# Feature Selection from Microarray Data via an Ordered Search with Projected Margin

**Saulo Moraes Villela, Saul de Castro Leite, Raul Fonseca Neto**
Computer Science Department, Federal University of Juiz de Fora
Juiz de Fora, Minas Gerais, Brazil
{saulo.moraes, saul.leite, raulfonseca.neto}@ufjf.edu.br

## Abstract

Microarray experiments are capable of measuring the expression level of thousands of genes simultaneously. Dealing with this enormous amount of information requires complex computation. Support Vector Machines (SVM) have been widely used with great efficiency to solve classification problems that have high dimension. In this sense, it is plausible to develop new feature selection strategies for microarray data that are associated with this type of classifier. Therefore, we propose, in this paper, a new method for feature selection based on an ordered search process to explore the space of possible subsets. The algorithm, called Admissible Ordered Search (AOS), uses as evaluation function the margin values estimated for each hypothesis by a SVM classifier. An important theoretical contribution of this paper is the development of the projected margin concept. This value is computed as the margin vector projection on a lower dimensional subspace and is used as an upper bound for the current value of the hypothesis in the search process. This enables great economy in runtime and consequently efficiency in the search process as a whole. The algorithm was tested using five different microarray data sets yielding superior results when compared to three representative feature selection methods.

## 1 Introduction

Microarray hybridization experiments can simultaneously measure the expression level of thousands of genes. This technology has become an important tool for understanding gene function since it enables researchers to observe and compare the behavior of genes in a genome scale. Microarray data analyses often means dealing with few samples in a very high dimensional space. Therefore, for discrimination studies, it is crucial that we have some way of selecting differently expressed genes or features.

The main purpose of feature selection methods is to eliminate irrelevant features in order to produce relevant subsets that are able to achieve a better generalization on classification tasks with a significant number of variables [Ng, 1998]. In fact, there is an expectation that the generalization error decreases as the dimension of the problem decreases, since the VC-dimension of the classifier is reduced. In this sense, it is possible to find, for each dimension, a classifier with a best subset of variables that generates a better generalization, establishing a better compromise between the expected error and the classifier capacity. This subset represents a subset that shows a larger value of margin considering a set of constraints. Feature selection methods search through subsets of $n$ features and try to obtain the best one of the $2^n$ possible candidate subsets according to some evaluation criterion.

Numerous feature selection methods based in SVM have been proposed for classification tasks in literature [Bradley and Mangasarian, 1998], [Weston *et al.*, 2000], [Guyon *et al.*, 2002]. In general, methods that use SVM can be classified in two categories, named wrapper and embedded methods. As example of wrapper method, we can cite the Recursive Feature Elimination (RFE) algorithm [Guyon *et al.*, 2002]. This method is successfully used in microarray data analysis such as gene selection. Unfortunately, the recursive elimination of one variable at a time is not sufficient to detect the most promising subset. Almost always, the algorithm finds suboptimal solutions instead of the optimal one.

The embedded methods use large margin classifiers with a regularization method that restricts or shrinks the magnitude of the features. These methods generate sparse solutions and have been employed as an alternative to methods that use greedy strategies or explore the search space of feature subsets. [Weston *et al.*, 2003] introduced a $L_0$ formulation for SVM which minimizes, in an approximate way, the number of non-zero variables or components of the normal vector. The algorithm, at each iteration and after a SVM training, rescales the components of the input vectors. A variant method consists of introducing a set of scaling factors that are used for tuning the variables relevance [Weston *et al.*, 2000]. In this sense, the scaling parameters are adjusted by model selection or by direct optimization [Vishwanathan *et al.*, 2010]. The main drawback of these methods is in the adjustment of the scaling factors and in solving the optimization problem that is more difficult than a standard SVM.

In this sense, we propose the development of a new wrapper method, the AOS algorithm. This method uses an ordered backward search process based on margin values, allowing the discovery of feature subsets with superior generalization

capacity. Different from a greedy strategy, that irrevocably removes one or a group of variables at time and obtains nested subsets of features in a suboptimal way, the AOS algorithm explores the space of candidates with a beam search. A beam search reduces the space of candidates, avoiding the combinatorial explosion but preserving the chances of finding the best constrained subset [Gupta *et al.*, 2002]. This non-monotonic strategy allows the identification of the correlation among the set of variables that could be lost by greedy strategies, such as the RFE algorithm. In order to avoid the exploration of an exponential number of subsets, the AOS algorithm employs a set of procedures related to branching factors and pruning mechanisms. To the best of our knowledge, this is the first wrapper method based on SVM that uses the margin value as evaluation measure to control the search in subset space.

