ISSN: 2277-9655
Impact Factor: 1.852

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Proposed Algorithm for Network Traffic Classification Based On DB Scan

Shevali Agarwal[*1], Anurag Punde[2], Shubhi Kesharwani[3]
[*1,2] Department of Computer Science, Acropolis Institute of Technology & Research, Indore, India
[3]Department of Computer Science, IET-DAVV, Indore, India
shevaliagarwal@gmail.com

Abstract

The trend of using internet is increasing rapidly. Everyone wants to share their information on the urgent basis but in the secured manner. Security issues have posed the giant problems within the organization. Many researchers have given their solution for intrusion detection system. In this they are unable to find labeled and unlabeled data. Due to this the efficiency of IDS gets decreased and it generates the wrong alarm in place. In this paper we have given an algorithm called DBSCAN that increases the efficiency of the IDS system.

Keywords: DBSCAN, IDS, Network Security, Confidentiality, Network Traffic

Introduction

IDS detect intruders. It is a software and/or hardware that monitor the traffic for unauthorized and/or unwanted access to computer systems and networks. Unlike firewalls and other security measure that works to stop intruders, IDS identifies intruders that are accessing the network or those that are already in the network. At the same time, the regular security measures detect activity from outside source while IDS detects both internal and external attacks. IDS can be either passive or reactive. Passive systems are those that detect intrusion activities, create logs and report to the administrators. No actions are taken. It is up to the administrators to determine the type of response that should take place to solve the problem. Reactive systems are more commonly known as intrusion prevention systems (IPS). These systems take a further step by resetting the network connection or blocking the network intruder, either automatically or by an operator. The choice of system depends on the needs of the business.[1]

The outline of desertion is organized in six sections as follows.

Section 2: In this section we will see all the terminologies, Section 3: In this section we will see the literature work, Section 4: In this section we will see the architecture and proposed algorithm, Section 5: In this section we will see conclusion and future enhancement, Section 6: Finally the references.

Terminologies

1. **Burglar Alert/Alarm:** A signal suggesting that a system has been or is being attacked.[2]
2. **True Positive:** A legitimate attack which triggers IDS to produce an alarm.[2]
3. **False Positive:** An event signaling IDS to produce an alarm when no attack has taken place.[2]
4. **False Negative:** A failure of IDS to detect an actual attack.[2]
5. **True Negative:** When no attack has taken place and no alarm is raised.
6. **Noise:** Data or interference that can trigger a false positive.[2]
7. **Site policy:** Guidelines within an organization that control the rules and configurations of IDS.[2]
8. **Site policy awareness:** An IDS's ability to dynamically change its rules and configurations in response to changing environmental activity.[2]
9. **Confidence value:** A value an organization places on an IDS based on past performance and analysis to help determine its ability to effectively identify an attack.[2]
10. **Alarm filtering:** The process of categorizing attack alerts produced from an IDS in order to distinguish false positives from actual attacks.[2]
11. **Detection Rate:** The detection rate is defined as the number of intrusion instances detected by the system (True Positive) divided by the total number of intrusion instances present in the test set.[3]
12. **False Alarm Rate:** defined as the number of 'normal' patterns classified as attacks (False Positive) divided by the total number of 'normal' patterns.[3]
13. **Intruder:** An entity who tries to find a way to gain unauthorized access to information, inflict harm or engage in other malicious activities.
14. **Masquerader:** They are generally outside users. They don't have right to access the system, but tries to admittance the information as the authorized user.[3]

ISSN: 2277-9655
Impact Factor: 1.852

15. **Misfeasor:** They are commonly internal users and can be of two types:
    - An authorized user with limited permissions.
    - A user with full permissions and who misuses their powers.
16. **Clandestine user:** A user who acts as a administrator and tries to use his rights so as to avoid being captured.[3]
17. **Clandestine user:** A user who acts as a administrator and tries to use his rights so as to avoid being captured.[3]

Literature Review

A. Network Traffic Classification Using K-means clustering" [4] In this paper Network traffic classification and application identification provide important benefits for IP network engineering, management and control and other key domains. Current popular methods, such as port-based and payload-based, have shown some disadvantages, and the machine learning based method is a promising one.

