ISSN:0975-9646

Improved Intrusion Detection Systemusing C4.5 Decision Tree and Support Vector Machine

Vaishali Kosamkar, Sangita S Chaudhari
Computer Department
A.C Patil College of Engg, Mumbai, India.

Abstract— **In today's era the security of computer system is of great concern. Because the last few years have seen a dramatic increase in the number of attacks, intrusion detection has become the mainstream of information assurance. While firewalls do provide some protection, they do not provide full protection and still need to be complimented by an intrusion detection system (IDS). Data mining techniques are a new approach for Intrusion detection. IDS system can be developed using individual algorithms like classification, neural networks, clustering etc. Such system yields good detection rate and less false alarm rate. Recent studies show that as compared to the single algorithm, cascading of multiple algorithms gives much better performance. False alarm rate was also high in such system. Therefore combination of different algorithms is performed to solve this problem. This paper we uses two hybrid algorithms for developing the intrusion detection system. C4.5 Decision Tree and Support Vector Machine (SVM) are combined to achieve high accuracy and diminish the wrong alarm rate. The experimental result shows that there is increase in the accuracy and detection rate and less false alarm rate. An attempt will be made to classify intrusions in various classes like Normal, DOS, R2L and U2R.Intrusion detection with Decision trees and SVM were tested with benchmark NSL- KDD, which is the advanced version of KDD Cup 99 for intrusion detection. Correlation- Based Feature Selection (CFS) algorithm is used of feature selection.**

Keywords— **Intrusion Detection System (IDS), Data Mining, Decision Tree, Support Vector Machine, NSL- KDD, Correlation- Based Feature Selection (CFS).**

I. INTRODUCTION

Intrusion detection can be defined as the process of detecting actions that attempt to compromise the confidentiality, integrity or availability of a systems resources. Mahoney defined Six types of attacks[1]. They are viruses, worms, server attacks, client attacks, network attacks and root attacks. Strong security policies such as anti- virus software, firewalls or other mechanism exist. Still it is difficult to detect these attacks, because every system has its own weakness and bugs. That's why the IDS are very much significant in today's world and can detect the new attacks. The goal of an IDS is to detect unwanted attempts at accessing, manipulating, and disabling of computer system. There are several techniques for the implementation of an IDS. The two most popular of them are:

Anomaly detection: This technique builds models for "normal" behavior of system by using either data mining or machine learning. Any deviation from this build model is considered to be an intrusion.

Advantage: can detect previous unseen attack.

Disadvantage: difficult to train system for dynamic changing environment, high false positives rates.

Misuse/Signature detection: These techniques apply the knowledge gathered about specific attacks. The IDS contains information about these attacks. Any action that is not explicitly recognized as an attack is considered normal. A constantly updated database is usually used to store the signatures of known attacks. The way this technique deals with intrusion detection resembles the way that anti-virus software operates.

Advantages: low false positive rates.

Disadvantages: need to update the signatures often, unable to detect novel attacks.

On the basis of detecting data target, the intrusion detecting system can be classified as host-based and Network-based [2].

Hosed Based IDS: Its data come from the records of various host activities, including audit record of operation system, system logs, application programs information, and so on.

Advantages: It givesbetter visibility of behavior of each applications running on the host.

Disadvantage: IDS is needed for every machine. If attacker takes over machine, can tamper with IDS binaries and modify audit logs.

Network Based IDS (NIDS): Collects audit data from the network traffic, such as: Internet Packets.

Advantages: single NIDS is needed which can protect many hosts and look for widespread patterns of activity.

Disadvantages: all attacks do not arrive from the network. High-speed links required to monitor record and process huge amount of traffic.

This paper is divided into four primary areas. The first section gives an overview of different approaches for intrusion detection. The second section describes various data mining techniques for intrusion detection. The third section represents the proposed system. Finally the results and comparative analysis are presented.

