Feature Selection in High Dimensional Spaces

Ghazaal Sheikhi

Submitted to the
Institute of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Engineering

Eastern Mediterranean University
September 2020
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Prof. Dr. Ali Hakan Ulusoy
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Doctor of Philosophy in Computer Engineering.

Prof. Dr. Hadi Işık Aybay
Chair, Department of Computer Engineering

We certify that we have read this thesis and that in our opinion it is fully adequate in scope and quality as a thesis for the degree of Doctor of Philosophy in Computer Engineering.

Prof. Dr. Hakan Altınçay
Supervisor

Examining Committee

1. Prof. Dr. Aydın Alatan ______________________
2. Prof. Dr. Hakan Altınçay ______________________
3. Prof. Dr. Tolga Çiloğlu ______________________
4. Prof. Dr. Ekrem Varoğlu ______________________
5. Asst. Prof. Dr. Ahmet Ünveren ______________________

ABSTRACT

In this study, two novel filter feature selection approaches are proposed as alternatives to state-of-the-art. The first proposed approach is a greedy-based feature selection method where redundancy is replaced by diversity to quantify the complementarity of a candidate feature with respect to the already selected subset. Both relevance and diversity are computed in terms of the ranks of positive instances, which is analogous to the computation of the area under the receiver operating characteristic curve (AUC). In the second approach, a novel dissimilarity metric based on Feature-to-Feature (F2F) scatter frequencies is proposed for clustering-based filter feature selection. The proposed metric is computed by obtaining feature-dependent ranks of samples and identifying the features which assign close ranks to each sample. Samples are represented as a set of affinity sets containing features having rank differences within a predefined proximity window size. The F2F dissimilarity of a pair of features is computed using the frequency of their appearance in different affinity sets. Features are then clustered into distinct groups using F2F dissimilarity metric. From each cluster, the feature having the highest relevance score is selected. The experiments conducted on 10 UCI and microarray gene expression data sets have confirmed that the proposed feature selection approaches provide better performance scores when compared to other competing methods. The proposed method outperforms the widely-used mutual information-based schemes in terms of classification accuracy, AUC and stability.

Keywords: feature selection, ranks of instances, relevance, diversity, dissimilarity, scatter frequency, representative feature.

ÖZ

Bu çalışmada, en son teknolojiye alternatif olarak iki yeni öznitelik yaklaşımı seçme önerilmiştir. Önerilen ilk yaklaşım, seçilmiş olan alt kümeye göre bir aday özniteliğin tamamlayıcılığını ölçmek için artıklığı çeşitleme ile değiştiren özyineli bir öznitelik seçim yöntemidir. Hem ilgililik hem de çeşitlilik, alıcı çalışma karakteristik eğrisi (AUC) altındaki alanın hesaplanmasına benzer olan pozitif örneklerin sıralarına göre hesaplanır. İkinci yaklaşımda, kümeleme tabanlı filtre öznitelik seçimi için öznitelikler arası (F2F) dağılım frekanslarına dayanan yeni bir benzemezlik metriği önerilmektedir. Önerilen metrik, özniteliğe bağlı örnek grupları elde edilerek ve her bir örneğe yakın düzeyler atanan özniteliklerin tanımlanmasıyla hesaplanır. Örnekler, önceden tanımlanmış bir yakınlık penceresi boyutu içinde sıra farklılıklarına sahip öznitelikler içeren bir yakınlık kümesi olarak temsil edilir. Bir çift özniteliğın F2F benzemezliği, farklı benzesim kümelerinde görünümlerinin sıklığı kullanılarak hesaplanır. Öznitelikler daha sonra F2F benzemezlik metriği kullanılarak farklı gruplara kümelenir. Her kümeden, ilgililik düzeyi en yüksek olan öznitelik seçilir. 10 UCI ve mikrodizi gen ekspresyon veri setleri üzerinde yapılan deneyler, önerilen öznitelik seçim yaklaşımlarının diğer rakip yöntemlere kıyasla daha iyi performans skorları sağladığını göstermiştir. Önerilen yöntem, sınıflandırma doğruluğu, AUC ve kararlılık açısından yaygın olarak kullanılan karşılıklı bilgi tabanlı tekniklerdan daha iyi performans göstermektedir.

Anahtar Kelimeler: öznitelik seçimi, örnek sıraları, ilgililik, çeşitlilik, farklılık, benzemezlik dağılım frekansı, temsilcisi öznitelik.

To my Parents

FARZANEH AND REZA

ACKNOWLEDGMENT

I would like to first and foremost express my gratitude to my PhD thesis supervisor Prof. Dr. Hakan Altınçay for all the support and encouragement he gave me. He is the best adviser and a better PhD supervisor beyond my imagination. I humbly admit that without his insightful guidance, endless patience and immense knowledge, this PhD would not have been achievable.

I gratefully acknowledge my follow-up jury committee members, Prof. Dr. Ekrem Varoğlu and Asst. Prof. Dr. Ahmet Ünveren for their time, interest, feedbacks and helpful comments.

I would like to say a very big thank you to my friend Dr. Payam Zarbakhsh for his empathy, moral support and sense of humor that helped me overcome the difficulties.

Many thanks also to my friends John and Barbara Grundey for being a source of friendship and for their willingness to help me in polishing and proofreading this dissertation.

My sincere thank goes to my respected friend, Mr. Majid Mohseni for his friendship, and for the kind advices he gave me.

The last but not the least, my deepest gratitude goes to my family, my sisters Mehrnoush and Setareh and my parents Frazaneh and Reza for their unconditional love, caring and encouragements.

TABLE OF CONTENTS

ABSTRACT iii

ÖZ iv

DEDICATION v

ACKNOWLEDGMENT vi

LIST OF TABLES ix

LIST OF FIGURES xi

LIST OF SYMBOLS AND ABBREVIATIONS xiii

1 INTRODUCTION 1

1.1 Background 1

1.2 Problem Definition 2

1.3 Motivation 3

1.4 Contributions 4

1.5 Outline 5

2 RELATED WORK 6

2.1 Introduction 6

2.2 Overview of Filter Methods 6

2.2.1 Mutual Information-based Feature Selection Methods 8

2.2.2 Clustering-based Feature Selection Methods 11

2.3 Conclusions 13

3 PROPOSED FEATURE SELECTION METHODS 14

3.1 MRMD 15

3.1.1 Relevance Measure 15

3.1.2 Diversity Measure 19

3.1.3 MRMD Algorithm 20

3.1.4 Artificial Example 21

3.2 Clustering-based Feature Selection using F2F Scatter Frequencies 28

3.2.1 Formulation of F2F Scatter Matrix 29

3.2.2 Artificial Example 31

3.2.3 Extension to Multi-class 34

3.2.4 Clustering-based Feature Selection using F2F Measure 34

3.3 Conclusions 37

4 EXPERIMENTAL RESULTS 38

4.1 Data Sets 38

4.2 Experimental Setting 39

4.3 Experimental Results of MRMD 40

4.4 Experimental Results of Clustering-based Feature Selection using F2F Matrix 52

4.4 Discussions 59

5 CONCLUSIONS AND FUTURE WORK 63

5.1 Conclusions 63

5.2 Future Work 65

REFERENCES 66

LIST OF TABLES

Table 3.1: Example decision values and related terms 17

Table 3.2: Artificial data set... 23

Table 3.3: Ranks of positive instances and relevance values.................................... 23

Table 3.4: Differences of ranks of positives with respect to $x1$ and diversity values 24

Table 3.5: Distribution of positive and negative instances in bins for each feature .. 27

Table 3.6: The sum of ranks of positives and validity of Equation 3.10 32

Table 3.7: The affinity sets extracted from the toy example. (The empty sets are discarded) 33

Table 4.1: Datasets used in the experiments 38

Table 4.2: Average accuracy of classifiers (in percentage) across all feature subsets 41

Table 4.3: Average AUC of classifiers (in percentage) across all feature subsets 42

Table 4.4: Average AUC and ACC (in percentage) of classifiers on different feature subset groups (in percentage).. 51

Table 4.5: Average accuracy of classifiers (in percentage) across all feature subsets 55

Table 4.6: Average AUC of classifiers (in percentage) across all feature subsets 56

Table 4.7: Average AUC and ACC (in percentage) of classifiers on different feature subset groups.. 57

Table 4.8: Average accuracy and AUC of classifiers (in percentage) across all data sets and all feature subsets 58

Table 4.9: Average accuracy of classifiers for each data set across all feature subsets (comparing the best performing methods) 60

Table 4.10: Average AUC of classifiers for each data set across all feature subsets (comparing the best performing methods) 61

Table 4.11: Average accuracy and AUC of classifiers across all data sets and across all feature subsets for 3NN and SVM (comparing the best performing methods)..... 62

LIST OF FIGURES

Figure 2.1: The general framework of clustering-based feature selection methods [52] 12

Figure 3.1: Scatter plot of samples in $(x1, x2)$ subspace 25

Figure 3.2: Scatter plot of samples in $(x1, x3)$ subspace 26

Figure 3.3: Scatter plot of samples in $(x1, x4)$ subspace 26

Figure 4.1: Average classification accuracy (in percentage) of SVM and 3-NN on Sonar data set 44

Figure 4.2: Average classification accuracy (in percentage) of SVM and 3-NN on Urban Land Cover data set 44

Figure 4.3: Average classification accuracy (in percentage) of SVM and 3-NN on Musk data set 45

Figure 4.4: Average classification accuracy (in percentage) of SVM and 3-NN on Colon data set 45

Figure 4.5: Average classification accuracy (in percentage) of SVM and 3-NN on Lymphoma data set 46

Figure 4.6: Average classification AUC (in percentage) of SVM and 3-NN on Sonar data set 47

Figure 4.7: Average classification AUC (in percentage) of SVM and 3-NN on Urban Land Cover data set 47

Figure 4.8: Average classification AUC (in percentage) of SVM and 3-NN on Musk data set 48

Figure 4.9: Average classification AUC (in percentage) of SVM and 3-NN on Colon data set 48

Figure 4.10: Average classification AUC (in percentage) of SVM and 3-NN on Lymphoma data set 49

Figure 4.11: Box plots of stability scores obtained for different filters 50

Figure 4.12: Stability measure for different feature selection methods. 59

LIST OF SYMBOLS AND ABBREVIATIONS

|.|	Cardinality operator
∩	Intersection operator
10-CV	Ten Fold Cross Validation
3NN	3 Nearest Neighbors
ACC	Accuracy
AUC	Area Under Receiver Operator Curve
CMIM	Conditional Mutual Information Maximization
Cov(.)	Covariance
d	Number of features in the data set
D(.)	Diversity
d_{ij}	Scatter frequency of feature *i* and *j*
DISR	Double Input Symmetrical Relevance
F	Feature set
F	Feature subset
F2F	Feature-to-Feature (scatter frequency)
H(,)	Entropy
HC	Hierarchical Clustering
I(.)	Mutual information
J(.)	Objective function
JMIM	Joint Mutual Information Maximization
K	Number of samples
MI	Mutual Information
MRMD	Maximum-Relevance and Maximum-Diversity

mRMR	Minimum Redundancy-Maximum Relevance
N	Number of negative samples
n	A negative sample
NJMIM	Normalized Joint Mutual Information Maximization
P	Number of positive samples
p	A positive sample
PAM	Partitioning Around Medoids
R	Rank matrix
$\tilde{R}$	Initial rank matrix
$R(.)$	Relevance
R_i	Relevance of feature *i*
${r^p}_k$	Rank of sample p in feature *i*
S	Data set
SI	Stability Index
s_i	Sample *i*
SU	Symmetric Uncertainty
SVM	Support Vector Machines
${v^m}_k$	Affinity set *m* extracted from sample *k*
x_i	Feature *i*
y	Class label
κ	Size of feature subset
ρ	Pearson’s correlation coefficient
ρ_r	Spearman’s rank correlation coefficient
σ	Standard deviation

Chapter 1

INTRODUCTION

1.1 Background

Recent advances in data acquisition and storage have generated huge amount of data in a wide range of domains such as bioinformatics [1], [2], computer vision [3], text categorization [4], [5] and natural language processing [6]. The resulting high dimensional data sets pose big challenges in the very first stage of data analysis, information retrieval, clustering, classification, data mining, and decision making. To address the issue, dimensionality reduction approaches namely, feature extraction and feature selection are applied [7]. Feature extraction approaches such as principal component analysis (PCA) or linear discriminant analysis (LDA) project the original feature space into a lower dimensional space. A drawback of these methods is that features in the new subspace are not interpretable for the domain experts [8], [9]. Unlike feature extraction techniques which transform the feature space, feature selection methods preserve the original feature space by selecting a compact subset of discriminative attributes. In general, feature selection not only reduces the computational load of underlying machine learning algorithm, but also enhances the performance of the classifier [10], [11].

There are various feature selection methods proposed in the literature that are generally grouped into three main categories as wrapper, embedded and filter approaches [12]. Wrapper methods are classifier-dependent and the attributes are selected to minimize a specific classifier's predictive error. The weaknesses of this category of feature selection methods are the computational overhead and over-fitting [12]. In embedded methods, feature selection and classification are interlocked in the learning algorithm. Embedded feature selection methods are efficient but the excessive computational burden restricts their application in high dimensions. They are also known to suffer from limited generalization capability [2]. On the other hand, in filter approaches, the evaluation criterion is totally independent from the learning algorithm. In particular, the features are evaluated in terms of their individual discriminative powers and pairwise redundancies [13]–[15]. Filter feature selection methods are generally characterized by scalability, low computational complexity, and high levels of generalization [8], [12]. Because of these reasons, they are more suitable in high dimensional spaces than wrappers and embedded methods [16], [17].

1.2 Problem Definition

Filter feature selection algorithms are generally distinguished by the relevance and redundancy metrics. The relevance of a feature quantifies its individual ability to predict the class labels. Univariate feature selection methods only rely on the relevance of features [18] and thus fail in addressing the redundancy among features [19]. In multivariate feature selection methods, similarity (or, dissimilarity) of features is also considered to address redundancy in the selected subset. Vast majority of feature filtering methods employ information theory-based metrics, namely mutual information (MI) for both relevance and redundancy. However, these metrics suffer from some inherent presumptions [20]. Continuous features are required to be discretized so that the samples are grouped into bins. These approaches are criticized for being negligent about the orders of the samples in the bins which can lead to loss of information, specifically when the number of samples is small [21]. Moreover, reliable estimation of mutual information is challenging in high dimensional spaces [13]. There is an indisputable fact that feature selection is principally aimed at improving the classification performance while reducing the dimensionality of the feature space. For the case of MI-based feature selection methods, given a specific value for MI, just lower and upper bounds for Bayesian error rate (or classification accuracy) can be defined [22]. Hence, it can be argued that these MI-based feature selection approaches are susceptible to fail in ranking the features for optimized classification performance.

1.3 Motivation

Alternative relevance measures based on area under the ROC curve (AUC) which can be computed using ranks of instances has been considered very effective in estimating the relevance of features [18], [19], [23]. For instance, the feature assessment by sliding threshold algorithm directly exploits the classification performance metric, AUC, as a measure of relevance [23]. However, it fails to address redundancy due to being univariate. Spearman's rank correlation coefficient is another rank-based measure used for quantifying the redundancy between different features by considering the dissimilarities in the ranks assigned to the training samples [18]. As a matter of fact, the rank-based metrics have been strongly advocated in data analysis for discovering associations among features [19], [24], [25]. Rank-based measures are of high reliability as ranks are less sensitive to outliers and measurement noise [24]. Estimating dissimilarity of the features using ranks is known to be almost independent from probability distributions and capable of capturing nonlinear relationships among features [26]. It should be noted that adapting an objective function for feature selection entails compatible relevance and redundancy metrics. Having adapted consistent relevance and redundancy metrics, there are two general strategies for searching the feature space [27]–[29]. The first technique is the conventional forward search or generally speaking, greedy search which has a long history in feature selection [30], [13], [14], [27]. The second one is the clustering-based feature selection which recently attracted the interest of researchers [31]–[34].

These facts motivated us to develop two feature selection methods using novel rank-based relevance and redundancy (more precisely diversity or dissimilarity) measures. AUC is exploited as the relevance measure and computed using ranks of positive instances. Two compatible rank-based metrics are adapted to measure 1) diversity of features in the proposed greedy-based method and 2) dissimilarity of features in the proposed clustering-based method.

1.4 Contributions

In this study, two novel multivariate filter approach are proposed. The first is a greedy-based that is method emerged from the widely-used performance measure for classification, namely area under receiver operating characteristic curve (AUC). The algorithm is based on the ranks of positive instances which is motivated by the fact that AUC can be determined by those ranks. Additionally, for a given pair of features, differences in ranks of the positive instances is considered as an indicator of complementarity between these features, that is highly valuable for classification. The proposed score includes two terms: *relevance* and *diversity*. The former estimates the discriminative ability of an individual feature while the latter determines its complementarity to the already selected subset. The maximum-relevance and maximum-diversity (MRMD) feature selection algorithm proposed in this study aims to maximize the sum of these two.

The second proposed method utilizes a novel rank-based dissimilarity metric that employs feature-to-feature scatter frequencies for clustering-based feature selection. Firstly, the local feature similarity information in each sample is captured by applying a proximity window on the ranks assigned by different features. More specifically, using the differences of the ranks and the size of proximity window, feature affinity sets of each sample are computed. For instance, if the ranks assigned to a sample by two different features are close, the features are expected to co-occur in an affinity set of that sample. The size of the proximity window can be adjusted based on data characteristics such as the number of samples. Using the affinity sets, the Feature-to-Feature (F2F) scatter frequencies of all feature pairs are calculated in the following phase to define the F2F dissimilarity matrix that represents global dissimilarity between each feature pair. Having utilized F2F dissimilarity matrix for clustering features, a representative feature is selected from each cluster.

