# A Comparative Study of Intrusion Detection Algorithms

Ruchi Jain
ME Student,
Department of Computer Science
Rajeev Gandhi Proudyogiki Vishwavidyalaya
Indore, M.P.

Anand Singh Rajawat
Assist. Professor,
Department of Computer Science
Shri Vaishnav Institute of Technology and Science
Indore, M.P.

## ABSTRACT

Intrusion detection system (IDS) is a kind of security management model that can be installed in computers and networks. IDS gather information from the network and computer and analyses it to find the possible security breaches into the system, which contain both intrusions and misuse. If we see modern IDS they also have few vulnerabilities, these systems also have drawbacks of false detection. So we aim to make a new model which can decrease the possibilities of false detection. In this paper we have done a survey of different algorithm used in intrusion detection system and then designed a new model that will use correlation based feature selection algorithm and SVM. This model will be tested on KDDCup99 dataset.

## Keywords

NIDS, HIDS, KDDCUP9 DATASET, CFS, SVM, U2R, R2L, PROBE, DoS

## 1. INTRODUCTION

Intrusion Detection System (IDS) is a software application that actively tests and monitors the system activities and the network traffic. It also finds the malicious activities that operate on the system or on the network. Growth in internet increase in the usage increased the concerns in protecting the digital media in a safe way. Now in modern time, hackers have also moved to new ways and new attacks to get the secret information. Many intrusion detection systems have evolved to help to detect the attacks. [1]

Different fields like business, finance, security and healthcare sectors, the LAN and WAN applications are progressing. All these areas are becoming the major target of attacks. These areas affect the whole community which gave attackers the target to get the important information. The attacker can either be inside the organization or outside the organization. Internally the malicious user use the internal system to collect the information and cause problems to internal software’s, or hack internal details or lapse the administration or leaving systems to its default configuration. With the advent growth the hackers are also advancing. New techniques and concept have risen to steal the important information. An intrusion detection system can be useful to safeguard our important information from the outer world like hackers, attackers. Crucial data of government and intelligence department are very crucial, so such system can be really helpful to provide security and correctness of data.

An intrusion detection system can be of various types for networks and for workstations. Different algorithms and rule base is used in this case. Similarly for different requirement the specifications of the IDS can also vary. So, before developing any system it is required to completely study the requirements and area of use. The Intrusion detection main targets on detecting the intrusion that had happened in the network or in the workstation, but the IDS doesn’t detect the intrusion before happening, it only tells that an attack has happened with the data being affected.

## 2. TYPES OF INTRUSION DETECTION SYSTEM

### 2.1 Network Intrusion Detection System

NIDS is a system that analyses the network and the ongoing traffic at all the OSI layers. It also checks for the suspicious activities that are going into the network. The intrusion detection system checks for attacks or strange behaviour of any packet by inspecting their contents and header while moving across the network. Most of the NIDSs are easy to deploy on a network and at one time they can view traffic from many systems. A NIDS is responsible to decide whether to pass the traffic into the network. Even wireless IDS to analyse the wireless network are widely used. The NIDS performs measures and monitors the network traffic for different networks segments and devices. It also analyses the network and application protocol activity to find the suspicious activity in the network. Basically there can be passive IDS and active IDS. Passive IDS only inform the administrator system about the attacks. It only detects the attack and do not take any measures to block the attack.

### 2.2 Host Based Intrusion Detection Systems

HIDS is used for single device or servers. Host based systems analyse the network traffic and system-specific settings. They are majorly used for publicly accessed servers or devices. HIDS needs to be installed on specific system; also all the configuration done is according to the specific operating system i.e. the HIDS are OS specific. [2] Host-Based systems monitors the characteristics and events of a single host occurring within that host to determine the suspicious activity. The HIDs works by collecting the data about the events taking place on a system. Or it also refers to the system logs, OS logs to analyse the event. The data recorded for the HIDS is recorded by operating system mechanisms called audit trails. Host-based systems mostly rely on audit trails.