II. INTRUSION DETECTION APPROACHES

Intrusion detection system uses many approaches. A brief review of different approaches is described below which are considered for the development of intrusion detection systems [3].

A. Statistical approach

This approach involves statistical comparison of specific events based on a predetermined set of criteria. The data was collected from the system and the network. Statistical model tests this collected data for attack analysis. This was much laborious and time consuming work.

B. Rule based approach

Rule based approach uses a set of "if-then" implication rules to characterize attacks. Each rule is mapped to a specific operation in the system. The intrusion detection mechanism continuously checked rules that are in the audit record. If the required conditions of a rule are satisfied by user activity the specified operation is executed. The drawback of this approach was that it unable to detect new intrusion. This approach requires a frequent update of rules which is time consuming.

C. Expert System approach

Expert system is a system of software or hardware/software that is capable of competently executing a specific task performed by the human expert. The major drawback of Expert Systems is it requires frequent updates by a System Administrator. The lack of maintenance or update is the weakness of this approach.

D. Pattern recognition approach

In this approach, a series of penetration scenarios are coded into the system. This approach is effective in reducing the need to review a large amount of audit data. This is also unable to detect new attacks.

E. Artificial neural network approach

This approach is a substitute to other approaches. This approach may learn from examples. After training or learning the system is able to detect intrusion. This approach offers the potential to resolve a number of the problems encountered by the other present approaches such as varying nature of attacks. The first advantage in the use of a neural network in the intrusion detection would be the flexibility that the network would provide. A neural network would be able of analyzing the data from the network, even if the data is incomplete or partial. In the same way, the network would have the ability to conduct analysis with data in a non-linear fashion. Further, because some attacks may be conducted against the network in a coordinated attack by multiple attackers, the capability to process data from a number of sources in a non-linear fashion is particularly important. The problem of regularly updating of traditional intrusion detection systems is also reduced by ANN. It has generalization property and hence able to detect unknown and even variation of known attacks. Another reason to employ ANN in intrusion detection is that, ANN can cluster patterns which share similar features, thus the classification problem in intrusion detection can be solved by this approach. The natural speed of neural networks is another advantage.

III. Data Mining Techniques And Ids

Data mining is defined as the process of extracting useful information from the large databases. Data mining analyses the observed sets to discover the unknown relation and sum up the results of data analysis to make the owner of data to understand. Hence data mining problems are considered as a data analysis problem. Data mining framework automatically detect patterns in our data set and use these patterns to find a set of malicious binaries.ie, Data mining techniques can detect patterns in large amount of data, such as byte code and use these patterns to detect future instances in similar data. In intrusion detection system, information comes from various sources like host data, network log data, alarm messages etc. Since the variety of different data sources is too complex, the complexity of the operating system also increases. Also network traffic is huge, so the data analysis is very hard. The data mining technology have the capability of extracting large databases; it is of great importance to use data mining techniques in intrusion detection. By applying data mining technology, intrusion detection system can widely verify the data to obtain a model, thus helps to obtain a comparison between the abnormal pattern and the normal behavior pattern. Manual analysis is not required for this method. One of the main advantages is that same data mining tool can be applied to different data sources. How effectively it can separate the attack patterns and normal data patterns from a large number of network data and how effectively it generates automatic intrusion rules after collected raw network data is the major problem of current intrusion detection. To accomplish this various data mining techniques are used such as classification, clustering, association rule mining etc. [4].

Classification: Classification is a form of data analysis which takes each instance of a dataset and assigns it to a particular class.

Clustering: The amount of available network data is too large hence human labelling is time-consuming and also expensive. The process of labelling data and assigning into groups (clusters) is known as clustering. Clustering is a division of data into groups of similar objects. The members of same cluster are quite similar and members from the different clusters are different from each other.

Association Rule: The association rule considers each attribute/value pair as an item. Collection of items referred as an item set in a single network request. The algorithm searches to find an item set from large number of dataset that frequently appears in network. The main aim of association rule is to derive multi-feature correlations from a database table. Association rule mining determines association rules and/or correlation relationships among large set of data items.

