A Comprehensive Survey on Data Mining Techniques to Extent Intrusion Detection System

Sudhanshu Gupta1,Jaswinder Singh2
Dept. of Computer Science and Engineering, SRM University, NCR Campus.
Email address:
shudhansu.gupta6@gmail.com (Sudhanshu Gupta1), jaswinder.s@ncr.srmuniv.ac.in (Jaswinder Singh2)

Abstract: Information security is most important work in current environment. Internet is widely using in each corner of the world. So to protect the data from unauthorized access effective intrusion detection system have to implement. To improve the accuracy and decrease the false alarm rate of intrusion detection system we are applying data mining techniques. The main objective of data mining technique to retrieve the data from huge data set and convert into understandable structure so that it can be use in future also. So in this paper various data mining techniques such as clustering, classification and association rules are explained that are being used for such purpose to be useful for analyzing network traffic. Positive points and negative points of data mining techniques are discussed in this paper.

Keywords: data mining, intrusion detection, genetic algorithm, neural network, Fuzzy logic, K-mean, C4.5 decision tree, KDD cup 99 Dataset.

1. Introduction:

An intrusion detection system (IDS) analyzes network traffic and monitors for unauthorized activity and alerts to the network administrator. In some cases the intrusion detection system may also reply to attack by taking action such as deny the user from using the network .Intrusion detection system come in a different of ways and approach the goal of detecting malicious traffic in different ways.

NIDS:

Network Intrusion Detection Systems are deploy at a crucial points within the network to monitor traffic in the Network. We should scan all incoming and outgoing traffic; however doing this it might generate a bottleneck that would damage speed of the network.

HIDS:

Host intrusion detection systems are executed on individual devices. A HIDS check the incoming and outgoing packets and alert the user or administrator of attack is identify.

Signature Based:

Signatures based IDS check packets on the network and compare them to a signatures or attributes from known suspicious threats. The problem is that when the new threat is come with new signature which is not in IDS database then ids is unable to identify to detect.

.

Data mining:

Data mining is defined as the process of extracting useful information from the large databases. Data mining analyses the observed sets to discover the unknown relation and sum up the results of data analysis to make the owner of data to understand. Hence data mining problems are considered as a data analysis problem. Data mining framework automatically detect patterns in our data set and use these patterns to find a set of malicious binaries.ie, Data mining techniques can detect patterns in large amount of data, such as byte code and use these patterns to detect future instances in similar data [3].

Classification:

Classification is a form of data analysis which takes each instances of a dataset and assigns it to a particular class.

Clustering:

The amount of available network data is too large hence human labeling is time-consuming and also expensive. The process of labeling data and assigning into groups (clusters) is known as clustering. The members of same cluster are quite similar and members from the different clusters are different from each other.

3. Various data mining techniques to enhance intrusion detection system:

A. Genetic Algorithm:

Genetic algorithms are search technique used to find solution to problem. Operation analogues to biological mutation selection and crossover are used to evolve and improve the solution [1].

Mutation is where random bits in an individual are randomly changed .Crossover is where two individual swap sequences of bits to form to new individuals.

Selection is where individuals that have better fitness are chosen to be parents. The fitness of an unfitness function, which determines the quality of particular individuals. Calculate the fitness of the fuzzy rule the proposed fitness function is used to optimize the rules.

To classify the data author use fuzzy algorithm instead of decision tree because fuzzy rule has an advantage over decision tree. In this paper genetic algorithm is used which is well known searching technique it will optimize the result of fuzzy rule set by the basic operations crossover and mutation. Testing on the KDDCUP 99 dataset false positive is 0.13, false negative is 1.55 and detection rate is 98.72.

B. Inductive Rule Generation:

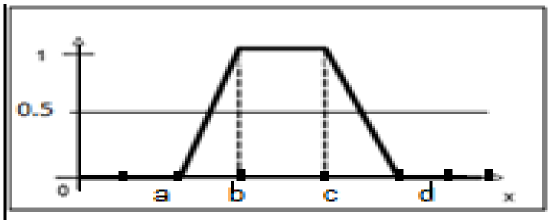
It is a rule learning program, quickly and Generates concise rule sets. One of the attractive feature of this approach is rule set which is generated by this approach is simple to understand; therefore a security analyst can verify it.

C. Neural Networks:

The application of neural networks for IDSs has been defined by a number of researchers. Neural networks give a solution to the problem of modeling the users' behavior in anomaly detection as they not require any external user model. Neural networks for IDS were first introduced as an alternative to statistical techniques in the intrusion detection expert system (IDES) to model. With earning by example approaches, attack"signatures" can be extracted automatically from labeled traffic data.

