A SVM and K-means Clustering based Fast and Efficient Intrusion Detection System

Alka Shrivastava
Dept of CSE
Samrat AshokTechnological
Institute, Vidisha M.P (India)

Ram Ratan Ahirwal
Dept of CSE
Samrat Ashok Technological
Institute, Vidisha M.P(India)

ABSTRACT

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 analyzed the effect of different input parameters on classification accuracy.

Keywords

Intrusion Detection System (IDS), KDD99 dataset, Support Vector Machine, K-means clustering.

1. INTRODUCTION

Computer network is now an essential part of modern life and used for almost every field like news, mail, audio and video communication, ecommerce, data exchanging and many other applications hence the security of network is a very important aspect. The computer network is also a soft and useful target for hackers because it can be easily accessed from any place and with minimal resources and it is also very difficult to trace the intruder and ones the intruder takes control over network it can get access to important information, bank transition and can completely shut down the network. The intrusion detection systems are used to detect such type of attack on a network. Mainly two techniques are used for development of IDS systems one is called signature based technique and other is anomaly based methods both the algorithms have their advantages and limitations such that the signature based methods provides better detection accuracy but it needed signatures of all intrusions in advance which is not possible for newly developed intrusions on the other hand anomaly based technique is much simpler but it suffers from false alarming and lower accuracy. This paper presents an anomaly based technique for IDS here the support vector machine is used for classification. The paper also analyzed the effect of different feature selection on the detection characteristics.

The rest of the paper is arranged as that the second section presents an overview of some recent work on the same field, the third section presents details about support vector machines (SVMs), in fourth section the KDD99 dataset is analyzed the fifth section explains the proposed algorithm followed by the simulation results and conclusions in sixth and seventh section respectively.

2. LITERATURE REVIEW

This section presents an overview of some recent IDS techniques available on the literatures. Shan Suthaharan et al [2] presented a probability distribution based approach that extract appropriate information from the intrusion data and supplies that information to the RST (Rough Set Theory) implementation so that the relevance features can be selected automatically. The proposed automatic feature selection approach simplifies and automates the detection of intrusion attacks with added advantages of high accuracy and less computing time. The Fuzzy Genetic Algorithm (FGA) is presented by Dalila BOUGHACI et al [3]. The FGA system is a fuzzy classifier, whose knowledge base is modeled as a fuzzy rule such as "if-then" and improved by a genetic algorithm. Unsupervised Network IDS (UNIDS) capable of detecting unknown network attacks without using any kind of signatures, labeled traffic, or training is presented by Pedro Casas et al [4]. UNIDS uses novel unsupervised outlier's detection approach based on Sub-Space Clustering and Multiple Evidence Accumulation techniques to pin-point different kinds of network intrusions and attacks. Kok-Chin Khor et al [5] presented a technique to increase the detection rates of the attack with insufficient data samples. Their approach relies on the training of cascaded classifiers using a dichotomized training dataset in each cascading stage. The training dataset is dichotomized based on the rare and non-rare attack categories. Hesham Altwaijry et al [6] proposed a naive Bayesian classifier to identify possible intrusions. Co-clustering approach is discussed by Evangelos E. Papalexakis [7] they look at the effectiveness of using two different co-clustering algorithms to both cluster connections as well as mark which connection measurements are strong indicators of what makes any given cluster anomalous relative to the total data set.

3. SUPPORT VECTOR MACHINE (SVM)

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].

3.1 Basic Theory [11]

Let we have L training points, where each input x_i has D attributes (i.e. is of dimensionality D) and is in one of two classes $y_i = -1 \, or \, +1$, i.e our training data is of the form:

$$\{x_i, y_i\} \, where \, i = 1 \ldots L, y_i \in \{-1,1\}, x \in R^D$$

Here we assume the data is linearly separable, meaning that we can draw a line on a graph of $x_1 \, vs \, x_2$ separating the two classes when $D = 2$ and a hyper-plane on graphs of $x_1, x_2 \ldots x_D$ 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.

