Attribute Reduction based Hybrid Anomaly Intrusion Detection using K-Means and SVM Classifier

Ujwala Ravale
Lecturer Professor
SIESGST, Nerul

Nilesh Marathe
Assistant Professor
R. A. I. T. Nerul

Puja Padiya
Assistant Professor
R. A. I. T. Nerul

ABSTRACT
In Information Security, intrusion detection is the act of detecting actions that attempt to compromise the confidentiality, integrity or availability of a resource. One of the primary challenges to intrusion detection is the problem of misjudgment, misdetection and lack of real time response to the attack. Various data mining techniques such as clustering, classification and association rule discovery are being used for intrusion detection. The proposed hybrid technique combines data mining approaches like K Means clustering algorithm and Support Vector Machine classification module. The main purpose of proposed technique is to decrease the number of attributes associated with each data point. So that, the proposed technique can perform better in terms of Detection Rate and Accuracy when applied to KDD'99 Data Set.

Keywords
Intrusion detection system, K Means clustering, Support Vector Machine, KDD'99 Data Set

1. INTRODUCTION
With the advent of Internet, information retrieval has become more accessible but it has exposed our information system to various intrusions which may compromise our information security. Even in the network environment, intrusions pose a security risk as they are exposed to both host based intrusions as well as network-based intrusions such as Bandwidth Theft and Denial of service. In order to combat these attacks an Intrusion Detection System (IDS) is needed, such that it can help in detecting harmful intrusions.

IDS are defined as a process or device that analyzes system and network activity for unauthorized access and/or malicious activity.

1.1 Classification of IDS

1.1.1 Misuse Based System
In misuse based IDS [18], detection is performed by looking for the exploitation of known weak points in the system, which can be described by a specific pattern or sequence of events or data. That means these systems can detect only known attacks for which they have a defined signature.

1.1.2 Anomaly Based System
In anomaly based IDS [18], detection is performed by detecting changes in the patterns of utilization or behavior of the system. The main advantage of anomaly detection system is that they can detect previously unknown attacks.

2. RELATED WORK
Data mining is the latest technology introduced in network security environment to find regularities and irregulaties in large datasets. KDD CUP '99 dataset is the dominating evaluation dataset used by most of researcher to test their proposed techniques. The best possible accuracy and detection rate can be achieved by using Hybrid learning approaches. However, the work to improve false alarm rate is an ongoing affair. Different classifiers can be use to formed a hybrid learning approaches such as combination of clustering and classification technique.

In 2009, Meng Jianliang, Shang Haikun, Bian Ling had presented the K-means algorithm [3] for intrusion detection. Experimental results on a subset of KDD-99 dataset showed the stability of efficiency and accuracy of the algorithm. With different setting, the detection rate stayed always above 96% while the false alarm rate was below 2%.The time complexity is low.

In 2009, Jingwen Tian, Meijuan Gao have proposed this technique which contains the ability of strong function approach and fast convergence of radial basic function neural network[12], the network intrusion detection method based on radial basic function neural network can detect various intrusion behaviors rapidly and effectively by learning the typical intrusion characteristic information.

In 2011, Preecha Somwang and Woraphon Lilakiatsakun have proposed the new intrusion detection technique by using hybrid methods of unsupervised/supervised learning scheme. The technique integrates the Principal Component Analysis (PCA) with the Support Vector Machine (SVM) [8]. The results show that the proposed technique can improve the performance of anomaly intrusion detection, the intrusion detection rate and generate fewer false alarms.

In 2012, XIE Yang, ZHANG Yilai have proposed Anomaly Intrusion Detection based on SVM. The application of support vector machine (SVM) for network intrusion detection was researched, Although SVM was an effective abnormal analysis for intrusion detection with a small sample, and there were two deficiencies in traditional SVM: slow in training, low detection rate. An intelligent anomaly analysis algorithm for intrusion detection based on SVM is presented [3]. This algorithm can intelligently select learning vector samples during the training state, and effectively reduce the number of training samples and training time, and also can obtain a higher detection rate classifier in the case of small samples.

