Computing, Information Systems, Development Informatics & Allied Research Journal
Vol. 7 No. 4, December, 2016 - www.cisdijournal.net

Heterogeneous Ensemble Methods Based On Filter Feature Selection

Ameen A. O., Balogun A. O. & Usman G.,
Department of Computer Science
University of Ilorin
Ilorin, Nigeria.
ahmedameeny2k4@yahoo.com
bharlow058@gmail.com
ganmuslimat@yahoo.co.uk

Fashoto S. G.
Department of Decision Sciences
University of South Africa
Pretoria, South Africa
gbengafash@yahoo.com

ABSTRACT

While certain computationally expensive novel methods can construct predictive models with high accuracy from high-dimensional data, it is still of interest in many applications to reduce the dimension of the original data prior to any modeling of the data. Hence, this research presents a précis of ensemble methods (Stacking, Voting and Multischeme) and Multilayer perceptron, K Nearest Neighbour and NBTree with a framework on the performance measurement of base classifiers and ensemble methods with and without feature selection techniques (Principal Component Analysis, Information Gain Attribute Selection and Gain Ratio Attribute Selection). The enhancement is based on performing feature selection on dataset prior to classification. The notion of this study is to evaluate the performances of the ensemble methods on original and reduced datasets. A 10-fold cross validation technique is used for the performance evaluation of the ensemble methods and base classifiers (Root to Local) R2L KDD cup 1999 dataset and UCI Vote dataset using Waikato environment for knowledge analysis (WEKA) tool. The experiment revealed that the reduced dataset yielded improved results than the full dataset after using the ensemble methods based on stacking, voting and multischeme. On the R2L dataset, Multischeme ensemble method gave accuracy of 98.76% with PCA as feature selection on R2L dataset while 98.58% accuracy was given without feature selection. Using the gain ratio attribute selection, the Multischeme gave 98.93% accuracy over 98.76% without feature selection while using information gain attribute selection gave accuracy 98.85% over 98.76% without feature selection. For the Vote Dataset, Multischeme ensemble method proved best with an accuracy of 92.18% with PCA feature selection over 89.88% without feature selection, 95.40% accuracy with information gain as feature selection over 93.10% without feature selection and 95.40% accuracy with gain ratio as feature selection over 93.10% without feature selection. Inarguably, it can be concluded that ensemble methods works well with feature selection.

Keyword: Data mining, Classification techniques, Feature selection techniques, Ensemble methods.

CISDI Journal Reference Format
Ameen A. O., Balogun A. O., Usman G. & Fashoto, S.G. (2016): Heterogenous Ensemble Methods Based On Filter Feature Selection. Computing, Information Systems, Development Informatics & Allied Research Journal. Vol 7 No 4. Pp 63-78.
Available online at www.cisdijournal.net

1. INTRODUCTION

Data mining is a vital concept that all organization need for their daily work to discover the hidden and undiscoverable data that needed to be processed for immediate purpose. The idea is to build computer programs that sift through databases automatically, seeking regularities or patterns. Strong patterns, if found, will likely generalize to make accurate predictions on future data. In many of our organization such as Banking sector, financial sector, Communication, Retail, Marketing organization, Educational sector etc. due to the large amount of data kept in their databases found it difficult in retrieving vital information which has been hidden (Valliappan, Sundresan, & Putra, 2015).

For useful information needed to be obtained from the database, there is need to separate the non-useful (noise) from the database. This process will further enhance the quality of useful information to be extracted and also the time taken to get the processing done (Balogun, Mabayoje, Arinze & Salihu, 2015). The aforementioned process explained is called feature selection. Feature selection, when applied to data, it selects relevant attributes or variable from databases based on some metric specified by the analyst. Feature selection is itself useful, but it mostly acts as a filter, muting out features that aren't useful in addition to the existing features. It helps to create an active predictive model and to also select features that give better accuracy whilst requiring less data (Balogun, Mabayoje, Arinze & Salihu, 2015).

A method or learner that combine multiple classifiers together which leverages the power of multiple models to achieve better prediction accuracy than any of the individual models could on their own is known as Ensemble method which can also be refers to as committee-based learning or multiple classifier learning (Zhou, 2012). Ensemble contains multiple learners called individual learners which are generated from training data set by a based leaning algorithms such as decision tree, genetic algorithm, and neural network etc. homogeneous base learners are generated from a single based learning algorithm while multiple learning algorithm are used to produce heterogeneous ensemble (Zhou, 2012)

In other to enhance the performance of the ensemble method, feature selection is needed to choose a subset of input variables by eliminating the feature with little or no predictive information. Feature selection has been proven effectively theoretically and practically in enhancing learning efficiency, increasing predictive accuracy and reducing complexity of learned results. Supervised learning in feature selection has a key objective of finding a feature subset that generates higher classification accuracy.

Various methods and algorithms have been either as a base classifier or a hybrid in other to get the best result (Balogun & Jimoh, 2015; Mabayoje, Akintola, Balogun & Ayilara, 2015; Ajayi, Idowu, & Anyaehie, 2013). Fortunately, some methods always perform better on data with low dimension than high dimension as irrelevant or redundant attributes often interfere with useful ones. Consequently, this has called for a method to remove the irrelevant information which is feature selection (Mahdi & Fazekas, 2011; Mabayoje, Balogun, Akintola & Ayilara, 2015). Feature selection is a technique of data mining which is used for selecting some features of a particular dataset for data mining process(es)(Mahdi & Fazekas, 2011; Balogun, Mabayoje, Salihu & Arinze, 2015). Feature selection has been deployed alongside some classification algorithms with high accuracy but there is still room for research on whether the removal of irrelevant or redundant attributes could assist classification algorithms in performing better since some studies reported classification process(es) with high accuracy and didn't perform feature selection on the dataset (Aneetha & Bose,2012; Ajayi, Idowu, & Anyaehie, 2013; Hashem, Muda, & Yassin, 2013). Hence, in this study, the performance evaluation of feature selection methods on the classification accuracy of heterogeneous ensemble method will be investigated using KDD'99 dataset and Vote date set from UCI.

This paper is organized into 5 sections; section 1 serves as introduction to the study. Section 2 provides an overview of data mining techniques, feature selection, ensemble methods and related literatures. Section 3 provides a brief overview of methodology used for this study. Section 4 discusses the proposed methodology and experimental results while conclusions and future recommendations derived were given in section 5.

2. DATA MINING

2.1 Data Mining Techniques

There are different Data mining techniques used in discovering knowledge from databases namely clustering, Classification, Regression and Association.

2.1.1 Classification

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier. In machine learning and statistics, **classification** is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. An example would be assigning a given email into "spam" or "non-spam" classes or assigning a diagnosis to a given patient as described by observed characteristics of the patient (gender, blood pressure, presence or absence of certain symptoms, etc. (Alpaydin & Ethem 2010).

Types of classification models:

1. Classification by decision tree induction
2. Bayesian Classification
3. Neural Networks
4. Support Vector Machines (SVM)
5. Classification Based on Associations

2.1.2 Clustering

Clustering is a division of data into groups of similar objects. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification. It models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. Clustering is the subject of active research in several fields such as statistics, pattern recognition, and machine learning (Wanner, 2004).

Types of clustering methods

1. Partitioning Methods
2. Hierarchical Agglomerative (divisive) methods
3. Density based methods
4. Distance-based methods
5. Model-based methods

2.1.3 Prediction

Regression technique can be adapted for prediction. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply predictable. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks too can create both classification and regression models.

