Stability of Feature Selection: Sample Size Effect on Conventional, Ensemble, and Group-based Algorithms

Keywords: feature selection, stability, high-dimensional data, small samples, classification

Abstract

ICML 2009 full paper submissions are due January 26, 2009. Reviewing will be blind to the identities of the authors, and therefore identifying information should not appear in any way in papers submitted for review. Submissions must be in PDF or Postscript, 8 page length limit.

1. Introduction

Many feature selection algorithms have been developed with a focus on improving classification accuracy while reducing dimensionality. A relatively neglected issue is the *stability of feature selection* - the insensitivity of the result of a feature selection algorithm to variations in the training set. Suppose we perform feature selection on a data set D with n samples and p features (dimensions). If we randomly split the data into two sets D_1 and D_2 with half samples each, and run a feature selection algorithm on them, the results R_1 and R_2 usually do not agree with each other.

Stability of feature selection is the main topic of this paper. This issue is important in many applications with high-dimensional data, where feature selection is used as a knowledge discovery tool for identifying characteristic markers for the observed phenomena (Pepe et al., 2001). For example, in microarray data analysis, a feature selection algorithm may select largely different subsets of features (genes) under variations to the training data (Yu et al., 2008; Kalousis et al., 2007). Such instability dampens the confidence of domain experts in investigating any of the various subsets of selected features for biomarker identification.

There are a number of factors which affect the stability of feature selection results: (1) sample sizes, (2) feature selection algorithms, and (3) stability measures.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

To our knowledge, sample size effect has not been investigated previously. It is the focus of this paper. We next review existing studies on feature selection algorithms w.r.t. stability. Stability measures will be discussed in Section 4.

There exit very limited studies on the stability of feature selection algorithms. One early work in this direction was done by (Kalousis et al., 2007). Their work compared the stability of a number of conventional feature ranking and weighting algorithms under training data variation based on three stability measures on high-dimensional data. More recently, two techniques were proposed to explicitly achieve stable feature selection without sacrificing classification accuracy: ensemble feature selection (Saeys et al., 2008) and group-based feature selection (Yu et al., 2008). An extensive comparison of both techniques with conventional feature selection algorithms on equal footing would be interesting.

The above studies have not addressed an important issue: how sample sizes affect the stability of feature selection algorithms. In particular, whether increasing the sample size (e.g., doubling or tripling the number of training samples) would be an effective way of improving the stability of feature selection algorithms, and how different algorithms compare under increasing sample size. We note that in many applications, increasing the number of training samples could be very costly. In gene expression microarray data, each sample is from the tissue of a cancer patient, which is usually hard to obtain and costly to perform experiments. Our study attempts to address this issue.

In this paper, we conduct a large-scale and comprehensive study of 15 feature selection algorithms, including 5 representative conventional feature selection algorithms (i.e., F-statistic, ReliefF, SVM-RFE, SVM, and mRMR), and their extensions by ensemble and group-based techniques, respectively. We evaluate the stability of these algorithms based on 3 different stability measures introduced later. To address the effect of sample size on stability, we develop various synthetic

high-dimensional data sets, and evaluate the performance of all 15 algorithms on varying training sample sizes from 100 to 1000. Since stability of results should not be considered per se, we also evaluate the correctness of the selected subsets by these algorithms based on precision and recall of true relevant features (on synthetic data) and classification accuracy (on both synthetic and real-world data).

The empirical results are surprising. To preview:...

2. Feature Selection Algorithms

Conventionally, the result of feature selection is determined by a feature selection algorithm on a single training set, and the algorithm performs relevance and redundancy analysis treating each feature as an individual entity. Very recently, two different directions have been proposed in order to increase the stability of feature selection results. One is ensemble feature selection, and the other is group-based feature selection. We first identify representatives from conventional feature selection algorithms, and introduce the new ideas and show how they can be applied to couple with the new ideas.

2.1. Conventional Algorithms

Univariate ranking: Various measures (e.g., Information Gain, Symmetrical Uncertainty, F-statistic, or Pearson correlation) can be used to evaluate and rank the relevance of individual features, and they have shown similar performance in terms of generalization accuracy (Li et al., 2004) and stability under training data variations for high-dimensional data (Kalousis et al., 2007). In this study, we choose F-statistic based feature ranking as a representative algorithm.

ReliefF: ReliefF is a simple and efficient feature weighting algorithm which considers all features together in evaluating the relevance of features (Robnik-Sikonja & Kononenko, 2003). Its key idea is to estimate the relevance of features according to how well their values distinguish between samples that are similar to each other. Like univariate ranking, ReliefF assigns similar relevance scores to highly correlated features, and hence do not minimize redundancy in the top-ranked features.

