International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Fuzzy K-Means Based Intrusion Detection System Using Support Vector Machine

Aman Mudgal[1], Rajiv Munjal[2]

[1]M. Tech Scholar, CBS Group of Institutions, Haryana, India

[2]Assistant Professor, Department of CSE, CBS Group of Institutions, Haryana, India

Abstract: ***Intrusion Detection System (IDS) is an important tool to identify various attacks to secure the networks. The goal of an Intrusion Detection System (IDS) is to provide a layer of defense against malicious users of computer systems by sensing a misuse and alerting operators to on-going attacks. Most real-world data, especially data available on the web, possess rich structural relationships. Most of the clustering algorithms neglect the structural relationships between the individual data types. We proposed Fuzzy K-Means clustering, which integrates two sources of information into a single clustering framework. Our main objective is to complete analysis of intrusion detection Dataset. In this paper we combine two of the efficient data mining algorithms and make a hybrid technique for the detection of intrusion called fuzzy k-means and Support vector machine.***

Keywords: Intrusion Detection, Fuzzy K-Mean, SVM.

1. Introduction

The intrusion or attack in the computer network is one of the most important issues creating problems for the network managers. However many countermeasures are taken for the security of the network but continuous growth of hackers requires to maintain the defending system up to data. This paper presents a K-means and support vector machine based intrusion detection system. The support vector machine is optimal partitioning based linear classifier and at least theoretically better other classifier also because only small numbers of classes required during classification SVM with one against one technique can be the best option and the K-means clustering filters the un-useful similar data points hence reduces the training time also hence provides an overall enhanced performance by reducing the training time while maintaining the accuracy. The proposed algorithm is tested using KDD99 dataset and results show the effectiveness of the algorithm. The paper also analysed the effect of different input parameters on classification accuracy.

2. Intrusion Detection System

Instrusion detection System monitors the violation of management and security policy and malicious activities in the computerized network [1]. The intrusion can be caused by inside (legal users), or outside (illegal users) in the system [2]. Nowadays recognition and prevention of intrusion is one of the most important mechanisms that provides security in networks and computer systems, and generally is used as a complemented security for firewalls [3]. IDS systems created as a software and hardware system that each one has its specific properties [1]. Hardware systems have been preferred to software system because of their speed and accuracy. But software systems are more common because of high compatibility with several operating systems [4]. James P. Anderson is known as a first person who propounded the investigation about registered events in the system in the field of security. Anderson demonstrated a report in 1980 which was the first activity about the recognition of intrusion [5, 6]. IDS generally have three main functions: monitoring and evaluation, detection and response [7].

Intrusion detection techniques are usually classified into misuse detection and anomaly detection. Anomaly detection focuses on detecting unusual activity patterns in the observed data. Misuse detection methods are intended to recognize known attack patterns.

2.1 Intrusion Detection Techniques

Intrusion detection techniques are divided into two groups and there are several algorithms which are described for supervised and unsupervised learning

2.2 Supervised Learning Algorithms

1. k-Nearest Neighbour - The k-Nearest neighbour is a classical algorithm [8] that finds k examples in training data that are c1losest to the test example and assigns the most frequent label among these examples to the new example. The only free parameter is the size k of the neighbourhood.
2. Multi-Layer Perceptron - Training of a multi-layer perceptron involves optimizing the weights for the activation function of neurons organized in a network architecture. The global objective function is minimized using the RPROP algorithm [9]. The free parameter is the number of hidden neurons.
3. Regularized discriminant analysis - Assuming both classes of examples are normally distributed, a Bayes-optimal separating surface is a hyperplane (LDA), if covariance matrices are the same, or a quadratic surface otherwise (QDA). A gradual morph between the two cases can be implemented by using a regularization parameter [10]. Another free parameter , controls the addition of identity matrix to covariance matrices.
4. Fisher Linear Discriminate - Fisher Linear Discriminate constructs a separating hyper plane using a direction that maximizes inter-class variance and minimized the

intra-class variance for the projection of the training points on this direction [8]. The free parameter is the trade-off between the norm of the direction and the "strictness" of projection.

5. Linear Programming Machine and Support Vector Machine - Linear Programming Machine (LPM) and Support Vector Machine (SVM) construct a hyper plane of the minimal norm which separates the two classes of training examples [11]. LPM uses the 1-norm, SVM uses the 2-norm. Furthermore, SVM apply a non-linear mapping to construct a hyper plane in a feature space. In our experiments, radial basis functions are used, their complexity controlled by the width parameter w. Another parameter C controls the trade-off between the norm of a hyper plane and the separation accuracy.

