

Western Kentucky University
TopSCHOLAR®

Masters Theses & Specialist Projects

Graduate School

5-1-2012

Ensemble of Feature Selection Techniques for High Dimensional Data

Sri Harsha Vege

Western Kentucky University, sriharsha.vege495@topper.wku.edu

Follow this and additional works at: <http://digitalcommons.wku.edu/theses>



Part of the [Databases and Information Systems Commons](#)

Recommended Citation

Vege, Sri Harsha, "Ensemble of Feature Selection Techniques for High Dimensional Data" (2012). *Masters Theses & Specialist Projects*. Paper 1164.

<http://digitalcommons.wku.edu/theses/1164>

This Thesis is brought to you for free and open access by TopSCHOLAR®. It has been accepted for inclusion in Masters Theses & Specialist Projects by an authorized administrator of TopSCHOLAR®. For more information, please contact connie.foster@wku.edu.

ENSEMBLE OF FEATURE SELECTION TECHNIQUES FOR HIGH
DIMENSIONAL DATA

A Thesis
Presented to
The Faculty of the Department of Mathematics and Computer Science
Western Kentucky University
Bowling Green, Kentucky


In Partial Fulfillment
Of the Requirements for the Degree
Master of Science

By
Sri Harsha Vege

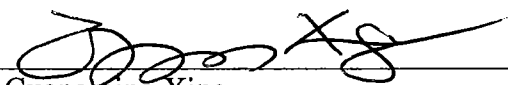
May 2012

ENSEMBLE OF MULTIPLE FEATURE SELECTION TECHNIQUES FOR HIGH
DIMENSIONAL DATA

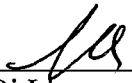
Date Recommended 05/02/2012



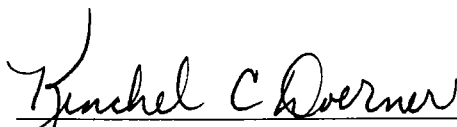
Dr. Huanjing Wang, Director of Thesis



Dr. Guangming Xing



Dr. Qi Li



Dean, Graduate Studies and Research Date 21-May-2012

ACKNOWLEDGEMENTS

This thesis would not have been possible without the guidance and the help of several individuals who in one way or another contributed and extended their valuable assistance in the preparation and completion of this study.

I especially want to thank my professor, Dr. Huanjing Wang for her guidance during my thesis. Her perpetual energy and enthusiasm motivated me to put in my best effort. In addition, she was always accessible and willing to help me with all the queries and difficulties I faced.

It would not have been possible to write this thesis if Dr. Guangming Xing hadn't kindled an interest in me to start my thesis. My utmost gratitude goes to him whose sincerity and encouragement has been my driving force.

My thesis found a strong foundation and gained strength from the lectures and seminars given by Dr. Qi Li. He gave in valuable inputs for my thesis and guided me. His suggestions and advice have played a vital role and steered my path.

TABLE OF CONTENTS

List of Figures	vi
List of Tables	vii
Abstract	viii
1. Introduction to Feature Selection	1
1.1 Introduction	1
1.2 Overview of Thesis	2
1.3 Outline	3
2. Related Work	6
3. Feature Ranking	9
3.1 Feature Selection	9
3.2 Feature Ranking Techniques	10
3.2.1 Information Gain	11
3.2.2 Gain Ratio	12
3.2.3 Symmetrical Uncertainty	13
3.2.4 ReliefF	14
3.2.5 OneR Attribute Evaluation	15
4. Classification	16
4.1 K Nearest Neighbour	16
4.2 Naïve Bayes	17
4.3 Support Vector Machines	17
4.4 Random Forest	19
4.5 Logistic Regression	19

4.6	C4.5 Classifier	19
5.	Evaluation Criteria	21
6.	Ensemble Feature Ranking Techniques	23
6.1	Ensemble of Single Feature Ranking Technique	24
6.2	Ensemble of Multiple Feature Ranking Techniques	24
6.3	Existing Ensemble Methods for Multiple Feature Ranking Techniques	24
6.4	Proposed Algorithm	25
7.	Experimental Design and Evaluation	28
7.1	Datasets	28
7.2	Experimental Design	28
7.3	WEKA	29
7.4	Case Study 1: Lung Cancer Dataset	30
7.4.1	Experimental results	30
7.4.2	Analysis of results	33
7.5	Case Study 2: Lymphoma Dataset	39
7.5.1	Experimental results	39
7.5.2	Analysis of results	42
8.	Conclusions and Future Research	43
	References	45

LIST OF FIGURES

1. Fig 4.1 Support vector machine	18
2. Fig 7.1 Model performance for top 20 features for Lung Cancer Dataset	35
3. Fig 7.2 Model performance for top 15 features for Lung Cancer Dataset	36
4. Fig 7.3 Model performance for top 10 features for Lung Cancer Dataset	36
5. Fig 7.4 Model performance for top 5 features for Lung Cancer Dataset	37
6. Fig 7.5 ANOVA tests on AUC for six rankers	38

LIST OF TABLES

1. Table 7.1 AUC values for top 20 features for Lung Cancer Dataset	31
2. Table 7.2 AUC values for top 15 features for Lung Cancer Dataset	31
3. Table 7.3 AUC values for top 10 features for Lung Cancer Dataset	32
4. Table 7.4 AUC values for top 5 features for Lung Cancer Dataset	32
5. Table 7.5 Average model performance for top k features	34
6. Table 7.6 Average model performance for classifiers	34
7. Table 7.7 Average model performance for rankers	35
8. Table 7.8 AUC values for top 25 features for Lymphoma Dataset	39
9. Table 7.9 AUC values for top 50 features for Lymphoma Dataset	40
10. Table 7.10 AUC values for top 100 features for Lymphoma Dataset	40
11. Table 7.11 AUC values for top 500 features for Lymphoma Dataset	41
12. Table 7.12 AUC values for top 1000 features for Lymphoma Dataset	41

ENSEMBLE OF FEATURE SELECTION TECHNIQUES FOR HIGH DIMENSIONAL DATA

Sri Harsha Vege

May 2012

48 pages

Directed by: Dr. Huanjing Wang, Dr. Guangming Xing, and Dr. Qi Li

Department of Mathematics and Computer Science

Western Kentucky University

Data mining involves the use of data analysis tools to discover previously unknown, valid patterns and relationships from large amounts of data stored in databases, data warehouses, or other information repositories. Feature selection is an important pre-processing step of data mining that helps increase the predictive performance of a model. The main aim of feature selection is to choose a subset of features with high predictive information and eliminate irrelevant features with little or no predictive information. Using a single feature selection technique may generate local optima.

In this thesis we propose an ensemble approach for feature selection, where multiple feature selection techniques are combined to yield more robust and stable results. Ensemble of multiple feature ranking techniques is performed in two steps. The first step involves creating a set of different feature selectors, each providing its sorted order of features, while the second step aggregates the results of all feature ranking techniques. The ensemble method used in our study is frequency count which is accompanied by mean to resolve any frequency count collision.

Experiments conducted in this work are performed on the datasets collected from Kent Ridge bio-medical data repository. Lung Cancer dataset and Lymphoma dataset are selected from the repository to perform experiments. Lung Cancer dataset consists of 57 attributes and 32 instances and Lymphoma dataset consists of 4027 attributes and 96

instances. Experiments are performed on the reduced datasets obtained from feature ranking. These datasets are used to build the classification models. Model performance is evaluated in terms of AUC (Area under Receiver Operating Characteristic Curve) performance metric. ANOVA tests are also performed on the AUC performance metric. Experimental results suggest that ensemble of multiple feature selection techniques is more effective than an individual feature selection technique.

Chapter 1: Introduction to Feature Selection

1.1 Introduction

Data mining involves the use of data analysis tools to discover previously unknown, valid patterns and relationships from large amounts of data stored in databases, data warehouses, or other information repositories. Data mining has two approaches. The first approach tries to produce an overall summary of a set of data to identify and describe main features. The second approach, pattern detection, seeks to identify small unusual patterns of behavior. The data mining analysis tasks typically fall into the following categories: data summarization, segmentation, classification, prediction, dependency analysis.

Various models have been developed to help explain the data mining process. One of the models is CRISP-DM [1]. It is a De Facto standard for industry. The CRISP-DM project began in mid-1997 to define and validate an industry and tool-neutral data mining process model. The six steps developed in this model are: business understanding, data understanding, data preparation, modeling, evaluation and deployment. Business understanding is the phase of understanding objectives and requirements of a project. Data understanding is the phase of becoming familiar with the data like identifying data quality problems, discover first insights into data. Data preparation phase describes the entire activities essential in constructing a final dataset from raw data. In the modeling phase various modeling techniques are selected and applied to the model. Evaluation is the phase where the project is thoroughly evaluated before the final deployment. Deployment is the phase where the knowledge discovered will be organized and presented in a way a client can use.

1.2 Overview of Thesis

Feature selection is an important pre-processing step in data mining that helps in increasing the predictive performance of a model. Feature selection can be categorized into feature ranking and feature subset selection. Feature ranking ranks the features in accordance with their predictive scores. Feature subset selection groups attributes which can collectively have good predictive scores. Feature ranking techniques can be classified into three categories: filters, wrappers and hybrids [7]. In this thesis we will be using four filter based feature ranking techniques and one wrapper based feature ranking technique. Classification is a data mining technique used to classify or predict group membership for data instances. One of the commendable features of classifier is its ability to tolerate noise. Its difficulty lies in handling quantitative data appropriately.