The AOS algorithm was tested on five microarray problems and the results were compared to three representative feature selection methods applied to microarray data analysis. The first is a variable ranking or filter method based on Golub criteria [Golub *et al.*, 1999]. The second is a wrapper method known as RFE-SVM [Guyon *et al.*, 2002] and the last one is an embedded statistical method known as Nearest Shrunken Centroids (NSC) [Tibshirani *et al.*, 2003]. Although in this paper the AOS method has been used only on microarray data sets, your usage can be extended to any feature selection classification problem.

The remainder of this paper is organized as follows. Section 2 briefly describes some preliminary concepts related respectively to the problem of binary classification and to the theoretical basis of Support Vector Machines. Section 3 addresses the feature selection problem and exposes the three most usual approaches to solve feature selection problems: filter, embedded and wrapper methods and its related algorithms. Section 4 presents the AOS algorithm including the branching and pruning strategies. In section 5 all results and computational experiments are presented and, finally, in section 6, some considerations about the work are reported.

# 2 Classification

## 2.1 The Binary Classification Problem

Let $Z = \{z_i = (x_i, y_i) : i \in \{1, \cdots, m\}\}$ be a training set composed of points $x_i \in \mathbb{R}^d$ and labels $y_i \in \{-1, 1\}$. In addition, let $Z^+$ and $Z^-$ be defined as the sets $\{(x_i, y_i) \in Z : y_i = +1\}$ and $\{(x_i, y_i) \in Z : y_i = -1\}$, respectively. A binary classification problem consists of finding a hyperplane, which is given by its normal vector $w \in \mathbb{R}^d$ and a constant $b \in \mathbb{R}$, such that the points in $Z^+$ and $Z^-$ lie separated in the two half spaces generated by it. That is, we look for $(w, b)$ such that:

$$y_i (w \cdot x_i + b) \geq 0, \text{ for all } (x_i, y_i) \in Z.$$

Clearly, this hyperplane may not exist for some training sets $Z$. When it exists, $Z$ is usually called linearly separable. We suppose that $Z$ is linearly separable, otherwise it will be linearly separable in a projected space with higher dimension, commonly called feature space, represented by $F$.

We say that $Z$ accepts a margin $\gamma \geq 0$ when there is a hyperplane $\mathcal{H} := \left\{x \in \mathbb{R}^d : w \cdot x + b = 0\right\}$ such that:

$$y_i (w \cdot x_i + b) \geq \gamma, \text{ for all } (x_i, y_i) \in Z.$$

In this case, we define two additional hyperplanes parallel to $\mathcal{H}$, given by $\mathcal{H}^+ := \left\{x \in \mathbb{R}^d : w \cdot x + (b - \gamma) = 0\right\}$ and $\mathcal{H}^- := \left\{x \in \mathbb{R}^d : w \cdot x + (b + \gamma) = 0\right\}$. The distance between these two parallel hyperplanes is given by:

$$\text{dist}\left(\mathcal{H}^-, \mathcal{H}^+\right) = \frac{-(b - \gamma) + (b + \gamma)}{||w||} = \frac{2\gamma}{||w||}.$$

Let $\gamma_g := \text{dist}\left(\mathcal{H}^-, \mathcal{H}^+\right)/2$, we call this $\gamma_g$ the geometric margin between the two hyperplanes $\mathcal{H}^+$ and $\mathcal{H}^-$. This way, we say that $Z$ accepts a geometric margin $\gamma_g \geq 0$ when there exists a hyperplane with $(w, b)$ such that:

$$y_i (w \cdot x_i + b) \geq \gamma_g ||w||, \text{ for all } (x_i, y_i) \in Z.$$

## 2.2 Support Vector Machines – SVM

SVM are maximal margin classifiers which were first introduced by [Boser *et al.*, 1992]. This technique aims to separate the training set by a hyperplane that maximizes the distance from members of opposite classes. In order to obtain the maximal margin hyperplane that correctly classifies all patterns in the training set, it is necessary to solve the following optimization problem:

$$\max_{(w,b)} \left( \min_i \frac{y_i (w \cdot x_i + b)}{||w||} \right)$$
$$\text{s.t. } y_i (w \cdot x_i + b) > 0, \text{ for all } (x_i, y_i) \in Z,$$

which can also be written as the equivalent problem:

$$\max \gamma_g$$
$$\text{s.t. } y_i (w \cdot x_i + b) \geq ||w|| \gamma_g, \text{ for all } (x_i, y_i) \in Z.$$