Intrusion detection using Artificial Neural Network (ANN) and fuzzy clustering shows that ANN gives improved performance as compared with other traditional models. A new approach for intrusion detection system based on hierarchical clustering and Support Vector Machine (SVM) provides high qualified training instance to SVM and reduces the training time. So the overall performance of the SVM increases. Hybridization of neural network and C4.5 for misuse detection shows that the system cannot able to detect U2R and R2L network attack. Hybrid approaches improves the detection rate as compared to single approaches. Research on hybrid methodology is being carried out for developing IDS as it combines the advantages of two algorithms.

IV. Proposed System

The proposed system represents two hybrid algorithms for developing IDS, C4.5 and SVM. The advantage of C4.5 is that it gives maximum accuracy. The low false alarm rate is the advantage of SVM. Fig 1.shows the proposed framework of the intrusion detection system.

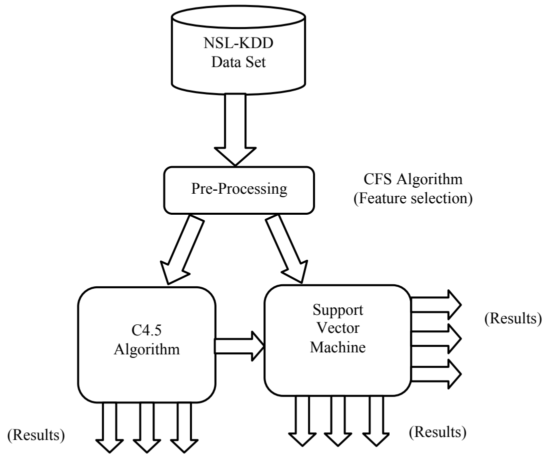


Fig 1: Proposed framework for IDS

A. Experiment Data Set

In this project the dataset that used is NSL- KDD. It is the advanced version of KDD Cup 99 for intrusion detection. The most important deficiencies of KDD data set is the huge number of redundant records, which causes the learning algorithms to be biased towards frequent records. The advantage of using NSL-KDD is that it does not include redundant records in the train set, so the classifier will not be biased towards more frequent records [5].

B. Preprocessing of Data and Feature Selection

Before analysis all the captured data needs to be organized in a particular format or pattern for the classification purpose. This whole process of organizing data is known as preprocessing. Data preprocessing is found to predominantly rely on expert domain knowledge for identifying the most relevant parts of network traffic and for constructing the initial candidate set of traffic features. On the other hand, automated methods have been widely used for feature extraction to reduce data dimensionality, and feature selection to find the most relevant subset of features from this candidate set. The main objective of our preprocessing module is to reduce ambiguity and provide accurate information to detection engine. Preprocessing of data results in reduction in false positive rate.

The Intrusion detection system deals with huge amount of data which contains irrelevant and redundant features causing slow training and testing process, higher resource consumption as well as poor detection rate. Feature selection, therefore, is an important issue in intrusion detection.

There are two algorithms used for feature selection and they are: - Correlation- Based Feature Selection (CFS) and Consistency- Based Feature Selection (CON) [6]. In this proposed system, we propose one of the most important method, the Correlation Feature Selection measure proposed by M. Hall. The CFS measures consider correlation between a feature and a class and inter correlation between features at the same time. This measure is used successfully in test theory for predicting an external variable of interest. It is desirable to get globally optimal subset of relevant features by means of the CFS measure with the hope of removing more redundant features and still keeping classication accuracies or even getting better performances. The Feature Selection Algorithm is given below [7].

