An Empirical Study of Stability of Feature Selection Algorithms

Lei Yu, Yue Han, and Steven Loscalzo

Department of Computer Science
Binghamton University
P.O. Box 6000
Binghamton, NY 13902-6000, USA
{lyu,yhan1,sloscal1}@binghamton.edu

Abstract. Stability is an important yet under-addressed issue for feature selection from high-dimensional data. In this paper, we investigate two causes of instability of feature selection: small sample size and elimination of redundant features. We propose a general stability measure which takes into account feature correlation when assessing the similarity of two feature subsets or two sets of feature groups. We empirically evaluate the stability of several representative feature selection and feature groups selection algorithms, and discuss the merits of the proposed measure and the impact of feature redundancy on stability based on stability profiles of these algorithms on microarray data sets.

Key words: Feature selection, stability, high-dimensional data, classification

1 Introduction

A great variety of feature selection algorithms have been developed and shown to be effective in reducing data dimensionality and improving predictive accuracy for classification in many applications [15]. However, a relatively neglected issue is the stability of feature selection algorithms. The issue is particularly important for knowledge discovery from high-dimensional data, where the goal is often to identify features best explaining the differences between classes or subsets of samples from thousands of features. For example, in biological applications (e.g., microarrays, mass spectrometry), the primary goal of domain experts in conducting high-throughput experiments is often to detect leads for some biologically relevant "marker" genes or proteins, rather than building models for predicting diseases or phenotypes of novel samples [16]. Although many feature selection algorithms are effective in selecting a subset of predictive features for sample class prediction, they are not necessarily reliable to identify candidate features for subsequent costly biological and clinical validation. A feature selection algorithm may select largely different subsets of features with similarly good prediction when different subsets of samples are used as training data [4, 11]. Such instability of a feature selection algorithm under training data

variations dampens the confidence of domain experts in investigating any of the various subsets of predictive features selected by the same algorithm.

The reason for the instability of feature selection algorithms is bifold. On one hand, due to the relatively small number of samples in high-dimensional data, the set of available samples provide a very rough approximation to original data distribution. Therefore, the estimate of feature relevance is greatly influenced by the subset of samples in the training set. A different subset of features may be selected by the same algorithm when there is a slight variation in the training set. On the other hand, the classic goal of feature selection aims to select a minimum subset of features necessary for constructing a classifier of best predictive accuracy [12, 13, 20]. Many feature selection algorithms thus discard features which are relevant to the class but redundant to the selected ones. Among a set of highly correlated features, different ones may be selected under different settings of a feature selection algorithm.

Some recent papers have studied the stability of feature selection algorithms under variations in the training data with a small number of samples [4, 11], but have not addressed the effect of different ways of handling redundant features on the stability of feature selection results. In this paper, we systematically study the stability of feature selection and feature groups selection algorithms, and investigate both causes of instability of feature selection algorithms identified above. In order to do so, a general stability measure is needed which takes into account feature correlation when assessing the similarity of two feature subsets or two sets of feature groups.

The rest of the paper is organized as follows. In Section 2, we review previous work in feature selection with contrast to our work. In Section 3, we propose a general measure of stability. In Section 4, we introduce representative algorithms studied in this paper. Section 5 describes data sets used, evaluation procedure, and algorithm settings. Section 6 provides experimental results and discusses the stability and predictive accuracy of the representative algorithms. Finally, Section 7 concludes this paper and identifies some future research directions.

2 Related Work

Feature selection has been generally viewed as a problem of searching for an optimal subset of features guided by some evaluation measures. Various feature selection algorithms can broadly fall into the filter model and the wrapper model depending on their evaluation measures [12]. Filter algorithms use measures of intrinsic data characteristics [15], and wrapper algorithms rely on the performance of a predefined learning algorithm to evaluate the goodness of a subset of features [12]. For high-dimensional data, filter algorithms are often preferred due to their computational efficiency. As to search strategy, a simple way of search is to evaluate each feature independently and form a subset based on top-ranked features. Such univariate methods have been shown effective in some applications [7, 14]. However, they do not work well when features highly correlate or interact with each other. Various algorithms based on feature subset

search evaluate features in a subset together and select a small subset of relevant but non-redundant features [1, 5, 20]. They have shown improved classification accuracy over univariate methods. Another way of search is to weight all features together according to the concept of maximum margin and form a subset based on top-ranked features [8, 17]. An advantage is that optimal weights of features can be estimated by considering features together. All work discussed above focuses on the generalization ability of feature selection algorithms without addressing their stability. In contrast, our paper systematically studies the stability of feature selection algorithms by defining a general stability measure and evaluating representative algorithms in each category.

Clustering has recently been applied to feature selection, by clustering features and then selecting one (or a few) representative features from each cluster [2, 3, 9], or simultaneously clustering and selecting features [10], to form a final feature subset. Intuitively, clustering features can illuminate relationships among features and facilitate feature redundancy analysis; a feature is more likely to be redundant to some other features in the same cluster than features in other clusters. Some of these algorithms [9, 10] produce feature selection result in the form of a set of feature groups each consisting of features relevant to the class but highly correlated to each other, instead of the traditional form of a single subset of features. Such set of predictive feature groups not only generalizes well but also provides additional informative group structure for domain experts to further investigate. Moreover, the stability may be improved by retaining redundant features in the same predictive feature group. To our knowledge, our paper is the first to investigate the stability of clustering based selection algorithms and compare them with other feature selection algorithms.

