Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 13, Number 9 (2017), pp. 5635-5653
© Research India Publications
http://www.ripublication.com

# Features Contribution for Detecting Attacks of an Intrusion Detection System

**Shilpa Bahl* and Dr. Deepak Dahiya**

*CSE Department, Ansal University, Gurgaon, 122003, Haryana, India.*

## Abstract

Monitoring network traffic for any anonymous activity using intrusion detection system is a keen area of interest. For building intrusion detection system all the features are not essential, due to their redundant and irrelevant nature. However all the features do not have the ability to prove their relevance for distinguishing normal data from an attack classes**.** For the verdict of some malicious behavior, we assert that feature reduction is essential for modeling network traffic into a detection system. To substantiate this we reduce the feature space into a minimum and the most relevant set of features. It is also claimed that all the features does not contribute in improving the detection process. It therefore becomes the need of the hour to extract the most contributing features that improves the performance of an intrusion detection system. In this paper we have focused to affirm the relevance and contribution of each feature in the detection process. The proposed subset of 12 features in our previous work is proved here to be the most promising set of features by evaluating them on Naïve Bayes classifier, in comparison to other features subsets having a selection count of less than six and the complete set of 41 features. This subset of 12 features is chosen by extracting them among six different correlation based feature selection algorithms and allocated with a maximum selection count of 6 (i.e. commonly selected by all the 6 search algorithms). The objective of this paper is to justify the proposed minimal subset of 12 features against rest of the features using two different scenarios, emphasizing on their contribution in enhancing the classification accuracy and some other basic metrics, and computational time. It is observed that the features that are not frequently selected (Selection count 5,4,3,2,1) by six search algorithms are less relevant and declines the classifiers performance comparative to the 12 features that are selected by all the six search algorithm(Selection Count 6). Moreover there is a subset of features that are not selected at least once by any of the search algorithm (Selection Count:0). They are not at all involved in the detection process.

**Keywords:** Feature Subset Selection, Correlation Feature selection, Intrusion Detection System, Selection Count, Machine learning.

## I. INTRODUCTION

The modern world of today shows an extensive dependency on information and communication technologies (ICT), which have tremendously created a strong necessity of various ICT functionalities. ICT with it brings a high risk of very sophisticated and diverse threats from the network traffic carrying confidential information. Classical approaches to intrusion detection system were efficient for detecting the malicious activities for the well known patterns of the traffic, but they are not effective in detecting the novel attacks. Moreover the amount of data flowing in the network is approximately ten of gigabits per second i.e. around 15 millions frames are travelling per second. Therefore putting such a humongous data through high speed network on monitoring is a challenging task. In order to analyze each packet for any anonymous behavior effectively, substantial data reduction to an acceptable volume is required [1-3].

As intrusion detection systems (IDS) complement various network security techniques like firewalls, data encryption, user authentication etc. It basically detects any anonymous or malicious activity taking place in the network, host or at the server as an endeavor to misuse the confidential information by breaking the security walls [4, 5]. Majority of IDS made use of all the features of the data. It is a keen observation that all the features are not of equal relevance for the detection of attacks. Moreover every feature does not contribute in enhancing the system performance significantly i.e. for detection of any anomalous or suspicious behavior patterns. Some of the features may be eliminated in the pre-processing step, without radically deteriorating the performance of the IDS. IDS can be designed using various machine learning approaches i.e. classification as signature based misuse detection approach, flagging exceptions as anomaly based detection approach and grouping the same types as in clustering based detection approach. Classical signature based techniques are not able to detect novel attack patterns, therefore ML approaches effectively distinguish any unexpected behavior from the normal pattern by learning from already monitored network data. In this paper we have addressed supervised classification approach for modeling effective and efficient IDS [6-9].

NSL-KDDCup 99 dataset is used to evaluate the IDS model which is publically available standard dataset. This dataset is an enhanced version of KDDCup99 dataset, as it have huge number of duplicate examples, level of difficulty was high (imbalaced class distribution), and attacks were not properly defined. Therefore a new dataset called NSL-KDDCup99 dataset was introduced that is free from all these discrepancies. The dataset is broadly divided into four attack classes i.e. Probe, Denial of Service attack ( DoS), User to Root Attack(U2R), Remote to Local attack (R2L) and normal class, with 41 attributes and 42nd class attribute[10-12].

The quality of outcomes of the ML algorithm depends on the promising features of the input dataset. Therefore it is necessary to forward the input dataset with the most relevant features that can effectively detect the attacks in the network or machine. Feature selection (FS) is a well known dimensionality reduction technique that aims to select a subset of features contributing in the detection process. Feature selection is undertaken as a pre-processing step in data mining before the learning phase. This declines the negative effect of irrelevant and redundant features that tends to confuse the ML algorithm. FS not only improves the over accuracy of the classifier but also reduces the computational time of the ML process which makes it an extensible area of research and investigation [13-15].