B. "Traffic anomaly detection is using k-means clustering" [5] in this paper, Data mining techniques make it possible to search large amounts of data for characteristic rules and patterns. If applied to network monitoring data recorded on a host or in a network, they can be used to detect intrusions, attacks and/or anomalies. This paper gives an introduction to Network Data Mining, i.e. the application of data mining methods to packet and flow data captured in a network, including a comparative overview of existing approaches.

C. "Offline/Online Traffic Classification Using Semi-Supervised Learning" [6] in this paper, Identifying and categorizing network traffic by application type is challenging because of the continued evolution of applications, especially of those with a desire to be undetectable. The diminished effectiveness of port-based identification and the overheads of deep packet inspection approaches motivate us to classify traffic by exploiting distinctive flow characteristics of applications when they communicate on a network. In this paper, we explore this latter approach and propose a semi-supervised classification method that can accommodate both known and unknown applications.

D. "Network Traffic Classification is using Semi-Supervised Approach", [7] in this paper, A semi-supervised approach for classification of network flows is analyzed and implemented. This traffic classification methodology uses only flow statistics to classify traffic. Specifically, a semi-supervised method that allows classifiers to be designed from training data consisting of only a few labeled and many unlabeled flows. The approach consists of two steps, clustering and classification. Clustering partitions the training data set into disjoint groups ("clusters"). After making clusters, classification is performed in which labeled data are used for assigning class labels to the clusters.

Architecture & Proposed Algorithm

Proposed method follows the necessary steps required to perform in Intrusion Detection

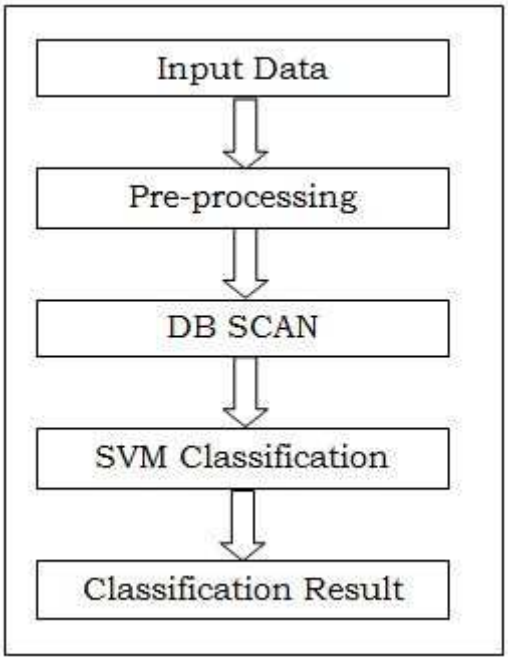

Figure 1 : Architecture

Proposed Algorithm

Existing techniques are used the Clustering classifier, neural network classifier and Bayesian classifier. These approaches have the problem of over-fitting. SVM classifier removes the problem of over-fitting, but it uses the all features or maximum features of dataset to find the accuracy rate, so there are increases the problem of data redundancy and it consumes more computer recourses.

To remove the drawbacks of existing approaches, DBSCAN approach is proposed. In this proposed approach, Intrusion Detection System is implemented by using DB SCAN and Support Vector Machine technique. [8,9,10,11] SVM is used as classifier in this system.

In the proposed approach, accuracy rate, false positive rate and attack detection rate of intrusion

detection using DB SCAN and support vector machine are tried to find out. Since support vector machine classifier support only numeric data, so in the proposed approach, firstly text features of dataset are converted into numeric data. Then redundant features are reduced to get small volume size dataset. Finally, selected features are passed to the support vector machine to get the accuracy rate, false positive rate and attack detection rate of dataset. The proposed algorithm has following steps.