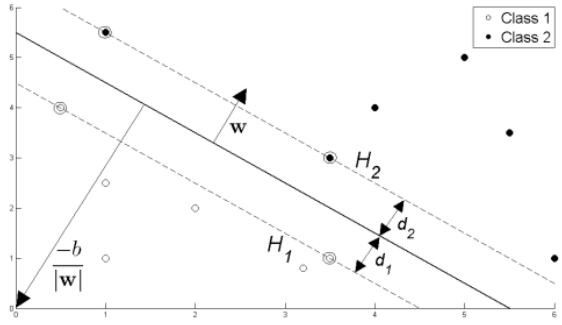

Figure 1: Hyper-plane through two linearly separable classes [11].

Referring to Figure 1, implementing a SVM boils down to selecting the variables w and b so that our training data can be described by:

$$w \cdot x_i + b \geq +1 \, for \, y_i = +1 \ldots\ldots (1.1)$$

$$w \cdot x_i + b \leq +1 \, for \, y_i = -1 \ldots\ldots (1.2)$$

$These \, equation \, can \, be \, combine \, into$:

$$y_i(w \cdot x_i + b) - 1 \geq 0 \, , \forall_i \ldots\ldots\ldots (1.3)$$

3.2 SVM for Multiclass Classification

The idea of using a hyper-plane to separate the feature vectors into two groups works well when there are only two target categories, but for more than two categories different approach required and many have been suggested, but two are the most popular: (1) "one against many" and, (2) "one against one" .

3.2.1 One against All

The earliest used implementation for SVM multiclass classification is probably the one-against-all method. It constructs SVM models where is the number of classes. The i^{th} SVM is trained with all of the examples in the i^{th} class with positive labels, and all other examples with negative labels. Thus given training data $(x_1, y_1), \ldots\ldots, (x_l, y_l)$ where $x_i \in R^n, i = 1, \ldots, l$ and $y_i \in (1, \ldots, k)$ is the class of x_i, the i^{th} SVM.

3.2.2 One against One

The 1A1 approach on the other hand involves constructing a machine for each pair of classes resulting in $N(N-1)/2$ machines. When applied to a test point, each classification gives one vote to the winning class and the point is labeled with the class having most votes. This approach can be further modified to give weighting to the voting process. From machine learning theory, it is acknowledged that the disadvantage the 1AA approach has over 1A1 is that its performance can be compromised due to unbalanced training datasets (Gualtieri and Cromp, 1998), however, the 1A1 approach is more computationally intensive since the results of more SVM pairs ought to be computed.

4. K-means Clustering

In data mining, k-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.

The most common algorithm uses an iterative refinement technique. Due to its ubiquity it is often called the k-means algorithm; it is also referred to as Lloyd's algorithm, particularly in the computer science community.

Given an initial set of k means $m_1^{(1)}, \ldots, m_k^{(1)}$ (see below), the algorithm proceeds by alternating between two steps [7]:

Assignment step: Assign each observation to the cluster whose mean is closest to it.

$$S_i^{(t)} = \left\{x_p : \left\| x_p - m_i^{(t)} \right\| \leq \left\| x_p - m_j^{(t)} \right\| \forall 1 \leq j \leq k\right\}$$

where each x_p is assigned to exactly one $S(t)$, even if it could be is assigned to two or more of them.

Update step: Calculate the new means to be the centroids of the observations in the new clusters.

$$m_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{x_j \in S_i^{(t)}} x_j$$

The algorithm has converged when the assignments no longer change.

5. KDD99 DATASET

This data set was used for The Third International Knowledge Discovery and Data Mining Tools Competition, which was held in conjunction with KDD-99 the Fifth International Conference on Knowledge Discovery and Data Mining. The competition task was to build a network intrusion detector, a predictive model capable of distinguishing between ``bad" connections, called intrusions or attacks, and ``good" normal connections. KDD training dataset consists of approximately 4,900,000 single connection vectors each of which contains 41 features and is labeled as either normal or an attack, with exactly one specific attack type. The simulated attacks fall in one of the following four categories [20][21]:

• Denial of Service (dos): Attacker tries to prevent legitimate users from using a service.

• Remote to Local (r2l): Attacker does not have an account on the victim machine, hence tries to gain access.

• User to Root (u2r): Attacker has local access to the victim machine and tries to gain super user privileges.

• Probe: Attacker tries to gain information about the target host.

The 41 feature set can be divided into four classes as listed below [21]:

Basic Features: Basic features can be derived from packet headers without inspecting the payload.

Content Features: Domain knowledge is used to assess the payload of the original TCP packets. This includes features such as the number of failed login attempts.

