International Journal of Enhanced Research in Management & Computer Applications
ISSN: 2319-7471, Vol. 5 Issue 6, June-2016, Impact Factor: 1.544

Data Mining Based Classification Using Machine Learning

Suman Hooda[1], Savita Bishnoi[2]

[1] M. tech Scholar, Department of Computer Science & Engineering, R I E M, Rohtak
[2] Head of department, Department of Computer Science & Engineering, R I E M, Rohtak

ABSTRACT

An intrusion detection system is activity that observes a network or system activities for malicious activities. IDS come in various variant and the main goal of detecting suspicious traffic in wide variety of ways. There are network based and host based intrusion detection systems. An intrusion detection system is the process for identifying attacks on network. Intrusion detection system is categorized into two types: Anomaly based and misuse based detection. The data mining techniques make it possible to observe the network and separate from the intruders such as machine learning. Different researcher works for the detection of intrusion on network. In this NSL KDD dataset is used a source of classification. The main aim is to recognize signature pattern of known attacks with better detection rate.

Keywords: Machine learning Intrusion Detection

I. INTRODUCTION

An intrusion detection system is an phenomenon or device that analyses system and network activity for unauthorized activity. Intrusion Detection System is any process or software that monitors a system or network of systems against any intrusion activity. The ultimate goal of any IDS is to catch immoral action before they do real damage. An IDS safeguard a system from attack, misuse, and any nasty activity. It will observe network activity, and configurations of a system for assailable, analyse data integrity and more. Intrusion detection system is a vital component in the security toolbox. An IDS provides three functions of monitoring, detecting and generating an Alert for nasty activity or any unauthorized access. Before ids, there was audit. Audit is defined as a process of generating, recording, and reviewing a chronological record of system events. People audit systems to accomplish a variety of goals. These goals include to assign and maintain personal accountability for system activities, to reconstruct events, to assess damage, to monitor the problem areas of the system.

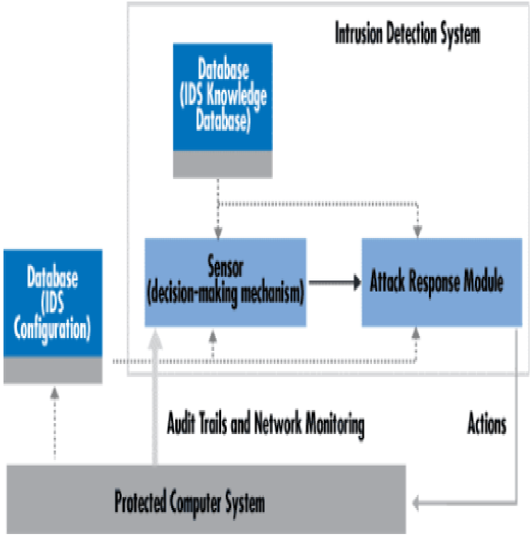

Intrusion Detection System
Database (IDS Knowledge Database)
Database (IDS Configuration)
Sensor (decision-making mechanism)
Attack Response Module
Audit Trails and Network Monitoring
Actions
Protected Computer System

Fig.1

2. CATEGORIZATION OF INTRUSION DETECTION

An intrusion detection system reviews all arriving and outbound network activity and recognizes guarded patterns that indicate a network or system attack from someone attempting to break into or compromise a system. Various classification of the Intrusion Detection System are possible as per the different criteria.