Defining $\gamma_g ||w|| = 1$ as the value of the minimal functional margin, [Vapnik, 1995] derived the classic SVM formulation that minimizes Euclidean norm of the vector:

$$\min \frac{1}{2} ||w||^2$$
$$\text{s.t. } y_i (w \cdot x_i + b) \geq 1, \text{ for all } (x_i, y_i) \in Z.$$

In order to facilitate the solution of this problem, it is convenient to relax the inequality constraints introducing a set of nonnegative Lagrangean multipliers $\alpha_i$, where $i \in \{1, \cdots, m\}$. Incorporating the relaxed constraints, we obtain the Lagrangean function:

$$L(w, b, \alpha) = \frac{1}{2} ||w||^2 - \sum_i \alpha_i y_i (w \cdot x_i + b) + \sum_i \alpha_i.$$

This function needs to be minimized with respect to $w$ and $b$, and maximized with respect to $\alpha$, subject to $\alpha_i \geq 0, \forall i \in \{1, \cdots, m\}$. This solution can be found through the maximization of a strictly dual function, where the parameters $w$ and $b$ are substituted. This particular dual formulation of the problem is called Wolfe's dual. In this way, we obtain the dual formulation of the SVM problem written only as a function of the multipliers $\alpha$:

$$\max L(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$
$$\text{s.t.} \begin{cases} \sum_i \alpha_i y_i = 0 \\ \alpha_i \geq 0, \forall i \in \{1, \cdots, m\}. \end{cases}$$

Solving this problem and obtaining the values $\alpha^*$, we can construct the normal vector solution:

$$w^* = \sum_i \alpha_i^* y_i x_i, \forall i \in \{1, \cdots, m\}.$$

Support Vector Machines in their strict dual formulation are a quadratic optimization problem which requires an optimization solver for batch solutions. However, SVM can also be solved through an iterative analytical process called Sequential Minimal Optimization (SMO) [Platt, 1999], which we decided to employ throughout this work.

# 3 Feature Selection

## 3.1 Feature Selection as Heuristic Search

A convenient paradigm to deal with the combinatorial aspect of the feature selection problem is the heuristic search, where each hypothesis in the search space represents a subset of features. According to [Jain and Zongker, 1997] these methods can be classified as deterministic or stochastic and optimal or suboptimal. The optimal methods seek to find a global optimal solution. Clearly, an exhaustive search in the space is unfeasible due to memory and time constraints, since there exist $2^n$ subsets, where $n$ represents the number of features of original set. The branch and bound algorithm, proposed by [Narendra and Fukunaga, 1977], can be used as alternative to an exhaustive search. This method requires a monotone evaluation function and produces a single deterministic solution. However, the worst case complexity of the algorithm is exponential.

The suboptimal methods can be divided in those that maintain just one current subset and those that maintain a set of candidates. The most important group of feature selection methods that maintain one current subset is the sequential search, which starts with one solution and iteratively adds or removes one feature until a stop criterion is achieved. These methods can be classified into two groups: forward or backward elimination. The RFE algorithm [Guyon *et al.*, 2002] is a classical example of sequential backward elimination method.

On the other hand, there are those methods that maintain a set of candidates. The main representative method of this group is the best-first search as well as its restricted version called beam search, which preserves a "beam" or promising set of hypothesis in order to maintain the tractability of the problem in large search spaces. These methods maintain a queue of candidate solutions. Also, we have the stochastic methods, where the most representatives are the genetic algorithms, introduced by [Siedlecki and Sklansky, 1989].

According to [Blum and Langley, 1997], there may be three types of interaction, defining three classes of methods: those that embed the selection with the induction algorithm, those that use selection to filter features passed to induction and those that consider the feature selection process as a wrapper around the induction process.

## 3.2 Embedded Methods

The strategies of these methods are based on the fact that the inductor algorithm can internally promote its own choice of relevant variables. Thus, the process of variable selection is part of the training process as a whole. According to [Guyon and Elisseeff, 2003], these methods can be more efficient than wrapper methods in two important aspects: it does not need to divide the training data in two subsets, training and validation, and it reaches a solution faster by avoiding retraining the predictor.