In 2012, Susheel Kumar Tiwari and Mahendra Singh Sisodiai have proposed a model of NIDS based on K-Means Clustering via Naive Bayes algorithm. The model builds the patterns of the network services over data sets labeled by the services. With the built patterns, the model detects attacks in the datasets using the k-means clustering via naive Bayes Classifier algorithm. Compared to the Naive based approach, this approach achieve higher detection rate. However, it generates somewhat more false positive rate.

In 2012, Roshan Chitrakar and Huang Chuanhe proposed a hybrid approach [7] to anomaly based intrusion detection by using k- Medoids clustering with Naïve Bayes classification and produced better performance compared to k-Means with Naïve Bayes classification. The approach, using KDD datasets, showed around 2% of improvement in both Accuracy and Detection Rate while reducing False Alarm Rate by 1%.

3. PROPOSED METHODOLOGY

Various hybrid techniques are used for intrusion detection. Each technique has its own advantages and disadvantages. As well as performance of each technique is varies in terms of Accuracy, Detection rate & False Positive Rate. The proposed technique combines unsupervised learning with supervised learning. The proposed technique combines unsupervised learning with supervised learning. The proposed intrusion detection technique initially clusters training data set into 'K' clusters where k is the no. of clusters. In the next step support vector is generated. As a last step classification is performed using SVM to detect intrusion is happened or not.

Clustering [2] is the method of grouping objects into meaningful subclasses so that the members from the same cluster are quite similar, and the members from different clusters are quite different from each other. Therefore clustering methods can be useful for classifying log data and detecting intrusions.

When clustering is done on a data set having each data point of N-dimensional. To cluster the data into some K clusters same attributes of each data point are considered. This scenario is shown in Fig. 1. It shows the general scenario where all N attributes of a data point are considered to cluster it into cluster-1, cluster-2, cluster-3, ... , cluster-K. That is all the attributes are used to find the distance between data point and a cluster.

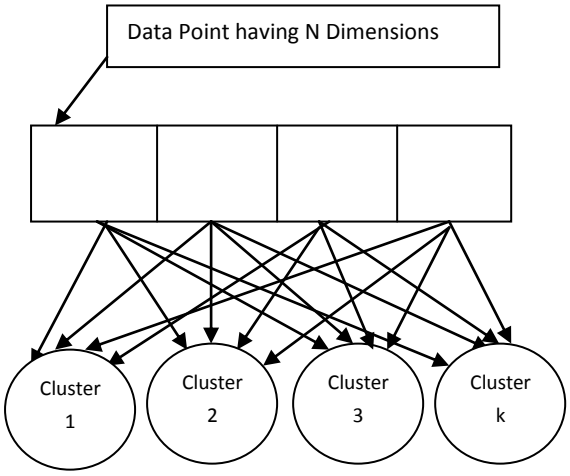

Figure 1: General Clustering Scenario

But in our proposed system we need only selected attributes say P attributes to cluster data into one cluster and some q attribute to cluster into another cluster. These attributes are selected based on the different attack classes present in the dataset and their characteristics.

The similar concept has been used here to detect different types of attacks. There is no need to consider all 21 attributes present in KDD99 data set [15] to cluster them into different types of attack. To cluster the data only some attributes out of total 21 attributes are needed. Table 2 shows that which attributes we can consider to cluster data point into particular cluster.

Table 1. Selected Attribute Set

Different Clusters	Selected Attributes
NORMAL	{1,3,5-10,12, 15,17,20-23,25-29,33,35,36,38-20}
DOS	{1-8,23,25,27,30,32-35,38-21 }
PROBE	{1,2,3,5,6,19,20,23,25,27,30,32}
U2R	{3,5,6,8, 12-19,21,22,22}
R2L	{1,2,3,2,5,6,10-12,12,17-19,21,22}

3.1 Proposed Architecture

The block diagram of proposed hybrid technique is given below.