1.5 Outline

The rest of this thesis is organized as follows. In Chapter 2, related studies are reviewed. Proposed methods are described in Chapter 3. Chapter 4 represents experimental results. The thesis is concluded in Chapter 5.

Chapter 2

RELATED WORK

2.1 Introduction

Feature selection methods are generally categorized as filter methods, wrapper methods and embedded methods [12]. Filter approaches rank features by using a predefined measure. These methods do not interact with the classifier and thus have the lowest computational complexity. Nevertheless, filter approaches do not take into account the level of interactions between the features and the classifier. Wrappers are based on evaluating the predictive performance of different feature subsets for a given classifier. In other words, these approaches compute the best-fitting feature set for the selected classifier. They may suffer from over-fitting. Wrappers are also the most computationally demanding ones. On the other hand, embedded methods employ an internal learning algorithm to find the optimal subset of features. Although they suffer less form computational complexity compared to wrappers, they do not have an acceptable generalization capability [12]. Obviously, a feature selection method needs to be efficient and at the same time simple and fast, specifically when it is supposed to handle high dimensional data sets [2].

2.2 Overview of Filter Methods

Filter feature selection methods typically perform by searching the feature space to optimize an objective function that is based on the relevance and the redundancy of features [13], [14], [27]–[29]. To solve the corresponding combinatorial problem, conventional greedy search methods such as forward and backward selection are generally used. Forward selection is the most popular algorithm which sequentially ranks the features according to a predefined quality measure [7], [13], [28], [29], [35]. Greedy search-based feature selection approaches are very effective in high dimensional feature spaces. A vast majority of greedy-based methods have utilized information theory-based measures to determine the objective function [14], [27]–[29], [36]–[41].

Feature selection based on mutual information has been criticized for some of its inherent presumptions [13], [20]. Firstly, continuous features which comprise majority of real world data sets are required to be discretized for entropy calculation and it is generally challenging to identify the best discretization method [21]. Discretization may lead to loss of information when simple unsupervised discretisers such as equal-width or equal frequency are used [21]. On the other hand, finding the optimal supervised discretization technique is NP-complete [42], which may highly increase the computational complexity.

Greedy-based search methods although very popular suffer from nesting problem [27], [43]. More precisely, each step of the search algorithm in estimating feature interactions highly relies on the result of previous steps. In the *k*th iteration of forward search, the algorithm searches for the next best feature with respect to the already selected subset of (*k-1*) features. Thus, the search strategy is rather atomistic than holistic, generating nested ranks of features. Moreover, forward or backward selection algorithms perform poorly for non-monotonic objective functions which is almost always the case for filter feature selection methods [43]. These issues can lead to a sub-optimal solution [27], [43]. It should be noted that floating search, an alternative for forward or backward search, addresses these issues to some extent but at the cost of increased computational load.

An alternative to greedy search-based feature selection is clustering-based one, which has recently attracted the interest of researchers in pattern recognition and data mining [31]–[34], [44]. The general framework of clustering-based feature selection is to firstly group the attributes into a set of distinct clusters including highly correlated features and then select a representative feature from each cluster. The key advantage of clustering-based approaches over greedy search methods is that features are grouped in a holistic manner as the similarity or dissimilarity measures are estimated through a global scheme [31]. The asset of this holistic approach in estimating feature interactions is two-fold. It not only addresses the nesting problem arisen by greedy search, but also benefits the domain experts for further investigations. In other words, by grouping features, salient patterns of associations among them are discovered, allowing for better interpretation of the results [7], [17], [45]. Cluster analysis is in fact one of the most widely-used approaches for data analysis, visualization and inspection, specifically in exploratory analysis of high dimensional gene expression data sets [7], [46]. Moreover, feature selection methods which employ clustering have been proved to be more stable than greedy search-based approaches [47], [48].

2.2.1 Mutual Information-based Feature Selection Methods

Information theory has been the keystone of filter approaches for feature selection in recent decade [14], [27]–[29], [36]–[41]. The mutual information (MI) between two variables is widely used as an effective measure for correlation. More specifically, MI is defined as the reduction in the level of uncertainty of the dependent variable, given an independent variable. Assuming two variables x and y, the mutual information between them is calculated as

$$I(x;y) = H(y) - H(y|x), \tag{2.1}$$

where $H(.)$ is entropy of the variable and $H(y|x)$ denotes the conditional entropy. Typically, discriminative potential of a feature, that is also named as relevance, is approximated as the mutual information between the feature and the class label. In addition, mutual information between the feature and the previously selected feature subset is considered as the redundancy measure. The basic paradigm of MI-based feature selection methods is to maximize an objective function that is defined as the difference of relevance and redundancy [13].

In recent years, several variations of MI-based feature selection methods have been proposed. The well-known minimum redundancy-maximum relevance (mRMR) feature selection method maximizes the objective function that is defined as [40]

$$J(x_k) = I(x_k;y) - \frac{1}{|\Gamma|}\sum_{x_j \in \Gamma} I(x_k;x_j), \tag{2.2}$$

where x_k is the k^{th} candidate feature, y is the class label, Γ is the already selected subset and $|.|$ is the cardinality operator.

The first term quantifies the relevance of x_k whereas the second represents its redundancy. It is argued that cumulative averaging of redundancies may result in overestimation of the significance of the features which are correlated with a fewer number of features in selected subset [28]. Conditional mutual information maximization (CMIM) has been proposed to address this problem [49]. CMIM ensures that the selected features are both individually discriminative and weakly dependent in a pairwise manner. The objective function to be maximized in CMIM is defined as

$$J(x_k) = \min_{x_j \in \Gamma}(I(x_k; y|x_j)). \tag{2.3}$$

On the other hand, the concept of variable complementarity [50] suggests another definition for the relevance as *the predictive information a feature adds on to the already selected subset*. Variable complementarity has been the motivation for the double input symmetrical relevance (DISR) method proposed by Meyer and Bontempi [51]. The objective function is the cumulative sum of the pairwise joint symmetrical relevances. In fact, it searches for the maximal value of

$$J(x_k) = \sum_{x_j \in \Gamma} \frac{I(x_k, x_j; y)}{H(x_k, x_j; y)}, \tag{2.4}$$

where $I(x_k, x_j; y)$ is a measure of relevance in the context of selected feature subset, Γ, that is calculated as

$$I(x_k, x_j; y) = I(x_j; y) + I(x_k; y|x_j). \tag{2.5}$$

Variable complementarity has also been the motivation for Bennaser et. al [28] who have proposed a maximum of minimum approach for feature selection. Their method has two variants named joint mutual information maximization (JMIM) and normalized joint mutual information maximization (NJMIM). The criterion for JMIM is defined as

$$J(x_k) = \min_{x_j \in \Gamma}(I(x_k, x_j; y)). \tag{2.6}$$

Obviously, JMIM can be viewed as a modification of CMIM that is made by adding the standard relevance term to the objective function. NJMIM is very similar to

JMIM except that it is defined using *normalized* joint mutual information given in Equation 2.7. It should also be noted that NJMIM is a modification of DISR where accumulative sum is replaced by minimum operator as

$$J(x_k) = \min_{x_j \in \Gamma}\left(\frac{I(x_k, x_j; y)}{H(x_k, x_j; y)}\right). \tag{2.7}$$

More recently, there have been ongoing attempts to modify MI-based filter approaches aiming at improved classification performance using linear and nonlinear classifiers. Wang et. al. [27] have conducted a comprehensive set of experiments on a wide range of data sets to prove that their maximum-relevance maximum-independence (MRI) method is superior to the state-of-the-art MI-based and non-MI-based filter approaches. Dynamic relevance and joint mutual information maximization (DRJMIM) is another variant of MI-based filter approaches, affirmed to function competitively in comparison to standard feature selection methods [29]. Nonetheless, no matter how efficiently these approaches are claimed to perform, the foundation of all of them is mutual information.

2.2.2 Clustering-based Feature Selection Methods

Clustering-based feature selection methods are characterized by the dissimilarity metric, clustering algorithm and relevance metric. The general framework of these techniques is as illustrated in Figure 2.1. In recent years, these methods have attracted the interest of researchers as an alternative to conventional greedy search-based methods which suffer from nesting problems. For instance, FAST is a clustering-based feature selection technique which applies minimum spanning tree to create feature clusters [31]. The similarity of features is estimated by pairwise symmetrical uncertainty (SU). From each cluster, a feature with maximum SU value with respect to class labels is selected.

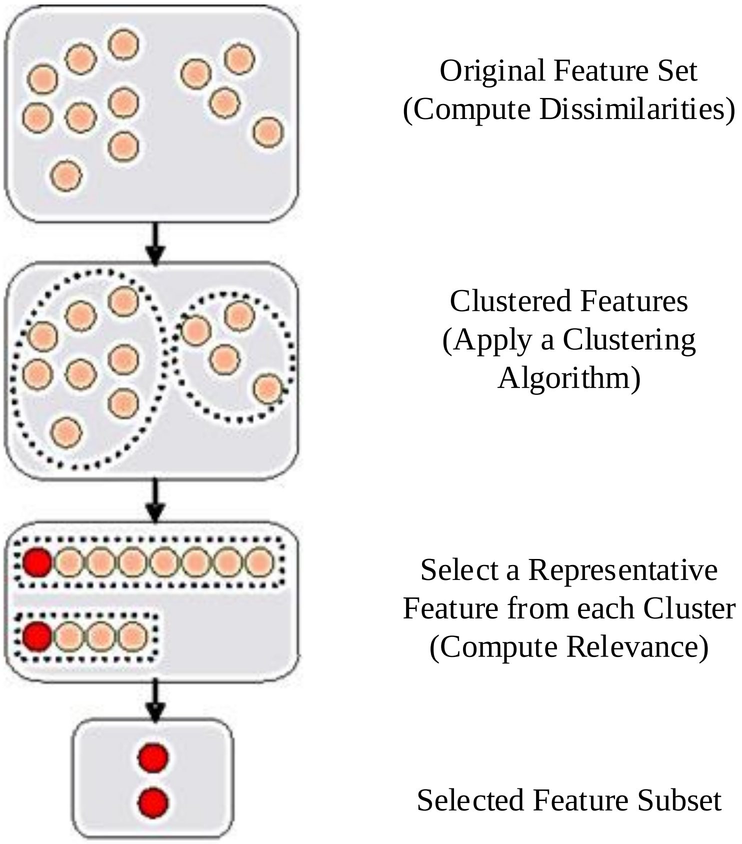

Figure 2.1: The general framework of clustering-based feature selection methods [52]

Yu *et al.* have investigated utilizing distinct peaks in the distributions of feature values [47]. They suggest that features close to the core region of the density function are highly correlated, forming a feature group. F-statistic is then used as the relevance measure to discard irrelevant groups and consequently select the representative features from the remaining groups. Relevance of the features in each group is obtained as the average relevance across the members of the group. Sotoca and Pla employed hierarchical clustering (HC) to group features using a similarity metric based on the conditional mutual information (CMI) [43]. From each cluster, the feature which has the maximum mutual information (MI) value with respect to the class labels is used to form the selected subset.

Another clustering-based feature subset selection (CFSS) is proposed in [34] which employs agglomerative HC for clustering and MI as similarity metric. The clustering algorithm starts by assuming all features as distinct clusters and then merging them in a bottom-up manner until a preselected number of clusters are obtained. MI between each feature and class labels is then used to select the most relevant feature from each cluster. Feature clustering is also claimed to improve the performance of the well-known support vector machine recursive feature elimination (SVM-RFE) algorithm in classifying gene expression data sets [33]. Feature clustering by SVM-RFE performs by *k*-means clustering of genes using Euclidean distance. Representative gene of each cluster is the closest one to the cluster center. These representative features are then ranked using SVM-RFE. In [44], clusters are formed by a community detection algorithm which employs Pearson's correlation coefficient. Having formed the feature correlation network from each subgraph of the network, the feature corresponding to the maximum MI with the class labels is added to the feature subset.

2.3 Conclusions

Filter feature selection plays a critical role in a wide range of applications of machine learning and pattern recognition, specifically in high dimensional spaces. The review of related literature reveals the potential drawbacks of popular MI-based feature selection methods and suggests that there is still a huge room for studying alternative feature selection approaches. It can be argued that performance improvements can be achieved by considering novel relevance and redundancy measures as well as alternative searching strategies.

Chapter 3

PROPOSED FEATURE SELECTION METHODS

Two of the most widely-used metrics to assess the predictive performance of learning algorithms are accuracy and AUC. AUC is defined as the area under the curve of true positive rate versus false positive rate. The intuitive interpretation of the AUC is the probability that a randomly chosen positive sample is ranked higher than a randomly chosen negative one [53]. Although accuracy is known as the conventional classification performance metric in machine learning and pattern recognition, since its shortcomings were highlighted, scholars have convinced that AUC is a more reliable measure [54]. Accuracy depends on the decision threshold and it ignores the way samples above or below the threshold are ranked. On the other hand, AUC is independent from the decision threshold and it considers the ranking of the output scores given by the learning algorithm. Unlike accuracy, AUC is also insensitive to the class distribution which makes it the most-widely used classifier evaluation measure in the case of imbalanced data sets [55], [56].

Inspired by the fact that AUC can be computed using ranks of positive instances and the effectiveness of rank-based measures in data analysis, we propose two feature selection methods named as maximum relevance and maximum divesity of ranks of positives (MRMD), and clustering-based feature selection using feature-to-feature (F2F) scatter frequencies. Both proposed techniques employ the same relevance measure which resembles AUC computation. MRMD is a greedy earch-based technique with a diversity measure obtained from difference of ranks of positive instances. In the proposed clustering-based technique, F2F scatter frequencies are utilized as the dissimilarity measure for feature clustering.

3.1 MRMD

In greedy search-based filter approaches, the feature subsets are iteratively computed by evaluating the candidate features in terms of their relevance with the target class and pairwise redundancies. MRMD is a novel filter approach based on ranks of positive instances. In this approach, redundancy is replaced by diversity to quantify the complementarity of a candidate feature with respect to the already selected subset.

3.1.1 Relevance Measure

Consider a two-class classification problem where the positive and negative classes include P and N samples, respectively. The feature values can be undertaken as the probability scores of a hypothetical classifier as practiced in [19]. For a particular feature x_k, assume that d_p^k denotes its numerical value for the pth positive sample. Similarly, let d_n^k denote the value of the feature for the nth negative sample. Under the assumption that the higher decision values are produced for positive instances, in overall, the maximum likelihood estimate of AUC can be obtained from the ROC curve as [57], [58]

$$AUC_k = \frac{\sum_{p=1}^{P}\sum_{n=1}^{N} f(d_p^k, d_n^k)}{PN} \tag{3.1}$$

where

$$f(d_p, d_n) = \begin{cases} 1 & d_p > d_n \\ 0 & d_p < d_n \\ 0.5 & d_p = d_n \end{cases} \tag{3.2}$$

Assuming no ties for simplicity, $\sum_{n=1}^{N} f(d_p, d_n)$ can be interpreted as the number of negative instances having lower feature values than the value of the pth positive instance.

Consider a ranking system which assigns higher ranks to the samples having larger feature values and lower ranks to those with smaller feature values. The rank (r) of the sample having the largest feature value is $r = N + P$ whereas it is $r = 1$ for the sample with the smallest value. Let the positive and negative samples also be numbered separately. More specifically, the p value of the positive sample that has the highest feature value is set as P and the sample having the smallest feature value has $p = 1$. Table 3.1 presents an exemplar feature, including the feature values and r values of all ten samples and p values of five positive instances. It should also be noted that the r value of pth positive sample is represented as r_p^k. Using this notation, the inner summation term in Equation 3.1 can be rewritten as [56]

$$F_k(p) = \sum_{n=1}^{N} f(d_p^k, d_n^k) = r_p^k - p \quad \forall p = 1, \ldots, P. \tag{3.3}$$

Replacing Equation 3.3 in Equation 3.1, the expression for AUC becomes [56]

$$AUC_k = \frac{\sum_{p=1}^{P} F_k(p)}{PN} = \frac{\sum_{p=1}^{P} r_p^k - p}{PN} = \frac{\sum_{p=1}^{P} r_p^k - \frac{P(P+1)}{2}}{PN}. \tag{3.4}$$

Thus, AUC is a function of ranks of positives biased by a fixed term determined by the total number of positives. Using Equation 3.4, AUC of the exemplar feature presented in Table 3.1 can be computed as $\frac{2+3+3+3+5}{5\times 5} = 0.64$.

Table 3.1: Example decision values and related terms

$Value$	$Label$	r	p	$r_p - p$
0.10	-	1	NA	-
0.20	-	2	NA	-
0.29	+	3	1	2
0.30	-	4	NA	-
0.40	+	5	2	3
0.58	+	6	3	3
0.71	+	7	4	3
0.81	-	8	NA	-
0.95	-	9	NA	-
0.00	+	10	5	5

Equation 3.4 shows that ranking the features according to their AUC values is equivalent to ranking them according to the sum of ranks of positive samples. Based on this, the relevance score of the k^{th} feature, x_k is defined as

$$R(x_k; y) = \sum_{p=1}^{P} r_p^k \qquad (3.5)$$

where $r_{p,k}$ is the rank of p^{th} positive instance in the k^{th} feature.