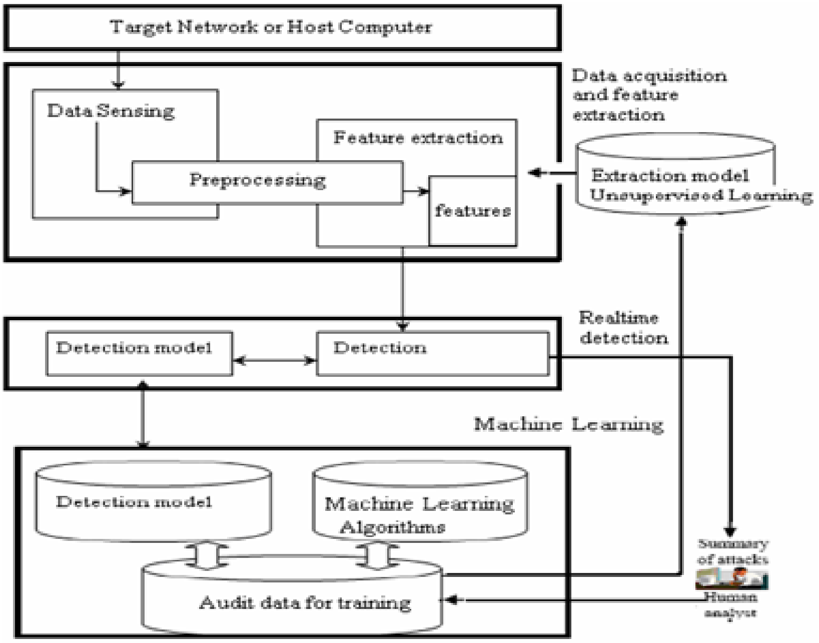


Fig. 2

2.1 Network Based Intrusion Detection System

Network based Intrusion Detection System (NIDS) monitors the traffic. It should be capable of standing against large amount of network traffic to remain effective. As network traffic increases exponentially it must grab all the traffic and analyse in a timely manner.

2. 2 Host Based Intrusion Detection System

Host based Intrusion Detection System (HIDS) keeps record of the traffic that is originated or is targeted to originate on a particular host.

2. 3 Anomaly based intrusion detection system

It will observe the ongoing traffic activity, behaviour in order to identify intrusions by detecting anomalies.

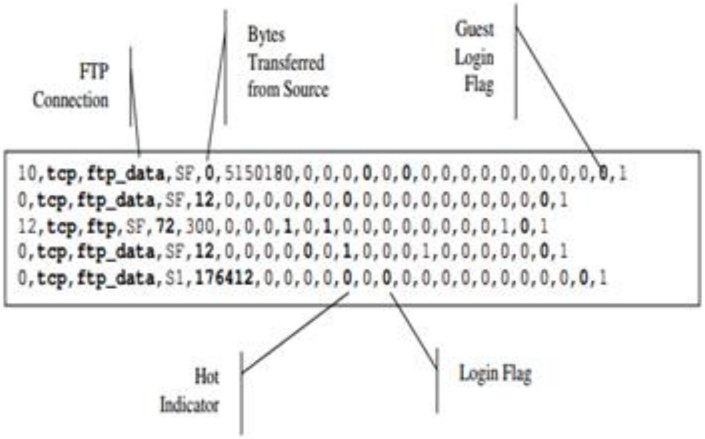
2. 4 Signature based intrusion detection system

Signature based Intrusion Detection System use a set of rule to identify patterns of events specific to known and documented attacks.

3. THE DEFICIENCIES OF CURRENT DATASET AND SOLUTION APPROACH

Dataset: NSL KDD

This dataset is basically derived to solve the problems and ambiguity found in previous KDD 99 however this new variant will also suffers from serious problems but with great advantages over 99’. This dataset will more applicable for real networks as well. Further this dataset contain number of records and test set and train data which are more reasonable over 99’. Evaluation result will be a far better and more comparable .It doesn’t include redundant records that are found in 99’.here classification will not biased for more frequent records. There is no redundant records in the proposal sets. The performance of classifiers is not biased by the method over 99’. The number of records found on the train and set are comparable make it affordable to run experiments on the complete dataset without the need to randomly selection of known small patterns. The total number of records in the given group is found inversely proportional to the percentage of records in the dataset found previously which makes it possible to have an accurate evaluation of abundant learning techniques.



```
10,tcp,ftp_data,SF,0,5150180,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1
0,tcp,ftp_data,SF,12,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1
12,tcp,ftp,SF,72,300,0,0,0,1,0,1,0,0,0,0,0,0,0,0,1,0,1
0,tcp,ftp_data,SF,12,0,0,0,0,0,0,1,0,0,0,1,0,0,0,0,0,0,1
0,tcp,ftp_data,S1,176412,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1
```