SVM-RFE: SVM-RFE is a popular algorithm in this category (Guyon et al., 2002). The main process of SVM-RFE is to recursively eliminate features based on SVM, using the coefficients of the optimal decision boundary to measure the relevance of each feature. 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. SVM-RFE differentiates the weights of highly correlated features and hence can minimize redundancy in the top-ranked features.

SVM: We also evaluated SVM ranking without the RFE procedure, that is the ranking at the first iteration of SVM-RFE.

mRMR: Algorithms above produce a ranking which can be used to pick subsets. Many algorithms apply subset search strategies to select a subset of features which are highly correlated to the class but lowly correlated with each other. We choose a well-known algorithm, mRMR (minimum redundancy and maximum relevance) due to its computational efficiency and flexibility in specifying the number of selected features (Ding & Peng, 2003). mRMR aims to select a subset of features which simultaneously maximize V_F and minimize W_r. In practice, mRMR combines the two criteria into a single criterion function by maximizing the difference or quotient of the two, 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.

2.2. Ensemble Algorithms

It is well known that ensemble methods () for classification which aggregate the predictions of multiple classifiers can achieve better generalization than a single classifier. Similarly, for feature selection, we can apply a feature selection algorithm on various bootstrapped training sets to obtain an ensemble of feature sets. In (Saeys et al., 2008), they apply a feature selection algorithm on various bootstrapped training sets to obtain an ensemble of feature sets. How to aggregate results for ranking or subset selection algorithms. Depending on the results format,... We did ensemble version for each of the algorithm introduced above.

2.3. Group-based Algorithms

The idea of stable feature selection via dense feature groups was first proposed and tested in (Yu et al., 2008). The idea was motivated by two main observations. in the sample space (i.e, the feature space defined by a set of samples in a training set), the dense core regions (peak regions), measured by probabilistic density estimation, are stable with respect to sampling of the dimensions (samples). For example, a spherical Gaussian distribution in the 100-dimensional space will likely be a stable spherical Gaussian in any of the subspaces. The features near the core of the spherical

Gaussian, viewed as a core group are likely to be stable under sampling, although exactly which feature is closest to the peak could vary. Another observation is that the features near the core region are highly correlated to each other, and thus should have similar relevance scores w.r.t. some class labels, assuming the class labels are locally consistent. Thus these features can be regarded as a single group in feature ranking. And we can pick any one of them in final classification. In this sense, the feature group is a stable entity.

The original DRAGS algorithm first finds dense feature groups based on ... It then uses the average relevance (F-statistic value) of features in the same group to rank each dense feature groups and selects top ones. It uses a representative feature (the one with the highest average similarity to all other features in the same group, that is, the one closes to the mean of all features) from each selected group for classification.

As ensemble feature selection, the idea of group-based feature selection can be applied to many existing feature selection algorithms. Therefore, we extend DRAGS by using different conventional feature selection algorithms introduced before in the second stage of the algorithm where features groups are selected. We have group-based F-statistics, ReliefF, SVM-REF, SVM-One, and mRMR algorithms. For each group-based algorithm, we apply DGF to get the groups first, then (1) take the representative feature (the one closest to the median) and use it to represent the group and trim the data or (2) take the median as a virtual feature to represent the group and convert and trim the data. After that we apply each conventional algorithm on the trimmed data to produce final feature selection results.

3. Evaluation Methodology

3.1. Performance Measures

Evaluating the stability of feature selection algorithms requires some similarity measures for two sets of feature selection results. Let R_1 and R_2 denote two sets of results by a feature selection algorithm from two different training sets, where R_1 and R_2 can be two vectors of feature weights, or two vectors of feature ranks, or two feature subsets, depending on the output of the algorithm. For feature weighting, the Pearson correlation coefficient can be used to measure the similarity between R_1 and R_2. For feature ranking, we use the Spearman rank correlation coefficient as in (Kalousis et al., 2007) and (Saeys et al., 2008):

$$Sim_R(R_1, R_2) = 1 - 6\sum_{i=1}^{d} \frac{(R_1^i - R_2^i)^2}{d(d^2-1)}, \qquad (1)$$

where d is the total number of features, and R_1^i and R_2^i are the ranks of the ith feature in the two rank vectors, respectively. Sim_R takes values in [-1,1]; a value of 1 means that the two rankings are identical and a value of -1 means that they have exactly inverse orders.

For feature subset selection, the Jaccard index was used in both Kalousis and Saeys's papers. A variation of Jaccard index was used in (Yu et al., 2008):

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

Like Jaccard index, Sim_{ID} takes values in [0,1], with 0 meaning that there is no overlap between the two subsets, and 1 that the two subsets are identical. When $|R_1| = |R_2|$ (i.e., evaluating an algorithm under the same subset cardinality), the value of Sim_{ID} directly reflects the percentage of overlapping features. Therefore, we adopt Eq. (2) in our study.