Feature Selection Algorithm

```
1.  // Remove irrelevant features
2.  Input original data set $D$ that includes features $X$
    and target class $Y$
3.  For each feature $X_i$
      Calculate mutual information $SU(Y; X_i)$
4.  Sort $SU(Y; X_i)$ in descending order
5.  Put $Xj$ whose $SU(Y; X_i) > 0$ into relevant feature
      set $R_{XY}$
6.  // Remove redundant features
7.  Input relevant feature set $R_{XY}$
8.  For each feature $X_j$
      Calculate pair wise mutual information
      $SU(X_j; X_k)\ \forall j \neq k$
9.  $S_{XX} = \Sigma\ (SU(X_j; X_k))$
10. Calculate means $\mu_R$ and $\mu_S$ of $R_{XY}$ and $S_{XX}$,
      respectively. $w = \mu_S / \mu_R$
11. $R = w.R_{XY} - S_{XX}$
12. Select $X_j$ whose $R > 0$ into final set $F$
```

C. Intrusion Detection Algorithm Based on C4.5

Intrusion detection algorithm based on C4.5 can be divided into three stages [10]:

Stage 1: Construct decision tree

Algorithm: C4.5 Tree generates a decision tree from the given training data.
Input: training sample set T, the collection of candidate attribute. attribute-list.
Output: A decision tree.
Create a root node N;

- if T belong to the same category C, then return N as a leaf node, and mark it as a class C.
- if atrribute-list is empty or the remainder sample of T is less than a given value, then return N as a leaf node, and mark it as a category which appears most frequently
- In attribute-list, for each attribute, calculate its information gain ratio.
- Suppose test-attribute is the testing attribute of N, then test attribute=the attribute which has the highest information gain ratio in attribute-list
- if the testing attribute is continuous, then find its division threshold
- for each new leaf node grown by node N.
- Calculate the classification error rate of each node, and then prune the tree.

Stage 2: Extract classification rules
For decision tree, each branch represents a test output, and each leaf node represents category or category distribution. We just need to follow every path from root node to leaf node, the conjunction of each attribute-value constitutes the antecedent of rules, and the leaf node constitutes the consequent of rules. So decision tree can easily be converted into IF-THEN rules.

Stage 3: Determine network behaviour
For new network behaviour, determine whether it intrudes or not according to classification rules.

D. Support Vector Machine

SVM is used to solve Binary classification problems.SVM maps linear algorithm into non-linear space. For this mapping purpose it uses kernel function. There are various kernel functions like polynomial, radial basis function. This kernel functions can be used at the time of training of the classifier to selects support vector along the surface of this function. These support vectors are used by SVM to classify data that outline the hyper plane in the feature space. The linear classifier has the form

$$F(x) = w^t + b$$

where w is weight vector
b is bias

The distance between the data and hyper plane is shown in fig 2.below. The idea is that SVM maps inputs vectors nonlinearly into the high dimensional feature space and construct the optimum separating hyperplane.

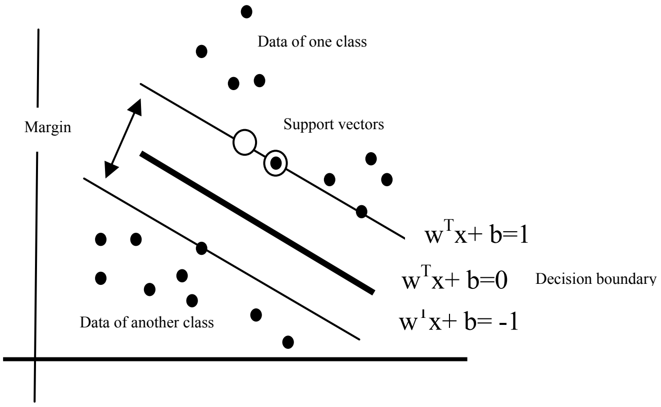
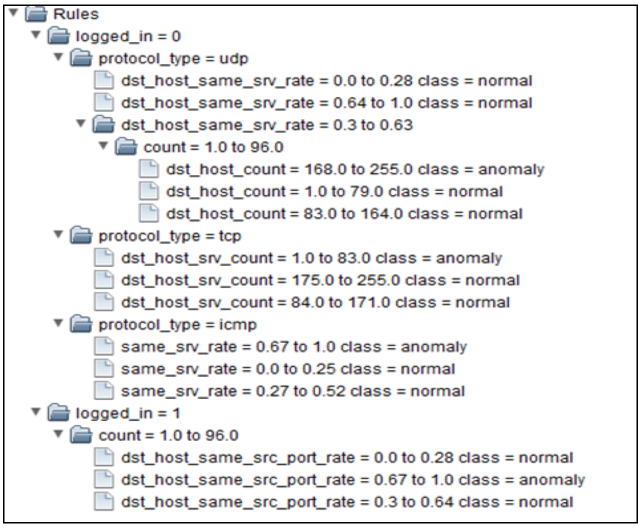


