# A Counting Technique based on SVM-RFE for Selection and Classification of Microarray data

Jose Crispin Hernandez Hernandez, Béatrice Duval, and Jin-Kao Hao

LERIA, Université d'Angers,
2 Boulevard Lavoisier, 49045 Angers, France
{josehh,bd,hao}@info.univ-angers.fr

**Abstract.** Selecting informative genes for the classification of tumor tissues from microarray data is a challenging problem in bioinformatics. Recursive Feature Elimination with Support Vector Machines (SVM-RFE) is known to be an effective method for selection and classification of microarray data. In this paper, we introduce a novel gene ranking method based on a frequency analysis in gene subsets processed by SVM-RFE. This novel ranking leads to select as informative genes the top ranked genes. The effectiveness of this method is assessed using two well-known benchmark data sets from the literature, showing highly interesting results.

**Key words:** Microarray data analysis, Support vector machines, Adatron, Recursive feature elimination

## 1 Introduction

Recent advances in DNA microarray technologies enable to consider molecular cancer diagnosis based on gene expression. Classification of tissue samples from gene expression levels aims to distinguish between normal and tumor samples, or to recognize particular kinds of tumors [9, 1]. Gene expression levels are obtained by cDNA microarrays and high density oligonucleotide chips, allowing to monitor and measure simultaneously gene expressions for thousands of genes in a sample. So, data that are currently available in this field concern a very large number of variables (thousands of gene expressions) relative to a small number of observations (typically under one hundred samples). This characteristics, known as the "curse of dimensionality", is a difficult problem for classification methods and requires special techniques to reduce the data dimensionality in order to obtain reliable predictive results.

Feature selection is one of the most popular ways for reducing data dimensionality. Feature selection aims at selecting a (small) subset of informative features (genes) from the initial data in order to obtain high classification accuracy [10, 6, 18]. A feasible approach to feature selection is to rank all the features according to their interestingness for the classification problem and to select the


Advances in Computer Science and Engineering.
Research in Computing Science 23, 2006, pp. 99-107

Received 10/07/06
Accepted 03/10/06
Final version 12/10/06

top ranked features. The feature score can be obtained independently for each feature, as it is done in [9] which relies on correlation coefficients between the class and each feature. The drawback of such a method is to score each feature independently while ignoring the relations between the features. More recently, [11] presents recursive feature elimination using support vector machines (SVM-RFE), which relies also on ranking criteria but takes into account the relations between features.

SVM-RFE is a backward feature elimination method [14] that searches among the $n$ initial features a subset of $d$ features that maximizes the performance of a SVM classifier. To achieve this, one starts with all the features and iteratively removes one feature until $d$ of them are left. The removal is based on a ranking criterion obtained from a SVM classifier trained on the current subset of features. One of the difficulties with such an approach is the choice of the appropriate number $d$ of selected features [17]. Moreover, SVM-RFE is basically a greedy method that studies nested subsets of features, since the selected subset of size $m$ is included in the previously selected subset of size $m+1$. As pointed out in [11], there is no guarantee that this search strategy leads to optimal results.

The most important motivation of this work is to better understand how the selected features occur during the iterative steps of SVM-RFE. So this paper presents a novel approach of gene selection that combines the recursive feature elimination algorithm with a counting method that allows to identify genes that have a recurrent importance at different stages of the elimination process.

The paper is organized as follows: in Section 2, we review the main characteristics of SVM, and explain how RFE ranks and selects the genes according to their importance for class discrimination. In Section 3, we describe our counting technique for gene ranking. Experimental results are presented in section 4, before the conclusion given in Section 5.

## 2 Feature ranking with SVM-RFE

### 2.1 Support Vector Machines (SVM)

The principle of a SVM classifier is to find an optimal hyperplane as a decision function in a high-dimensional space [3]. Thus, let us consider a training data set $\{x_k, y_k\} \in \Re^n \times \{-1, 1\}$ where $x_k$ are the $m$ training examples and $y_k$ the class labels. At first, the method consists in mapping $x_k$ in a high dimensional space owing to a function $\Phi$. Then, it looks for a decision function of the form:

$$f(x) = w \cdot \Phi(x) + b \tag{1}$$

and $f(x)$ is optimal in the sense that it maximizes the distance between the nearest points $\Phi(x_k)$ and the hyperplane (this distance is called margin). The class label of $x$ is then obtained by considering the sign of $f(x)$. This optimization problem can be transformed, in the case of L2 soft-margin SVM classifier (misclassified examples are quadratically penalized), in this following one:

$$\min_{w,\xi} \frac{1}{2} \|w\|^2 + C \sum_{k=1}^{m} \xi_k^2 \tag{2}$$

under the constraint $\forall k,\ y_k f(x_k) \geq 1 - \xi_k$. The solution of this problem is obtained using the Lagrangian theory and one can find that the vector $w$ is of the form:

$$w = \sum_{k=1}^{m} \alpha_k^* y_k \Phi(x_k) \tag{3}$$

where $\alpha_k^*$ is the solution of the following quadratic optimization problem:

$$\max_{\alpha} W(\alpha) = \sum_{k=1}^{m} \alpha_k - \frac{1}{2} \sum_{k=1,l=1}^{m} \alpha_k \alpha_l y_k y_l \left( K(x_k, x_l) + \frac{1}{C} \delta_{k,l} \right) \tag{4}$$

subject to $\sum_{k=1}^{m} y_k \alpha_k = 0$ and $\forall k,\ \alpha_k \geq 0$, with $\delta_{k,l}$ being the Kronecker symbol and $K(x_k, x_l) = \langle \Phi(x_k), \Phi(x_l) \rangle$ is the Gram matrix of the training examples.

The SVM is commonly trained using either mathematical programming (MP) approaches such as quadratic programming or by strategies that avoid the use of the MP techniques. The latter techniques have the advantage of being easier to implement, while providing similar levels of performance as their MP counterparts. One non-MP based approach is the kernel-Adatron (KA) algorithm [5, 13]; this algorithm iteratively updates the Lagrange multipliers, $\alpha$, associated with each example. The Adatron has been demonstrated as a useful tool for analyzing microarray data (see [4]).

### 2.2 Recursive feature elimination (RFE)

Molecular biologists and oncologists seem to be convinced that only a small subset of genes are responsible for particular biological properties, so they want to identify the most important or informative genes. Guyon et al. [11] propose an SVM-based learning method, called SVM Recursive Feature Elimination (SVM-RFE) for gene selection. The motivating idea is that the orientation of the separating hyperplane found by a linear SVM can be used to select informative features: if the plane is orthogonal to a particular gene dimension, then that gene is informative, and vice versa. Specially, given a SVM with weight vector $w = \sum_{k=1}^{m} \alpha_k y_k x_k$, the ranking criterion for gene $i$ is $c(i) = (w_i)^2$. Thus, one starts with all the genes; at each step of the RFE procedure, a classifier is trained on the given sample set, which determines a ranking value for each gene; then the gene with the smallest value is discarded and the process is iterated. RFE thus removes one feature at a time until a single gene is left.

The computational cost of RFE is a function of the number of features, because a classifier must be trained each time a feature is removed. The removal of chunks of features at each loop represents a feasible approach. Another alternative is to eliminate a chunk of least important features per step until a small number of features is left, from which point, one feature is removed per step [6, 7].

SVM-RFE is a powerful method for selection and classification, but the reasons of its effectiveness are not completely known yet. Several authors have studied this technique to better understand its properties. In [15], the authors

present experiments of recursive and non recursive selection procedures applied with different types of classifier in order to see whether the success depends on the specific classifier or on the recursive procedure.

## 3 A combined technique for gene ranking using SVM-RFE

As explained in the introduction, SVM-RFE works on nested subsets: from the selected subset of size $m$, the system builds a SVM classifier associated to an hyperplane determined by a vector of weights $w$. The coefficients $w_i^2$ are the ranking coefficients of the genes and the gene with the smallest coefficient is removed to obtain the subset of size $m-1$ on which this process is iteratively applied. Even if one can determine a unique feature that gives the best separation between two classes, there is no guarantee that the best pair of features for the classification must contain this feature. So the search strategy of SVM-RFE does not guarantee an optimal result.

This work proposes to analyze the different subsets of genes selected during the successive steps of SVM-RFE. More precisely, we propose a model where the ranking of genes during each step of SVM-RFE determines an importance weight for each gene (the first gene has the greatest importance weight); then we compute the sum of these weights which assigns to each gene an importance value that takes into account the role of this gene during the whole process of SVM-RFE. Finally, this counting, based on weighted occurrences, is used to establish a new ranking of the genes. This model is presented in Algorithm 1.