Two recent papers [4, 11] have studied the stability issue of feature selection under small sample size, and compared a few feature selection algorithms. According to [11], the stability of a feature selection algorithm is defined as the robustness of the results (e.g., selected subsets, feature weights) the algorithm produces to differences in training sets drawn from the same generating distribution $P(X, C)$. Both papers concluded that algorithms which performed equally well for classification had a wide difference in terms of stability, and suggested empirically choosing the best feature selection algorithm based on both accuracy and stability. Compared with these papers, our paper has three unique contributions. First, previous studies focus on the stability of feature selection algorithms on small training sets, without addressing the effect of feature redundancy on stability. Our study investigates both causes of instability of feature selection identified in Introduction. Second, in addition to individual feature ranking and feature weighting algorithms evaluated in previous studies, our paper includes feature subset search algorithms which explicitly eliminate redundant features and clustering based selection algorithms which retain redundant features in a feature group. Third, the stability measures defined in previous studies are based on overlap between two feature subsets. The stability measure proposed in our study is more general as described in the next section.

3 Stability Measures

Measuring the stability of feature selection algorithms requires some similarity measures for two sets of feature selection results. Let $R_1 = \{S_i\}_{i=1}^{|R_1|}$ and $R_2 = \{S_j\}_{j=1}^{|R_2|}$ denote two sets of feature selection results, where each S_i and S_j represents a group of features. In a special case when each S_i and S_j only contains a single feature, R_1 and R_2 become two subsets of features. In such case, the similarity between R_1 and R_2 can be simply measured by:

$$Sim_{ID}(R_1, R_2) = \frac{2|R_1 \cap R_2|}{|R_1| + |R_2|}, \tag{1}$$

where the subscript $_{ID}$ indicates that the similarity is decided by matching feature indices between the two subsets. Measures of similar forms have been used for assessing the stability of selected feature subsets in related papers discussed above [4, 11]. In our study, we develop a measure which extends existing similarity measures in two aspects. First, it is directly applicable to assess the similarity between two sets of feature groups in a general case. Second, it considers the similarity of feature values in addition to feature indices, which makes it informative when two feature subsets contain a large portion of different but highly correlated features. This general similarity measure for two sets of feature selection result is defined based on maximum weighted bipartite matching.

Given a bipartite graph $G = (V, E)$, with vertex partition $V = V_1 \cup V_2$, and edge set $E = \{(u, v)|u \in V_1, v \in V_2\}$. G is called a weighted bipartite graph if every edge (u, v) is associated with a weight $w_{(u,v)}$, and a complete bipartite graph if every u in V_1 is adjacent to every v in V_2. A matching M in G is a subset of non-adjacent edges in E. The problem of maximum weighted bipartite matching (also known as the assignment problem) is to find an optimal matching where the sum of the weights of all edges in the matching have a maximal value. There exist various algorithms, for instance, the Hungarian algorithm, for finding an optimal solution in polynomial time $O(|V|^3)$.

Given two sets of feature selection results, $R_1 = \{S_i\}_{i=1}^{|R_1|}$ and $R_2 = \{S_j\}_{j=1}^{|R_2|}$, and a user-defined similarity measure $r(S_i, S_j)$ for any pair of S_i and S_j, we model R_1 and R_2 as a weighted complete bipartite graph $G = (V, E)$, where $V = R_1 \cup R_2$, and $E = \{(S_i, S_j)|S_i \in R_1, S_j \in R_2\}$, and $w_{(S_i,S_j)}$ is determined by $r(S_i, S_j)$. The similarity between R_1 and R_2 is defined as:

$$Sim^M(R_1, R_2) = \frac{\sum_{(S_i,S_j)\in M} w_{(S_i,S_j)}}{|M|}, \tag{2}$$

where M is a maximum matching in G.

Depending on how to decide $w_{(S_i,S_j)}$, we differentiate two forms of Sim^M: Sim^M_{ID} and Sim^M_V, where the subscripts $_{ID}$ and $_V$ respectively indicate that each weight is decided based on feature indices or feature values. In the general case when S_i and S_j represent feature groups, for Sim^M_{ID}, each weight $w_{(S_i,S_j)}$ can be decided by the simple measure Sim_{ID} in (1); For Sim^M_V, each weight can

be decided according to the center or the most presentative feature of each group by Pearson correlation. In the special case when S_i and S_j represent individual features, for Sim^M_{ID}, since $w_{(S_i,S_j)} = 1$ for matching features and 0 otherwise, Sim^M_{ID} becomes Sim_{ID}; for Sim^M_V, each feature is simply represented by itself and used for Pearson correlation. Therefore, the similarity measure in (2) is a general measure for studying the stability of feature selection algorithms.

Given the general similarity measure, we define the stability of a feature selection algorithm as the average similarity of various sets of result produced by the same feature selection algorithm under training data variations. Let $Sim^M(R, R_i)$ denote the similarity between two sets of result R and R_i from the full set of samples and a subset of samples, respectively. Each subset of samples can be obtained by randomly sampling or bootstrapping the full set of samples. The stability over q subsets of samples is given by:

$$\overline{Sim}^M(R, R_i) = \frac{1}{q} \sum_{i=1}^{q} Sim^M(R, R_i) \ . \tag{3}$$

It is worth to note that the stability can also be measured based on pair-wise similarity of result from different subsets of samples. We use formula (3) because it is more efficient to compute than pair-wise comparison. Moreover, it directly captures how different the result will be from the result obtained based on the full data, when some training samples are randomly removed.