It should be noted that for some features, the positive class may be assigned larger values than negatives whereas the opposite might be the case for some others. In order to have compatible ranking among features, the ranking system is defined to be feature-dependent. This dependency is determined in the context of AUC and its relationship with the ranks of instances using Wilcoxon-Mann-Whitney test as $AUC = \frac{U}{PN}$ [56]. This test checks if one of the two random variables is stochastically larger than the other one [59]. AUC estimation using Equation 3.4 is thus under the assumption that ranks assigned to the positive class are generally higher than the ones assigned to the other class. This assumption is valid only if a larger feature value corresponds to a higher probability of belonging to the positive class. Otherwise, feature values need to be converted into probability scores to ensure that they represent the degree of being from the positive class. This conversion requires pre-processing of features as practiced in [18], [19]. The assumption of Wilcoxon-Mann-Whitney test will be valid if the AUC given in Equation 3.6 is greater than 0.5 and accordingly, the relevance value determined by Equation 3.5 is correct only if

$$AUC_k > 0.5 \Rightarrow \sum_{p=1}^{P} r_p^k > (0.5 \times PN) + \frac{P(P+1)}{2} \Rightarrow \sum_{p=1}^{P} r_p^k > \frac{P(P+N+1)}{2} \qquad (3.6)$$

is satisfied. Otherwise, both the ranking system and the numbering system should be switched to perform the opposite.

Another important issue to be addressed in computing ranks of samples is ties. One general practice which is also undetaken in this study is to assign the average rank to all the samples in a tie set [60]. It is worth mentoning that Equation 3.4 is the precise estimate of AUC provided that there are no ties in the ranks. Although this limitation has been ignored in some of the previous studies [18], [19], it is well-known that the estimated AUC would be affected by the way ties are treated [60]. Considering average rank in case of ties implies that AUC is estimated as the area under a ROC curve with diagonal moves.

As discussed in Chapter 2, MI-based feature selection methods apply an initial discretization process that merges some instances into a single bin. Hence, the ranks of instances inside each bin are inevitably ignored. Similarly, the orders of the bins do not influence the score assigned to the feature. In other words, MI-based feature selection methods exhibit the same characteristics if the order of samples inside the bins or the orders of bins are randomly changed. However, taking into account the orders of instances as in $R(x_k, y)$ in computing the relevance of each feature is expected to result in more accurate evaluation of discriminative potential of different features.

3.1.2 Diversity Measure

As it can be seen in Equation 2.2 in the general form of the objective function of feature selection methods, another important component is redundancy. In the proposed approach, MRMD, this term is estimated as the diversity of ranks of positives. Consider the two dimensional space constructed by features x_j and x_k and a hypothetic decision boundary. We assume that the feature x_j contains classification-complementary information with respect to x_k if the low-ranked instances in x_k are the high-ranked instances in x_j. It should be mentioned that the ranks of positives are identified by the orders of both positive and negative instances and consequently, considering positive instances in calculations does not mean ignoring the negative class. Diversity is approximated as the sum of the absolute differences of the ranks of positives in x_j and x_k. The larger this value is, the more scattered instances appear in the two dimensional space. Thus, diversity of ranks of positives can be utilized to assess the complementarity of a pair of features. Diversity score between two features is computed as

$$D(x_k, x_j; y) = \sum_{p=1}^{P} |r_p^k - r_p^j| \tag{3.7}$$

where $|.|$ stands for absolute value. Note that the p values assigned to positive instances are determined from the first selected feature and they remain unchanged during the rest of feature selection procedure.

3.1.3 MRMD Algorithm

The objective function to be maximized in MRMD is the combination of relevance and diversity terms. As discussed in Chapter 2, there are two schemes in MI-based feature selection methods for defining the redundancy term: average and minimum. The *avg* scheme, as in mRMR, computes the redundancy as the total/average mutual information between the candidate feature and all of the already selected features. On the other hand, the *min* scheme utilizes the minimum mutual information value when all previously selected features are considered. In this study, we similarly adapted two variants of MRMD. However, unlike redundancy which is a competing term to relevance, diversity is a contributing factor. Hence, either average or minimum of the diversity is added to the relevance. The objective functions of MRMD^{avg} and MRMD^{min} are given in Equations 3.8 and 3.9, respectively.

$$J_{avg}(x_k; y) = R(x_k; y) + \frac{1}{|\Gamma|} \sum_{x_j \in \Gamma} D(x_k, x_j; y) \quad (3.8)$$

$$J_{min}(x_k; y) = R(x_k; y) + \min_{x_j \in \Gamma} D(x_k, x_j; y) \quad (3.9)$$

where Γ refers to the already selected subset of features which is initialized as an empty set and |Γ| is the cardinality of the Γ, i.e. the number of already selected features.

The proposed method searches for the maximal value the objective function. The overall algorithm used for MRMD^{avg} is presented in Algorithm I. The greedy search-based MRMD^{avg}, iteratively searches for the candidate feature which maximizes the summation of relevance to the class label and average diversity with respect to the previously selected features. MRMD^{min} can be obtained by replacing $J_{avg}(x_k; y)$ with $J_{min}(x_k; y)$.

Algorithm I. MRMD^{avg} Feature Selection

```
Inputs:
    $F$: feature set
    $d$: data set dimension
    $\kappa$: target dimension
    $P$: number of positives
Output:
    $\Gamma$: selected subset
Begin:
    $\Gamma = \phi$
    $x_k \in F$
    Assign $r_p^k$ to satisfy Equation 3.6
    Compute $R(x_k; y) = \sum_{p=1}^{P} r_p^k$
    $f = \underset{x_k \in F}{arg\,max}\, R(x_k; y)$
    $\Gamma = \{f\}$
    Compute p values according to $f_1$
    $while\ |\Gamma| < \kappa$
        $x_k \in F - \Gamma$
        $x_j \in \Gamma$
        $D(x_k, x_j; y) = \sum_{p=1}^{P} |r_p^k - r_p^j|$
        $J_{avg}(x_k; y) = R(x_k; y) + \frac{1}{|\Gamma|} \sum_{x_j \in \Gamma} D(x_k, x_j; y)$
        $f = \underset{x_k \in F - \Gamma}{arg\,max}\, J_{avg}(x_k; y)$
        $\Gamma = \Gamma \cup \{f\}$
    $endwhile$
Return $\Gamma$
```

3.1.4 Artificial Example

In order to demonstrate the efficiency of proposed algorithm, an artificial example of selecting a two-dimensional feature subspace is considered. The feature subspace computed using MRMD is compared with that of mRMR. The artificial data set is given in Table 3.2 with 20 samples identified $s_1, \dots, s_{20}$. y refers to the class label where the positive samples are labeled using $y = 1$. The data set comprises 4 features, $x_1, \dots, x_4$. For simplicity, 10 positive and 10 negative instances are employed. Without loss of generality, we assumed the same range of values for all of the features to simplify the discretization procedure and the calculations of entropies for mutual information. We firstly conduct MRMD feature selection on this data set to select a subset of two features. The selected subset is initialized as $\Gamma = \phi$. According to Equation 3.6, the instances need to be ranked appropriately to satisfy $\sum_{p=1}^{10} r_p > 105$. When higher ranks are assigned to larger values, the relevance scores of the features are computed as

R(x$_1$*;y) = 136, R(x*$_2$*;y) = 129, R(x*$_3$*;y) = 1113, R(x*$_4$*;y) = 99.*

Except for x_4, the relevance scores satisfy the condition. For x_4, the higher ranks should be assigned to smaller values which returns $R(x_4; y) = 111$. Since the maximum relevance score belongs to x_1, the selected subset is updated to $S = \{x_1\}$ in the first stage. Table 3.3 shows the final ranks of positive instances and the corresponding p values for all features. Note that x_1 is the reference feature to number positive instances since it is the first feature added to the subset.

Table 3.2: Artificial data set

	x_1	x_2	x_3	x_4	y
s_1	0.05	0.05	0.45	0.35	0
s_2	0.10	0.10	0.05	0.20	0
s_3	0.15	0.15	1.00	0.65	0
s_4	0.20	0.20	0.85	1.00	0
s_5	0.25	0.25	0.70	0.40	0
s_6	0.30	0.45	0.65	0.55	0
s_7	0.45	0.50	0.50	0.60	0
s_8	0.65	0.65	0.25	0.15	0
s_9	0.70	0.75	0.30	0.90	0
s_{10}	0.85	0.95	0.10	0.75	0
s_{11}	0.35	0.30	0.95	0.05	1
s_{12}	0.40	0.35	0.90	0.10	1
s_{13}	0.50	0.40	0.80	0.25	1
s_{14}	0.55	0.55	0.75	0.30	1
s_{15}	0.60	0.60	0.60	0.45	1
s_{16}	0.75	0.70	0.55	0.50	1
s_{17}	0.80	0.80	0.40	0.70	1
s_{18}	0.90	0.85	0.35	0.80	1
s_{19}	0.95	0.90	0.20	0.85	1
s_{20}	1.00	1	0.15	0.95	1

Table 3.3: Ranks of positive instances and relevance values

p	r_p^1	r_p^2	r_p^3	r_p^4
1	7	6	19	20
2	8	7	18	19
3	10	8	16	16
4	11	11	15	15
5	12	12	12	12
6	15	14	11	11
7	16	16	8	7
8	18	17	7	5
9	19	18	4	4
10	20	20	3	2
R	**136**	**129**	**113**	**111**

In the second step, diversity values with respect to x_1 are to be computed. In this example, MRMD$^{\text{avg}}$ is considered. However, it should be noted that there is no difference between the two variants of MRMD i.e. MRMD$^{\text{avg}}$ and MRMD$^{\text{min}}$, when two features are selected. The reason is that diversity part is employed only in selecting the second feature, where $|\Gamma| = 1$. Table 3.4 shows $(r_p^1 - r_p^k)$ values and diversity scores, $D(x_k, x_1; y)$ of the candidate features, x_2, x_3 and x_4. The overall scores of these features are calculated as

$J_{avg}(x_2;y)$ = 129 + 7 = 136

$J_{avg}(x_3;y)$ = 113 + 87 = 200

$J_{avg}(x_4;y)$ = 111 + 75 = 186

According to MRMD, x_3 carries the most complementary information to the selected subset, $\Gamma = \{x_1\}$. For visual interpretation of the discrimination potential of the three possible feature subsets, scatter plots of samples are illustrated in Figures 3.1, 3.2 and 3.3. Notice that the subset selected by MRMD (i.e $S = \{x_1, x_3\}$) constructs a linearly separable subspace.

Table 3.4: Differences of ranks of positives with respect to x_1 and diversity values

p	$\lvert r_p^1 - r_p^2 \rvert$	$\lvert r_p^1 - r_p^3 \rvert$	$\lvert r_p^1 - r_p^4 \rvert$
1	1	12	13
2	1	10	11
3	2	6	6
4	0	4	4
5	0	0	0
6	1	4	4
7	0	8	9
8	1	11	13
9	1	15	15
10	0	17	18
$\boldsymbol{D(x_k, x_1; y)}$	**7**	**87**	**75**

mRMR is also applied to the artificial data set for comparison. In the first step, feature values are discretized into 5 equally frequency bins. In Table 3.5, distribution of positive and negative samples in each bin is shown for all features. In the table, 1's denote positives and 0's denote negatives. For instance, in bin1, there are three positives and one negative where, the negative sample has the smallest feature value. The mutual information between the class label and the features are computed as follows:

$I(x_1;y) = H(y) – H(y|x_1) = 1 – 0.72 = 0.28$

$I(x_2;y) = H(y) – H(y|x_2) = 1 – 0.72 = 0.28$

$I(x_3;y) = H(y) – H(y|x_3) = 1 – 1 = 0$

$I(x_4;y) = H(y) – H(y|x_4) = 1 – 1 = 0$

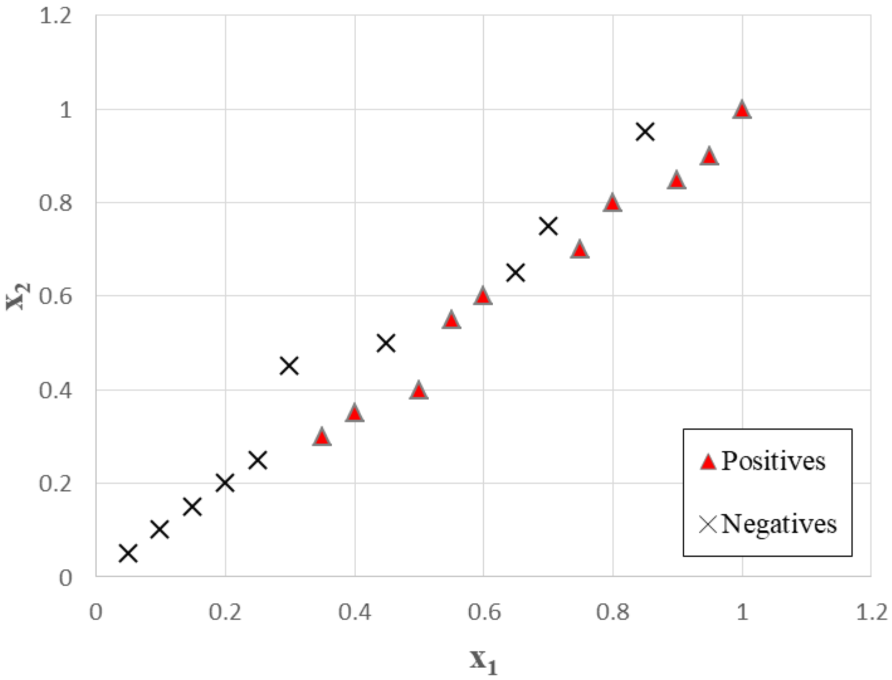

Figure 3.1: Scatter plot of samples in (x_1, x_2) subspace

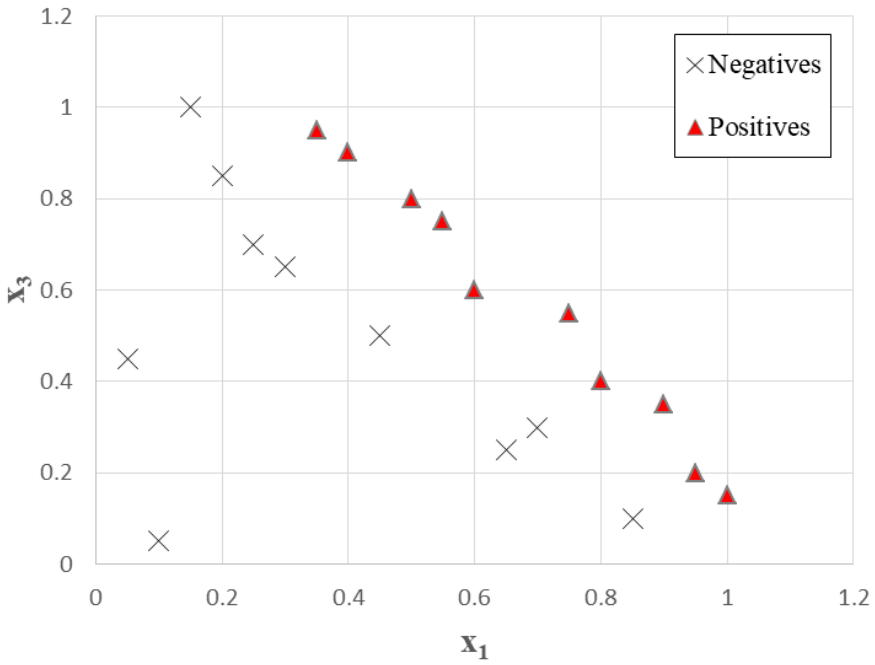

Figure 3.2: Scatter plot of samples in (x_1, x_3) subspace

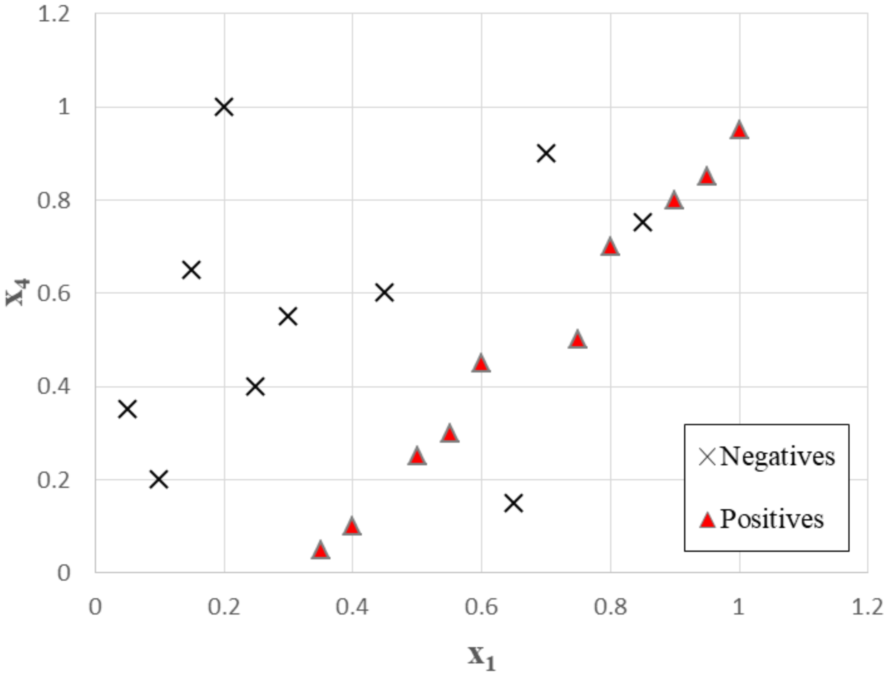

Figure 3.3: Scatter plot of samples in (x_1, x_4) subspace

Table 3.5: Distribution of positive and negative instances in bins for each feature

bin#	x_1	x_2	x_3	x_4
bin1	1	1	0	0
	1	0	1	1
	1	1	1	0
	0	1	0	1
bin2	1	1	1	1
	1	0	1	0
	0	1	0	1
	0	0	0	0
bin3	1	1	1	0
	1	1	1	0
	1	0	0	1
	0	0	0	1
bin4	1	1	1	0
	1	1	1	0
	0	1	0	1
	0	0	0	1
bin5	0	0	1	0
	0	0	1	0
	0	0	0	1
	0	0	0	1

Although the features have different AUC values, mutual information between the class labels is not representative for the differences in their discriminative potential. This may foster the generalization capability of MI-based feature selection approaches and add some levels of nonlinearity, but it may result in a suboptimal feature subset for classification.