Types of regression methods

- Linear Regression
- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

2.1.4 Association rule

An association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships. Support is an indication of how frequently he items appear in the data base. Confidence indicates the number of times the if/then statements have been found to be true. An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data while consequent is an item that is found in combination with the antecedent.

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value.

Types of association rule

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

2.3 Feature Selection

One of the factors which have been considered to improve the accuracy of ensemble is feature selection. Feature selection has been an active and fruitful field of research area in pattern recognition, machine learning, statistics and data mining communities. Feature selection is the process of selecting a subset of relevant features for use in model construction and also by way of selecting those features in the data that are most useful or most relevant for the problem you are working on. Feature selection also called variable selection or attribute selection. Feature selection methods can be used to identify and remove unneeded, irrelevant and redundant attributes from data that do not contribute to the accuracy of a predictive model or may in fact decrease the accuracy of the model. Feature selection has proven in both theory and practice to be effective in enhancing learning efficiency, increasing predictive accuracy and reducing complexity of learned results. The achievement of combining attribute subset evaluator with a search method is done by searching the space of attribute subsets, evaluating each one (Ramaswami & Bhaskaran, 2009).

The characteristics of the search method used are important with respect to the time efficiency of the feature selection methods. Search methods include BestFirst, Exhaustive, FCBF, Genetic, GreedyStepwise, Race, Random and Ranker. Ranker is more appropriate for attribute evaluation methods. Ranking of features determines the importance of any individual feature, neglecting their possible interactions. Ranking methods are based on statistics, information theory, or on some functions of classifier's outputs (Pitt & Nayak, 2007). Application of feature selection is done in ensemble in order to find the subsets of feature for the classifiers of the ensemble. It is use to reduce the redundancy of the features as well as to increase diversity of the classifiers of an ensemble (Santana et al., 2007). There are three categorizations of feature selection algorithms:

1. Filter methods
2. Wrapper methods
3. Embedded methods.

2.3.1 Filter Methods.

Filter methods carry out the feature selection process as a pre-processing step with no induction algorithm. The general characteristics of the training data are used to select features (for example, distances between classes or statistical dependencies). This method is faster than Wrapper approach and results in a better generalization because it acts independently of the induction algorithm. However, it tends to select subsets with a high number of features (even all the features) and so a threshold is required to choose a subset. Filter feature selection methods apply a statistical measure to assign a scoring to each feature. The features are ranked by the score and either selected to be kept or removed from the dataset. The methods are often univariate and consider the feature independently, or with regard to the dependent variable. Example of some filter methods include the Chi squared test, information gain and correlation coefficient scores (Sanchez-Marono & Alonso-Betanzos, 2007).

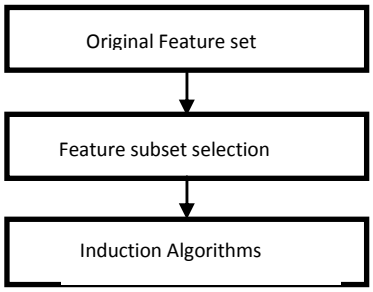

Figure 2.1: Filter methods for feature selection (Sanchez-Marono & Alonso-Betanzos, 2007).

2.3.2 Wrapper Methods

Wrapper methods consider the selection of a set of features as a search problem, where different combinations are prepared, evaluated and compared to other combinations. A predictive model us used to evaluate a combination of features and assign a score based on model accuracy. The search process may be methodical such as a best-first search, it may stochastic such as a random hill-climbing or it may use heuristics, like forward and backward passes to add and remove features. An example if a wrapper method is the recursive feature elimination algorithm. Wrapper methods search for an optimal feature subset tailored to a particular algorithms and a domain. The wrapper approach conducts a search in the space of possible parameters. A search required a state space, an initial state a termination condition and a search engine (Kohavi & Jorn, 2011).

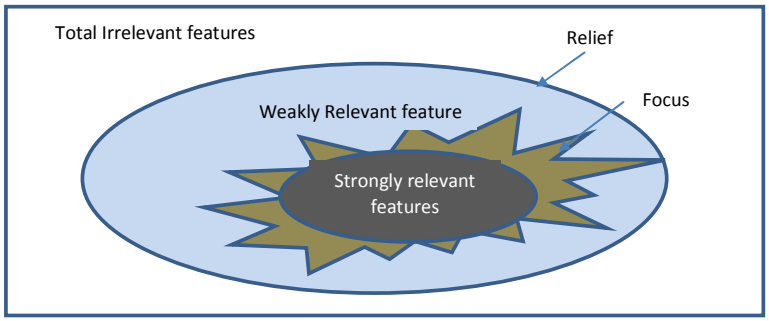

Figure 2.2: A view of relevant features (Kohavi & Jorn, 2011).

2.3.3 Embedded Methods

Embedded methods learn which features best contribute to the accuracy of the model while the model is being created. The most common type of embedded feature selection methods are regularization methods. Regularization methods are also called penalization methods that introduce additional constraints into the optimization of a predictive algorithm (such as a regression algorithm) that bias the model towards lower complexity (less coefficient). Example of regularization algorithms are LASSO, Electric Net and Ridge Regression. They also perform variable selection as part of the learning procedure and are usually specific to given learning machines. Examples are classification trees, random forests, and methods based on regularization techniques (e.g. lasso). In embedded methods the learning part and the feature selection part cannot be separated the structure of the class of functions under consideration plays a crucial role (Lai et al, 2014).

2.4 Ensemble Method

Classifier ensembles are known to be very useful methods for improving the classification accuracy as well as diversity. They combine multiple classifiers together to get a single stronger one whose performance is more precise and accurate as compared to its individual members. A variety of factors have been considered in literature to improve the accuracy of the ensemble. These include classifier selection, feature selection, diversity creation in ensemble of classifiers and combination methods. The importance of Ensemble methods is to greatly combining the output of individual classifiers which have been immensely successful in producing accurate predictions for many complex classification tasks. The major goal of these methods is attributed to their ability to both consolidate accurate predictions and correct errors across many diverse base classifiers. Diversity is key to ensemble performance: If there is complete consensus the ensemble cannot outperform the best base classifier, yet an ensemble lacking any consensus is unlikely to perform well due to weak base classifiers. Successful ensemble methods establish a balance between the diversity and accuracy of the ensemble.

However, utilize a single type of base classifier to build the ensemble such homogeneous ensembles may not be the best choice for problems where the ideal base classifier is unclear. A good way is by building ensemble from the predictions of a wide variety of heterogeneous base classifiers such as support vector machines, neural networks, and decision trees. There are two major well known heterogeneous ensemble methods such as stacking which is known as meta-learning, and other one is ensemble selection. Stacking constructs a higher-level predictive model over the predictions of base classifiers, while ensemble selection uses an incremental strategy to select base predictors for the ensemble while balancing diversity and performance. Due to their ability to utilize heterogeneous base classifiers, these approaches have superior performance across several application domains (Whalen & Pandey , 2013). Some of the Ensemble Techniques are Boosting, Bagging and Stacking

2.4.1 Bagging Methods

Bagging is one of the Meta learning algorithm use in combining several machine learning techniques into one predictive model in order to decrease the variable. Bagging also (stands for Bootstrap Aggregation) is use in decreasing variance of prediction by generating additional data for training from the original datasets using combination with repetition to produce multisets of the same cardinality/size as the original data. By increasing the size of the training set the predictive model force can't be improve but can only decrease the variance, narrowly tuning the prediction to expected outcome.

