CHAPTER – 6
Feature Reduction and N-Fold Validation

6.1 Introduction

Objective of this chapter is to address feature reduction and N-Fold Validation techniques. Need of the Feature reduction technique, various feature reduction techniques used by the researchers on KDD CUP 1999 dataset are also addressed at middle part of this chapter. At the end, the feature selection and N-fold validation model used in the proposed model is discussed in details.

6.2 N-Fold Validation

![Diagram of N-Fold Validation]

**FIGURE 6.1:**
N- Fold Validation [19]

As per authors of [19], validation is used for the following reasons:
- Model Selection
- Variable Selection
- Variable Optimization
- Performance Estimation

The proposed model uses N-fold validation for the variable selection and performance estimation. In N-fold validation, the model is executed for N rounds. For each round, as per Fig.6.1, the entire dataset is divided in two subsets: training subset and testing subset.
The model is trained and tested by training and testing subsets and performance statistics are recorded. Average performances of N rounds of different models are compared and the best model is selected.

### 6.3 Feature Reduction

Feature reduction or selection is a method which improves performance and reduces processing time and complexity of the model [1] [7] [10] [11]. But selecting the features which has the best discrimination ability between the classes is very challenging task [6] [7]. There are two approaches for feature reduction or selections: Filter and Wrapper [8]. Wrapper approach is based on the performance of the learning algorithm, while Filter approach evaluates features according to statistical characteristics of the data. As compared to the Filter approach, Wrapper approach is generally considered to generate better feature subsets, but runs much slowly and requires more computing resource [9].

### 6.4 Need for Feature Reduction in the Proposed Model

For the detection of anomalies, the proposed model takes real time network traffic as input which contains very large number of features. All of these features might not be useful for IDS. IDS extracts specific set of the features from real-time network traffic. Selection of such specific set of features depends upon the type of behavior layer and the type of attacks which suppose to be detected. For example, if IDS uses connection based behavior layer to detect attacks as mentioned in [4] then 41 features is used to build behavior layer. These selected features can still be reduced by applying various feature reduction techniques. Such reduced set of features improves performance and reduces processing time as well as complexity of the model [1] [2] [3] [4]. However, while in the selection of feature reduction technique, topmost care should be taken so that it does not reduce the accuracy.

For the experiments related to reduction of response time, the proposed model uses KDD CUP 1999 dataset which has 41 features while for the experiments related to multi class attack detection, the proposed model uses expert dataset which has 3 features. As there are very less number of features in the experiments related to the multi class attack detection, application of feature reduction technique will have very marginal impact or no impact on performance of the model. However, for the experiments related to reduction of response time where number of features is high, application of feature reduction technique may have visible impact.
6.5 Literature Review

For reducing the features of NSL KDD CUP dataset, authors of [7] used Bayesian Network as a feature selection technique. They used Weka as a tool and compared their feature reduction technique with IG [14], GR [15], ReliefF [16] [17] and ChiSquare [18]. Their own technique was able to reduce features from 41 to 11 as compared to 20, 20, 20 and 20 reduced features of IG, GR, ReliefF, and ChiSquare respectively. Their technique reduces the complexity, model building time and improves the accuracy.

Similarly, Tesfahun et al. in [1] used wrapper approach and reduced features of KDD CUP 1999 dataset from 41 to 22. Authors used Information Gain (IG) as a feature reduction technique, Random Forest as a classifier and Weka as a tool. Authors were able to improve the classification rate of minority classes. Authors of [1] had selected features with number 1, 3, 4, 5, 6, 10, 12, 14, 23, 24, 25, 26, 29, 30, 32, 33, 34, 35, 36, 37, 38 and 39.

Authors of [13] had used Information Gain as feature reduction technique, centroid-based and SVM based machine learning algorithm as a classification technique, KDD CUP 1999 as a dataset, and Weka as a tool. LibSVM was used as a support vector machine which was embedded in Weka. Authors were able to reduce the features from 41 to 10 with improved accuracy, precision and recall values also.

6.6 Selection of Feature Reduction Technique (Tesfahun et al.)

During the literature review, we came across various feature reduction techniques which were applied on KDD CUP 1999 dataset. These techniques reduce features from 41 into the range of 10 to 22. Some of the authors who had implemented these techniques had claimed for significant improvement in accuracy. During the literature review, we came to know that all of these authors had used data mining based techniques. They had not tested their techniques on the real-time network traffic. It might be possible that these techniques might perform well during testing on dataset but may perform poor during processing the real-time network traffic. Conversely, KDD CUP 1999 dataset is designed by the domain experts at MIT Lincon Lab [12]. The dataset was generated by creating the environment similar to US Air force [12]. For such dataset, by reducing more features, internal structure of the dataset can be broken. Classifier which is well trained on such tiny dataset might perform well during testing but poor during processing the real-time network traffic. Hence, to take the benefit and to minimize the limitations of feature reduction technique on
KDD CUP 1999 dataset, we had selected the technique presented in [1] which uses 22 features.