It is investigated in the literature that each individual feature gives its contribution towards the detection of an attack and normal on the basis of degree of dependency and information gain computed. Moreover for some particular attack a single feature or combination two or more than two features are required for predicting the classes. It also observes that there are some features that not at all contribute in detecting any of the attack having least dependency ratio [16, 17].

The objective of the work done in this paper is to show the most contributing subsets of the features among the proposed subset 12 features commonly selected by six search algorithms of feature selection, having a selection count(SC) of 6[18], and the subsets having a count of 5,4,3,2,1 . It is observed that features that have count less than 6 tends to improve the classifiers performance comparative to complete set of features but not in competence to the subset of 12 features having count 6. This shows that subset of 12 features is the most relevant and contributing set of features in detecting the maximum number of attacks. Naïve Bayes ML algorithm is used to evaluate all the possible subets of features [18].

## II. RELATED WORK

As the research on data mining techniques for building an intrusion detection system is growing rapidly. The performance and efficiency of IDS depends upon the input data log forwarded. The input data log contains a large number of features; all the features are not relevant and contributing for the detection of attacks. Therefore it becomes necessary to reduce the feature space using a suitable feature reduction technique. Feature selection has been carried as a requisite pre-processing step. It is a process of reducing the number of suitable traffic features without an adverse effect on classification accuracy. Dimensionality reduction also enhances the efficiency of IDS [19-21].Feature selection is broadly divided into two categories: Filter approach which is independent of the induction algorithm and depends on some characteristics (e.g. distance, correlation etc) of the features, while the wrapper approach depends on the induction algorithm. Feature selection improves the performance of the classifier by reducing the feature space and lowering the computational time to build IDS [22]. Lot of work has been conducted in the literature for constructing IDS by reducing the dimensionality of the input data. Ten well known classifiers from Bayesian network family, Decision trees, and rule based family have been implemented to evaluate selected subset of features for classifying four attack classes against normal behavior [23]

In other related work using Classification and Regression Trees (CART) and Bayesian Networks (BN) the authors have given ensemble feature selection algorithms which yielded in lightweight IDS [24]. [25] used the ranking technique to select significant features for intrusion detection. It sequentially deleted the features from the dataset one at a time and the reduced dataset, which is then used for the training and testing of the Support Vector Machine (SVM). Features are ranked on the basis of their importance according to the rules constructed using fuzzy logic [17]. Zaman, presented one of the ranker methodologies and correlation between the features technique for feature selection using enhanced support vector decision function for feature selection [26]

It is well discussed in the literature that feature are selected on the basis of their relevance and involvement in effective classification of the attacks to their respective classes. There are various platforms on the basis of which these features are evaluated for their usefulness like dependency ratio, information gain, distance, and correlation etc. [27, 28]

The author in his work discussed that feature selection is carried out by analyzing their redundancy and relevance, therefore the features are grouped into three categories: Strong relevance, weak relevance and irrelevant. Also while constructing an optimal subset; the features under first group (strongly relevant) are always mandatory to be included unless and until there is some change in conditional class distribution. However weak relevant features are necessary at some certain conditions, while irrelevant features are not to be included at all. On the basis of characteristics of the features, some features have similar characteristics are redundant and can be eliminated. Results for 10 UCI datasets on Naïve bayes and C4.5 classifier, using five feature selection techniques have been presented in [29]. The author implemented Euclidean Distance for selecting the subset of 29 features which is evaluated on support vector machine. The literature also presented that inter-class distance measures result in better performance compared to probabilistic measures, i.e. correlation based feature selection is one of the most promising FS technique among the other implemented for some particular classifiers e.g. Naïve Bayes and decision trees.

Olusola in his work shows the relevance of each feature as their contribution in detection of some particular attack on the basis of their dependency ratio i.e. each attack is dependent on a single or combination of two or more features for their detection. Moreover the attacks (spy, rootkit) with very less representative features are difficult to detect. Similarly there are some features also that have no contribution in classification process. The author have also mentioned some features that have no involvement for predicting the attack classes e.g. 13,15,17,20,21,22, and 40[16]. In our work these particular are mentioned under the least selection count subsets (Count 0,1)[18]. Kayacik also discussed the relevancy of features computed on the basis of information gain for detecting some particular attacks into their respective classes. Also as discussed in the above work done by Olusola features 20 and 21 are not at all contributing in detecting any of the attacks. Moreover there are 9 features that have information gain less than 0.0001 which have almost negligible contributing nature [17].