Assume that x_1 is added to the subset. The next step of mRMR is to calculate redundancies with respect to x_1. Using the contingency tables of pairs of features, the mutual information between each candidate feature and x_1 can be calculated as

$I(x_2; x_1) = H(x_2) - H(x_2|x_1) = .32 - 0.32 = 2.00$

$I(x_3; x_1) = H(x_3) - H(x_3|x_1) = 2.32 - 1.02 = 1.30$

$I(x_4; x_1) = H(x_4) - H(x_4|x_1) = 2.32 - 1.50 = 0.82$

Then, the scores are obtained using mRMR are as follows:

$J(x_2; y) = 0.28 - 2 = -1.72$

$J(x_3; y) = 0 - 1.30 = -1.30$

$J(x_4; y) = 0 - 0.82 = -0.82$

Consequently, x_4 is added to the feature subset as it achieves the maximum score. In fact, from Figures 3.1, 3.2 and 3.3, we know that this is not the best-fitting subspace. It should be noted that this example is not considered as an unconditional proof to superiority of the proposed method over mRMR, but it clarifies the procedure of MRMD and elucidates probable deficiencies of MI-based methods.

3.2 Clustering-based Feature Selection using F2F Scatter Frequencies

One of the well-known rank-based similarity measures proposed in the literature is Spearman's Correlation Coefficient [26]. In computing the similarity of features, the Spearman's rank correlation coefficient takes into account the squared differences of ranks over all training samples. While computing this metric, the samples whose ranks are highly different may dominate the redundancy score. Consider a set of five training samples. Let the ranks be *{1,2,3,4,5}* using the feature x_1, *{2,1,4,3,5}* using x_2 and *{5,2,3,4,1}* using x_3. When the squared differences of ranks are considered, the dissimilarity between x_1 and x_3 is found to be much higher than the dissimilarity between x_1 and x_3. However, most of the instances are assigned the same rank by x_1 and x_3. It can be argued that smoothing the rank differences may help to define a more reasonable metric. One way to achieve this is to set an upper limit for the differences in ranks. In other words, the differences above a certain upper threshold will be set to a constant value that depends on the number of the training samples. Alternatively, the dissimilarity may be defined to take into account the number of times that the training samples are closely ranked by the features.

The proposed dissimilarity metric for feature clustering, Feature-to-Feature (F2F) scatter frequencies is based on the assumption that the difference between ranks assigned by two features to an arbitrary sample is proportional to their dissimilarity. In order to quantify the dissimilarity between a pair of features, each sample is encoded in terms of affinity sets. Affinity sets are in fact the local clusters of features containing features with rank differences smaller than a predefined value. Considering all affinity sets, the global dissimilarity is obtained as the pairwise scatter frequencies of features. The more a pair of features co-occur in the affinity sets, the more similar they are. Correspondingly, the fewer two features co-occur, the more dissimilar they are.

3.2.1 Formulation of F2F Scatter Matrix

Let $S \in \mathcal{R}^{K \times d}$ be a matrix representing the data set of K samples in d-dimensional feature space. Each row of S is a sample denoted by s_k $(k = 1, 2, \dots, K)$ and each column is a feature denoted by x_i $(i = 1, 2, \dots, d)$. The first step to compute the dissimilarity of two features is to convert the sample values into ranks to obtain the rank matrix, $R \in \mathcal{R}^{K \times d}$. For each feature x_i, the samples are firstly ranked based on their feature values. As mentioned above, a good feature is expected to assign larger scores to one class when compared to the other class.

Assume that the number of positive and negative samples are P and N, respectively. For a particular feature x_i, the ranks are denoted by r^i where the sample having the largest feature value is assigned $r^i = P + K$ and the sample with the smallest feature value is assigned $r^i = 1$. We call this rank ordering ascending. Let $r_k^i \in [1, K]$ denote the feature-dependent rank of sample s_k when all samples are ranked using the feature value of x_i. The data matrix S is converted to the rank matrix R where r_k^i is the kth row and ith column of R.

The formulation of F2F is based on the presumption that the rank ordered values of each feature are generally higher for the positive class than that of the negative class. This concept resembles the assumption of Wilcoxon-Mann-Whitney test as given in Equation 3.6. Thus, the following condition is to be satisfied for each feature x_i.

$$\alpha_i = \sum_{p=1}^{P} r_p^i > \frac{P(P+N+1)}{2} \tag{3.10}$$

Otherwise, the ranking system is reversed to descending by assigning higher ranks to samples having smaller values and lower ranks to samples with larger values [58]. This strategy guarantees the consistency of rank matrix among all features.

After computing R, the affinity sets of features are constructed. In order to identify the affinity sets of sample s_k, a proximity window of size w ($w \ll K$) is applied on $r_k^i,\ i = \{1, \dots, d\}$. The window captures the neighboring features denoted by x_i and x_j by letting the value $\Delta r_k^{ij} = |r_k^i - r_k^j|$ be always less than w. It is worth mentioning that for ties, average rank is used to ensure that the proximity window captures the neighboring features on both sides of a tied set. However, the average is truncated to the closest integer since the size of the proximity window is integer.

The mth affinity set derived from s_k is defined as

$$v_k^m = \bigcup_{|r_k^i - m| < w} x_i \tag{3.11}$$

where $m = 1,2,\ldots,(K - w + 1)$. The number of sets obtained from each sample depends on the total number of samples and the window size. It should be noted that some affinity sets of a given sample may be empty and each feature may appear in multiple affinity sets of the same sample. The empty affinity sets are discarded.

The dissimilarity of features can be quantified by taking into account their scatter frequencies among different affinity sets. Frequencies of features and their co-occurrences are firstly computed for this purpose. Let X_i be the number of sets which contain feature x_i and X_{ij} be the number of sets in which x_j co-occurs with x_j. Then, the dissimilarity of x_i and x_j is computed as their F2F scatter frequency using

$$d_{ij} = \begin{cases} X_i + X_j - 2 \times X_{ij} & i \neq j \\ 0 & i = j \end{cases} \tag{3.11}$$

where $d_{ij} = d_{ji}$. The F2F scatter frequency of x_i and x_j is the total number of affinity sets containing either of them and not both. d_{ij} corresponds to the ith row and jth column entry of the F2F dissimilarity matrix, $D \in \mathcal{R}^{d \times d}$.

3.2.2 Artificial Example

In order to clarify the steps of calculating F2F dissimilarity matrix, an artificial example is provided. Assume a sample data set with 5 samples $(K = 5)$ and 10 features $(d = 10)$ as

$$S = \begin{bmatrix} 0.9 & 5 & 1.0 & 0.3 & 4 & 0.10 & 4 & 6 & 0.4 & 4 \\ 0.8 & 3 & 0.0 & 0.4 & 2 & 0.03 & 2 & 5 & 0.3 & 3 \\ 0.1 & 7 & 1.5 & 0.7 & 3 & 0.02 & 7 & 1 & 0.1 & 2 \\ 0.2 & 2 & 1.2 & 0.6 & 1 & 0.01 & 1 & 2 & 0.2 & 1 \\ 0.3 & 4 & 0.2 & 0.7 & 5 & 0.09 & 6 & 3 & 0.1 & 0 \end{bmatrix} \tag{3.12}$$

In order to obtain R, the class labels are required to determine the direction of ranking (descending or ascending). According to Equation 3.10, the rank matrix is consistent among all features if the sum of ranks of positive instances is not smaller than $\frac{P(P+N+1)}{2}$. This is to assure that negative and positive correlations among features are represented identically. Suppose that the labels are $[+,+,+,-,-]^T$. The α_i values (Equation 3.10) of positive samples (i.e. first three) should be greater than or equal to $\frac{3(3+2+1)}{2} = 9$. For each feature, the ranks are firstly computed by assigning the largest rank ($P + N = 5$) to the sample having the largest feature value. Note that this choice has no effect on the final result. The preliminary form of the rank matrix denoted by $\tilde{R}$ is

$$\tilde{R} = \begin{array}{c} \begin{array}{cccccccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 & x_9 & x_{10} \end{array} \\ \begin{bmatrix} 5 & 4 & 3 & 1 & 4 & 5 & 3 & 5 & 5 & 5 \\ 4 & 2 & 1 & 2 & 2 & 3 & 2 & 4 & 4 & 4 \\ 1 & 5 & 5 & 4 & 3 & 2 & 5 & 1 & 1 & 3 \\ 2 & 1 & 4 & 3 & 1 & 1 & 1 & 2 & 3 & 2 \\ 3 & 3 & 2 & 4 & 5 & 4 & 4 & 3 & 1 & 1 \end{bmatrix} \end{array} \tag{3.13}$$

Then, α_i value of each feature is computed as presented in Table 3.6. The order of ranking is reversed in cases where the condition in Equation 3.10 is violated. In the given example, the only case is x_4.

Table 3.6: The sum of ranks of positives and validity of Equation 3.10

x_i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
α_i	10	11	9	7	9	10	10	10	10	12
$\alpha_i \geq 9$	✓	✓	✓	×	✓	✓	✓	✓	✓	✓

Accordingly, the ranking of x_4 is reversed and the fourth column of $\tilde{R}$ is updated to obtain the consistent rank matrix R as

$$R = \begin{array}{c} \begin{array}{cccccccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 & x_9 & x_{10} \end{array} \\ \begin{bmatrix} 5 & 4 & 3 & 5 & 4 & 5 & 3 & 5 & 5 & 5 \\ 4 & 2 & 1 & 4 & 2 & 3 & 2 & 4 & 4 & 4 \\ 1 & 5 & 5 & 1 & 3 & 2 & 5 & 1 & 1 & 3 \\ 2 & 1 & 4 & 3 & 1 & 1 & 1 & 2 & 3 & 2 \\ 3 & 3 & 2 & 1 & 5 & 4 & 4 & 3 & 1 & 1 \end{bmatrix} \end{array} \qquad (3.14)$$

Based on R, the updated $\alpha_4 = 10$ and other α_i values are remain as given in Table 3.10. In order to obtain the F2F dissimilarity matrix, affinity sets, v_k^ms are firstly formed assuming a window size of $w = 2$. In other words, each affinity set includes features with rank differences equal to 0 or 1. Table 3.7 shows the affinity sets obtained from the artificial data set.

Table 3.7: The affinity sets extracted from the toy example. (The empty sets are discarded)

Sample	Affinity sets
s_1	$\{x_2, x_3, x_5, x_7\}$; $\{x_1, x_2, x_4, x_5, x_6, x_8, x_9, x_{10}\}$
s_2	$\{x_2, x_3, x_5, x_7\}$; $\{x_2, x_5, x_6, x_7\}$; $\{x_1, x_4, x_6, x_8, x_9, x_{10}\}$
s_3	$\{x_1, x_4, x_6, x_8, x_9\}$; $\{x_5, x_6, x_{10}\}$; $\{x_2, x_3, x_7\}$
s_4	$\{x_1, x_2, x_5, x_6, x_7, x_8, x_{10}\}$; $\{x_1, x_4, x_8, x_9, x_{10}\}$; $\{x_3, x_4, x_9\}$
s_5	$\{x_3, x_4, x_9, x_{10}\}$; $\{x_1, x_2, x_3, x_8\}$; $\{x_1, x_2, x_6, x_7, x_8\}$; $\{x_5, x_6, x_7\}$

Elements of matrix D are computed by simply counting all X_i and X_{ij} values and replacing them in Equation 3.11 to compute the F2F scatter frequencies. For instance, $X_1 = 7$, $X_2 = 8$ and $X_{1,2} = 4$, resulting in $d_{1,2} = 7$. Note also that for x_1 and x_8 which are highly correlated, we get $d_{1,8} = 0$.

3.2.3 Extension to Multi-class

The proposed F2F scatter metric structurally requires the labels of positive and negative samples. However, it can be easily adapted for the case of multi-class problems by means of one-versus-all approach. Consider a data set with K samples, d features and C classes Let K_c denote the number of samples in class c, $c \in \{1,2, \dots, C\}$. We decompose the problem into C one-versus-all subproblems. Consider the cth subproblem where the instances in the cth class are defined to belong to the positive class and the remaining instances form the negative class. The rank matrix, R_c for this subproblem is first obtained. From R_c, only the K_c rows belonging to class c are retained. Having obtained R_c for all classes, they are concatenated to form the final rank matrix $R \in \mathcal{R}^{K \times d}$. Accordingly, affinity sets and the F2F dissimilarity matrix, D, are obtained from R. In multi-class problems, the α_i of a feature is computed by averaging the corresponding values obtained from C subproblems.

3.2.4 Clustering-based Feature Selection using F2F Measure

The proposed clustering-based feature selection method using F2F dissimilarity matrix is described in Algorithm II. It starts by finding feature-dependent rank matrix. In this phase, in addition to R, the α_i values are obtained. These values are not only employed in computing a proper rank matrix, but also used as the relevance measure to find the representative feature of each cluster. The affinity sets are captured using R and F2F scatter matrix is computed from affinity sets. In the clustering phase, feature clusters are formed based on F2F dissimilarity metric.

The choice of clustering algorithm is crucial in obtaining meaningful clusters of features [45], [61]. There are several clustering algorithms proposed in the literature [45], [61]. However, according to the short guideline developed via a comprehensive investigation of the performance of several clustering methods for data analysis, it has been suggested that partitioning around medoid (PAM), and hierarchical clustering (HC) are amongst the best performing clustering methods [61]. Therefore, these two clustering approaches are considered as the alternatives for implementing proposed clustering-based feature selection method.

Algorithm II. Clustering-based Feature Selection using F2F Scatter Frequencies

```
Input:
    $S$ - data matrix of $K$ samples and $d$ features
    $w$ - proximity window size
    $\kappa$ - cardinality of the selected feature subset
Output:
    $\Gamma$ - selected feature subset
Begin:
    Compute the feature-dependent rank matrix $R$
     $for\ \ k = 1{:}K$
        find all $v_k^m$ in $s_k$
        discard empty $v_k^m$
    $endfor$
    $for\ \ i = 1{:}d$
        $for\ \ j = i + 1{:}d$
            compute $d_{ij}$ using $v_k^m$ and Equation 3.11
        $endfor$
    $endfor$
    cluster features into $\kappa$ distinct clusters using $D$
    $for\ k = 1{:}\kappa$
        find the representative feature of cluster $c_k$ using $\alpha_i$
        $f = x_{\underset{x_i \in c_k}{argmax}(\alpha_i)}$
        $\Gamma = \Gamma\ \cup \{f\}$
    $endfor$
Return $\Gamma$
```

Partitioning around medoids (PAM) is the most popular realization of k-medoids clustering [62] that is a modified form of k-means clustering. Given κ as the number of clusters, k-means starts by assigning each data point to the cluster with nearest mean. The algorithm iterates by recomputing cluster means and reassigning cluster members until a stopping criterion is met. k-medoids algorithm, on the other hand, selects one of the cluster members called medoid as the center. This algorithm is less sensitive to outliers and noise than k-means [63]. In PAM implementation of k-medoids, the cost function is defined as the sum of distances of data points to their corresponding medoids. In searching for optimum solution which minimizes the cost function, data points and medoids are swapped and the cost function is recalculated. If the swap reduces the cost, it will be kept. Otherwise, the algorithm redoes the swap and continues as long as the cost function is decreased. Since PAM algorithm is based on greedy search, it is faster than conventional k-means [63].

In hierarchical clustering (HC), a series of nested clusters are generated that represent a hierarchical structure [64]. This hierarchy portrays a dendrogram representing how clusters are formed at different levels. By cutting the dendrogram at some specific height, κ clusters are obtained. There are two approaches for forming the dendrogram namely, agglomerative and divisive [45]. Agglomerative approach is bottom-up which considers each single data point as a distinct cluster at the beginning and merges pairs of close clusters at each level until one cluster containing all data points is obtained. Divisive approach is top-down which initially regards all data points as a single cluster. Moving down to the hierarchy at each level, clusters are split until each data point forms a cluster. In this study, HC is implemented by using agglomerative approach based on complete linkage for merging clusters [65].

Having formed the feature clusters, α_i is used as the relevance metric to select the representative feature of each cluster. As mentioned above, in computing the dissimilarity matrix and α_i scores of the features, the ranks of ties are computed by truncating average ranks.

3.3 Conclusions

In this chapter, the two proposed feature selection methods, namely MRMD and clustering-based feature selection using F2F scatter frequencies are explained in details. The metrics adopted in the proposed methods are principally developed based on ranks of instances. The relevance metric in both methods is the same and the feature ranking it offers is identical to that of AUC. MRMD employs greedy search to maximize an objective function defined as the summation of the relevance and relative diversity. The compatible relevance and diversity terms are formulized based on ranks of positive instances to address the shortcomings of the popular MI-based feature selection techniques. In clustering-based feature selection method using F2F scatter frequencies, a novel dissimilarity measure related to the co-occurrence of features captured by closeness of ranks is utilized.

Chapter 4

EXPERIMENTAL RESULTS

4.1 Data Sets

The experiments are conducted on four data sets from UCI machine learning repository and six Microarray gene expression data sets. Table 4.1 lists the data sets considered in this study. The number of features, samples, classes, and the source of each data set are also given in the table. These data sets are amongst the popular ones in feature selection literature. The number of features in data sets varies between 44 and 12533 which cover a wide range of dimensions. It should be noted that computing the best-fitting feature subsets of Microarray data sets is very challenging due to having small number of samples and large number of features.

