International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 10 | Oct -2017 www.irjet.net p-ISSN: 2395-0072

Network based Intrusion Detection System using Filter based Feature Selection Algorithm

SARANYA.K[1] , PRABHU.R[2], Dr.RAMESH KUMAR.M[3],PREETHI.P[4]

[1,2,4]*Assistant Professor, Department of computer science Engineering,V.S.B Technical Campus,Tamilnadu,India*
[3] *Associate Professor, Department of computer science Engineering,V.S.B Technical Campus,Tamilnadu,India*

Abstract – *An Redundant and irrelevant features in data have caused a long-term problem in network traffic classification. These features not only slow down the process of classification but also prevent a classifier from making accurate decisions, especially when coping with big data. In this paper, we propose a mutual information based algorithm that analytically selects the optimal feature for classification. This mutual information based feature selection algorithm can handle linearly and nonlinearly dependent data features. Its effectiveness is evaluated in the cases of network intrusion detection. Most of the previously proposed methods suffer from the drawback of K-means method with low detection rate and high false alarm rate. This paper presents a hybrid data mining approach encompassing feature selection, filtering, clustering, divide and merge and clustering ensemble. A method for calculating the number of the cluster centroid and choosing the appropriate initial cluster centroid is proposed in this paper. The IDS clustering ensemble is used to achieve high accuracy.*

Key Words: **selection, filtering, clustering, clustering ensemble, IDS**

1. INTRODUCTION

It is essential to find an effective way to protect it as more and more sensitive information is being stored and manipulated online. The network based attacks can also be referred as some kind of intrusion. An intrusion can be defined as "any set of actions or a type of attack that attempt to compromise the confidentiality, availability, or integrity availability of a resource". For controlling intrusions, intrusion detection systems are introduced. An Intrusion Detection System (IDS) [1] is a defense system that plays an important role to protect or secure a network system and its main goal is to monitor network activities automatically to detect malicious attacks. Intrusion detection system (IDS) is increasingly becoming a vital and critical component to secure the network in today's world of Internet

1.1 Feature Selection Process

1.1.1 Steps for feature selection

The four key steps of a Feature selection process are feature subset generation, subset evaluation, stopping criterion and result validation. The feature subset generation is a heuristic search process which results in the selection of a candidate subset for evaluation. It uses searching strategies like complete, sequential and random search to generate subsets of features. Dunne et al. [5] stated that these searching strategies are based on stepwise addition or deletion of features. The goodness of the generated subset is evaluated using an evaluation criterion. If the newly generated subset is better than the previous subset, it replaces the previous subset with the best subset. These two processes are repeated until the stopping criterion is reached. The final best feature subset is then validated by prior knowledge or using different tests.

1.1.2 Feature Selection Algorithm

The feature selection algorithm removes the irrelevant and redundant features from the original dataset to improve the classification accuracy. The feature selections also reduce the dimensionality of the dataset; increase the learning accuracy, improving result comprehensibility. The feature selection avoid over fitting of data. The feature selection also known as attributes selection which is used for best partitioning the data into individual class. The feature selection method also includes the selection of subsets, evaluation of subset and evaluation of selected feature. The two search algorithms forward selection and backward eliminations are used to select and eliminate the appropriate feature. The feature selection is a three step process namely search, evaluate and stop. Different kinds of feature selection algorithms have been proposed. The feature selection techniques are categorized into three Filter method, Wrapper method, and Embedded method. Every feature selection algorithm uses any one of the three feature selection techniques. According to the class label present or not the feature selection can be further classified into two categories. Supervised and unsupervised feature selections. In the supervised method the relevance between the feature and the class is evaluated by calculating the correlation between the class and the feature. The relevance is evaluated by checking some property of the data in an unsupervised method. When we consider the microarray datasets it has thousands of features and also has high dimensional dataset. The feature selection algorithm plays an important role to maximizing the performance of such high dimensional datasets. To maximize the accuracy of high dimensional dataset we must follow the steps:

step1. Select the appropriate feature selection method

step2. Select the suitable classification algorithm

Originally we have two categories in datasets, Binary and multiclass datasets. The selected feature selection method and the classification algorithm must support to classify the

data into both binary and multi classes and maximize it accuracy.

2. Filter Method

The filter attribute selection method is independent of the classification algorithm. Filter method is further categorized into two types

1. Attribute evaluation algorithms
2. Subset evaluation algorithms

The algorithms are categorized based on whether they rate the relevance of individual features or feature subsets. Attribute evaluation algorithms rank the features individually and assign a weight to each feature according to each feature's degree of relevance to the target feature. The attribute evaluation methods are likely to yield subsets with redundant features since these methods do not measure the correlation between features. Subset evaluation methods, in contrast, select feature subsets and rank them based on certain evaluation criteria and hence are more efficient in removing redundant features [2]. The main disadvantage of the filter method is it ignores the dependencies among the features and treats the features individually.