Fig. 3

4. THE PROPOSED SECURITY SOLUTION APPROACH

Learning a Bayesian networks from data requires the construction of the structure and CPT from a given database of cases. It requires to learn the structure, the parameters for the structure(i.e., the conditional probabilities among variable),hidden variables and missing values. Learning the structure is a much more challenging problem than estimating parameters for known attacks. Given a network set D={x1,x2,_ xn}, the goal is to find a Bayesian network B that approximates the join distribution *p(X)*. The network B can be found by maximizing the likelihood or log –likelihood of the data. Learning the structure of Bayesian networks be generally divided into 2 categories independence analysis or score based analysis. Exhaustive search for the best network is Np hard, even from a small sample dataset and when earch node has at most 2 parents. Identifying high scoring DAG’s from a large datasets when using a consisting criterion is also NP hard. Score based methods are more robust for small datasets, and it works with a wide range of probabilistic methods. Independence based methods do not require computation of the parameters of the model during the structure discovery process, thus are efficient they are generally more efficient than score based for sparse networks. However most of these algorithms need an exponential number of conditional independence test.

4.1 Score based methods

The Score based Bayesian networks methods such as K2, B deu. The Bayesian networks measure a score which is equivalent to the marginal likelihood of the model given the data. For most criteria, Bayesian networks structures are interpreted as independence constraints in some distribution from which data was generated. The k2 applied to DAG evaluated the relative posterior probabilities that the generative distribution has the same independence

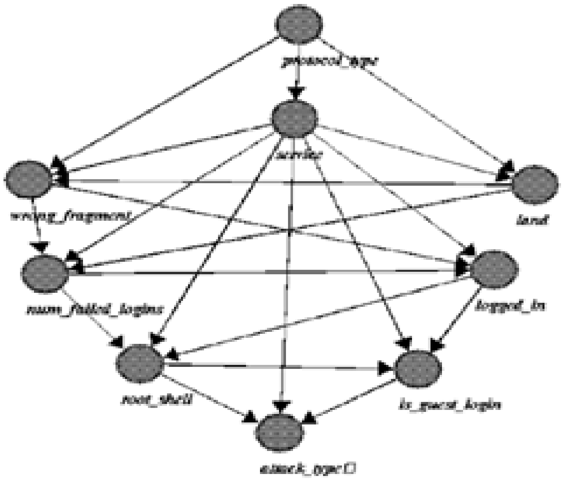
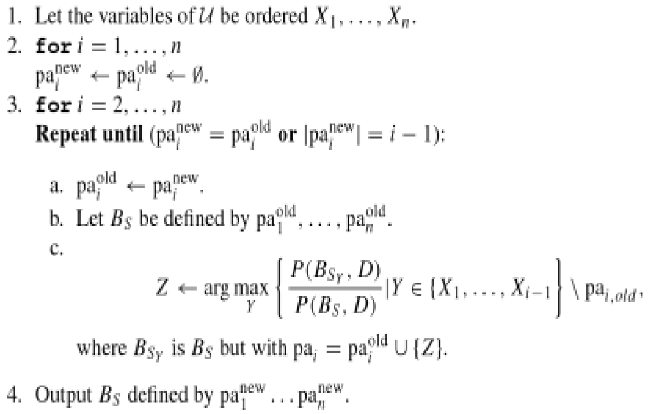


