INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

**ISSN: 2454-132X**
**Impact factor: 4.295**
***(Volume 5, Issue 5)***
Available online at: www.ijariit.com

# A review on integrating intrusion detection model using classifiers and chi-square feature selection

*Rayees Ahmad Sheikh*
*sheikhrayees24@gmail.com*
*CT Group of Institutions, Jalandhar, Punjab*

*Abhishek Bhardwaj*
*bhardwajabhishek786@gmail.com*
*CT Group of Institutions, Jalandhar, Punjab*

## ABSTRACT

**With the increasing use of IT technologies for maintaining the information, there is huge need for stronger intrusion detection mechanisms. The old usages like firewalls are not much so effective there are few new techniques that are used for the detection of intrusions. The various techniques of Intrusion Detection have been studied briefly. The main aim of this paper is to increase the accuracy with maximum and decrease with fewer false alarms. Also concern about reducing the disadvantages of the present IDS for large data. The intrusions enter the systems without any allowances to replicate files and running pirated software's on their own to access systems and takes the very much confidential data like sensitive information. Intrusion detection can take care of the systems and if something suspicious happens the prompt will be given and monitored.**

***Keywords***— *Intrusion detection, Chi-square*

## 1. INTRODUCTION

Intrusion are in brief intruders that violate the rules and regulations of the systems and networks to teal confidential information. Intrusions enter systems with the motives to copying files and running duplicate software and also defacement on the servers took place. There are two types Host IDs and NIDs. In Host IDS the single system had unique IDS and used for best detection. Host-based is first type as well as first layer of IDS and each individual system had one IDS. The second layer is Network IDS and it very much easy to deploy and one IDS will be used at a time with many systems. The second type IDS has deployed one IDS for many networks and has much benefits than HIDS. Intrusion detection has been mainly classified into two optimal categories. Misuse/signature is technique in which specific rules have been defined and according to that rules suspicious activities have been detected very well but only those which are familiar. The Anamoly has capability with better rules it can detect both known and unfamiliar. The Mutual information is actually amount of the information sending and received. It contained the conditional probability with symmetric properties and the positive values means non-negative. They are expressed to entropies and joint entropies. The IPS or IDS is important for security purposes and get better further strengthen for security.

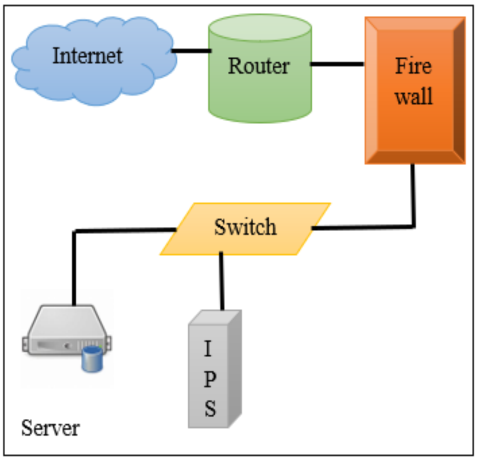


**Fig. 1: Instruction Detection system**

The chi square attribute selection is between the variables and would determines usually dependency. It is a feature optimisation and direct classification of features. Also, space will be reduced with this dimensionality gets reduced for the data and would make ranking of features.

