Network Intrusion Detection using Data Mining Technique

Sivaranjani S[1], Mr. Ravi Pathak[2], Vaidehi.V[3]
Anna University, MIT Campus, Chennai

Abstract - **In recent years, most of the research has been done in the field of Intrusion Detection System (IDS) to detect attacks in network traffic indicating malicious activity. IDS can be built in two different ways: Signature based and Anomaly based. This paper presents anomaly based network intrusion detection model using Density Based Spatial Clustering for Applications with Noise (DBSCAN) technique called DNIDS (DBSCAN based Network Intrusion Detection System) for the identification of abnormality in the network traffic dataset. In case of detecting attacks the rule generation algorithm is utilized to frame a new attack and attacks are identified using detection model. The generated rules for normal and abnormal data are used to form a new abnormal record in the dataset. DNIDS uses less memory space and gives better performance comparing with other techniques. This is achieved by reducing the features present in the network intrusion dataset. Feature extraction is done using the Correlation Feature Selection (CFS) method. The proposed DNIDS is validated using well known KDD Cup1999 Data (KDD 99) and the performance is compared with k-means, Expectation Maximization (EM) based on their accuracy in clustering the data and the time they take to detect intrusion. It is found that DNIDS gives better accuracy in detecting anomalies.**

Index Terms: Correlation Feature Selection, data preprocessing, density based clustering, EM, k-means.

I. INTRODUCTION

In the modern era, Computing, Communication and Networking play a vital role. From a small sensor node to very big high performance computer and super computer need network to exchange the information. Huge amount of information is created, stored, processed and communicated among computers using networks. Information communicated using networks are vulnerable to several types of attacks. Most common types of attacks are active attacks and passive attacks. Alteration of system resources or degrading the system performance comes under active attacks, whereas passive attacks tries to learn the information available in the system. Identifying these attacks is very important, since these attacks can degrade the performance of computers. Many network security and dynamic security policies are developed to overcome these types of attacks.

Intrusion Detection System (IDS) is used for monitoring and detecting the intrusion in the network. Intrusion detection [3] can be done in two different ways. One is based on the signatures whereas another one is based on anomaly. Signature-based IDS generate “signatures” based on characteristics of previous known attacks. Anomaly based IDS [6, 10, 11] can detect previously undocumented threats. In general, intrusion detection algorithm takes collection of object, record, event, sample, entity or points as input. The algorithm takes the collection of records as input data which are captured packets over the network. Now-a-days the sources of streaming data are increasing. Unlike static data mining, the data points in stream mining are continuous. In general, these data streams cannot be processed with limited amount of memory. It requires large memory to process it. This is not possible in real time. Hence optimization of data storage plays vital role in the identification of attacks over a network. To cope with this kind of stream environment a novel technique DNIDS has been proposed in this paper. This technique uses data preprocessing technique [4] and it applies the DBSCAN (Density based spatial clustering of applications with noise) for intrusion detection. This technique also utilizes the rule based classification technique [17] called PARTial decision tree (PART) for the creation and identification of new attacks. Both classification and clustering are applied in the proposed work for less memory storage with better performance.

This paper is organized as follows: Section II explains the related work which helps in making decision about what kind of Intrusion detection technique can be used. Section III discusses the proposed work and the architecture. Section IV presents the experimental evaluation and the conclusions are discussed in Section V.

II. RELATED WORK

This section describes about some of the existing techniques in the field of identifying attacks over network and its drawbacks.

a. Multistage Attack Detection System:

The multistage attack detection system [14] was developed to discover, visualize and predict behavior pattern of attackers in network based system which is based on alerts generated by Snort, one of the widely used open source IDS, had a simple architecture with four steps: Generation of Snort alerts, Clustering & Association rule generation, Profiling attacker’s behavior and finally the Prediction and Visualization of multistage attacks. The detection system is based on the alerts generated by snort and association rules created by the Apriori algorithm. The major drawbacks of the system is that it is not suitable for finding unknown attacks and it is not capable of managing real time data in a more vulnerable environment.

b. An Intrusion Detection System:

A decision tree model based IDS is proposed in [8]. It is a methodology towards developing a more-robust Intrusion Detection System through the use of data-mining techniques and anomaly detection. These data mining techniques dynamically model what a normal network should look like and reduce the false positive and false negative alarm rates in the process. This paper used classification-tree techniques to accurately predict probable attack sessions. The main drawbacks of the system are that the system is trained only for TCP packets and not suitable for real time network traffic.