4 Representative Feature Selection Algorithms

Due to the large number of feature selection algorithms in the literature, we need to narrow down our study to a set of representative algorithms. As discussed in Section 2, various feature selection algorithms can be broadly categorized according to their evaluation measures (filters or wrappers) and their search strategies (individual feature ranking, feature subset search, or feature weighting). Since a wrapper algorithm relies on the predictive accuracy of a predefined classification algorithm to evaluate the goodness of a subset of features, the features selected from a given training set depend on the tradeoff between the bias and variance of the classification algorithm. Therefore, the stability of a wrapper algorithm under training data variation is directly caused by the variance of the classification algorithm used. There have been extensive studies on the variance and stability of classification algorithms [6, 18]. In this paper, we focus on the stability of filter algorithms which rely on data characteristics of the training set to evaluate the goodness of features. We select the following five filter algorithms representing different search strategies: F-statistic ranking, mRMR, ReliefF, SVM-RFE, and K-means clustering based selection.

4.1 Individual Feature Ranking

For individual feature ranking, various measures (e.g., Information Gain, Gini index, Symmetrical Uncertainty, t or F-statistic, etc.) can be used to evaluate

the relevance of an individual feature to the target class, and have shown similar performance in terms of both predictive accuracy [14] and stability under training data variation [11] for high-dimensional microarray data. In our study, we choose F-statistic based feature ranking as a representative, and compare it with other filter algorithms with more sophisticated search strategies.

F-statistic is a general form of t-statistic for two classes, which measures the differences of means between samples of different classes [5]. The F-statistic of a feature is computed with information about a single feature and the class, and does not take into account the possible dependency among features. Selecting a subset of top-ranked features according to F-statistic scores (or other measures mentioned above) does not yield a compact subset because highly correlated features are assigned similar scores and hence selected or excluded together. Moreover, interacting features that individually do not separate well the class but together do are missed from the selected subset.

4.2 Feature Subset Search

For structural risk minimization, a minimum subset including only relevant but non-redundant features is preferred for learning a general model from the training data [20]. A majority of feature selection algorithms apply various subset search strategies to find a subset of features which are highly correlated to the class but lowly correlated with each other (See [15] for a comprehensive survey). Among such algorithms, we select a well-known mRMR (minimum redundancy and maximum relevance) algorithm due to its efficiency in handling feature redundancy and flexibility in specifying the number of desired features by users [5].

mRMR algorithm aims to select a subset of features which simultaneously optimizes the following two criteria:

$$\max_S \frac{1}{|S|} \sum_{f_i \in S} I(f_i, C), \quad \min_S \frac{1}{|S|^2} \sum_{f_i, f_j \in S} I(f_i, f_j),$$

where $I(f_i, C)$ denotes the relevance of feature f_i to the class variable C, and $I(f_i, f_j)$ denotes the correlation between two features in the subset S. $I(f_i, C)$ can be measured by mutual information or F-statistics, and $I(f_i, f_j)$ can be measured by mutual information or Pearson correlation. In practice, mRMR combines the two optimization criteria into a single criterion function by maximizing the difference or quotient of the two above functions, and applies heuristic sequential forward search to add one feature at a time to the current best subset until a desired number of features is reached.

mRMR selects features which have the highest relevance with the target class and are also minimally correlated to each other, and hence yield more compact subsets than algorithms based on individual feature ranking. However, like individual feature ranking, the mutual information or F-statistic of a feature to the class is individually computed, and therefore, mRMR does not identify interacting features.

4.3 Feature Weighting

ReliefF is a simple and efficient feature weighting algorithm which considers all features together in evaluating the goodness of features [17]. Its key idea is to estimate the goodness of features according to how well their values distinguish between samples that are similar to each other. For this purpose, given a randomly selected sample X from a training set, ReliefF searches its K nearest neighbors from the same class called nearest hits H, and also its K nearest neighbors from each of the different classes called nearest misses M. It then adjusts the weight W_i for each feature based on the feature values of X, H, and M. The weight W_i is decreased according to the difference of feature values between X and those in H and increased according to the difference of feature values between X and those in M. The process is repeated m times, where m is a user-defined parameter with default value as the size of the training set.

Since all features are used to compute distances in deciding the nearest neighbors for the selected sample in each iteration, ReliefF considers feature interaction while weighting features. A subset of top-ranked features is normally selected based on the final feature weights and a user-defined threshold. Detailed discussions on ReliefF can be found in [17].

SVM-RFE is another type of margin-based feature weighting algorithm [8]. The main process of SVM-RFE is to recursively eliminate features based on SVM. At each iteration, it trains a linear SVM classifier, ranks features according to the squared values of feature coefficients assigned by the linear SVM, and eliminates one or more features with the lowest scores. It is worth mentioning that although SVM-RFE applies a classification algorithm in evaluating the goodness of features, it does not rely on the predictive accuracy of the classifier as wrapper methods do.

Although SVM-RFE and ReliefF are both feature weighting algorithms, they are largely different in three aspects. First, in SVM-RFE the margin is defined by support vectors, while in ReliefF, the margin is defined by a randomly selected sample and its nearest neighbors. Second, SVM-RFE recursively shrinks the subset of top-ranked features and reassigns weights to the remaining features, while ReliefF forms a subset of top-ranked features based on the final weights of all features from a single run of the algorithm. Third, SVM-RFE differentiates the weights of highly correlated features and hence can lead to a compact subset of top-ranked features with limited feature redundancy, while ReliefF assigns similar weights to highly correlated features and hence does not eliminate feature redundancy in the subset of top-ranked features. Nevertheless, ReliefF is computationally more efficient than SVM-RFE. Therefore, we select both algorithms in our study.