## 2. LITERATURE REVIEW

Thaseen *et* al. [1] have proposed Chi-Square feature selection which reduces the dimensionality of the specific attributes. Their aim was to find the main critical features as well as better accuracy for the Intrusion detection model. In this they have used Majority voting technique in which it notices the votes of each and every classifier like election protocol. The classifier which gives more votes in comparison with the other classifiers has been selected. The classifier Support vector machine (SVM), Modified Naïve Bayes (MNB) and LP Boosting are used for building the optimal Intrusion detection. The NSL-KDD dataset performance has been analysed and it is new version of benchmark DARPA Intrusion detection. It shows good generalisation and improved accuracy i.e., our suggested model when combining with classifiers. The majority voting of final classifier has been predicted and the experimental results have shown Distibuted denial of service, R2L and normal have been detected. Kausar *et* al. [2] have suggested principal component analysis-based mechanism (PCA) for the SVM intrusion detection system. In present performance factor and training overhead are the drawbacks of several Intrusion detection. The very much increased efficiency and reduce very less false prompts with the maximising diagnosis rate. Main limitation in processing of the raw features of classifiers is much increasing complexity aritecture with accuracy reduce when detection increase. The processing of classifier overhead is a problem and for this motive have been use PCA. It can transform attributes into big dimension space and old minimising method by abstract several groups from attribute vector computing by the PCA. Now, these subsets separately train as well as test system through related resources and manipulate sensitivity. There is a lot of work shows maximising accuracy. Dhanabal *et* al. [3] have proposed the special Classification algorithm with the analysis of the data set. Basically, the refined version of NSL-KDD data set is KDD'99 of its predecessor dataset. The analysis of NSL-KDD dataset classification algorithms has the plenty of effectiveness and the study in network diagnosis anomalies. Therefore, the analysed relationship of protocols in network traffic are available patterned. Data mining classification algorithm are through the WEKA tool. The data set NSL-KDD of the analysis result best applicant data set to test feature IDS performance. Furthermore, CFS increases accuracy and the dimensionality minimizes as well as reduce the detection. Akashdeep *et* al. [4] have suggested intelligent system intrusion detection system which have potential to perform correlation and the information gain to perform well feature ranking firstly. So, to recognise the useless and useful attributes by using the Novel approach. The feedforward networks reduced the attributes of the networks for the testing on the KDD 99 dataset. Several instances of training before the preprocessing and then normalise the data. And the intelligently system behaves the categories into normal and attack classes. Their aim was that the attributes reduction system like in normal system to perform in same degree. They have tested five types of dataset so all average and unique results of data sates are reported. Zainal *et* al. [5] have proposed the ensemble in which sorting of the unique classes in which everyone have a learning model. The methods in this structure are LGP, Adaptive neural, ANFI and Random forest, integrating of many learning models shows increasing in diagnosis on the network traffic for accuracy. The limitation of this paper was many classification trees gather and results on unique address was capable to variance dataset with which machine knowledge method will unsuccessful to address it. Zhang *et* al. [6] have proposed that the well-structured of the Hybrid system which actually overcomes the drawbacks of more wrong positive rate in anomaly detection and cannot detect the unknown intrusions by misuse. The Random forest algorithm in both cases misuse and anomaly where been integrated. Their proposed result showed increased detection on NIDS with great performance on misuse and anomaly detection usage. Pietraszek *et* al. [7] have suggested two best optimal approaches of complementary and the orthogonal to minimize several wrong positives by with alert process Intrusion detection in machine learning and also in the data mining. The alert system and these both methods work together. In real and the simulated environment wrong positive have been much minimized. Mukherjee *et* al. [8] have suggested that the method for attribute reduction vitality based to recognise the essential input reduction attributes. In their work, they have used Naïve Bayes classifier and it is efficient dataset minimization for Intrusion detection. The minimized features of empirical results give performance to IDS as well as effective for network IDS. Panda *et* al. [9] have proposed that the Intrusion detection in anomaly based was given naïve Bayes for data mining algorithm is very much efficient. The experimental results show detection increases on network. They have shown wrong positives, computation time and cost better performances when compared with the neural network back propagation. Daejoon *et* al [10] have proposed that the neural network with cost ratio of wrong positive as well as wrong false errors. The neural network in first step develops and in second phase performance will be analysed, errors of asymmetric costs. In the experimental result of IDS network shows performance by higher accuracy.

## 3. COMPARATIVE ANALYSIS

Research scholars and professionals have used many techniques and methods In the Intrusion detection models.

**Table 1: Comparative analysis**

| S no. | Context of research | Techniques used | Problem Discussed | Advantages | References |
|---|---|---|---|---|---|
| 1 | Characteristics attributes for attack and normal traffic | Weighted majority voting technique | Combining several techniques with this detection reduces | Maximising accuracy when integrate with several classifiers. | [1] |
| 2 | Having possible feature reduction with maximum accuracy | (PCA) Principal Component Analysis | Accurateness of attack detection | performance increase when the attributes decrease | [2] |
| 3 | Integrity of algorithms | Classification Specific algorithm | Network traffic generate by attacks | Refining strengthen | [3] |
| 4 | Characteristics identified worthy and motive less | Intrusion detection system New intelligent | Complication and accumulation intensity | Efficiently System operates | [4] |
| 5 | Sorting and grouping in learning model | ANFIS, RF and LGP | False prompts and unsuccessful for results to address | Increasing in detection for accuracy | [5] |
| 6 | Integrating and framework of organisation | Random forest | Cannot detect unknown attacks | Maximising detection on NIDS | [6] |
| 7 | The complementary as | ID in Data mining and | Beneficial reduction in | false positives have | [7] |

