ACM Conf. on Bioinformatics, Computational Biology and Biomedical Informatics (ACM BCB), September 2013

# Stable Feature Selection with Minimal Independent Dominating Sets

Le Shu
Computer and Information Science, Temple University
1805 N. Broad St.
Philadelphia, PA
slevenshu@gmail.com

Tianyang Ma
Computer and Information Science, Temple University
1805 N. Broad St.
Philadelphia, PA
ma.tianyang@gmail.com

Longin Jan Latecki
Computer and Information Science, Temple University
1805 N. Broad St.
Philadelphia, PA
latecki@temple.edu

## ABSTRACT

In this paper, we focus on stable selection of relevant features. The main contribution is a novel framework for selecting most informative features which can preserve the linear combination property of the original feature space. We propose a novel formulation of this problem as selection of a minimal independent dominating set (MIDS). MIDS of a feature graph is a smallest subset such that no two of its nodes are connected and all other nodes are connected to at least one node in it. In this way, the diversity and coverage of the original feature space can be preserved.

Furthermore, the proposed MIDS framework complements standard feature selection algorithms like SVM-RFE, stability lasso and ensemble SVM RFE. When these algorithms are applied to feature subsets selected by MIDS as opposed to all the input features, they select more stable features and achieve better prediction accuracy, as our experimental results clearly demonstrate.

## Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

## General Terms

Algorithms

## Keywords

Feature Selection, Minimum Independent Dominating Sets, Stability

## 1. INTRODUCTION

One of the main challenges in feature selection is the selection stability, Since there are more and more high-dimensional data sets with limited samples. A small variation in the samples often leads to huge differences in selected features. This fact makes classifier learning biased and hence severely impairs the generalization capability.

In particular, gene expression data is inherently high dimensional, which is usually in the order of thousands and even tens of thousands. Meanwhile, due to the expense and difficulty in collecting the data, the number of samples can be even less than one hundred. This fact makes the following two tasks very challenging: 1) predict prognostic value given gene expression profile [17, 18, 19], i.e., make prediction on whether a patient has disease or not, and 2) identify the gene expression profiles associated with a certain disease. For task 1), the prediction accuracy may drop dramatically if overfitting happens when training classifiers, such as SVM or Adaboost. In order to enhance the generalization ability, it is very important to prevent the potential overfitting during the training phase. A common solution is to reduce the dimension of the data by filtering the features. Its efficiency has been proved by many feature selection algorithms, such as [3, 6, 10, 21, 1, 15]. Using these feature selection methods, task 2) seems to be also solved, since the output of those algorithms corresponds to the set of relevant features for prediction. However, this is only the case if the selected relevant features are stable, i.e., small change in the training samples, leads to only small variance in the selected features. Moreover, a feature selection algorithm with good stability will further decrease the danger of overfitting, since it is less sensitive to the training data, which is a key factor to ensure proper generalization power. Furthermore, by presenting the selected feature set to an expert with prior knowledge, such as biologist, this will enable a better analysis and understanding between the genes and disease. Although a lot of them have not been discovered yet, it is well known that several genes could have similar functionality, therefore selecting any one of those features in a group will deliver a good prediction.

In this paper, we propose a novel feature selection framework. Its focus is on stability of the selected features, which also delivers a good prediction on test data sets. We have two main observations for developing a stable feature selection method in high-dimensional and small-sample data. The first one is that highly correlated features with same functionality can be selected differently due to slight variation in the training data set, especially, when there are very few training samples. This problem have been confirmed by several recent works on feature selection [4, 9, 20]. Consequently, identifying highly correlated features and finding


*BCB* '13, September 22 - 25, 2013, Washington, DC, USA

out a representative feature subset should be part of any stable feature selection method. The second one is that many feature selection algorithms tend to select a minimum subset of features to predict the target label. This indicates that only one (or at most few features in the same highly correlated feature set) should be selected, with the other features in the same group ignored. Therefore, it is crucial to guarantee those ignored features can be truly represented by the selected features, without any information loss, otherwise we may lose some relevant features which will in turn harm the predication performance and the feature selection stability.

Motivated by the two observations, our goal is to identify the smallest set of representative features by suppressing features with similar functionality. The main contribution of this paper is a novel framework to effectively find the most representative features. We solve this problem as finding a Minimal Independent Dominating Set (MIDS). First, a graph $G = (V, E)$ is constructed, where vertices $V$ represent features and two features are linked by an edge in $E$ if their correlation is significant. Hence $G$ is a binary and undirected graph.

Finding MIDS is a classical problem in graph theory, in which a subset $S \subset V$ with the minimal number of vertices needs to be found such that no two vertices in $S$ are adjacent and every vertex not in $S$ is connected to at least one vertex in $S$. The main advantage of this framework for the proposed feature selection is that it has two guarantees: 1) For the selected representative features, i.e., the features in the MIDS, any two of those features have very low correlation. This promises that the *diversity* of the representative features in the MIDS. 2) For any feature not in MIDS, which is removed from consideration, there is at least one feature in the MIDS with a large correlation. This indicates that for the abandoned features, features with similar functionality will be very likely preserved in the MIDS. We refer to this property as *coverage*.