### 6.7 The Model for Feature Reduction and N-Fold Validation

To check effectiveness and feasibility of Tesfahun et al.’s feature reduction technique [1] on BPNN classifier, we had proposed a model which is shown in Fig. 6.2. The model has KDD CUP Dataset, Feature Selection, Normalization, N – Fold Validation, Classification (BPNN), Testing Evaluation and Performance Measures as sub modules.

![Feature Reduction Model for KDD CUP 1999 Dataset](image)

**FIGURE 6.2:**
Feature Reduction Model for KDD CUP 1999 Dataset

#### 6.7.1 KDD CUP 1999 Dataset

To train and test Network Based Intrusion Detection System, majority of the researchers use the KDD CUP 1999 dataset [23]. This dataset was generated by MIT Lincoln Lab by
creating the environment similar to the US Air force. As the KDD CUP 1999 dataset is generated by domain experts and is publicly available, the model uses it for training, validation and testing BPNN classifier. More details about this dataset are available in Chapter 5.

6.7.2 Feature Reduction

To reduce the number of the features of KDD CUP 1999 dataset, the model uses feature selection technique presented in [1]. Authors of [1] used Information Gain (IG) as a feature selection technique. To use IG as a feature selection, an entropy value of each attribute of the dataset has to be calculated. The entropy value is used for ranking the features that affect data classification. A feature which does not have much effect on the data classification, have very small information gain and can be ignored without affecting the detection accuracy of a classifier [1] [18]. By IG feature selection technique, authors of [1] had selected features with number 1, 3, 4, 5, 6, 10, 12, 14, 23, 24, 25, 26, 29, 30, 32, 33, 34, 35, 36, 37, 38 and 39. As the proposed model follows the feature selection technique of [1], we used above features and reduces the KDD CUP1999 dataset for further processing.

6.7.3 Normalization

Model uses KDD CUP 1999 dataset for training, N-fold validation and testing the BPNN classifier. The attributes of KDD CUP 1999 dataset are neither in numeric form nor properly scaled. If this dataset without normalization are applied to BPNN classifier, then processing time will be high, which is not acceptable for any real time anomaly detection system. To reduce the processing time, normalization module which is already developed in our previous work of [21] is used in the proposed model. It has encoding, scaling and size reduction sub modules. More detail about normalization is also available in Chapter - 7.

6.7.4 N-Fold Validation

The proposed model uses N-fold validation for the variable selection and performance estimation. The model uses 10-Fold cross validation in which the original dataset is divided into 10 subsets. Each time, one of the 10 subsets is used as the test set and the other 9 subsets are used as training subsets [20]. During each run, performance measures like accuracy, precision, recall and Fscore are calculated. On the basis of this performance measures, the variable selection and performance estimation can be done.
6.7.5 Classification (BPNN)
To classify the given network data into the five classes (Normal, DOS, Prob, U2R and R2L) is multiclass linear or nonlinear classification problem. The proposed model uses BPNN as a classifier. Reasons for the selection of BPNN as a classifier have been discussed in chapter related to BPNN. For optimum performance of BPNN, various parameters like Learning Rate, Initial Weights, Number of Hidden Units, Overtraining and Early Stopping Criteria, Number of Learning Samples, Activation function had been set as per our previous work [22].

6.7.6 Testing Evaluation and Performance Measures
The model for feature reduction and N-fold validation uses accuracy, precision, recall, and Fscore as performance measures. These performance measures are calculated using confusion matrix which is shown in Table-6.1. Formulas for calculating these performance measures are given in (6.1). These performance measures are recorded and compared during N-Fold Validation and Testing. On the basis of these performance measures, both the models (Model with reduced dataset and Model with full dataset) are compared. Comparison of N-fold validation shows which model has better generalization ability. To be more accurate, both the models are tested on testing dataset. The model which outer performs in validation and testing might have better generalization ability in real-time traffic. Such model has low false alarm rate.

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attack</td>
<td>Normal</td>
</tr>
<tr>
<td>True Positive (TP)</td>
<td>False Negative (FN)</td>
</tr>
<tr>
<td>False Positive (FP)</td>
<td>True Negative (TN)</td>
</tr>
</tbody>
</table>

As per [14], accuracy, precision, recall and Fscore can be calculated as follows:

Accuracy= \( \frac{(TP+TN)}{(TP+TN+FP+FN)} \)

Precision= \( \frac{TP}{(TP+FP)} \)

Recall= \( \frac{TP}{(TP+FN)} \)

Fscore= \( \frac{(2*TP)}{((2*TP)+(FP+FN))} \) ................................. (6.1)
6.8 References


10.1109/TAI.1995.479783.