Table 4.1: Datasets used in the experiments

No.	Data set	#Features	#Samples	#Classes	Source
1	SPECTF	44	267	2	UCI
2	Sonar (Connectionist)	60	208	2	UCI
3	Plant Leaf	64	1599	100	UCI
4	Urban Land Cover	147	675	9	UCI
5	Musk (Musk1 and Musk2)	166	7074	2	Microarray
6	Colon Cancer	2000	62	2	Microarray
7	Breast Cancer (NIH)	2905	168	2	Microarray
8	Leukemia (NIH)	3571	72	2	Microarray
9	Lymphoma	7129	77	2	Microarray
10	Lung cancer	12533	181	2	Microarray

4.2 Experimental Setting

In all experiments, 10-fold cross validation is applied for generating train and test splits. For each fold, given a predefine feature subset size, κ, the best-fitting feature subset is firstly obtained by using the corresponding training set. Then, classifier models are trained to predict the labels of the test samples. In order to reduce the classifier bias, two classifiers namely, 3 nearest neighborhood (3NN) and linear support vector machines (SVM) are employed. These two classifiers are widely-used in the state-of-the-art [27], [28] for evaluating feature selection techniques. All experiments are performed using different number of feature subsets ranged from 2 to 50 ($\kappa = \{2,3\ldots,50\}$). In cases where the data set have less than 50 features, experiments continue until *d* (the total number of features in the data set).

The performance of each classifier is measured using both AUC and accuracy, separately for all feature subsets. AUC and accuracy are averaged across the two classifiers to reduce the bias. Both performance metrics are reported as the average across all subsets. For instance, for $\kappa = 5$, the average performance of 3NN and SVM is computed for subsets of size 2, 3, 4 and 5. The reported performances are the averages of all subsets. Thus, the comparisons make based on these values correspond to the relative effectiveness of the competing methods for different sizes of the feature subset.

Another important criterion for assessing feature selection methods is stability. In this study, Kuncheva's stability index [66] is utilized to measure the stability of the feature subsets. This measure is computed as follows. Assume that two feature subsets of the same size, κ, from two folds named as F_i and F_j are obtained. Kuncheva's stability index is obtained for F_i and F_j as

$$SI(F_i, F_j) = \frac{m.d-\kappa^2}{\kappa(d-\kappa)}, \tag{4.1}$$

where d is the total number of features and m is the cardinality of the intersection set $m = |F_i \cap F_j|$. The larger the SI value is, the more stable the feature selection is considered. In this study, stability index is computed for $\kappa = min(50, d)$ for each data set. Stability index of all possible pairs of feature subsets for all folds is obtained. Average SI values across all 45 pairs of F_i and F_j and then over the folds of 10 fold cross validation are calculated to report the total stability index of each feature selection method for each data set.

4.3 Experimental Results of MRMD

The performance of the proposed method is compared with five MI-based feature selection methods, namely NJMIM [28], JMIM [28], CMIM [49], DISR [51], and mRMR [13], and a conventionally used non-MI-based filter, namely Relief [67]. These methods are implemented using R package, 'praznik'. For MI-based feature selection algorithms, numerical attributes are discretized into 10 equally-spaced bins. In MRMD, categorical attributes are converted into probabilities to obtain the ranks of training instances. In the case of multi-class problems, the features are ranked in terms of the average MRMD scores computed using multiple binary one-versus-all problems.

The average accuracies and AUC scores of SVM and 3-NN classifiers are presented in Table 4.2 and Table 4.3, respectively. The maximum value obtained on each data set is shown in boldface which corresponds to the best performing method or briefly, *best*. To test the paired significant difference between the best and the others, Wilcoxon signed ranked test [68] is applied between the filter resulting in the highest average value and all the other filters. The cases where null hypothesis is not rejected $(p > 0.05)$ are marked with an asterisk $(*)$. Hence, the cases marked with $*$ correspond to ties with the winner. The average AUC scores and the numbers of wins, ties and losses are given in the last two rows of the tables.

Table 4.2: Average accuracy of classifiers (in percentage) across all feature subsets

Data	**MRMD^{avg}**	**MRMD^{min}**	**NJMIM**	**JMIM**	**CMIM**	**DISR**	**mRMR**	**Relief**
1	**78.21**	77.97*	77.82*	76.69	76.78	77.87*	77.44	77.92*
2	77.72*	77.13	78.38*	**78.46**	77.99	77.90	76.08	77.83
3	**86.50**	86.15*	85.93	85.90	85.93	85.95	86.35*	85.96
4	77.13	76.89	79.45	80.38	79.46	77.34	**82.74**	76.71
5	91.83	91.70	**92.16**	91.91	91.69	92.07*	89.64	89.93
6	**79.81**	78.23	75.68	72.74	77.03	78.15	76.10	77.32
7	**78.22**	76.90*	73.02	71.01	73.11	71.36	71.34	72.20
8	**96.99**	96.43*	93.43	94.33	93.64	92.60	93.57	93.39
9	**98.05**	94.48	91.75	90.21	90.01	90.22	89.60	91.22
10	**99.03**	98.96*	97.77	98.92*	98.66	97.55	98.17	98.30
AVG.	**86.35**	85.48	84.54	84.06	84.43	84.10	84.10	84.08
W/T/L	**7/1/2**	0/5/5	1/2/7	1/1/8	0/0/10	0/2/8	1/1/8	0/1/9

* corresponds to p>0.05

The results confirm that both versions of MRMD outperform the reference methods both in terms of average accuracy and AUC. Specifically, there is a remarkable improvement in AUC. Among the reference schemes, CMIM provides the highest average AUC value (84.20%). MRMD^{avg} improves this AUC to 86.85%. Considering that AUC is the foundation of the proposed method, these results are consistent with the theoretical basic of MRMD approach. Similarly, the highest average accuracy achieved by the competing method NJMIM is improved from 84.54% to 86.35% MRMD^{avg}. When the average accuracies achieved for each data set are evaluated, it can be seen that MRMD^{avg} is the best or tied with the best on 7 data sets out of 10. Similarly, it is the best or tied with the best on 6 data sets when average AUC is considered while the second best performing method is Relief with only 2 wins. MRMD^{avg} surpasses all reference schemes in total numbers of wins and losses as well.

Table 4.3: Average AUC of classifiers (in percentage) across all feature subsets

Data	**MRMD^{avg}**	**MRMD^{min}**	**NJMIM**	**JMIM**	**CMIM**	**DISR**	**mRMR**	**Relief**
1	**79.30**	78.84	77.49	76.84	77.51	78.63	77.08	77.66
2	84.82	84.13	84.69	84.81	**85.55**	84.34	83.34	85.15
3	**68.77**	68.73	68.59	68.54	68.55	68.61	68.60	68.60
4	72.12	72.21	72.58	72.19	71.27	72.81*	71.70	**72.87**
5	**91.07**	91.00*	90.75	90.57	90.01	90.85	84.27	86.39
6	**86.53**	84.40*	61.38	65.13	76.75	58.81	72.38	67.50
7	**87.79**	85.92*	78.89	71.62	79.60	78.94	68.48	71.35
8	99.08	98.94	99.70*	99.87*	99.57*	99.70*	99.73*	**100.00**
9	**99.50**	96.65	95.67	97.46	93.63	87.17	94.25	90.58
10	99.50	99.34	99.34	**99.74**	99.58	99.34	99.49	99.64
AVG.	**86.85**	86.02	82.91	82.68	84.20	81.92	81.93	81.97
W/T/L	**6/0/4**	0/3/7	0/1/9	1/1/8	1/1/8	0/2/8	0/1/9	2/0/8

* corresponds to $p>0.05$

Although the second variant of proposed method, MRMD^{min} performs better than other feature selection methods in terms of average accuracy and AUC, it is not as efficient as MRMD^{avg}. It can be concluded that, contribution of diversity term in feature score is more consistent with relevance term when it is averaged over the selected subset.

As stated in Chapter 2, estimation of mutual information is not reliable in high dimensional spaces. Relief also suffers in high dimensions with limited number of samples. Because of this, MRMD^{avg} is expected to achieve significantly better scores on microarray data sets, especially on the ones having a few number of samples (i.e. data sets numbered as 6 to 10). Referring to Table 4.2, it can be seen that the proposed method achieves remarkable improvements in terms of accuracy scores on these data sets. It also provides the highest AUC values on majority of these data sets as given in Table 4.3. It can be argued that the proposed method is a strong candidate to be considered for high dimensional data sets having a limited number of samples compared to the number of features.

In order to have a clearer insight, the average accuracy and AUC scores for some data sets are presented for different numbers of selected features in Figures 4.1 to Figure 4.10. Accuracy values in the plots are obtained by adding 5 more features to the feature subset in each step. The accuracy curves of Sonar, Urban Land Cover, Musk, Colon and Lymphoma are shown in Figure 4.1 to 4.5. These data sets correspond to a case of ties (Sonar), two cases of loses (Urban Land Cover and Musk), and two cases of wins (Colon and Lymphoma) for MRMD^{avg}. Similarly, average AUC of SVM and 3-NN classifiers for the same data sets are shown in Figure 4.6 to 4.10. On Sonar data set, Relief provides better accuracy and AUC scores on some feature subsets. However, when averaged over all 50 feature subsets, it performs poor compared to JMIM in terms of average accuracy as it can be seen in

Table 4.2. Similarly, its performance is worse than that of CMIM when average AUC scores are considered. On the other hand, MRMD^{avg} hits the highest accuracy and AUC scores for a subset of 30 features.

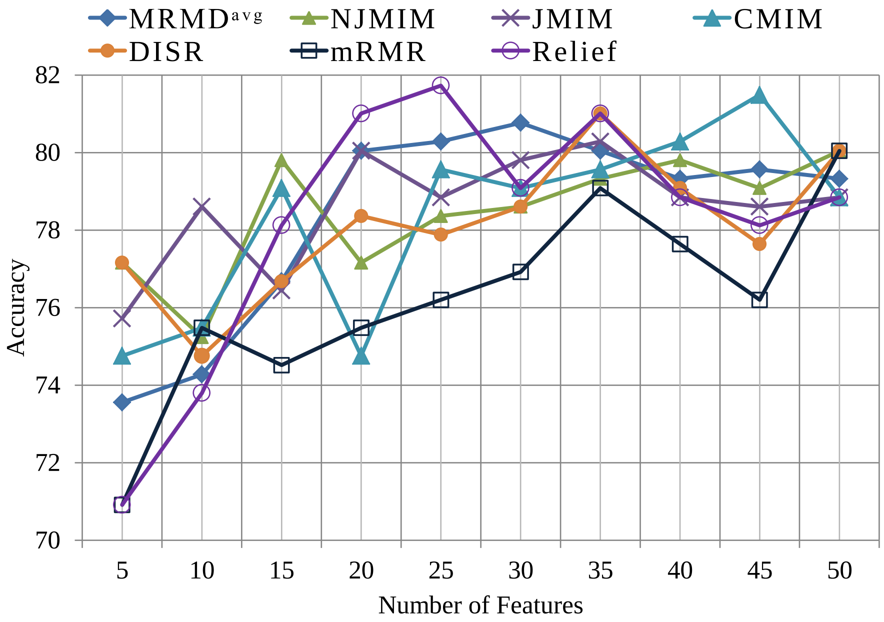

Figure 4.1: Average classification accuracy (in percentage) of SVM and 3-NN on Sonar data set

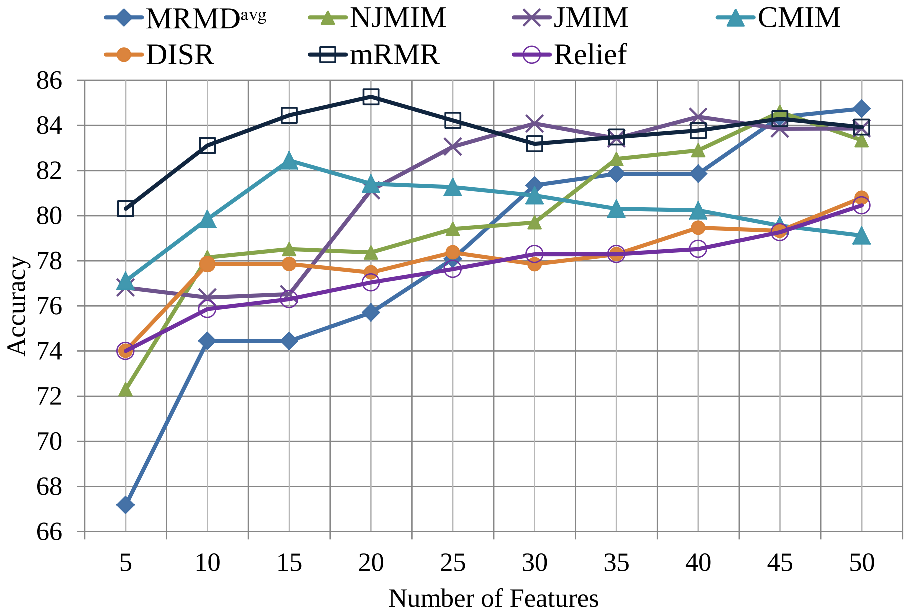

Figure 4.2: Average classification accuracy (in percentage) of SVM and 3-NN on Urban Land Cover data set

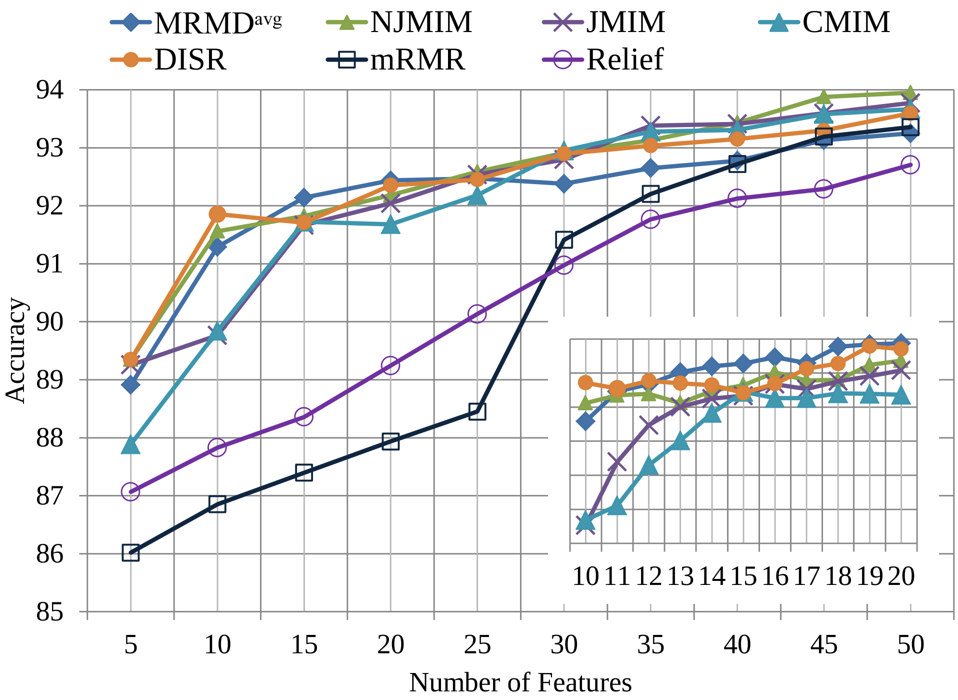

Figure 4.3: Average classification accuracy (in percentage) of SVM and 3-NN on Musk data set

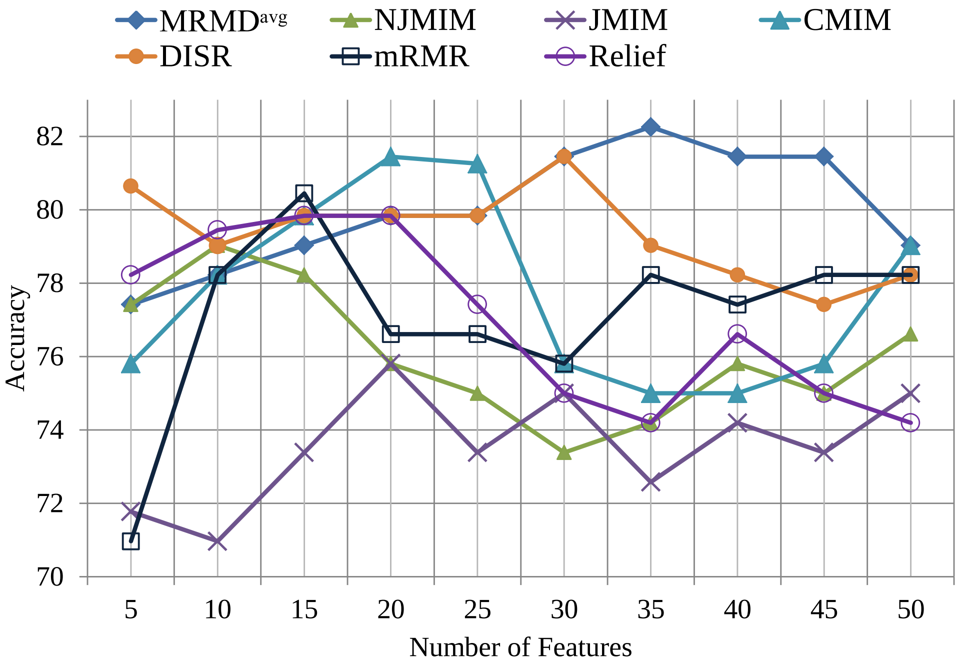

Figure 4.4: Average classification accuracy (in percentage) of SVM and 3-NN on Colon data set

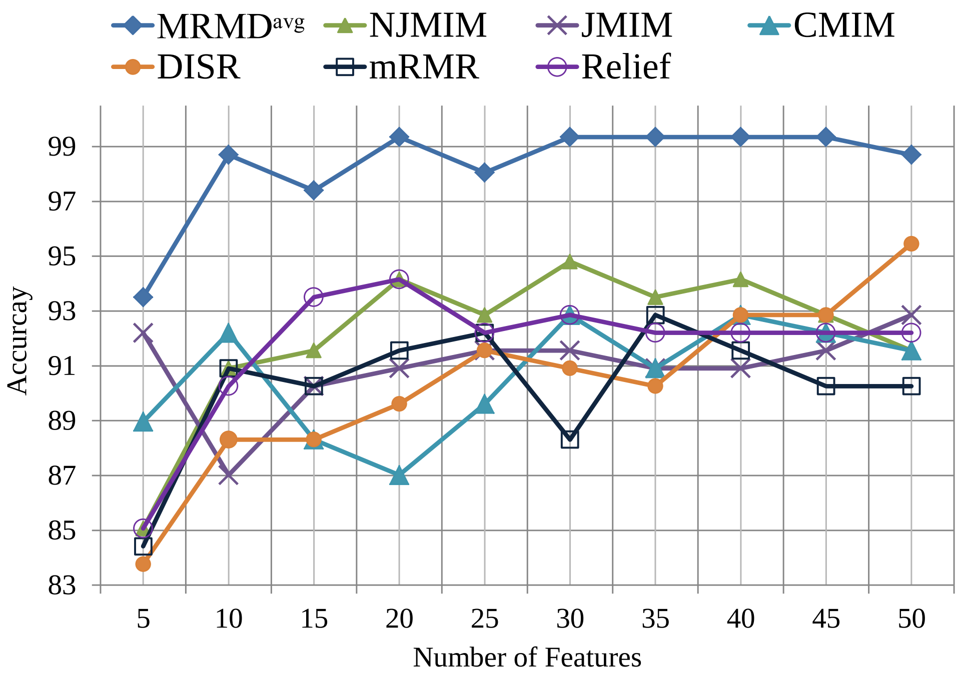

Figure 4.5: Average classification accuracy (in percentage) of SVM and 3-NN on Lymphoma data set

The Urban Land Cover is a multi-class data set with uneven distribution of samples over classes. The proposed algorithm is fundamentally defined for two-class problems where multi-class problems are addressed using one-versus-all approach as mentioned before. This may be the main reason for achieving poor performance on this data set. As an alternative approach, the use of one-versus-one should be investigated for such cases. The second case for which MRMD^{avg} performs poor is the Musk data set for which NJMIM and DISR provide significantly better results than the other methods when average accuracy using 50 feature sets (as given in Table 4.2) is considered. However, referring to Figures 4.3 and Figure 4.8, it is revealed that MRMD^{avg} is top-ranked for some feature subsets. For the two microarray data sets namely, Colon and Lymphoma, both accuracy and AUC curves confirm that the proposed algorithm functions superior to the reference methods.