Sim_{ID} does not take into account the similarity of feature values; two subsets of different features will be considered dissimilar no matter whether the features in one subset are highly correlated with those in the other subset. However, from classification's point of view, among a group of highly correlated features, which one is selected based on a training set will have similar effect on the resulting model. To capture the similarity in this sense, we adopt the similarity measure proposed by Yu (2008) which considers similarity of feature values. The measure is defined based on maximum weighted bipartite matching:

$$Sim_V(R_1, R_2) = \frac{\sum_{(X_i, X_j) \in M} w_{(X_i, X_j)}}{|M|}, \qquad (3)$$

where M is a maximum matching in the bipartite graph representing R_1 and R_2. Each weight $w_{(X_i, X_j)}$ is decided by the correlation coefficient between the two features X_i and X_j, where X_i...

Given each similarity measure above, the stability of a feature selection algorithm is then measured as the average of the pair-wise similarity of various feature selection results produced by the same algorithm from different training sets.

3.2. Data Sets

?? We perform our study on both synthetic data sets and real-world microarray data sets. The merit of syn-

thetic data sets is two-fold: first, it enables us to examine the relationship of stability and the correctness of feature selection based on prior knowledge about feature relevance and redundancy; second, it allows us to examine the effect of sample size on stability with data sets of increasing sample size but other properties being equal.

We generated a total of 16 data sets with different dimensionality d and sample size s. A summary of these data sets is provided in Table 1, which All represents the total number of features, True represents the number of features used to decide the binary class labels, and Rel. represents the total number of relevant features, including the strongly relevant features and features that are highly correlated to one of the true features. Given the same d, the same set of features are used to generate four data sets with increasing number of instances. For comparable results across different data sets, we use the following same procedure to generate data sets. For each data set, the average correlation among each correlated group is within the range of [] across all groups. The minimum correlation within each group is 0.5. The same linear function is used to decide the class label with coefficients of each true feature between []. The only difference is sample size and the number of irrelevant groups and group sizes (from D1 to D4 the number of correlated features in each group and the irrelevant groups grow), which can be reflected in the number of features in various categories.

Table 1. Summary of synthetic data sets. Each $D_{d,m}(d \in \{10, 100, 1K, 5K\}, m \in \{100, 200, 500, 1K\})$ represents a synthetic data set with d features and m instances.

Data Set	Features			Classes
	All	True	Rel.	
$D_{10,m}$	10	10	10	2
$D_{100,m}$	100	10	50	2
$D_{1K,m}$	1000	10	100	2
$D_{5k,m}$	5000	10	200	2

We experimented with six frequently studied public microarray data sets[1], characterized in Table **??**.

3.3. Experimental Procedures

To empirically evaluate the stability of a feature selection algorithm for a given data set, we can simulate different training sets drawn from the same distribution by a resampling procedure like bootstrapping or

[1]http://www.cs.binghamton.edu/∼lyu/ICML09/data/

Table 2. Summary of microarray data sets.

Data Set	Genes	Instances	Classes
Colon	2000	62	2
Leukemia	7129	72	2
Prostate	6034	102	2
Lung	12533	181	2
Lymphoma	4026	62	3
SRBCT	2308	63	4

N-fold cross-validation. We opted for N-fold (N=3) cross-validation, as previous papers did. For each microarray data set, each feature selection algorithm was repeatedly applied to 2 out of the 3 folds, while a different fold was hold out each time. Different stability measures were calculated. In addition, a classifier was trained based on the selected features from the same training set and tested on the corresponding hold-out fold. The CV accuracies of both linear SVM and KNN classification algorithms were calculated. The above process was repeated 10 times for different partitions of the data set, and the results were averaged.

For each synthetic data set, we followed the same 10 times 3-fold CV procedure above except that an independent test set of 500 instances randomly generated from the same distribution was used in replacement of the hold-out fold in each iteration. In addition to stability and accuracies measures, we also measured the recall and precision w.r.t. true features and relevant features during each iteration of the 10×3 CV.

Due to the large scale of the experiments (about 20 data sets, 15 algorithms, ensemble algorithms.) it is a very costly (explain how costly), we did not try different N values, instead, we used 3-fold cross-validation as in (Yu et al., 2008) to make the experiments more manageable and allow sufficient training sample size.

As to algorithms settings, for ReliefF, we set m to be the default value and K to be 10. For SVM-RFE, since it is computationally intensive, as in previous works, we eliminate 10 percent of the remaining features at each iteration. For mRMR, we use the quotient function to combine the two optimization criteria in mRMR. For SVM, we use linear kernel with default C parameter. For ensemble algorithms, we use 40 bootstrapped samples for each feature selection algorithm.