In this thesis, we use the Area under the ROC (receiver operating characteristic) curve to evaluate classification models. The ROC curve graphs true positive rates versus the false positive rates. Traditional performance metrics evaluate the classifiers with the default decision threshold of 0.5 only [2]. The AUC is a single value measurement whose value ranges from 0 to 1. When the value of AUC is high for a classification model, it suggests that the classification model has the highest probability for making a correct decision. It has also been shown that AUC has lower variance and is more reliable than other performance metrics such as precision, recall and F-measure.

Using a single feature ranking technique may generate local optima. Ensemble approach improves the classification performance by using a combination of feature ranking techniques. Ensemble of multiple feature ranking techniques is performed in two steps. The first step involves creating a set of different feature selectors, each

providing its sorted order of features, while the second step aggregates the results of all feature ranking techniques [3].

In this thesis we propose an ensemble of multiple feature ranking techniques. This method uses frequency count. It will also use mean to resolve any frequency count collision. It starts with counting the occurrence of individual feature in all ranking lists. This would be the frequency of each feature. The next step is to sort the features based on frequency count. The probable chances of features having the same frequency count are high. If more than one feature has the same frequency then we sort the features using mean. The mean value of a feature is obtained by calculating the average of feature's score in all ranking lists. We build classification models using the ensemble ranking list and evaluate the performance of ensemble.

The experimental results have shown that ensemble method performed better than individual ranker. The results have also shown that the selection of optimal feature subset not only depends on the performance of ensemble method but also on the size of feature subset selected.

1.3 Outline

This thesis has eight chapters with outlines provided below:

Chapter one provides an introduction to data mining. It also provides an overview of the thesis. This section explains feature ranking techniques, ensemble technique, classification models and performance metric.

Chapter two provides the related work performed in the area of ensemble of feature selection techniques. The chapter begins by explaining the studies performed by various researchers in this area. All the studies summarized in this section conclude by

stating ensemble feature selection techniques outperform individual feature ranking techniques.

Chapter three explains feature ranking techniques. This chapter begins by explaining the need for feature ranking and then moving on with filters, wrappers and hybrids. It explains the advantage of choosing feature ranking over feature subset selection. It explains four filter based feature ranking techniques and one wrapper based feature ranking technique.

Chapter four explains the classification models that are built using the results obtained from feature ranking techniques. The chapter starts by explaining the role of classification in data mining. This chapter explains six classifiers that will be used in our thesis.

Chapter five explains the performance metric used to evaluate the classification models. AUC is the performance metric used. This chapter tries to explain AUC and the benefit of using AUC over other measures.

Chapter six explains the ensemble technique. This chapter starts with an explanation of the ensemble of feature ranking techniques. It also explains the need to choose the ensemble of multiple feature ranking techniques over the ensemble of single feature ranking techniques and then describes the new ensemble approach we have proposed. The chapter provides a brief description of how the algorithm works and then the algorithm.

Chapter seven explains the experimental design that will be used in our thesis. It also provides results of the experiments in the form of tables and graphs. The experimental results are analysed.

Chapter eight provides the conclusion and future research opportunities. It concludes our thesis work by summarizing the concepts developed. It also explains the future research that could be done.

Chapter 2: Related Work

This section provides a brief coverage of the works performed in the area of ensemble feature ranking. These works assess how an ensemble of feature ranking techniques can improve robustness, performance and diversity. Feature ranking is a process of selecting the most relevant features from a large set of features. It is considered as one of the most critical problems researchers face today in data mining and machine learning. The main focus of ensemble feature ranking approach is on improving classification performance through the combination of feature ranking techniques. Very limited research exists on ensemble feature ranking.

Early studies on ensemble of feature ranking techniques were performed by Rokach et al. [28]. The experiments in this study are performed to check whether ensemble of feature subsets improve classification accuracy over individual rankers. The experiments are performed on datasets obtained from UCI machine learning repository. Five different feature selection algorithms were used to generate 10 ensembles. The combining methods used for ensemble are: majority voting, take-it-all, smaller is heavier. The ensembles were evaluated using C4.5 classification model. The experimental results have shown that ensemble method performed better than individual feature rankers.

Saeyns et al. [21] performed a study on ensemble of feature selection techniques. The study proves that ensemble methods provide more robust and stable results for high dimensional datasets when compared to individual feature selectors. The experiments are performed on datasets obtained from bioinformatics and biomedical domains. Two filter and two wrapper approaches were used as feature selection techniques. They are

symmetrical uncertainty, relief, random forests and linear support vector machines. The ensemble method used in this study is instance perturbation. The ensembles were evaluated using k-nearest neighbour, random forests and support vector machines. The experimental results have shown that robustness of feature ranking and feature subset selection could be improved by using ensemble of feature selection techniques.

Souza et al. [29] performed a study on a framework for combining feature selection techniques. The framework proposed for this study is STochFS. The STochFS framework works by combining the outcomes of feature selection technique in a stochastic manner. These outcomes form a single structure and acts as a seed which can be used for generating new feature selection subsets. The experiments were performed on 13 datasets obtained from the UCI repository. The feature selection techniques used in this study are: LVF, relief, focus and relieved algorithms. The outcomes were evaluated using C4.5, naive bayes and k-nearest neighbour classification models. The experimental results have showed that STochFS framework achieved high performance when compared to individual rankers.

Olsson and Oard [30] performed a study on combining feature selectors for text classification. The experiments were performed on two sets containing 23, 149 documents and 200,000 documents from RCV1-v2. The documents were combined using document frequency thresholding, information gain and the chi-square feature selection methods. The combination methods used are highest rank, lowest rank and average rank combination. The documents were classified using k-nearest neighbours with $k=100$. The evaluation criteria used for this study was R-precision. The

experiments showed that the ensemble approach could achieve higher peak R-precision than a non-combined feature ranker.

Wilker et al. [6] performed a study using six standard and eleven threshold based filter based feature ranking techniques. In this study six ensemble approaches were considered based on standard and threshold based filters. In addition, four other ensemble approaches were developed based on their robustness to class noise. This study used seven datasets from different domain applications, with different dimensions and different level of class imbalance. This work was evaluated on binary classification datasets. The experimental results showed that ensemble robustness can be predicated from the knowledge of individual components.

Chapter 3: Feature Ranking

This chapter explains the need of feature selection in data mining and explains various feature ranking techniques that are needed to perform the experiments.

3.1 Feature Selection

Feature selection is an important pre-processing tool in data mining. It has been an active field of research and development for the past three decades [4]. As the datasets are getting bigger both in terms of instances and feature count in the fields of biomedical research, intrusion detection and customer relationship management, this enormity causes scalability and performance issues in learning algorithms [4]. Feature selection solves the scalability issue and increases the performance of classification models by eliminating redundant, irrelevant or noisy features from high dimensional datasets [5].

Feature selection is a process of selecting a subset of relevant features by applying certain evaluation criteria. In general, feature selection process consists of three phases. It starts with selecting a subset of original features and evaluating each feature's worth in the subset. Secondly, using this evaluation, some features in the subset may be eliminated or enumerated to the existing subset. Thirdly, it checks whether the final subset is good enough using certain evaluation criterion.

Feature selection can be classified into feature subset selection and feature ranking. Feature ranking calculates the score of each attribute and then sorts them according to their scores. Feature subset selection selects a subset of attributes which collectively increases the performance of the model.

The process of feature selection can be supervised, unsupervised or semi-supervised based on class labels. In supervised feature selection, the evaluations of features are determined using their correlation with the class while unsupervised algorithm uses data variance or data distribution in its evaluation. In semi-supervised we use limited label information to improve unsupervised feature selection. Depending on how and when the worth of each feature in the subset is evaluated, three models can be proposed. They are filters, wrappers and hybrids. Filters evaluate the worth of a feature without any learning algorithm. Wrappers have a predetermined learning algorithm to evaluate the worthiness of an attribute in the subset. Hybrids are a combination of filters and wrappers.

Our work emphasis is mainly on filter based feature ranking techniques. The main advantage of using a filter model is that it is independent of the learning model and therefore it is unbiased. The second advantage is that it allows the algorithms to have a simple structure. Having a simple structure in the filter model generates two critical uses. The algorithms are easy to design and they are fast because of the simple design. We will also be using a wrapper based ranking technique.

3.2 Feature Ranking Techniques

In this work we focus primarily on four filter based feature ranking techniques and one wrapper based feature ranking technique. They are

1. Information gain
2. Gain ratio
3. Symmetrical uncertainty
4. ReliefF

5. OneRAttribute evaluation.

3.2.1 Information Gain

Information gain (IG) is based on the concept of entropy. The expected value of information gain is the mutual information of target variable (X) and independent variable (A). It is the reduction in entropy of target variable (X) achieved by learning the state of independent variable (A) [6]. The major drawback of using information gain is that it tends to choose attributes with large numbers of distinct values over attributes with fewer values even though the later is more informative.