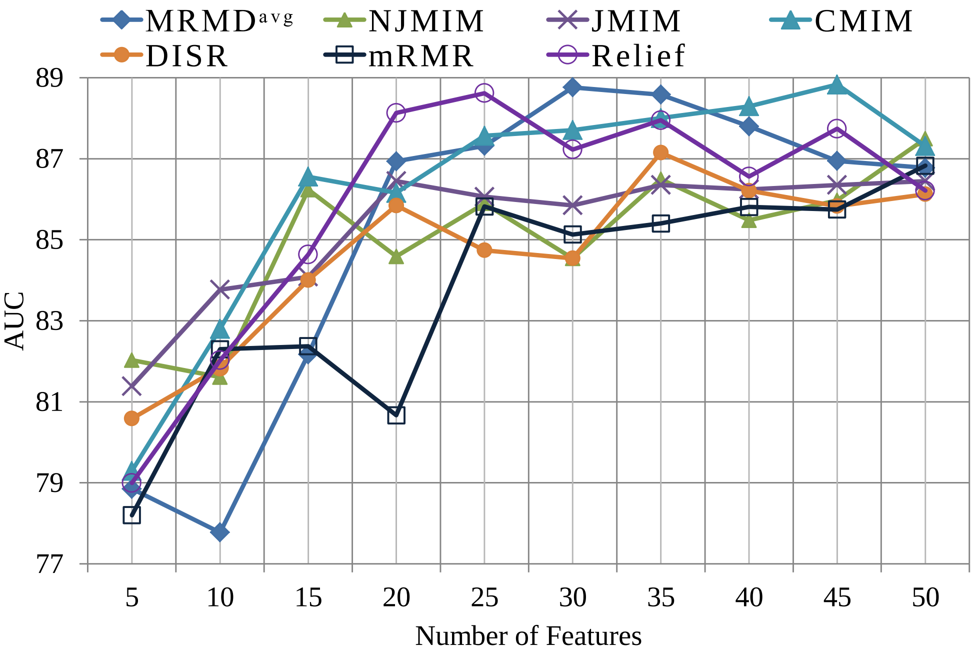

Figure 4.6: Average classification AUC (in percentage) of SVM and 3-NN on Sonar data set

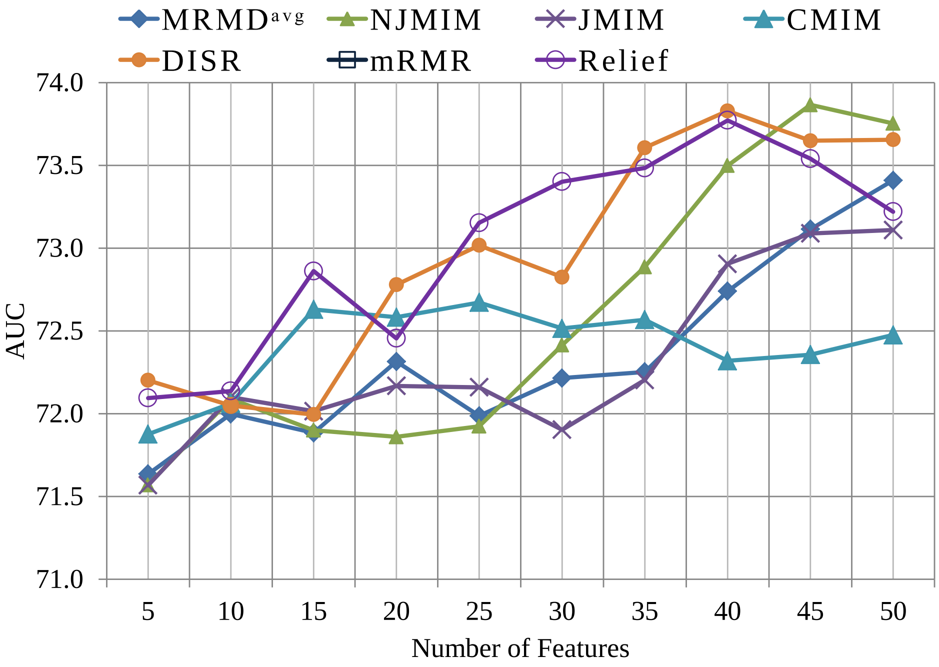

Figure 4.7: Average classification AUC (in percentage) of SVM and 3-NN on Urban Land Cover data set

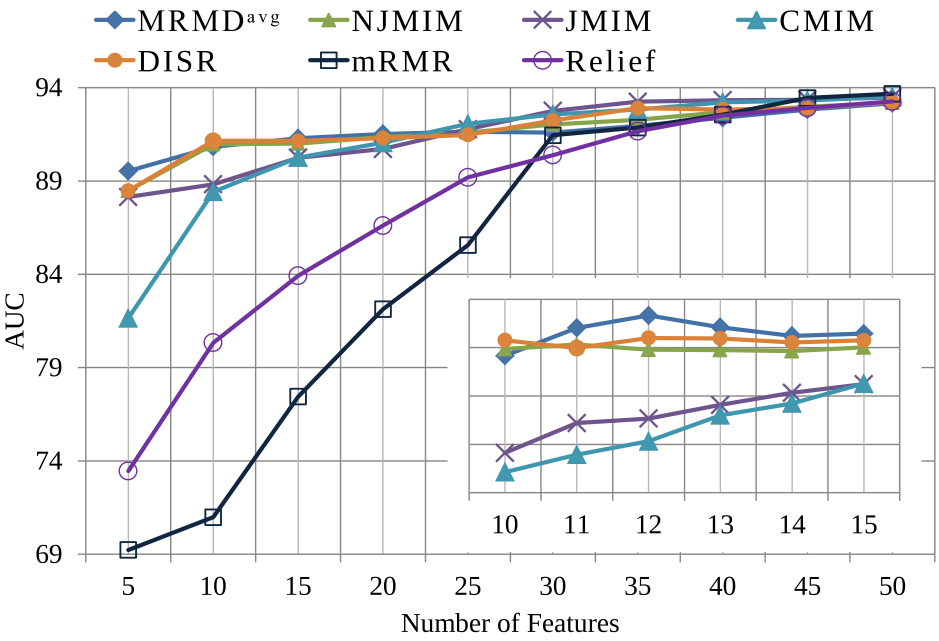

Figure 4.8: Average classification AUC (in percentage) of SVM and 3-NN on Musk data set

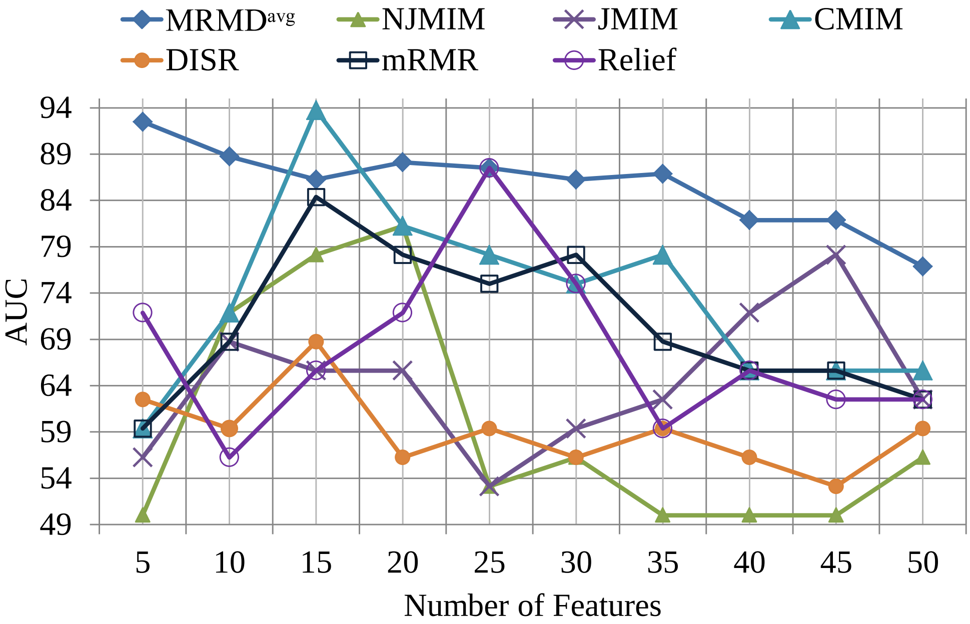

Figure 4.9: Average classification AUC (in percentage) of SVM and 3-NN on Colon data set

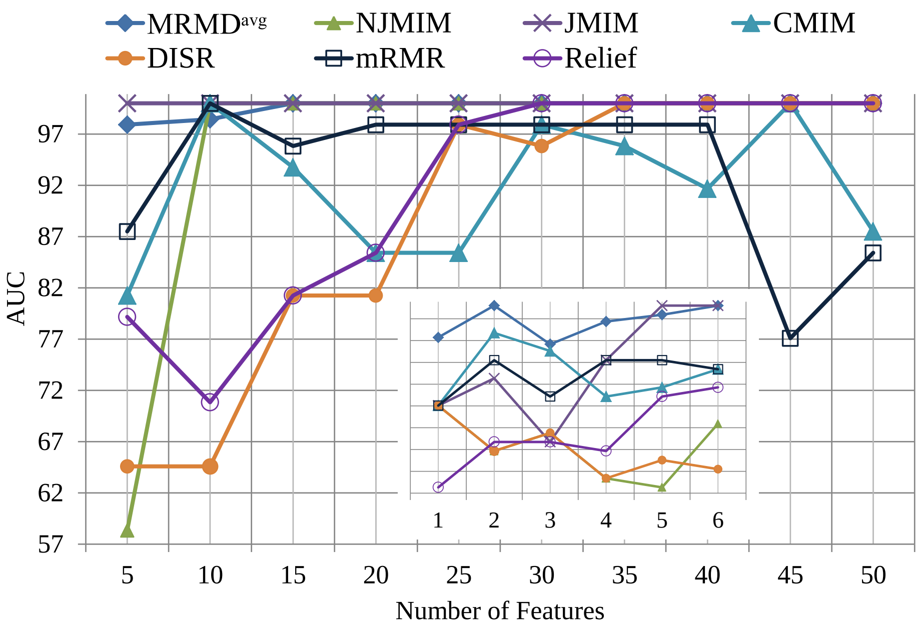

Figure 4.10: Average classification AUC (in percentage) of SVM and 3-NN on Lymphoma data set

In addition to the performance of the classifiers, the stability is also computed for each data set on feature subsets with cardinalities 10, 20 or 50. The main reason is the fact that there is a notable difference among the data sets in terms of their total number of features. Because of this, the cardinality of the original feature set is taken into consideration in determining the cardinality of the feature subsets. More specifically, subsets of κ features are considered where κ is 10 for data sets with less than 50 features (data set 1), 20 for data sets with more than 50 but less than 100 features (data sets 2 and 3) and 50 for data sets with more than 100 features (data sets 4 to 10). Figure 4.11 illustrates the box plot of stabilities representing maximum, median and minimum values as well as the first and the third quartiles. Both versions of the proposed approach provide stability scores comparably better than benchmark filters. Specifically, MRMD^{avg} is the most stable feature selection method.

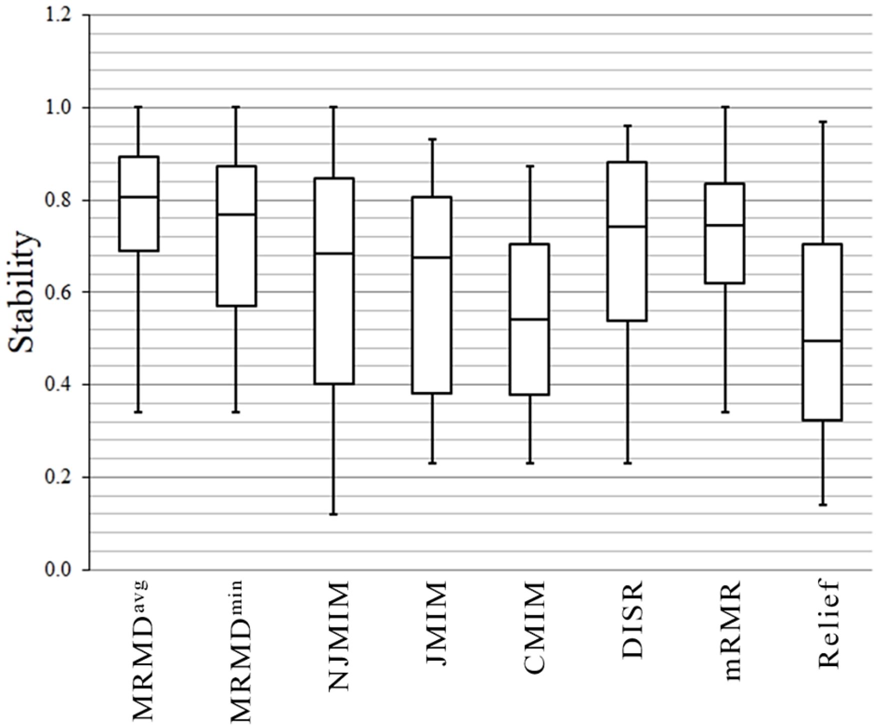

Figure 4.11: Box plots of stability scores obtained for different filters

The performance of MRMD^{avg} is also compared with three widely-used univariate filters, namely t-test, Chi square (χ^2) and mutual information maximization (MIM). The feature selection method proposed in [23] which is based on attributes' AUC obtained from the training data is also implemented. Both SVM and 3-NN are applied on different groups of feature subsets and performance metrics are averaged across classifiers and feature subset groups. Table 4.4 presents the average accuracy and AUC scores for different feature subsets. The cardinality of the feature subsets are given as κ in the table.

Table 4.4: Average AUC and ACC (in percentage) of classifiers on different feature subset groups (in percentage)

κ	**Metric**	**MRMD^{avg}**	**t-test**	χ^2	**MIM**	**AUC**
10	AUC	**84.58**	82.35	82.45	83.12	83.70
	ACC	**83.70**	77.89	78.21	78.27	78.65
20	AUC	**85.29**	83.41	83.68	83.64	84.21
	ACC	**84.06**	78.70	80.03	79.86	80.01
30	AUC	**85.90**	84.33	84.00	83.87	84.36
	ACC	**84.78**	79.83	80.34	80.47	80.45
40	AUC	**86.11**	84.41	84.41	84.03	84.71
	ACC	**85.40**	80.75	80.65	80.75	81.14
50	AUC	**86.85**	84.63	84.15	84.21	84.97
	ACC	**86.35**	81.31	81.22	80.90	81.50
AVG	AUC	**85.75**	83.83	83.74	83.77	84.39
	ACC	**84.86**	79.70	80.09	80.05	80.35

The proposed algorithm outperforms the univariate filter methods both in terms of accuracy and AUC for all feature subsets. Moreover, among the univariate filters, AUC outperforms t-test, χ^2 and MIM. Specifically, recalling that MIM is based on mutual information, it can be concluded that MI is not as effective as AUC in estimating the relevance of features. This provides another evidence on the competence of the proposed method which is principally an AUC-based alternative to MI-based feature selection approaches.

