Network Intrusion Detection system based on Feature Selection and Triangle area Support Vector Machine

Venkata Suneetha Takkellapati[1] , G.V.S.N.R.V Prasad[2]

[1] M.Tech(CSE), Gudlavalleru Engineering College, Gudlavalleru

[2]Professor&Head of the Dept,gutta1.prasad@gmail.com, Gudlavalleru Engineering College, Gudlavalleru

Abstract **:As the cost of the data processing and Internet accessibility increases, more and more organizations are becoming vulnerable to a wide range of cyber threats. Most current offline intrusion detection systems are focused on unsupervised and supervised machine learning approaches. Existing model has high error rate during the attack classification using support vector machine learning algorithm. Besides, with the study of existing work, feature selection techniques are also essential to improve high efficiency and effectiveness. Performance of different types of attacks detection should also be improved and evaluated using the proposed approach. In this proposed system, Information Gain (IG) and Triangle Area based KNN are used for selecting more discriminative features by combining Greedy k-means clustering algorithm and SVM classifier to detect Network attacks. This system achieves high accuracy detection rate and less error rate of KDD CUP 1999 training data set.**

*Keywords***— Intrusion, IDS,data mining.**

I. INTRODUCTION

Intrusion detection techniques using data mining basically get into among the 2 categories; misuse recognition and anomaly recognition. In misuse detection, each example within a information set is labeled as 'normal or 'intrusion' over the labeled data. These techniques can automatically retrain intrusion detection designs on different input information which include new kinds of attacks, as long as they have been labeled appropriately[1,2,3].

Unlike signature-based IDS, models of abuse are made automatically, and can feel more advanced and precise than manually created predefined signatures. A key advantage of abuse recognition techniques is their tall level of precision in detecting known attacks and their variants. Their apparent drawback is the inability to detect attacks whose times have not however been observed. Anomaly detection, however, builds designs of normal behavior, and automatically detects any deviation from this, flagging the last as suspect.

Network Intrusion detection contains identifying a group of harmful actions that compromise the integrity, confidentiality, and availability of information. Monitored events are matched against the signatures to detect intrusions. Traditional methods for intrusion recognition are based on extensive knowledge of signatures of known attacks. Anomaly detection skills thus identify new kinds of intrusions as deviations from normal use. While a very powerful and novel appliance, a potential draw- back among these skills is the speed of fake alerts. This can happen

generally because previously unseen (yet le- gitimate) program behaviors can also feel distinguished as anomalies, and therefore flagged as possible intrusions. The signature database has got to feel manually re-vised for every new type of intrusion that is discovered.A significant limitation of signature-based methods would be that they cannot identify emerging cyber risks, since by their very nature these risks are launched using previously unknown attacks. These restrictions have led to an growing interest in intrusion detection skills based upon information excavation .

Anomaly is a pattern that does not conform to the anticipated behavior. Anomaly detection pertains to detecting patterns wearing a given information set that do not conform with an established normal behavior. The patterns thus recognized are called anomalies. These non-conforming designs are usually called anomalies, outliers, discordant observations, exceptions.An anomaly detection approach usually consists of two phases: a training phase and a test phase. In the training phase, the normal traffic profile is defined and in the test phase the learned profile is applied to the new data. Anomaly detection is used for identifying attacks in a computer networks, malicious activities in a computer systems, misuses in a Web-based systems. A network anomaly by malicious or unauthorized users can cause severe disruption to networks.

For a signature-based IDS to detect attacks, it must possess an attack description that can be matched to sensed attack manifestations. If an appropriate abstraction can be found, signature-based systems can identify previously unseen attacks that are abstractly equivalent to known patterns. These are inherently unable to detect novel attacks and suffer from false alarms when signatures match both intrusive and nonintrusive outputs. Signatures can be developed in a variety of ways, from hand translation of attack manifestations to automatic training or learning using labeled sensor data. Because a given signature is associated with a known attack abstraction, it is relatively easy for a signature-based detector to assign names (such as Smurf or Ping-of-Death) to attacks[IEEEIDS].

This paper presents the scope and status of our work both in misuse detection and anomaly detection. After the brief overview of building predictive models for learning from rare classes, the paper gives a comparative study of several anomaly detection schemes for identifying novel network intrusions.