4.4 Feature Clustering based Selection

Algorithms introduced previously either do not identify redundant features, like F-statistic ranking and ReliefF, or eliminate redundant features from the selected feature subset, like mRMR and SVM-RFE. From the perspective of struc-

tural risk minimization, a minimum subset including only relevant but non-redundant features is preferred. Nevertheless, if the goal of knowledge discovery is to determine which features are important for the differences between classes and the relationship among them, eliminating redundant features misses some important knowledge about the features. As discussed in Section 2, some recent algorithms [9, 10] apply feature clustering to feature selection and produce feature selection result as a set of predictive feature groups, each consisting of features relevant to the class but redundant to each other. In our study, we evaluate a K-means based feature groups selection algorithm, which is similar to but computationally more efficient than the algorithm introduced in [10].

The algorithm clusters all features into K groups based on traditional K-means algorithm, and uses a representative feature, the one with highest average similarity to all other features in the same group, to represent each resulting feature group. As in [10], a group is considered relevant and selected for the subsequent classification task if its representative feature is among the top k (k <K) groups according to relevance ranking (e.g, based on F-statistic). Intuitively, when K is reasonably large, the resulting K feature groups will be coherent enough such that features in each group can be represented by a single feature in feature ranking and classification. For the sake of a simple model, the algorithm is able to provide a compact feature subset for classification by only using one representative feature from each relevant feature group. Besides, the algorithm produces coherent feature groups which provide valuable knowledge about how relevant features are correlated. We also evaluate a simpler version of this algorithm by using representative features from all K clusters for classification without relevance filtering as in [2].

5 Experiments Setup

Recall the two causes of instability of feature selection algorithms: small sample size and elimination of redundant features. High-dimensional microarray data sets, which contain many redundant features but limited samples, serve excellent testbeds for the stability of feature selection algorithms. We evaluated the representative feature selection algorithms introduced the previous section with six frequently studied public microarray data sets[1], characterized in Table 1.

In order to evaluate the stability and classification performance of each feature selection algorithm, each data set was randomly partitioned into 3 folds, with each fold containing 1/3 of all the samples. The same feature selection algorithm was repeatedly applied to 2 out the 3 folds, while each time a different fold was hold out. The selected features (or feature groups as in K-means based selection) were recorded for stability calculation. Meanwhile, a classifier was trained on the same subset of samples for feature selection and tested on the corresponding hold-out fold, based on the the selected features (or representative features as in K-means based selection). The predictive accuracy was recorded. We used

[1] http://www.cs.binghamton.edu/∼lyu/data/

Table 1. Summary of microarray data sets used in the study.

Data Set	# genes	# Samples	# Classes
Colon	2000	62	2
Leukemia	7129	72	2
Lung	12533	181	2
Prostate	6034	102	2
Lymphoma	4026	62	3
Srbct	2308	63	4

both sophisticated SVM (liner kernel) and simple KNN (K=1) classification algorithms (with Weka's implementation [19]) to test the predictive accuracy. The above process was repeated 10 times for different partitions of the data set. Overall, a total of 10×3 different subsets of samples were used to generate various sets of feature selection result by the same feature selection algorithm. The stability and classification performance of a feature selection algorithm were respectively measured by $\overline{Sim}^M(R, R_i)$ and average predictive accuracy over the 30 folds. To calculate $\overline{Sim}^M(R, R_i)$, the feature selection algorithm was also applied to the full set of data to produce R, a reference set of features (or feature groups).

As to algorithms settings, we use the mutual information quotient function to combine the two optimization criteria in mRMR, and denote this algorithm as mRMR(/). We set m to be the default value and K to be 10 for ReliefF. Since SVM-RFE is computationally intensive, as in [8], we first eliminate half of the remaining features at each iteration and then switch to one feature at a time when only a small number of features are left. For K-means based feature groups selection, without prior knowledge about the optimal number of clusters in each data set, the performance of the algorithm was evaluated under a wide range of K values. For each K value, K-means was repeated 50 times with random initial seeds, and the clustering result with minimum WSS (Within clusters Sum of Squared errors) was used for feature groups selection.

6 Results and Discussion

In this section, we evaluate and compare the stability and classification accuracy of the representative feature selection algorithms introduced in Section 4. We also discuss the merits of the proposed stability measure and the impact of feature redundancy on stability based on stability profiles of these algorithms.

6.1 Stability Results

For each algorithm, both forms of similarity measure, Sim_{ID}^M and Sim_V^M defined in Section 3, are used to assess the similarity of selected features or feature groups. For K-means based selection, to compute Sim_V^M, Pearson correlation of representative features is used as the weight in determining a maximum matching

between two sets of feature groups. To compute Sim_{ID}^{M}, a maximum matching is determined in two ways: using all features in each cluster or 5 features closest to each cluster center (up to 5 if there are less than 5 features in the cluster). The reported results are based on the former which generally shows better matching. For other feature selection algorithms, it's straightforward to compute Sim_{ID}^{M} and Sim_{V}^{M} based on feature indices or values, respectively.