In order to calculate information gain, consider an attribute X and a class attribute Y. The information gain of a given attribute X with respect to class attribute Y is the reduction in uncertainty about the value of Y when the value of X is known. The value of Y is measured by its entropy, $H(Y)$ [6]. The uncertainty about Y, given the value of X is given by the conditional probability of Y given X, $H(Y|X)$.

$$I(Y; X) = H(Y) - H(Y|X) \quad (3.1)$$

where Y and X are discrete variables that take values in $\{y_1, \dots, y_k\}$ and $\{x_1, \dots, x_l\}$ then the entropy of Y is given by:

$$H(Y) = - \sum_{i=1}^{i=k} P(Y = y_i) \log_2 P(Y = y_i) \quad (3.2)$$

The conditional entropy of Y given X is

$$H(Y|X) = - \sum_{j=1}^l P(X = x_j) H(Y|X = x_j) \quad (3.3)$$

Alternatively the information gain is given by:

$$I(Y; X) = H(X) + H(Y) - H(X, Y) \quad (3.4)$$

Where $H(X, Y)$ is the joint entropy of X and Y :

$$H(X, Y) = - \sum_{i=1}^k \sum_{j=1}^l P(X = x_j, Y = y_i) \log_2 P(X = x_j, Y = y_i) \quad (3.5)$$

when the predictive variable X is not discrete but continuous, the information gain of X with class attribute Y is computed by considering all possible binary attributes, X_θ , that arise from X when we choose a threshold θ on X [6]. θ takes values from all the values of X . Then the information gain is simply:

$$I(Y; X) = \operatorname{argmax}_{X_\theta} I(Y, X_\theta) \quad (3.6)$$

3.2.2 Gain Ratio

The information gain measure is biased towards tests with many outcomes. That is, it prefers to select attributes having a large number of possible values over attributes with fewer values even though the later is more informative [7]. For example consider an attribute that acts as a unique identifier, such as a student id in a student database. A split on student id would result in a large number of partitions; as each record in the database has a unique value for student id. So the information required to classify database with this partitioning would be $Info_{student_ID}(D) = 0$. Clearly, such a partition is useless for classification.

C4.5, a successor of ID3 [31], uses an extension to information gain known as gain ratio (GR), which attempts to overcome the bias. Let D be a set consisting of d data samples with n distinct classes. The expected information needed to classify a given sample is given by

$$I(D) = - \sum_{i=1}^n p_i \log_2(p_i) \quad (3.7)$$

where p_i is the probability that an arbitrary sample belongs to class C_i . Let attribute A have v distinct values. Let d_{ij} be number of samples of class C_i in a subset D_j . D_j contains those samples in D that have value a_j of A . The entropy based on partitioning into subsets by A , is given by

$$E(A) = - \sum_{i=1}^n I(D) \frac{(d_{1i} + d_{2i} + \dots + d_{mi})}{d} \quad (3.8)$$

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

$$Gain(A) = I(D) - E(A) \quad (3.9)$$

C4.5 applies a kind of normalization to information gain using a “split information” value defined analogously with $Info(D)$ as

$$SplitInfo_A(D) = - \sum_{j=1}^v \left(\frac{|D_j|}{|D|} \right) \log_2 \left(\frac{|D_j|}{|D|} \right) \quad (3.10)$$

This value represents the information computed by splitting the dataset D , into v partitions, corresponding to the v outcomes of a test on attribute A [7]. For each possible outcome, it considers the number of tuples having that outcome with respect to the total number of tuples in D . The gain ratio is defined as

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)} \quad (3.11)$$

The attribute with maximum gain ratio is selected as the splitting attribute.

3.2.3 Symmetrical Uncertainty

Correlation based feature selection is the base for symmetrical uncertainty (SU). Correlation based feature selection evaluates the merit of a feature in a subset using a hypothesis – “Good feature subsets contain features highly correlated with the class, yet uncorrelated to each other” [9]. Symmetric uncertainty is used to measure the degree of

association between discrete features. It is derived from entropy [8]. It is a symmetric measure and can be used to measure feature-feature correlation.

$$SU = 2.0 \times \frac{H(X) + H(Y) - H(X, Y)}{H(Y) + H(X)} \quad (3.12)$$

Symmetrical uncertainty is calculated by the above equation. $H(X)$ and $H(Y)$ represent the entropy of features X and Y . The value of symmetrical uncertainty ranges between 0 and 1. The value of 1 indicates that one variable (either X or Y) completely predicts the other variable [9]. The value of 0 indicates the both variables are completely independent.

3.2.4 ReliefF

Relief was proposed by Kira and Rendell in 1994. Relief is an easy to use, fast and accurate algorithm even with dependent features and noisy data [2]. The algorithm is based on a simple principle. Relief works by measuring the ability of an attribute in separating similar instances. The process of ranking the features in relief follows three basic steps:

1. Calculate the nearest miss and nearest hit.
2. Calculate the weight of a feature.
3. Return a ranked list of features or the top k features according to a given threshold.

ReliefF (RFF) is an extension to relief algorithm. It was extended by Kononenko so that it can deal with multi-class problems and missing values. The basic idea of ReliefF is to draw instances at random, compute their nearest neighbors, and adjust a feature weighing vector to give more weight to features that discriminate the instance

from neighbors of different classes [23]. It is also improved to deal with noisy data and can be used for regression problems.

3.2.5 OneR Attribute Evaluation

Rule based algorithms provide ways to generate compact, easy-to-interpret, and accurate rules by concentrating on a specific class at a time. One way of generating classification rules is to use decision trees. The disadvantage of using a decision tree is because it is complex and incomprehensible [18]. A classification rule can be defined as $r = (a, c)$ where a is a precondition which performs a series of tests that can be evaluated as true or false and c is a class that apply to instances covered by rule r . A general rule of a rule based algorithm tries to cover all instances belonging to a class. Rule base algorithms work on a specific class at a time. Rule based algorithms follow three steps: Generate rule R on training data S , remove the training data covered by rule and repeat the process.

OneR is the simplest approach to finding a classification rule as it generates one level decision tree. OneR constructs rules and tests a single attribute at a time and branch for every value of that attribute. For every branch, the class with the best classification is the one occurring most often in the training data.

Chapter 4: Classification

Classification is a data mining technique used to predict group membership for data instances. It is one of the important techniques in data mining and is used in various applications such as customer relationship management, pattern recognition, disease diagnosis and targeted marketing [14]. One of the commendable features of a classifier is its ability to tolerate noise. Its difficulty lies in handling quantitative data appropriately. Generally, a quantitative attribute domain is divided into a set of regions. This division leads to partitioning whole data space into corresponding regions of attribute domain. Each partition in data space corresponds to a classification rule [14]. This rule classifies the sample into the corresponding representative class of partition. Various classifiers used in our thesis will be studied in this chapter.

4.1 K – Nearest Neighbor

The K Nearest Neighbor (KNN) classifier is a non parametric lazy learning algorithm. A data sample in KNN is classified on the basis of a selected number of k nearest neighbors [15]. The assumptions followed in KNN are

1. KNN assumes that the data is in a feature space, so they have the concept of distance. Euclidean distance can be used to compute distance between vectors.
2. Each training vector is associated with set of vectors and class label.
3. K decides how many neighbors influence the classification

The following rule is the majority rule that is used extensively in KNN. The classification of the nearest neighbors can be decided by calculating the count of

individual class values from all k nearest neighbors. The class value with the majority count is classified to the sample. K is an odd number to avoid duplicate counts.

4.2 Naïve Bayes

A Naïve Bayes (NB) classifier is a simple probabilistic classifier based on Bayes theorem where every feature is assumed to be class-conditionally independent [16]. In naïve bayes learning, each instance is described by a set of features and takes a class value from a predefined set of values. Classification of instances gets difficult when the dataset contains a large number of features and classes because it takes enormous numbers of observations to estimate the probabilities [16]. When a feature is assumed to be class-conditionally independent, it really means that the effect of a variable value on a given class is independent of the values of other variables.

4.3 Support Vector Machines

A support vector machine (SVM) is a hyperplane that separates two different sets of samples with maximum distance of hyperplane to nearest samples from both sets [10]. The formula for the output of a linear SVM is

$$u = \bar{w} \cdot \bar{x} - b \quad (4.1)$$

In this equation w is the normal vector to the hyperplane and x is the input vector. The nearest points lie on the planes $u = \pm 1$. The distance d is

$$d = \frac{1}{\|\bar{W}\|_2} \quad (4.2)$$

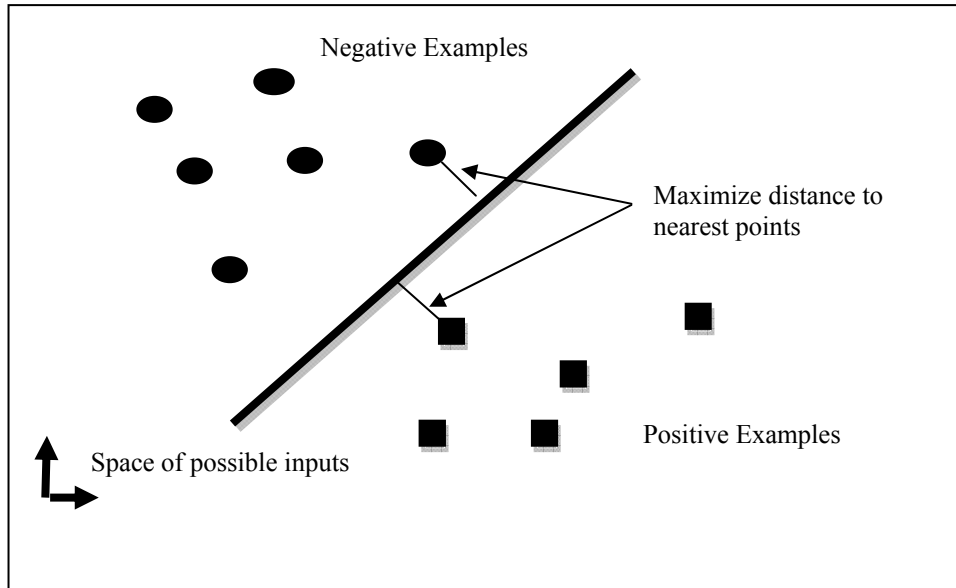
The maximum distance d can be expressed using optimization problem

$$\min_{\bar{w}, b} \frac{1}{2} \|\bar{w}\|^2 \text{ subject to } y_i(\bar{w} \cdot \bar{x}_i - b) \geq 1, \quad (4.3)$$

where x_i is the ith training sample and y_i is the correct output of the SVM for the ith

training sample. The value y_i is +1 for the positive samples and -1 for the negative samples.