Time-based Traffic Features: These features are designed to capture properties that mature over a 2 second temporal window. One example of such a feature would be the number of connections to the same host over the 2 second interval.

Host-based Traffic Features: Utilize a historical window estimated over the number of connections in this case 100 instead of time. Host based features are therefore designed to assess attacks, which span intervals longer than 2 seconds.

6. CLASSIFICATION PERFORMANCE MEASURES

There are many measures available for judgment of the quality of the classifier and all of them are derived from the following confusion matrix [22]:

Data class	Classified as pos	Classified as neg
pos	true positive (tp)	false negative (fn)
neg	false positive (fp)	true negative (tn)

Figure 2: TP, TN, FP and FN estimation.

Accuracy: Overall effectiveness of a classifier:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

Precision: the number of correctly classified positive examples divided by the number of examples labeled by the system as positive:

$$Precision = \frac{tp}{tp + fp}$$

Recall: the number of correctly classified positive examples divided by the number of positive examples in the data:

$$Recall = \frac{tp}{tp + fp}$$

F-score: a combination of the Precession and Recall:

$$FMeasure = \frac{2 * Precision * Recall}{Precision + Recall}$$

7. PROPOSED ALGORITHM

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 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.

Step 9: Apply the testing data to the trained SVMs and stores the classification results given by each SVM.

Step 10: Calculate the total classifications for each class and get the maximally classified class and declare the input vector of that class.

Step 11: Repeat the same process for the each feature vector in the testing dataset and compare it with their original class and finally estimate the classification characteristics of the proposed algorithm.

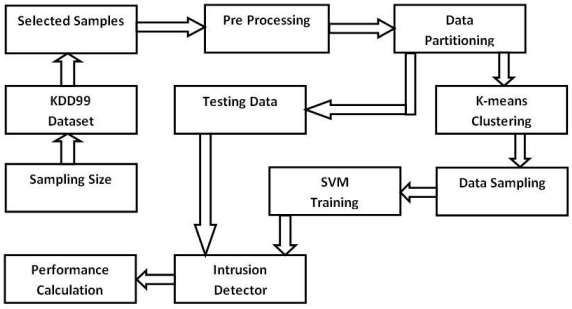

Figure 3: Block diagram of the proposed system

8. SIMULATION RESULTS

The simulation of the proposed algorithm is performed using MATLAB on Pentium 4 processor based PC with 2GB of RAM. The results are calculated for the different size of datasets and with different features. The results are also compared with the Navie Bayes based classification technique.

Simulation Parameters:

Training/Testing Data Ratio = 1;

Cluster data Selection Ratio = 0.5;

Table 2: Performance of Navie Bayes based System.

Dataset	TPR	TNR	FPR	FNR	Acc.	Prec.	Recall	F-meas
1000	0.9021	0.7934	0.2066	0.0979	0.8050	0.3417	0.9021	0.4956
2000	0.7832	0.9061	0.0939	0.2168	0.8102	0.9674	0.7832	0.8656
3000	0.6837	0.9885	0.0115	0.3163	0.9540	0.8836	0.6837	0.7709
4000	0.7846	0.9035	0.0965	0.2154	0.8259	0.8914	0.7846	0.8156
5000	0.9021	0.7934	0.2066	0.0979	0.9167	0.7167	0.0833	0.0833

Table 3: Performance of Proposed System (K-means and SVM)

Dataset	TPR	TNR	FPR	FNR	Acc.	Prec.	Recall	F-meas
1000	0.9680	0.9926	0.0074	0.032	0.9900	0.9397	0.9680	0.9536
2000	0.9931	0.9754	0.0246	0.0069	0.9892	0.9931	0.9931	0.9931
3000	0.9594	0.9984	0.0016	0.0406	0.9940	0.9873	0.9594	0.9731
4000	0.9866	0.9798	0.0202	0.0134	0.9898	0.9867	0.9866	0.9866
5000	0.9735	0.9888	0.0112	0.0265	0.9911	0.9734	0.9735	0.9733

8. CONCLUSION

The model of the Intrusion detector is presented in this paper is not only capable of attack situation but can also classifying the individual attacks. The Detection accuracy of the system is up to 90% which is excellent also the algorithm have very low FPR (max 8.3%) hence reduces the chances of false alarming. The results also shows that it takes only 0.0075 seconds to identify the intrusion hence fast enough to prevent any loss due to delayed action. Further it could achieve much better performance by increasing the number of samples taken and increasing the number of characteristics parameter selected.