The proposed MIDS feature selection framework is different from the dense group finder algorithms in [20], which is motivated by a key observation that in the sample space, the dense core regions (peak regions), measured by probabilistic density estimation, are stable with respect to sampling of the dimensions (samples). While the methods show that the features can be effectively grouped together, it is hard to control the inter and intra cluster differences, which means that it is hard to guarantee *diversity* and *coverage*. In our approach, we use Pearson's correlation as the similarity measure. Due to the non-transitivity of Pearson's correlation, we only need to consider direct neighbors when finding representative features and its corresponding high correlated features.

Finding the MIDS in a graph is in general NP-complete [5], but for small problems it can be solved with mathematical programming solvers such as CPLEX[8]. Due to large size of the considered problems (large number of features), we first map this problem to the problem of finding Maximum Weight Independent Set (MWIS), which is then solved in a relaxed setting by a recently proposed algorithm [2]. The algorithm is very effective and fast. After we obtain the MIDS,any other feature selection methods, such as LASSO, can be used to further determine which of features in MIDS are relevant to the target label.

The pipeline of our method is shown in Figure 1. The main goal of computing MIDS is to identify independent dominating features, which allows us to represent the original feature space without information loss. The main benefit is that the number of independent dominating features is usually much smaller than that of the number of the original features. By doing this, the high dimensional data can be reduced to a much lower dimension while minimizing the possible loss of useful features. In the second step, we can use any standard stability feature selection method to identify the relevant features among the dominating features. In this paper, we consider the following three algorithms: SVM-RFE [6],a recursive feature elimination algorithm using support vector machines, and Stability LASSO [14], a variance of LASSO with bagging embedded, which is shown to be one of the state-of-the art feature selection approaches. The third algorithm is ensemble SVM-RFE [16], which aggregates results from SVM-RFE on a number of bootstrapping samples.

We present experiments on both synthetic and benchmark data sets, and there are performance improvements in both the stability of the selected features and in the prediction accuracy. This demonstrates the efficiency and effectiveness of the proposed method.

The rest of the paper is organized in the following way. In Section 2, we review the related works, especially those approaches which put effort on stable feature selection. In Section 3, the details of the proposed dominating set based feature selection algorithm are elaborated. Experimental settings and results are proposed in section 4, detail analysis about the stability and classification accuracy can also be found in section 4. In section 5, we conclude with some final marks.

## 2. RELATED WORK

For many years, feature selection was simply considered as a dimensionality reduction method to improve the prediction accuracy, without serious need in understanding which features are selected. However, in recent years, a stable feature selection draws more attention. In particular, there are several recent works showing that traditional feature selection approaches are sensitive to the training data, especially in the setting with small number of samples but with high feature dimensionality [4, 9, 20]. This means that small variation in the training samples may lead to huge variation in the selected features. Consequently, the features selected on training data do not generalize well to the test data, due to the overfitting problem. Meanwhile, a quickly growing number of researchers agree that stable feature selection is critical for a further study of the understanding between the genes and diseases [7].

There are two main research directions aiming at achieving a more stable feature selection without reducing the prediction accuracy. One direction is ensemble based stability feature selection approaches. These methods combine subsampling (bootstrapping) with traditional feature selection algorithms. Saeys et al. [16] studied bagging-based ensemble feature selection, which aggregates the results from a conventional feature selection algorithm, such as SVM-RFE, and is repeatedly applied on a number of bootstrapped samples of the same training set. Their results show that stability ensemble SVM-RFE can improve a lot compared to a single run of SVM-RFE. Meinshausen et al. [14] introduce stability selection based on subsampling in combination with random LASSO. They perturbed the data many times and

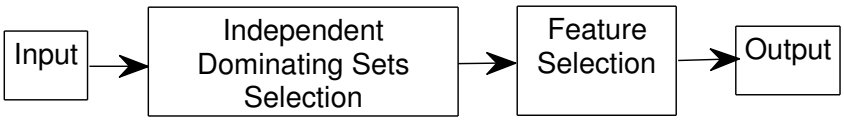


Figure 1: The processing flow of the proposed framework for stable feature selection based on MIDSs.

choose all variables that occur in a large fraction of the resulting selection sets. By doing the subsampling, a more stable and with better generalization feature set is identified. However, this group of methods only consider reducing variance in the feature selection algorithm, but without putting efforts in explicitly exploring the structure among the features, such as features which are highly correlated and function similarly.

The other direction is feature grouping based approaches. It is known that there exists highly correlated feature groups among high dimensional data, especially for gene expression data. The key insight of these approaches is to identify consensus feature groups by explicitly exploring the structure among the features. Then within each feature group, only one feature is taken as representative. By doing this, the number of the features can be significantly reduced so that the selected feature set is less likely to suffer from the overfitting problem during the training phase. An example approach is DRAGS proposed by Loscalzo et al. [20], which exploits the intrinsic correlations among a large number of features to identify consensus feature groups and treat features in the each dense groups as a coherent entity for feature selection. The later work of this group (CGS) [11] combines subsampling and grouping features together. They first identify consensus feature groups by subsampling training samples and then select the relevant features by treating the consensus feature groups as entities.