Fig. 1 reports the stability profiles (stability scores across different numbers of selected features or feature groups) of various representative algorithms (Random, F-statistic ranking, ReliefF, mRMR(/), SVM-RFE, and K-means based selection) based on $\overline{Sim}_{ID}^{M}$ (left column) and $\overline{Sim}_{V}^{M}$ (right column) for four microarray data sets (Colon, Leukemia, Lymphoma, and Srbct). Due to space limit, the results of two other data sets (Lung and Prostate) which show similar trends as the ones presented are excluded from the paper. For each data set, the stability profile of Random selection which randomly selects k features from the entire set of features is included as a baseline for comparing other algorithms.

Stability Profiles based on $\overline{Sim}_{ID}^{M}$ From the left column of Fig. 1, we can clearly observe from every data set that the stability profiles of the six algorithms fall into two distinct groups. F-statistic ranking, ReliefF, and mRMR(/) are significantly more stable than SVM-RFE, K-means, and Random in terms of $\overline{Sim}_{ID}^{M}$. Our observation of F-statistic ranking and ReliefF being more stable than SVM-RFE under variations of small training sets is consistent with the conclusions about the stability of these algorithms in previous study [11]. In the following, we verify that different ways of handling feature redundancy also affect the stability of feature selection algorithms by taking a close look at the stability profiles of all six algorithms on Colon data. The observations from Colon data are consistent with those from other data sets used in the study.

The stability scores of ReliefF are generally around 0.6, indicating that on average about 60% of the top k features selected based on the full set of samples will appear in the top k features selected based on a random subset (2/3) of the full set. The stability profile of F-statistic ranking is similar to ReliefF. Since both algorithms select top k features only based on feature relevance without considering feature redundancy, the variations of the top k features selected by each algorithm are mainly controlled by the variations in the training set. Different subsets of samples may result in different relevance scores of the same feature and hence different feature rankings.

mRMR(/) is very similar to F-statistic ranking when $k \geq 20$, but much less stable for smaller k. Since F-statistic ranking and mRMR(/) apply the same F-statistic measure to decide the relevance of individual features and are evaluated based on the same set of random data partitions, variations in the training set will have the same effect on the variations of F-statistic scores. However, mRMR(/) applies the mutual information quotient criterion to penalize features highly correlated with those in the current subset during the sequential forward subset search. The penalty is more aggressive when the current subset is small. Among a set of highly correlated features, different ones may be selected by

mRMR(/) based on different subsets of samples. Consequently, the stability of mRMR(/) significantly reduces compared with F-statistic ranking when a small number of features are selected (e.g, $k \leq 10$). Such observation verifies the effect of eliminating redundant features on the stability of feature selection algorithms.

In contrast, SVM-RFE and K-means are very unstable in terms of $\overline{Sim}^M_{ID}$, and similar to Random selection which shows almost no overlap between the subset of features selected from the full data and the subset of features selected from any random subset of data. While both mRMR(/) and SVM-RFE eliminate redundant features, SVM-RFE is much less stable under the same variations of the training data. This can be explained by the fact that SVM-RFE applies multiple iterations of eliminating features and reassigning weights to remaining features. Since different subsets of samples may result in different feature rankings and elimination of different features at a given iteration, multiple iterations will cause higher instability on the top k selected features.

K-means based feature groups selection handles feature redundancy in a different way. Instead of eliminating redundant features, highly correlated features are grouped together by K-means clustering, and the top k relevant feature clusters are determined by F-statistic scores of the representative features of K feature clusters (K=50 for Fig. 1). Its stability profile indicates that almost no overlap between any pair of matching clusters (considering either all features in each cluster or several closest features to each cluster center). Such instability could result from the instability of both feature clustering and F-statistic ranking of representative features.

Summary: (1) Measure $\overline{Sim}^M_{ID}$ enables us to evaluate the stability of K-means based selection algorithm and compare it with other feature selection algorithms. (2) The limited number of samples in each data set (less than 100) clearly affects the stability of feature selection algorithms under training data variations. According to $\overline{Sim}^M_{ID}$, the best performing algorithms, ReliefF and F-statistic ranking, show stability scores about 0.6 when 1/3 of the samples are excluded from each training set. (3) Due to the abundance of redundant features in each data set, algorithms which deal with feature redundancy differently, mRMR(/), SVM-RFE, and K-means based selection, show degraded stability to different extents, when measured by $\overline{Sim}^M_{ID}$ which only considers the overlap between two sets of features.

Stability Profiles based on $\overline{Sim}^M_V$ We now examine the stability profiles of all algorithms based on $\overline{Sim}^M_V$ (the right column of Fig. 1). The first clear observation is that for every data set, the stability profile of every algorithm based on $\overline{Sim}^M_V$ consistently shows higher stability than the corresponding profile based on $\overline{Sim}^M_{ID}$. This conforms to our similarity definition (2) in Section 3. For any two sets of feature selection result R_1 and R_2, $Sim^M_V(R_1, R_2) \geq Sim^M_{ID}(R_1, R_2)$ if every pair of features matched by feature values has non-negative correlation value. The increase of stability is at different scales for different algorithms. In

the following, we examine the merit of the proposed stability measure $\overline{Sim}_V^M$ by taking a close look at the stability profiles of the algorithms on Lymphoma data.

The stability profile of Random selection based on $\overline{Sim}_V^M$ is in general about 0.4 on Lymphoma data. Such value exhibits a significant increase from the almost zero value based on $\overline{Sim}_{ID}^M$, and defines the baseline for judging the significance of Pearson correlation of matching features by other algorithms. Note that as k increases, the stability score of Random selection steadily increases. This can be explained by the fact that when k increases, each randomly picked feature will have a better chance to be matched with a correlated feature.