4. Results

4.1. Results from Synthetic Data

Table 3 reports a comprehensive set of results for the three categories (conventional, ensemble, and group-based) and 15 feature selection algorithms w.r.t. various evaluation measures (stability, precision, and accuracy) on the 16 synthetic data sets. As a baseline, the accuracy of the full set of features is also recorded for each data set. Since KNN consistently results in lower accuracy than SVM on all data sets, KNN results are not included due to space limit. Meanings of columns are described in table caption.

At a first look, it is clear that the accuracies of SVM on the full feature set drop from a near-perfect 97% at $D_{10,1K}$ to 60% at $D_{5K,100}$ while the dimensionality increases and sample size decreases. This shows that SVM experiences trouble in finding the true linear function used to define the binary class boundary (hence, the optimal weights of the features) when the number of redundant features and irrelevant features increases. The same problem occurs when the sample size decreases.

We next examine how different feature selection algorithms compare w.r.t to (1) various stability measures (S_{ID}, S_V, and S_R) and (2) effectiveness in identifying the truly relevant features to improve SVM accuracy. More importantly, we examine how their behaviors change as sample size increases. In each block of Table 3, the value of the best algorithm for each measure is **boldfaced**. Other algorithms, whose performance are not statistically distinguishable from the best algorithm at $p = 0.01$ using two-sample paired t-tests on the 10 random trials, are marked by *. Entries in each column that are neither bold nor starred indicate performance that is significantly lower than the best algorithm at $p = 0.01$.

Let us take a close look at the block for data set $D_{1K,100}$. Among the five conventional feature selection algorithms, F-statistic and ReliefF are more stable than mRMR, SVMone, and SVMRFE w.r.t. S_{ID} and S_R, which is not surprising since the former two algorithms do not eliminate redundant features. S_V shows slightly different trend as it considers feature value correlation in measuring subsets similarity. The precision values w.r.t. truly relevant features (P) are very low for all algorithms, which is consistent with the very low SVM accuracy based on selected features for this data set.

Surprising trends can be observed when we compare these conventional algorithms under increasing sample size or when we compare them with their ensemble and group-based versions. Figure 1 visualizes such trends from $D_{1K,n}$ data sets. To enhance clarity, for each of the two picked measures (S_{ID} and SVM accuracy), three separate views ((a), (b), and (c)) of the 15 feature selection algorithms are shown in Figure 1. As shown in view (a) which compares the 5 conventional algorithms, the stability S_{ID} values for F-statistics and ReliefF in general do not improve, however, S_{ID} values for the other three algorithms steadily improve as sample sizes increase from 100 to $1K$. Surprisingly, the stability of SVMRFE and SVMone win over F-statistics and ReliefF due to their dramatic improvement. The accuracy of these algorithms in general improves as the stability improves.

Views (b) and (c) in Figure 1 compare 3 versions of F-statistics and ReliefF, and 3 versions of mRMR, SVMone, and SVMFRE, respectively. Ensemble F-statistics and ReliefF do not show clear improvement of stability over their conventional versions, however, ensemble versions of the other three algorithms do. This indicates that bagging ensemble approach may only benefit algorithms which are individually correct but instable. In contrast to ensemble, group-based versions of all five algorithms show significant improvement of stability over their conventional versions, and the improvement is much more significant than that of ensemble versions. A closer look at the figures reveal another surprising trend - the stability of the group-based version for all algorithms (except ReliefF) at sample size 200 are either significantly better than (as in F-statistic and mRMR) or comparable to (as in SVMone and SVMRFE) their conventional and ensemble versions at sample size 1000. Such observation indicates that group-based approach is potentially more effective in improving the stability of feature selection algorithms than increasing the sample size. The latter could also be very costly in certain domains like microarray analysis.

The trends observed from Figure 1 are generally applicable to other data sets.

4.2. Results from Real-World Data

Main observations and conclusions from Colon data: (1) on traditional algorithms same generalization accuracy, but very different stability, which confirms the observations of previous papers; higher stability do not contribute to higher accuracies in traditional algorithms. Conclusion: choosing algorithms is simple. Just prefer algorithms with better stability (2) ensemble approaches works more effectively for improving stability for some less stable algorithms than those that are more stable by themselves; (3) ensemble ap-

Table 3. Comparison of various feature selection algorithms based on **synthetic** data sets. Each block reports the stability S_{ID}, S_V, and S_R, precision w.r.t. truly relevant features (P) and all relevant features (P′), and SVM classification accuracy for 15 different feature selection algorithms as well as the SVM classification accuracy for the full set (without feature selection) on a data set $D_{p,n}$ (with dimensionality p and sample size n). Note: (1) feature selection algorithms are not compared on $D_{10,n}$ since its full set is exactly the optimal set of 10 features; (2) S_R is not directly applicable to mRMR and Group-based feature selection algorithms.