The proposed framework is inspired by works in both directions. Particularly, our main contribution is a novel framework for identifying a set of representative features by explicitly exploring the features structure through finding the smallest dominating set. The main advantage of the proposed framework is its ability to explicitly guarantee the diversity and coverage properties, which proved to be a good guidance for selecting representative features, as demonstrated by the experimental results. In the second phase we can run any feature selection algorithm on the selected, representative features. Since the feature dimension have been reduced with the functionality of the original feature space preserved, the stability and classification accuracy of the feature selection algorithms can be improved a lot.

# 3. METHODOLOGY

In this section, we introduce our stable feature selection framework in detail. we first formulate the selection of smallest set of representative features as the minimum independent dominating set. Then we formulate our problem to maximum weight independent set problem and give the solution.

## 3.1 Problem formulation

Given is a training data set containing $N$ data points in $p$ dimensional feature space with $N \ll p$. We represent it with a data matrix $X = (x_{i,j}) \in \mathbb{R}^{N\times p}$, where each data point $X_i = (x_{i,1}, \ldots, x_{i,p}) \in \mathbb{R}^p$ is a row vector. We also given a corresponding set of labels $\mathbf{y} = (y_1, y_2, \cdots, y_N)^T$, i.e., $\mathbf{y}$ is a column vector. We will denote these data with $D_{X,\mathbf{y}}$. The columns of data matrix $X$ represent the $p$ features. Hence we can write $X = (F_1, F_2, \cdots, F_p)$, where each $F_i \in \mathbb{R}^N$ is a column vector.

Given a pair of features $F_i$ and $F_j$,we use the Pearson's correlation coefficient (PCC) to measure their correlation:

$$\rho(i,j) = \frac{\sum_k (F_{i,k}-\overline{F_{i,k}})(F_{j,k}-\overline{F_{j,k}})}{\sqrt{\sum_i (F_{i,k}-\overline{F_{i,k}})^2}\sqrt{\sum_i (F_{j,k}-\overline{F_{j,k}})^2}} \quad (1)$$

Where $\overline{F_{i,k}}$ is the mean of $F_i$ and $\overline{F_{j,k}}$ is the mean of $F_j$.

PCC captures linear correlation between pairs of features with efficient computation. The value of $\rho(i,j)$ lies between $-1$ and 1, inclusive. If $F_i$ and $F_j$ are completely correlated, $\rho(i,j)$ can take value of 1 (positively correlated) or $-1$ (negatively correlated). If $F_i$ and $F_j$ are independent, $\rho(i,j)$ is 0.

We construct an undirected, binary graph $G = (V, E)$, which we call feature graph. The vertex set $V$ represents the features, so $|V| = p$, and $E$ corresponds to the correlation between two features. Given a threshold $\epsilon > 0$ two features are connected if their correlation is above the threshold, i.e.,

$$E(i,j) = \begin{cases} 1 & \text{if} \quad |\rho(i,j)| > \epsilon \ \ and \ \ i \neq j \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Unlike many common mathematical binary relations, Pearson's correlation is not transitive. This non-transitivity property has been pointed out by McNemar [13] in 1949. A more detailed way of looking at the non-transitivity property is that, given three quantitative features $F_i$, $F_j$, and $F_k$, a positive correlation between $F_i$ and $F_j$ and a positive correlation between $F_j$ and $F_k$ (in terms of Pearson's correlation coefficients, $\rho(i,j) > 0$ and $\rho(j,k) > 0$), not necessarily mean that $F_i$ and $F_k$ will be positively correlated. In fact, $F_i$ and $F_k$ might be uncorrelated ( $\rho(i,k) = 0$ ) or even negatively correlated ($\rho(i,k) < 0$). This property holds even when feature pair $F_i, F_j$ is highly correlated and feature pair $F_j, F_k$ is also highly correlated.

For the above reasons, it make sense that features can only be represented by its direct neighbors. In order to guarantee that the selected features can preserve the functionality of all the removed features. We formulate the problem as independent dominating set.

A subset of graph nodes $S \subset V$ is called *independent dominating set (IDS)* if

$$\forall v_i, v_j \in S, \quad E(i,j) = 0, \text{ and} \tag{3}$$
$$\forall v_i \notin S \quad \exists v_j \in S \quad E(i,j) = 1 \tag{4}$$

Consequently, the correlation between any two features in the $S$ set must be smaller or equal to $\epsilon$, and any feature not in $S$ must have at least one feature in $S$ with correlation larger than $\epsilon$. The threshold $\epsilon$ can be determined empirically with certain prior knowledge, or through cross-validation. In all of our experiments, $\epsilon$ is determined through an exhaustive search on a validation set in the range between 0.5 and 0.8.

Our aim is to identify a MIDS of features from high dimensional feature space given only a small number of data samples. The intuition is that the MIDS is able to represent all the other features in the original feature space with high diversity and coverage.

However, finding a MIDS is a known NP-complete problem [5]. In order to solve it, we first reformulate it to a NP-complete problem, which we then solve in a relaxed setting.