## III. RELATED THEORY

### Dataset

The dataset used in this paper is customized KDDCup 99 dataset i.e. NSL-KDDCup data. As seen in the literature KDDCup 99 dataset had some discrepancies of redundant and unevenly distributed class problem. Though the dataset have some serious structural issues in the dataset still it remains to be the standard for building an intrusion detection model. The dataset have 21 attacks in training file and 37 attacks in testing file respectively. There are some novel attacks present in testing file that do not have their signatures in training file. The dataset have various types of structural files, we have implied 20% training dataset and full testing dataset in our work. The dataset is divided into five classes Normal, Probe, DoS, U2R and R2L [10-12].The details of the dataset are shown in the table 1 below.

**Table 1:** Details of Attacks instances in NSL-KDDCup 99 dataset

| **KDDTrain$^+$** | | **KDDTest$^+$** | |
|---|---|---|---|
| Normal | 12828 | Normal | 9712 |
| DoS | 8819 | DoS | 7461 |
| Probe | 2237 | Probe | 2422 |
| R2L | 205 | R2L | 2886 |
| U2R | 12 | U2R | 68 |
| Total | 25192 | Total | 22544 |

### WEKA Data mining Tool

WEKA is a JAVA based benchmark tool used in the area of data mining. It is an open source tool available through general public license. It is user friendly software even for novice users. It has various working platforms for classification, visualization, clustering, regression and pre-processing etc. In this work we have used WEKA 3.7.11 on windows7. It has four application programs in the package i.e. explorer, experimenter, knowledge flow and command line interface [30].

### Feature Subset Selection

Feature selection is one of the pre-processing phase to carry an effective and efficient data mining process. It basically eliminates the irrelevant and redundant set of features which are not at all important for predicting the target class. That is before extracting some useful information or analyzing the trend of the data , pre-processing step should be added. Feature selection is divided into two broad categories: filter and wrapper methodologies, independent of the induction algorithm and dependent on the same respectively. Filter method is entirely dependent on the characteristics of the features. Therefore former technique is preferred to be used for various machine learning processes. Feature selection also follows a directional pattern i.e it eliminates the features either in forward direction starting with an empty set and in backward direction starting with the full set of features called sequential forward selection and sequential backward elimination respectively[31, 32].

Feature selection process is followed by four basic steps:

- **Subset Generation**: It is a process of generating the candidate feature subsets for evaluation phase, using various search strategies (Best first, Random search. Greedy search etc.)

- **Subset Evaluation**: Numbers of subsets generated in the previous phase are evaluated by comparing it with the previous best one based on some evaluation criteria. The better turned out subset will replace the previously chosen subset.
- **Stopping Criteria**: The above two processes continues repeatedly until some stopping criteria is not reached.
- **Result validation**: The best selected subset is validated by using some kind of test e.g. implementing the classifiers algorithm or some clustering technique etc.

In this paper we have followed correlation based feature selection technique for reducing the feature space to a relevant subset of features for detecting the attacks on in the network.

### Correlation Based Features Selection

A major challenge in the IDS feature selection process is to choose appropriate measures that can precisely determine the relevance and the relationship between features of a given data set and the predictive class attribute. Therefore one of the promising filter based feature subset selection technique which is based on the principal of Pearson's coefficient of correlation. According to Pearson, features that are extremely coupled with the projecting class and loosely with each other are the most relevant for ML process. It states that the feature should be tightly coupled with the target class attributes and loosely coupled among each other. In simple words statistical dependency tends to quantify how tightly two features are associated with each other and with the attack class attribute i.e. by knowing one, we can predict the other [33]. The correlation coefficient computed for the two attribute sets $X_i$, $X_j$ and k is the number of features, by the equation (1)

$$R_{X(i)X(j)}= \frac{\sum_k^m\left(x_k^i-\overline{x^i}\right)(x_k^j-\overline{x^j})}{(m-1)S_i\,S_j} \tag{1}$$

$$Ms=R_{FC}=\frac{K\,r_{fc}}{\sqrt{K+k(K-1)r_{ff}}} \tag{2}$$

where $R_{FC}$= Correlation among the Attack class and the attributes.

$r_{fc}$= Average value of attribute– Attack class correlation.