2.4.2 Boosting Methods

Boosting is a machine learning ensemble meta-algorithm for primarily reducing bias, and also variance in supervised learning, and a family of machine learning algorithms which convert weak learners to strong ones. Boosting is based on the question : Can a set of **weak learners** create a single **strong learner**? A weak learner is defined to be a classifier which is only slightly correlated with the true classification (it can label examples better than random guessing). In contrast, a strong learner is a classifier that is arbitrarily well-correlated with the true classification. Boosting methods is a two-step approach in which the subsets of the original data is first use to produce a series of averagely performing models and then "boosts" their performance by combining them together using a particular cost function (=majority vote). Unlike bagging in the classical boosting the subset creation is not random and depends upon the performance of the previous model: every new subsets contains the elements that were (likely to be) mis-classified by previous model.

2.4.3 Stacking Methods

Stacking is similar to boosting because several models also apply to the original data. In stacking there is an introduction of metal-level and use another model/approach to estimate the input together with outputs of every model to estimate the weights or to determine what models perform well and what badly given these input data.

2.5 Related Literature

According to Kumar and Minz (2014), in their research, they gave an introduction into the concepts of feature relevance, general procedures, evaluation criteria, and the characteristics of feature selection. A comprehensive overview, categorization, and comparison of existing feature selection methods are also done, and the guidelines are also provided for user to select a feature selection algorithm without knowing the information of each algorithm. They described in details the general approach of feature selection methods such as filter, wrapper and embedded methods and their pseudo code is also presented also review the interesting facts regarding the advantages and disadvantages of feature selection which to handle the different characteristics of the real world application are also enumerated.

Tuarob et al. (2014) proposed the use of five heterogeneous features in combination with ensemble machine learning techniques to address the limitations posed by the traditional bag-of-word based methods and to discover health-related information, which could prove to be useful to multiple biomedical applications, especially those needing to discover health-related knowledge in large scale social media data. The analysis of the parameter sensitivity of the extraction algorithms were also done to obtain the best possible features from each feature type. Each base classifier is trained using standard ensemble methods. Combination of classifiers that learn different characteristics of the data is use to reduce the limitation of the N-gram features on the social media domain.

According to Tsoumakas, Katakis, & Vlanhavas (2004), they carried out an experiment and examine the use of classifier Evaluation and Selection (ES) method on combination of classification models to evaluate each of the models by using 10-fold cross-validation and select the best. The performance of this method in comparison with the oracle selecting the best classifier for the test set and show that 10-fold cross-validation has problems in detecting the best classifier then extend ES by applying a statistical test to the 10-fold accuracies of the models and combining through voting the most significant one. Effective Voting performs comparably with the state-of-the-art method of stacking with Multi-Response Model Tree.

Singh, Appavu, and Jebamalar (2016) analyzed several feature selection methods which observed that the feature ranking-based methods are better than the subset-based methods in terms of memory space and computational complexity and the ranking-based methods do not reduce the redundancy. Feature selection can be developed for high-dimensional data using filter approach with ranking method for selecting the significant features from the high-dimensional space.

While in Chandrashekar and Sahin (2013), they used Classifier accuracy and the number of reduced features to compare the feature selection techniques. They successfully used feature selection to improve predictor performance and for fault prediction analysis of Fault Mode data. Borji(2007) proposed a combining classification approach for intrusion detection with the uses of three combination strategies: majority voting, Bayesian averaging and a belief measure that produce output of four base classifiers ANN, SVM, KNN and decision trees in which the outcome support the superiority of the proposed approach compared with single classifiers for the problem intrusion detection. From all of the aforementioned which serve as motivation for this study, there need to make known the extent of feature selection on classification process. Inarguably, many researches has pointed to this process as a better process in data mining, that is, selecting and classifying only the features that are important to the task at hand.

3. RESEARCH METHODOLOGY

Our aim is to investigate whether selecting the best attributes will improve the performance of these ensemble methods. This section discusses the algorithms extensively and outlines the proposed system architecture involved in the analysis and evaluation of the proposed algorithms before and after feature selection is applied. As stated earlier, the data for analysis will be drawn from the UCI Vote and KDD'99 Remote-to-local (R2L) datasets

3.1 Feature Selection

Reducing dimensionality of a dataset is an essential step before any analysis of the data can be performed in many real world problems. The importance for reducing the dimension is the aim to preserve most of the relevant information of the original data according to some optimality criteria. Methods such as Principal component analysis (PCA), Info gain Attribute Evaluation (IGA) and Gain ratio Attribute Evaluation (GAE) are most used in general classification and pattern recognition which have been extensively used for this research as feature selection methods.

3.1.1 Principal Component Analysis

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. PCA combines similar (correlated) attributes and creates new ones. Superior to original attributes. PCA is a useful statistical technique that has found application in fields such as face recognition and image compression, and is a common technique for finding patterns in data of high dimension. Technically, a principal component can be defined as a linear combination of optimally-weighted observed variables(Smith, 2002).

PCA is used frequently in all forms of analysis - from neuroscience to computer graphics due to its simplicity, non-parametric method of extracting relevant information from confusing data sets. With minimal additional effort PCA gives ways of reducing a complex data set to a lower dimension in order to open hidden data sometimes and also to simplify dynamics that often underlie it. The purpose of PCA is to reduce the original data set of two or more sequentially observed variables by identifying a small number of "meaningful" components or modes (Fashoto et al., 2016). The method therefore permits the identification of Coherent structures as dominant or recurring patterns in a time series. To start, the data is organized in an m x n matrix $X^T(x,t)= [u_1 \ldots u_n]$, where the $u_i(x)$ vector describe the m pixels in an image observed at n different times. To be precise in this discussion, and without loss of generality, we can assume that the Data matrix has dimension $m \geq n$ with rank $r \leq n$. The data is then centered by subtracting the average from each image. The formal procedure consists in solving the eigenvalue problem for the two-point m x m covariance matrix

$$C(x , x') = \sum x^T(x,t)X(x',t). \qquad (1)$$

In general the number of eigenvalues will be the same as the rank $r \leq n$ of the data matrix. When dealing with linearly-independent observations we obtain r = n eigenvalues λ_1. Otherwise the number of eigenvalues is equal to the rank of the data matrix.

3.1.2 Information Gain Attribute Evaluation

Subsequent to preprocessing of data, the features of the data set are identified as either being significant to the intrusion detection process, or redundant. This process is known as feature selection. Redundant features are generally found to be closely correlated with one or more other features. As a result, omitting them from the intrusion detection process does not degrade classification accuracy. In fact, the accuracy may improve due to the resulting data reduction, and removal of noise and measurement errors associated with the omitted features. Therefore, choosing a good subset of features proves to be significant in improving the performance of the system.

They thus define Information Gain: In this method, the features are filtered to create the most prominent feature subset before the start of the learning process. Mathematically, it can be stated that information gain for a dataset (D) containing S_i tuples of class C_i for i = (1,…..,m) is defined as:

Information measures info required to classify any arbitrary tuple

$$\text{Info}(D) = -\sum \frac{S_i}{S} \log_2 (S_i/S) \qquad (1)$$

Where, S_i is the total value of a feature X and S is the number of possible value a feature X can take.