c. *Network Anomaly Detection Based on TCM-KNN Algorithm:*

Transductive Confidence Machines (TCM) [7] introduced the computation of the confidence using algorithmic randomness theory. It uses K-Nearest Neighbors algorithm for detecting anomalies. It showed good performance even when there is interference by the presence of “noise” in training set. The major drawback of the system is although the TCM-KNN algorithm is good for detecting intrusions according to the network flow. It is not suitable for our anomaly detection for the following reasons: In TCM-KNN, it is sure that the point examined belongs to one of the classes. In most cases of intrusion detection, it is very difficult to get purely “clean data” or “attack data” for training phase. We are only sure if the data we handle are normal or abnormal. It is unrealistic that we may acquire exact characteristics and patterns of all the attacks such as DoS, Probe, U2R, etc.

d. *Anomaly Extraction in Backbone Networks Using Association Rules:*

Daniela Brauckhoff, et al [4] proposed a novel method to identify anomalous flows that combines and consolidates information from multiple histogram-based anomaly detectors. The basic assumption behind this approach is that frequent item-sets in the pre-filtered data are often related to the anomalous event. A large part of evaluation results is devoted to the verification of this assumption and shows that this is indeed true. This model uses meta-data provided by several histogram-based detectors to identify suspicious flows, and then apply association rule mining to find and summarize anomalous flows. The drawback involved is optimizing the scalability and efficiency of frequent item-set mining for dealing with big network traffic data including stream processing, which is an open problem.

The above mentioned papers discuss about the Intrusion Detection techniques [13, 16] which can be applied to detect the network intrusions but the real problem starts when we preprocess the given dataset. The input data should be preprocessed before they enter into the detection section. The existing techniques do not deal with data preprocessing. Hence there is a need for novel data preprocessing method with known attack detection in a better manner.

Cluster Analysis

The process of categorizing a large number of data points into groups where all members in the group are similar in some manner is called clustering. In other words, clustering is the process of grouping the data objects based on maximizing the intra-class similarities and minimizing the inter-class similarities.

This section describes about the performance of three different clustering techniques namely Simple k-means [1, 2], Expectation Maximization (EM) and Density based clustering approaches. In one hand Simple k-means uses a parameter ‘k’ and divide ‘n’ data points into ‘k’ clusters, to create relatively high similarity in the cluster and, relatively low similarity between the clusters. And minimize the total distance between the values in each cluster to the cluster centre. The cluster centre of each cluster is the mean value of the cluster. The calculation of similarity is done by mean value of the cluster objects. In another hand EM is used for unrecorded (missed) data in a statistical model. It is an iterative model in which two steps called expectation and maximization are carried out. Expectation step calculates for log-likelihood and in the maximization step the log-likelihood value calculated in the expectation step for unobserved data is maximized. The maximization step is repeatable one. In density based techniques, clusters are identified by looking at the density of points. Regions with a high density of points depict the existence of clusters which are denoted as normal points whereas regions with a low density of points indicate clusters of noise or clusters of outliers. Hence distance based clustering technique is used for intrusion detection.

III. PROPOSED NETWORK INTRUSION DETECTION SYSTEM

Problem Statement: The major problem in the Signature based NIDS is that they are capable of identifying attacks for which signature is already available but identification of unknown threats are not possible. The current anomaly detection system lacks ability to deal with gradual or abrupt change in data flows. To overcome these issues, a density based clustering of applications with noise (DBSCAN) algorithm is used with the data preprocessing technique for utilizing less memory and giving a better performance.

In the proposed design, both classification method called rule based classifiers and density based clustering technique are used to detect the intrusions over the network.