Fig 4.1 Support vector machine



The Sequential Minimal Optimization (SMO) is an algorithm that solves quadratic programming (QP) problem which occurs in support vector machine without involving extra matrix space [10]. SMO decomposes the overall QP problem into smallest possible QP sub-problems at every step using Osuna's theorem. At every step, SMO tries to find optimum value of the two Lagrange multipliers and updates the SVM to reflect the new optimum values [10].

4.4 Random Forest

Random forests (RF) are the generalization of recursive partitioning which combines a collection of trees called an ensemble. The random forest was first proposed by Tin Kam Ho of Bell Labs in 1995 which was later extended by Leo Breiman, who also coined the term "Random Forest". Random forests [27] are a collection of identically distributed trees whose class value is obtained by a variant on majority vote.

The classifier consists of a collection of tree like classifiers which uses a large number of decision trees, all of which are trained to tackle the same problem. There are three factors that govern the individuality of the trees:

1. Each tree is trained using a random subset of trained samples.
2. When the tree is growing the best split on each node in the tree is found by searching through n randomly selected features. For a data set with N features, n is selected and kept smaller than that of N .
3. Each tree is made to grow to the fullest so that there is no pruning.

Random forests are tree classifiers that are trained in randomly choosing the subset of input data where the final classification is based on the majority vote by the trees in the forest.

4.5 Logistic Regression

Logistic regression (LR) can be best explained by considering a scenario. Given a set of features in a system space S_p , and an input x_q , the classifier tries to approximate the probability $P(y_q | S_p, x_q)$ for the output y_q [11]. A two dimensional space can be considered as input to the system S_p . The output of this two dimensional space is boolean. Consider an unlabelled point in the two dimensional space is (x_q, y_q) . In order to approximate the probability $P(y_q | S_p, x_q)$, we need some knowledge of the system S_p . Now, the approximation itself is a classification problem.

4.6 C4.5

C4.5 is a variant and extension of an ID3 decision tree algorithm [31]. It is based on the concept of a decision tree. A decision tree is a hierarchical collection of rules that describe how to divide a large collection of data into groups based on the regularities of

the data [17]. It is a tree-like structure used for classification, regression, clustering and prediction function. The objective of a decision tree is to produce an accurate classifier and develop understandable patterns that can be interpreted as interesting knowledge. Decision tree is interesting as it describes a clear relationship between input data and target outputs.

The ID3 algorithm uses gain ratio as the evaluating test. The classification accuracy of a decision tree depends on the test selected to evaluate the training samples. The decision tree algorithms are greedy algorithms. If a test has been selected to partition the training sample, the consequences of alternative choices are not explored. In order to ensure a final predictable tree the choice of tests must be correct. C4.5 contains mechanisms to propose three types of tests: standard test, complex test and binary test. All tests are based on a discrete attribute. These tests are evaluated using gain ratio.

Chapter 5: Evaluation Criteria

A classification algorithm is a function that given a set of training samples and their classes constructs a classifier. A classifier is a function that given an instance assigns it to one of the predefined classes. There are a variety of classifiers that have been developed. The main question that arises in the development and application of these algorithms is about the accuracy of the classifiers they produce. We will be using AUC as evaluation criteria in our thesis which will be discussed in this chapter.

AUC is an acronym for Area under Receiver Operating Characteristic Curve [19]. An ROC graph is a technique for visualizing, organizing and selecting classifiers based on their performance. Given a classifier and an instance, there are four possible outcomes for the instance. If the instance is positive and it is classified as positive, then it is counted as true positive (TP). If it is classified as negative, then it is counted as false negative (FN). If the instance is negative and it is classified as negative, then it is counted as true negative (TN). If it is classified as positive, then it is counted as false positive (FP). If we consider a whole training set we can build a confusion matrix from this methodology [19].

$$\text{Confusion matrix} = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \quad (5.1)$$

The diagonal (upper left to lower right) of the confusion matrix represent the correct decisions made and the elements of the diagonal (upper right to lower left) represent the errors. The true positive rate of a classifier can be estimated as

$$TP \text{ rate} = \frac{\text{Positives correctly classified}}{\text{Total positives}} \quad (5.2)$$

The false positive rate can be defined as

$$FP\ rate = \frac{\textit{Negatively incorrectly classified}}{\textit{Total negatives}} \quad (5.3)$$

ROC graphs are two-dimensional graphs in which TP rate is plotted on the Y-axis and FP rate is plotted on X-axis. An ROC graph depicts relative trade-offs between true positives and false positives. To find a clear dominating relation between two ROC curves we use AUC which provides a single-number summary for the performance of learning algorithms.

Chapter 6: Ensemble Feature Ranking Techniques

Ensemble of feature ranking techniques is an approach where multiple feature ranking lists obtained from corresponding feature ranking techniques are combined to generate a single ranking list. Ensemble of multiple feature ranking techniques is performed to improve the classification performance [7]. Two steps are performed in ensemble of feature ranking techniques. The first step is to create a set of n ranking lists using corresponding rankers and the second is to select the combination function i.e. the function that will transform the ranking lists obtained in the first step into one single ranking list. The second step is the crucial step as it contains the combining method. There are three types of combination methods: fusion based, selection based, and hybrid. Fusion based makes use of all the information obtained from individual rankers to produce a final outcome [7]. Selection based methods chooses a single ranker from the list to become the final outcome. In hybrid, the final outcome is obtained after both selection and fusion methods have been used.

We can illustrate the above method more formally. Let us consider a dataset D with N instances having M features. The first step is to obtain a set of n ranking lists $\{F_1, F_2, F_3 \dots F_n\}$. The second step is to determine a combination method T . Let f_i^j denotes the rank of feature i from ranking list j , such that the set of rankings of feature i is given by $S_i = \{f_i^1, f_i^2 \dots f_i^n\}$. The new score obtained by feature i using the combination method T is

$$\check{f}_i = T(f_i^1, f_i^2 \dots f_i^n) \quad (6.1)$$

There are two ways in which an ensemble can be performed on a ranking list. They are

1. Ensemble of a single feature ranking technique.

2. Ensemble of multiple feature ranking techniques.

We will be using ensemble of multiple feature ranking techniques in our thesis.

6.1 Ensemble of A Single Feature Ranking Technique

In ensemble of a single feature ranking technique boot strap aggregation and some other algorithms can be used to generate different bags of data. For each of the bags, a separate feature ranking was performed, and the ensemble was performed by aggregating the single rankings by weighted voting, using linear aggregation [21].

Bootstrap aggregating, also known as bagging, is a technique used to generate multiple versions of data [20]. The multiple versions are formed by making bootstrap replicates of the data set and using these as data sets for model fitting.

6.2 Ensemble of Multiple Feature Ranking Techniques

Ensembles of multiple feature ranking techniques combine outcomes of various feature selection techniques. This technique yields more stable and robust results. Two steps are essential in creating a single feature ranking list from multiple feature ranking lists. First a set of different ranking lists is created using corresponding rankers and in the second step these ranking lists are combined to using rank ordering of features.

6.3 Existing Ensemble Methods for Multiple Feature Ranking Techniques

The whole ensemble process is the same for all the existing ensemble methods except for the combination method. Every ensemble method differs in combination method. The existing ensemble methods use various aggregate functions such as mean, median etc [2]. In ensemble mean, each feature's score is determined by the average of

the ranking scores of the feature in each ranking list [3]. In ensemble median, each feature's combining score is the median score in all ranking lists.

6.4 Proposed Algorithm

The proposed algorithm is based on the ensemble approach. It determines a feature's importance or score by determining the presence of a feature in the given ranking lists. It also uses an aggregate function mean to avoid frequency collisions. The proposed algorithm can be extended to any number of ranking lists.