Entropy of feature X with values (x1, x2, …….,xv)

$$E(X) = \sum_{J=1}^{v} \frac{S1j + \ldots\ldots\ldots + Smj}{S} \text{Info}(D) \qquad (2)$$

Information gained by branching on feature X

$$\text{Gain}(X) = \text{Info}(D) - E(X) \qquad (3)$$

3.1.3 Gain Ratio

Gain ratio is a modification of the information gain that solves the issue of bias towards features with a larger set of values, exhibited by information gain.as Information Gain filters the features to create the most prominent feature subset before the start of the learning process. Gain ratio should be Large when data is evenly spread and small when all data belong to one branch attribute. Gain ratio takes number and size of branches into account when choosing an attribute as It corrects the information gain by taking the intrinsic information of a split into account (i.e. how much information do we need to tell which branch an instance belongs to) where Intrinsic information is the entropy of distribution of instances into branches (Han & Kamber, 2001).
For a given attribute (Priyadarsini, Valarmathi & Sivakumari, 2011), it is calculated as follows:

$$\text{Gain ratio} = \frac{Gain(Attribute)}{intrinisic\_info(Attribute)} \qquad (4)$$

It can be explained further mathematically as Let S be set consisting of s data samples with m distinct classes. The expected information needed to classify a given sample is given by

$$I(S) = -\sum_{i=m}^{m} Pi \log_2 Pi \quad (5)$$

where Pi is the probability that an arbitrary sample belongs to class Ci and is estimated by Si /S.

Let attribute A has v distinct values. Let Sij be number of samples of class Ci in a subset Sj.Sj contains those samples in S that have value aj of A. The entropy, or expected information based on the partitioning into subsets by A, is given by

$$E(A) = -\sum_{i=1}^{m} I(S) \frac{S1i+S2i+S3i+\cdots_n+Smi}{S} \quad (6)$$

The encoding information that would be gained by branching on A is

$$Gain\ (A) = I(S) - E(A) \quad (7)$$

Gain ratio which applies normalization to information gain using a value defined as

$$Splitinfo_A = -\sum_{i=m}^{v} \frac{|Si|}{S} * \log_2 \frac{|Si|}{|S|} \quad (8)$$

The above value represents the information generated by splitting the training data set S into v partitions corresponding to v outcomes of a test on the attribute A.

The gain ratio is defined as

$$Gain\ Ratio(A) = \frac{Gain\ (A)}{SplitInfoA(S)} \quad (9)$$

It should be noted that the $Splitinfo_A$ is also the intrinisic_info as in (1) The attribute with the highest gain ratio is used as the splitting factor (Han & Kamber,2001).

3.2 Ensemble Methods

An ensemble is a set of classifiers that learn a target function, and their individual predictions are combined to classify new examples. Ensemble is a technique for combining many weak learners in an attempt to produce a strong learner. Evaluating the prediction of an ensemble typically requires more computation than evaluating the prediction of a single model, so ensembles may be thought of as a way to compensate for poor learning algorithms by performing a lot of extra computation. In this research, Stacking, Majority voting and Multischeme ensemble methods are considered.

3.2.1 Stacking Ensemble Method

Stacked generalization (or stacking) is a way of combining multiple models from classifiers. It introduces the concept of a meta-learner. Unlike bagging and boosting, stacking may be (and normally is) used to combine models of different types. Voting are used in combining the outputs result and only make sense when learning schemes performs comparably well. (Witten, et al, 2011).
Stacking is concerned with combining multiple classifiers generated by different learning algorithms $L_1,\ldots\ldots,L_N$ on a single dataset S, which is composed by a feature vector $S_i = (x_i, y_i)$.

The stacking process can be broken into two phases:

1. Generate a set of base-level classifiers $C_1,\ldots\ldots\ldots, C_N$
   Where $C_i = L_i(S)$
2. Train a meta-level classifier to combine the outputs of the base-level classifiers

If an ensemble has M base models having an error rate e < 1/2 and if the base models' errors are independent, then the probability that the ensemble makes an error is the probability that more than M/2 base models misclassify the example. The simple idea behind stacking is that if an input–output pair (x, y) is left out of the training set of hi, after training is completed for hi, the output y can still be used to assess the model's error. In fact, since (x, y) was not in the training set of hi, hi(x) may differ from the desired output y. A new classifier then can be trained to estimate this discrepancy, given by y –hi(x). In essence, a second classifier is trained to learn the error the first classifier has made. Adding the estimated errors to the outputs of the first classifier can provide an improved final classification decision

3.2.2 Voting Ensemble Method

Voting based methods operate on labels only, where dt,j is 1 or 0 depending on whether classifier t chooses j, or not, respectively. The ensemble then chooses class J that receives the largest total vote:

Majority (plurality) voting

$$\sum_{t=1}^{T} d_{t,J}(\mathbf{x}) = \max_{j=1,\cdots,C} \sum_{t=1}^{T} d_{t,j} \quad (10)$$

Under the condition that the classifier outputs are independent, it can be shown the majority voting combination will always lead to a performance improvement. If there are a total of T classifiers for a two-class problem, the ensemble decision will be correct if at least ⌊T/2+1⌋ classifiers choose the correct class. Now, assume that each classifier has a probability p of making a correct decision. Then, the ensemble's probability of making a correct decision has a binomial distribution, specifically, the probability of choosing k>⌊T/2+1⌋ correct classifiers

$$\text{out of } T \text{ is } P_{ens} = \sum_{k=(T/2)+1}^{T} \binom{T}{k} p^k (1-p)^{T-k} \quad (11)$$

Then, $P_{ens} \rightarrow 1$, as $T \rightarrow \infty$ if $p>0.5$
$P_{ens} \rightarrow 0$,as $T \rightarrow \infty$ if $p<0.5$

Note that the requirement of p>0.5 is necessary and sufficient for a two class problem, whereas it is sufficient, but not necessary for multi class problems (Robbi, 2009).

3.2.3 Multi-Scheme Ensemble Method

Multischeme ensemble method is an ensemble technique for selection by cross validation. New examples are classified by the base-level algorithm which has the least cross validation error on training data.

3.3 Base Classifier

3.3.1 Multilayer Perceptron

The neural network gains the experience initially by training the system to correctly identify pre-selected examples of the problem. The most popular static network is the MLP. MLP are feed-forward neural networks trained with the standard back propagation algorithm. They are supervised networks so they require a desired response to be trained. They are widely used for pattern classification. With one or two hidden layers, they can approximate virtually any input–output map (Ibrahim et al, 2012, Fashoto et al., 2013).

Adsul, Danke, Jagdale, Chaudhari, and Jadhav (2014) explained that Artificial Neural Network (ANN) is the network of individual neurons. Each neuron in a neural network acts as an independent processing element. Each processing element (neuron) is fundamentally a summing element followed by an activation function.

3.3.2 k-Nearest Neighbour

K-Nearest Neighbor (k-NN) is an instance based learning method for classifying objects based on the closest training examples in the feature space (Lee, Stolfo & Mok, 1999). It is a type of lazy learning where the function is only approximated locally and all computations are deferred until classification. The k-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors. If k=1, then the object is simply assigned to the class of its nearest neighbor (Lane, 2000). The k-NN algorithm uses all labeled training instances as a model of the target function. During the classification phase, k-NN uses a similarity-based search strategy to determine a locally optimal hypothesis function.