2.3 Unsupervised Learning Algorithms

1. k-Means Clustering - k-Means clustering is a classical clustering algorithm [8]. After an initial random assignment of example to k clusters, the centres of clusters are computed and the examples are assigned to the clusters with the closest centres. The process is repeated until the cluster centres do not significantly change. Once the cluster assignment is fixed, the mean distance of an example to cluster centres is used as the score. The free parameter is k.
2. Single Linkage Clustering - Single linkage clustering [12] is similar to k-Means clustering except that the number of clusters is controlled by the distance parameter W: if the distance from an example to the nearest cluster center exceeds W a new cluster is set.
3. Quarter-sphere Support Vector Machine - The quarter-sphere SVM [13] is an anomaly detection method based on the idea of fitting a sphere onto the center of mass of data. An anomaly score is defined by the distance of a data point from the center of the sphere. Choosing a threshold for the attack scores determines the radius of the sphere enclosing normal data points.

3. Related Work

Evaluation of Fuzzy K-Means And K-Means Clustering Algorithms In Intrusion Detection Systems [14]
According to the growth of the Internet technology, there is a need to develop strategies in order to maintain security of system. One of the most effective techniques is Intrusion Detection System (IDS). This system is created to make a complete security in a computerized system, in order to pass the Intrusion system through the firewall, antivirus and other security devices detect and deal with it. The Intrusion detection techniques are divided into two groups which includes supervised learning and unsupervised learning. Clustering which is commonly used to detect possible attacks is one of the branches of unsupervised learning. Fuzzy sets play an important role to reduce spurious alarms and Intrusion detection, which have uncertain quality. This paper investigates k-means fuzzy and k-means algorithm in order to recognize Intrusion detection in system which both of the algorithms use clustering method.

3.1 An Improved Techniques Based on Naive Bayesian for Attack Detection [15]

With the enormous growth of computer networks and the huge increase in the number of applications that rely on it, network security is gaining increasing importance. Moreover, almost all computer systems suffer from security vulnerabilities which are both technically difficult and economically costly to be solved by the manufacturers. Therefore, the role of Intrusion Detection Systems (IDSs), as special-purpose devices to detect anomalies and attacks in a network, is becoming more important. The naive Bayesian Classification is use for intrusion detection system. One of the most important deficiencies in the KDD99 data set is the huge number of redundant records, which causes the learning algorithms to be biased towards the frequent records, and thus prevent them from learning infrequent records, which are usually more harmful to networks such as U2R and R2L attacks.NSL KDD data set have less redundant record .

3.2 Data Mining for Network Intrusion Detection [16]

This paper gives an overview of our research in building rare class prediction models for identifying known intrusions and their variations and anomaly/outlier detection schemes for detecting novel attacks whose nature is unknown. Experimental results on the KDDCup'99 data set have demonstrated that our rare class predictive models are much more efficient in the detection of intrusive behavior than standard classification techniques. Experimental results on the DARPA 1998 data set, as well as on live network traffic at the University of Minnesota, show that the new techniques show great promise in detecting novel intrusions. In particular, during the past few months our techniques have been successful in automatically identifying several novel intrusions that could not be detected using state-of-the-art tools such as SNORT. In fact, many of these have been on the CERT/CC list of recent advisories and incident notes.

3.3 Intrusion Detection based on Boosting and Naïve Bayesian Classifier [17]

In this paper, we introduce a new learning algorithm for adaptive intrusion detection using boosting and naïve Bayesian classifier, which considers a series of classifiers and combines the votes of each individual classifier for classifying an unknown or known example. The proposed algorithm generates the probability set for each round using naïve Bayesian classifier and updates the weights of training examples based on the misclassification error rate that produced by the training examples in each round.

3.4 Network Intrusion Detection Using Tree Augmented Naive-Bayes [18]

Computer networks are nowadays subject to an increasing number of attacks. Intrusion Detection Systems (IDS) are designed to protect them by identifying malicious behaviours or improper uses. Since the scope is different in each case (register already-known menaces to later recognize them or model legitimate uses to trigger when a

variation is detected), IDS have failed so far to respond against both kind of attacks. In this paper, we apply two of the ancient data mining algorithms called Naive Bayes and tree augmented Naive Bayes for network intrusion detection and compares them with decision tree and support vector machine. We present experimental results on NSL-KDD data set and then observe that our intrusion detection system has higher detection rate and lower false positive rate.

4. Proposed Methodology

Intrusion detection is the major task in networking. There are so many solution provided by the researchers for the detection of intruder in the network. Like Pattern Matching, Measure Based method, Data Mining method and Machine Learning Method.

Here we are detecting intrusion through data mining method. Basically we are combining two data mining technique called Fuzzy K-means and Naive Bayes classification and form hybrid technique. We are combining this technique because the existing rules are the knowledge from experts knowledge or other system. The different methods will measure different aspects of intrusions. Combine these rules may find the intruder attack like Denial of Service more quickly from the existing one and give less false detection.