Proposed Architecture (NIDS)

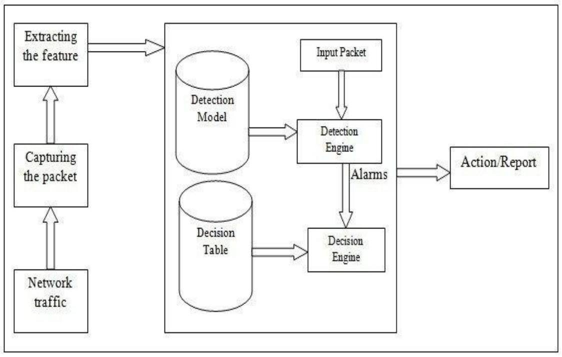

Figure 1 Network Intrusion Detection System

Figure 1 shows the architecture of the proposed NIDS system. Packets are captured from the network traffic and features are extracted. The known attack detection module is used to detect attacks using known profiles. The extracted features can be given as input to the Detection model which is formulated by using density based clustering technique. Detection engine can generate the result for the given input as to find which cluster (normal/abnormal) it belongs to and then the decision engine can take actions and report the attack.

Feature Extraction

The Correlation Feature Selection (CFS) measure evaluates subsets of features on the basis of the following hypothesis: "Good feature subsets contain features highly correlated with the classification, yet uncorrelated to each other". The following equation (1) gives the merit of a feature subset *S* consisting of *k* features:

$$CFS = \max_{S_k} \left[\frac{r_{cf_1} + r_{cf_2} + \cdots + r_{cf_k}}{\sqrt{k + 2(r_{f_1f_2} + \cdots + r_{f_if_j} + \cdots + r_{f_kf_1})}} \right] . \quad Merit_{S_k} = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \qquad (1)$$

Here, $\overline{r_{cf}}$ is the average value of all feature-classification correlations, and $\overline{r_{ff}}$ is the average value of all feature-feature correlations. The CFS criterion is defined as follows:

(2)

In equation (2) the r_{cf_i} and $r_{f_if_j}$ variables are referred to as correlations, but are not necessarily Pearson's correlation coefficient or Spearman's ρ. Dr. Mark Hall's dissertation uses neither of these, but uses three different measures of relatedness, Minimum Description Length (MDL), symmetrical uncertainty, and relief.

Let x_i be the set membership indicator function for feature f_i;. Then the above can be rewritten in equation (3) as an optimization problem:

$$CFS = \max_{x \in \{0,1\}^n} \left[\frac{(\sum_{i=1}^{n} a_i x_i)^2}{\sum_{i=1}^{n} x_i + \sum_{i \neq j} 2b_{ij} x_i x_j} \right] \qquad (3)$$

These combinatorial problems are, in fact, mixed 0–1 linear programming problems that can be solved by branch-and-bound algorithms.

The whole intrusion dataset can be reduced into a small set consisting of the following features protocol type, service, flag, src_bytes, dst_bytes, land,wrong_fragment, root_shell, count, diff_srv_rate, dst_host_same_src_port_rate.

The experiment starts by building the learning model for the given input dataset. The learnt model is being used to group the dataset into normal and abnormal clusters.

Detection Model

Algorithm

Let X-be the finite set of data points to be clustered. DBSCAN requires two user defined parameters: ε (eps) and the minimum number of points to form a cluster (minPts).

Core Object: Object with at least MinPts objects within a radius, eps - neighborhood (all the points within ε-distance).

Border Object: Object that is on the border of a cluster

The steps of DNIDS are as follows:

1. Select an arbitrary point ***p*** from the set X.
2. Retrieve all points density-reachable from the point ***p*** with respect to e*ps* and m*inPts*.
3. If ***p*** is a core point, a cluster is formed.
4. If ***p*** is a border point, no points are density-reachable from ***p*** and DBSCAN visits the next point of the database.
5. Continue the process until all of the points have been processed.

From the detection model, two clusters are formed namely normal and abnormal. Rule based technique PART is utilized to generate rules for both normal and abnormal data inside the dataset. The detection model is used to detect the new kind of attack in the dataset.

IV. EXPERIMENTAL EVALUATION

The proposed DNIDS is validated using WEKA [9, 12, 15]. WEKA is a machine learning tool that implements data mining algorithm using the JAVA language.

A dataset is roughly equivalent to a two-dimensional spreadsheet or database table. Dataset consists of 5036 instances with 12 attributes and the performance is analyzed using Weka. We can also visualize the dataset in the form of graph with red and blue code. By experimental analysis we can decide which technique is suitable for noise detection.

This experimental evaluation has been done by running the k-means, EM and DNIDS techniques on the KDD dataset. The performance of each method is analyzed and the best suitable technique for attack detection is identified. The records in the KDD dataset are modified in order to make a new attack by analyzing the rule set generated by PART algorithm. The modified dataset is checked with the generated DNIDS technique to identify the new attacks that are inserted inside the dataset.