Test instances are compared to the stored instances and are assigned the same class label as the k most similar stored instances. Generally it is used for intrusion detection in combination with statistical schemes (anomaly detection) (Lee, Stolfo & Mok, 1999). An advantage of the K-Nearest Neighbor Algorithm as a classifier for an IDS is that it is analytically tractable. KNN is simple in implementation and it uses local information, which can yield highly adaptive behavior. Finally, a major strength of the KNN algorithm is that it lends itself very easily to parallel implementations (Lee, Stolfo & Mok, 1999). One of the weaknesses of the K-Nearest Neighbor Algorithm as a classifier for an IDS is its large storage requirements. KNNs are also known to be highly susceptible to the curse of dimensionality and slow in classifying test tuples.

3.3.3 NBTree

The NBTree algorithm is a hybrid of the Naïve Bayes and the Decision Tree algorithm. The learned knowledge is represented in the form of a tree. This Tree is constructed recursively (Rupali & Brajesh, 2014). However, the leaf nodes are Naive Bayes categorizers rather than nodes predicting a single class. For continuous attributes, a threshold is chosen so as to limit the entropy measure. The utility of a node is evaluated by discretizing the data and computing the fivefold cross-validation accuracy estimation using Naive Bayes at the node. The utility of the split is the weighted sum of utility of the nodes and this depends on the number of instances that go through that node (Sandeep & Andreas, 2008). The NBTree Algorithm strives to approximate whether the generalization accuracy of Naive Bayes at each leaf is higher than a single Naive Bayes classifier at that node. A split is considered to be significant if relative reduction in the error is greater than 5% and there are a minimum of 30 instances in the node (Rupali & Brajesh, 2014)

3.4. Proposed System Architecture

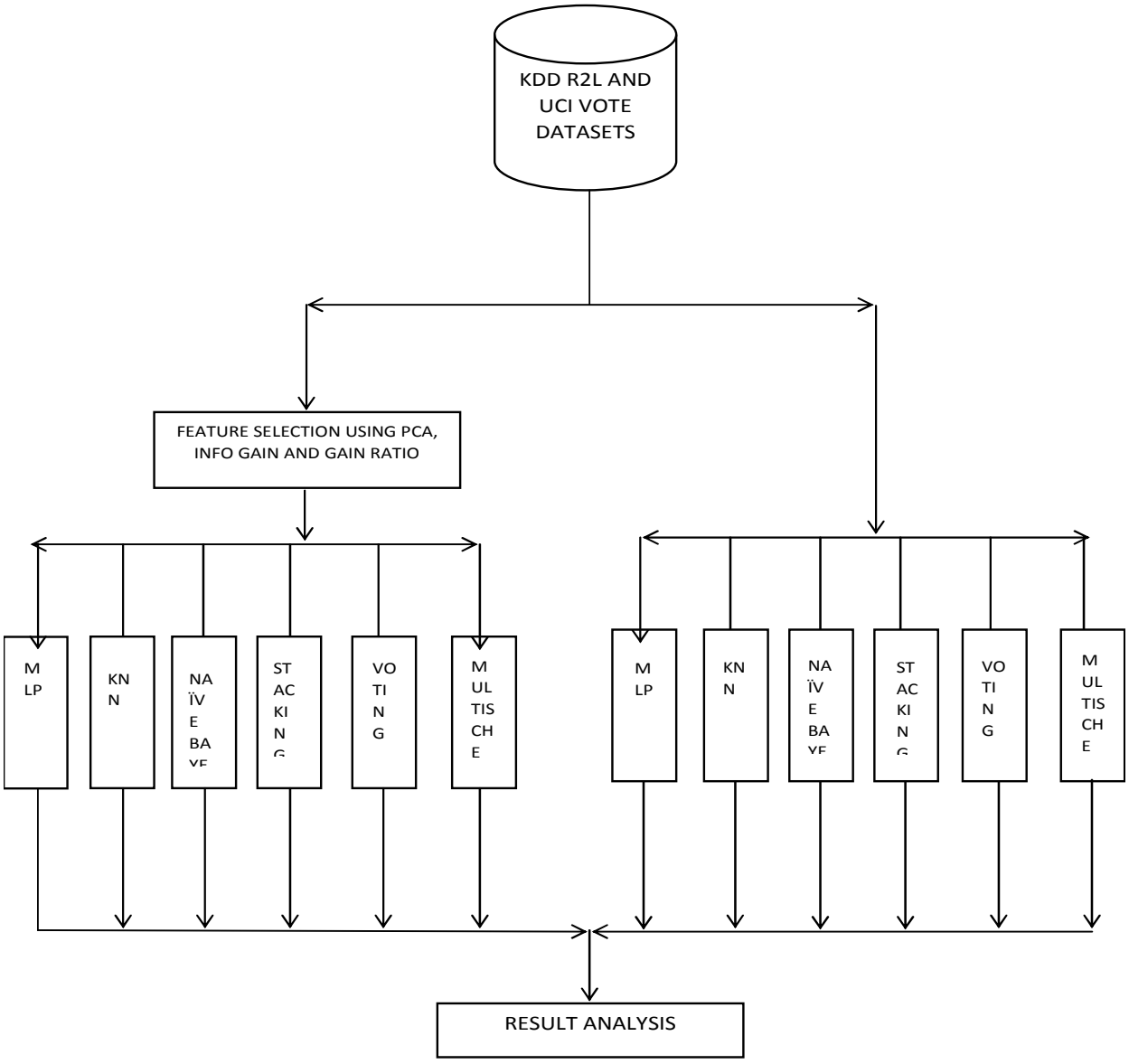

Figure 3.1: Proposed Experimental Architecture

3.4.1 Dataset

For the purpose of this study, the datasets that were used are the Root-to-Local (R2L) dataset in the KDD'99 dataset and the Vote dataset from the Congressional Quarterly Almanac (CQA). Remote to Local Attack (R2L) occurs when an attacker who has the ability to send packets to a machine over the network but does not have an account on that machine exploits some vulnerability to gain root access as a user of that machine. While the vote dataset includes votes for each of the U.S. House of Representatives Congressmen on the 16 key votes identified by the Congressional Quarterly Almanac (CQA). The CQA lists nine different types of votes: voted for, paired for, and announced for (these three simplified to yea), voted against, paired against, and announced against (these three simplified to nay), voted present, voted present to avoid conflict of interest, and did not vote or otherwise make a position known (these three simplified to an unknown disposition) (retrieved from http://archive.ics.uci.edu/ml, 2016).

Afterwards for the purpose of achieving the aim of this study, feature selection of the dataset shall be performed via principal component analysis (PCA), information gain attribute selection and Gain ratio attribute selection independently. This will help in the selection of the number of features that the base classifiers and the ensemble methods will use in making correct and accurate classification. The feature selection technique will best select some attribute from the original 41 features (plus the label feature) from the R2L datasets and original 17 features from the Vote dataset.

The respective datasets (reduced and full datasets of R2L and Vote datasets) will be used to train base classifiers and the ensemble methods for correct classification and it is subsequently tested to see if learning algorithm is able to correctly classify the data. This training and testing process will be carried out using 10-fold cross validation technique. Conclusively, the results derived from both separate operations will be subjected to analysis and comparison. Their performance shall be evaluated and measured respectively via precision, recall, training time, kappa statistics, accuracy and many more. Basically, the true positive and accuracy of each operations shall be revealed as they are the variables that will be used to evaluated the performance of the research methodology.

4. EXPERIMENTAL RESULTS ANALYSIS

The main objective of our simulations is to demonstrate the influence of performing feature pre-selection to the enhancement of the performance of ensemble methods and base classifiers in classification technique. The experiment was carried out using WEKA (Waikato Environment for Knowledge Analysis), a data mining software.