II LITERATURE SURVEY

Rule based intrusion representation . This is the most common approach to representation of intrusion detection knowledge. In an If…Then formatted rule, the condition of the rule records the match criteria for the intrusion, and the action of the rule records the reaction for the intrusion.

Pattern oriented intrusion representation: Many intrusions may not be completed in a single step, and this is also true of intrusion detection. With only a single rule, only intrusions with a single step or intrusions with a significant feature, e.g., a BO intrusion can be represented. Therefore, for intrusions with several steps, a pattern oriented intrusion representation of intrusion behavior is needed. A pattern oriented intrusion representation can represent an intrusion, for example, a state machine or a state diagram in a sequence of states.

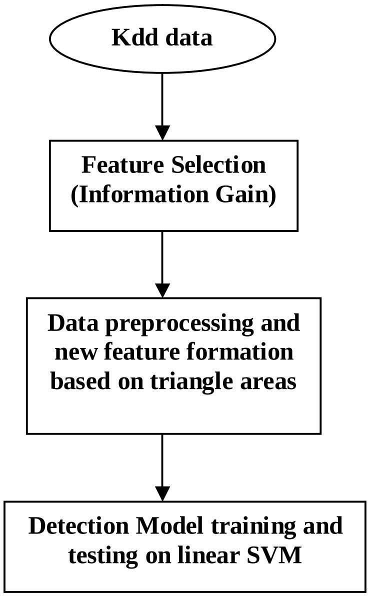
Specific intrusion representation: Many researches have tried to define a specific model together with a corresponding specific intrusion representation. For example, goal tree , which may achieve good performance for some specific target intrusions, is used to represent intrusion patterns. However, the specific representations sometimes lack extendibility since they may be not suitable for all kinds of intrusions[2].

Nowadays, the anomaly intrusion detection is the major research direction and has become a valuable technology to protect systems from outer malicious attacks [3]. KDD CUP 1999 dataset is the dominating evaluation dataset nowadays. In addition, there are several major evaluation metrics such as detection rate (DR), false alarm rate (FAR) are often utilized [4] [5].
Generally, most articles choose clustering as a method to reprocess datasets first for improving the following classifiers' performance. It might shorten the time of support vectors' building and analyze data in clustering. Then a sort of classifiers can be applied to classification. Large amount of researchers began to study ensemble of supervised and unsupervised machine learning algorithms in recent years. Khan et al. [6] used hierarchical clustering analysis with Dynamically Growing Self-Organizing Tree (DGSOT) algorithm to find out the boundary points which are the most significant points in training SVM, Comak et al. [7] proposed the KNN support vector machines (CKSVMs) based on a Gaussian function and Euclidean distance to reduce the long time for training, Li et al. [8] utilized Kmeans clustering to assign the data of each class to k clusters, and then used the new dataset consisting of only the centers of clusters to train SVM, in which k is the upper bound of the number of support vectors in each class.

EXISTING SYSTEM

DRAWBACKS:

- Existing system does not use attribute selection techniques in order to get effective attributes.
- Dynamically Growing Self-Organizing Tree (DGSOT) algorithm to find out the boundary points which are the most significant points in training SVM.
- Support vector machines (SVMs) based on a Gaussian function and Euclidean distance to reduce the long time for training.
- Kmeans clustering to assign the data of each class to k clusters, and then used the new dataset consisting of only the centers of clusters to train SVM, in which k is the upper bound of the number of support vectors in each class.
- All these approaches computational complexity and result in lower detection precision and more false positives.


Kdd data
Feature Selection (Information Gain)
Data preprocessing and new feature formation based on triangle areas
Detection Model training and testing on linear SVM