4.4 Experimental Results of Clustering-based Feature Selection using F2F Matrix

The proposed method is evaluated for different number of clusters denoted by κ. For a given κ value, the clustering algorithms are run for all numbers of clusters in the set $\{2,3,\dots,\kappa\}$ and the average performance over all subsets is reported. In the first stage of experiments, the proposed methods namely, *PAM-D*$,\alpha$ and *HC-D*$,\alpha$ are compared with the reference clustering-based feature selection methods. It is worth mentioning that in these experiments, irrelevant features are firstly discarded by applying a threshold on the corresponding relevance measure (α_i) on the train data to ensure that totally irrelevant clusters are not created. Moreover, the window size (w) is set as one tenth of the number of samples in the data set (i.e. $= \frac{K}{10}$), which is truncated to the closest integer.

The clustering-based feature selection methods are characterized by three assets: the dissimilarity (distance) metric they rely on for finding clusters, the clustering algorithm and the relevance metric employed to capture the representative features from clusters. Proposed clustering-based method utilizes the novel F2F scatter frequencies as the dissimilarity measure and the AUC-like relevance measure computed by means of ranks of positives. In order to evaluate the effectiveness of the proposed method, different scenarios are taken into account by considering alternative dissimilarity metrics and different clustering algorithms.

The performance of the proposed dissimilarity metric is compared with some popular metrics namely, Pearson's correlation coefficient (ρ), Spearman's rank correlation coefficient (ρ_r), and symmetric uncertainty (SU). All these similarity measures are utilized for clustering features using either PAM or HC algorithm resulting in six alternative scenarios of clustering-based feature selection. It should be noted that for each reference approach, the same metric is used for both relevance and redundancy. We name these clustering-based feature selection methods as *PAM-*ρ, *PAM-*ρ_r, *PAM-SU*, *HC-*ρ, *HC-*ρ_r, and *HC-SU*, respectively. A brief description of these methods is given below. Note that the number of clusters κ is known in advance since it is equal to the target number of features to be selected.

***PAM-*$\boldsymbol{\rho}$*, HC-*$\boldsymbol{\rho}$**: The similarity between the features x_i and x_j is computed using absolute value of Pearson's correlation coefficient as

$$\rho(x_i, x_j) = \frac{|Cov(x_i,x_j)|}{\sigma_i \sigma_j} \tag{4.2}$$

where $Cov(.,.)$ is the covariance, σ_i and σ_j are the standard deviations of x_i and x_j, respectively. After clustering features by applying PAM/HC algorithm, Pearson's correlation coefficient between each feature and class label y denoted by $\rho(x_i, y)$ is calculated. From each cluster, the feature having the maximum $\rho(x_i, y)$ value is selected as the most relevant.

***PAM-*$\boldsymbol{\rho_r}$*, HC-*$\boldsymbol{\rho_r}$**: The formula for Spearman's rank correlation coefficient is very similar to that of Pearson's but, it is calculated using ranks instead of values. The similarity between x_i and x_j is computed using absolute value of Spearman's correlation coefficient as

$$\rho_r(x_i, x_j) = \frac{|Cov(r_{x_i},r_{x_j})|}{\sigma_{r_{x_i}} \sigma_{r_{x_j}}} \tag{4.3}$$

where the ranks of feature values i.e. r_{x_i} and r_{x_j} are used in computing covariance and standard deviations. Having clustered the features using this similarity measure, the feature with maximum relevance value i.e. $\rho_r(x_i, y)$ is selected as the representative feature.

PAM-SU, HC-SU: Symmetric uncertainty is a normalized version of mutual information (MI), a widely-used correlation measure in information theory. This measure computes the amount of information shared between two features. Unlike MI which is ranged $[0, \infty)$, SU lies in $[0,1]$ and thus can be used as a similarity measure for clustering. Symmetric uncertainty between features x_i and x_j is defined as

$$SU(x_i, x_j) = \frac{2I(x_i; x_j)}{H(x_i) + H(x_j)}, \tag{4.4}$$

where $I(x_i, x_j)$ is the mutual information between features and $H(.)$ stands for entropy. It should be noted that the representative features of clusters are identified using $I(x_i, \mathrm{y})$.

The accuracy rates achieved using different feature subsets are given in Table 4.5. The maximum values in each row are shown in bold and considered as the winners. The ties with the winner are marked by a '*'. It can be observed from the table that the proposed method, *HC-D*$,\alpha$, outperforms the reference approaches on average. More specifically, the accuracy achieved using *HC-D*$,\alpha$ is 84.21% while, by using the second best method, *PAM-D*$,\alpha$, 82.42% accuracy is achieved. This method also provides superior accuracy scores for 7 data sets out of 10 and only loses 2 times without being tied with the winner. Among all methods it can be argued that HC clustering provides relatively higher scores when compared to PAM. Among the dissimilarity measures, SU seems to be superior to the other two alternatives ρ and ρ_r. Nonetheless, the numbers in each row of the table imply that even in the cases when $HC\text{-}D,\alpha$ is not the winner method, it takes a place above the average among all 8 competing schemes.

Table 4.5: Average accuracy of classifiers (in percentage) across all feature subsets

Data	**$PAM\text{-}D,\alpha$**	**$PAM\text{-}\rho$**	**$PAM\text{-}\rho_r$**	**$PAM\text{-}SU$**	**$HC\text{-}D,\alpha$**	**$HC\text{-}\rho$**	**$HC\text{-}\rho_r$**	**$HC\text{-}SU$**
1	78.09	76.57	78.13	76.96	**79.82**	78.75*	77.31	77.20
2	77.33	76.05	76.16	77.19	**78.99**	76.94	77.38	78.13
3	85.93	82.25	83.27	84.41	**87.16**	84.98	84.08	87.15*
4	78.99	80.35	80.74*	79.50	**83.34**	79.94	82.04	80.72*
5	**90.80**	85.27	87.59	87.51	90.25*	82.77	89.64*	90.33
6	**79.81**	78.23*	75.68	72.74	77.16	78.15*	76.10	77.32
7	70.22	67.46	68.28	68.94	72.00	69.64	70.53	**73.02**
8	89.65	88.90	90.07	89.11	**91.46**	86.64	87.81	90.52*
9	80.28	84.84	81.24	84.38	87.21*	**87.40**	84.47	86.22*
10	93.09*	89.99	91.37	92.92	**94.56**	93.83*	91.89	91.30
AVG.	82.42	80.99	81.25	81.37	**84.21**	81.90	82.13	83.19
W/T/L	2/1/7	0/1/9	0/1/9	0/0/10	6/2/2	1/3/6	0/1/9	1/4/5

* corresponds to $p>0.05$

Table 4.6 shows the average AUC scores for different data sets averaged across all subsets. Numbers typed in boldface are the maximum values of each row in the table. It can be seen that AUC values obtained by $HC\text{-}D,\alpha$ are superior to other methods. The second best method is *HC-SU* with average AUC equal to 83.13%. Proposed $HC\text{-}D,\alpha$ improves this AUC by 1% on average reaching 84.17%. When referring to the number of wins, losses and ties, $HC\text{-}D,\alpha$ stands higher than competing approaches by losing only once. The other methods are not very satisfactory in terms of wins. The table shows that proposed dissimilarity measure is generally more effective than the other three alternatives. In other words, the highest average AUC for the methods based on PAM is acheived by $PAM\text{-}D,\alpha$. This is also the case for the methods based on HC as $HC\text{-}D,\alpha$ is superior to the competing methods. Another important consequence of the results given in Table 4.5 and 4.6 is that HC is more successful as a clustering algorithm for clustering-based feature selection. Accuracy and AUC scores of the methods based on HC are relatively higher than their counterparts using PAM.

Table 4.6: Average AUC of classifiers (in percentage) across all feature subsets

Data	$PAM\text{-}D,\alpha$	$PAM\text{-}\rho$	$PAM\text{-}\rho_r$	$PAM\text{-}SU$	$HC\text{-}D,\alpha$	$HC\text{-}\rho$	$HC\text{-}\rho_r$	$HC\text{-}SU$
1	77.52	75.87	78.11*	77.76	**78.44**	76.46	78.05	77.98
2	85.30*	84.41	85.43*	85.41	**86.58**	86.13*	84.06	86.17*
3	71.22	68.21	69.41	70.37	69.40	68.19	68.77	**73.63**
4	72.99	73.08	73.45	73.06	**75.73**	73.00	74.12*	73.57
5	83.98	84.23	**85.34**	82.38	85.22*	84.94	83.06	84.43
6	76.36	75.20	71.87	67.56	**78.08**	74.15	74.64	77.46*
7	80.83*	81.12*	79.74	79.19	80.42	78.42	**81.18**	79.24
8	94.10	94.97*	93.28	93.37	93.78	94.82	94.49	**95.17**
9	93.57	93.98	**96.82**	89.60	96.08*	94.27	95.26*	92.75
10	92.77	91.19	93.54	95.60	**98.01**	94.61	90.43	90.92
AVG.	82.86	82.23	82.70	81.43	**84.17**	82.50	82.41	83.13
W/T/L	0/2/8	0/2/8	2/2/6	0/1/9	5/2/3	0/1/9	1/2/7	2/2/6

* corresponds to p>0.05

In addition to the alternative clustering-based methods, $HC\text{-}D,\alpha$ feature selection method is also compared with some univariate filter appraoches including t-test, χ^2, MIM and AUC. Table 4.7 represents accuracy and AUC scores for different sizes of feature subsets given as κ. This table shows that clustering-based feature selection is more effective on the data sets of small sizes. In fact, feature selection using AUC generally outperforms proposed *HC-D*$,\alpha$ in terms of AUC for all feature subset sizes. Although the highest AUC in the table (84.97%) belongs to feature selection by using AUC score of individual features at $\kappa = 50$, *HC-D*$,\alpha$ achieves comparable AUC, i.e. 84.85% at $\kappa = 30$. It can be argued that if *HC-D*$,\alpha$ provides a good trade-off between the size of the subset, AUC and accuracy.

Table 4.7: Average AUC and ACC (in percentage) of classifiers on different feature subset groups

$\boldsymbol{\kappa}$	**Metric**	***HC-D*$,\boldsymbol{\alpha}$**	**t-test**	$\boldsymbol{\chi^2}$	**MIM**	**AUC**
10	AUC	83.55	82.35	82.45	83.12	**83.70**
	ACC	**83.20**	77.89	78.21	78.27	78.65
20	AUC	84.14	83.41	83.68	83.64	**84.21**
	ACC	**83.91**	78.70	80.03	79.86	80.01
30	AUC	**84.85**	84.33	84.00	83.87	84.36
	ACC	**84.69**	79.83	80.34	80.47	80.45
40	AUC	84.53	84.41	84.41	84.03	**84.71**
	ACC	**84.33**	80.75	80.65	80.75	81.14
50	AUC	84.17	84.63	84.15	84.21	**84.97**
	ACC	**84.20**	81.31	81.22	80.90	81.50
AVG	AUC	84.25	83.83	83.74	83.77	**84.39**
	ACC	**84.07**	79.70	80.09	80.05	80.35

For the low dimensional data sets with relatively larger number of samples, clustering-based methods provide superior results as also observed in Table 4.5 and 4.6. For the high dimensional data sets however, it can be argued that univariate methods are superior to clustering-based approaches specifically in terms of AUC. The seemingly inconsistent scores in Table 4.7 when accuracy and AUC are taken into account can also be explained by referring to the data set characteristics. Note that $HC\text{-}D, \alpha$ performs better on small data sets for which both AUC and accuracy can be assumed as reliable measures. On the other hand, the high dimensional data sets considered in this study are microarray data sets having small number of samples per class. Thus, even a slight difference in the number of samples per class results in remarkable changes in the interpretation of accuracy and AUC. It can be argued that for those data sets, AUC is a more reliable performance measure.

To provide a clearer insight, the HC-based feature selection approaches are compared with 4 MI-based methods in terms of accuracy, AUC and stability. Table 4.8 represents the average accuracy and AUC across all data sets. The test of significance is applied to check if the winner method (shown in bold) meaningfully outperforms the competing ones. This table reveals that, when both AUC and accuracy are considered, proposed clustering-based approach performs comparable to CMIM.

Table 4.8: Average accuracy and AUC of classifiers (in percentage) across all data sets and all feature subsets

	$\boldsymbol{HC\text{-}D, \alpha}$	$\boldsymbol{HC\text{-}\rho}$	$\boldsymbol{HC\text{-}\rho_r}$	$\boldsymbol{HC\text{-}SU}$	**mRMR**	**CMIM**	**JMIM**	**NJMIM**
AUC	84.17*	82.50	82.41	83.13	81.97	**84.20***	82.68	82.91
ACC	84.21*	81.90	82.13	83.19	84.10*	84.43*	84.06	**84.54**

* corresponds to p>0.05

Figure 4.12 illustrates the box plots of stability index. Referring to the figure, it can be seen that the proposed method *HC-D,α* and *mRMR* are the most stable approaches. It is also worth mentioning that CMIM which performs well in terms of performance metrics accuracy and AUC, is not satisfactory in terms of stability index. It is in fcat one of the least stable feature selection methods in Figure 4.12.

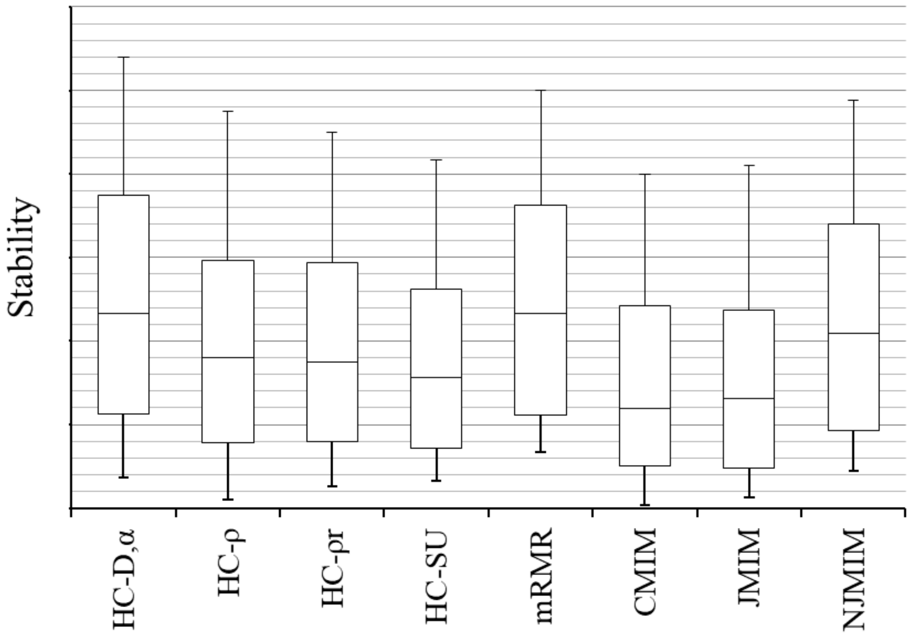

Figure 4.12: Stability measure for different feature selection methods.

4.4 Discussions

The experimental results are summarized in Table 4.9 and 4.10 for accuracy and AUC respectively. The aim is to shed light on the overall performance of MRMD^{avg} and *HC-D,α* compared to one another and also compared to the best performing methods implemented for comparison. An alternative MI-based method and a clustering-based method are taken into account in these tables. The selected methods for concluding the results are the best performing methods in terms of average accuracy and AUC for Table 4.9 and 4.10 respectively. In the tables, the number of features (dimensionality) of the data sets is also given. The highest accuracy and AUC of each row is marked in bold.

Table 4.9: Average accuracy of classifiers for each data set across all feature subsets (comparing the best performing methods)

Data	**#Features**	**MRMD^{avg}**	***HC-D,α***	**NJMIM**	***HC-SU***
1	44	78.21	**79.82**	77.82	77.20
2	60	77.72	**78.99**	78.38	78.13
3	64	86.50	**87.16**	85.93	87.15
4	147	77.13	**83.34**	79.45	80.72
5	166	91.83	90.25	**92.16**	90.33
6	2000	**79.81**	77.16	75.68	77.32
7	2905	**78.22**	72.00	73.02	73.02
8	3571	**96.99**	91.46	93.43	90.52
9	7129	**98.05**	87.21	91.75	86.22
10	12533	**99.03**	94.56	97.77	91.30
AVG.	2862	**86.35**	84.21	84.54	83.19

Table 4.9 reveals that the method resulting in the highest accuracy is one of the proposed methods for 9 data sets out of 10. As stated before, it can be argued that for lower dimensional data sets, i.e. data set 1 to 5, proposed *HC-D*, α is superior to the competing methods. For the high dimensional data sets however, MRMD^{avg} performs significantly better than other methods. On average, MRMD^{avg} is the best performing method in terms of accuracy.

Table 4.10: Average AUC of classifiers for each data set across all feature subsets (comparing the best performing methods)

Data	**#Features**	**MRMD^{avg}**	***HC-D, α***	**CMIM**	***HC-SU***
1	44	**79.30**	78.44	77.51	77.98
2	60	84.82	**86.58**	85.55	86.17
3	64	68.77	**69.40**	68.55	73.63
4	147	72.12	**75.73**	71.27	73.57
5	166	**91.07**	85.22	90.01	84.43
6	2000	**86.53**	78.08	76.75	77.46
7	2905	**87.79**	80.42	79.60	79.24
8	3571	99.08	93.78	**99.57**	95.17
9	7129	**99.50**	96.08	93.63	92.75
10	12533	99.50	98.01	**99.58**	90.92
AVG.	2862	**86.85**	84.17	84.20	83.13

Considering AUC scores in Table 4.10 reveals that a similar pattern exists for the AUC. Note that NJMIM is replaced by CMIM since CMIM provides higher AUC scores when compared to NJMIM. It can be seen that, for the low dimensional data sets, *HC-D, α* provides the highest AUC scores and as the dimensionality increases, MRMD^{avg} takes the winner part resulting in superior results. On average terms, however, MRMD^{avg} outperforms the alternative approaches.

