© 2016 IJSRSET | Volume 2 | Issue 2 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Anomaly Detection in Network Using Data Mining Algorithms

Amardeep Singh, Sharanjit Singh, Simmy

G.N.D.U Regional Campus, Gurdaspur, Punjab, India

ABSTRACT

In today's world the security of computer system is of great concern. Because the last few years have seen a dramatic increase in the number of attacks, intrusion detection has become the mainstream of information insurance. Firewalls provide some protection. They do not provide full protection and still need to be complimented by an intrusion detection system. Data mining techniques are a new approach for intrusion detection. Recent studies show that as compared to the single algorithm, cascading of multiple algorithm's gives much better performance. False alarm rate was also high in such system. Therefore, combination of different algorithms is performed to solve this problem. In this paper, we use two hybrid algorithms for developing the intrusion detection system. C4.5 decision tree and Support Vector Machine are combined to achieve high accuracy and diminish the wrong alarm rate. Data mining, or knowledge discovery, is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining tools predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were too time consuming to resolve. Data mining derives its name from the similarities between searching for valuable information in a large database and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find where the value resides.

Keywords: Data Mining, Support Vector Machine, VPN, SVC, SVR

I. INTRODUCTION

In recent year's computer have been utilized by many people all over the world in several areas. With the development of internet technology, network security has become a global focus in the world. Traditional security such as firewall, VPN and data encryption is insufficient to detect attacks by malicious users. Data mining is the process of automatically discovering useful information in large data repositories. However, intrusion detection is dynamic in nature, which in turn provides dynamic security to the network in various fields like monitoring, attack and counter attack [1].

Generally data mining which is even termed as data or knowledge discovery is the process of evaluating the data from different viewpoints and summarizing it to necessary information. Data mining software is one the best analytical tools for evaluating data. It permits the users to view data from much different perspective, categorize it, and conclude with the relationships identified. Thus, in technical terms data mining is the technique of finding correlations or patterns among huge fields in large relational databases [2].

Data mining is the computer dependent technique of deeply searching through and analyzing large data sets and atlas involves the extraction of the meaning of the data. Data mining tools helps to foresee the behaviors and future scope, supporting companies to make proactive, knowledge-driven decisions. Data mining tools can easily answer questions that traditionally involved a lot of time and various other valuable resources to resolve. Data mining derives its name from the similarities between searching for valuable information in a large database and mining a mountain for valuable ore. Both processes require intelligent probing to find where the data of importance is placed [3].

II. METHODS AND MATERIAL

A. Applications of Data Mining

Data mining has wide range of applications in various diverse areas. There are numerous commercial data mining system available in the market.

Here's the list of areas where data mining is widely used [4]:

- Financial data analysis
- Retail industry
- Telecommunication industry
- Scientific application
- Biological data analysis
- Intrusion detection

Anomaly detection techniques are bifurcated into two categories [5]:

1. Anomaly Detection: It mainly refers to storing features of user's usual behaviors into database, then comparing user's current behavior with those in database. If the deviation is huge enough, we can say that there is something abnormal.
2. Misuse Detection: Misuse Detection refers to confirming attack incidents by matching features through the attacking feature library.

The data mining can be used for solving the problem of network intrusion because of following reasons:

- Data mining can process huge amount of data.
- It is beneficial to find out the ignored and hidden information.

Data mining algorithms are help to carry out data summarization and visualization which helpful in various areas and also promises security.

B. Various Types of Algorithm

- ID3 algorithm
- K-nearest neighbor
- Naïve bayes classifier
- Support Vector Machine
- C4.5 algorithm
- Apriori algorithm
- EM algorithm
- Page rank algorithm
- Adaboost algorithm
- CART(classification and regression tress)

C. Literature Survey

Denning [6] was the first person to visualize in the area of application of data mining to network security. He brought into implementation a model of a real-time intrusion-detection expert system. The concept behind the model is that exploitation of a system's vulnerabilities involves abnormal usage of system and this abnormality can be detected by looking for the abnormal patterns in the audit records. The model proposed is capable of detecting break-ins, penetrations, and other forms of computer anomaly. In this paper, we are using two methods of anomaly detection SVM (Support Vector Machine) and C4.5 that is extended version of classification algorithm ID3. Both the methods are supervised algorithm.

Xiang M.Y. Chang et.al(2004)[7],designed a multiple level tree classifier for intrusion detection system and increase the detection rate. Classifier is more efficient in case of known attacks but for unknown vulnerabilities, it gives low detection rate.

Peddabachigiri S.et.al (2007)[8],proposed a model of intrusion detection system combining decision tree and support vector machine classification techniques and produces high detection rate.

Mohammadreza Ektela et.al (2010)[9],used support vector machine and classification tree data mining technique for intrusion detection in network. They compared C4.5 and support vector machine by experimental result and found that C4.5 algorithm has better performance in term of detection rate and false alarm rate than SVM, but for U2R attack SVM performs better.