The algorithm implements the different steps of a recursive feature elimination. At each step, we use the kernel-Adatron algorithm to compute the coefficients of the SVM classifier and we obtain the ranking coefficients $c_i$. The feature associated to the smallest ranking value is eliminated from the subset of "surviving features".

Moreover, the ranking is used to determine the importance coefficient assigned to each feature: at each step, the feature that is top ranked receives a weight of $n$, where $n$ is the initial number of features, the second ranked gene receives a weight of $n-1$, and so on... Finally the sum of these weights records the different positions of the feature during the process and produces a novel ranking of the set of features.

## 4 Experimental results

### 4.1 Data sets

We applied our approach on two well-known data sets: the leukemia data set and The colon cancer data set. The leukemia data set consists of 72 tissue samples, each with 7129 gene expression values. The samples include 47 acute lymphoblastic leukemia (ALL) and 25 acute myeloid leukemia (AML). The original data are divided into a training set of 38 samples and a test set of 34 samples. The

1. The data set is initialized with the $m$ examples.
Each example is described by $n$ features (gene expression values)
Training samples : $X = [x_1,\ x_2,\ ...\ x_k,\ ...x_m]^T$
Class labels : $y = [y_1,\ y_2,\ ...\ y_k,\ ...y_m]^T$
2. Initialize the following variables:
Subset of surviving features : $s = [1,\ 2,\ ...\ i,\ ...n]$
Feature ranked list : $r = [\,]$
Importance of features : $Imp\,[1,\ 2,\ ...\ i,\ ...n] = 0$
3. **repeat**
   Restrict training samples to good feature indices : $X = X(:,\ s)$
   Train the classifier : $\alpha = Adatron\_train(X,\ y)$
   Compute the weight vector of dimension $length(s)$ : $w = \sum_k \alpha_k y_k x_k$
   Compute the ranking criteria : $c_i = w_i^2$, for all i in $s$
   Sort the features according to $c_i$; let $position(i)$ be the place of feature $i$ in this sorted list: 1 for the last feature associated to the smallest $c_i$ and length(s) for the best feature associated to the greatest $c_i$
   Find the feature with smallest ranking criterion: $f = argmin(c)$
   Update the importance value of all features

$$Imp(i) = \begin{cases} Imp(i) + (n - length(s)) + position(i) & : \quad i \in s \\ Imp(i) & : \quad i \in r \end{cases}$$

   Update feature ranked list : $r = [s(f),\ r]$
   Eliminate the feature with smallest ranking criterion :
   $s = s[1 : f-1,\ f+1 : length(s)]$
**until** $s = [\,]$ ;
4. Sort the $n$ features, in decreasing order, according to their importance value $Imp(i)$: the first feature is associated to the greatest value $Imp(i)$ and it is the index of the top ranked gene

**Algorithm 1**: Feature Ranking by frequency counting and SVM-RFE

data were produced from Affymetrix gene chips. The data set was first used in [9] and is available at http://www-genome.wi.mit.edu/cancer/.

The colon cancer data set contains 62 tissue samples, each with 2000 gene expression values. The tissue samples include 22 normal and 40 colon cancer cases. The data set is available at http://www.molbio.princeton.edu/colondata and was first studied in [1].

A linear normalization procedure transforming the gene expression to mean value 0 and standard deviation 1 was applied to each dataset.

### 4.2 Experiments and results

We applied our model on each data set. In each case, we obtained a ranked list of genes, from which we could select a subset of interesting genes. For a selected subset, we used a leave-one-out cross-validation (LOOCV) estimate of the misclassification rate, which means that each observation of the training set is successively omitted from the data and then classified by a classifier trained on

the remaining observations, the mean of these experiments gives the estimated misclassification rate. For the classification task, we used a SVM-classifier.