## 3.2 Finding the Minimal Independent Dominating Set

We first show that the problem of finding a MIDS set can be expressed as a problem of finding a maximum weight independent set (MWIS). We define the weight of node $v_i$ of graph $G$ as the cardinality of its neighborhood

$$w_i = |\mathcal{N}(v_i)|, \tag{5}$$

and consider the weight vector $\mathbf{w} = (w_1, \ldots, w_p)^T$.

Let $\mathbf{x} = (x_1, \ldots, x_p)^T \in \{0,1\}^p$ be an indicator vector of a subset $S$ of the vertices of graph $G$, i.e., $x_i = 1$ if and only if $v_i \in S$. The problem of finding MWIS can be expressed as integer program (IP):

$$\begin{aligned} &\text{maximize} \quad g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} \\ &\text{subject to} \quad \mathbf{x^T}E\mathbf{x} = 0 \quad \text{and} \quad \mathbf{x} \in \{0,1\}^p. \end{aligned} \tag{6}$$

It is easy to see that condition $\mathbf{x^T}E\mathbf{x} = 0$ is equivalent to the independence condition (3).

Let $\mathbf{x}$ be a solution of (6) and let $S$ be the corresponding MWIS. Since no two points of $S$ are neighbors, we have that

$$g(\mathbf{x}) = |\mathcal{N}(S)| = |\sum_{i \in S} \mathcal{N}(v_i)| = |V \setminus S|. \tag{7}$$

Consequently, by solving (6) we have selected an independent set $S$ such that $V \setminus S$ has the maximal cardinality. However, this means that $S$ has the minimal cardinality among all independent sets. We have just proved the following proposition.

**Proposition 1.** Any solution of problem (6) is a minimal independent dominating set (MIDS) of graph $G$.

In order to solve problem (6), we reformulate it as the following integer quadratic program (IQP):

$$\begin{aligned} &\text{maximize} \quad h(\mathbf{x}) = \frac{1}{||\mathbf{w}||}\mathbf{w}^T\mathbf{x} - \mathbf{x^T}E\mathbf{x} \\ &\text{subject to} \quad \mathbf{x} \in \{0,1\}^p. \end{aligned} \tag{8}$$

To show that the reformulation from (6) to (8) always holds [2], let us assume independence condition (3) does not hold given $\mathbf{x}$, which means that $\mathbf{x^T}E\mathbf{x} \geq 1$. Considering $\frac{1}{||\mathbf{w}||}\mathbf{w}^T\mathbf{x}$ is upper-bounded by 1, therefore $h(\mathbf{x}) \leq 0$. However, given $\mathbf{w}$, it is easy to derive a solution $\mathbf{x}$ with only one element equal to 1, i.e., $\mathbf{x_i} = 1$ with $\mathbf{w_i} > 0$, with all other elements to be 0, this will give $h(\mathbf{x}) > 0$. This implies that if a *discrete* solution of (8) is optimal, independence condition (3) must hold.

Since problem (8) is still NP-complete, we relax its binary constraints to continuous ones and obtain

$$\begin{aligned} &\text{maximize} \quad f(\mathbf{x}) = \frac{1}{||\mathbf{w}||}\mathbf{w}^T\mathbf{x} - \mathbf{x^T}E\mathbf{x} \\ &\text{subject to} \quad \mathbf{x} \in [0,1]^p. \end{aligned} \tag{9}$$

Let $A$ be a diagonal matrix with diagonal entries equal to $\frac{1}{||\mathbf{w}||}\mathbf{w}^T$, i.e., $A = diag(\frac{1}{||\mathbf{w}||}\mathbf{w})$. We obtain an equivalent formulation to (9):

$$\begin{aligned} &\text{maximize} \quad f(\mathbf{x}) = \mathbf{x^T}A\mathbf{x} - \mathbf{x^T}E\mathbf{x} \\ &\text{subject to} \quad \mathbf{x} \in [0,1]^p. \end{aligned} \tag{10}$$

We solve problem (10) with the algorithm proposed in [12], which is a general case for the algorithm in [2]. As is the case of the experimental results reported in [12], also in our settings the algorithm always yields a discrete solution, which in turn guarantees that $\mathbf{x^T}E\mathbf{x} = 0$.

We denote with $X_S$ the data set obtained as the solution of (10), i.e., feature $F_i$ is a column vector of $X_S$ if and only if $i \in S$ if and only if $\mathbf{x}^*_i = 1$, where $\mathbf{x}^*$ is the solution of (10).

Hence $X_S$ is a $N \times q$ matrix composed of the dominating features (each feature is a column vector and each data point is a row vector). Usually we have $q \ll p$. The ratio between the number of features in minimal independent dominating set $S$ and the number of original features may vary for different data sets. In our experiments (in Table 2), the ratio range from %0.5 to %25. Only the dominating features in $S$ are taken for further processing. As our experimental results demonstrate, this reduction of features has a very positive effect on both stability of feature selection as well as prediction accuracy.

# 4. EMPIRICAL STUDY

In this section, results of our analysis of stability feature selection framework on large feature and small sample size domains are presented. First, the data sets and the experiment settings used in this analysis are briefly described. Second, we analyze two key parts of stability feature selection techniques: stability and prediction accuracy.