According to $\overline{Sim}_{ID}^M$, ReliefF is the most stable algorithm with stability scores about 0.6 across various numbers of selected features, which indicates that for a given k, on average about 60% of the features selected based on the full set of samples are selected again based on each random subset. An interesting question is whether the 40% of non-matching features (according to feature indices) are highly correlated with some features selected based on each random subset. The measure $\overline{Sim}_V^M$ enables us to answer this question. When measured by $\overline{Sim}_V^M$, the stability scores of ReliefF increase to about 0.9, which indicates that for a given k, on average each of the no-matching features according to feature indices has Pearson correlation about 0.8 to its matching feature according to feature values (i.e., given Pearson correlation value being 1 for matching features, $0.6k \times 1 + 0.4k \times 0.8)/k \approx 0.9$). Such Pearson correlation is much higher than the baseline value of 0.4.

For all the other four algorithms, we can also observe that non-matching features according to $\overline{Sim}_{ID}^M$ on average show high Pearson correlations when measured by $\overline{Sim}_V^M$. Such trend is particularly interesting for mRMR(/), SVM-RFE, and K-means based selection, which show low stability profiles due to the treatment of redundant features. For example, according to $\overline{Sim}_{ID}^M$, SVM-RFE is very instable with stability scores less than 0.1 across various numbers of selected features, however, it shows much better stability with stability scores about 0.6 (clearly better than the baseline 0.4 from Random selection) when measured by $\overline{Sim}_V^M$.

The observations on Lymphoma data verify that by taking into account the similarity of feature values, $\overline{Sim}_V^M$ is more informative than traditional similarity measures based on subset overlap. Such merit of $\overline{Sim}_V^M$ can also be appreciated on Leukemia and Srbct data, although the margin from Random selection to SVM-RFE and K-means is less pronounced. The margin becomes hard to tell on Colon data where the stability profile of Random selection based on $\overline{Sim}_V^M$ is in general about 0.6.

Summary: (1) Measure $\overline{Sim}_V^M$ enables us to evaluate the stability of K-means based selection algorithm and other feature selection algorithms from a different aspect than $\overline{Sim}_{ID}^M$ by taking into account the similarity of feature values. (2) Algorithms which deal with feature redundancy show significantly improved stability when measured by $\overline{Sim}_V^M$.

6.2 Accuracy Results

Table 2. Average classification accuracies (%, with standard deviation ±) of SVM classification algorithm based on features selected by Random selection, F-statistic ranking, ReliefF, mRMR, SVM-RFE, and K-means based selection.

Method	k							
	4	6	8	10	20	30	40	50
Random	64.7±1.6	63.9±1.4	66.6±3.8	65±3.4	68±3.8	74±4.2	74.5±7.2	77.1± 3.3
F-statistic	80±3.3	80.3±3.2	82.1±2.1	82.9±2	84.4±2.7	84.5±3.1	84.9±2.2	84.9±2.3
ReliefF	79.3±2.5	80.6±1.7	81.8±1.5	82.9±2	83.7±2.7	84.1±3.7	83.3±2.5	82.3±2.5
mRMR(/)	78.5±4.4	82.4±2.8	83.4±2	84.4±1.7	85.2±2.5	84.5±3.2	84.5±2.2	83.2±2.3
SVM-RFE	64.8±3.3	65.5±4.1	68.9±5.1	68.3±5.6	76.1±5.3	78.6±5.1	80.2±6.7	81.9±7.6
Kmeans	77.6±10.7	79.3±8.7	81.7±7.7	82.7±7.0	84.2±6.3	84.2±5.8	81.9±8.2	82.2±7.8
Colon								
Random	66.1±1.2	69.2±3.3	69.7±2.9	69.9±3.3	78.3±4.5	77.8±6.3	77.6±6.1	81.4±6.9
F-statistic	90.7±2.1	91.4±1.4	92.9±1.2	92.8±1.4	93.9±1.6	95±1.5	96±1	96.4±0.7
ReliefF	91.2±0.9	92.5±1	93.1±1.6	93.6±2.1	95±1.3	96±1	96.1±1.6	96.1±1.7
mRMR(/)	91.4±1.6	92.4±1.3	92.9±1.4	94±1.1	95.6±0.6	96±1.2	96.2±0.9	96.7±1.3
SVM-RFE	78.2±5	85.4±3.7	87.2±5.5	89.9±4.9	95.3±2.9	95.4±2.7	96.8±0.9	97.5±0.9
Kmeans	89.4±5.9	91.8±4.3	92.9±4.1	93.6±4.3	95.7±4	95.4±3.5	95.6±3.6	97.4±3.4
Leukemia								
Random	74.4±4.5	76.1±5.5	82.9±5.3	86.8±4.1	94±2.4	95.3±3.3	96.3±2.3	96.8±1.8
F-statistic	87.3±3.7	90.2±2.9	90.9±2.6	93±3.6	96.2±1.7	97.3±1.5	97.1±2.2	97.4±2.3
ReliefF	82.9±1.6	84.7±1.9	85.5±1.4	86.8±3	95±2.1	97.7±1.6	97.6±1.4	97.9±1.7
mRMR(/)	95.5±1.7	96.9±1.8	97.4±2.2	97.6±2	97.9±1.5	98.7±1	99.2±0.9	99.4±0.8
SVM-RFE	87.5±3.7	95.8±4	97.3±3.1	98±3.1	99±2.4	99.3±1.7	99.5±1.5	99.5±1.5
Kmeans	87.9±9.6	91.9±9.4	94.4±7.6	94.4±7.7	97.3±3.7	97.6±3.7	98.2±2.9	98.4±2.9
Lymphoma								
Random	42.5±6.5	44.3±6	53.2±5.4	55.2±6.3	68.9±6.9	77.9±5	77.6±6.4	83.2±3.8
F-statistic	76.7±6	90.6±4.1	93.5±4.4	95.1±2.3	98.6±1.4	98.6±1.2	98.6±1.2	98.3±1.2
ReliefF	80.3±2.1	86.7±4.4	93.3±2.7	96.7±2.5	97.8±1.9	97.8±2	97.6±1.3	98.3±0.9
mRMR(/)	71.7±5.1	88.6±4.8	95.4±3.9	97.5±2.6	98.9±1.8	98.4±1.7	97.8±1.5	98.3±0.9
SVM-RFE	56.8±10.5	70.8±9.9	81.3±8.8	89.5±6.5	95.1±3.6	97±3.4	97.3±3.7	97.9±3.2
Kmeans	61.4±15.3	72.7±16.2	79±15.6	85.4±11.3	90.8±6.4	93±5.6	94±6	93.2±6.8
Srbct								