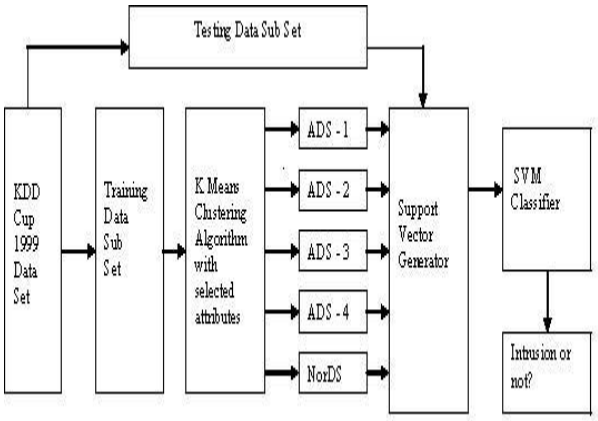

Figure 2 : Proposed System Block Diagram

3.2 K Means Clustering Algorithm

The objective function is [1]:

$$J = \sum_{i=1}^{k} \sum_{j=1}^{n} dij(Xj, Ci)$$

Where dij(Xj,Ci) is a chosen distance measure (Euclidean distance) between a data point xj and the cluster center ci, is an indicator of the distance of the data points from their respective cluster centers.

The algorithm is composed of the following steps [1]

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into group's i. e. clusters.

3.3 Support Vector Generation Module

SVM has recently been introduced as a new technique for solving a variety of learning, classification and prediction problems. The basic SVM deals with two-class problems, known as Binary classification problems in which the data are separated by a hyperplane defined by a number of support

vectors [11]. Support vectors are a subset of training data used to define the boundary between the two classes. In situations where SVM cannot separate two classes, it solves this problem by mapping input data into high-dimensional feature spaces using a kernel function.

In the two-dimensional case, the SVM action can be illustrated using Fig. 3. In Fig. 3, a series of points for two different classes of data are shown, circles (class A) and squares (class B). The SVM attempts to place a linear boundary (solid line) between the two different classes and orients this line in such a way that the margin (space between dotted lines) is maximized. The nearest data points used to define the margin are known as support vectors (gray circles and square). Support vectors, not the number of input features, contain all of the information needed to define the classifier. One remarkable property of SVM is its ability to learn can be independent of the feature space dimensionality. This means that SVM can generalize well in the presence of many features. Fig. 3 presents the simplest model of SVM called the maximal margin classifier. It works only for data that are linearly separable in the feature space.

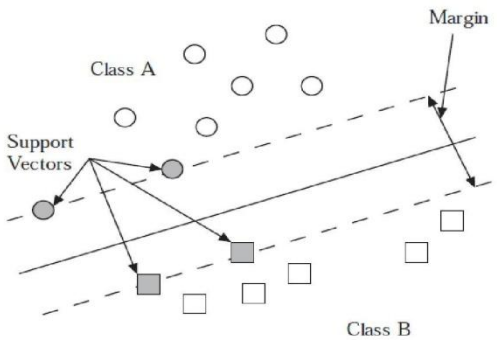

Figure 3: Seperation of two classes by SVM [11]

Mathematically, the linear boundary can be expressed in terms of

$w^T x \square b \square 0$

In a classification problem, we try to estimate a function $f : R^n \rightarrow \{\pm 1\}$ using training data[16]. Let us denote the class A with $x \varepsilon A$, $y = 1$ and class B with $x \varepsilon B$, $y = -1$ and

$(x_i, y_i) \varepsilon R^n \rightarrow \{\pm 1\}$. If the training data are linearly separable then there exists a pair

$(w, b) \varepsilon \Re^n \times R$ such that

$w^T x \square b \geq \square 1$ for all $x \varepsilon A$

$w^T x \square b \leq -1$ for all $x \varepsilon B$

with the help of above equations decision function is given by

$f_{w,b}(x) = \text{sign}(w^T x \square b)$

w is termed the weight vector and b the bias. The above inequality constraints can be combined to give y

$(w^T x + b) \geq 1$ for all $x \varepsilon A \cup B$.