## 4.1 Data sets and Experiment Setting

In this section, we empirically study the framework of stable feature selection based on MIDS. Especially, we focus on the study of the stability of the selected features. The stability of feature selection algorithms can be defined as the variation in feature selection results due to variations in the data set. To measure the features selection stability, we compute the similarity between two features sets $R_1$ and $R_2$ as in [4, 9]:

$$Sim_{ID}(R_1, R_2) = \frac{2|R_1 \bigcap R_2|}{|R_1| + |R_2|}, \tag{11}$$

where $|R_1 \bigcap R_2|$ is decided by the number of overlapping features in two feature sets. A higher similarity means a

larger overlap between two feature sets, which indicates that the selected features are more stable against the changes in the training samples.

We evaluate three feature selection algorithms with respect to stability and classification accuracy with or without our MIDS framework. The first one is SVM-RFE [6], a baseline gene selection algorithm for cancer classification using support vector machines. The second one is stability LASSO [14], a variance of LASSO with bagging embedded, which can select stable features. The third one is ensemble SVM-RFE[16], which applies SVM-RFE on a number of bootstrapped samples of the same training set. The ensemble algorithms can improve the performance a lot when compared to a single run of a given algorithm.

The test data sets include two part: The first part includes two synthetic data sets, the details can be found in Table 1. The second part includes seven benchmark data sets, which were taken from the bioinformatics and biomedical domain. Those seven benchmark data sets are characterized in Tables 2. For synthetic data sets, we generate the data in the same way as in [11]. The synthetic data set consists of 1000 training samples randomly drawn from the same distribution $D_{X,y}$. The feature matrix $F$ contains 1000 features, including 100 mutually independent features, $F_1, F_2, ..., F_{100}$, and a number of $(10 \pm 5)$ highly correlated features to each of these 100 features. In each correlated group, the Pearson correlation coefficient of each feature pair is within $(0.5, 1)$ so that the average pairwise correlation is below 0.75. The target label $\mathbf{y}$ is decided based on the first 10 features $F_1, F_2, ..., F_{10}$ only using a linear function of equal weight to these 10 truly relevant features. We also follow the same procedure to generate another data set, in which the number of relevant feature groups remains to be 10, but with the number of independent feature groups increased to 250 and the number of highly correlated features increased to $(20 \pm 5)$. A summary of these data sets is provided in Table 1. In our algorithm, the threshold $\epsilon$ in Function 2 is set to 0.6, the corresponding MIDS feature number for the 7 real world is listed in Table 2.

For every data set, the evaluation is performed in 10 fold cross-vaidation. The data set is divided into 10 parts, the features are selected based on 9 folds and the prediction accuracy is evaluated on the remaining hold-out fold. We repeat the above procedure for 10 times, each time obtaining the feature ranking. Then we vary the selected feature sets, and the feature number varies from 1 to 50 based on the ranking of features. For each of these sets we compute the stability of selected features as the average $Sim_{ID}$ between every two sets of the 10 sets of selected features produced by the 10 fold cross validation. For example, we obtain 10 sets composed of 20 selected features, and compute the average stability of these sets with Eq 11. Finally, we repeat 10 times the 10 fold cross validation and report the average stability and accuracy for each data set. In order to obtain the prediction accuracy on the benchmark data sets, a linear SVM classifier is trained based on the selected features from the same training set (9 folds) and tested on the corresponding hold-out fold. For the synthetic data set, we train the SVM classifier on 9 out of the 10 folds, and test on an independent test set of 500 samples, which is generated from the same distribution of the training data set.

It is important to note that the stability of feature selection results should be considered in combination with classification accuracy. An algorithm that yields very stable feature sets makes no sense if it returns a badly performing model.

As the proposed method is designed for a binary classification problem, one-vs-all strategy is used when dealing with multi-class classification task, such as Lymphoma and SRBCT.

## 4.2 Result and Discussion

For synthetic data sets, besides stability and prediction accuracy, we are able to measure the precision of the selected feature as the percentage of the ground-truth relevant features among the selected features.

In Figure 2, we can see that when the size of samples becomes larger, all algorithms perform better in all three performance metrics. However, with relatively small number of training samples, SVM-RFE, ensemble SVM-RFE and stability Lasso have better performance with MIDS framework. By applying SVM-RFE, stability LASSO or ensemble SVM-RFE alone without our MIDS method, the performance becomes much worse. Especially, they may fail in the task of selecting truly relevant features, i.e., these three algorithms without MIDS have very low precision when the sample number is 100.

The results on synthetic data sets clearly demonstrates that for a large number of features, many of which have similar functionality, SVM-RFE, stability LASSO and ensemble SVM-RFE cannot work effectively to select relevant features without first identifying those representative features and removing the redundant features.

Figure 3 compares SVM-RFE, ensemble SVM-RFE, Stability LASSO with and without MIDS on stability and prediction accuracy performance on the 7 benchmark data sets. Figures with '$Sim_{ID}$' as y-axis show the stability varies with the number of selected features. Figures with 'Accuracy' as y-axis show the SVM classification accuracy for various numbers of selected features.