We now compare the classification performance of the six algorithms studied in the previous section. Table 2 reports the average predictive accuracies (over 30 folds) for SVM classification based on selection results of these algorithms under a wide range of k, where k stands for the number of selected relevant feature groups for K-means based selection, and the number of features for other feature selection algorithms, respectively. As mentioned before, we also evaluated these algorithms based on 1NN classification. Since the table for predictive accuracies for 1NN shows similar trends as Table 2, it is not included due to space limit.

For the ease of discussion, we exclude Random selection when referring to all algorithms from now on. It is clear that for every data set, the accuracies

of SVM resulted from all algorithms are significantly better than those from Random selection, which verifies that all algorithms are effective (to some extent) in selecting predictive features. However, there is no clear winner among these algorithms in terms of best predictive accuracy. If one needs to empirically choose a feature selection algorithms among a few, as suggested in [4, 11], predictive accuracy and stability profiles can be considered together to guide the selection.

For Colon and Leukemia data, all algorithms in general perform equally well according to the means and standard deviations of the average accuracies, with the exception of SVM-RFE which shows significantly lower accuracies than other algorithms under small k values. According to the stability profiles in Fig. 1, both SVM-RFE and K-means based selection should be excluded from consideration due to its poor stability in terms of both $\overline{Sim}_{ID}^{M}$ and $\overline{Sim}_{V}^{M}$. Srbct data shows similar trend as Colon and Leukemia data, except that K-mean based selection can also be excluded due to its inferior classification accuracy than other algorithms.

For Lymphoma data, we can observe a different trend. All algorithms in general perform equally well, however, mRMR(/) and SVM-RFE show significantly higher accuracies than F-statistic ranking and ReliefF for small k values. Such observation is consistent with the fact that both mRMR(/) and SVM-RFE eliminate redundant features. The subsets of features selected by mRMR(/) and SVM-RFE contain more relevant information to the class than subsets of the same size selected by F-statistic ranking and ReliefF which do not consider feature redundancy. If we further examine the stability profiles, mRMR(/) is clearly a better choice than SVM-RFE. It is worthy of mentioning that measure $\overline{Sim}_{V}^{M}$ allows us to observe the high stability of mRMR(/) when a small number of features are selected (about 0.8 at $k \leq 10$).

7 Conclusion

In this paper, we have systematically investigated the impact of small sample size and feature redundancy on the stability of feature selection algorithms. We have proposed and shown the merits of a general stability measure. We have empirically evaluated several representative feature selection and feature groups selection algorithms, and compared their stability profiles and classification performance based on microarray data sets.

According to experimental results, algorithms which only consider feature relevance, such as F-statistic ranking and ReliefF, in general show both good classification performance and stability under training data variations. Among algorithms which dealt with feature redundancy, mRMR(/) is in general a better choice than SVM-RFE, and K-means based selection when considering both accuracy and stability. A future research direction is to develop solutions for improving the stability of feature selection algorithms, in particular, feature groups selection algorithms, without sacrificing accuracy. The general measure proposed in this paper can also be used to evaluate newly developed algorithms.