The proposed ensemble approach is performed in two steps. It starts with creating a set of different ranking lists obtained using the rankers selected and then applies the ensemble approach to form a single feature ranking list. The ensemble approach used in our study is frequency count which is accompanied by mean to resolve any feature count collision. The first step is to select a fixed number of features from every ranking list. The second step is to count the occurrence of an individual feature in all the ranking lists. This would be the frequency of each feature. Then we sort the features based on frequency. The probable chances of features having the same frequency are high. To resolve the issue of frequency collision, we have introduced mean ordering; each feature's score is determined by the average of ranking scores in all the ranking lists. The sorting is performed in an increasing order.

The input to our algorithm would be a list containing n ranking lists with top k features. The variables n and k can be altered. It starts with initializing an array F containing features and their rank in each ranking list, count and mean rank. It also initializes an ensemble list E . The algorithm starts with selecting the first feature in the ranking list and then searches for the corresponding feature in the remaining ranking

lists. It assigns rank for the feature obtained in each ranking list to the list F. After searching all the ranking lists the feature count is updated and mean is calculated. This process is repeated for all the features in all the lists. Once the process completes, the list F is sorted based on frequency. If the list contains features with the same frequency, then the corresponding features will be sorted based on their mean values. The output of this algorithm would be a list E containing top k features from the list F obtained from the ensemble method.

Algorithm: Ensemble of Multiple Feature Selection Techniques

Input:

n ranking lists (list 0 to n-1) and each list has k features.

Output:

1. An array F containing features and their rank in each ranking list, count, and mean rank.
2. An ensemble list E.

Initialize E and F to empty

FOR each ranking list i

 FOR each feature in ith ranking list

 IF the feature is not in F

 Add the feature and its rank in list i to F

 FOR list j, j is from i+1 to n-1

 IF the feature is in the list j

 Add the rank of the feature in list j to F

 ENDIF

 ENDFOR

 ENDIF

 ENDFOR

ENDFOR

FOR each feature in F

 Calculate frequency and mean rank of the feature

ENDFOR

Sort the features in F based on their frequency, if same frequency, sort by mean rank;
select the top k features and assign the features to list E.

Chapter 7: Experimental Design and Evaluation

7.1 Datasets

Experiments conducted in this study were performed on the datasets collected from Kent Ridge bio-medical data repository [26]. We have chosen Lung Cancer dataset and Lymphoma dataset from the repository. Lung Cancer dataset consists of 57 attributes and 32 instances. All the attributes are nominal. A class attribute for Lung Cancer dataset has 3 distinct values. Lymphoma dataset consists of 4027 attributes and 96 instances. All the attributes are numeric except for class attribute which is nominal. A class attribute for Lymphoma dataset has 9 distinct values.

7.2 Experimental Design

In our design we will be using four filter based feature ranking techniques and one wrapper based feature ranking technique. They are information gain (IG), gain ratio (GR), symmetrical uncertainty (SU), reliefF (RFF) and oneR attribute evaluation (OneR). We will also be using the ensemble approach we have proposed. The experiments were performed to evaluate the predictive performance of individual rankers over ensemble approach. The experiments are also performed on the entire dataset to evaluate the performance between rankers, ensemble and base dataset. In order to evaluate the performance of ensemble approach and individual rankers, we have built classification models using k-nearest neighbor (KNN), naïve bayes (NB), random forest (RF), logistic regression (LR), support vector machines (SVM) and decision trees (C4.5). The classification models used in our study has 10-fold cross validation as a default setting in WEKA. The classification models are evaluated using the AUC performance metric.

7.3 WEKA

The experimental results are obtained using WEKA data mining tool. WEKA is a collection of machine learning algorithms for data mining tasks. The algorithms can be directly applied to dataset from Java code. It contains tools for data visualization, data analysis and predictive modelling [24]. The input files to the WEKA are datasets that are in ARFF format. Aside from algorithms, WEKA also provides a list of options to partition the data between training and testing sets [25]. The steps for using WEKA are

1. Start the WEKA program.
2. It provides a list of options such as explorer, experimenter, knowledge flow and simple CLI.
3. Our experiments are limited to the explorer section. This section has various options that are extensively used in this thesis. It has pre-process, classify, cluster, associate, attribute selection and visualize options.
4. Open the file of the dataset to be mined. Data can be imported from the file in various formats such as ARFF, CSV, C4.5 and binary.
5. We will be using pre-process tab to open the dataset. It gives a detail description of the dataset by displaying all the features in the dataset. It also displays all the values available for a feature.
6. We will mainly use classify and attribute selection tabs where we can use different classifiers and feature selection techniques.

The proposed algorithm was implemented in JAVA using WEKA developer version 3.3. The subversion repository server of WEKA is used for developing the proposed algorithm.

7.4 Case Study 1: Lung Cancer Dataset

7.4.1 Experimental results

We have applied the six feature ranking techniques (GR, RFF, SU, OneR, IG, and Ensemble) to the Lung Cancer dataset. We have selected the top k (k is set to 20, 15, 10, and 5) feature subsets for the experiments. After the feature selection, we used six learners, KNN, C4.5, NB, RF, LR, and SVM, to build classification models on the datasets with various selected subset of features. The classification models are evaluated in terms of the AUC performance metric. The results of the experiments are displayed in Table 7.1, 7.2, 7.3 and 7.4. Each value in the table is determined by the row (ranker) and the column (learner) in which the value is loaded. It also depends on the value of k used for the table. The process of calculating AUC value for a table is performed in three steps:

1. Identify the row and column for which the AUC needs to be calculated. This helps in selecting a ranker and a learner.
2. Ranker is applied to the dataset to get the ranking list. The top k features are selected from the ranking list. The value of k can be determined by checking the table for which the AUC is calculated.
3. Classification model is built using the dataset with selected features from the previous step.

The last row of each table represents the results obtained for base datasets. These results can be used as a baseline for comparison.

Table 7.1: AUC values for rankers with top 20 features for Lung Cancer dataset

Rankers	KNN	C4.5	NB	RF	LR	SVM	Average
GR	0.7583	0.7871	0.8678	0.8098	0.6386	0.7529	0.7690
RFF	0.7991	0.7848	0.8329	0.7636	0.7215	0.7530	0.7758
SU	0.7892	0.7883	0.8552	0.7671	0.7116	0.7676	0.7798
OneR	0.7290	0.7273	0.8288	0.7865	0.6332	0.7553	0.7433
IG	0.7296	0.7759	0.8449	0.7741	0.7770	0.7543	0.7759
Ensemble	0.7755	0.7883	0.8673	0.8070	0.7454	0.7676	0.7918
Base Dataset	0.5970	0.6620	0.7130	0.6630	0.5810	0.6240	0.6400

Table 7.2: AUC values for rankers with top 15 features for Lung Cancer dataset

Rankers	KNN	C4.5	NB	RF	LR	SVM	Average
GR	0.7302	0.8043	0.8750	0.7772	0.6682	0.7813	0.7727
RFF	0.7061	0.8204	0.8297	0.7770	0.7366	0.7187	0.7647
SU	0.7986	0.7927	0.8693	0.7808	0.6932	0.7295	0.7773
OneR	0.5993	0.8038	0.8234	0.6446	0.5916	0.7157	0.6964
IG	0.7986	0.7927	0.8693	0.7808	0.6932	0.7295	0.7773
Ensemble	0.7504	0.8146	0.8707	0.7967	0.7150	0.7670	0.7857
Base Dataset	0.5970	0.6620	0.7130	0.6630	0.5810	0.6240	0.6400

Table 7.3: AUC values for rankers with top 10 features for Lung Cancer dataset

Rankers	KNN	C4.5	NB	RF	LR	SVM	Average
GR	0.7573	0.7637	0.8512	0.7498	0.7182	0.7002	0.7567
RFF	0.7987	0.8089	0.8636	0.8305	0.7349	0.7148	0.7919
SU	0.7767	0.8040	0.8553	0.8097	0.6840	0.7612	0.7818
OneR	0.7843	0.7896	0.8303	0.7566	0.6525	0.7477	0.7601
IG	0.7977	0.7913	0.8370	0.7923	0.6287	0.7760	0.7705
Ensemble	0.7767	0.8040	0.8553	0.8097	0.6840	0.7612	0.7818
Base Dataset	0.5970	0.6620	0.7130	0.6630	0.5810	0.6240	0.6400

Table 7.4: AUC values for rankers with top 5 features for Lung Cancer dataset

Rankers	KNN	C4.5	NB	RF	LR	SVM	Average
GR	0.7356	0.7495	0.7849	0.7349	0.6657	0.7739	0.7407
RFF	0.8442	0.7691	0.8317	0.8485	0.8042	0.8357	0.8222
SU	0.8771	0.7431	0.8279	0.8355	0.7537	0.8212	0.8097
OneR	0.7585	0.7573	0.7812	0.7642	0.7204	0.7739	0.7592
IG	0.8771	0.7431	0.8279	0.8355	0.7537	0.8212	0.8097
Ensemble	0.8807	0.7726	0.8357	0.8335	0.8051	0.8404	0.8280
Base Dataset	0.5970	0.6620	0.7130	0.6630	0.5810	0.6240	0.6400

7.4.2 Analysis of Results

The tables 7.1, 7.2, 7.3 and 7.4 summarize the classification performance in terms of AUC for the five selected rankers and ensemble method with top k features. The tables also display model performance on base dataset. All these results are mapped into a group of features as shown in Figures 7.1 through 7.4. The results from the above experiments can be summarized in terms of size of feature subset, classifiers and rankers in the following tables. Table 7.5 shows that selecting top 5 features subset generates highest classification accuracy when compared to other feature subsets while the top 15 features subset performed lowest. Table 7.6 suggests that NB has the highest classification accuracy over other classifiers while LR performed worst. Table 7.7 shows that ensemble ranker performed best over other rankers in terms of AUC performance metric, while OneR performed worst.