**Leukemia classification** We ran our gene ranking method on the entire data set of 72 samples. The resulting misclassification rate is 0/72 (100% of correct prediction) when we use the subset of the four top ranked genes. This result is similar to the best published results: this maximal rate of correct classification is obtained with 6 genes in [16], but only with 2 genes in [11]. Table 1 shows the four top ranked genes obtained with our method. The first two ranked genes have been reported in [6], the third has been reported in [8] and the fourth has been reported in [9].

| Ranking | Gene Accession Number | Description |
|---|---|---|
| 1 | M27891 | CST3 Cystatin C (amyloid angiopathy) |
| 2 | M19507 | MPO Myeloperoxidase |
| 3 | M26708 | PTMA Prothymosin alpha |
| 4 | M63138 | CTSD Cathepsin D (lysosomal aspartyl protease) |

**Table 1.** The 4 best-ranked genes of Leukemia data

**Colon classification** We put all the Colon cancer data into one training set of 62 samples. The LOOCV misclassification rate obtained is 1/62 when we use the subset of four top ranked genes (see table 2). The first ranked gene has been reported in [12], the second and third ranked genes have been reported in [11], and the fourth ranked gene has been reported in [7]. The accuracy of our method is equal to the best result known for this data set and reported in [11]; the unique misclassified sample is sample number 6, which is a normal tissue reported by [1]. As we will show in the next section, the misclassification rate of 1/62 can be improved by combing our frequency counting technique with an exhaustive search, leading to a 100% correct classification.

| Ranking | Gene Accession Number | Description |
|---|---|---|
| 1 | H06524 | Gelsolin Precursor, plasma (HUMAN). |
| 2 | T62947 | 60S ribosomal protein L24 (Arabidopsis thaliana) |
| 3 | H64807 | Placental folate transporter (Homo sapiens) |
| 4 | R62549 | Putative serine/threonine-protein kinase b0464.5 |

**Table 2.** The 4 best-ranked genes of colon cancer data

**More experiments on colon data set**

When we further analyzed the gene subsets processed by RFE, we observed a significant change at the stage that considers 32 genes. For the previous steps, i.e. from the step that considers 2000 genes up to the step that considers 32

genes, a particular gene, X69550, is always the best ranked one according to the $c_i$ coefficients. But in the following steps (with 32 genes and less than 32 genes), this particular gene is no longer in the first position and is eventually eliminated when only 16 genes are retained.

This fact represents thus an important change in the subsets of relevant genes and indicates that it may be useful to take into account the role of a gene along the different steps of a RFE process. This consideration goes with what our counting-based ranking method does. In order to understand what really happens from the step which considers the subset of 32 top ranked genes, we decided to have a closer look at these genes and performed a combinatorial search based on them.

Concretely, from these 32 genes, we examine all the subsets of $k$ genes with increasing values of $k$=1,2... until a 100% classification rate is reached. This experiment proves to be successful since we found that it is possible to classify the whole data set without error (0/62) with only $k$=4 genes (using a leave-one-out cross-validation estimation). Once this result is reached, the exhaustive search process is terminated.

Quantitatively, this 62/62 classification rate with only 4 genes for the colon cancer data set constitutes the best possible result and has never been reported in the literature before. The set of 4 genes leading to this result contains (T62947, J02854, T57882 and D14812). T62947 has been reported in [11], J02854 has been reported in [1], T57882 has been reported in [6], and D14812 in [7].

According to our ranking, these genes have respectively received the positions (2,6,10 and 19). This result is very interesting since it shows that the 20 top ranked genes using our ranking contain a subset of genes that enables a perfect classification of the colon data set. Notice that this is not the case if we consider the 20 top ranked genes using SVM-RFE alone. This indicates that some important information is obtained by our ranking method when observing the multiple occurrences of a gene across different selection processes.

**Comparison with SVM-RFE with correction of bias selection**

The classification rates given in the previous parts are obtained by a LOOCV process that assesses the performance of the classifiers on a set of features previously selected, as it was done in [11]. As it was later pointed out in [2], error estimation and gene selection are not independent processes because both are based on the same training set and this way to conduct LOOCV induces a bias in the results, that are too optimistic. The proper way to evaluate a feature selection method is to perform a cross validation that is external to the selection process.