Data		$D_{p,100}$						$D_{p,200}$						$D_{p,500}$						$D_{p,1K}$					
	Algorithm	S_{ID}	S_V	S_R	P	P′	Ac.	S_{ID}	S_V	S_R	P	P′	Ac.	S_{ID}	S_V	S_R	P	P′	Ac.	S_{ID}	S_V	S_R	P	P′	Ac.
$D_{10,n}$	Full Set						.87						.90						.94						.97
$D_{100,n}$	Full Set						**.81**						.85*						.88						.91
	F-Stat.	.51*	.61	**.63**	.27	**.97**	.70	.56	.65	**.75**	.37	**1.0**	.73	.60	.65	**.85**	.49	**1.0**	.77	.59	.65	**.87**	.52	1.0*	.79
	ReliefF	.47	.56	.47	.35	.88	.71	.57	.64	.53	.51	.97*	.76	.67	.70	.70	.58	1.0*	.80	.66	.71	.70	.64	1.0*	.81
	mRMR	.38	.61		.30	.92*	.74	.44	.67		.50	.97*	.79	.45	.70		.60	.99*	.84	.51	.79		.68	1.0*	.86
	SVMone	.41	.55	.49	.37	.88	.74	.50	.62	.52	.58	.96	.80	.71	.74	.57	.79	1.0*	.86	.96*	.96*	.62	.98*	1.0*	.96*
	SVMRFE	.42	.62	.48	.48	.95*	.77	.57	.73	.51	.65	.99*	.82	.74	.80	.57	.79	1.0*	.87	.99*	.99*	.64	1.0*	1.0*	.97*
	E-*F*-Stat.	.50*	.60	.59	.28	.97*	.70	.56	.65	.73*	.35	1.0*	.73	.59	.65	.83*	.48	1.0*	.77	.59	.65	.86*	.52	1.0*	.79
	E-ReliefF	.43	.54	.46	.38	.85	.71	.53	.60	.55	.52	.92	.77	.67	.69	.71	.62	.99*	.81	.72	.74	.74	.72	.99*	.83
	E-mRMR	.47	.59		.29	.93*	.72	.55	.69		.47	1.0*	.77	.66	.75		.66	1.0*	.84	.76	.81		.82	1.0*	.88
	E-SVMone	.46	.57	.54	.36	.93*	.73	.54	.66	.58	.58	.96	.81	.70	.74	.61	.78	.99*	.86	.99*	.99*	.64	1.0*	1.0*	.97*
	E-SVMRFE	.43	.58	.54	.40	.93*	.75	.57	.70	.58	.62	.99*	.81	.78	.81	.62	.84	1.0*	.87	1.0*	1.0*	.66	1.0*	1.0*	.97*
	G-*F*-Stat.	.55*	.67*		.62	.80	.78*	.75	.81		.86	.89	.85*	.88	.88		.93	.93	.90*	1.0*	1.0*		1.0*	1.0*	.97*
	G-ReliefF	.50*	.65*		.63	.77	.78*	.65	.71		.75	.78	.82	.76	.77		.78	.78	.85	.89	.90		.92	.92	.91
	G-mRMR	.56*	.69*		.63	.80	.78*	.77	.81		.85	.88	.85*	.88	.89		.93	.93	.90*	1.0*	1.0*		1.0*	1.0*	.97*
	G-SVMone	.60*	.71*		**.70**	.83	.80*	.82*	.86*		**.92**	.95	.86*	**.94**	**.94**		**.98**	.98*	.92*	1.0*	1.0*		1.0	1.0*	.97*
	G-SVMRFE	**.60**	**.72**		.70*	.83	.80*	**.83**	**.87**		.91*	.94	**.87**	.92*	.92*		.97*	.97*	**.92**	**1.0**	**1.0**		**1.0**	**1.0**	**.97**
$D_{1K,n}$	Full Set						.66						.72						.81						.84
	F-Stat.	.44*	.58*	.23	.23	**.92**	.66	.47	.59	.30	.28	**1.0**	.69	.52	.59	.35	.36	**1.0**	.73	.46	.54	.38	.45	1.0*	.76
	ReliefF	.39	.52*	.27*	.27	.87*	.67	.60	.69	.29	.31	.99*	.69	.53	.62	.38	.48	1.0*	.76	.44	.50	.39	.48	1.0*	.76
	mRMR	.15	.42		.17	.61	.67	.17	.53		.27	.86	.75	.30	.63		.40	.92	.79	.36	.72		.54	.97*	.84
	SVMone	.36	.53*	.21	.29	.83	.68	.38	.52	.23	.45	.96*	.75	.57	.66	.27	.67	1.0*	.82	.75	.80	.26	.81	.99*	.87
	SVMRFE	.29	.48	.20	.34	.71	.70	.46	.66	.21	.59	.95	.81	.70	.78	.25	.77	.99*	.86	.95*	.96*	.24	.98*	1.0*	.96*
	E-*F*-Stat.	.42*	.57*	.21	.21	.91*	.66	.41	.55	.28*	.27	1.0	.69	.51	.60	.33	.36	1.0*	.73	.47	.56	.37	.45	1.0*	.76