## 3. DETECTION TECHNIQUES

### 3.1 Signature-Based Detection

An IDS can use signatures or specific pattern based detection, relying on known traffic data to find out the potentially unwanted traffic. This type of detection technique is very fast and easy. However, little change in the attack can fail the signature based system. Irrespective of this disadvantage the signature based system can be very accurate for specific attacks. A signature based systems are helpful in small co-operations and small areas like LAN, and MAN.

### 3.2 Anomaly-Based Detection

Anomaly detection system are those systems that detects new attacks that is those attacks that are unknown to the system itself. Signature based systems have a rule base with which the system checks and detects the attack but in anomaly based system algorithms are so designed that the system become capable of detecting even detecting new attacks. This method detects the unwanted traffic packets that is not specifically known by the system. These system would not detects where the attack has affected.

### 3.3 Stateful Protocol Inspection

Stateful protocol inspection is a detection technique in which the system detects unknown attacks similar to anomaly system. But the main advantage is it analyses the traffic at network, transport layer and application layer [3].

## 4. METHODOLOGY

### 4.1 Correlation Based approach

Feature selection algorithms are used to find and select the most salient features from a dataset. Its main purpose is learning of a classification algorithm. Feature selection is a process of eliminating redundant and irrelevant features from a data set. Correlation based Feature Selection (CFS) technique works on the hypothesis. CFS aims at having a good feature set that includes features that are highly correlated to the class and least correlated to each other. The algorithm ranks a feature on the basis of the extent to which it predicts a class and its correlation with other features. Correlation based feature selection algorithm can be found very well when talking about false detection in IDS.

### 4.2 Dataset

The main aim of ‘KDDCUP99 Data Sets’ is to use it in the KDDCUP ’99 Classifier-Learning Competition. The competition task was to build a network intrusion detector, a model capable of distinguishing between intrusions or attacks, and normal activates. This data set consists of a standard set of data, which includes a wide variety of intrusions simulated in a military network environment. The training and testing data from this data set were made available by Stolfo and Lee and the set consisted of a pre-processed data of the 1998 DARPA Evaluation Data. The KDD Cup 99 dataset was created by processing the tcp dump portions. The tcp portion was of the 1998 DARPA Intrusion Detection System Evaluation dataset. It was created by Lincoln Lab under contract to DARPA. The pre-processed version of DARPA dataset which contains only network data is known as KDD dataset. [4]

The input KDD Cup 1999 dataset is divided into two subsets one is training dataset and other is testing dataset. The training dataset is classified into five subsets Denial of Service, Remote to Local, User to Root, Probe and normal data using classification technique

KDD training dataset consists of approx. 4,900,000 single connection vectors in which each single connection vectors contain 41 features and is categorized as either normal or an attack, with exactly one particular attack type. KDDCup99 dataset has been good for the IDS systems. Various types of attacks described in four major categories [5]. KDDcup dataset is the best dataset to use for the training and testing purpose while developing the new algorithms by the researchers. It acts as a trace data having different attributes that can be tested on this dataset. A KDD dataset came into being after the DARPA dataset. But the DARPA dataset was not having this much number of attributes. So, to train and test the new implemented model a new dataset evolved which was later named as KDDCUP99 dataset. It has 41 attributes, which we will use in our new model for training and testing.