Fig. 4

constraints as those entailed. The basic idea of Bayesian network is to maximize the probabilities of network structure of given dataset i.e. to maximize the P(BS/d) over all possible networks. The algorithm will consider atmost O (n3) different structures for node n the major drawback of previous algorithm is highly dependent on the ordering on the variables taken as point of departure. The revised algorithm is as described:



```
1. Let the variables of $\mathcal{U}$ be ordered $X_1, \ldots, X_n$.
2. for $i = 1, \ldots, n$
   $\mathrm{pa}_i^{\mathrm{new}} \leftarrow \mathrm{pa}_i^{\mathrm{old}} \leftarrow \emptyset$.
3. for $i = 2, \ldots, n$
   Repeat until ($\mathrm{pa}_i^{\mathrm{new}} = \mathrm{pa}_i^{\mathrm{old}}$ or $|\mathrm{pa}_i^{\mathrm{new}}| = i - 1$):
     a. $\mathrm{pa}_i^{\mathrm{old}} \leftarrow \mathrm{pa}_i^{\mathrm{new}}$.
     b. Let $B_S$ be defined by $\mathrm{pa}_1^{\mathrm{old}}, \ldots, \mathrm{pa}_n^{\mathrm{old}}$.
     c.
        $Z \leftarrow \arg\max_{Y} \left\{ \frac{P(B_{S_Y}, D)}{P(B_S, D)} | Y \in \{X_1, \ldots, X_{i-1}\right\} \setminus \mathrm{pa}_{i,old},$
       where $B_{S_Y}$ is $B_S$ but with $\mathrm{pa}_i = \mathrm{pa}_i^{\mathrm{old}} \cup \{Z\}$.
4. Output $B_S$ defined by $\mathrm{pa}_1^{\mathrm{new}} \ldots \mathrm{pa}_n^{\mathrm{new}}$.
```

Fig. 4

For smaller database ,the MDL measures assign equal quality to networks that represent the same set of independencies wile Bayesian network measure does not.

4.3 Junction tree

This paradigm which can be used for measure any query via passing message on the tree followed by step of this paradigm creates an DAG from input via procedure called moralization. These processes keep the same edges but drop the direction and then will approach to the parents of every child thereafter.

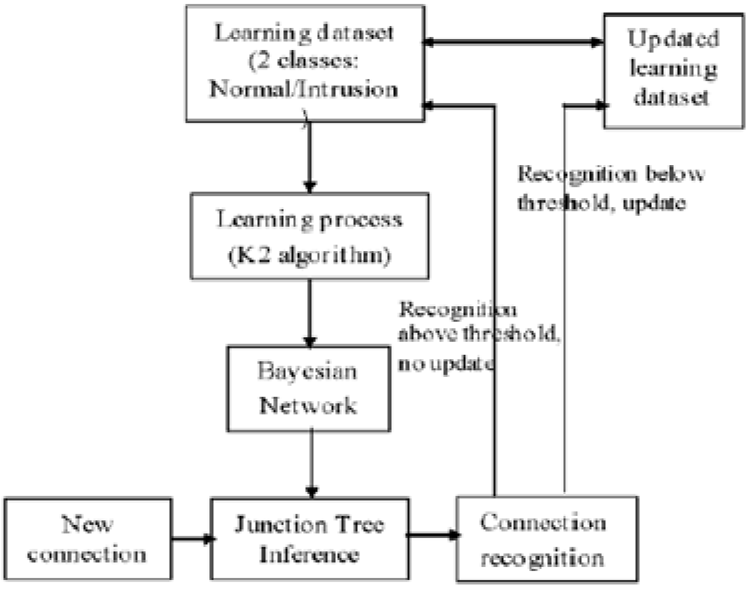

Learning dataset (2 classes: Normal/Intrusion)
Updated learning dataset
Recognition below threshold, update
Learning process (K2 algorithm)
Recognition above threshold, no update
Bayesian Network
New connection
Junction Tree Inference
Connection recognition

Fig. 5

5. Self adaptive system solution approach

In this paper we propose a environment for intrusion detection system which works on data mining source. The learning dataset can be updated by inserting new intrusions signatures for future point of view for normal connection or known attacks. The environment process begins for classifying connections of learning dataset in to 2 classes normal or abnormal by using associations rule applied. This will definitely accelerate both the learning and classifying the process for the deduction into normal or anamoly. Anamoly connection patterns will already registered and will definitely provides adaptive learning for malicious activity.