Table 4.1: Accuracy of base classifiers and ensemble methods on R2L dataset with Gain Ratio as feature selection.

	MLP	KNN	NBTree	STA	VOT	MULTI
25 Attributes	98.49	98.49	98.14	97.78	98.58	98.49
30 Attributes	98.93	98.93	98.14	98.67	99.02	98.93
35 Attributes	99.11	98.76	98.31	98.4	98.93	98.76
Full datasets	99.11	98.76	98.67	98.85	99.02	98.76
	Accuracy on Gain ratio on R2L Dataset					

Table 4.2: Accuracy of base classifiers and ensemble methods on R2L dataset with Principal Component analysis as feature selection.

	MLP	KNN	NBTree	STA	VOT	MULTI
25 Attributes	98.05	98.58	97.78	97.96	98.49	98.58
30 Attributes	98.14	98.76	97.87	97.78	98.4	98.77
35 Attributes	98.4	98.76	97.69	97.6	98.67	98.76
Full datasets	98.49	98.58	97.78	96.71	98.58	98.58
	Accuracy on PCAonR2L dataset					

Table 4.3: Accuracy of base classifiers and ensemble methods on R2L dataset with Information Gain as feature selection.

	MLP	KNN	NBTree	STA	VOT	MULTI
25 Attributes	98.67	98.76	98.22	97.56	98.86	98.76
30 Attributes	98.14	98.76	97.86	98.66	98.93	98.85
35 Attributes	98.4	98.76	97.69	98.76	98.93	98.85
Full datasets	98.49	98.76	98.22	98.4	98.85	98.76
	Accuracy on Info Gain on R2L Dataset					

From the tables 4.1, 4.2 and 4.3 above, the ensemble methods (Stacking, Voting and Multischeme) gave a very good accuracy result on the average over the base classifiers with the better results coming from the usage of feature selection techniques such as Principal Component analysis, Information gain and gain ratio. Individually, some of the base classifiers with feature selection techniques, outperformed some of the ensemble methods, such as in Table 4.1, Table 4.2 and Table 4.3, where stacking ensemble method gave a slightly low accuracy compared to the base classifiers when the 25 attributes are selected. The Majority Voting Ensemble method and the Multischeme ensemble methods gave the best results on the R2L dataset with or without feature selection but the results were more impressive for the ensemble methods with feature selection.

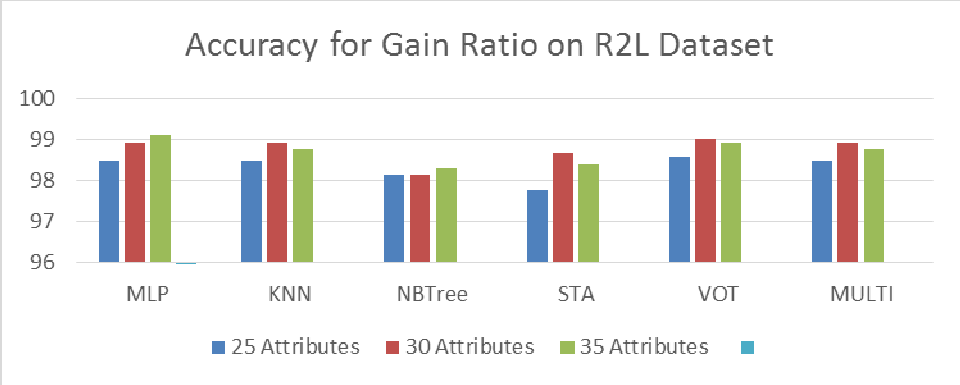

Figure 4.1: A Graphical representation of the accuracy value of the base classifiers and the ensemble methods on R2L dataset using Gain Ration as Feature Selection.

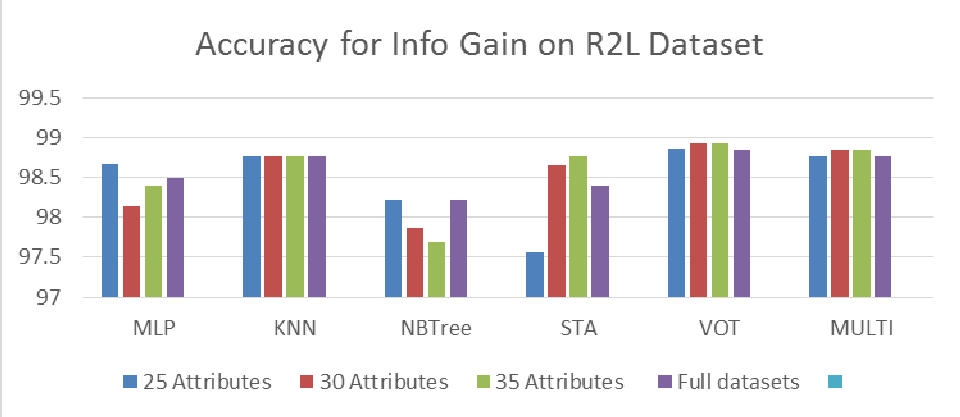

Figure 4.2: A Graphical representation of the accuracy value of the base classifiers and the ensemble methods on R2L dataset using Information Gain as Feature Selection.

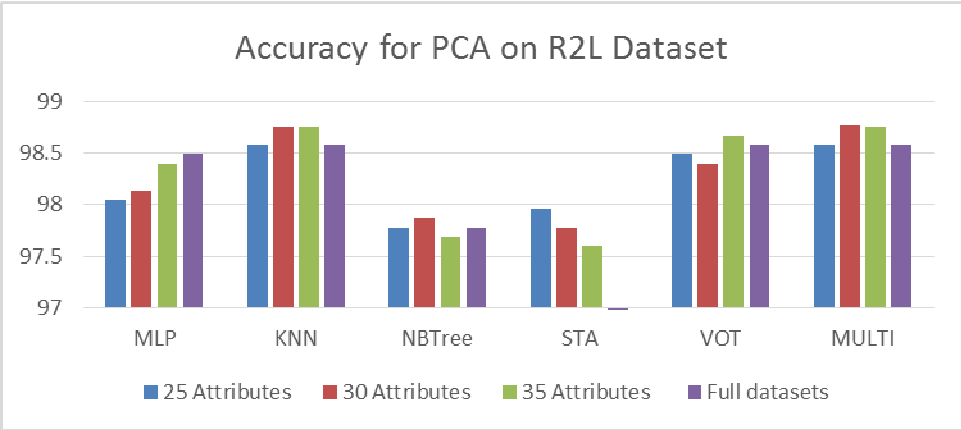

Figure 4.3: A Graphical representation of the accuracy value of the base classifiers and the ensemble methods on R2L dataset using Principal Component Analysis as Feature Selection.