**Table 1. List of features given in KDD Cup 99 dataset**

| Feature index | Feature name | Description |
|---|---|---|
| 1 | duration | length (number of seconds) of the connection |
| 2 | protocol_type | type of the protocol, e.g. tcp, udp, etc. |
| 3 | service network | service on the destination, e.g., http, telnet, etc. |
| 4 | flag | normal or error status of the connection |
| 5 | src_bytes | number of data bytes from source to destination |
| 6 | dst_bytes | destination to source |
| 7 | Land | 1 if connection is from/to the same number of data bytes from host/port; 0 otherwise |
| 8 | wrong_fragment | number of ``wrong" fragments |
| 9 | urgent | number of urgent packets |
| 10 | hot | number of ``hot" indicators |
| 11 | num_failed_logins | number of failed login attempts |
| 12 | logged_in | 1 if successfully logged in; 0 otherwise |
| 13 | num_compromised | number of ``compromised" conditions |
| 14 | root_shell | 1 if root shell is obtained; 0 otherwise |
| 15 | su_attempted | 1 if ``su root" command attempted; 0 otherwise |
| 16 | num_root | number of ``root" accesses |
| 17 | num_file_creations | number of file creation operations |
| 18 | num_shells | number of shell prompts |
| 19 | num_access_files | number of operations on access control files |
| 20 | num_outbound_cmds | number of outbound commands in an ftp session |
| 21 | is_hot_login | 1 if the login belongs to the ``hot' list; 0 otherwise ' |
| 22 | is_guest_login | 1 if the login is a ``guest" login; 0 otherwise |
| 23 | count | number of connections to the same host as the current connection in the past two seconds |
| 24 | srv_count | number of connections to the same service as the current connection in the past two seconds |

| | | |
|---|---|---|
| 25 | serror_rate | % of connections that have ``SYN" errors |
| 26 | srv_serror_rat e | % of connections that have ``SYN" errors |
| 27 | rerror_rate | % of connections that have ``REJ" errors |
| 28 | srv_rerror_rate | % of connections that have ``REJ" errors |
| 29 | same_srv_rate | % of connections to the same service |
| 30 | diff_srv_rate | % of connections to different services |
| 31 | srv_diff_host_ rate | % of connections to different hosts |
| 32 | dst_host_coun t | count for destination host |
| 33 | dst_host_srv_c ount | srv_count for destination host |
| 34 | dst_host_same _srv_rate | same_srv_rate for destination host |
| 35 | dst_host_diff_ srv_rate | diff_srv_rate for destination host |
| 36 | dst_host_same _src_port_rate | same_src_port_rate for destination host |
| 37 | dst_host_srv_ diff_host_rate | diff_host_rate for destination host |
| 38 | dst_host_serro r_rate | serror_rate for destination host |
| 39 | dst_host_srv_s error_rate | srv_serror_rate for destination host |
| 40 | dst_host_rerro r_rate | rerror_rate for destination host |
| 41 | dst_host_srv_r error_rate | srv_serror_rate for destination host |

**Class Labels that Appears in "10% KDDCUP99" Dataset Attack Number of Samples Category**

**Table 2. Category of attacks**

| **Attack** | **Number of Samples** | **Category** |
|---|---|---|
| smurf. | 280790 | DOS |
| neptune. | 107201 | DOS |
| back. | 2203 | DOS |
| teardrop. | 979 | DOS |
| pod. | 264 | DOS |
| land. | 21 | DOS |
| normal. | 97277 | Normal |
| satan. | 1589 | Probe |
| ipsweep. | 1247 | Probe |
| portsweep. | 1040 | Probe |
| nmap. | 231 | Probe |
| warezclient. | 1020 | R2L |
| guess_passwd. | 53 | R2L |
| warezmaster. | 20 | R2L |
| imap. | 12 | R2L |
| ftp_write. | 8 | R2L |
| multihop. | 7 | R2L |
| phf. | 4 | R2L |
| spy | 2 | R2L |
| buffer_overflow. | 30 | U2R |
| rootkit. | 10 | U2R |
| loadmodule. | 9 | U2R |
| perl. | 3 | U2R |

# 5. LITERATURE REVIEW

## 5.1 Previous work

Te-Shun Chou el at [6] proposes a new scheme for correlation based feature selection for intrusion detection design. In this paper, the author aim to reduce the dimensionality of the original feature space by removing irrelevant, redundant features. The author has proposed a correlation-based feature selection algorithm for selecting a most informative subset of the total features. The author has retrieved six data sets from UCI databases and an intrusion detection benchmark data set, DARPA KDD99, are used to train and to test C4.5 and naive bayes machine learning algorithms. The author used Correlation analysis to be employed between feature and feature and between feature and class for removing irrelevant, redundant features from both low and high dimensional feature spaces. The experimental results demonstrate that algorithm has a superior performance in six UCI databases and KDD99 data set. In this paper the author's main focus is on correlation based algorithm