	E-ReliefF	.32	.47	**.28**	.27	.83	.67	.47	.60	**.31**	.34	.94	.70	.47	.55	**.48**	.49	.94	.76	.51	.57	**.56**	.56	.97*	.79
	E-mRMR	.30	.52*		.18	.79	.66	.35	.58		.31	.97*	.73	.48	.67		.45	1.0*	.79	.59	.70		.73	1.0*	.85
	E-SVMone	.44*	**.59**	.24	.28	.90*	.68	.46	.57	.27*	.44	.99*	.73	.61	.69	.36	.68	1.0*	.82	.84	.86	.37	.87	1.0*	.89
	E-SVMRFE	.41*	.57*	.24	.31	.87*	.69	.53	.65	.27*	.57	.99*	.79	.72	.76	.35	.83	1.0*	.87	.97*	.97*	.36	.99*	1.0*	.96*
	G-*F*-Stat.	.46*	.55*		.58*	.62	.73*	.68	.72		.79	.83	.83*	.84	.85		.90	.90	.89	1.0*	1.0*		1.0*	1.0*	.97*
	G-ReliefF	.40*	.49		.47	.53	.70	.45	.51		.59	.61	.76	.64	.66		.66	.66	.83	.76	.77		.79	.79	.86
	G-mRMR	.45*	.54*		.54	.56	.73*	.69*	.73*		.79	.81	.83*	.84	.85		.90	.90	.89	.97*	.97*		.98*	.98*	.96*
	G-SVMone	.42*	.50*		.57*	.59	.73*	.74*	.77*		.82*	.84	.84*	**.98**	**.98**		**.99**	.99*	**.93**	1.0*	1.0*		1.0	1.0*	.97*
	G-SVMRFE	**.47**	.55*		**.62**	.63	**.73**	**.76**	**.78**		**.83**	.85	**.85**	.92	.93		.96	.96	.92*	**1.0**	**1.0**		**1.0**	**1.0**	**.97**
$D_{5K,n}$	Full Set						.60						.67						.74						.79
	F-Stat.	.33*	.50*	.18*	.16	**.93**	.64	.32	.47	.21	.22	**1.0**	.68	.29	.40	.25	.34	**1.0**	.72	.42	.50	.27	.38	1.0	.74
	ReliefF	.31	.48	**.22**	.23	.80	.65	.33	.47	.24*	.32	.97*	.70	.46	.55	.26	.43	1.0*	.75	.46	.56	.29	.40	1.0*	.74
	mRMR	.10	.35		.08	.44	.62	.05	.43		.12	.71	.73	.21	.58		.32	.87	.80	.22	.68		.34	.96	.83
	SVMone	.35*	.54*	.17	.24	.85	.66	.35	.49	.17	.41	.99*	.74	.53	.63	.19	.53	1.0*	.78	.61	.68	.20	.63	1.0*	.81
	SVMRFE	.14	.38	.17	.21	.51	.64	.36	.61	.16	.46	.88	.79	.62	.75	.18	.69	.98*	.86	.89	.93	.19	.93	1.0*	.95
	E-*F*-Stat.	.28	.45	.16	.13	.90*	.63	.27	.42	.19	.19	.99*	.68	.30	.42	.23	.35	1.0*	.72	.40	.49	.25	.36	1.0*	.73
	E-ReliefF	.25	.44	.22*	.20	.77	.65	.26	.40	**.27**	.29	.91	.69	.45	.54	**.37**	.47	.97*	.76	.45	.55	**.42**	.42	.99*	.75
	E-mRMR	.17	.40		.13	.58	.63	.20	.52		.21	.96*	.73	.36	.61		.40	.99*	.78	.39	.61		.48	.10	.80
	E-SVMone	**.39**	**.56**	.18*	.22	.91*	.66	.35	.48	.19	.38	.99*	.73	.55	.64	.22	.55	1.0*	.79	.67	.73	.24	.62	1.0*	.82
	E-SVMRFE	.36*	.56*	.18*	.23	.82	.65	.44	.59	.19	.48	.99*	.77	.67	.74	.22	.72	1.0*	.85	.91	.92	.24	.95	1.0*	.94
	G-*F*-Stat.	.35*	.46		.47*	.52	.68*	.64*	.67*		.76*	.77	.82*	.78	.79		.85	.85	.87	.97*	.97*		.99*	.99*	.96*
	G-ReliefF	.30	.42		.39	.43	.66	.39	.45		.53	.54	.74	.61	.63		.62	.62	.82	.65	.67		.73	.73	.85
	G-mRMR	.33*	.45		.45	.48	.69*	.62*	.66*		.74	.74	.82*	.81	.82		.87	.87	.88	.93	.93		.96	.96	.94
	G-SVMone	.36*	.47		.44*	.48	.69*	.61	.65		.72	.72	.81*	.87*	.88*		.93*	.93	.91*	.99*	.99*		1.0*	1.0*	.97*
	G-SVMRFE	.37*	.48		**.50**	.52	**.70**	**.69**	**.71**		**.79**	.79	**.83**	**.89**	**.89**		**.94**	.94	**.91**	**1.0**	**1.0**		**1.0**	**1.0**	**.97**