Table 4.4: Accuracy of base classifiers and ensemble methods on Vote dataset with Principal Component Analysis as feature selection

	MLP	KNN	NBTree	STA	VOT	MULTI
5 Attributes	93.79	92.18	92.41	92.18	93.1	92.18
10 Attributes	93.79	91.26	92.87	93.79	93.79	91.26
15 Attributes	94.48	90.8	93.1	94.25	94.25	90.8
Full datasets	94.02	89.88	93.79	92.87	94.25	89.88
	Accuracy on PCA on vote dataset					

Table 4.5: Accuracy of base classifiers and ensemble methods on Vote dataset with Information Gain as feature selection

	MLP	KNN	NBTree	STA	VOT	MULTI
5 Attributes	95.17	95.86	94.94	93.79	94.94	95.4
10 Attributes	94.71	94.48	94.02	93.56	94.02	95.17
15 Attributes	95.17	93.56	94.48	94.75	95.17	92.87
Full datasets	95.17	92.41	95.63	95.56	95.63	93.1
	Accuracy on Info Gain on vote Dataset					

Table 4.6: Accuracy of base classifiers and ensemble methods on Vote dataset with Gain Ratio as feature selection

	MLP	KNN	NBTree	STA	VOT	MULTI
5 Attributes	95.17	95.86	94.94	93.79	94.94	95.4
10 Attributes	94.71	94.48	94.02	93.56	94.02	95.17
15 Attributes	95.17	93.56	94.48	94.25	95.17	92.87
Full datasets	95.17	92.41	95.63	95.86	95.63	93.1
	Accuracy on Gain Ratio on vote dataset					

From the tables 4.4, 4.5 and 4.6 above, the ensemble methods (Stacking, Voting and Multischeme) gave a very good accuracy result on the average over the base classifiers with the better results coming from the usage of feature selection techniques such as Principal Component analysis, Information gain and gain ratio just like in case of using R2L dataset. Individually, some of the base classifiers with feature selection techniques, outperformed some of the ensemble methods, such as in Table 4.4, Table 4.5 and Table

4.6, where stacking ensemble method and Voting ensemble methods gave a slightly low accuracy compared to the base classifiers when the 5 and 10attributes are selected but reverse is the case when 15 attributes were selected. The Multischeme ensemble methods gave the best results on the Vote dataset with or without feature selection but the results were more impressive for the ensemble methods with feature selection. The poor results of some of the ensemble methods (Stacking and Voting) is due to the few attributes (5 and 10 attributes) selected. Unlike in the case of R2L which has 42 attributes, the Vote dataset has 17 datasets and in the results of the R2L datasets, the ensemble methods worked very due to the reason that 25, 30 and 35 attributes are selected.

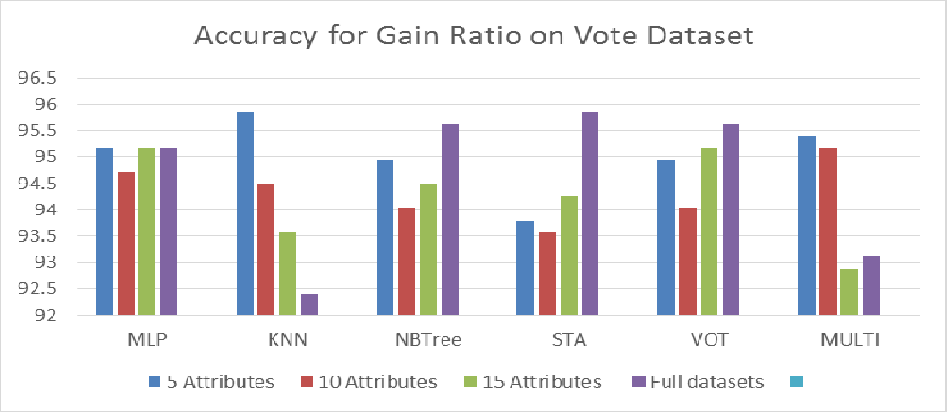

Figure 4.4: A Graphical representation of the accuracy value of the base classifiers and the ensemble methods on Vote dataset using Gain Ratio as Feature Selection.

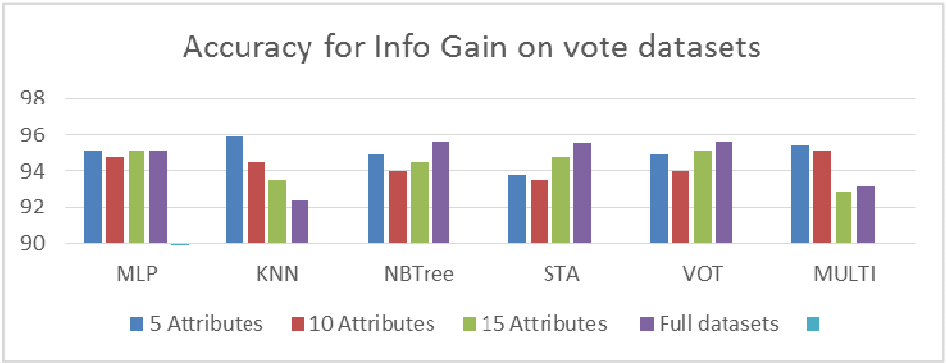

Figure 4.5: A Graphical representation of the accuracy value of the base classifiers and the ensemble methods on Vote dataset using Information Gain as Feature Selection.

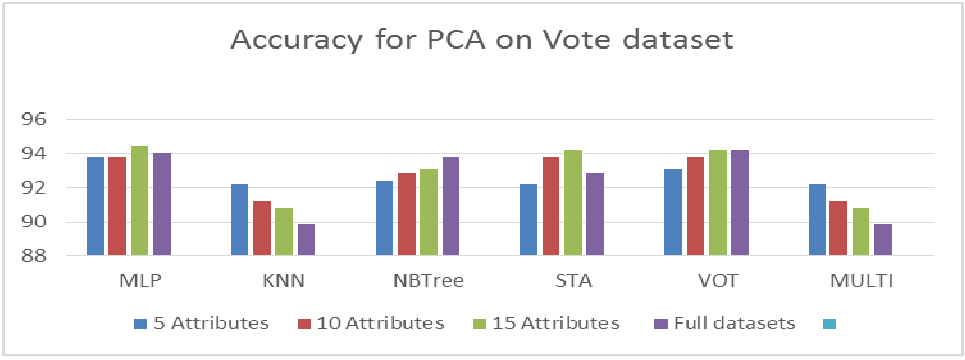

Figure 4.6: A Graphical representation of the accuracy value of the base classifiers and the ensemble methods on Vote dataset using Principal Component Analysis as Feature Selection.

5. CONCLUSIONS AND FUTURE RECOMMENDATION

From the analysis, the experimental result revealed that feature selection improved the performance of the ensemble methods (Stacking, Majority Voting and Multischeme) in both datasets, though Staking ensemble method didn't perform well on the reduced datasets but it was still better than the individual base classifiers. Perhaps the stacking ensemble method may have considered the selected features too little but there were significant improvement in the Kappa statistic in both dataset. These experiments recorded improvements on Voting and Multischeme ensemble methods.

For the R2L datasets, with the difference of accuracy results of the reduced and full datasets, it can be deduced that feature selection should be used in such process, so as to increase the accuracy of the classification process and also reduce the time taken for the computation. For the Vote datasets, with the difference of accuracy results of the reduced and full datasets, it can be deduced that feature selection should not be used in such process; so as to increase the accuracy of the classification process and the time taken for the computation won't be much since there are few attributes.

Therefore considering the general outcome of the experiment, we can say that the optimization performed on the dataset caused an improvement on ensemble method, hence such feature selection technique should be carried before data analysis. From this research, it was discovered that data pre-processing such as feature selection caused an improvement in the accuracy of the base classifiers and ensemble methods but researchers should exercise caution in the choice of feature selection to use before data analysis is carried out. Also, the search method to be used should be considered carefully because different search method produces different results on the dataset during feature selection. Perhaps, stacking ensemble could have performed better if a different search method were used. Further study should be carried out on these ensemble methods, applying different feature selection techniques with varied search methods.