The quantitative results are summarized in Tables 3 and 4. Table 3 reports the best prediction performance for all the compared algorithms as the function of the selected feature number, which is reported in brackets. Table 4 reports the stability for the feature subsets which achieve the highest prediction accuracy.

From Table 3 and Table 4, it is easy to observe that SVM-RFE with MIDS can achieve higher classification accuracy as well as higher stability. The average classification accuracy improves from 0.899 to 0.926 with MIDS framework. The average similarity between selected features also increase from 0.473 to 0.700. The performance enhancement for SVM-RFE algorithms with MIDS is the most significant among all the three algorithms. For ensemble SVM-RFE and stability LASSO, The average classification accuracy also improves 0.02. Particularly, stability LASSO with MIDS got the best classification accuracy by only selecting 17 features, which is much higher than the other algorithm. The proposed MIDS framework with the other two stability feature selection algorithms also has a higher stability on average. Stability LASSO with MIDS achieved the best stability in 5 of the 7 data sets. On average, stability LASSO with MIDS achieves average stability as 0.789 with only 17 features, which is the best among the compared approaches. SVM-RFE with MIDS and ensemble SVM-RFE with MIDS have more stable results when compares to the algorithms

Table 1: Summary of Synthetic Data Sets

| Datasets | Features | Groups | Rel. Feat. |
|---|---|---|---|
| $D_1$ | 1000 | 100 $(size10 \pm 5)$ | 10 |
| $D_2$ | 5000 | 250 $(size20 \pm 5)$ | 10 |

Table 2: Summary of Benchmark Data Sets

| Datasets | Classes | Instances | Genes | MIDS Genes |
|---|---|---|---|---|
| Colon | 2 | 62 | 2000 | 65 |
| Leukemia | 2 | 72 | 7120 | 1235 |
| Ovarain | 2 | 253 | 15154 | 72 |
| Lung | 2 | 181 | 12533 | 1599 |
| Lymphoma | 3 | 62 | 4026 | 1175 |
| Pancreatic | 2 | 181 | 6771 | 134 |
| SRBCT | 4 | 63 | 2308 | 765 |

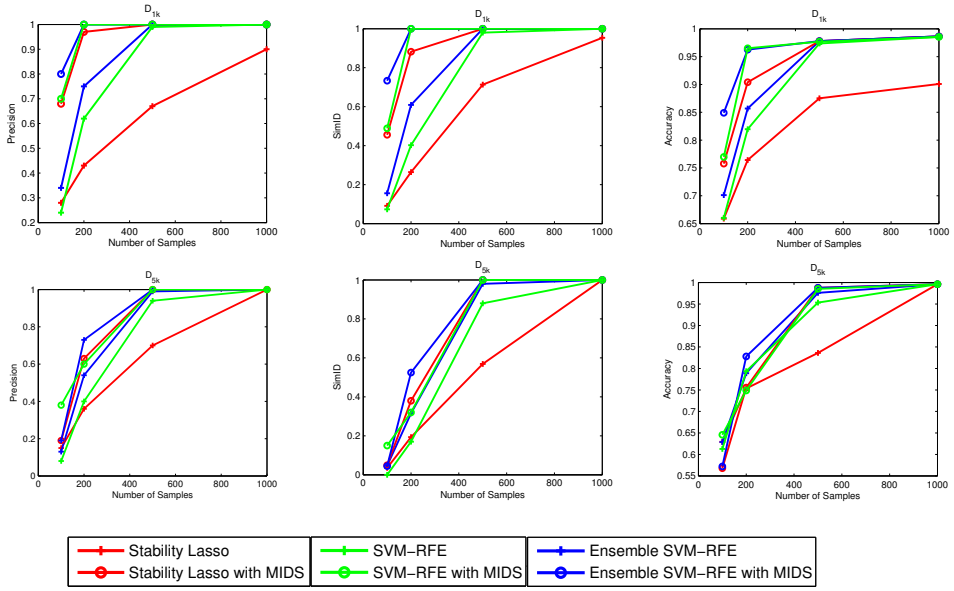


Figure 2: Comparison of algorithms with or without MIDS on synthetic data sets. Column 1: precision w.r.t. the truly relevant features. Column 2: stability of the selected features measured with $Sim_{ID}$. Column 3: SVM classification accuracy, for the top 10 selected features.

| Datasets | SVM-RFE[6] | | Ensemble SVM-RFE[16] | | Stability LASSO[14] | |
|---|---|---|---|---|---|---|
| | Without MIDS | With MIDS | Without MIDS | With MIDS | Without MIDS | With MIDS |
| Colon | 0.83(40) | **0.94(39)** | 0.82(9) | 0.93(50) | 0.83(5) | 0.94(50) |
| Leukemia | 0.96(43) | **0.99(14)** | 0.96(32) | 0.98(34) | 0.96(18) | 0.98(26) |
| Ovarain | 0.99(20) | 1.00(12) | 1.00(9) | 1.00(9) | 1.00(4) | **1.00(3)** |
| Lung | 0.98(19) | 0.99(16) | 0.98(16) | 0.99(20) | 0.992(6) | **0.995(3)** |
| Lymphoma | **1.00(27)** | 0.99(9) | 0.99(24) | 0.98(24) | 0.99(36) | 0.99(11) |
| Pancreatic | 0.56(50) | 0.61(43) | 0.56(15) | 0.60(29) | 0.63(9) | **0.64(4)** |
| SRBCT | 0.97(48) | 0.96(50) | 0.99(40) | 1.00(27) | 1.00(28) | **1.00(24)** |
| **Average** | 0.899(35.3) | 0.926(24.7) | 0.900(20.7) | 0.926(27.6) | 0.913(15) | **0.935(17)** |

Table 3: Performance comparison of classification accuracy of different feature selection algorithms with or without MIDS on different data sets. The best results are highlighted in bold.

without MIDS.