Hossein Gharaeeand Hamid Hosseinvand[7] Proposed a New Feature Selection IDS based on Genetic Algorithm and SVM. The new model has used a feature selection method based on Genetic with an innovation in fitness function, reduce the dimension of the data, increase true positive detection and simultaneously decrease false positive detection. The proposed model in this paper first selects the best feature subset in each class in terms of resulting the highest classification accuracy and true positive rate with the lowest false positive rate. Then collects the selected features in each class and gives priority based on the repetition of the features. The experiments and numeric results developed with the KDD CUP 99 and UNSW-NBI5 datasets showed the efficiency through classification accuracy, FP, TP rates and ROC curves. The results obtained with the GF-SVM model improve the detection accuracy to 99.05 % for normal traffic.

Prof. Amit Saxena et. al [8] proposes a survey on Intrusion Detection System on KDDCup99 Dataset. In this paper the author has done a survey of all the techniques implemented for the discovery and categorization of intrusions on KDDCup 99 dataset. By identifying various issues of dataset a new efficient algorithm is implemented by author who classifies and detects intrusions in KDDCup 99 dataset. In this paper the author has analysed different algorithms like Naïve Bayes which uses a probabilistic approach for the pattern generation of rules, K-means clustering is applied on the network traffic flow and hence classify the normal and abnormal instance traffic. C4.5 when applied enhances the classified instances of the dataset; SVM & Decision Tree effectively classify intrusions. Genetic Algorithm is mainly used to generate intrusive rules from the network traffic and then Fuzzy Logic is applied for the optimization of classified values. PSO-SVM is applied first for the grouping of similar features of the dataset and then support vector machine is used for the classification of the datasets.

Urvashi Modi, Prof. Anurag Jain [9] proposes a survey of IDS classification using KDD CUP 99 dataset in WEKA. This paper provides a detailed survey of intrusion detection techniques. It represents a study of Intrusion Detection and data mining techniques to classify different Intrusion attacks. This survey also focuses on WEKA Tool and its various algorithms of classification. This paper worked on machine learning techniques through which the author declare that machine learning algorithms employed as classifiers for the KDD CUP 1999 dataset don't offer a lot assure for detecting U2R and R2L attacks within the misuse detection context. But

Multivariate Adaptive Regression Splines give improved results in the detection of U2R and L2R attacks. The author worked on decision tree also to obtain an improved intrusion detection rates up higher than the 96% level and less false alerts from the rest of classifier data mining algorithms.

## 5.2 Literature summary

We have studied many papers and analysed different algorithms to find the best classification and feature selection algorithms. From the above papers we see that correlation feature selection algorithm set up a correlation with the features. It categorizes similar features into sets. And then gives a rank to these sets based on the quality of the sets. This way the algorithm may take more time for feature selection but resolves the problem of false alarm rate. IDS based on genetic algorithm and SVM performs the feature selection and the classification properly but the issue comes at local maxima. The genetic and SVM based IDS uses a feature selection algorithm based on genetic algorithm which results in improving the fitness function, detection rate and lowers the false alarm. Also with the survey it has been concluded that KDDCup99 proves to be a better set for the research of intrusions. It has 41 attacks with complete details to make it easier to be tested by IDS. Also different data mining techniques, machine learning techniques are discussed above also brief about their advantages and drawbacks are discussed. The major

Issue now days are false alarm rate. Our aim is to reduce the false alarm rate.

## 5.3 Comparison

In this paper, we have made a comparison of six algorithms. The parameters are authors, the classification and feature selection algorithms, year and performance of the algorithm.