2.1 Basic Feature Selection Algorithm

Input:

S - Data sample f with features X, |X| = n

J - Evaluation measure to be maximized

GS – successor generation operator

Output:

Solution – (weighted) feature subset

L: = Start Point(X);

Solution: = {best of L according to J };

Repeat L: = Search Strategy (L, GS (J), X);

X':= {best of L according to J};

If J (X') =J (Solution) or (J (X') =J (Solution) and |X'|

< |Solution|)

then Solution: =X';

Until Stop (J, L).

The filter method uses the discriminating criteria for feature selection. The correlation coefficient or statistical test like t-test or f-test is used to filter the features in the filter feature selection method.

3. Cluster Method

3.1 Hybrid approaches (Combination of Supervised and Unsupervised Learning)

In general, hybrid models are built by combining two or more data mining techniques in order to use the strength of different classifiers and increase the performance of the basic classifiers. For instance, supervised and unsupervised techniques can be serially integrated." That is, clustering can be used as a pre-processing stage to identify pattern classes for subsequent task of supervised prediction or classification(Jain 1999)". Hence, clustering can be used to identify homogenous populations in the data set. Then each cluster becomes a training set to train and finally a classification model can be created for the desired clusters. (Tsai 2009). The combination of supervised and unsupervised learning algorithms is a recent approach in the field of machine learning. Such a combination can either be used to improve the performance of a supervised classifier by biasing the classifier to use the information coming from the supervised classifier or it can be used to incorporate large amount of unlabeled data in the supervised learning process.

3.2 K-Means Clustering Algorithm

K-Means is one of the simplest unsupervised learning algorithms that solves the well-known problem of clustering. K-Means requires the number of clusters as an input.

K-Means steps for clustering are listed below:

1. Decide on a value for k

2. Initialize the K cluster centers

3. Form K clusters by assigning each point to the nearest cluster.

4. Recompute the center of each cluster

Repeat step 3 and step 4 until the centers no longer change

3.3 Two Step Clustering Algorithm

The algorithm can handle very large datasets. It is also designed to handle both continuous and categorical variables. It uses an agglomerative hierarchical clustering and involves two steps: including Preclustering and Clustering. Preclustering step: The first step makes a single pass through the data. Two Step applies a log-likelihood distance measure to calculate distance between cases. In preclustering step records are scanned one by one and based on the distance measure decides if the current record should be merged with the previously formed cluster or starts a new cluster. Clustering step: The second step uses a hierarchical clustering method to progressively merge the sub clusters into larger clusters, without requiring another

pass through the data. In this step, In order to increase the accuracy of clustering, the dataset was first classified with C&R Tree classification algorithm and then the predicted "Confidence Value" of the C&R Tree output node was added as an additional input node to the data set.

$$\text{Confidence} = \frac{N_{f,j}(t) + 1}{N_{f,j}(t) + k} \quad (1)$$

Where N f,j(t) is the sum of frequency weights for records in node t in category j, N f is the sum of frequency and K is the number of categories for target field. In our approach in order to determine the optimal number of clusters, first K-means clustering technique was applied for the initial clustering of the customers. Therefore, 9 different models ranging from 2 to 10 were generated and the standard deviation (SD) of each model was calculated. Lower SD means better clustering. Results from K-means clustering shows that SD lowers when the number of clusters increases.

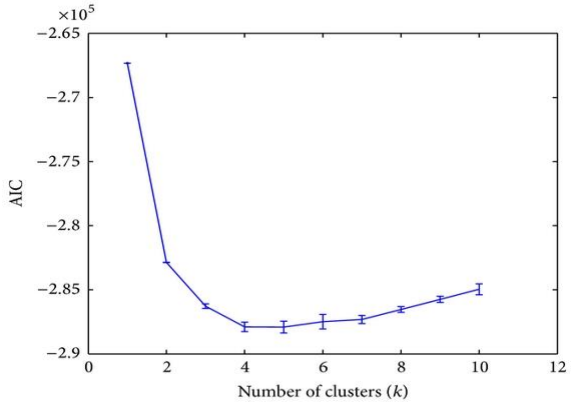

Chart -1: Standard deviation of different number of clusters

4. Intrusion Detection System

An intrusion detection system (IDS) is a device or software application that monitors a network or systems for malicious activity or policy violations. Any detected activity or violation is typically reported either to an administrator or collected centrally using a security information and event management (SIEM) system. A SIEM system combines outputs from multiple sources, and uses alarm_filtering techniques to distinguish malicious activity from false alarms.