The comparative analysis of DNIDS, EM and K-means is shown in Table 1. It is found that the density based clustering performs well and it runs faster than the other algorithms. In Simple K-means the number of clusters needs to be specified in advance. When we increase the number of cluster the time taken to run the K - means algorithm is more. And so the running time for large dataset is high. The main drawback of K-means is that it does not handle noise in the dataset. Expectation Maximization (EM) requires more time to cluster the data instances. Also its accuracy is lesser than the other methods.

Table 1 Comparative analysis of clustering techniques

Technique	Running Time (Sec)	Accuracy (%)	Incorrectly clustered instances (%)
Simple K-means	12.64	71.32	28.67
EM	1525.29	49.58	50.41
DNIDS	279.43	78.33	21.66

Figure 2 shows the accuracy of each clustering technique.

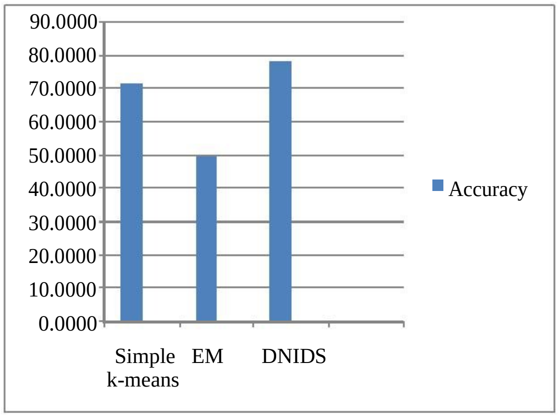

Figure 2 Performance comparisons of clusters

When we look at the performance of DNIDS , even though the running time is more and accuracy is moderate, it can generate arbitrary clusters and there is no need to define the number of clusters before we run the algorithm. It is well suitable for identifying new attacks compared to K-means and EM.

Once the best technique has been identified, the dataset can be given as input for the cluster model to be generated. This learning model can cluster the data instances as normal or abnormal.

Rule Generation:

The experiment utilizes the classification technique for the identification of new attacks in the dataset. PART is one of the classification algorithms that generate the rules for the normal and abnormal data.

The following are the some of the rules generated by the PART for the reduced dataset with 494021 data instances:

rule (1) : Normal
src_bytes > 28 AND src_bytes <= 40494
AND dst_host_same_src_port_rate <=
0.99 AND wrong_fragment <= 0 AND
flag = SF AND src_bytes
<= 1011 AND dst_bytes
> 0 AND root_shell <= 0
AND
diff_srv_rate <= 0.69 AND
Service = http: normal.
(55510.0) **rule (2) : Abnormal**
diff_srv_rate > 0.03 AND
src_bytes <= 0 AND count <= 303 AND
dst_host_same_src_port_rate <= 0.06
AND diff_srv_rate <= 0.89 AND
dst_host_same_src_port_rate <= 0.01: abnormal. (106928.0).

For the experimental analysis DataSample database is taken with 20000 instances and 13 attributes. The rules are given as follows.

Abnormal
Rule (1):
Count > 66 AND
Instance_Number Numeric > 7810: abnormal (9313.0)
Rule (2):
Count > 1: abnormal
(683.0/1.0) : abnormal (4.0)
Normal
Rule (1):
Instance_Number Numeric <= 38738
AND Count <= 67 AND
dist_bytes > 0 AND
Service = http: normal. (9159.0)
Rule (2)
Instance_Number Numeric <= 7793
AND Flag = SF AND
dist_bytes <= 1260: normal.
(805.0) Root_shell <= 0: normal.
(36.0/1.0) : abnormal (4.0)

From these three rules we can conclude that the record whose Instance_Number Numeric is greater than 7810 and Count is greater than 66 becomes abnormal while count is less than or equal to 67 is classified as normal.

Insertion of new attacks

Based on the rules generated in the previous step, new abnormal records are inserted. After inserting one record with,

Instance Number: 7900;
Count: 67
as an attack inside the DataSample, the abnormal record can be identified (clustered) using the detection model.

The result derived from DNIDS technique can be used as a model for the dataset. This model can help in the identification of the clusters to which the input data belongs. The new attack generated can also be identified by this model.