4.1 Fuzzy K-means Algorithm

The clusters produced by the k-means procedure are sometimes called "hard" or "crisp" clusters, since any feature vector **x** either is or is not a member of a particular cluster. This is in contrast to "soft" or "fuzzy" clusters, in which a feature vector **x** can have a degree of membership in each cluster

- Make initial guesses for the means $\mathbf{m}_1$, $\mathbf{m}_2$,..., $\mathbf{m}_k$
- Until there are no changes in any mean:
  - Use the estimated means to find the degree of membership u(j,i) of $\mathbf{x}_j$ in Cluster i;
    for example, if a(j,i) = exp(- $\| \mathbf{x}_j - \mathbf{m}_i \|^2$), one might use u(j,i) = a(j,i) / sum_j a(j,i)
  - For i from 1 to k
- Replace $\mathbf{m}_i$ with the fuzzy mean of all of the examples for Cluster i --

$$\mathbf{m}_i = \frac{\sum_j u(j,i)^2 \mathbf{x}_j}{\sum_j u(j,i)^2}$$

  - end_for
- end_until It has the advantage that it more naturally handles situations in which subclasses are formed by mixing or interpolating between extreme examples, so that it makes more sense to say that **x** is 40% in Cluster 1 and 60% in Cluster 2, rather than having to assign **x** completely to one cluster or the other.

4.2 Support Vector Machine

Support Vector Machines (SVM's) are a relatively new learning method used for binary classification. The basic idea is to find a hyper-plane which separates the d-dimensional data perfectly into its two classes. However, since example data is often not linearly separable, SVM's introduce the notion of a "kernel induced feature space" which casts the data into a higher dimensional space where the data is separable [9].

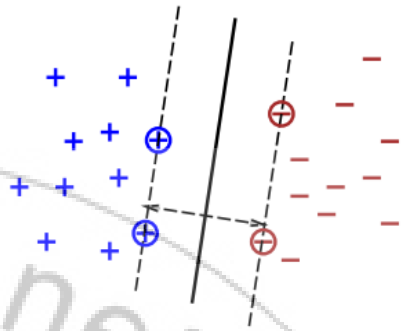

Figure1: SVM Hyper plane

4.3 Algorithm

Let we have L training points, where each input xi has D attributes (i.e. is of dimensionality D) and is in one of two classes $yi = -1$ *or* $+1$, i.e our training data is of the form:
xi, yi *where* $i = 1 \ldots L, yi \in -1,1 , x \in RD$
Here we assume the data is linearly separable, meaning that we can draw a line on a graph of $x1$ *vs* $x2$ separating the two classes when $D = 2$ and a hyper-plane on graphs of $x1, x2 \ldots xD$ for when $D > 2$.
This hyper-plane can be described by $w \cdot x + b = 0$, *where*:
• w is normal to the hyperplane.
• b w is the perpendicular distance from the hyperplane to the origin.
Support Vectors are the examples closest to the separating hyper-plane and the aim of Support Vector Machines (SVM) is to orientate this hyper-plane in such a way as to be as far as possible from the closest members of both classes. Iimplementing a SVM boils down to selecting the variables w and b so that our training data can be described by:
$w \cdot x\, i + b \geq +1$ *for* $yi = +1 \ldots \ldots \ldots$ (1.1)
$w \cdot x\, i + b \leq +1$ *for* $yi = -1 \ldots \ldots \ldots$ (1.2)
These equation can be combine into:

5. Hybrid Approach using Fuzzy K means and Support vector machine

Combining two data mining technique called Fuzzy K-means and Support vector machine and form hybrid technique. We are combining this technique because the existing rules are the knowledge from experts knowledge or other system. The different methods will measure different aspects of intrusions.

5.1 Working

Firstly we trained the database by using KDD database and fuzzy k means algorithm. Fuzzy k means detect outlier and extract the useful information which is relevant to our dataset. Using fuzzy k means clustering has been categorized .After extracting the useful information classify the dataset using Support Vector Machine algorithm. If

dataset is useful or relevant than classified into system related database otherwise classified as an intruder.

5.2 Proposed System Architecture

The proposed algorithm uses the support vector machine for the IDS and can be describe as follows:
Step 1: Read the KDD99 dataset.
Step 2: Preprocess the data by selecting the only attributes which are needed for testing from the feature vectors.
Step 3: Group the feature vectors according to their attack type.
Step 4: Now partition the above feature vectors into training and testing groups.
Step 5: Now cluster the training data using fuzzy K-means Clustering.
Step 6: From each cluster select the given percentage of data points as possible as away from the centroid of the cluster.
Step 7: Estimate the total classes in the Training dataset and form $N * (N - 1)/2$ (N is the number of classes in dataset) feature vectors group.
Step 8: Train the SVMs for $N * (N - 1)/2$ datasets and form similar numbers of SVM.