REFERENCES

[1] Bing Wu, Jianmin Chen, Jie Wu and Mihaela Cardei "A Survey of Attacks and Countermeasures in Mobile Ad Hoc Networks", wireless/mobile network security, 2006 Springer.

[2] Abhay Kumar Rai, Rajiv Ranjan Tewari and Saurabh Kant Upadhyay "Different Types of Attacks on Integrated MANET-Internet Communication", International Journal of Computer Science and Security (IJCSS) Volume (4): Issue (3).

[3] Dr Karim KONATE and GAYE Abdourahime "Attacks Analysis in mobile ad hoc networks:Modeling and Simulation", 2011 Second International Conference on Intelligent Systems, Modelling and Simulation, 2011 IEEE.

[4] Farah Jemili, Dr. Montaceur Zaghdoud and Pr. Mohamed Ben Ahmed "A Framework for an

Adaptive Intrusion Detection System using Bayesian Network", 2007 IEEE.

[5] Jingbo Yuan , Haixiao Li, Shunli Ding and Limin Cao "Intrusion Detection Model based on Improved Support Vector Machine", Third International Symposium on Intelligent Information Technology and Security Informatics, 2010 IEEE.

[6] Z. Muda, W. Yassin, M.N. Sulaiman and N.I.Udzir "Intrusion Detection based on K-Means Clustering and OneR Classification", 2011 IEEE.

[7]http://link.springer.com/chapter/10.1007%2F978-3-642-14400-4_50?LI=true#.

[8] Martin Schütte " Detecting Selfish and Malicious Nodes in MANETs", SEMINAR: SICHERHEIT IN SELBSTORGANISIERENDEN NETZEN, HPI/UNIVERSITÄT POTSDAM, SOMMERSEMESTER 2006.

[9]http://www.personal.reading.ac.uk/~sis01xh/teaching/CY2D2/Pattern3.pdf

[10]http://voyagememoirs.com/pharmine/2008/06/22/probabilistic-neural-network-pnn/

[11] S. Nascimento, B. Mirkin and F. MouraPires "A Fuzzy Clustering Model of Data and Fuzzy c-Means", Fuzzy Systems, FUZZ IEEE 2000. The Ninth IEEE International Conference on 7-10 May 2000.

[12] R Rangadurai Karthick, Vipul P. Hattiwale and Balaraman Ravindran "Adaptive Network Intrusion Detection System using a Hybrid Approach", Communication Systems and Networks (COMSNETS), 2012 Fourth International Conference on 3-7 Jan. 2012.

[13] Chih-Wei Hsu and Chih-Jen Lin "A Comparison of Methods for Multiclass Support Vector Machines", IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 13, NO. 2, MARCH 2002.

[14] Eddy Mayoraz and Ethem Alpaydm "Support VectorMachines",http://www.cmpe.boun.edu.tr/~ethem/files/papers/iwann99.pdf

[15] Shigeo Abe "Analysis of Multiclass Support Vector Machines",http://www.lib.kobeu.ac.jp/repository/90000226.pdf

[16] Gidudu Anthony, Hulley Gregg and Marwala Tshilidzi "Image Classification Using SVMs: One-against-One Vs One-against-All", Proccedings of the 28th Asian Conference on Remote Sensing, 2007.

[17] P Amudha, H Abdul Rauf "Performance Analysis of Data Mining Approaches in Intrusion Detection", Process Automation, Control and Computing (PACC), 2011 International Conference on 20-22 July 2011.

[18] Sungmoon Cheong, Sang Hoon Oh and Soo-Young Lee"Support Vector Machines with Binary Tree Architecture for Multi-Class Classification", Neural Information Processing – Letters and Reviews Vol. 2, No. 3, March 2004.

[19] POWERS, D.M.W. "EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION", Journal of Machine Learning Technologies ISSN: 2229-3981 & ISSN: 2229-399X, Volume 2, Issue 1, 2011, pp-37-63.

[20] 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).

[21] H. Günes Kayacık, A. Nur Zincir-Heywood, Malcolm I. Heywood "Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion DetectionDatasets" https://web.cs.dal.ca/zincir/bildiri/pst05-gnm.pdf