

A Support Vector Machine Classifier based on Recursive Feature Elimination for Microarray Data in Breast Cancer Characterization.

R.Campanini, D. Dongiovanni, N. Lanconelli, G. Palermo, A. Riccardi, M. Roffilli
Dipartimento di Fisica, Università degli Studi di Bologna
campanini@bo.infn.it

Abstract. An effective approach to cancer classification based upon gene expression monitoring using DNA microarray was introduced by [1]. Here they used DNA microarray analysis on primary breast tumours of 78 young patients without tumour cells in local lymph nodes at diagnosis, 34 from patients who developed distant metastasis within 5 years (poor prognosis group), 44 from patients who continued to be disease-free after a period of at least 5 years (good prognosis group) and applied a three step supervised classification (based on correlation methods) to identify a gene expression signature strongly predictive of a short interval to distant metastasis (“poor prognosis” signature).

We use a Support Vector Machine (SVM) to face the same problem because such a method has already done well in cancer classification problems, and we think that we could obtain slightly better results.

In addition we also address the problem of selection of a small subset of genes from the initial number of genes (~25000). We use a method of gene selection utilising SVM methods based on Recursive Feature Elimination (RFE) instead of the feature ranking with correlation method used in [1], because the last method doesn't take into account mutual information between features in the feature selection process, and this could impact classification performance [2].

1 Introduction

Breast cancer is one of the most common malignant tumours affecting women of the U.S. and European populations. Hereditary breast cancer, which accounts for less than 10 % of all cases, is due mainly to germline mutations in the tumour suppressor genes BRCA1 and BRCA2. Most cases of breast cancer occur in sporadic forms, in which the nature of their genetic determinants remain elusive[3]. Breast cancer patients with the same stage of disease can have markedly different treatment responses and overall outcome. The strongest predictors for metastasis (for example, lymph nodes status and histological grade) fail to classify accurately breast tumours according to their clinical behaviour. Chemotherapy or hormonal therapy reduces the risk of distant metastasis by approximately one-third; however, 70-80 % of patients receiving this treatment would have survived without it [1]. Since these therapies use pharmaceutical agents, such as oestrogen modulators or cytotoxic drugs, that reach cancer cells through the bloodstream, it is not surprising that these treatments frequently have toxic side effects.

Finally diagnosis of cancer must be accurate in order for the patient to receive the correct treatment and so have the best chance of survival. It would be very important to find a strategy to select patients who would benefit (or not) from such therapies as chemotherapy, and gene expression profiles are revealing a really powerful weapon to face this kind of problems (compare to the traditional way of cancer identification based on the location of tumour and its morphological appearance), especially for their ability to subtype disease. All we need now is a classification method that can afford to manage a such huge amount of data, because expression datasets contain

measurements for thousands of genes which proves problematic for many traditional methods. Also for this reason we decided to use Support Vector Machine, a supervised machine learning technique, that have been shown to perform well in evaluating microarray gene expression [4, 5, 6]. SVMs, though, are well suited to working with high dimensional data such as this. In this work a systematic and principled method is introduced that analyses microarray expression data from thousands of genes. The primary goal is the proper classification of new samples. We do this by training the SVM on samples classified by experts, and then testing the SVM on samples it has non seen before.

We also look at the method introduced in [1] to focus the analysis on a smaller subset of genes that appear to be the best diagnostic indicators. This amounts to a kind of dimensionality reduction on the dataset. If one can identify particular genes that are diagnostic for the classification one is trying to make, then there is also hope that some of these genes may be found to be of value in further investigations of the disease and in future therapies. For this purpose we use a method of gene selection utilising SVM methods based on RFE, because we are confident that it could work better than correlation methods[2]

2 Materials and methods

In recent years several methods have been developed for performing gene expression experiments. Measurements from these experiments can give expression levels for genes in tissue or cell samples. Datasets used for our experiments [11] consist of a relatively small number of tissue samples (less than 100) each with expression measurements for thousands of genes. Previous methods used in the analysis of similar datasets start with a procedure to extract the most relevant features. Most learning techniques do not perform well on datasets where the number of feature is large compared to the number of examples. SVMs are believed to be an exception. We are able to begin with tests using the full dataset, and systematically reduce the number of features selecting those we believe to be the most relevant. To understand our method a familiarity with SVM is required, and a brief introduction follows.