We also compared the results from the subset of features with the results from the complete set of features (base dataset). We found that the classification performance is improved even after a significant number of features were removed from the original dataset. This demonstrates that feature selection was successfully applied to the Lung Cancer dataset.

Table 7.5: Average model performances for top k features using Lung Cancer dataset

Top k features	AUC
20	0.7726
15	0.7623
10	0.7738
5	0.7949

Table 7.6: Average model performances for classifiers using Lung Cancer dataset

Classifier	AUC
KNN	0.7762
C4.5	0.7823
NB	0.8423
RF	0.7860
LR	0.7054
SVM	0.7633

Table 7.7: Average model performances for rankers using Lung Cancer dataset

Ranker	AUC
GR	0.7598
RFF	0.7887
SU	0.7871
OneR	0.7397
IG	0.7834
Ensemble	0.7968

Fig 7.1 Model performance for top 20 features for Lung Cancer dataset

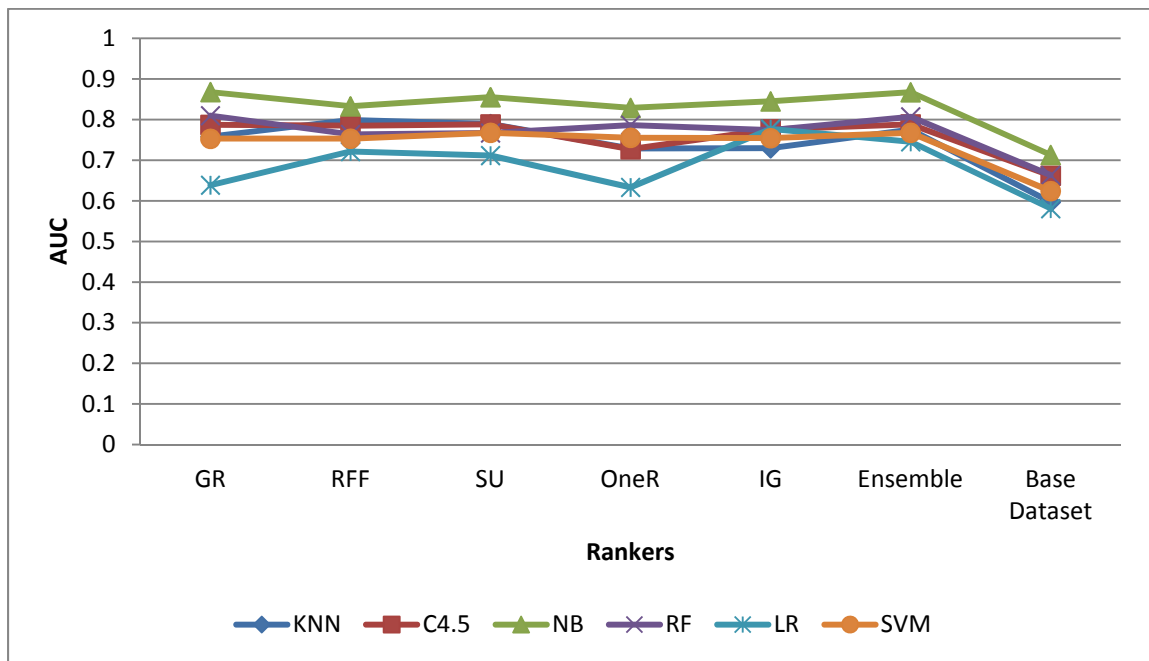


Fig 7.2 Model performance for top 15 features for Lung Cancer dataset

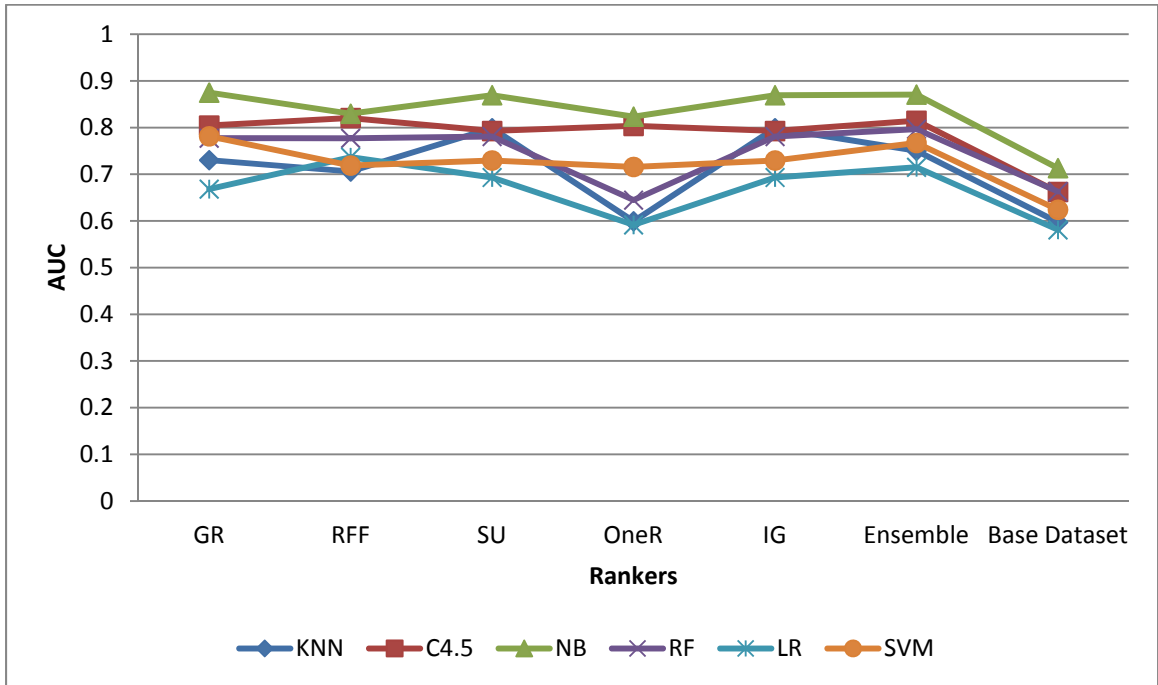


Fig 7.3 Model performance for top 10 features for Lung Cancer dataset

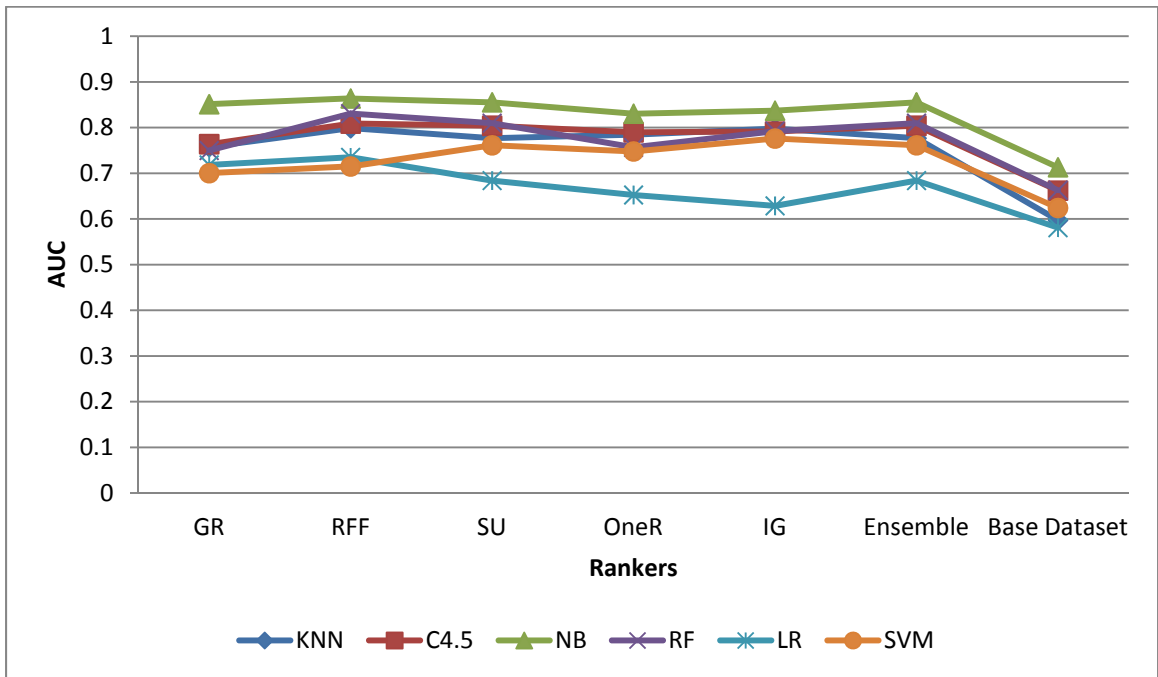
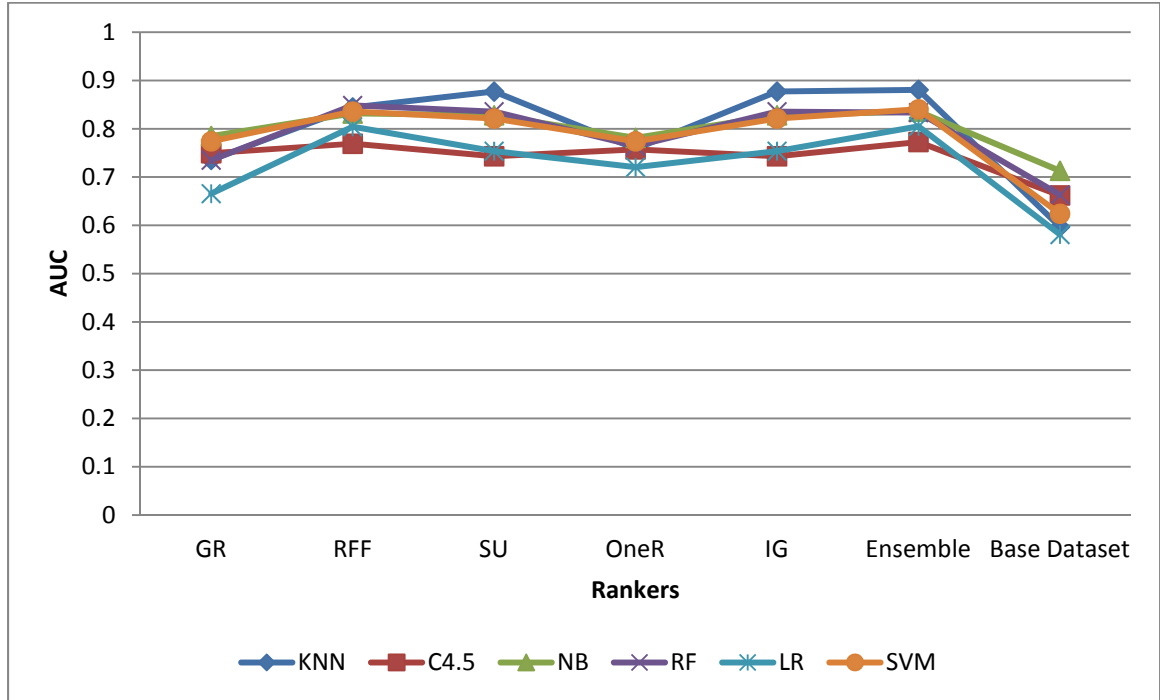