Fig 2: Hyper plane of Support Vector Machines

V. Experiment Results And Analysis

The outcomes of the implementation of the algorithm are shown below. The NSL-KDD Intrusion detection contest data is used in our experiments.First; the algorithms are trained with the preprocessed dataset. Dataset was separated into two parts. Through the first part, the model was prepared and with the remaining of the dataset, the model was tested. CFS algorithm was used for feature selection. Out of 42 features only 12 features were selected that are count, dst_host_count, dst_host_same_srv_rate, same_srv_rate, dst_host_srv_count, dst_host_same_src_port_rate, protocol_type, serror_rate, dst_host_srv_serror_rate, dst_host_serror_rate, srv_ serror_rate and logged_in.
Number of leaves produced by C4.5 algorithm is shown in fig 3.



- Rules
  - logged_in = 0
    - protocol_type = udp
      - dst_host_same_srv_rate = 0.0 to 0.28 class = normal
      - dst_host_same_srv_rate = 0.64 to 1.0 class = normal
      - dst_host_same_srv_rate = 0.3 to 0.63
        - count = 1.0 to 96.0
          - dst_host_count = 168.0 to 255.0 class = anomaly
          - dst_host_count = 1.0 to 79.0 class = normal
          - dst_host_count = 83.0 to 164.0 class = normal
    - protocol_type = tcp
      - dst_host_srv_count = 1.0 to 83.0 class = anomaly
      - dst_host_srv_count = 175.0 to 255.0 class = normal
      - dst_host_srv_count = 84.0 to 171.0 class = normal
    - protocol_type = icmp
      - same_srv_rate = 0.67 to 1.0 class = anomaly
      - same_srv_rate = 0.0 to 0.25 class = normal
      - same_srv_rate = 0.27 to 0.52 class = normal
  - logged_in = 1
    - count = 1.0 to 96.0
      - dst_host_same_src_port_rate = 0.0 to 0.28 class = normal
      - dst_host_same_src_port_rate = 0.67 to 1.0 class = anomaly
      - dst_host_same_src_port_rate = 0.3 to 0.64 class = normal

Fig 3: Number of leaves produced by C4.5 algorithm

Detection of attack can be measured by following metrics [9]:
False positive (FP): Or false alarm, Corresponds to the number of detected attacks but it is in fact normal.
False negative (FN): Corresponds to the number of detected normal instances but it is actually attack, in other words these attacks are the target of intrusion detection systems.
True positive (TP): Corresponds to the number of detected attacks and it is in fact attack.
True negative (TN): Corresponds to the number of detected normal instances and it is actually normal.
The accuracy of an intrusion detection system is measured regarding to detection rate and false alarm rate.
The efficiency of IDS System is calculated by using following two terms:
Detection rate:
Detection rate refers to the percentage of detected attack among all attack data, and is defined as follows:

$$\text{Detection rate} = TP/ (TP+TN) \times 100$$

False alarm rate:
False alarm rate refers to the percentage of normal data which is wrongly recognized as attack, and is defined as follows:

$$\text{False alarm rate} = FP/ (FP+TN) \times 100$$

The results that we got for C4.5, SVM and SVM (CFS) Algorithm for 1000 records are shown below in Table 1 with their corresponding values.