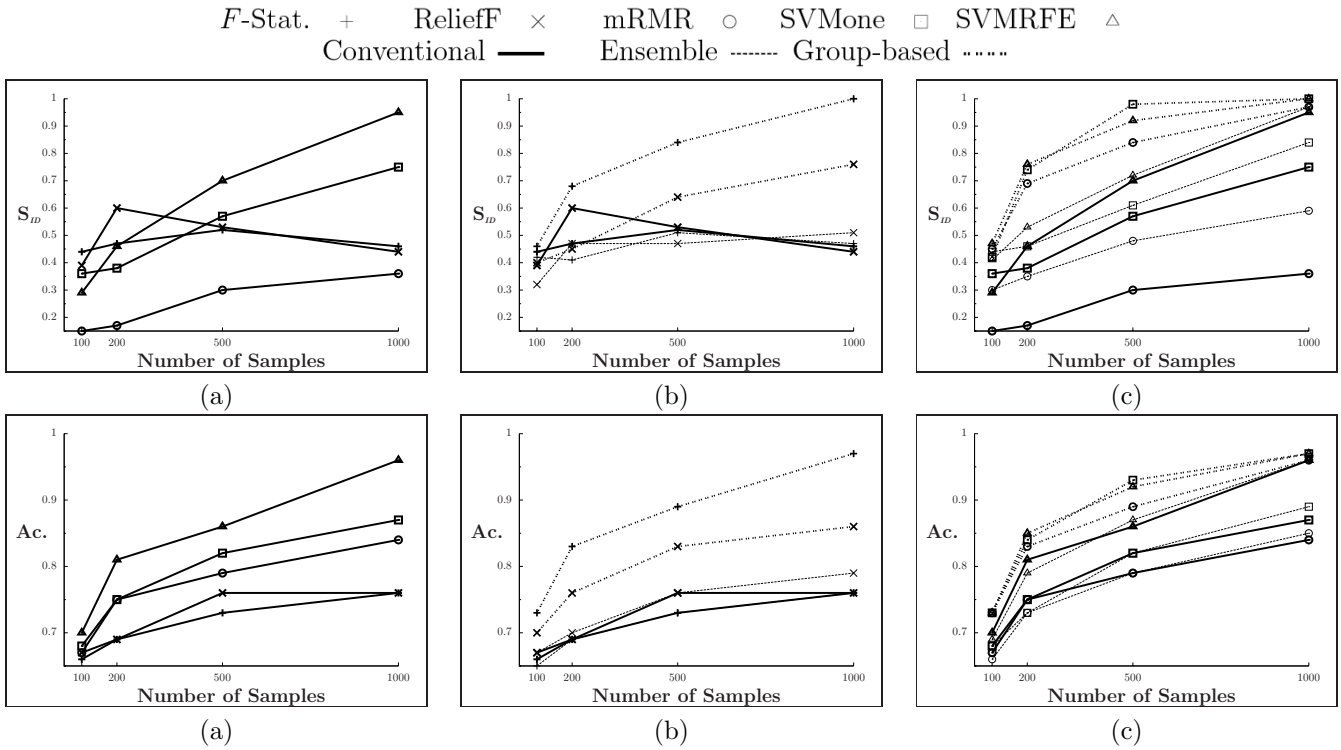

Figure 1. Comparison of various feature selection algorithms based on $D_{1K,n}$ data sets. Figures in the top row depict the stability S_{ID} values of the 15 different feature selection algorithms: (a) conventional version of 5 algorithms, (b) 3 versions of F-statistics and ReliefF, and (c) 3 versions of mRMR, SVMone, and SVMFRE. Figures in the bottom row depict the corresponding SVM classification accuracy values. The same set of values are also reported in Table 3.

proach increasing accuracies while increasing stability, which is different from the observation in (1). conclusion: specifically improving stability and balance it in the relevance and redundancy framework can/may lead to both increased stability and accuracies, hence suggests a new paradigm of feature selection. shift from traditional feature selection which focuses on the trade-off of relevance and redundancy. (4) there are special cases, when stability increases, accuracies drops. domain experts may have to choose algorithm using a weighting function for the two criteria. (5) group based group works even better than ensemble approach.

5. Conclusion

Stability is an important problem. What we did in the paper. We conclusions we draw from the results.