Therefore we carried out a comparison between SVM-RFE and our counting method with the following evaluation process. The initial data set is splitted into two subsets, a training set and a test set. Each method (SVM-RFE and our counting method) is applied on the training set to select a subset of relevant genes and to build a classifier; then this classifier is used on the test set to estimate the performance. 50 trails are performed, with a new split of the data

into a training set and a test set each time. Table 3 summarizes the results for our counting method, for SVM-RFE and two other methods from the literature. We report the average and the standard deviation of the correct classification rate when $n$ genes are selected; we give the results for the number $n$ that gives in these experiments the best expected classification rate for our counting method and for SVM-RFE. For the colon data set, each of the 50 experiments splits the data into two groups of 31 samples, while ensuring that each group has 11 normal and 20 cancerous tissues. For the leukemia data set, the data are splitted into two groups, one of 38 samples (25 acute lymphoblastic leukemia and 13 acute myeloid leukemia) and one of 34 samples (22 acute lymphoblastic leukemia and 12 acute myeloid leukemia) for each of the 50 splits.

As it can be observed, on the two data sets, our counting method provides the best classification rate.

| | Our Counting Method | SVM-RFE | [2] | [12] |
|---|---|---|---|---|
| Leukemia $n = 64$ | 98.18% ±1.45% | 97.47% ±1.88% | ≈95% | - |
| Colon $n = 32$ | 89.36% ±0.03% | 88.71% ±0.04% | 82.5% | 88.84% |

**Table 3.** Mean and standard deviation for classification rate over 50 splits

## 5 Conclusions and future work

This paper deals with the problem of selection of relevant genes for the classification of microarray data. We propose a novel gene ranking method, based on the SVM-RFE process, that provides encouraging results. Our ranking relies on an analysis of the different subsets of gene selected by RFE and takes into account the role of a gene along the different steps of the selection process. With the top ranked genes, we obtain classification results, which is better than that SVM-RFE. Moreover, on the colon cancer data set, we show that we can extract from the 32 top ranked genes a subset of 4 genes that gives a 100% classification accuracy. This classification accuracy (with a bias selection) was never reported before in the literaure.

## References

1. 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. *Proc Natl Acad Sci USA*, 96:6745–6750, 1999.
2. C. Ambroise and G.J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci USA*, 99(10):6562–6566, 2002.

3. B. E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *Computational Learing Theory*, pages 144–152, 1992.
4. T. S. Davison, S. Mehta, I. J. Burgetz, and O. Huner. Support vector machine classification of data quality in microarray experiments. In *Critical Assessment of Microarray Data Analysis*, 2001.
5. T. Friess, N. Cristianini, and C. Campbell. The kernel adatron algorithm: a fast and simple learning procedure for support vector machine. In *15th International Conference on Machine Learning, Morgan Kaufman Publishers*, 1998.
6. L.M. Fu and C.S. Fu-Liu. Evaluation of gene importance in microarray data based upon probability of selection. *BMC Bioinformatics*, 6, 2005.
7. C. Furlanello, M. Serafini, S. Merler, and G. Jurman. Entropy-based gene ranking without selection bias for the predictive classification of microarray data. *BMC Bioinformatics*, 4, 2003.
8. G. Getz, E. Levine, and E. Domany. Coupled two-way clustering analysis of gene expression data. *Proc Natl Acad Sci USA*, 97(22):12079–12084, 2000.
9. 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*, 286:531–537, 1999.
10. I. Guyon, S. Gunn, A. Ben-Hur, and G. Dror. Result analysis of the nips 2003 feature selection challenge. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 545–552. MIT Press, 2005.
11. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.
12. T. M. Huang and V. Kecman. Gene extraction for cancer diagnosis by support vector machines - an improvement. *Artificial Intelligence in Medicine*, 35(1-2):185–194, 2005.
13. V. Kecman, M. Vogt, and T. M. Huang. On the equality of kernel adatron and sequential minimal optimization in classification and regression tasks and alike algorithms for kernel machines. In *11th European Symposium on Artificial Neural Networks ESANN*, 2003.
14. R. Kohavi and G.H. John. Wrappers for feature subset selection. *Artif. Intell.*, 97(1-2):273–324, 1997.
15. F. Li and Y. Yang. Analysis of recursive gene selection approaches from microarray data. *Bioinformatics*, 21(19):3741–3747, October 2005.
16. S. Peng, Q. Xu, X.B. Ling, X. Peng, W. Du, and L. Chen. Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines. *FEBS Lett.*, 555(2):358–362, 2003.
17. A. Rakotomamonjy. Variable selection using svm-based criteria. *Journal of Machine Learning Research*, 3:1357–1370, 2003.
18. J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik. Feature selection for SVMs. In *NIPS*, pages 668–674, 2000.