TABLE 1 RESULTS OF C4.5, SVM AND SVM (CFS) ALGORITHM FOR 1000 RECORDS

Parameters	**C4.5**	**SVM**	**SVM(CFS)**
Accuracy	94.7 %	90.4%	96.8
False Alarm Rate	9.46 %	6.12	2.34
Detection Rate	99.36	87.65	95.73

The results that we got for C4.5, SVM and SVM (CFS) Algorithm for 20% records are shown below in Table 2 with their corresponding values.

TABLE 2 RESULTS OF C4.5, SVM AND SVM (CFS) ALGORITHM FOR 20% OF RECORDS

Parameters	C4.5	SVM	SVM(CFS)
Accuracy	95.29 %	92.40%	98.30
False Alarm Rate	9.79 %	4.94	1.01
Detection Rate	100	90.09	98.623

VI. CONCLUSION & FUTURE WORK

After implementation we are able to increase the detection rate and decrease false alarm rate of Intrusion Detection System by combining two data mining algorithms C4.5 Decision Tree and Support Vector Machine. Comparison is done using various parameters like Accuracy, Detection rate, False Alarm Rate.

Building an effective intrusion detection models with good accuracy and real-time performance are essential.However, other kinds of preprocessing techniques and data mining approach like artificial intelegence,neural network models may be tested for a better detection rate in the future research in IDS System. An attempt will be made in future to classify types of attack into different categories like DOS, Probe, U2R and R2L. A more efficient feature selection algorithm can be used in future.

REFERENCES

[1] M. Mahoney, Computer security: A survey of attacks and defences, 2000. http://www.cs.fit.edu/~mmahoney/ids.html

[2] S. Wu, E. Yen. “Data mining-based intrusion detectors,” Elsevier computer Network, 2009.

[3] Iftikhar Ahmad, Azween B Abdullah and Abdullah S Alghamdi.”Comparative Analysis of Intrusion Detection Approaches”. In 12th International Conference on Computer Modelling and Simulation, 2010.

[4] Deepthy K Denatious and Anita John. “Survey on data mining techniques to enhance intrusion detection”. In Computer Communication and Informatics (ICCCI), 2012 International Conference on Digital Object Identifier, p. 1–5. IEEE, 2012.

[5] http://nsl.cs.unb.ca/NSL-KDD/

[6] Lei Yu and Huan Liu. Feature Selection for High-Dimensional Data: Fast Correlation-Based Filter Solution. Proceedings of the twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.

[7] T. S. Chou, K. K. Yen, and J. Luo.”Network Intrusion Detection Design Using Feature Selection of Soft Computing Paradigms”. World Academy of Science, Engineering and Technolog, 2008

[8] M. Tavallaee, E. Bagheri, W. Lu, A. A. Ghorbani, “A Detailed Analysis of the KDD CUP 99 Data Set”, Proceedings of the 2009 IEEE Symposium on Computational Intelligence, Ottawa, Canada, p. 53-58, 2009.

[9] Mohammadreza Ektefa, Sara Memar, Fatimah Sidi and Lilly Suriani Affendey.”Intrusion Detection Using Data Mining Techniques”. International conference on Digital Object Identifier, p.200-203, IEEE, 2010.

[10] Ron Kohavi and Ross Quinlan. Decision Tree Discovery. In Handbook of Data Mining and Knowledge Discovery.

[11] Liu Wu,Ren Ping,Liu Ke and Duan Hai-xin.”Intrusion Detection Using SVM”. 7th International conference on Digital Object Identifier, p.1-4, IEEE, 2011.

[12] Sandhya Peddabachigari, Ajith Abraham, and Johnson Thomas. Intrusion detection systems using decision trees and support vector machines. International Journal of Applied Science and Computations, USA, 11(3):p.118–134, 2004.