In this paper, we have systematically studied three important issues that affect the generalization ability of feature selection: relevance, redundancy, and stability. We have provided theoretical explanations on why stable feature selection is feasible and how to design stable feature selection algorithms by balancing the trade-off of maximum relevance, minimum redundancy, and maximum stability. We have compared six feature selection algorithms representing a variety of feature selection algorithms, based on widely used relevance and redundancy measures and a general stability measure. Experimental results on public microarray data sets have revealed several interesting findings about the trade-off of relevance, redundancy, and stability.

A future research direction is to theoretically study the relationship between stability and feature selection accuracy. Another direction is to study the effect of noise on stable feature selection. Third, we can study the

Table 4. Comparison of various feature selection algorithms based on **microarray** data sets.

Algorithms	Colon				Leukemia				Prostate				Lung			
	S_{ID}	S_V	S_R	Ac.	S_{ID}	S_V	S_R	Ac.	S_{ID}	S_V	S_R	Ac.	S_{ID}	S_V	S_R	Ac.
Full Set				.82				.97				.90				.99
F-Stat.	.42*	.71*	**.52**	**.86**	**.56**	**.80**	.54*	.96*	.54*	.78*	.62*	.90*	.68*	.87*	.53	.99*
ReliefF	.43*	.71*	.43	.84*	.53*	.78	.52*	.97*	**.56**	**.81**	.52	.91*	.60	.83	.82	.99*
mRMR	.42*	**.75**		.85*	.49	.76		.97*	.47	.73		**.92**	.66	.85		.99*
SVMone	.24	.67	.24	.82*	.32	.62	.36	.96*	.34	.64	.37	.91*	.27	.58	.53	.99*
SVMRFE	.20	.69	.24	.80	.25	.59	.36	.97*	.24	.59	.36	.92*	.29	.64	.52	.99*
E-F-Stat.	.42*	.71*	.51*	.85*	.55*	.80*	**.55**	.96*	.53*	.79*	**.63**	.92*	**.70**	**.88**	.56	.99*
E-ReliefF	.40*	.70*	.44	.84*	.51*	.78	.54*	.96*	.55*	.81*	.60*	.92*	.66	.86	**.84**	.99*
E-mRMR	**.44**	.75*		.85*	.53*	.78		.96*	.50	.75		.90*	.68*	.87*		.99*
E-SVMone	.30	.70*	.41	.83*	.44	.72	.46	.96*	.44	.72	.57	.91*	.41	.69	.66	.99*
E-SVMRFE	.28	.71*	.41	.82*	.42	.71	.46	.97*	.41	.70	.56	.92*	.47	.75	.65	.99*
G-F-Stat.	.17	.71*		.85*	.16	.70		.94	.26	.72		.91*	.26	.85		.99*
G-ReliefF	.12	.66		.82*	.12	.67		.95*	.26	.76		.91*	.20	.73		.99*
G-mRMR	.16	.67		.85*	.14	.65		.95*	.22	.64		.90*	.20	.78		.99*
G-SVMone	.12	.67		.78	.10	.50		.94	.20	.59		.90*	.15	.55		.99*
G-SVMRFE	.09	.68		.82*	.08	.51		.95*	.12	.54		.90*	.17	.66		.99*

combination of ensemble with group-based methods.

References

Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, *2*, 499–526.

Ding, C., & Peng, H. (2003). Minimum redundancy feature selection from microarray gene expression data. *Proceedings of the Computational Systems Bioinformatics conference (CSB'03)* (pp. 523–529).

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*, 389–422.

Kalousis, A., Prados, J., & Hilario, M. (2007). Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and Information Systems*, *12*, 95–116.

Li, T., Zhang, C., & Ogihara, M. (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, *20*, 2429–2437.

Nilsson, R., Pena, J. M., Bjorkegren, J., & Tegner, J. (2006). Evaluating feature selection for svms in high dimensions. *Proceedings of the ECML Conference* (pp. 719–726).

Pepe, M. S., Etzioni, R., Feng, Z., & et al. (2001). Phases of biomarker development for early detection of cancer. *J Natl Cancer Inst*, *93*, 1054–1060.

Robnik-Sikonja, M., & Kononenko, I. (2003). Theoretical and empirical analysis of Relief and ReliefF. *Machine Learning*, *53*, 23–69.

Saeys, Y., Abeel, T., & Peer, Y. V. (2008). Robust feature selection using ensemble feature selection techniques. *Proceedings of the ECML Confernce* (pp. 313–325).

Witten, I. H., & Frank, E. (2005). *Data mining - pracitcal machine learning tools and techniques*. Morgan Kaufmann Publishers.

Yu, L., Ding, C., & Loscalzo, S. (2008). Stable feature selection via dense feature groups. *Proceedings of the 14th ACM International Conference on Knowledge Discovery and Data Mining (KDD-08)* (pp. 803–811).