Nearest Shrunken Centroid (NSC) is a statistical method that uses denoised version of the centroids as prototypes for each class for classification tasks. It was proposed by [Tibshirani *et al.*, 2002] and applied on multiple cancer types diagnostic with the discriminate analysis of microarray data sets. Nearest centroid classification takes a new sample and compares to the Euclidean distance of each class centroids. The class whose centroid is closest is the predicted class for this new sample. NSC makes one important modification compared to unsrhunken nearest centroid classification. It shrinks each class centroid toward the overall centroid for all classes by a threshold $\Delta$ before making the prediction. This amount can be determined by cross-validation technique.

This shrinkage has two advantages: it can make the classifier more accurate by reducing the effect of noisy attributes and it makes automatic feature selection. A feature is eliminated if the same is shrunk to zero for all classes.

## 3.3 Filter Methods and Statistical Scores

These methods introduce a separate process in feature selection that occurs before the induction process and is independent of the chosen predictor. Hence, [John *et al.*, 1994] named then as filter methods, since they filter the irrelevant attributes. The preprocessing step can be used to reduce the space dimensionality and overcome overfitting [Guyon and Elisseeff, 2003]. Filter methods generally establish a ranking of variables based on a statistical score. The most popular is the criterion proposed by [Golub *et al.*, 1999] also called signal-to-noise. This measure expresses the difference between the expected values of one variable of two classes divided by the sum of their standard deviation. The greater this value, the greater is the importance of the associated variable. If we define $\mu_1$ and $\mu_2$ as the mean of the two classes, and $\sigma_1$ and $\sigma_2$ as their standard deviation, we can define the Golub's $G$ score as follows:

$$G = \frac{|\mu_1 - \mu_2|}{\sigma_1 + \sigma_2}$$

The major drawback of this method is that the score values are computed for each feature individually without taking into account the interdependence between them. Also, this approach can produce better results for one predictor and worse for the others.

However, also exists a different filter approach to feature selection that uses a correlation heuristic to evaluate the merit of feature subsets, named Correlation-based Feature Selection (CFS), coupled with a heuristic search. CFS starts from the empty set of features and employs a forward best-first search to reach the best subset [Hall and Smith, 1999]. This final subset of features is not redundant among them and the same time correlated with the class to be predicted.

### 3.4 Wrapper Methods

In contrast to the approach of the filter methods, wrapper methods are based on the fact that the inductors can generate important information regarding the relevance of the set of features. Therefore, it makes use of the inductor algorithm for evaluate the candidate subsets. In general, this method can be expensive because the inductor algorithm is executed for each candidate subset that needs to be evaluated.

In order to implement this method, a wide range of search strategies can be used, as we have seen before. These strategies include greedy, best-first, branch-and-bound, genetic algorithm and others. The validation of subset candidates can be done by using a validation set or a model evaluation such as $k$-fold cross-validation test.

An example of wrapper method is the RFE that uses as inductor the SVM algorithm [Guyon *et al.*, 2002]. The basic idea of this method is the recursive elimination of features associated with the smallest component of the normal vector, since they do not have much influence on the positioning of the hyperplane. At each step of the process, a fixed number of features are eliminated and the SVM classifier is retrained. In [Guyon *et al.*, 2002] is mentioned that the recursive elimination of one feature at a time generates nested classifiers with less expected error.

## 4 Admissible Ordered Search – AOS

### 4.1 State Space and Heuristic Search

The AOS algorithm is a wrapper method that uses as predictor a large margin classifier, such as SVM, and explores the subsets space with an ordered beam search that we simply call best-first. It finds for each dimension the single classifier that has the largest margin. The feature subset of this classifier is a minimal set according to the criterion of the margin maximization. Therefore, there is no alternative subset of the same cardinality able to provide a classifier with a larger margin.

In a process of ordered search, we ensure the admissibility of the algorithm if the evaluation function is monotone [Hart *et al.*, 1968]. For maximization problems this function needs to be monotone decreasing. Therefore, since we are looking for maximizing the margin, we consider as merit of each candidate hypothesis the final value of the margin obtained by the classifier associated with the respective subset of variables. The admissibility of the process is preserved by the fact that the margin values are always decreasing when the dimension is reduced. Let $\gamma_{g_j}^{d-1}$ represents the real value of the maximal or largest margin associated with the child hypothesis that excludes the $j^{th}$ variable. Let $\gamma_g^d$ represents the real value of the maximal margin associated with the father's hypothesis. Then, in a lower dimension space with $d-1$ variables we have: $\gamma_{g_j}^{d-1} \leq \gamma_g^d, \forall j$.

