Name : Nushrat Jahan
ID : 2016-1-60-050

Our group project is about comparative analysis of different Machine Learning algorithms. So I selected two papers related to our topic so that they come as helpful to apply in our project in future.

First paper : Feature Selection and Intrusion Detection in Cloud Environment based on Machine Learning Algorithms

The first paper is about feature selection and intrusion detection in Cloud environment using Machine Learning algorithms. It provided a new method based on intrusion detection systems and its various architectures aimed at increasing the accuracy of intrusion detection in cloud computing. Pearson linear correlation and mutual information these two techniques are used in this paper.

A simple pseudo code for feature selection used in this study is explained briefly below.
step1 : feature selection using linear correlation method .
step2 : feature selection using mutual information method .
step3 : application of classification methods including neural network random forest .
step4 : comparing the results and selecting the best results.
step5 : comparing the results with the use of feature selection and without using it.

Neural networks, CART algorithm, ID3Decision tree algorithm, Random forest algorithm are applied to pre-processed data and feature selection and the evaluation criteria include the criteria of accuracy, recall and result precision. KDD99 database is used for solving the problem for the study which contains 42 features that represent the last attribute of the data class. Pearson correlation property selection method is required as a feature selection method to find the criteria for removing plug in features.
Additional features are found by applying the mutual information feature selection algorithm which can have a negative effect on the accuracy of the classifier. It is shown comparison of numerical accuracy of obtained accuracy in classification methods without feature selection, proposed method and the evaluation criteria derived from methods applied by feature selection with percentage of random forest, decision tree, neural network and CART.
It is observed that the accuracy of the applied methods on data is improved in the case when the feature selection method is used and the method of classification of the neural network using the proposed method introduced in the paper has a higher accuracy than other methods. One of the suggestion was to use data balancing techniques for continuing the work.

The paper have proper procedure of using various machine learning algorithms, which will help us to decide which algorithms we might work on. But the authors suggested to use data balancing techniques which sometimes cause high accuracy rate with imbalanced data without fetching errors. So estimating the comparison of various algorithms might get difficult for us.

Second Paper : On Efficiency Enhancement of the Correlation based Feature Selection for Intrusion Detection Systems

The authors of the second paper proposed the reduction of dimensionality by an efficient feature selection algorithm that considers the correlation between a subset of features and the behavior class label. Correlation-based feature selection(CFS) and symmetrical uncertainty(SU) are the two correlation metrics used to measure the dependency level between features and class labels.

NSL-KDD dataset is used where fewer features significantly outperforms the existing schemes in terms of training time. It is derived from KDDCUP 99 dataset with a fewer number of instances. There are 1,25,973 instances in the NSL-KDD training dataset while the testing set has 82,332 records.

A pseudo code is given to explain the feature selection algorithm followed in the study.
Step 1 : subsets of features are built with a forward selection technique that starts with an empty subset and adds features one by one.
Step 2 : CFS is calculated for every subset and is compared to the rest. The features with highest CFS will be selected.
Step3 : SU is calculated for every two features to remove those features that are not good fit for the subset of features.
Step4 : backward elimination technique is used to remove features with least SU. This elimination will continue until with elimination of more features results in significant decrease in accuracy. Accuracy is calculated with the same training set.
Step5 : the iteration of feature removal will stop when there can be more than one feature found that with its removal the accuracy decrease equally. In this way, 4 best discriminative features were selected from 42 features.

The work was compared with six other filter-based feature selection technique such as information gain, gain ratio, chi-square and CFS. The performance of each scheme is represented in terms of miss rate, false alarm rate, precision and accuracy and random forest, J48, PART and C4.5 classifiers are used in the study. The metrics that are used for evaluation in this study are detection rate(DR), accuracy(Acc), precision(Pr), false alarm rate(FAR) and miss rate(MR). In addition, Received Operating Characteristics of the system is obtained in this paper. The area under the ROC curves is calculated for different classifiers in tables. The training time for several feature selection techniques (CFS, chi-squared, IG and GR) are compared with random forest classifiers. J48 classifier gives the highest accuracy and precision and the lowest miss and false alarm rate is considered to be the best classifier for the proposed feature selection technique.

The described paper had no suggestions but they compared on six different feature selection technique which we might use in our work according to our imparted knowledge. Moreover, proposed feature selection technique is tested on different classification algorithm. CFS and SU are used as the feature selection schemes. The most informative subset is selected which can be based on distance, divergence, consistency, classification or dependency.
In this paper, dependency is used for measure of relevance. Best feature subset is detected through a combination of both mutual information and Pearson correlation coefficient. CFS is used to keep relevant features. It increases due to their low correlation which might select more number of features than required. So non-linear correlation could be considered.
Symmetrical uncertainty is used to remove the features that do not collaborate with other features in the selected subset. Here mutual information is chosen over other statistical correlation techniques. It might be biased toward the features having greater values so SU overcomes this by dividing MI by the sum of the entropy of the features. These information about the algorithms are really helpful for us to work on the betterment of our project.

Therefore, the above brief discussion of the background work of two different papers are stated to take suggestions and proper help for the group project.