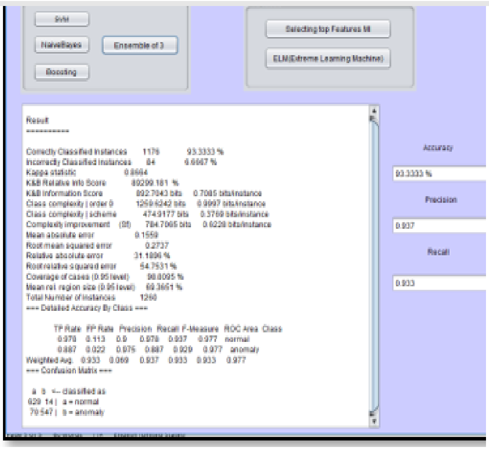
| | | | | | |
|---|---|---|---|---|---|
| | well as orthogonal approach | machine learning. | simulated and real environment | been reduced | |
| 8 | Statistically ID systematized | Attribute Reduction and Naïve Bayes | well-planned reduction | Exactness optimal | [8] |
| 9 | Neural network comparison | Data mining algorithms | Several attacks discussed | Better detection and will take time less | [9] |
| 10 | Performances of the neural networks developed | Accuracy shown in Intrusion detection | Minimization of the errors | Structure establish to achieve great accuracy | [10] |

The Netbeans with JDK and Java programming

**Table 2: Experimental Setup**

| | |
|---|---|
| **Platform Used** | Net Beans IDE 8.0 |
| **Database** | MySQL |

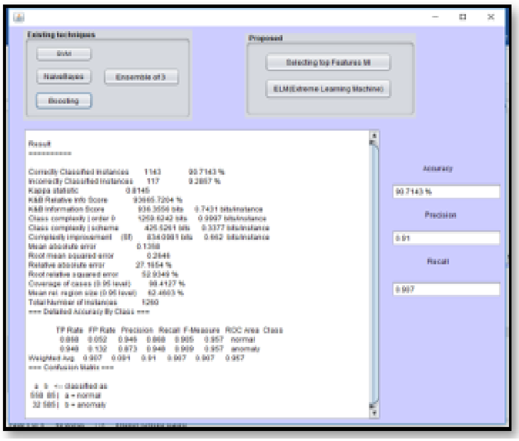
## 4. DISCUSSIONS



```
Result
===========

Correctly Classified Instances      1176          93.3333 %
Incorrectly Classified Instances      84           6.6667 %
Kappa statistic                        0.8664
K&B Relative Info Score            80299.181  %
K&B Information Score                892.7043 bits     0.7085 bits/instance
Class complexity | order 0          1259.6242 bits     0.9997 bits/instance
Class complexity | scheme            474.9177 bits     0.3769 bits/instance
Complexity improvement     (Sf)      784.7065 bits     0.6228 bits/instance
Mean absolute error                    0.1559
Root mean squared error                0.2737
Relative absolute error               31.1896 %
Root relative squared error           54.7531 %
Coverage of cases (0.95 level)        98.8095 %
Mean rel. region size (0.95 level)    69.3651 %
Total Number of Instances           1260
=== Detailed Accuracy By Class ===

              TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
               0.978    0.113    0.9        0.978   0.937      0.977     normal
               0.887    0.022    0.975      0.887   0.929      0.977     anomaly
Weighted Avg.  0.933    0.069    0.937      0.933   0.933      0.977
=== Confusion Matrix ===

   a   b   <-- classified as
 629  14 |  a = normal
  70 547 |  b = anomaly
```

```
Accuracy
93.3333 %
Precision
0.937
Recall
0.933
```

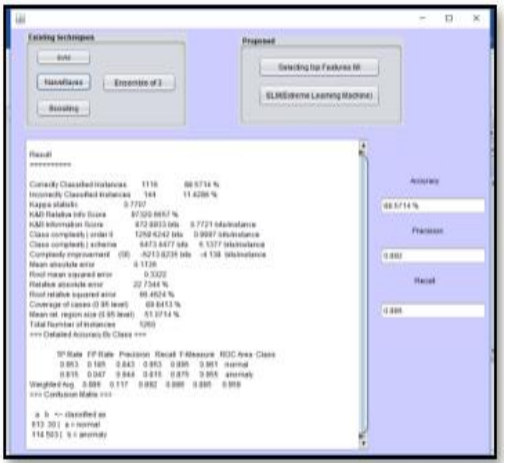
**Fig. 2: Base paper hybrid technique results**



```
Result
===========

Correctly Classified Instances      1143          90.7143 %
Incorrectly Classified Instances     117           9.2857 %
Kappa statistic                        0.8145
K&B Relative Info Score            93665.7204 %
K&B Information Score                936.3556 bits     0.7431 bits/instance
Class complexity | order 0          1259.6242 bits     0.9997 bits/instance
Class complexity | scheme            425.5261 bits     0.3377 bits/instance
Complexity improvement     (Sf)      834.0981 bits     0.662  bits/instance
Mean absolute error                    0.1358
Root mean squared error                0.2646
Relative absolute error               27.1654 %
Root relative squared error           52.9349 %
Coverage of cases (0.95 level)        98.4127 %
Mean rel. region size (0.95 level)    62.4603 %
Total Number of Instances           1260
=== Detailed Accuracy By Class ===

              TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
               0.868    0.052    0.946      0.868   0.905      0.957     normal
               0.948    0.132    0.873      0.948   0.909      0.957     anomaly
Weighted Avg.  0.907    0.091    0.91       0.907   0.907      0.957
=== Confusion Matrix ===

   a   b   <-- classified as
 558  85 |  a = normal
  32 585 |  b = anomaly
```

```
Accuracy
90.7143 %
Precision
0.91
Recall
0.907
```