The control strategy of the best-first algorithm is implemented with the insertion of candidate hypotheses, called states, in a priority queue structure, often called list of open states, ranked by the value of margins. Since the order of the selected features of a hypothesis does not matter, there will be some redundancy. This redundancy that generates replicated hypothesis is prevented by the use of a hash table.

### 4.2 Projected Margin

As stated before, we use the margin's value as the evaluation criteria. However, if we use the real value of the margin we have to solve a margin maximization problem for each generated hypothesis. Instead, we used an upper bound or an optimistic estimate of this margin that maintain the admissibility. This value is the projected margin, i.e., the value obtained by performing the projection of the margin using as direction the normal vector in a lower dimension space associated with the elimination of one feature. In this sense, if we insert in the priority queue a state with projected margin value we can, if appropriate, prune the same, without affecting the admissibility of the process.

First, we define the geometric margin in $\mathbb{R}^d$ space as:

$$\gamma_g^d = \gamma_g^d \frac{w}{||w||}.$$

Let $\gamma_{p_j}^{d-1}$ denotes the projected margin in $\mathbb{R}^{d-1}$ space of a candidate hypothesis with $d$-1 features considering that the $j^{th}$ feature is excluded. This value is computed as follows:

$$\gamma_{p_j}^{d-1} = \frac{\gamma_g^d}{||w||} \left( \sum_{k \neq j} w_k^2 \right)^{\frac{1}{2}}.$$

We can easily observe that if the $j^{th}$ component of the normal vector is zero the projected margin in $\mathbb{R}^{d-1}$ space will have the same value of the superior margin associated with the father hypothesis in $\mathbb{R}^d$ space. Therefore, we maintain the admissibility of the process, ensuring that:

$$\gamma_{g_j}^{d-1} \leq \gamma_{p_j}^{d-1} \leq \gamma_g^d, \forall j.$$

Hence, we can use the projected margin as an upper limit for the real geometric maximal margin of the hypothesis in the associated space. This use can be considered as an admissible heuristic for the evaluation function of the states, since its value will be always an optimistic estimate against the real value. For each iteration of the algorithm, the hypothesis related to the current largest margin value is selected from priority queue, regardless of its dimension, to be expanded and generate new hypothesis in a space of one less dimension. This way, we can verify two possible situations:

First: for the case where the value of the margin is the projected margin, we calculate its real value through the solution of a margin maximization problem and compare it to the second highest value of the priority queue. If it is still the best option, we remove it from the queue, and generate its child hypothesis. Otherwise we replace the projected margin value by the real value computed and reinsert it into the queue.

Second: for the case where the margin value of the chosen hypothesis is already the real value, we remove this state and generate its child hypothesis.

### 4.3 Branching Factor

In order to select which hypothesis would be generated, we adopted the elimination of the smallest components of the normal vector as used in RFE-SVM [Guyon *et al.*, 2002]. However, instead of eliminating just one feature, generating

only one nested hypothesis, we select two or three smallest features to eliminate, obtaining the respective number of child hypotheses. Therefore, using a branching factor higher than one allows the algorithm to explore the variables interdependence. It is possible to verify that, removing only the smallest component related to the least projected margin will not always result in the best geometric margin for a problem with dimension reduced to $d-1$.

We can observe that the magnitude of the normal vector components, $|w_j|$, has a direct relation with the projected margin value in a space where the $j^{th}$ feature is removed. Therefore, if we choose a variable related to a normal vector component with zero value, we have the same value between the projected margin and the real maximum margin in the associated subspace. In this case, it is not necessary to solve the margin maximization problem.

### 4.4 Pruning Strategies

In order to control the combinatorial explosion, we introduce an adaptive heuristic pruning scheme based on a greedy strategy that uses the RFE algorithm as a deep-first search to generate a lower bound for margins values at determined pruning depth. Therefore, every time we close the first state of each dimension, we recompute the value of the margin in a lower dimension in order to define a new lower bound that is used to eliminate candidate hypothesis that has inferior projected or real margin values.