$r_{ff}$ = Average value of attribute-attribute correlation

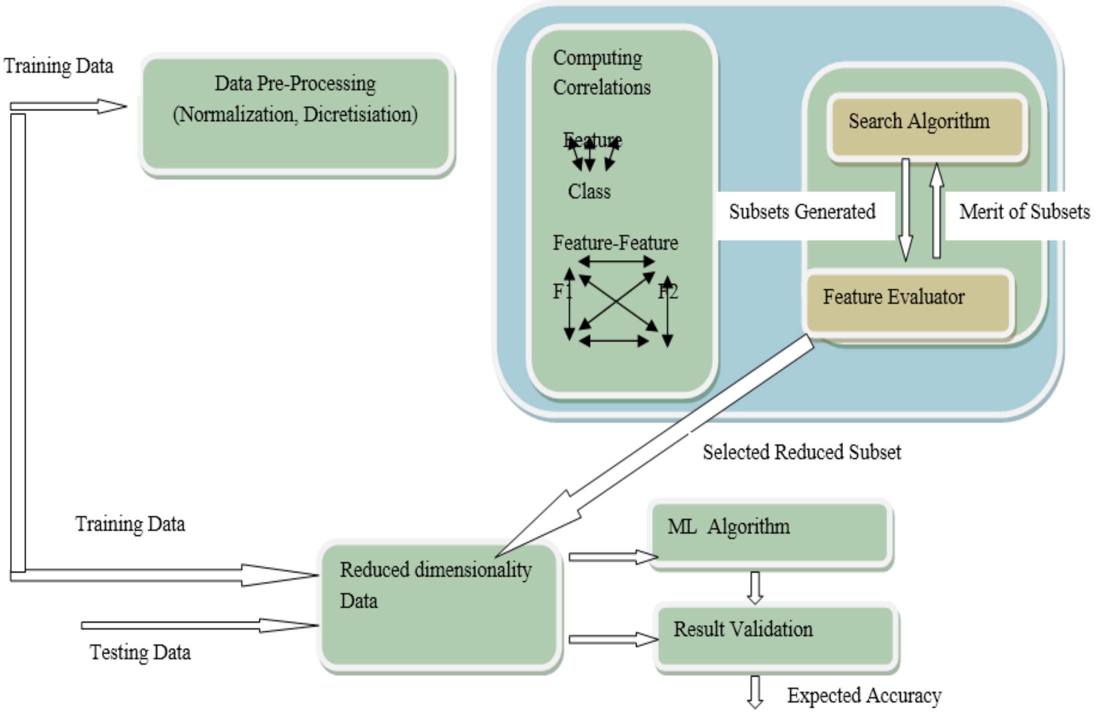

**Figure 1:** Filter based Correlation Feature Selection Technique

The six search algorithms implemented in this work are: Best first, Greedy Stepwise, Exhaustive search, Genetic search, Random search and Scatter search VI. The details of the features reduced using these search algorithms are given in Table:2

## Machine learning Algorithm

### Naïve Bayes Algorithm

It is a general probabilistic classifier which is independent of the class conditions. It works on “Bayes theorem” which calculates subsequent probabilities from the previous probabilities by recording both of them. Counting the frequency of occurrence is used to make an estimate of the two probabilities. Naïve Bayes algorithm is used to show a comparative study among the reduced subsets and complete set of 41 features. As an outcome it gives the class labels having maximum probabilities for making the decision. The algorithm works on the principal of “conditional independence” (Naive) i.e. the probability (occurrence or not) of one attribute is independent of the probability (occurrence or not) of other attributes and also to the known value of the target class attribute. It is an observation reported in the literature that the merit of Naïve Bayes, decision tree family some of the neural network classifiers is equivalent[7,8].

**Advantages of Naïve Bayes**

- Have very simple structure
- Construction of the model is very simple.
- Classification is achieved in linear time.
- Naïve bayes can be updated very easily as its construction is linear.
- It works on the principal of strong independence assumptions.

## IV. RESEARCH METHODOLGY

In the previous work done we have proposed a minimal subset of 12 features [18] which are the most contributing features in significantly enhancing the performance of the classifier and to detect the rare class attack as well on WEKA 3.7.11 data mining tool. The empirical outcomes for naïve bayes classifier had shown a noticeable enhancement for various metrics (accuracy, true positive rate RMSE, false positive rate, time to build the model etc.) on this reduced feature subset.

**Table 2:** Details of Feature Selection Algorithm with selected Subsets

| Sr. No. | Search Method | Selected Subset of Features | #Features Selected |
|---|---|---|---|
| 1 | Best First | [2,3,4,5,6,8.10,12,23,25,29,30,35,36,37,38,40] | 17 |
| 2 | Greedy Stepwise | [2,3,4,5,6,8,10,12,23,26,29,30,35,36,37,38,40] | 17 |
| 3 | Genetic Search | [2,3,4,5,6,8,12,13,22,23,25,26,27,29,30,31,32,33,36,37,38] | 21 |
| 4 | Scatter Search V 1 | [2,3,4,5,6,8,11,12,14,23,25,29,30,35,36,37,38,40] | 18 |
| 5. | Exhaustive Search | [2,3,4,5,6,8,10,12,14,23,25,29.30,31,35,36,37,38,40] | 19 |
| 6. | Random Search | [2,3,4,5,6,8,10,11,23,25,26,27,29,30,36,37,38] | 17 |
| 7. | Proposed subset | [2,3,4,5,6,8,23,29,30,36,37,38] | 12 |

On the basis of the selection count of six, i.e. these 12 features are commonly selected features by six feature selection algorithms mentioned in table: 3, these 12 features are meant to be the important subset of features for detecting the attacks [18]. The focus of the work done in this paper is to show the involvement or contribution of rest of the subets that have a selection count less than six i.e. the features that are not very frequently or not at all selected by six search algorithms. These all subsets of features (SC: 6,5,4,3,2,1,0) constructed with combinational logic of features selection frequency among six search algorithms. These combinations of subsets are given a selection count for priority. Features with selection count less than 6 lead to turn down the accuracy of the classifier, and degrades the performance of the intrusion detection system. For the justification of the aforementioned statement we have implemented two following Scenarios.