The average accuracy and AUC for 3NN and SVM are shown in Table 4.11 for MRMD^{avg}, *HC-D, α*, and the three competing methods considered in Tables 4.9 and 4.10. This table represents a clear picture of how the two proposed methods contribute to the performance improvement for each individual classifier. MRMD^{avg}

is the best performing method for both 3NN and SVM in terms of accuracy and AUC. In general, 3NN results in higher accuracy and SVM in higher AUC. It can be explained by the nonlinear characteristics of AUC which may result in an inappropriate order of classification scores and degraded AUC. In addition, since the some of the data sets considered in this study are not linearly separable, accuracy of SVM in general is lower than that of 3NN.

Table 4.11: Average accuracy and AUC of classifiers across all data sets and across all feature subsets for 3NN and SVM (comparing the best performing methods)

Metric	**Classifier**	**MRMD^{avg}**	***HC-D*, $\boldsymbol{\alpha}$**	**NJMIM**	**CMIM**	***HC-SU***
ACC	3NN	**88.11**	85.68	86.39	86.18	84.89
	SVM	**84.58**	82.74	82.68	82.73	81.50
AUC	3NN	**83.74**	82.54	81.22	81.31	81.05
	SVM	**89.95**	85.80	84.60	87.08	85.22

Chapter 5

CONCLUSIONS AND FUTURE WORK

5.1 Conclusions

In this study, novel feature selection approaches based on the ranks of positive instances are proposed. In MRMD, feature score is defined as the summation of two terms namely, relevance and diversity. Analogous to the calculation of AUC, both relevance and diversity terms are based on ranks of positive instances. Relevance is an absolute term indicating the discriminate power of an individual feature. On the other hand, diversity is a relative term measuring the complementarity of each candidate feature with respect to the already selected feature. The so-called maximum-relevance and maximum-diversity (MRMD) algorithm searches for the maximal feature score in a greedy scheme similar to the other multivariate MI-based feature selection methods. We proposed two variants called MRMD^{avg} and MRMD^{min}. In MRMD^{avg}, diversity of a feature is computed as the average of the diversities between the feature and the already selected feature set. MRMD^{min} computes the diversity score as the minimum diversity between the candidate and the already selected set of features.

Experiments are conducted on 10 widely-used data sets from UCI machine learning repository and microarray gene expression data sets. In order to explore the efficiency of the proposed method, SVM and 3-NN are applied on various feature subsets selected by each of the feature selection methods and average values of both accuracy and AUC are reported. Proposed method is compared with 6 different multivariate (NJMIM, JMIM, CMIM, DISR, mRMR and Relief) and 3 different univariate filters (t-test, χ^2, and MIM). Experimental results confirm that MRMD^{avg} generally outperforms other algorithms, especially on data sets having high dimensional feature vectors and small number of samples. In addition, the proposed algorithm achieves the highest level of stability when compared to other multivariate filters.

The second approach proposed in this study is a new dissimilarity metric for clustering-based feature selection. The metric aims to reduce the dominance of highly different ranks to the overall dissimilarity score by focusing on similarities of the features in a smaller range of rank differences. The notion of F2F scatter matrix is adapted by creating local clusters of neighboring features at the sample level and then obtaining scatter frequencies. After clustering features using the proposed metric, the feature having the highest AUC score (measured in terms of ranks of positive samples) is selected from each cluster. Two clustering algorithms, namely PAM and HC are taken into account in the experiments. Alternative dissimilarity measures include Pearson's correlation coefficient, Spearman's rank correlation coefficient and symmetric uncertainty. Proposed method is also compared with mRMR, CMIM, JMIM and MJMIM. Experimental results have verified the effectiveness of the proposed metric when compared to state-of-the-art measures in terms of average accuracy and AUC of 3NN and SVM classifiers.

The two schemes proposed in this study are also compared to one another considering the accuracy and AUC achieved for each data sets. The comparisons reveal that for lower dimensional data sets $HC\text{-}D, \alpha$ is preferred to MRMD$^{\text{avg}}$. However, as the number of features increases, MRMD$^{\text{avg}}$ outperforms $HC\text{-}D, \alpha$.

5.2 Future Work

There are different paths which can be taken as the future research. Firstly, the proposed F2F dissimilarity metric can be combined with a consistent relevance measure to estimate its performance when it is used in greedy search-based selection schemes. The major issue that needs to be addressed will be normalization of the metric such that it becomes compatible with the range of relevance, i.e. α_i. Similarly, the diversity measure in MRMD objective function can be used as dissimilarity metric in a clustering-based feature selection method.

Another path that is going to be followed in the future studies is to develop dimensionality reduction methods by encoding neighborhood information. By employing the context of proximity in terms of ranks, neighborhood information can be encoded into specific patterns. These patterns will represent the neighboring samples within a proximity window. By adopting a proper coding, a learning method then can be applied to learn a mapping which projects neighboring samples into close coordinates in the reduced subspace.

REFERENCES

[1] N. S. Mohamed, S. Zainudin, and Z. A. Othman, “Metaheuristic approach for an enhanced mRMR filter method for classification using drug response microarray data,” *Expert Syst. Appl.*, vol. 90, pp. 224–231, 2017.

[2] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, “Recent advances and emerging challenges of feature selection in the context of big data,” *Knowledge-Based Syst.*, vol. 86, pp. 33–45, 2015.

[3] K. Zhang, L. Zhang, and M. H. Yang, “Real-Time Object Tracking Via Online Discriminative Feature Selection,” *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4664–4677, 2013.

[4] D. Agnihotri, K. Verma, and P. Tripathi, “Variable Global Feature Selection Scheme for automatic classification of text documents,” *Expert Syst. Appl.*, vol. 81, pp. 268–281, 2017.

[5] D. S. Guru, M. Suhil, L. N. Raju, and N. V. Kumar, “An alternative framework for univariate filter based feature selection for text categorization,” *Pattern Recognit. Lett.*, vol. 103, pp. 23–31, 2018.

[6] M. Abdelwahab and C. Busso, “Ensemble feature selection for domain adaptation in speech emotion recognition,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5000–5004.

[7] H. Bhaskar, D. C. Hoyle, and S. Singh, “Machine learning in bioinformatics: A brief survey and recommendations for practitioners,” *Comput. Biol. Med.*, vol. 36, no. 10, pp. 1104–1125, 2006.

[8] Y. Saeys, I. Inza, and P. Larrañaga, “A review of feature selection techniques in bioinformatics,” *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[9] Y. Cui, C.-H. Zheng, J. Yang, and W. Sha, “Sparse maximum margin discriminant analysis for feature extraction and gene selection on gene expression data,” *Comput. Biol. Med.*, vol. 43, no. 7, pp. 933–941, 2013.

[10] K. Shin and S. Miyazaki, “A Fast and Accurate Feature Selection Algorithm Based on Binary Consistency Measure,” *Comput. Intell.*, vol. 32, no. 4, pp. 646–667, 2016.

[11] H. Wei and S. A. Billings, “Feature Subset Selection and Ranking for Data Dimensionality Reduction,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 162–166, 2007.

[12] G. Chandrashekar and F. Sahin, “A survey on feature selection methods,” *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16–28, 2014.

[13] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[14] Z. Zeng, H. Zhang, R. Zhang, and C. Yin, “A novel feature selection method considering feature interaction,” *Pattern Recognit.*, vol. 48, no. 8, pp. 2656–2666, 2015.

[15] J. M. Arevalillo and H. Navarro, “Exploring correlations in gene expression microarray data for maximum predictive–minimum redundancy biomarker selection and classification,” *Comput. Biol. Med.*, vol. 43, no. 10, pp. 1437–1443, 2013.

[16] A. K. Shukla, “Identification of cancerous gene groups from microarray data by employing adaptive genetic and support vector machine technique,” *Comput. Intell.*, pp. 1–30, 2016.

[17] X. Zhu, Y. Wang, Y. Li, Y. Tan, G. Wang, and Q. Song, “A new unsupervised feature selection algorithm using similarity-based feature clustering,” *Comput. Intell.*, vol. 35, no. 1, pp. 2–22, 2019.

[18] R. Wang and K. Tang, “Feature Selection for Maximizing the Area Under the ROC Curve,” 2009, pp. 400–405.

[19] L. Sun, J. Wang, and J. Wei, “AVC: Selecting discriminative features on basis of AUC by maximizing variable complementarity,” *BMC Bioinformatics*, vol. 18, no. 3, p. 50, 2017.

[20] S. Meyen, “Relation between classification accuracy and mutual information in equally weighted classification tasks,” University of Hamburg, 2016.

[21] S. Garc\'\ia, J. Luengo, J. A. Sáez, V. López, and F. Herrera, “A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning,” *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 4, pp. 734–750, Apr. 2013.

[22] I. Kononenko and I. Bratko, “Information-based evaluation criterion for classiffer’s performance,” *Mach. Learn.*, vol. 6, no. 1, pp. 67–80, 1991.

[23] X. Chen and M. Wasikowski, “FAST: A Roc-based Feature Selection Metric for Small Samples and Imbalanced Data Classification Problems,” 2008, pp. 124–132.

[24] Y. S. Son and J. Baek, “A modified correlation coefficient based similarity measure for clustering time-course gene expression data,” *Pattern Recognit. Lett.*, vol. 29, no. 3, pp. 232–242, 2008.

[25] A. K. Shukla and D. Tripathi, “Identification of potential biomarkers on microarray data using distributed gene selection approach,” *Math. Biosci.*, vol. 315, p. 108230, 2019.

[26] M. Kotlyar, S. Fuhrman, A. Ableson, and R. Somogyi, “Spearman Correlation Identifies Statistically Significant Gene Expression Clusters in Spinal Cord Development and Injury,” *Neurochem. Res.*, vol. 27, no. 10, pp. 1133–1140, Oct. 2002.

[27] J. Wang, J.-M. Wei, Z. Yang, and S.-Q. Wang, “Feature selection by

maximizing independent classification information," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 4, pp. 828–841, 2017.

[28] M. Bennasar, Y. Hicks, and R. Setchi, "Feature selection using Joint Mutual Information Maximisation," *Expert Syst. Appl.*, vol. 42, no. 22, pp. 8520–8532, 2015.

[29] L. Hu, W. Gao, K. Zhao, P. Zhang, and F. Wang, "Feature selection considering two types of feature relevancy and feature interdependency," *Expert Syst. Appl.*, vol. 93, pp. 423–434, 2018.

[30] N. D. Cilia, C. De Stefano, F. Fontanella, and A. Scotto di Freca, "A ranking-based feature selection approach for handwritten character recognition," *Pattern Recognit. Lett.*, vol. 121, pp. 77–86, Apr. 2019.

[31] Q. Song, J. Ni, and G. Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 1, pp. 1–14, 2013.

[32] M. García-Torres, F. Gómez-Vela, B. Melián-Batista, and J. M. Moreno-Vega, "High-dimensional feature selection via feature grouping: A Variable Neighborhood Search approach," *Inf. Sci. (Ny).*, vol. 326, pp. 102–118, 2016.

[33] X. Huang, L. Zhang, B. Wang, F. Li, and Z. Zhang, "Feature clustering based support vector machine recursive feature elimination for gene selection," *Appl. Intell.*, vol. 48, no. 3, pp. 594–607, 2018.

[34] Z. Dehghan and E. G. Mansoori, “A new feature subset selection using bottom-up clustering,” *Pattern Anal. Appl.*, vol. 21, no. 1, pp. 57–66, Feb. 2018.

[35] M. Rahmaninia *et al.*, “Individual Comparisons by Ranking Methods,” *Pattern Recognit. Lett.*, vol. 29, no. 6, pp. 848–855, Mar. 2017.

[36] M. Rahmaninia and P. Moradi, “OSFSMI: Online stream feature selection method based on mutual information,” *Appl. Soft Comput.*, vol. 68, pp. 733–746, 2018.

[37] Z. Chen *et al.*, “Feature selection with redundancy-complementariness dispersion,” *Knowledge-Based Syst.*, vol. 89, pp. 203–217, 2015.

[38] A. Das and S. Das, “Feature weighting and selection with a Pareto-optimal trade-off between relevancy and redundancy,” *Pattern Recognit. Lett.*, vol. 88, pp. 12–19, 2017.

[39] J. Che, Y. Yang, L. Li, X. Bai, S. Zhang, and C. Deng, “Maximum relevance minimum common redundancy feature selection for nonlinear data,” *Inf. Sci. (Ny).*, vol. 409--410, pp. 68–86, 2017.

[40] X. Tang, Y. Dai, P. Sun, and S. Meng, “Interaction-based feature selection using Factorial Design,” *Neurocomputing*, vol. 281, pp. 47–54, 2018.

[41] H. Peng and Y. Fan, “Feature selection by optimizing a lower bound of

conditional mutual information," *Infromation Sci.*, vol. 418--419, pp. 652–667, 2017.

[42] B. S. Chlebus and S. H. Nguyen, "On Finding Optimal Discretizations for Two Attributes," in *Rough Sets and Current Trends in Computing*, 1998, pp. 537–544.

[43] J. M. Sotoca and F. Pla, "Supervised feature selection by clustering using conditional mutual information-based distances," *Pattern Recognit.*, vol. 43, no. 6, pp. 2068–2081, 2010.

[44] M. Savić, V. Kurbalija, Z. Bosnić, and M. Ivanović, "Feature selection based on community detection in feature correlation networks," *Computing*, vol. 101, no. 10, pp. 1513–1538, Oct. 2019.

[45] Daxin Jiang, Chun Tang, and Aidong Zhang, "Cluster analysis for gene expression data: a survey," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 11, pp. 1370–1386, Nov. 2004.

[46] T. Kim, I. R. Chen, Y. Lin, A. Y.-Y. Wang, J. Y. H. Yang, and P. Yang, "Impact of similarity metrics on single-cell RNA-seq data clustering," *Brief. Bioinform.*, 2018.

[47] L. Yu, C. Ding, and S. Loscalzo, "Stable feature selection via dense feature groups," pp. 803–811.

[48] R. Jörnsten and B. Yu, "Simultaneous gene clustering and subset selection for sample classification via MDL," *Bioinformatics*, vol. 19, no. 9, pp. 1100–1109, 2003.

[49] F. Fleuret, "Fast Binary Feature Selection with Conditional Mutual Information," *J. Mach. Learn. Res.*, vol. 5, pp. 1531–1555, 2004.

[50] I. Kojadinovic, "Relevance measures for subset variable selection in regression problems based on k-additive mutual information," *Comput. Stat. Data Anal.*, vol. 49, no. 4, pp. 1205–1227, 2005.

[51] P. E. Meyer and G. Bontempi, "On the Use of Variable Complementarity for Feature Selection in Cancer Classification," in *Applications of Evolutionary Computing*, 2006, pp. 91–102.

[52] W. Liu, S. Liu, Q. Gu, X. Chen, and D. Chen, "FECS: A Cluster Based Feature Selection Method for Software Fault Prediction with Noises," in *Proceedings - International Computer Software and Applications Conference*, 2015, vol. 2, pp. 276–281.

[53] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[54] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 3, pp. 299–310, 2005.

[55] R. C. Prati, G. E. A. P. A. Batista, and M. C. Monard, “A Survey on Graphical Methods for Classification Predictive Performance Evaluation,” *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 11, pp. 1601–1618, 2011.

[56] D. J. Hand and R. J. Till, “A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems,” *Mach. Learn.*, vol. 45, no. 2, pp. 171–186, Nov. 2001.

[57] L. Yan, R. Dodier, M. C. Mozer, and R. Wolniewicz, “Optimizing Classifier Performance via an Approximation to the Wilcoxon-Mann-Whitney Statistic,” in *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, 2003, pp. 848–855.

[58] H. A. Güvenir and M. Kurtcephe, “Ranking Instances by Maximizing the Area under ROC Curve,” *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 10, pp. 2356–2366, 2013.

[59] H. B. Mann and D. R. Whitney, “On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other,” *Ann. Math. Stat.*, vol. 18, no. 1, pp. 50–60, 1947.

[60] S. J. Mason and N. E. Graham, “Areas beneath the relative operating characteristics {(ROC)} and relative operating levels {(ROL)} curves: Statistical significance and interpretation,” *Q. J. R. Meteorol. Soc.*, vol. 128, no. 584, pp. 2145–2166, 2002.

[61] C. Wiwie, J. Baumbach, and R. Röttger, “Comparing the performance of biomedical clustering methods,” *Nat. Methods*, vol. 12, no. 11, pp. 1033–1038, 2015.

[62] L. Kaufman and P. J. Rousseeuw, *Multivariate analysis: methods and applications*. Wiley series in probability and mathematical statistics, Wiley, 1990.

[63] T. S. Madhulatha, “Comparison between k-means and k-medoids clustering algorithms,” 2011, pp. 472–481.

[64] J. H. W. Jr., “Hierarchical Grouping to Optimize an Objective Function,” *J. Am. Stat. Assoc.*, vol. 58, no. 301, pp. 236–244, 1963.

[65] D. Defays, “An efficient algorithm for a complete link method,” *Comput. J.*, vol. 20, no. 4, pp. 364–366, 1977.

[66] L. I. Kuncheva, “A Stability Index for Feature Selection,” in *Proceedings of the 25th Conference on Proceedings of the 25th IASTED International Multi-Conference: Artificial Intelligence and Applications*, 2007, pp. 390–395.

[67] K. Kira and L. A. Rendell, “The Feature Selection Problem: Traditional Methods and a New Algorithm,” in *Proceedings of the Tenth National Conference on Artificial Intelligence*, 1992, pp. 129–134.

[68] F. Wilcoxon, “Individual Comparisons by Ranking Methods,” *Biometrics*

Bull., vol. 1, no. 6, pp. 80–83, 1945.