D. Data Mining Algorithms

It is targeted at supervised learning. Given an attribute valued data set where instances are described by collections of attributes and belong to one of a set of mutually exclusive classes, C4.5 learn a mapping from attribute values to classes that can be applied to classify

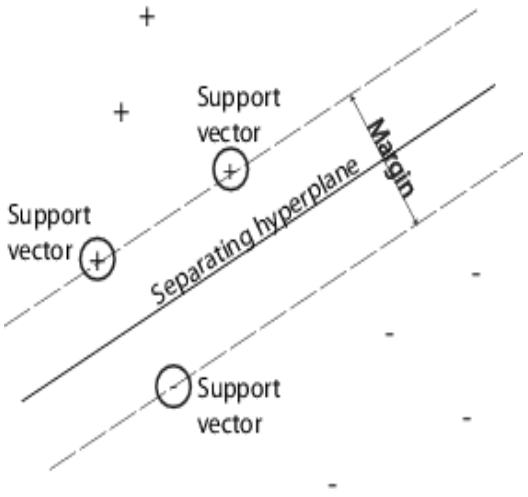
new , unseen instances. This algorithm is more applicable for continuous and discrete value attributes.

Algorithm C4.5 (D)

Input: an attribute-valued dataset *D*
1: Tree = *{}*
2: if *D* is "pure" OR other stopping criteria met then
3: terminate
4: end if
5: for all attribute *a* ∈ *D* do
6: Compute information-theoretic criteria if we split on *a*
7: end for
8: *abest* = Best attribute according to above computed criteria
9: Tree = Create a decision node that tests *abest* in the root
10: *Dv* = Induced sub-datasets from *D* based on *abest*
11: for all *Dv* do
12: Tree*v* = C4.5 (*Dv*)
13: Attach Tree*v* to the corresponding branch of Tree
14: end for
15: return Tree

Support Vector Machine

Support vector machines (SVMs), including support vector classifier (SVC) and support vector regressor (SVR), are among the most robust and accurate methods in all well-known data mining algorithms. The aim of SVC is to find a hyper plane that can separate two classes of given samples with a maximal margin which has been proved able to offer the best generalization ability.



Figure[10] : SVM

Hyper plane can be written as: $\mathbf{w}^T \mathbf{x} + b = 0$
Where W = {w1, w2, …, wn } are weight vectors for n attributes A = { A1, A2, …, An }; b is a scalar, and X ={x1, x2, …, xn} are values of attributes.

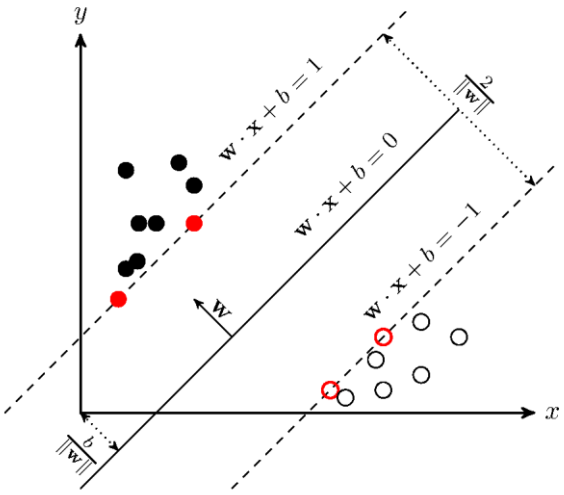


Figure [11]:

III. RESULTS AND DISCUSSION

Comparison

We have two algorithms namely C4.5 and SVM and comparison is done on the basis of detection rate and false alarm rate.
1. Detection rate: Detection rate refers to the percentage of detected attack among all attack data, and is defined as:
Detection rate = detected attack *100%All Attack data
2. False alarm rate: False alarm rate refers to the percentage of normal data which is wrongly recognized as attack, and defined as:

False alarm rate= false positive *100 % All attack data.

Algorithm	Dos	U2R	Probe	R2L
SVM	92.85	68.88	88.19	16.89
C4.5	92.87	34.44	94.48	17.44

Figure [12]:

IV. CONCLUSION

According to necessary parameter, we conclude that execution time of support vector machine is less and produces high accuracy with smaller dataset, while decision tree has high accuracy rate in case of large dataset. In this paper, two data mining algorithm techniques namely C4.5 and SVM used to detect anomaly in network. This paper conclude that C4.5 algorithm has better result than SVM in both detection rate as well as in false alarm rate but in U2R attack, SVM performs better than C4.5. In future some more exploration can be done on this area. Some other feature selection algorithm can be used that can select the more significant feature and make system more effective. Dataset should be collected for testing the system. System should be trained with the new dataset regularly so that it becomes capable of recognizing the new attacks. Algorithms can be tested with different set of options in order to achieve more effective results.

V. REFERENCES

[1] M.Xue,C.Zhu,"Applied Research on DataMiningAlgorithm in Network Intrusion Detection," International joint conference on artificial intelligence,2009.

[2] T.Bhavani et.al"Data Mining for security Application,"proceedings of the 2008 IEEE/IFIP international conference on embeded and ubiquitous computing-ol 02,IEEE computer society,2008

[3] Dorothy E.Denning."A Intrusion Detection Model" 1986 IEEE computer society symposium on research in security and privacy

[4] Xiang, M.Y.Chong and H.L.Zhu,"Design of multiple-leel tree classifiers for intrusion detection system",IEEE conference on cybernetics and intelligent system,2004.

[5] Peddabachigiri S., A.Abraham, Modeling of intrusion detecton system using hybrid intelligent systems," journals of network computer application,2007

[6] Mohammadreza Ektefa, et.al "intrusion detection using data mining techniques", pp200-203, IEEE,2010.

[7] www.pixshark.com.

[8] www.pengyifan.com.

[9] Sushil Kumar Chaturvedi, "anomaly detection in network using data mining techniques" IJETAC,2012

[10] www.anderson.ucla.edu/faculty/jason.fraud/teacher.technologies/palace/datamining.htm.

[11] www.laids.utexas.edu

[12] www.dataminingtools.net