**Table 3:** Feature selection details for six search algorithms under CFS with Selection count

| Label | Feature | Best First | GreedyStepwise | Genetic Search | Scatter search V1 | Exhaustive Search | Random Search | Total(SC) |
|---|---|---|---|---|---|---|---|---|
| 1 | Duration | | | | | | | 0 |
| 2 | Protocol-type | √ | √ | √ | √ | √ | √ | 6 |
| 3 | Service | √ | √ | √ | √ | √ | √ | 6 |
| 4 | Flag | √ | √ | √ | √ | √ | √ | 6 |
| 5 | src_bytes | √ | √ | √ | √ | √ | √ | 6 |
| 6 | dst_bytes | √ | √ | √ | √ | √ | √ | 6 |
| 7 | Land | | | | | | | 0 |
| 8 | wrong_fragment | √ | √ | √ | √ | √ | √ | 6 |
| 9 | Urgent | | | | | | | 0 |
| 10 | Hot | √ | √ | | | √ | √ | 4 |
| 11 | num_failed_logins | | | | √ | | √ | 2 |
| 12 | logged_in | √ | √ | √ | √ | √ | | 5 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 13 | num_comromised | | | √ | | | | 1 |
| 14 | root_shell | | | | √ | √ | | 2 |
| 15 | su_attempted | | | | | | | 0 |
| 16 | num_root | | | | | | | 0 |
| 17 | num_file_creation | | | | | | | 0 |
| 18 | num_shells | | | | | | | 0 |
| 19 | num_access_files | | | | | | | 0 |
| 20 | num_outbound_cmds | | | | | | | 0 |
| 21 | is_host_login | | | | | | | 0 |
| 22 | is_guest_login | | | √ | | | | 1 |
| 23 | Count | √ | √ | √ | √ | √ | √ | 6 |
| 24 | srv_count | | | | | | | 0 |
| 25 | serror_rate | √ | | √ | √ | √ | √ | 5 |
| 26 | srv_serror_rate | | √ | √ | | | √ | 3 |
| 27 | rerror_rate | | | √ | | | √ | 2 |
| 28 | srv_rerror_rate | | | | | | | 0 |
| 29 | same_srv_rate | √ | √ | √ | √ | √ | √ | 6 |
| 30 | diff_srv_rate | √ | √ | √ | √ | √ | √ | 6 |
| 31 | srv_diff_host_rate | | | √ | | √ | | 2 |
| 32 | dst_host_count | | | √ | | | | 1 |
| 33 | dst_host_srv_count | | | √ | | | | 1 |
| 34 | dst_host_same_srv_rate | | | | | | | 0 |
| 35 | dst_host_diff_srv_rate | √ | √ | | √ | √ | | 4 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 36 | dst_host_same_src_port_rate | √ | √ | √ | √ | √ | √ | 6 |
| 37 | dst_host_srv_diff_host_rate | √ | √ | √ | √ | √ | √ | 6 |
| 38 | dst_host_serror_rate | √ | √ | √ | √ | √ | √ | 6 |
| 39 | dst_host_srv_serror_rate | | | | | | | 0 |
| 40 | dst_host_rerror_rate | √ | √ | | √ | √ | | 4 |
| 41 | dst_host_srv_rerror_rate | | | | | | | 0 |
| Total | | 17 | 17 | 21 | 18 | 19 | 17 | |

In the first scenario subset containing 12 features [18] is compared using naïve Bayes classifier by sequentially adding to it other subsets of features having count 5,4,3,2 and 1 one at a time. In the second scenario subsets having a selection count of 5,4,3,2,1 and 0 are individually crosschecked against the complete set of features and subset of 12 features. The outcomes of the two aforementioned scenarios shows that all the subsets having a selection count less than six are enhancing the accuracy of the classifier as compared to the complete set of features but not to extent comparable with12 subset of features having a maximum count of six. This minimal subset of 12 features is giving the highest classifiers accuracy, smallest computational time to build the model and good merit of subset selected.

## EMPERICAL OUTCOMES AND DISCUSSION

The detailed outcomes of classifier accuracy, Root mean square error (RMSE), Kappa Statistics and time to build the model for various combinations of Subsets of selection count less than 6 in table:4 in columns 3,4,5 and 6 respectively. As discussed earlier also selection count is that how many times a feature is selected among six search algorithms shown in table: 3, this is also called as selection frequency. The significance of selection count is, the more number of times a feature is selected the more is its involvement in the detection process for an effective and efficient IDS functioning.