**Table 3. Comparison between algorithms**

| **Author** | **Year** | **Algorithm** | **Accuracy** |
|---|---|---|---|
| Manjiri V. Kotpalliwar Rakhi Wajgi [10] | 2015, Classification of Attacks Using Support Vector Machine (SVM) on KDDCUP'99 IDS Database | SVM | 99.9 |
| Saurabh Fegade, Amey Bhadkamkar, Kamlesh Karekar, Jaikishan Jeshnani, Vinayak Kachare [11] | 2016, Network Intrusion Detection System Using C4.5 Algorithm | c4.5 plus svm | 96.71% |
| Basant Subba , Santosh Biswas, Sushanta Karmakar[12] | 2016, A Neural Network Based System for Intrusion Detection and Attack Classification | Neural Network | 94.74 |
| Mr. Vijay D. Katkar, Mr. Siddhant Vijay Kulkarni [13] | 2013, Experiments on Detection of Denial of Service Attacks using Naive Bayesian Classifier | Bayesian Classifier | 99.26% |
| Vidit Pathak,Dr. Ananthanarayana [14] | 2012, A Novel Multi-Threaded K-Means Clustering Approach for Intrusion Detection | Multi-Threaded K-Means Clustering | 99.18% |
| Hossein Gharaee, Hamid Hosseinvand [15] | 2016, A New Feature Selection IDS based on Genetic Algorithm and SVM | GF-SVM | 99.05% |

## 6. PROPOSED WORK

In this paper, we presented a new algorithm to detect the intrusions. We have used the KDDCup99 datasets for training the system and testing it. KDDCup99 dataset have 41 attributes, it has four classed U2R, R2L, probe and DoS. The given set is pre-processed i.e. noisy data is removed from the data. The major aim is to remove missing values and null values from the dataset. Further at this stage, symbolic-valued attributes of the datasets are mapped to numerical values. The encoding of symbolic data to numerical form is known as encoding. After encoding the feature selection process is carried. In feature selection our aim is to reduce the dimensions, size and complexity of data. Only the required amount of data in appropriate form is required. In this algorithm we are using correlation based feature selection algorithm (CFS). Because CFS makes use of all the training data at once, it can give better results than the wrapper on small datasets. CFS can be applied to larger datasets as compared to wrapper and CFS is much faster than the wrapper. Experiments on artificial datasets showed that CFS quickly identifies and screens irrelevant, redundant, and noisy features, and identifies relevant features as long as their relevance does not strongly depend on other features. Further, the dataset is divided into training and testing datasets in different ratios. The training data is given to the SVM whereas the testing data is given to the trained model. This process is continued with different ratios of training and testing datasets. Finally the performance of the system will be noted. The new system will be better in terms of false alarm rate. Our main aim of this new system with new algorithm is to improve the performance and decrease the false alarm rate. As, in the previous papers the major concern was about the false alarm rate. Many system have been developed that detects the attack completely but they also wrongly detects the right packets due to little change in the format. So, our major concern is to make a system that can be completely correct with minimum false detection.