Existing system Architecture

III PROPOSED MODEL

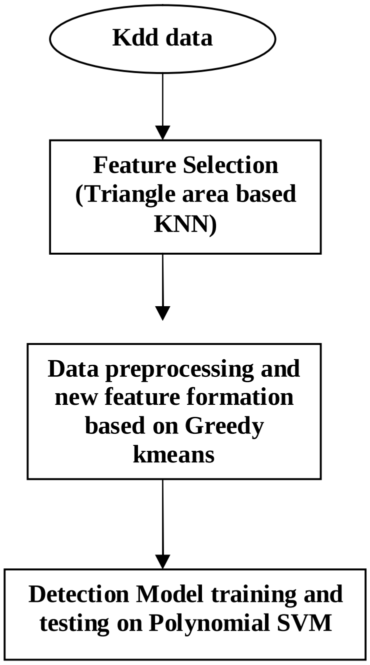
As the network environment has grown rapidly, so has the problem of intrusions. MIT kdd99 dataset is Currently available approaches to dealing with intrusions can be categorized as follows: Reconnaissance/Snooping/Information Gathering (Probing): This kind of intrusion tries to gather useful information, including public or private data, using the powerful computing capability of a computer.
Gaining Access (User to Root): Intruders or hackers try to get access rights from a victim host, e.g., the access right of the root account.
Remote Control (Remote to Local): The intruder uses a back door program or takes advantage of application vulnerability to control remote victims through a network.
Denial of Service (DoS): The basic goal of DoS intrusion is to overwhelm the victim host with a huge number of requests. DoS intrusion is easy to achieve, and it can cause the host to crash.

In addition to the intrusions mentioned above, other intrusions may use physical or social strategies to intrude into a system by taking advantage of the vulnerability of the system or application.

Advantanges:

- In our detection model, we select the most relevant feature for each attack type from original features by ranking each feature's information gain of this attack and. Next, we create new low-dimensional feature vector for each data by mapping the most relevant but still high-dimensional feature space into a triangle area based feature space.
- This system uses efficient Information Gain(IG) for attribute selection.
- This system uses Triangle area based KNN feature selection. Polynomial based SVM to avoid problems caused by its high-dimensional feature space in classifying result.

Architecture of proposed model:


Kdd data
Feature Selection (Triangle area based KNN)
Data preprocessing and new feature formation based on Greedy kmeans
Detection Model training and testing on Polynomial SVM

Proposed System Architecture

*Main algorithms in this proposed approachs are***:**

A) Feature Selection Algorithms
    1) Triangle area based KNN
B) Clustering Algorithm
    1) Greedy Kmeans
C) Classification Algorithm
    2)Polynomial Support Vector Machine

Feature Selection:

Feature selection, knowledge discovery in databases (KDD), data mining and database mining are terms used to express the growing field of interest about selecting proper information from databases. This field of interest gains daily more regards due to the increasing availability of collecting and accessing large amount of information. The advances in the technology of computing, storage, and communications enlarge continuously the available amount of information. Large information systems exist in many institutions and other information is available via networks or aerospace. Many examples are arising form the banking, diagnosis, finance, health care, manufacturing, marketing, retail sales' fields and establishments. However, the proper technology for analyzing, understanding, visualization of the existing data does not yet exist [1].

Information Gain:

Attribute selection decide which attribute is the best using a statistical property, called information gain, is used. Gain measures how well a given attribute separates training examples into targeted classes. The one with the highest information (information being the most useful for classification) is selected. Given a collection S of c outcomes The expected information needed to classify a tuple in D is given by

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i),$$

where pi is the probability that an arbitrary tuple in D belongs to class Ci. A log function to the base 2 is used, because the information is encoded in bits. Info(D) is just the average amount of information needed to identify the class label of a tuple in D. Info(D) is also known as the entropy of D.

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j).$$

The term Dj /D acts as the weight of the jth partition. InfoA(D) is the expected information required to classify a tuple from D based on the partitioning by A. The smaller the expected information (still) required, the greater the purity of the partitions. Information gain is defined as the difference between the original information requirement (i.e., based on just the proportion of classes) and the newrequirement .That is,

$$Gain(A) = Info(D) - Info_A(D).$$

Inforamtion Gain Attribute Selection Results for KDD99 cup dataset:

Selected attributes:
5,6,3,4,8,18,17,14,19,16,15,2,11,13,12,7,10,9,1 : 19

Greedy KMeans:

The fast greedy k-means algorithm uses the concepts of the greedy global k-means algorithm for a global solution. The intermediate convergence is done using the restricted k-means algorithm instead of the naïve k-means algorithm.