- Data Preprocessing.
- DB SCAN Algorithm.
- Intrusion SVM Classification

Conclusion & Future Work

As computers are becoming increasingly used by businesses, security issues have posed a big problem within organizations. Firewalls, antivirus software, password control are amongst the common steps people take towards protecting their systems. However, these preventive measures are not perfect. Firewalls are vulnerable; they may be improperly configured or may not be able to prevent new types of attacks. Antivirus software works only if the virus is known to the public. Passwords can be stolen and therefore, systems can be easily hacked into. Hackers can change the system on initial access and manipulate it so that their future access will not be detected2. In these situations, intrusion detection systems (IDS) come into play. Even though security issues are more commonly linked to internal management concern, auditors should also be aware of these issues with clients' businesses. With security problems, intruders may be able to change the data so that it is not representational of the client business. It is possible that discrepancies in data are a result of foul play rather than fraud.

The aim of the project is to design and implement a semi-supervised learning approach for network traffic classification and it has been achieved successfully. A DB SCAN approach to design a Network Traffic Classifiers is implemented successfully. Algorithm permits both labeled and unlabeled data to be used in training the network. While performing training and testing of the classifier for a dataset, it is observed that a test error rate depends on the number of clusters which is randomly used in training phase. We have used the KDD Data Set as a training data set and improved the accuracy rate of the semi supervised algorithm.
Proposed algorithm is very apt and reliable for finding the supervised and unsupervised data. The algorithm has been proved that in the future Of course, we can improve accuracy rate, false positive rate and attack detection rate of intrusion detection by providing the some improved form of the DB SCAN algorithm.

References

[1] S. V. Sabnani, "Computer Security: A Machine Learning Approach", 2008.
[2] Nitin.; Mattord, Verma (2008). Principles of Information Security. Course Technology. pp. 290–301.ISBN 978-1-4239-0177-8.
[3] wwusers.cs.york.ac.uk / ~jac / PublishedPapers/ AdhocNetsFinal.pdf.
[4] Liu Yingqiu, Network Traffic Classification Using K-means clustering, published in Computer and Computational Sciences, 2007. IMSCCS 2007. Second International Multi-Symposiums on 13/Aug/2007. ISBN: 978-0-7695-3039-0
[5] Münz, Gerhard, Sa Li, and Georg Carle. "Traffic anomaly detection using k-means clustering." Proc. of Leistungs-, Zuverlässigkeits-und Verlässlichkeitsbewertung von Kommunikationsnetzen und Verteilten Systemen 4 (2007).
[6] J. Erman, A. Mahanti, M. Arlitt, I. Cohen, and C. Williamson, "Offline/Online Traffic Classification Using Semi-Supervised Learning", Technical report, University of Calgary, 2007.
[7] Amita Shrivastav, Aruna Tiwari, Network Traffic Classification using Semi-Supervised Approach published in Machine Learning and Computing (ICMLC), 2010 Second International Conference on 9/feb/2010.
[8] Amita Shrivastav, Network Traffic Classification using Semi-Supervised Approach published in 2010 Second International Conference on Machine Learning and Computing.
[9] Martin Ester,Han-peter Kriegel,Jorg Sander, Xiaowei Xu,"A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", 2nd International conference on Knowledge Discovery and Data Mining (KDD-96).
[10] http://en.wikipedia.org/wiki/MATLAB.
[11] Amita Shrivastav, Aruna Tiwari, Network Traffic Classification using Semi-Supervised Approach published in Machine Learning and Computing (ICMLC), 2010 Second International Conference on 9/feb/2010.
[12] The Study of Network Traffic Identification Based on Machine Learning Algorithm

ISSN: 2277-9655
Impact Factor: 1.852

published in Computational Intelligence and Communication Networks (CICN), 2012 Fourth International Conference on Nov, 2012.

[13] Application of Clustering Algorithms in Ip Traffic Classification published in Intelligent Systems, 2009. GCIS '09. WRI Global Congress on May 2009.

[14] The Anomaly Detection by Using DBSCAN Clustering with Multiple Parameters published in Information Science and Applications (ICISA), 2011 International Conference on April 2011.

[15] Internet Traffic Classification Using DBSCAN published in Information Engineering, 2009. ICIE '09. WASE International Conference on July 2009.