Fig 7.4 Model performance for top 5 features for Lung Cancer dataset



We also perform ANOVA test on the AUC performance metric. ANOVA is an acronym for Analysis of Variance. It is defined as a procedure for assigning sample variance to different sources and making a decision if the variation is within or among different population groups [22]. Samples are described in terms of variation around group means and variation of group means around an overall mean. If variations within groups are small relative to variations between groups, a difference in group means may be inferred. Hypothesis Tests are used to quantify decisions.

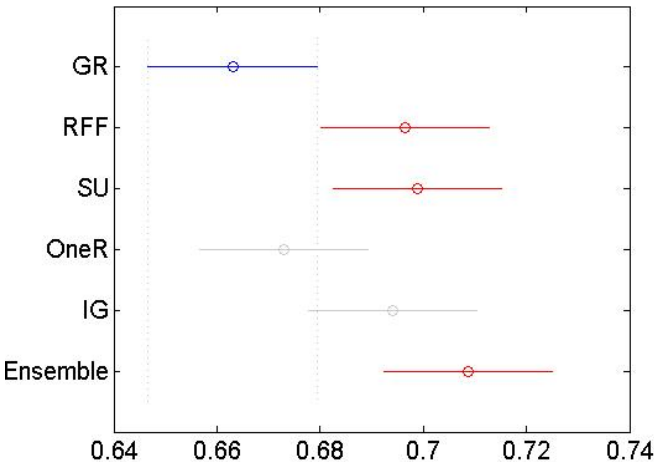
N-way ANOVA determines if the means in a set of data differ when grouped by multiple factors. If they do differ, you can determine which factors or combinations of factors are associated with the difference [22]. N-way ANOVA is a generalization of two-way ANOVA. For three factors, the model can be written

$$y_{ijkl} = \mu + \alpha_j + \beta_i + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijkl}$$

In this notation parameters with two subscripts, such as $(\alpha\beta)_{ij}$, represent the interaction effect of two factors. The parameter $(\alpha\beta\gamma)_{ijk}$ represents the three-way interaction. An ANOVA model can have the full set of parameters or any subset, but conventionally it does not include complex interaction terms unless it also includes all simpler terms for those factors.

A one-way ANOVA is performed in this study. The factor A represents six rankers. In this ANOVA test, the results from all ten-folds were taken into account together. A significance level of $\alpha = 5\%$ was used for all statistical tests. The p-value is 0.004, indicating that the classification performances of six rankers (Factor A) were significantly different from each other. The multiple comparison results are presented in Figure 7.5. The figure shows the following facts: GR performed worst and ensemble approach performed best. OneR, IG, RFF, and SU sit between them and these four rankers are ordered by their performances from worst to best. Ensemble approach performed significantly better than GR and OneR.

Fig 7.5 ANOVA tests on AUC for six rankers



7.5 Case Study 2: Lymphoma Dataset

7.5.1 Experimental results

We also conducted experiments on a high dimensional dataset, Lymphoma dataset which has 4027 features and 96 instances. Sizes of feature subsets are set as 25, 50, 100, 500, and 1000. The results of the experiments are displayed in Table 7.5, 7.6, 7.7, 7.8 and 7.9. The description of tables is similar to the tables of Lung Cancer dataset. The last row of each table represents the results obtained for base dataset. This row can be used as a baseline for comparison.

Table 7.8: AUC values for rankers with top 25 features for Lymphoma dataset

Rankers	KNN	C4.5	NB	RF	LR	SVM	Average
GR	0.8714	0.8170	0.9412	0.9374	0.9523	0.8541	0.8955
RFF	0.8522	0.8360	0.9275	0.9723	0.9741	0.9660	0.9213
SU	0.9366	0.9037	0.9216	0.9758	0.9832	0.9276	0.9414
OneR	0.8357	0.8028	0.9045	0.9699	0.9535	0.9185	0.8974
IG	0.9388	0.8665	0.9260	0.9756	0.9755	0.9411	0.9372
Ensemble	0.9419	0.9303	0.9243	0.9828	0.9819	0.9381	0.9498
Base Dataset	0.8600	0.8920	0.7640	0.9640	0.9820	0.9800	0.9070

Table 7.9: AUC values for rankers with top 50 features for Lymphoma dataset

Rankers	KNN	C4.5	NB	RF	LR	SVM	Average
GR	0.9096	0.8829	0.9529	0.9705	0.9774	0.9656	0.9431
RFF	0.8707	0.8947	0.9221	0.9790	0.9730	0.9646	0.9340
SU	0.9341	0.9113	0.9235	0.9858	0.9943	0.9693	0.9510
OneR	0.8211	0.8192	0.9167	0.9679	0.9703	0.9548	0.9083
IG	0.9509	0.8830	0.9211	0.9636	0.9878	0.9680	0.9457
Ensemble	0.9579	0.9389	0.9261	0.9747	0.9630	0.9537	0.9523
Base Dataset	0.8600	0.8920	0.7640	0.9640	0.9820	0.9800	0.907

Table 7.10: AUC values for rankers with top 100 features for Lymphoma dataset

Rankers	KNN	C4.5	NB	RF	LR	SVM	Average
GR	0.9331	0.9000	0.9273	0.9716	0.9881	0.9769	0.9495
RFF	0.9065	0.9127	0.9174	0.9664	0.9702	0.9632	0.9394
SU	0.9578	0.9083	0.9346	0.9865	0.9968	0.9692	0.9588
OneR	0.8438	0.8748	0.9037	0.9521	0.9800	0.9689	0.9205
IG	0.9643	0.9222	0.9198	0.9771	0.9949	0.9692	0.9579
Ensemble	0.9516	0.9095	0.9336	0.9760	0.9878	0.9638	0.9537
Base Dataset	0.8600	0.8920	0.7640	0.9640	0.9820	0.9800	0.9070

Table 7.11: AUC values for rankers with top 500 features for Lymphoma dataset

Rankers	KNN	C4.5	NB	RF	LR	SVM	Average
GR	0.9516	0.9259	0.8977	0.9530	0.9771	0.9751	0.9467
RFF	0.9289	0.9197	0.8877	0.9763	0.9752	0.9770	0.9441
SU	0.9767	0.9034	0.9074	0.9806	0.9868	0.9773	0.9553
OneR	0.8575	0.8668	0.8677	0.9859	0.9910	0.9590	0.9213
IG	0.9703	0.9092	0.9018	0.9736	0.9959	0.9793	0.9550
Ensemble	0.9279	0.9086	0.8976	0.9636	0.9933	0.9775	0.9447
Base Dataset	0.8600	0.8920	0.7640	0.9640	0.9820	0.9800	0.9070

Table 7.12: AUC values for rankers with top 1000 features for Lymphoma dataset

Rankers	KNN	C4.5	NB	RF	LR	SVM	Average
GR	0.9611	0.9087	0.8734	0.9667	0.9781	0.9770	0.9441
RFF	0.9511	0.9145	0.8878	0.9640	0.9902	0.9781	0.9476
SU	0.9739	0.9085	0.8780	0.9789	0.9518	0.9777	0.9448
OneR	0.8611	0.9153	0.8629	0.9837	0.9358	0.9769	0.9226
IG	0.9576	0.9088	0.8821	0.9597	0.9449	0.9778	0.9334
Ensemble	0.9455	0.9088	0.8836	0.9735	0.9978	0.9777	0.9478
Base Dataset	0.8600	0.8920	0.7640	0.9640	0.9820	0.9800	0.9070

7.5.2 Analysis of Results

The above tables show the following facts:

1. Among the five sizes of feature subset, overall subset with 100 features performed best.
2. For the six classifiers, LR performed best on average while C4.5 performed worst in terms of AUC performance metric.
3. Among the six rankers, on average SU and Ensemble performed best while OneR performed worst.
4. In general, the classification performance is improved even after a significant number of features were removed from the original dataset.