- Construct an appropriate set of positions/locations which can act as good candidates for insertion of new clusters;
- Initialize the first cluster as the mean of all the points in the dataset;
- In the Kth iteration, assuming K-1 clusters after convergence find an appropriate position for insertion of a new cluster from the set of points created in step 1 that gives minimum distortion;
- Run k-means with K clusters till convergence. Go back to step 3 if the required number of clusters is not yet reached.

TRIANGLE AREA BASED KNN

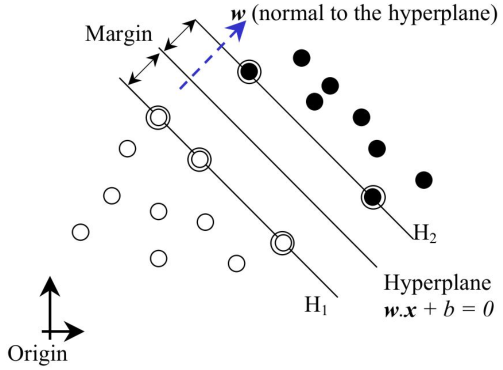
Let us start with KNN algorithm for classification. K-nearest neighbor is a supervised learning algorithm where the result of new instance is classified based on majority of K-nearest neighbor category. The purpose of this algorithm is to classify a new object based on attributes and training samples. The classifiers do not use any model to fit and only based on memory. Given a point object, we find K number of objects closest to the query point. The classification is using majority vote among the classification of the K objects.

Using KNN approach choose k data points randomly. The third stage is to calculate the triangle area. The Euclidean distance formula is employed to calculate the length of three edges in a triangle. Therefore, we can get the triangle perimeter and compute area by Heron's formula as well. We assume a, b, c are the length of three edges of a triangle. P is the semiperimeter of the triangle. S is the triangle area. The Heron's formula is[BASE PAPER]:

$$\sqrt{s(s-a)(s-b)(s-c)}$$

LINEAR AND POLYNOMIAL SVM

A support vector machine is primarily a two-class classifier. It is possible to solve multi class problems with
support vectors by treating each single class as a separate problem. It aims to maximise the width of the margin between classes, that is, the empty area between the decision boundary and the nearest training patterns.
Given a set of points $\{x_i\}$ in n-dimensional space with corresponding classes $\{y_i : y_i \varepsilon \{-1,1\}\}$ then the training algorithm attempts to place an hyperplane between points where $y_i = 1$ and points where $y_i = -1$. Once this has been achieved a new pattern x can then be classified by testing which side of the hyper-plane the point lies on.



IV. EXPERIMENTAL RESULTS

All experiments were performed with the configurations Intel(R) Core(TM)2 CPU 2.13GHz, 2 GB RAM, and the operation system platform is Microsoft Windows XP Professional (SP2). Kddcup 99 dataset is used for network intrusion detection.

Experimental results:

TRIANGLE KNN BASED FEATURE SELECTION

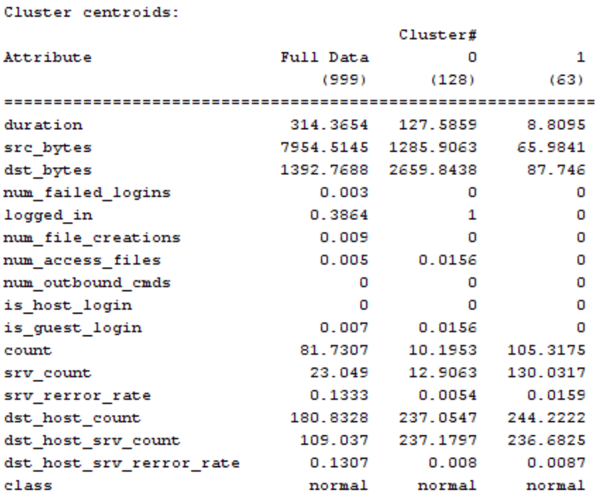
```
Attribute Evaluator (supervised, Class (nominal): 20 class):
         KNN TRIANGLE Filter

ROBUST selected attributes:
 0.864852   5 src_bytes
 0.695479   6 dst_bytes
 0.678822   3 service
 0.648504  18 dst_host_srv_count
 0.625962  14 count
 0.549952   4 flag
 0.402497   8 logged_in
 0.363826  17 dst_host_count
```