REFERENCE

1. Ajayi, A., Idowu, S.A., & Anyaehie, A.A, (2013). Comparative study of the selected data mining algorithms used for intrusion detection. International Journal of Soft computing Engineering (IJSCE).www.ijsce.org/attachments/File/v3i3/C1662073313.pdf
2. Alpaydin, E. (2010). Introduction to Machine Learning. MIT Press. p. 9. ISBN 978-0-262-01243-0..
3. Aneetha, A.S. & Bose, S.(2012). The combined approach for anomaly detection using neural networks and clustering techniques. Comput. Sci. Eng. Int. J., 2: 37-46
4. Balogun, A. O., & Jimoh, R.G. (2015): Anomaly Intrusion Detection Using An Hybrid Of Decision Tree And K-Nearest Neighbor. Journal of Advances in Scientific Research & Applications (JASRA). 2(1): 67-74. Published by Faculty of Science, Adeleke University, Ede, Osun state. Available online at http://www.cisdijournal.net/uploads/JASRA-V2N1P7.pdf
5. Balogun, A. O., Mabayoje M.A., Salihu, S. & Arinze, S.A. (2015):Enhanced Classificatio Via Clustering Using Decision Tree for Feature Selection. International Journal of Applied Information Systems (IJAIS). 9(6):11-16. Published by Foundation of Computer Science FCS, New York, USA. Available online at http://www.ijais.org/research/volume9/number6/abdullateef-2015-ijais-451425.pdf.
6. Borji, A.,(2007). Combining heterogeneous classifiers for network intrusion detection. Annual Asian Computing Science Conference, 254-260, 2007. Publisher: springer Berlin Heidelberg. Chandrashekar, G., & Sahin, F., (2014). A survey on feature selection methods. Computers and Electrical Engineering: 40(2014) 16-28
7. Fashoto,S.G., Akinnuwesi, B., Owolabi,O. & Adelekan, D.(2016). Decision support model for supplier selection in healthcare service delivery using analytical hierarchy process and artificial neural network. African Journal of Business Management, Vol. 10(9), pp. 209-232. DOI: 10.5897/AJBM2016.8030
8. Fashoto,S.G., Gbadeyan, J.A. & Sadiku, J.S.(2013) Application of Data Mining technique to fraud detection in Health Insurance Scheme using Multilayer Perceptron. In the Proceedings of the 2nd annual International Conference on E-leadership of IEEE in University of Pretoria, Pretoria, South Africa, pp. 107 - 116.
9. Han, J. & Kamber, M. (2006). Data Mining: Concepts and Techniques (2nd ed., pp 1-10).
10. Hashemi, V.M., Muda, Z., & Yassin, W.(2013). Improving intrusion detection using genetic algorithm. Information Technology Journal., 12: 2167-2173.
11. Kohavi, R. & John, G. H. (1997). Wrappers for feature subset selection. Artificial Intelligence, 97, pp. 273–324. https://intranet.cs.aau.dk/fileadmin/user_upload/Education/Courses/2010/DWML/papers/kohavi-john-wrappers.pdf
12. Kumar, V. & Minz, S. (2014). Feature Selection: A Literature Review. Smart Computing review, Vol 4., No. 3, June, 2014.

13. Lai, T. N., Chapelle, O., Weston, J., & Elisseeff, A., (2014). Embedded Methods www.cyberneum.de/fileadmin/user_upload/files/publications/pdf3012.pdf
14. Lane, T. D. (2000). "Machine Learning Techniques for the computer security domain of anomaly detection", Ph.D. Thesis, Purdue Univ., West Lafayette, IN, USA.
15. Lee,W., Stolfo,S.,& Mok, K., (1999). A data mining framework for building intrusion detection models.Security and Privacy, 1999. Proceedings of the 1999 IEEE Symposium. DOI:10.1109/SECPRI.1999.766909
16. Mabayoje M.A., Akintola, A. G., Balogun, A. O & Ayilara, O. (2015): Gain Ratio and Decision Tree Classifier for Intrusion Detection. International Journal of Computer Applications (IJCA). 126(1):56-59. Published by Foundation of Computer Science FCS, New York, USA. Available online at http://www.ijcaonline.org/research/volume126/number1/modinat-2015-ijca-905983.pdf
17. Mahdi, E. & Fazekas, G. (2011). Feature Selection as an Improving Step for Decision Tree Construction. International Conference on Machine Learning and Computing IPCSIT Vol. 3. p. 35.
18. Sanchez-Marono, N., Alonso-Betanzos A., & Tombilla-Sanromán, M. (2007). Filter methods for Feature Selection- A Comparative Study. Intelligent Data Engineering and Automated Learning - IDEAL 2007. 8th International Conference, Birmingham, UK, December 16-19, 2007. Proceedings pp 178-187 DOI: 10.1007/978-3-540-77226-2_19
19. Pitt, E., & Nayak, R., (2007). The Use of Various Data Mining and Feature Selection Methods in the Analysis of a Population Survey Dataset.
20. Ramaswami, M., & Bhaskaran, R. (2009). A Study on Feature Selection Techniques in Educational Data Mining. JOURNAL OF COMPUTING, VOLUME 1, ISSUE 1, DECEMBER 2009, ISSN: 2151-9617
21. Santana, L. E., de Oliveira, D. F., Canuto, A. M. P., & de Souto, C. P.M. (2007) Comparative Analysis of Feature Selection Methods for Ensembles with Different Combination Methods. Proceedings of International Joint Conference on Neural Networks, Orlando, Florida, USA, August 12-17, 2007.
22. Silwattananusarn, T. & Tuamsuk, K. (2012). Data Mining and Its Applications for Knowledge Management : A Literature Review from 2007 to 2012 International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.5, September 2012.
23. Singh, A. G., Appavu, S., & Jebamalar, E. L. (2016). Literature Review On Feature Selection Methods for High-Dimensional Data. International Journal of Computer Applications (0975 –8887)Volume 136 –No.1, February 2016
24. Smith, I. L. (2002). A tutorial on Principal Components Analysis.https//arvix.org/pdf/1404.1100
25. Tuarob S., Tucker, C. S., Salathe, M., & Ram, N. (2014). An ensemble heterogeneous classification methodology for discovering health-related knowledge in social media messages. Journal of Biomedical Informatics 49 (2014) 255–268
26. Witten, I. H., Frank, E. & Hall, A. M. (2011). Data Mining: Practical Machine Learning Tools and Techniques (3rd edition). Morgan Kaufmann Publishers Inc., Inc., San Francisco, CA, USA.
27. Valliappan, R., Sundresan, P., & Putra, S. (2015). Interactive Data Mining: A Short Background Study on Effective Interaction and Visualization by Association Rules. 2nd International conference on Innovative Engineering Technologies (ICIET'2015) August Bangkok (Thailand)
28. Venkatadri.M & Reddy, L. C. (2011). A Review on Data mining from Past to the Future. International Journal of Computer Applications 15(7):19–22, February 2011..
29. Wanner, L.(2004). Introduction to Clustering Techniques", International Union of Local Authorities, July, 2004.
30. WEKA at http://www.cs.waikato. ac.nz/~ml/weka.
31. Whalen, S., & Pandey, G., (2013). A Comparative Analysis of Ensemble Classifiers: Case Studies in Genomics. arxiv: 1309.5047.pdf
32. Zhou, Z.,(2012) Ensemble Method Foundation and Algorithms. 236 P. Boca Ration, FL: Chapman & Hall/CRC. ISBN:978-1-439-830031