V. CONCLUSION

In this paper, anomaly based Network IDS based on DBSCAN is presented. The performance of DNIDS is compared with simple k-means, Expectation Maximization (EM). It is observed that density based method can be applied for noisy data. Also by preprocessing the data the whole dataset has been reduced into small dataset and few features which are important in the identification of attacks are extracted. DNIDS technique identifies new kind of attacks in the dataset. Hence it is concluded that the combination of classification and clustering can lead to a better performance of identification of attacks over a network.

REFERENCES

[1] Awan Dhawan, “Data mining with Improved and efficient mechanism to detect the Vulnerabilities using intrusion detection system”, ISSN: 2278 – 1323 International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 2, February 2013.

[2] Chakraborty, S., Nagwan, N. K., & Dey, L. (2011), “Performance Comparison of Incremental K-means and Incremental DBSCAN Algorithms”, International Journal of Computer Applications, 27.

[3] Chandola, V., Banerjee, A., & Kumar, V. (2009). “Anomaly detection: A survey”, ACM Computing Surveys (CSUR), 41(3), 15.

[4] Daniela Brauckhoff, Xenofontas Dimitropoulos, Arno Wagner, and Kavé Salamatian, “Anomaly Extraction in Backbone Networks Using Association Rules” - IEEE/ACM Transactions On Networking, Vol. 20, No. 6, December 2012.

[5] Jonathan J. Davis, C3I Division, DSTO, Andrew J. Clark, Information Security Institute, QUT, “Data preprocessing for anomaly based network intrusion detection: A review”, computers & security 30 (2011) 353-375.

[6] K. Wang, “Network Payload-Based Anomaly Detection and Content-Based Alert Correlation,” - Columbia University, New York, 2006.

[7] Li, Y., Fang, B., Guo, L., & Chen, Y. (2007, March), “Network anomaly detection based on TCM-KNN algorithm” In Proceedings of the 2nd ACM symposium on Information, computer and communications security (pp. 13-19)

[8] LTC Bruce D. Caulkins, “A Dynamic Data Mining Technique for Intrusion Detection Systems” - ACM (2008).

[9] Narendra Sharma, Aman Bajpai, Mr. Ratnesh Litoriya, Department of computer science, Jaypee University of Engg. &Technology., "Comparison the various clustering algorithms of weka tools", International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, Volume 2, Issue 5, May 2012).

[10] M. Mahoney and P. Chan, "Learning Non Stationary Models of Normal Network Traffic for Detecting Novel Attacks," ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, July 2002, pp. 376-385.

[11] M. Mahoney, "Network Traffic Anomaly Detection Based on Packet Bytes," ACM-SAC, Melbourne, 2003 pp. 346- 350.

[12] Nidhi Singh and Divakar Singh, "Performance Evaluation of K-Means and Hierarchal Clustering in Terms of Accuracy and Running Time" - (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (3), 2012, 4119-4121.

[13] R. Perdisci, D. Ariu, P. Fogla, G. Giacinto and W. Lee, "McPAD: A Multiple Classifier System for Accurate Payload-based Anomaly Detection," Computer Networks, Special Issue on Traffic Classification and Its Applications to Modern Networks, Vol. 5 No. 6, 2009, pp. 864-881.

[14] Rajeshwar Katipally, Wade Gasior Dept.of Comp. Sci. and Eng., University 2010 ,"Multistage Attack Detection System for Network Administrators Using Data Mining", ACM 978-1-4503-0017-9.

[15] Sanjay Chakraborty and Prof. N.K.Nagwani Lopamudra Dey National Institute of Technology (NIT) Raipur, CG, India, International Journal of Computer Applications (0975 – 8887), "Performance Comparison of Incremental K-means and Incremental DBSCAN Algorithms", Volume 27– No.11, August 201114.

[16] Wang, K., & Stolfo, S. J. (2004, January), "Anomalous payload-based network intrusion detection In Recent Advances in Intrusion Detection" (pp. 203-222) - Springer Berlin Heidelberg.

[17] Yang Li, Binxing Fang Institute of Computing Technology Chinese Academy of Sciences No.6 South of Kexueyuan Road Beijing, P.R. China, 2007, "Network Anomaly Detection Based on TCM-KNN Algorithm", ACM 1-59593-574-6/07/0003.

[18] Zhang, G. P. (2000). Neural networks for classification: a survey. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 30(4), 451-462.