Executing greedy kmeans algorithm for KDD dataset

```
GreedykMeans
======

Number of iterations: 9
Within cluster sum of squared errors: 184.52699037162498
Missing values globally replaced with mean/mode
```



```
Cluster centroids:
                                         Cluster#
Attribute                    Full Data          0           1
                                 (999)      (128)        (63)
==============================================================
duration                      314.3654   127.5859      8.8095
src_bytes                    7954.5145  1285.9063     65.9841
dst_bytes                    1392.7688  2659.8438      87.746
num_failed_logins                0.003          0           0
logged_in                       0.3864          1           0
num_file_creations               0.009          0           0
num_access_files                 0.005     0.0156           0
num_outbound_cmds                    0          0           0
is_host_login                        0          0           0
is_guest_login                   0.007     0.0156           0
count                          81.7307    10.1953    105.3175
srv_count                       23.049    12.9063    130.0317
srv_rerror_rate                 0.1333     0.0054      0.0159
dst_host_count                180.8328   237.0547    244.2222
dst_host_srv_count             109.037   237.1797    236.6825
dst_host_srv_rerror_rate        0.1307      0.008      0.0087
class                           normal     normal      normal
```

```
=== Clustering stats for training data ===

Clustered Instances
0      128 ( 13%)
1       63 (  6%)
2      297 ( 30%)
3       16 (  2%)
4       56 (  6%)
5      120 ( 12%)
6       67 (  7%)
7       85 (  9%)
8      167 ( 17%)
```

Comparision of Linear SVM and Polynomial SVM

Detection model	Evalution Index		
	Accuracy	False rate	Time taken on trained data(sec)
Exsisting model	83.68	16.31	0.9
Proposed model	86.38	13.61	0.06

V CONCLUSION

Feature selection is an important task of Network Intrusion application.Now a days, large amount of attacks are threatening network and information security. Previous single intrusion detection methods are substituted by ensemble of various machine learning algorithms detection models. This project presented a hybrid machine learning intrusion detection model using triangle area knn, poly kernel svm and greedy kmeans algorithm. Using Feature selection approach kdd attacks are detected with less error rate and high accuracy.

In future this work can be extended to implement svm with high optimized kernel functions.This work need to implement in real time web analytics for intrusion detection.

REFERENCES

[1] Protecting Against Cyber Threats in Networked Information Systems L. Ertoz.

[2] W. Lee, S. J. Stolfo, Data Mining Approaches for Intrusion Detection, Proceedings of the 1998 USENIX Security Symposium, 1998.

[3]. E. Bloedorn, et al., Data Mining for Network Intrusion Detection: How to Get Started, MITRE Technical Report, August 2001.

[4] D. Barbara, N. Wu, S. Jajodia, Detecting Novel Network Intrusions Using Bayes Estimators, Proceedings of the First SIAM Conference on Data Mining, Chicago, IL, 2001.

[5]. S. Manganaris, M. Christensen, D. Serkle, and K. Hermix, A Data Mining Analysis of RTID Alarms, Proceedings of the 2nd International Workshop on Recent Advances in Intrusion Detection (RAID 99), West Lafayette, IN, September 1999.

[6]. Cohen, W. W., ”Fast effective rule induction”, In A. Prieditis and S. Russell (Eds.), Proc. of the 12th International Conference on Machine Learning, Tahoe City, CA, pp. 115123. Morgan Kaufmann, 9-12 July, 1995.

[7]. S. Stolfo, A. L. Prodromidis and P. K. Chan, ”JAM: Java Agents for Meta- Learning over Distributed Databases”, in Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, D. Heckerman, H. Mannila, D. Pregibon, and R. Uthurusamy, editors, AAAI Press, Menlo Park, 1997.

[8]. Lee, W., S. J. Stolfo, and K. W. Mok, ” Mining in a data-flow environment: Experience in network intrusion detection,” In S. Chaudhuri and D. Madigan (Eds.), Proc. of the Fifth International Conference on Knowledge Discovery and Data Mining (KDD-99), San Diego, CA, pp. 114124. ACM, 12-15 August 1999.

[9]. Lee, W., S. J. Stolfo, and K. W. Mok, ”Adaptive intrusion detection: A data mining approach,” Artificial Intelligence Review 14 (6), 533567, 2000.