**Scenario I**

In table: 4 we have shown the empirical outcomes of the subset with maximum SC:6 and sequentially added other subsets with SC:5,4,3,2,1 one at a time and evaluated them on the basic metrics for Naïve Bayes classifier. A comparative study is done on the 41 attributes with other subsets formed with combinations in table: 4 below.

- It is observed that the subset with selection count 6 reveals the highest classifier accuracy.
- Lowest Root Mean Square Error.
- Best Kappa Statistics.
- Least computational time to build the model.
- It is also a keen observation that there is a declining pattern in all the metrics for Naïve Bayes classifier as we sequentially add-up the subsets one at a time with lower selection count.
- The results justifies that the subets with selection count lower than 6 have less contribution towards the detection of attacks i.e. the contribution graph declines sharply as the SC declines.

**Table 4:** Feature Subset **C**ombinations Evaluation for Naïve Bayes Classifier, Scenario I

| S.No | Selection Frequency | Accuracy | RMSE | Kappa statistic 0.5735 | Time to build Model(Sec) |
|---|---|---|---|---|---|
| 1. | 41 Attribute | 33.91 | 0.24 | 0.221 | 3.9 |
| 2. | Selection Count:6 | 65.43 | 0.168 | 0.608 | 0.09 |
| 3. | Selection Count:6,5 | 62.81 | 0.172 | 0.573 | 0.13 |
| 4. | Selection Count:6,5,4 | 61.74 | 0.175 | 0.555 | 0.14 |
| 5. | Selection Count:6,5,4,3 | 60.59 | 0.182 | 0.533 | 0.14 |
| 6. | Selection Count:6,5,4,3,2 | 40.86 | 0.216 | 0.279 | 0.19 |
| 7 | Selection Count:6,5,4,3,2,1 | 50.23 | 0.194 | 0.36 | 0.25 |

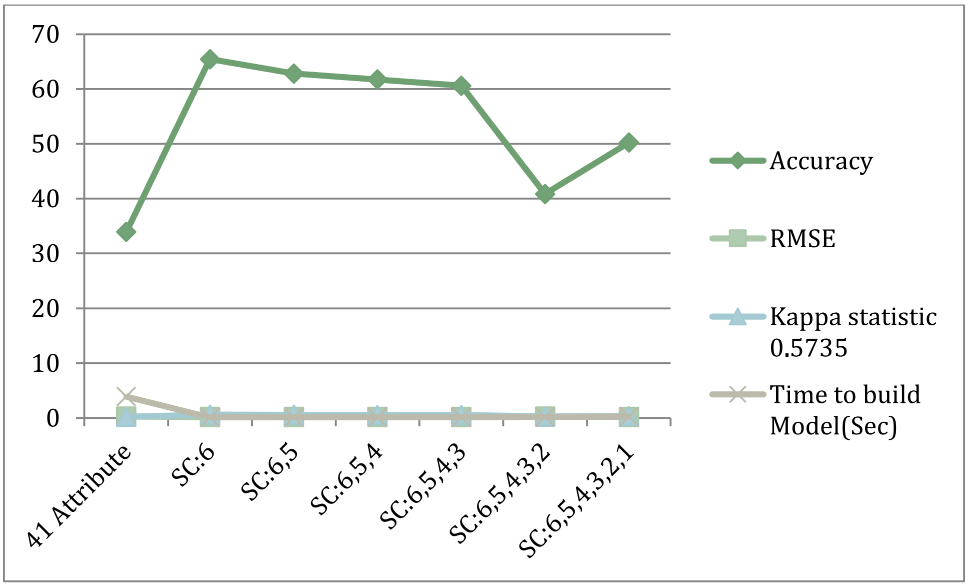

**Figure 2:** Graphical representation for Sequential combinational SC subsets

**Scenario II**

In table: 5 we have represented the comparative analysis of individual subsets with selection count 6,5,4,3,2,1 and 0 with 41 attributes. The metrics on which the results are evaluated are the same i.e. accuracy, RMSE, Kappa statistics and time to build the model in column 3,4,5,and 6. In this scenario also the comparison between 41 attributes and the individual subsets with selection count 6,5,4,3,2,1 are shown. The outcome reveals that:

- The accuracy with selection count 6 is maximum.
- Except the subset with count 4 shows accuracy greater than the subset with count 5 else it again forms a declining accuracy pattern from higher to lower.