In addition, as second scheme, we keep in the priority queue only a current subset of promising hypotheses. Therefore, when a new lower bound is generated, we remove states related to the hypothesis that have a dimension value higher than the last dimension reached plus a cut depth parameter.

When we use these strategies and a reduced branching factor, the admissible property, in theory, is lost. However, we gain in practice much computational speed. This is necessary to reduce the search space and deal with combinatorial explosion avoiding an exponential search [Gupta *et al.*, 2002]. The obtained results demonstrated that the use of a higher branching factor, greater than three, does not make possible to obtain better results. If we define the pruning depth to be one, the AOS algorithm becomes equivalent to RFE strategy. Hence, when adopting this pruning schema we are implementing a flexible RFE. It is important to observe that, when we close the first hypothesis that reaches a certain dimension we find the classifier of largest margin related to this number of variables. Also, a cross-validation estimate of the expected error could be computed with this classifier. It could be used as a stop criterion of the algorithm since it represents the generalization performance of this hypothesis.

### 4.5 Pseudocode

Pseudocode 1 describes the AOS algorithm.

**Algorithm 1:** Admissible Ordered Search – AOS

**Input**: training set $Z = \{(x_i, y_i)\} : i \in \{1, \ldots, m\}$;
features set $F = \{1, 2, 3, \ldots, d\}$;
branching factor $b$;
pruning depth $p$;
cut depth $c$;
stop level $s$;
**Output**: last opened state;

```
begin
    initiate heap H and hash table HT;
    compute the solution using SVM for the initial state
    S_init with feature set F;
    level ← d;
    insert S_init into H;
    while level > s and H is not empty do
        get the best hypothesis S from H;
        if S only has a projected margin then
            compute the real margin for S using SVM;
            if solution using SVM has converged then
                insert S into H;
            end if
        else
            if dimension of state S is equal to level then
                use RFE to depth level-p and find the
                new value for lower;
                cut from H every state with margin value
                less then lower;
                cut from H every state with level more
                then level+c;
                level ← level − 1;
            end if
            for i ← 1 to b do
                create new state S' with feature set
                F' = F − {f_i};
                compute γ_pj for S';
                if γ_pj > lower and F' is not in HT then
                    insert F' into HT;
                    insert S' into H;
                end if
            end for
        end if
    end while
    return last opened state;
end
```

### 4.5 Pseudocode

Pseudocode 1 describes the AOS algorithm.

## 5 Experiments and Results

### 5.1 Data Sets

In order to analyze the results, we used five data sets derived from microarray experiments and were related to cancer studies. All data sets used in this paper are contained in the UCI Machine Learning Repository [Bache and Lichman, 2013], or referenced by [Golub *et al.*, 1999], [Alon *et al.*, 1999] or [Singh *et al.*, 2002], and are shown in table 1.

### 5.2 Performance Comparison on Data Sets

We compared the performance of the AOS algorithm to three different methods of feature selection. We used the filter method that uses Golub score coupled with a SVM classifier, the statistical NSC method and the wrapper monotone method RFE. Also, we report the results about validation and test errors and the margin values using the SMO algorithm without variables elimination, to compare with original data.

Table 1: Data sets information.

| Set | Features | Samples | | |
|---|---|---|---|---|
| | | Pos. | Neg. | Total |
| Colon | 2000 | 22 | 40 | 62 |
| Leukemia | 7129 | 47 | 25 | 72 |
| Prostate | 12600 | 50 | 52 | 102 |
| Breast | 12625 | 10 | 14 | 24 |
| DLBCL | 5468 | 58 | 19 | 77 |

For Golub's method we eliminated one feature at a time and the process was interrupted when we reach a non separable instance of data set. Hence, the margin value associated with the algorithm is the value produced by the SMO solution with the referred subset of features.

For NSC method we used the 'pamr' package, the method's version for the $R$ language, referenced by [Tibshirani *et al.*, 2003]. For the $\Delta$ threshold, we used a value that finds the minimal subset of features without affect the validation and test error. These choices represent a trade-off between the capacity and the complexity of the classifier. If we overly shrink the centroids, we can produce hypotheses that do not correctly classify the data.

For AOS and RFE methods we used as classifier the SMO algorithm. For RFE we proceed in a similar fashion of Golub's method eliminating one feature at a time. For AOS tests, we set the pruning and cut depths to 5 and the branching factor to 3. As stop criterion for AOS and RFE, we use the same procedure adopted in Golub's method. That is, we interrupted the process when we reached a nonlinear instance.