In this case, we can conclude that the framework, which using MIDS to remove features with same functionality, does not reduce the classification accuracy and stability of all the three algorithms. The selected feature number is on par with the algorithms without MIDS frame work. Instead,

| Datasets | SVM-RFE[6] | | Ensemble SVM-RFE[16] | | Stability LASSO[14] | |
|---|---|---|---|---|---|---|
| | Without MIDS | With MIDS | Without MIDS | With MIDS | Without MIDS | With MIDS |
| Colon | 0.45(40) | 0.87(39) | 0.39(9) | 0.96(50) | 0.61(5) | **0.99(50)** |
| Leukemia | 0.39(43) | **0.70(14)** | 0.38(32) | 0.53(34) | 0.56(18) | 0.69(26) |
| Ovarain | 0.80(20) | 0.79(12) | 0.70(9) | 0.79(9) | 0.79(4) | **1.00(3)** |
| Lung | 0.44(19) | 0.79(16) | 0.54(16) | 0.60(20) | 0.68(6) | **0.83(3)** |
| Lymphoma | 0.38(27) | **0.60(9)** | 0.47(24) | 0.52(24) | 0.29(36) | 0.53(11) |
| Pancreatic | 0.26(50) | 0.57(43) | 0.33(15) | 0.55(29) | 0.60(9) | **0.80(4)** |
| SRBCT | 0.59(48) | 0.58(50) | 0.62(40) | 0.64(27) | **0.71(28)** | 0.68(24) |
| **Average** | 0.473(35.3) | 0.700(24.7) | 0.490(20.7) | 0.656(27.6) | 0.605(15) | **0.789(17)** |

Table 4: Performance comparison of stability ($Sim_{ID}$) of different feature selection algorithms with or without MIDS on different data sets. The stability for feature subsets with the best classification accuracy is reported for each data set. The best results are highlighted in bold.

for data set with highly correlated feature groups, such as colon cancer, MIDS framework can improve the classification accuracy and stability of algorithms significantly.

## 5. CONCLUSION

In this paper, we propose a novel stability feature selection framework that utilizes the idea of minimal independent dominating sets. The minimum independent dominating sets selected by our framework can preserve the functionality of the original feature space. The performance of state-of-art feature selection algorithms can be improved due to the removed redundant features. Empirical study shows that the features selected in our framework are not only stable but also lead to better prediction accuracy as compared to directly applying feature selection algorithms to all input features.

## 6. ACKNOWLEDGMENTS

This work has been supported by Johnson & Johnson Pharmaceutical Research & Development, L.L.C. We want particularly thank Fred Baribaud and Shannon Telesco for their support and discussions.

## 7. REFERENCES

[1] A. Appice, M. Ceci, S. Rawles, and P. Flach. Redundant feature elimination for multi-class problems. In *Proceedings of the twenty-first international conference on Machine learning*, pages 5–13, 2004.

[2] W. Brendel and S. Todorovic. Segmentation as maximum-weight independent set. In *NIPS*, pages 307–315, 2010.

[3] X.-w. Chen and J. C. Jeong. Minimum reference set based feature selection for small sample classifications. In *Proceedings of the 24th international conference on Machine learning*, pages 153–160, 2007.

[4] C. A. Davis, F. Gerick, V. Hintermair, C. C. Friedel, K. Fundel, R. Küffner, and R. Zimmer. Reliable gene signatures for microarray classification: assessment of stability and performance. *Bioinformatics*, 22:2356–2363, 2006.

[5] M. R. Garey and D. S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA, 1990.

[6] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.

[7] Z. He and W. Yu. Review article: Stable feature selection for biomarker discovery. *Comput. Biol. Chem.*, 34:215–225, 2010.

[8] ILOG, Inc. ILOG CPLEX: High-performance software for mathematical programming and optimization, 2006. http://www.ilog.com/products/cplex/.

[9] A. Kalousis, J. Prados, and M. Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowl. Inf. Syst.*, 12:95–116, 2007.

[10] 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*, pages 2429–2437, 2004.

[11] S. Loscalzo, L. Yu, and C. Ding. Consensus group stable feature selection. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 567–576, 2009.

[12] T. Ma and L. J. Latecki. Maximum weight cliques with mutex constraints for video object segmentation. In *CVPR*, pages 670–677, 2012.

[13] Q. mcnemar. *Psychological statistics*. New York: Wiley., New York, NY, USA, 1949.

[14] N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B*, 72:417–473, 2010.

[15] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *TPAMI*, 27:1226–1238, 2005.