6. Conclusion

In this thesis we proposed a method for classification of intruder in system Intrusion detection is the major task in networking. There are so many solution provided by the researchers for detection of intruder in the network. Like Pattern Matching, Measure Based method, Data Mining method and Machine Learning Method. Here we detected intrusion through data mining method by combining two data mining technique fuzzy K means and Support Vector machine classification and formed a hybrid technique. We combined these different methods for measured different aspects of intrusions. Combined these rules find the intruder attack more quickly from the exiting one.

7. Future Aspect

In future, an association rule based approach or IF-THEN rules could be effective in classified the traffic in different classes. However accuracy of the algorithms plays an important role to correctly cluster the datasets. Standalone algorithms may not be able to provide efficient results. An another hybrid approach to data clustering can also be applied for analysis and to obtain low inter-cluster similarity.

References

[1] Pormohseni , Review and identify the computer Network intrusion detection systems , 2011 (Language in Persian).
[2] R. Heady , G. Luger , A. Maccabe , M. Sevilla.‖ The Architecture of a Network - level Intrusion Detection System, Technical report , CS90-20. Dept. of Computer Science, University of New Mexico, Albuquerque, NM 87131.pp:1-18, 1990.
[3] K . Scarfone and p . Mell , Guid to intrusion detection and prevention systems (idps), National Institude of Standard and Technology , Special publication 800 - 94 , page 127, 2007 . Availabel : http://csrc.nist.gov/publications/nistpubs/800-94/SP800-94.pdf ,Last Available: 23.08.2012.
[4] Bro IDS homepage, Available: www.bro-ids.org Last Available: 23.07.2012.
[5] A . A . horbani , W. Lu, M.Tavallaee, Network Intrusion Detection and Prevention : Concepts and Techniques , Springerpublisher , pages 234, 2009.
[6] J . P Anderson , Computer Security Threat monitoring and surveillance , (1980), Availabel: http://csrc.nist.gov/publications/history/ande80.pdf Last Availabel: 05.08.2012.
[7] A . hamidi , M . rezai , Introduction to Intrusion Detection System (Part I) , Technical report, Mashad University ,Iran,(language in Persian)
[8] Duda, R., P.E.Hart, D.G.Stork: Pattern classiffication. second edn. John Wiley & Sons (2001)
[9] Rojas, R.: Neural Networks: A Systematic Approach Springer-Verlag, Berlin, Deutschland (1996)
[10] Friedman , J. Regularized discriminant analysis . Journal of the American Statistical Association 84 (1989) 165-175
[11] SchÄolkopf, B., Smola, A.: Learning with Kernels. MIT Press, Cambridge, MA(2002)
[12] Portnoy, L., Eskin, E., Stolfo, S.: Intrusion detection with unlabeled data using clustering. In: Proc. ACM CSS Workshop on Data Mining Applied to Security. (2001)
[13] Laskov, P., SchÄafer, C., Kotenko, I.: Intrusion detection in unlabeled data with quarter sphere support vector machines. In: Proc. DIMVA. (2004) 71-82
[14] P.Garcia -Teodoro, J.Diaz- Verdejo, " Anomalyy network intrusion detection : Techniques, systems and challenges", www.elsevier.com , 2009.
[15] Mr. Manish Jain , Prof. Vineet Richariya " An Improved Techniques Based on Naïve Bayesian for Attack Detection " International Journal of Emerging Technology and Advanced Engineering Website : www.ijetae.com (ISSN 2250-2459, Volume 2, Issue 1, January 2012)
[16] Paul Dokas, Levent Ertoz, Vipin Kumar, Aleksandar Lazarevic, Jaideep Srivastava, Pang- Nig Tan " Data Mining for Network Intrusion Detection" Computer Science Department, 200 Union Street SE, 4-192, EE/CSC Building University of Minnesota, Minneapolis, MN 55455, USA
[17] Dewan Md. Farid, Mohammad Zahidur Rahman, Chowdhury Mofizur Rahman " Intrusion Detection based on Boosting and Naïve Bayesian Classifier International Journal of Computer Applications (0975 – 8887) Volume 24– No.3, June 2011.
[18] R.Naja, Mohsen Afsharsfi , " Network Intrusion Detection Using Tree Augmented Naïve Bayes" CICIS'12, IASBS, Zanjan, Iran, May 29-31, 2012.
[19] "Network Intrusion Detection Using Tree Augmented Naive-Bayes" ,CICIS'12, IASBS, Zanjan, Iran, May 29-31, 2012