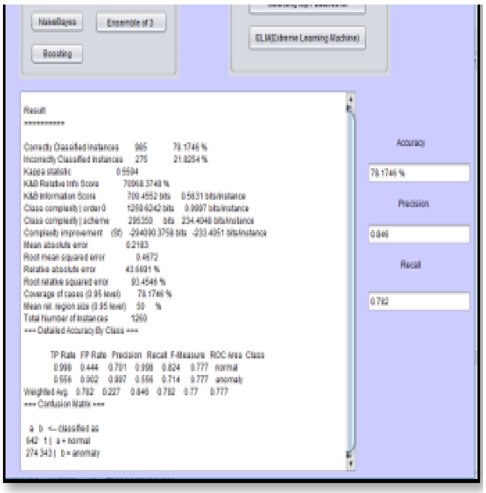
**Fig. 3: Results of boosting base technique**



```
Result
===========

Correctly Classified Instances      1116          88.5714 %
Incorrectly Classified Instances     144          11.4286 %
Kappa statistic                        0.7707
K&B Relative Info Score            87320.6657 %
K&B Information Score                872.8803 bits     0.7721 bits/instance
Class complexity | order 0          1259.6242 bits     0.9997 bits/instance
Class complexity | scheme           6473.6477 bits     5.1377 bits/instance
Complexity improvement     (Sf)    -5213.0235 bits    -4.138  bits/instance
Mean absolute error                    0.1126
Root mean squared error                0.3322
Relative absolute error               22.7344 %
Root relative squared error           66.4624 %
Coverage of cases (0.95 level)        89.6413 %
Mean rel. region size (0.95 level)    51.3714 %
Total Number of Instances           1260
=== Detailed Accuracy By Class ===

              TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
               0.953    0.185    0.843      0.953   0.895      0.961     normal
               0.815    0.047    0.944      0.815   0.875      0.955     anomaly
Weighted Avg.  0.886    0.117    0.892      0.886   0.885      0.958
=== Confusion Matrix ===

   a   b   <-- classified as
 613  30 |  a = normal
 114 503 |  b = anomaly
```

```
Accuracy
88.5714 %
Precision
0.892
Recall
0.886
```

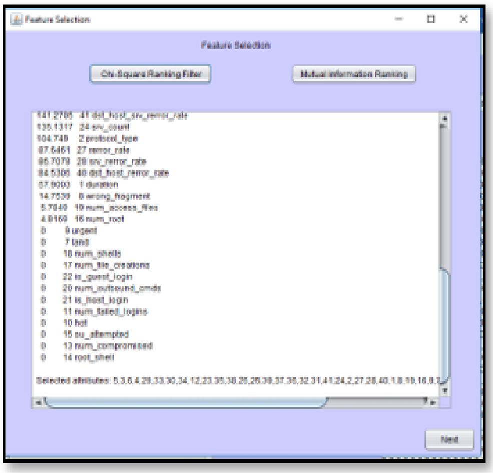
**Fig. 4: Results of Naïve base technique**



```
Result
===========

Correctly Classified Instances       985          78.1746 %
Incorrectly Classified Instances     275          21.8254 %
Kappa statistic                        0.5594
K&B Relative Info Score            70968.3748 %
K&B Information Score                700.4552 bits     0.5631 bits/instance
Class complexity | order 0          1259.6242 bits     0.9997 bits/instance
Class complexity | scheme         295350     bits   234.4048 bits/instance
Complexity improvement     (Sf) -294090.3758 bits  -233.4051 bits/instance
Mean absolute error                    0.2183
Root mean squared error                0.4672
Relative absolute error               43.6601 %
Root relative squared error           93.4546 %
Coverage of cases (0.95 level)        78.1746 %
Mean rel. region size (0.95 level)    50      %
Total Number of Instances           1260
=== Detailed Accuracy By Class ===

              TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
               0.998    0.444    0.701      0.998   0.824      0.777     normal
               0.556    0.002    0.997      0.556   0.714      0.777     anomaly
Weighted Avg.  0.782    0.227    0.846      0.782   0.77       0.777
=== Confusion Matrix ===

   a   b   <-- classified as
 642   1 |  a = normal
 274 343 |  b = anomaly
```

```
Accuracy
78.1746 %
Precision
0.846
Recall
0.782
```