Chapter 8: Conclusions and Feature Research

In this thesis, we have reviewed feature selection and explained the basic concept of different feature selection methods: filter, wrapper and hybrid model. We reviewed four filter based feature ranking techniques and one wrapper based feature ranking technique. They are information gain, gain ratio, symmetrical uncertainty, reliefF and oneRattribute evaluation. We examined classification models that are built using various classification techniques such as naïve bayes, k-nearest neighbor, random forest, support vector machine, logistic regression and decision trees. We took a brief review of the evaluation criteria used to evaluate the classification models. We have also introduced ensemble methods for feature ranking technique that can help build stable and robust classification models.

The experimental results showed that the performance of rankers may be significantly influenced by learner used in the classification. This study proposed and investigated ensemble technique with a unique combining method using rank ordering of features. The ensemble method used in our study is frequency count which is accompanied by mean to resolve any frequency count collision. The first step is to identify rankers that form a set of ranking lists and then select a fixed number of features from every ranking list. The second step is to count the occurrence of individual feature in all the ranking lists. This would be the frequency of each feature. Then we sort the features based on the frequency. The probable chances of features having the same frequency are high. To resolve the issue of frequency collision, we have introduced mean ordering. The experiments were conducted on two biomedical datasets. The results demonstrated that the ensemble technique performed better overall than any individual

ranker. The results also showed that the performances of classification models are improved even after 90% of the features are removed.

Future work will involve experiments on the datasets from different domains. The ensemble algorithm will be tested on more datasets with different backgrounds. The difference in performance and accuracy of different ensemble approaches will be evaluated. Statistical analysis tests can be extended to different tests. ANOVA tests will be performed on individual fold values for each classifier.

At present our thesis has mainly concentrated on filter based feature ranking techniques. In the future we would like to explore different approaches such as feature subset selection techniques and its applicability to our ensemble approach.

REFERENCES

- [1] J. Jackson, "Data Mining: A Conceptual Overview", Communications of the Association for Information Systems (Volume 8, 2002), pages 267- 296.
- [2] H. Wang, T. M. Khoshgoftaar, K. Gao, "Ensemble feature selection technique for software quality classification", Proceedings of the 22nd International Conference on Software Engineering & Knowledge Engineering, Redwood City, San Francisco Bay, CA, USA, July 1 - July 3, 2010, pages 215-220.
- [3] H. Wang, T. M. Khoshgoftaar, A. Napolitano, "A comparative study of ensemble feature selection techniques for software defect prediction", Proceedings of the Ninth International Conference on Machine Learning and Applications, Washington, DC, USA, December 12-14, 2010, pages 135-140.
- [4] H. Liu, H. Motoda, R. Setiono, Z. Zhao, "Feature Selection: An Ever Evolving Frontier in Data Mining", JMLR: Workshop and Conference Proceedings 2010, Volume: 4, Publisher: Citeseer, pages 4-13.
- [5] L. Yu, H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution", Proceedings of the Twentieth International Conference on Machine Learning, ICML-03, Washington, D.C., August, 2003, pages 856-863.
- [6] W. Altidor, T. M. Khoshgoftaar, J. Van Hulse, A. Napolitano, "Ensemble feature ranking methods for data intensive computing applications", Handbook of data intensive computing, Springer Science + Business media, LLC 2011, pages 349 -376.
- [7] Asha G. K, A. S. Manjunath, M. A. Jayaram, "A Comparative Study of Attribute Selection Using Gain Ratio and Correlation Based Feature Selection",

International Journal of Information Technology and Knowledge Management, July – December 2012, Volume 2, pages 271 – 277.

[8] Y. Chen, Y. Li, X. Cheng, L. Guo, H. Lipmaa, M. Yung, D. Lin, "Survey and Taxonomy of Feature Selection Algorithms in Intrusion Detection System", Inscrypt 2006, LNCS 4318 Springer-Verlag, Berlin, 2006, pages 153 – 167.

[9] D. Ienco, R. G. Pensa, R. Meo, "Context-based Distance Learning for Categorical Data clustering", IDA 2009, LNCS 5772, Springer, Berlin, 2009, pages 83 – 94.

[10] J. C. Platt, "A Fast Algorithm for Training Support Vector Machines", Technical Report MSR-TR-98-14, April 21, 1998, John Platt Microsoft Research.

[11] K. Deng, "OMEGA: On-line Memory-based general purpose system detector", Carnegie Mellon University, Chapter 4, pages 59-75.

[12] B. Clarke, E. Fokoué, H. H. Zhang, "Principles and Theory for Data Mining and Machine Learning", Chapter 5, Supervised Learning: Partition Methods, Springer Series in Statistics, 2009, pages 231- 254.

[13] S. R. Joelsson, J. A. Benediktsson, J. R. Sveinsson, "Random Forest Classification of Remote Sensing Data", Signal and Image Processing for Remote Sensing, Volume 978 Issues 0-5097, pages 327-338.

[14] M. W. Kim, J. W. Ryu, "Optimized Fuzzy Classification for Data Mining", 9th International Conference Y. Lee et al. (Eds): DASFAA 2004, LNCS 2973, pages 582 – 593.

[15] H. He, W. Graco, X. Yao, "Application of Genetic algorithm and k-nearest neighbor method in medical fraud detection", Simulated Evolution and Learning,

Second Asia-Pacific Conference on Simulated Evolution and Learning, SEAL' 98
Canberra, Australia, Springer, November 1998, pages 74 – 81.

[16] M. Narasimha Murty, V. Susheela Devi, “Pattern Recognition: An Algorithmic approach”, Springer, Chapter 4, pages 86 -97.

[17] O. A. Omitaomu, “Lecture Notes in Data Mining”, Chapter 4, World Scientific publishing Co, 2006, pages 39 – 51.

[18] S. Pulatova, “Covering (rule-based) algorithms Lecture Notes in Data Mining”, World Scientific publishing Co, 2006, pages 87-97.

[19] T. Fawcett, “ROC Graphs: Notes and Practical Consideration for Researchers”, HP Laboratories, March 16, 2004, Kluwer Academic Publishers, pages 1 -38.

[20] S. S. Qian, “Environmental and Ecological Statistics with R”, Simulation for Model checking and Statistical Inference, Chapter 9, CRC Press 2010, pages 334-335.

[21] Y. Saeys, T. Abeel, Y. V. Peer, “Robust feature selection using ensemble feature selection techniques”, W. Daelemans et al. (Eds.): ECML PKDD 2008, Part II, LNAI 5212, pages 313–325.

[22] ANOVA tests resource: www.mathworks.com and www.wikipedia.com

[23] Y. Wang, F. Makedon, “Application of ReliefF feature filtering algorithm to selecting informative genes for cancer classification using microarray data”, Computational systems bioinformatics conference, 2004 IEEE, pages 497 – 498.

[24] R. P.Schumaker, O. K. Solieman, H. Chen, “Sports Data Mining”, Chapter 9, “Open source data mining tools for sports”, Springer 2010, pages 89-90.

[25] I. H. Witten, E. Frank, M. A. Hall, “Data Mining: Practical Machine Learning Tools and Techniques”, Third edition, January 2011.

- [26] Kent Ridge bio-medical repository - <http://datam.i2r.a-star.edu.sg/datasets/krbd/>
- [27] T. D. Lemmond, B. Y. Chen, A. O. Hatch, W. G. Hanley, “An Extended Study of the Discriminant Random Forest”, Chapter 6, “Data Mining: A Special Issue in Annals of Information Systems”, Springer Science, LLC 2010.
- [28] L. Rokach, B. Chizi, O. Maimon, “Feature selection by combining multiple methods”, Advances in Web Intelligence and Data Mining, 2006, pages 295–304.
- [29] J. T de Souza, N. Japkowicz, S. Matwin, “Stochfs: A framework for combining feature selection outcomes through a stochastic process”, Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, 2005, pages 667–674.
- [30] J. O. S. Olsson, D. W. Oard, “Combining feature selectors for text classification”, CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management, New York, NY, USA, 2006, pages 798–799.
- [31] J. R. Quinlan, “Induction of decision trees”, Machine Learning, Kluwer Academic Publishers, 1986, pages 81-106.