**Table 5:** Feature Subset **C**ombinations Evaluation for Naïve Bayes Classifier, Scenario II

| S.No | Selection Frequency | Accuracy | RMSE | Kappa statistic | Time to build Model(Sec) |
|---|---|---|---|---|---|
| 1. | 41 Attributes | 33.91 | 0.24 | 0.221 | 3.9 |
| 2. | Selection Count 6 | 65.43 | 0.168 | 0.608 | 0.09 |
| 3. | Selection Count 5 | 50.03 | 0.171 | 0.27 | 0.2 |

| | | | | | |
|---|---|---|---|---|---|
| 4. | Selection Count 4 | 60.21 | 0.187 | 0.133 | 0.2 |
| 5. | Selection Count 3 | 58.06 | 0.155 | 0.41 | 0.1 |
| 6. | Selection Count 2 | 55.34 | 0.201 | 0.295 | 0.12 |
| 7. | Selection Count 1 | 16.12 | 0.228 | 0.066 | 0.08 |
| 8. | Selection Count 0 | 4.3 | 0.268 | 0.0098 | 0.13 |

- It is also a key observation that the features that belong to content categories are not selected at all i.e. they are part of feature subset with Selection Count 0. They have no involvement in detection process.
- For subset with SC:6 have the minimum computational time to build the model, highest accuracy, and least RMSE.

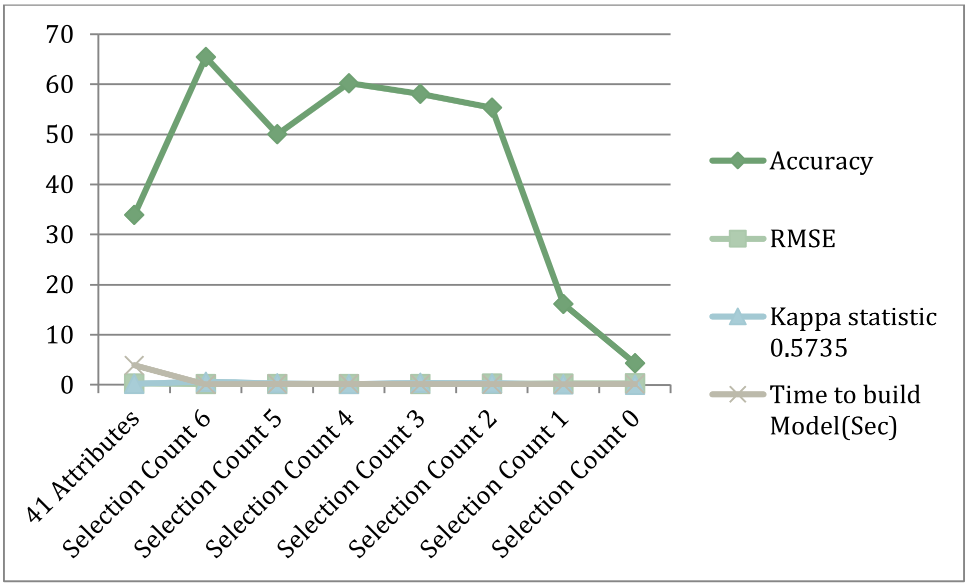

**Figure 3:** Graphical representation of Individual SC subsets

The overall conclusion from both the scenario is that the proposed subset of 12 features[18] with a selection count of 6 that is selected maximum number of times among six search algorithms. They turn out to be the most significant set of features in detecting the predictive classes to build effective and efficient IDS. As the selection count goes down the involvement of the respective subset decreases. The feature subset with selection count 0 shows negligible contribution in detecting the attack classes.

## CONCLUSION

The work presented here implemented correlation based feature selection techniques for reducing the dimensionality of NSL-KDDCup 99 dataset. Subsets of 12 features have been extracted among six search algorithms for feature selection. On the basis of selection frequency among these six search algorithms, the subset of 12 features is allocated with a selection count of 6. The 12 features show a significant enhancement in classifiers output as compared to a complete set of 41 attributes. The remaining subsets having a selection count less than 6 are also evaluated using two different scenarios against 12 proposed subset and complete set of features. It is clearly justified that these 12 feature subset is the most promising subset of features, providing their maximum contribution for distinguishing the normal and any anonymous activity taking place in the network. Moreover these 12 features take less computational time to build the IDS model detecting wider range of attacks for the network administrator.

## REFERENCES

[1] Niemelä, A. N. T. T. I. (2011). Traffic analysis for intrusion detection in telecommunications networks. *Master of Science Thesis, TAMPERE UNIVERSITY OF TECHNOLOGY, Finland*.

[2] J. McHugh. (2001). Intrusion and Intrusion detection. *International Journal of Information Security.vol. 1no. 1 pp.14-35.*

[3] k. Scarfone and P.Mell. (2007). *Guide to intrusion detection and prevention system (idps). National Institute of Strandads and technology (NSIT). Tech. Rep.2007.*

[4] Javidi, M. M., Rafsanjani, M. K., Hashemi, S., & Sohrabi, M. (2012). An overview of anomaly based database intrusion detection systems. *Indian Journal of Science and Technology*, *5*(10), 3550-3559.