As microarray data sets have higher dimensions, we adopted for AOS analysis the RFE until it reaches the dimension 100. This is necessary to make the process of feature selection feasible. Thenceforward, we use the search algorithm to explore the $2^{100}$ possible subsets. In order to validate this strategy we conduct an experiment with the Colon dataset that has 2000 attributes. The obtained results were the same as we restrict the dimension to 100, demonstrating that the feature selection problem has a monotone nature when the dimension is higher. In this situation it is not necessary to explore the variable correlation to reach best constrained subsets.

In order to test the algorithms on the data sets and validate their results, we divided the data sets in 2 subsets, 2/3 for the feature selection process and 1/3 for estimating the prediction or test error. In order to estimate the validation error related to the training set we used the $k$-fold cross-validation [Kohavi, 1995]. In order to estimate the final validation error, we maintain the percentage of data points of each class and compute the mean error from 10 executions of a 10-fold cross-validation. For more accurate comparisons we selected, for each base, always the same training and test sets and always the same 10 subsets for cross-validations, preserving the generating seed associated to the randomness process.

Table 2 shows the results related to performance comparison of AOS algorithm. The results demonstrate that AOS algorithm had a superior performance in almost all data sets reaching subsets of features with inferior cardinality and good power of generalization. In some experiments the AOS algorithm finds final subsets with very lower cardinality as solution. These solutions naturally can present validation and test errors with higher values compared to greedy strategies. In order to deal with this problem, in Leukemia dataset we made performance comparison between the AOS and RFE algorithms for the same number of features. In this modified experiment, the AOS algorithm reaches a higher margin value (2362.388), a lower 10-fold error (0.00%) and a same validation error (4.17%).

Table 2: Comparison on data sets.

| Set | Algorithm | $F$ | $\gamma$ | 10-fold | Test |
|---|---|---|---|---|---|
| Colon | SVM | 2000 | 1529.516 | 16.55% | 19.05% |
| | Golub | 9 | 59.396 | 19.30% | 19.05% |
| | NSC | 7 | — | 21.95% | 19.05% |
| | RFE | 4 | 44.748 | 11.95% | 14.29% |
| | AOS | 4 | 105.900 | 5.65% | 14.29% |
| Leukemia | SVM | 7129 | 18181.530 | 0.00% | 8.33% |
| | Golub | 5 | 97.609 | 8.60% | 12.50% |
| | NSC | 4 | — | 6.25% | 12.50% |
| | RFE | 3 | 1398.465 | 4.05% | 4.17% |
| | AOS | 2 | 647.368 | 2.05% | 8.33% |
| Prostate | SVM | 12600 | 529.011 | 14.82% | 5.88% |
| | Golub | 15 | 16.143 | 17.43% | 8.82% |
| | NSC | 3 | — | 13.24% | 5.88% |
| | RFE | 5 | 2.648 | 8.62% | 5.88% |
| | AOS | 4 | 18.912 | 2.69% | 2.94% |
| Breast | SVM | 12625 | 5222.205 | 26.50% | 12.50% |
| | Golub | 2 | 191.930 | 8.50% | 12.50% |
| | NSC | 2 | — | 43.75% | 50.00% |
| | RFE | 2 | 394.356 | 0.00% | 12.50% |
| | AOS | 2 | 413.664 | 0.00% | 12.50% |
| DLBCL | SVM | 5468 | 14688.808 | 2.93% | 7.69% |
| | Golub | 11 | 479.156 | 14.43% | 19.23% |
| | NSC | 10 | — | 15.68% | 15.38% |
| | RFE | 4 | 1323.384 | 2.70% | 23.08% |
| | AOS | 2 | 269.738 | 0.83% | 7.69% |

## 6 Discussion

The feature selection problem has no trivial solution. For each type of problem one method seems to be more adequate than the others. Consequently, it is possible that does not exist consensus in this issue, in order to affirm that one method is better than the other. Nevertheless, we proposed in this work a novel and promising wrapper method that uses a large margin classifier as predictor. As a compromise between greedy methods and exhaustive search we use an ordered beam search algorithm that allows the exploration of subsets of variables in the combinatorial space of the feature selection problem.