References

1. A. Appice, M. Ceci, S. Rawles, and P. Flach. Redundant feature elimination for multi-class problems. In *Proceedings of the 21st international conference on Machine learning*, 2004.
2. W. Au, K. C. C. Chan, A. K. C. Wong, and Y. Wang. Attribute clustering for grouping, selection, and classification of gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2(2):83–101, 2005.
3. R. Butterworth, G. Piatetsky-Shapiro, and D. A. Simovici. On feature selection through clustering. In *Proceedings of the Fifth IEEE International Conference on Data Mining*, pages 581 – 584, 2005.
4. C. A. Davis, F. Gerick, V. Hintermair, C. C. Friedel, K. Fundel, R. Kffner, and R. Zimmer. Reliable gene signatures for microarray classification: assessment of stability and performance. *Bioinformatics*, 22:2356–2363, 2006.
5. C. Ding and H. Peng. Minimum redundancy feature selection from microarray gene expression data. In *Proceedings of the Computational Systems Bioinformatics conference (CSB'03)*, pages 523–529, 2003.
6. P. Domingos. A unified bias-variance decomposition and its applications. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 231–238, 2000.
7. G. Forman. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289–1305, 2003.
8. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.
9. T. Hastie, R. Tibshirani, D. Botstein, and P. Brown. Supervised harvesting of expression trees. *Genome Biology*, 2:0003.1–0003.12, 2001.
10. R. Jornsten and B. Yu. Simultaneous gene clustering and subset selection for sample classification via MDL. *Bioinformatics*, 19:1100–1109, 2003.
11. A. Kalousis, J. Prados, and M. Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and Information Systems*, 12:95–116, 2007.
12. R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.
13. D. Koller and M. Sahami. Toward optimal feature selection. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pages 284–292, 1996.
14. T. Li, C. Zhang, and M. Ogihara. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, 20:2429–2437, 2004.
15. H. Liu and L. Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Trans. on Knowledge and Data Engineering*, 17(3):1–12, 2005.
16. M. S. Pepe, R. Etzioni, Z. Feng, and et al. Phases of biomarker development for early detection of cancer. *J Natl Cancer Inst*, 93:1054–1060, 2001.
17. M. Robnik-Sikonja and I. Kononenko. Theoretical and empirical analysis of Relief and ReliefF. *Machine Learning*, 53:23–69, 2003.
18. P. Turney. Technical note: bias and the quantification of stability. *Machine Learning*, 20:23–33, 1995.
19. I. H. Witten and E. Frank. *Data Mining - Pracitcal Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, 2005.
20. L. Yu and H. Liu. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5:1205–1224, 2004.

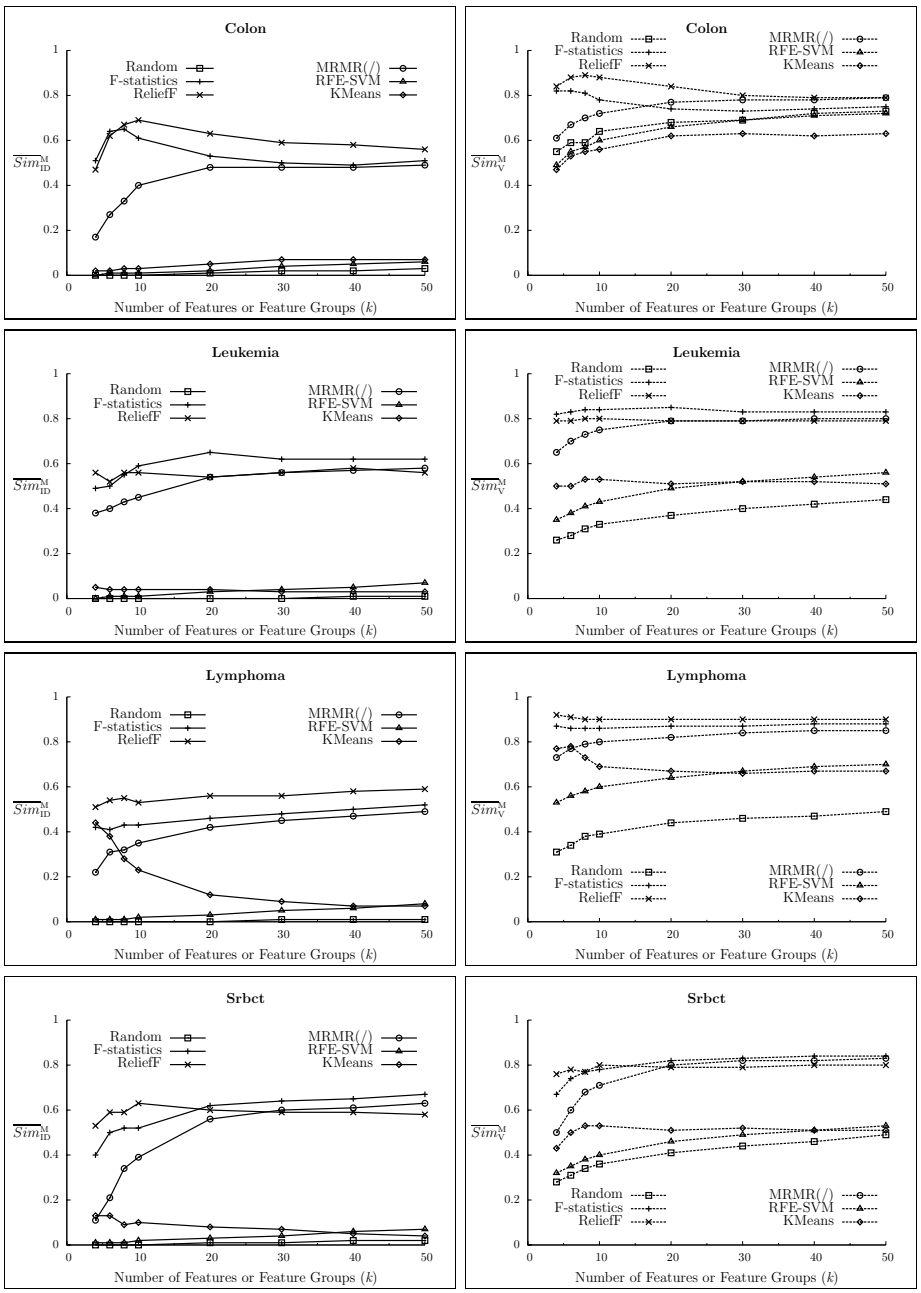

Fig. 1. Stability profiles based on $\overline{Sim}^{M}_{ID}$ (left column) and $\overline{Sim}^{M}_{V}$ (right column) for six representative algorithms on four microarray data sets.