4.1 Network intrusion detection systems

Network intrusion detection systems (NIDS) are placed at a strategic point or points within the network to monitor traffic to and from all devices on the network. It performs an analysis of passing traffic on the entire subnet, and matches the traffic that is passed on the subnets to the library of known attacks. Once an attack is identified, or abnormal behavior is sensed, the alert can be sent to the administrator. An example of an NIDS would be installing it on the subnet where firewalls are located in order to see if someone is trying to break into the firewall. Ideally one would scan all inbound and outbound traffic, however doing so might create a bottleneck that would impair the overall speed of the network. OPNET and NetSim are commonly used tools for simulation network intrusion detection systems. NID Systems are also capable of comparing signatures for similar packets to link and drop harmful detected packets which have a signature matching the records in the NIDS. When we classify the design of the NIDS according to the system interactivity property, there are two types: on-line and off-line NIDS, often referred to as inline and tap mode, respectively. On-line NIDS deals with the network in real time. It analyses the Ethernet packets and applies some rules, to decide if it is an attack or not. Off-line NIDS deals with stored data and passes it through some processes to decide if it is an attack or not.

4.2 Detection method

4.2.1 Signature-Based

Signature-based IDS refers to the detection of attacks by looking for specific patterns, such as byte sequences in network traffic, or known malicious instruction sequences used by malware.This terminology originates from anti-virus software, which refers to these detected patterns as signatures. Although signature-based IDS can easily detect known attacks, it is impossible to detect new attacks, for which no pattern is available.

4.2.2 Anomaly-Based

Anomaly-based intrusion detection systems were primarily introduced to detect unknown attacks, in part due to the rapid development of malware. The basic approach is to use machine learning to create a model of trustworthy activity, and then compare new behavior against this model. Although this approach enables the detection of previously unknown attacks, it may suffer from false positives: previously unknown legitimate activity may also be classified as malicious.

4.2.3 Intrusion Prevention

Intrusion detection and prevention systems (IDPS) are primarily focused on identifying possible incidents, logging information about them, and reporting attempts. In addition, organizations use IDPSes for other purposes, such as identifying problems with security policies, documenting existing threats and deterring individuals from violating security policies. IDPSes have become a necessary addition to the security infrastructure of nearly every organization.

IDPSes typically record information related to observed events, notify security administrators of important observed events and produce reports. Many IDPSes can also respond to a detected threat by attempting to prevent it from succeeding. They use several response techniques, which involve the IDPS stopping the attack itself, changing the security environment (e.g. reconfiguring a firewall) or changing the attack's content.

Intrusion prevention systems (IPS), also known as intrusion detection and prevention systems (IDPS), are network security appliances that monitor network or system activities for malicious activity. The main functions of intrusion prevention systems are to identify malicious activity, log information about this activity, report it and attempt to block or stop it.

Intrusion prevention systems are considered extensions of intrusion detection systems because they both monitor network traffic and /or system activities for malicious activity. The main differences are, unlike intrusion detection systems, intrusion prevention systems are placed in-line and are able to actively prevent or block intrusions that are detected.IPS can take such actions as sending an alarm, dropping detected malicious packets, resetting a connection or blocking traffic from the offending IP address. An IPS also can correct cyclic redundancy check (CRC) errors, defragment packet streams, mitigate TCP sequencing issues, and clean up unwanted transport and network layer options.

5. CONCLUSIONS

In this paper, we are replacing the classification algorithm. In existing system, we are using NIDS classifiers for classification. OPNET and NetSim are commonly used tools for simulation network intrusion detection systems. Accuracy is high and it is important to reduced input features in building IDS that is computationally efficient and effective. Finally, based on the experimental results achieved on dataset, it can be concluded that the proposed detection system has achieved promising performance in detecting intrusions over computer networks.

REFERENCES

[1] V.K.Pachghare,ParagKulkarni,Deven M. Nikam, “Intrusion Detection System Using Self Organizing Maps”, In Proceedings of IAMA 2009, IEEE, 2009.

[2] Ellen pitt, Richi nayak,”The use of various data mining and feature selection methods in the analysis of a population survey dataset”, Australlian computer socity inc 2007.

[3] L.Latha, T.deepa,”Feature selection methods and algorithms”, International journal on computer science and engineering, Vol. 3 No. 5 May 2011.

[4] And*erson, Ross* (2001). Security Engineering: A Guide to Building Dependable Distributed Systems. New York: John Wiley & Sons. pp. 387–388. ISBN 978-0-471-38922-4

[5] S. Mukkamala, A. H. Sung, A. Abraham, Intrusion detection using an ensemble of intelligent paradigms, Journal of network and computer applications 28 (2) (2005) 167–182.

[6] Lunt,Teresa F.,”Detecting Intruders in computer Systems”,1993 Conference on Auditing and Computer Technology,SRI International.

[7] M.S.Abadeh J.Habibi and C.Lucas,”Intrusion detection using a fuzzy genetics-based learning algorithm,”Journal of Network &Computer Applications.Vol.30,no.1,pp.414-428,2007

[8] Sandip Sonawane,Shailendra Pardeshi, and Ganesh Prasad”A Survey on intrusion detection techniques techniques”World Journal of Science and Technology 2012,2(3):127-133,ISSN:2231-2587.

[9] Pontarelli ,G.Bianchi,S.Toefili,”Trafficaware design of a high-speed fpga network intrusion detection System”,Computers, IEEE Transactions on 62(11)(2013) 2322 - 2334.