D. Fuzzy Logic:

We used a trapezoidal shape to measure a probability of being an attack identified by each attribute. The fuzzy logic is encoded into four parameter which a,b,c and d.

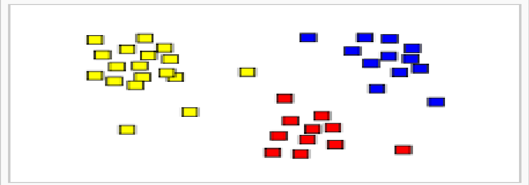


```
If b<=data<=c then
        Prob=1.0
Else if a<data<b then
        Prob=(data-a)/(b-a)
Else if c<data<d then
        Prob=(d-data)/(d-c)
Else
        Prob=0.0
End if
```

So detection rate is 98.56% and false alarm rate is 1.55%.

E.K-Mean:

K-means is one of the clustering method which is simplest unsupervised learning algorithm .The K-mean algorithm takes the input parameter, k, and partition a set of n objects into k-cluster so that the resulting intra cluster similarity is high but inter cluster similarity is low.



The result of a cluster analysis shown as the coloring of the squares into three clusters.

But the combination of the fuzzy k-mean gives the effective IDS to identify the attack. So the detection rate is 95.87% and false alarm rate is 2.124%.

F. Decision Tree:

- Classify the data using C4.5 decision tree algorithm.
- The advantage of C4.5 that it gives maximum detection rate.
- Steps followed in C4.5 decision tree algorithm…
- ➢ Construct a decision tree.
- ➢ Extract classification rule.
- ➢ Determine network behavior.

1. Constructdecisiontree:

At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other [3]. The splitting criterion is the information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision.

2. Extract classification rule:

For decision tree each branch represent a test output, and each leaf node represent category. We just follow each path from root node leaf node, the conjunction of each attribute value constitutes the antecedent of rules, and the leaf node constitutes the consequent of rules. So decision tree can easily be converted into IF-THEN rules.

3. Determine network behavior:

Check whether it intrudes or not according to classification rules.

So this algorithm give the detection rate is 99.36% and false alarm rate is 9.46%.

4. Result Evaluation Parameter:

Detection rate: correctly classify attack data from total attack data.

DR=TP/ (TP+FN)

False Alarm Rate: correctly classify normal data from total normal data.

FAR=FP/ (FP+TN)

Where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

5. Conclusion:

This survey paper presented about the various data mining technique which is apply on the intrusion detection system to enhance. Different –different data mining technique give the different results so after doing comprehensive survey we can take decision that which technique we have to apply on IDS so that it can detect attack correctly.

6. References:

[1]P.jongsuebsuk,N.Wattanapongdakorn,C.Charnsripinyo,” Improved IDS with Fuzzy Genetic Algorithm ” ,UMM Csci Senior Seminar conference,December2013 IEEE.

[2]G.V.Nadiammai,M.Hemalatha,”Effective approach towards IDS using Data Mining”- Egyptian Informatics journal Production and hosted by Elsevier. 2013

[3]VaishaliKosamkar,Sangita S chaudhari,” Improved IDS using C 4.5 decision tree and SVM “ International journal of computer science and information technology -2013 IJCSI.

[4]Idown S.A,Anyaechie Amarachi,Ajayi,” Comparartive Study of Selected Data mining algorithm used for IDS” 2013 IJSCE.

[6]Chih-Fong Tsai,Jung-Hsiang Tsai,Jui-Sheng Chou,” Centroid –Based Nearest Neighbor Feature representation for Intrusion detection”-2012-The Institute of Electronics,Information and Communication Engineeres 2012 IEICE.

[5] LI Han,” Research and Implementation of an Anomaly Detection Model Based on ClusteringAnalysis”,InternationalSymposium on Intelligence Information processing and trusted computing ,2010 IEEE

[7]RoyaEnsafi,SoheilaDehghanazadeh,Mohammad R. Akbarzadeh,” Optimizing Fuzzy K-means for Network anomaly detection using PSO.” ,International Conference on Hybrid information Technology,2008 IEEE

[8] Te-Shun and Kang K. Yen,” Fuzzy Belief K-nearest neighbor anomaly detection ” Workshop on Information Assurance United States Military academy .West point ,Ny ,June 2007 IEEE .

[9]KDD cup99 Intrusion detection dataset.http://kdd.ics.uci.edu/databases/kddcup99.html.

[10]“Data Mining Concept and Techniques” by Jiawei Han and Micheline Kamber,Publication-Elsevier, Second Edition 2006.

[11]“Data Mining Practical Machine Learning Tools and Techniques” by Ian H. Witten & Eibe Frank,Publication-Elsevier, Second Edition 2005.