**Fig. 5: Results of SVM base technique**



```
141.2705  41 dst_host_srv_rerror_rate
135.1317  24 srv_count
104.749    2 protocol_type
 87.6461  27 rerror_rate
 86.7079  28 srv_rerror_rate
 84.5306  40 dst_host_rerror_rate
 57.9003   1 duration
 14.7539   8 wrong_fragment
  5.7849  19 num_access_files
  4.9169  16 num_root
  0        9 urgent
  0        7 land
  0       18 num_shells
  0       17 num_file_creations
  0       22 is_guest_login
  0       20 num_outbound_cmds
  0       21 is_host_login
  0       11 num_failed_logins
  0       10 hot
  0       15 su_attempted
  0       13 num_compromised
  0       14 root_shell

Selected attributes: 5,3,6,4,29,33,30,34,12,23,35,38,26,25,39,37,36,32,31,41,24,2,27,28,40,1,8,19,16,9,1
```

**Fig. 6: Feature ranking using Chi-Squared features selection based technique**

## 5. CONCLUSION

Intrusions breach the systems with their own purposes. There is a lot of research is going for detection of Intrusions. With the proper management attacks can be look after through new machine learning classifiers rather than of traditional Intrusion detection systems. There are two categories of Intrusion detection Host and network-based IDs. The most beneficial techniques are misuse or signature based and anomaly based. The most familiar attacks have positively, detected by Misuse but have disadvantage this cannot diagnosis unfamiliar which are not known. But the anomaly has potential will smoothly detect both familiar and unfamiliar attacks. The group of a classifier of Support vector machine, MNB, Boosting and Hybrid of these has taken into consideration very well.

The Hybrid of these classifiers shows increased accuracy. In this paper we have make a comparative analysis of the several authentic Intrusion detection models in a simple way. So, in this there will be the optimal solution for the big data as well as large datasets. The final results of the classifiers will be degraded for huge amount of data, for this problem Extreme learning machine will be used to overcome that type of problem in the Intrusion detection and will analysis large data with efficiently. In future the proposed technique will integrate Mutual information priority ordered and the feature ranking technique with the learning technique of Extreme learning machine.

## 6. REFERENCES

[1] Thaseen S *et* al. Integrated Intrusion Detection Model Using Chi-Square Feature Selection and Ensemble of Classifiers. Arabian Journal for Science and Engineering 2018, 44(4); 3357-3368

[2] Kausar N *et al*. An Approach towards Intrusion Detection using PCA Feature Subsets and SVM. International Conference on Computer & Information Science (ICCIS) lEEE 2012; 978-1-4673-1938-6/12.

[3] Dhanabal L et *al*. A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms" International Journal of Advanced Research in Computer and Communication Engineering (June 2015) Vol. 4, Issue 6,

[4] Akashdeep et al. A feature reduced intrusion detection system using ANN classifier Expert Systems with Applications (2017); 88: S249–257

[5] Zainal A et al. Ensemble of One-class Classifiers for Network Intrusion Detection System. IEEE (2008); 978-0-7695-3324-7/08.

[6] Zhang J et al. A Hybrid Network Intrusion Detection Technique Using Random Forests. Proceedings of the First International Conference on Availability, Reliability and Security (ARES'06) IEEE 2006; 0-7695-2567-9/06.

[7] Pietraszek T, Tanner A. Data mining and machine learning Towards reducing false positives in intrusion detection. Information Security Technical Report (2005) 10, 169-183.

[8] Mukherjee S, Sharma N. Intrusion Detection using Naive Bayes Classifier with Feature Reduction. Procedia Technology (2012); 4:119 – 128.

[9] Mrutyunjaya Panda and Manas Ranjan NETWORK INTRUSION DETECTION USING NAÏVE BAYES IJCSNS International Journal of Computer Science and Network Security (December 2007), VOL.7 No.12.

[10] Daejoon Joo et al The neural network models for IDS based on the asymmetric costs of false negative errors and false positive errors Expert Systems with Applications 25 (2003) 69–75.