In situations where SVM cannot separate two classes, it solves this problem by mapping input data into high-dimensional feature spaces using a kernel function. Various kernel functions can be used, such as linear, polynomial or Gaussian [14].

3.3.1 Polynomial kernel

This Polynomial kernel is a non-stationary kernel. Polynomial kernels are well suited for problems where all the training data is normalized. Adjustable parameters are the constant term c and the polynomial degree d.

$$K(X_p, X_j) = (X_p, X_j)^d + c$$

3.3.2 Gaussian Kernel

The Gaussian kernel is an example of radial basis function kernel.

$$K(X_i, X_j) = \exp\{-(|X_i - X_j|) / 2\sigma$$

Where σ stands for window width.

3.3.3 Sigmoid kernel

Sigmoid Kernel is also called as Hyperbolic Tangent Kernel and the Multilayer Perception (MLP) kernel. The Sigmoid Kernel comes from the Neural Networks field, where the bipolar sigmoid function is often used as an activation function for artificial neurons.

$K(X_i, X_j) = \tanh(K(X_i, X_j) + r$

3.4 Feature Selection Using Information gain (IG) Method

In KDD dataset plenty of features are redundant or have little significance during the detection process. These features will affect computational efficiency and final classification effects. Therefore, selecting useful features by features relevance analysis [9] is a critical step for the whole system. We use IG [12] to decide the most relevant feature in discrimination. For a specified attack type, the feature with the highest IG is regarded as the most relevant feature which has a key role in determining the attack type.

Consider S is a set of training data set samples with labels. We assume the data set has m classes (attack classes and normal class) and Si is the number of samples in class 1. We set S as the total sample number in the training set. Now, expected information can be calculated to classify a given sample by

$I(S_1, S_2 \ldots\ldots S_m) = -\sum_{i=1}^{m} S_i / S \log \quad (Si/S)$

We assume a feature F possesses values {fl , f2, .. .,fv}, so the training set can be divided into v parts {SI,S2, ... ,Sv}according to these feature values.Sj is the part which containing features value. Besides, Sj includes Sij samples belonging to class i. We can obtain feature F's entropy value E (F) by:

$$\sum_{j=1}^{v} [\![S1j + \cdots + Smj]\!] — S \times I(S1j + \cdots + Smj)$$

Therefore, the IG for F is

Gain (F) = I (S1,S2,……Sm) – E (F)

Feature Information Gain for each class can be calculated according to the three above formulas. So the most relevant feature for each class label is available now.

Now, the **most relevant attribute for SVM** can be Protocol_type, service, Src_bytes, dst_bytes, logged_in, count, srv_diff_host_rate, dst_host_srv_count, dst_host_same_srv_rate, dst_host_same_src_port_rate, dst_host_srv_diff_host_rate, dst_host_srv_rerror_rate.

4. ANALAYSIS

To evaluate the effectiveness of proposed approach on KDD99 Dataset, it can describe the results using Detection Rate (DR), False Positive Rate (FPR) and Accuracy (ACC). Each metric is defined below:

Detection Rate (DR): Detection rate [2] is the rate of correctly classified intrusive examples to the total no. of intrusive examples.

Detection Rate (DR) = (TP) / (TP + FP)

False Positive Rate (FPR): The false positive rate is the number of normal connections that are misclassified as attacks divided by the number of normal connections in the data set.

False Positive Rate (FPR) = (FP) / (FP+TN)

Accuracy (ACC)
Accuracy [2] is the ratio of correctly classified to the total classified examples.

Accuracy (ACC) = (TP+TN) / (TP+TN+FP+FN)

Where,

FN is False Negative,

TN is True Negative,

TP is True Positive and

FP is False Positive.