Despite some good results, the RFE-SVM method remains as a greedy sub-optimal method generating nested subsets and likely missing the best correlation among variables. Feature ranking or filter methods are proposed as selection mechanisms because of its simplicity and scalability. However, the major drawback is that these methods evaluate each variable in independent form. In this sense, we emphasize the words of [Guyon and Elisseeff, 2003]: "A variable that is completely useless by itself can provide a significant performance improvement when taken with others".

The results obtained by the AOS algorithm were very satisfactory, both in obtaining subsets of inferior cardinality or

in quality of generalization. In this way, it is important to remember that the algorithm is able to find the maximal margin classifier at each dimension of the problem. Finally, we would like to mention that the algorithm employs techniques from Heuristic Search and Machine Learning.

## Acknowledgment

The authors would like to thank FAPEMIG for the support.

## References

[Alon *et al.*, 1999] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*, 96(12):6745–6750, 1999.

[Bache and Lichman, 2013] K. Bache and M. Lichman. UCI machine learning repository, 2013.

[Blum and Langley, 1997] Avrim L. Blum and Pat Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2):245–271, 1997.

[Boser *et al.*, 1992] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pages 144–152, New York, NY, USA, 1992. ACM Press.

[Bradley and Mangasarian, 1998] Paul S. Bradley and O. L. Mangasarian. Feature selection via concave minimization and support vector machines. In *Proceedings of the 15th International Conference on Machine Learning*, pages 82–90. Morgan Kaufmann, 1998.

[Golub *et al.*, 1999] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

[Gupta *et al.*, 2002] Puneet Gupta, David S. Doermann, and Daniel DeMenthon. Beam search for feature selection in automatic svm defect classification. In *ICPR (2)*, pages 212–215, 2002.

[Guyon and Elisseeff, 2003] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

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

[Hall and Smith, 1999] Mark A. Hall and Lloyd A. Smith. Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper. In Amruth N. Kumar and Ingrid Russell, editors, *FLAIRS Conference*, pages 235–239. AAAI Press, 1999.

[Hart *et al.*, 1968] Peter Hart, Nils Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968.

[Jain and Zongker, 1997] Anil Jain and Douglas Zongker. Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):153–158, 1997.

[John *et al.*, 1994] G. H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In *Proceedings of the 11th International Conference on Machine learning*, pages 121–129, New Brunswick, NJ, 1994.

[Kohavi, 1995] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th international joint conference on Artificial intelligence*, volume 2, pages 1137–1143, San Francisco, CA, 1995. Morgan Kaufmann Publishers Inc.

[Narendra and Fukunaga, 1977] P. Narendra and K. Fukunaga. A branch and bound algorithm for feature subset selection. *IEEE Trans. Comput.*, 26(9):917–922, 1977.

[Ng, 1998] Andrew Y. Ng. On feature selection: Learning with exponentially many irrelevant features as training examples. In *Proceedings of the 15th International Conference on Machine Learning*, pages 404–412. Morgan Kaufmann, 1998.

[Platt, 1999] John C. Platt. *Fast training of support vector machines using sequential minimal optimization*, pages 185–208. MIT Press, Cambridge, MA, 1999.

[Siedlecki and Sklansky, 1989] W. Siedlecki and J. Sklansky. A note on genetic algorithms for large-scale feature selection. *Pattern Recogn. Lett.*, 10(5), 1989.

[Singh *et al.*, 2002] Dinesh Singh, Phillip G. Febbo, Kenneth Ross, Donald G. Jackson, Judith Manola, Christine Ladd, Pablo Tamayo, Andrew A. Renshaw, Anthony V. D'Amico, and Jerome P. Richie. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2):203–209, 2002.

[Tibshirani *et al.*, 2002] Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572, 2002.

[Tibshirani *et al.*, 2003] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Class prediction by nearest shrunken centroids, with applications to dna microarrays. *Statistical Science*, 18(1):104–117, 2003.

[Vapnik, 1995] Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[Vishwanathan *et al.*, 2010] S. V. N. Vishwanathan, Nicol N. Schraudolph, Risi Kondor, and Karsten M. Borgwardt. Graph kernels. *Journal of Machine Learning Research*, 11:1201–1242, 2010.

[Weston *et al.*, 2000] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik. Feature selection

for svms. In *Advances in Neural Information Processing Systems 13*, pages 668–674. MIT Press, 2000.

[Weston *et al.*, 2003] Jason Weston, André Elisseeff, Bernhard Schölkopf, and Mike Tipping. Use of the zero norm with linear models and kernel methods. *Journal of Machine Learning Research*, 3:1439–1461, 2003.