[16] Y. Saeys, T. Abeel, and Y. Peer. Robust feature selection using ensemble feature selection techniques. In *Proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases - Part II*, pages 313–325, 2008.

[17] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.

[18] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts,

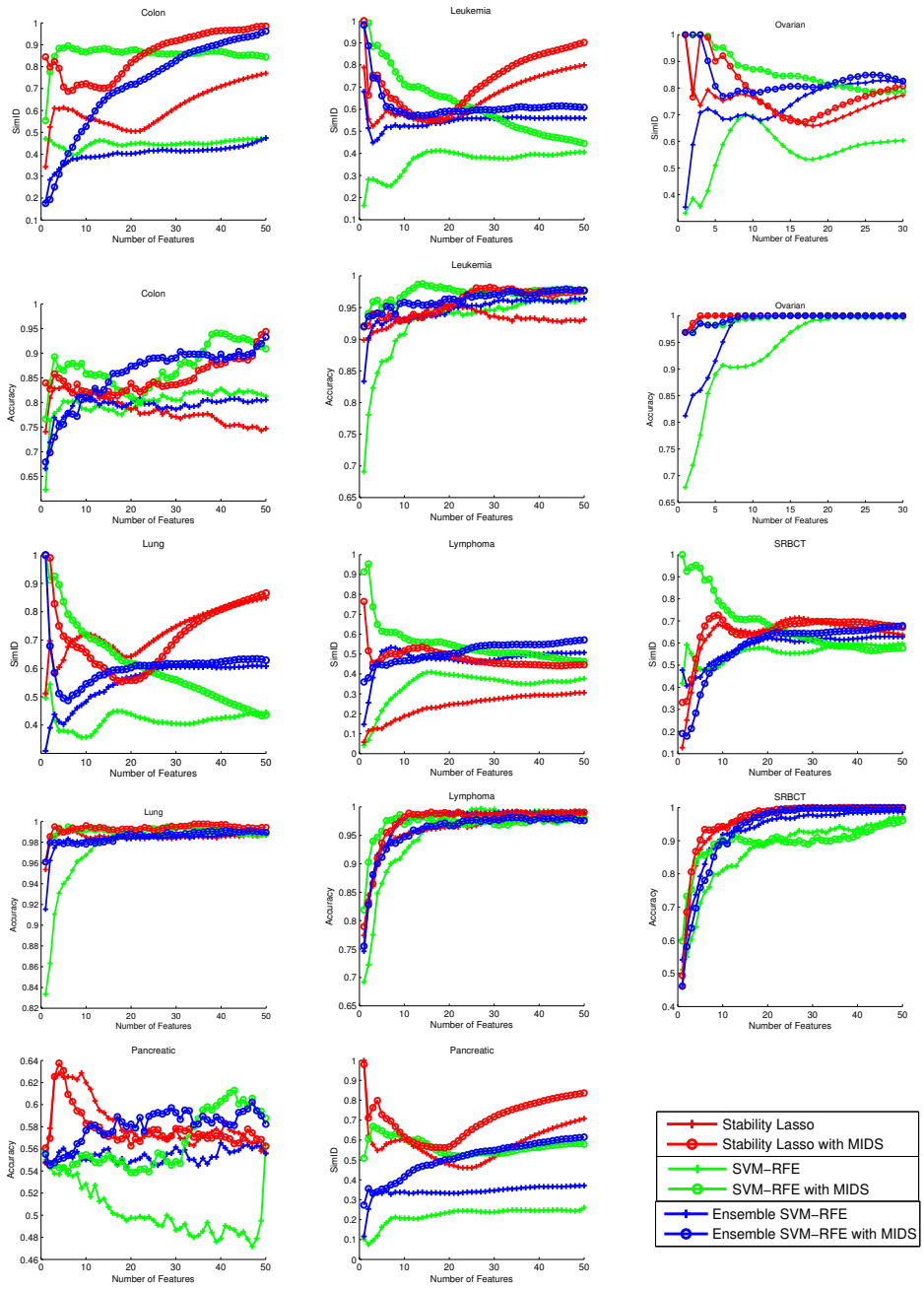


Figure 3: Comparison of three feature selection algorithms with or without MIDS on Colon, Leukemia, Lung, Overain,Lymphoma,SRBCT,Pancreatic. Figures with '$Sim_{ID}$' as y-axis show the stability of the selected representative features. Figures with 'Accuracy' as y-axis show the SVM classification accuracy for various numbers of selected features.

M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer-van Gelder, J. Yu, T. Jatkoe, E. M. Berns, D. Atkins, and J. A. Foekens. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365:671–679, 2005.

[19] J. L. L. L. Yijun Sun, Steve Goodison and W. G. Farmerie. Improved breast cancer prognosis through the combination of clinical and genetic markers. *Bioinformatics*, 23:30–37, 2007.

[20] L. Yu, C. Ding, and S. Loscalzo. Stable feature selection via dense feature groups. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 803–811, 2008.

[21] L. Yu and H. Liu. Efficient feature selection via analysis of relevance and redundancy. *J. Mach. Learn. Res.*, 5:1205–1224, 2004.