Confusion Matrix
A Confusion matrix[3] is a visualization tool typically used in supervised learning (in unsupervised learning it is usually called matching matrix). Each row in the matrix represents instances of predicted class, while each column represents instances of actual class. The Advantage of using this matrix is that it not only tells us how many got misclassified but also what misclassifications occurred.

5. CONCLUSION

Intrusion detection is an important component in network security. Feature selection is the major challenging issues in IDS in order to reduce the useless and redundant features among the attributes. In the proposed hybrid technique we are combining K Means clustering algorithm with classification algorithm i. e. Support Vector Machine. KDD CUP 99 dataset can be used for experimental purpose.

In clustering mainly attribute selection is based on the attack classes present in the dataset and their features. Only selected attributes can be used for cluster formation. As well as in SVM feature selection can be done using Information Gain (IG). By combining these concepts we can get better performance and computation time. Better performance can be gained by increasing accuracy rate and decreasing false positive rate.

6. REFERENCES

[1] Z. Muda, W. Yassin, M.N. Sulaiman, N.I. Udzir, "Intrusion Detection based on K-Means Clustering and Naïve Bayes Classification", 2011 7th International Conference on IT in Asia (CITA).

[2] Meng Jianliang Shang Haikun Bian Ling, "The Application on Intrusion Detection Based on K-means Cluster Algorithm", 2009 IEEE International Forum on Information Technology and Application.

[3] Sanjay Kumar Sharma, Pankaj Pande, Susheel Kumar Tiwari and Mahendra Singh Sisodia, "An Improved Network Intrusion Detection Technique based on k-Means Clustering via NaIve Bayes Classification", IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM - 2012) March 30, 31, 2012

[4] Deepthy K Denatious & Anita John, "Survey on Data Mining Techniques to Enhance Intrusion Detection", 2012 International Conference on Computer Communication and Informatics (ICCCI -2012), Jan. 10 – 12, 2012, Coimbatore, INDIA.

[5] Roshan Chitrakar and Huang Chuanhe, "Anomaly Detection using Support Vector Machine Classification with k-Medoids Clustering", In IEEE computer society symposium on research in security and privacy,2012.

[6] Preecha Somwang and Woraphon Lilakiatsakun, "Computer Network Security Based On Support Vector Machine Approach",11th International Conference on Control, Automation and Systems Oct. 26-29, 2011 in KINTEX, Gyeonggi-do, Korea.

[7] E.Raju, K.Sravanthi, "Network intrusion detection using Support Vector Machines", International Journal of Computer Science and Management Research Vol 2 Issue 1 January 2013 ISSN 2278-733X.

[8] Shyam Sunder, Balaram, P.Pavan kumar, "SVM & Decision Trees for High Attack Detection Ratio", IJCAE, Vol.3 Issue 3, November 2012,37 – 25.

[9] Jingwen Tian, Meijuan Gao, Fan Zhang, "Network Intrusion Detection Method Based on Radial Basic Function Neural Network", Computer Engineering and Design, vol. 29,no. 12, pp. 3022-3025, 2009 IEEE.

[10] Yogita B. Bhavsar1, Kalyani C.Waghmare, "Intrusion Detection System Using Data Mining Technique: Support Vector Machine", International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 3, March 2013.

[11] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set", Proceedings of the 2009 IEEE symposium on Computational Intelligence in Security and Defense Applications (CISDA 2009).

[12] Kazem Qazanfari, Minoo Sadat Mirpouryan, Hossein Gharaee, "A Novel Hybrid Anomaly Based Intrusion Detection Method", 6th International Symposium on Telecommunications (IST'2012).

[13] Hari Om, Aritra Kundu, "A Hybrid System for Reducing the False Alarm Rate of Anomaly Intrusion Detection System", 1st Int'l Conf. on Recent Advances in Information Technology (RAIT) 2012 IEEE.

[14] Jingwen Tian, Meijuan Gao, Fan Zhang, "Network Intrusion Detection Method Based on Radial Basic Function Neural Network", Computer Engineering and Design, vol. 29,no. 12, pp. 3022-3025, 2009 IEEE.