] Vladimir Bukhtoyarov, Eugene Semenkin, "Neural Networks Ensemble Approach for

Detecting Attacks in Computer Networks", WCCI 2012 IEEE World Congress on Computational Intelligence June, 10-15, 2012 - Brisbane, Australia

[6] Xiao Hang Yao, "A Network Intrusion Detection Approach combined with Genetic Algorithm and Back Propagation Neural Network", 2010 International Conference on E-Health Networking, Digital Ecosystems and Technologies, 978-1-4244-5517-1/10/$26.00 ©201O IEEE

[7] Sufyan T. Faraj, Al-Janabi and HadeelAmjedSaeed, "A Neural Network Based Anomaly Intrusion Detection System", 2011 Developments in E- systems Engineering, 978-0-7695-4593-6/11 $26.00 © 2011 IEEE, DOI 10.1109/DeSE.2011.19