[5] Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). A detailed analysis of the KDD CUP 99 data set. In *Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications 2009*.

[6] Lee, W., & Stolfo, S. J. (1998, January). Data mining approaches for intrusion detection. In *Usenix security*.

[7] Witten, I. H., & Frank, E. ( 2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

[8] Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques*. Elsevier.

[9] Garcia-Teodoro P, Diaz-Verdejo J, Maciá-Fernández G, Vázquez E. Anomaly-based network intrusion detection: Techniques, systems and challenges. *Computers and security*. 2009 Mar 31; 28(1):18–28.

[10] KDD Cup 1999 dataset [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, (Accessed 2014, Oct.).

[11] Nsl-kdd data set for network-based intrusion detection systems [Online]. Available on: http://nsl.cs.unb.ca/KDD/NSL-KDD.html, (Accessed March 2014). KDD Cup 1999 dataset [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, (Accessed 2014, Oct.)

[12] Revathi, S., & Malathi. (2013-December) A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection. In *International Journal of Engineering Research and Technology*(Vol. 2, No. 12 (December-2013)). ESRSA Publications.

[13] Singh, J., & Nene, M. J. (2013). A survey on machine learning techniques for intrusion detection systems. *International Journal of Advanced Research in Computer and Communication Engineering*, *2*(11).

[14] Aggarwal, C. C., & Reddy, C. K. (Eds.). (2014). Data classification: algorithms and applications. CRC Press.

[15] Sabhnani M, Serpen G. Application of Machine Learning Algorithms to KDD Intrusion Detection Dataset within Misuse Detection Context. MLMTA; 2003 Jun.p. 209–215.

[16] Olusola, A. A., Oladele, A. S., & Abosede, D. O. (2010, October). Analysis of KDD'99 Intrusion detection dataset for selection of relevance features. In Proceedings of the World Congress on Engineering and Computer Science (Vol. 1, pp. 20-22).

[17] Kayacik, H. G., Zincir-Heywood, A. N., & Heywood, M. I. (2005, October). Selecting features for intrusion detection: A feature relevance analysis on KDD 99 intrusion detection datasets. In *Proceedings of the third annual conference on privacy, security and trust*.

[18] Bahl, S., & Dahiya, D. (2016). Enhanced Intrusion Detection System for Detecting Rare Class Attacks using Correlation based Dimensionality Reduction Technique. Indian Journal of Science and Technology, 9(11).

[19] Chizi B, Maimon O. Dimension reduction and feature selection. Data Mining

and Knowledge Discovery Handbook. Springer US; 2005 Jan 1.p. 93–111.

[20] Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *Knowledge and Data Engineering, IEEE Transactions on*, *17*(4), 491-502.

[21] Bahl, S., & Sharma, S. K. ( 2015, February). Improving Classification Accuracy of Intrusion Detection System Using Feature Subset Selection. In*Advanced Computing & Communication Technologies (ACCT), 2015 Fifth International Conference on* (pp. 431-436). IEEE.

[22] Liu, H., & Motoda, H. (Eds.). (2007). *Computational methods of feature selection*. CRC Press.

[23] Nguyen, H. A., & Choi, D. (2008, October). Application of data mining to network intrusion detection: classifier selection model. In Asia-Pacific Network Operations and Management Symposium (pp. 399-408). Springer Berlin Heidelberg.

[24] Srilatha Chebrolu, Ajith Abraham, and Johnson P. Thomas"Hybrid Feature Selection for Modeling Intrusion Detection Systems "Springer, 2004,pp 1020-1025.

[25] Sung, A.H., Mukkamala, "Feature Selection for Intrusion Detection using Neural Networks and Support Vector In the earlier work Sung and Mukkamala Machines", Journal of the Transportation Research Board, 2003.

[26] Zaman, S., & Karray, F. (2009, January). Features selection for intrusion detection systems based on support vector machines. In 2009 6th IEEE Consumer Communications and Networking Conference (pp. 1-8). IEEE.].

[27] (Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. Journal of machine learning research, 5(Oct), 1205-1224.)

[28] H. John, R. Kohavi, and K. Pfleger. Irrelevant feature and the subset selection problem. In Proceedings of the Eleventh International Conference on Machine Learning, pages 121–129, 1994.

[29] Piramuthu,S.(2004). Evaluating feature selection methods for learning in data mining applications. European journal of operational research, 156(2), 483-494.

[30] Weka Data Mining Machine Learning Software .http://www.cs.waikato.ac.nz/ml/weka.

[31] Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. Data Classification: Algorithms and Applications, 37.

[32] Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research*, *5*, 1205-1224.

[33] Hall, M. A. (1999).*Correlation-based feature selection for machine learning*(Doctoral dissertation, The University of Waikato)