6. EXPERIMENTATION RESULTS

The main criterion we have considered in the experimentation is the detection rate.
Basically is it is equivalent to the number of examples correctly considered as normal vs predictive number of target groups.
We divide into 2 steps:

1) First step

This will identification of normal vs predictive number of cases either appear as normal or attack.

TABLE 1

Connection	Detection
Normal	87.68%
Intrusion	88.64%

2) Second step

Furthermore a better detection rated on the frequent record found as well.

(kdd train set) TABLE 2.1

Types		**Reduction rate**
Attacks		**93.32%**
Normal		**16.44%**
Total		**78.05%**

(kdd test set) TABLE 2.2

Types	**Reduction rate**
Attacks	**88.26%**
Normal	**20.92%**
Total	**75.15%**

In addition we analyzed the difficulty level of records in kdd set will provide a successful prediction up to 86% with learners.

3. Third step

TABLE 3

Intrusion type	Detection
DOS	88.64%
Probing	99.15%
R2L	20.88%
U2R/others	66.51%

TABLE 1 Shows a high staging on system in detection of Normal ,intrusion connections as well. Similarly TABLE 2 shows a high staging in detection of DOS, Probing and other connections. The low staging in r2l connections may be explained by the low fraction of R2L training connections.

CONCLUSION

In this paper, various research articles in the domain of the data mining with its application in Intrusion Detection System have studied. The three paradigm of the machine learning are used to model various real world problems, intrusion detection is one of them. In depth literature survey, it is observed this has been used in various ways in Intrusion Detection Systems like classifying TCP traffic, finding useful and high level alerts from the alerts generated by Intrusion Detection System, Modelling program behaviour using traces of system call for intrusion detection. From the comparative analysis on the various machine learning techniques for the intrusion detection, it is concluded that this method are a viable for the detection of malicious intrusions The comparison between various learning techniques will allow software professionals to find best machine learning technique to find clear, unambiguous knowledge about intrusion detection more effectively and efficiently. It can also assist in making acceptable tradeoffs among sometimes conflicting goals such as functionality, quality, cost, and time to market and to allocate resources based on the security requirements importance to the project as a whole.

REFERENCES

[1]. A. M. Chandrasekhar, K. Raghuveer " Intrusion Detection Techniques by using Fuzzy Neural network and SVM classifier", ICCCI- 2013, Jan. 04-06, 2013, Coimbatore, INDIA.

[2]. Annie George, ‗Anomaly Detection based on Machine Learning: Dimensionality Reduction using PCA and Classification using SVM‘, International Journal of Computer Applications (0975 – 8887) Volume 47– No.21, June 2012.

[3]. W.K. Lee, S. J. Stolfo. ―A data mining framework for building intrusion detection model‖, In: Gong L., Reiter M.K. (eds.): Proceedings of the IEEE Symposium on Security and Privacy. Oakland, CA: IEEE Computer Society Press, pp.120~132, 1999.

[4]. V. Jyothsna, V. V. Rama Prasad, K. Munivara Prasad, ‗A Review of Anomaly based Intrusion Detection Systems‘ International Journal of Computer Applications (0975 – 8887) Volume 28– No.7, August 2011.

[5]. Neethu B, ‗Classification of Intrusion Detection Dataset using machine learning Approaches‘ International Journal of Electronics and Computer Science Engineering 1044 ISSN- 2277-1956. Available Online at www.ijecse.org.

[6]. Lindsay I Smith, ―A tutorial on Principal Components Analysis‖.

[7]. CHEN Bo, Ma Wu, ―Research of Intrusion Detection based on Principal Components Analysis‖, Information Engineering Institute, Dalian University, China, Second International Conference on Information and Computing Science, 2009.

[8]. T. J. Hastie, R. J. Tibshirani, and J. H. Friedman. The elements of statistical learning: Data mining, inference, and prediction, Springer-Verlag, 2001.