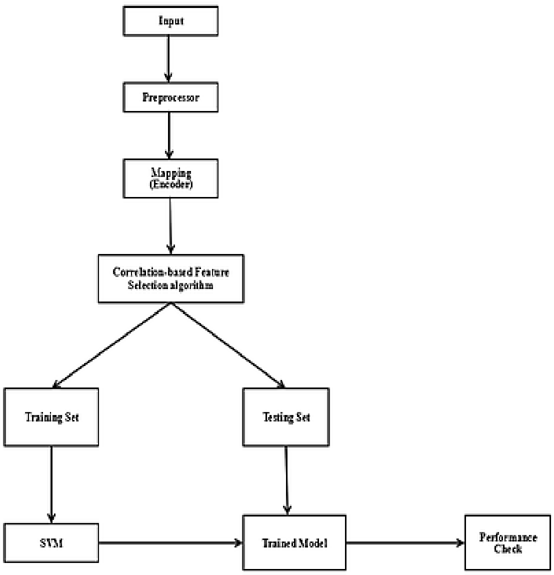

**Fig 1: Flowchart of the new Algorithm**

## 7. CONCLUSION

This survey paper aims at finding the best classification and feature mapping algorithm to design the best possible Intrusion detection system. As discussed in different papers, even the recent IDS systems have problem of false alarm. We studied different algorithm and in future we aims at creating a new algorithm which can improve the performance and decrease the false alarm rate. Our proposed algorithm will use the correlation based feature selection algorithm with SVM. This new algorithm will be trained and tested over KDDCup99 datasets. KDDCup99 dataset have 41 attributes and it is the best dataset for training and testing the new models. The data will be divided into testing and training dataset in different ratios. This will help to properly train and test the data. We will take 30 percent for training of the new model and 70 percent to test the attributes after the implementation of the new model. Our main of proposing a correlation based algorithm with SVM is to give a system with least false detection and best performance. This paper gives a brief about new model.

## 8. FUTURE SCOPE

Our model will be a combination of correlation based feature selection algorithm and SVM. This will give a model that will be best in terms of performance and false detection. In future we will aim to work on zero false detection system that can detect the data that is affected.

## 9. REFERENCES

[1] Dr. S.Vijayarani1 and Ms. Maria Sylviaa.S, “Intrusion Detection System – A Study” International Journal of Security, Privacy and Trust Management, Volume 4, No 1, February 2015

[2] Uman K Chaudhary, Ioannis Papapanagiotou, Flow classification using clustering and association rule mining.

[3] Jeffrey Erman, Martin Arlitt, Anirban Mahanti, “Traffic Classification Using Clustering Algorithms”, University of Calgary, 2500 University Drive NW, Calgary, AB, Canada

[4] MR Shanmugavadivu, “KDD Cup 99 Dataset” 2015

[5] Terry Brugger 15 Sep 2007 KDD Cup '99 dataset (Network Intrusion) considered harmful http://www.kdnuggets.com/news/2007/n18/4i.html

[6] Te-Shun Chou, Kang K. Yen, and Jun Luo Te-Shun Chou, Kang K. Yen, and Jun Luo “Correlation-Based Feature Selection For Intrusion Detection Design”, 2007

[7] Sonali Rathore et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (4) , 2015, 3345-3348, “Intrusion Detection System on KDDCup99 Dataset: A Survey”

[8] Ms. Urvashi Modi, Prof. Anurag Jain. A survey of IDS classification using KDD CUP 99 dataset in WEKA, International Journal of Scientific & Engineering Research, Volume 6, Issue 11, November-2015.

[9] Megha Aggarwal, Amrita, “Performance Analysis Of Different Feature Selection Methods In Intrusion Detection”, International Journal of scientific & technology research volume 2, issue 6, june 2013 ISSN 2277-8616 225 IJSTR©2013 www.ijstr.org

[10] Manjiri V. Kotpalliwar, Rakhi Wajgi, "Classification of Attacks Using Support Vector Machine (SVM) on KDDCUP’99 IDS Database", 2015, Fifth International Conference on Communication Systems and Network Technologies

[11] Building Efficient Intrusion Detection Model Based on Principal Component Analysis and C4.5 You Chen' ,Yang Li' , Xue-Qi Cheng1,Li Guo

[12] Basant Subba , Santosh Biswas, Sushanta Karmakar, "A Neural Network Based System for Intrusion Detection and Attack Classification", 2016

[13] Mr. Vijay D. Katkar, Mr. Siddhant Vijay Kulkarni, "Experiments on Detection of Denial of Service Attacks using Naive Bayesian Classifier" 2013

[14] Vidit Pathak,Dr. Ananthanarayana V. S., " A Novel Multi-Threaded K-Means Clustering Approach for Intrusion Detection" 2012

[15] Hossein Gharaee, Hamid Hosseinvand, “A New Feature Selection IDS based on Genetic Algorithm and SVM” 2016 8th International Symposium on Telecommunications (lST'2016)