2.1 Support Vector Machines

Support Vector Machines are learning machines used in pattern recognition and regression estimation problems [7]. They grow up from Statistical Learning Theory (SLT) [7]., which gives some useful bounds on the generalization capacity of machines for learning tasks. The SVM algorithm constructs a separating hypersurface in the input space. Its way to do that is:

- a) Mapping the input space into a high dimensional features space through some non linear mapping chosen a priori (kernel);
- b) Constructing in this features space the Maximal Margin Hyperplane

Hyperplane are defined by $\mathbf{w} \cdot \mathbf{x} + b = 0$: when training data $(\mathbf{x}_i, y_i), i = 1, \dots, l$ are separated by this hyperplane it happens that $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1$, where $y_i = \pm 1$ are the labels. It can be shown that the margin is $\frac{2}{\|\mathbf{w}\|}$, so finding the hyperplane which separates data with maximal margin is equal to:

$$\begin{cases} \text{minimize } \frac{\|\mathbf{w}\|^2}{2} \\ \text{with } y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \end{cases} \quad (1)$$

In order to allow for misclassification errors, constraints are relaxed to $y_i(\mathbf{w} \cdot \mathbf{x} + b) \geq 1 - \xi_i, \xi_i \geq 0$. (1) becomes then:

$$\begin{cases} \text{minimize } \frac{\|\mathbf{w}\|^2}{2} + C \cdot \sum \xi_i \\ \text{with } y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i \end{cases} \quad (2)$$

The dual formulation of (2) reduces to :

$$\begin{cases} \text{maximize } \sum \alpha_i - \frac{1}{2} \sum \alpha_i \alpha_j (\mathbf{x}_i, \mathbf{x}_j) y_i y_j \\ \text{with } \sum \alpha_i y_i = 0, 0 \leq \alpha_i \leq C \end{cases} \quad (3)$$

This formulation, where example vectors are present only in dot products, makes quite simple the execution of point a) because of a theorem by Mercer[7]. It gives an easy way to compute dot products in feature spaces, where vectors in input space are non-linearly mapped by a function $\phi(\mathbf{x})$. By using a suitable function K such that $\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) = K(\mathbf{x}_i, \mathbf{x}_j)$ we do not need to calculate each singular mapping $\phi(\mathbf{x})$. If instead of controlling the overall training error one wants to control the trade-off between false positive and false negative, it is possible to modify the primal in the following way:

$$\begin{cases} \text{minimize } \frac{\|\mathbf{w}\|^2}{2} + C^+ \cdot \sum \xi_i^+ + C^- \cdot \sum \xi_i^- \\ \text{with } y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i^+, (\mathbf{w} \cdot \mathbf{x}_i + b) \leq -1 + \xi_i^- \end{cases} \quad (4)$$

where C^+ and C^- give different costs to false-positive and false-negative errors.

We use an SVM software implemented by ourself. In order to verify the reliability of our software we tested it (before of using it in this specific problem of breast cancer classification) using gene expression data from [4]. In this case the goal of using SVM was to functionally classify genes based on their expressions (we used expression data from 2467 genes of the budding yeast *Saccharomyces Cerevisiae* measured in 79 different DNA microarray hybridisation experiments). Our results are comparable with the results obtained in [4].

2.2 Feature Selection

Classical gene selection methods select the genes that individually classify best the training data. These methods include correlation methods. Evaluating how well an individual feature contributes to the separation (e.g. “poor patients” vs. “good patient”) can produce a simple feature (gene) ranking. Various correlation coefficients are used as ranking criteria. For instance the coefficient used in [8] is defined $w_i = \frac{\mu_i^+ - \mu_i^-}{\sigma_i^+ + \sigma_i^-}$ where μ_i e σ_i are the mean and standard deviation of the gene expression values of gene i for all patients of class(+) or class(-). Large positive w_i values indicate strong correlation with class(+) whereas large negative w_i values indicate strong correlation with class(-).

What characterize feature ranking with correlation methods is the implicit orthogonality assumptions that are made. Each coefficient w_i is computed with information about a single feature and does not take into account mutual information between features. Several authors have suggested using the change in objective function when one feature is removed as a ranking criterion [9]. For classification problems, the ideal objective function is the expected value of the error, that is the error rate computed on an infinite number of examples. For the purpose of training this ideal objective is replaced by a cost function J computed on training examples only. Hence the idea is to compute the change in the cost function $DJ(i)$ caused by removing a given feature or, equivalently, bringing its weight to zero. The OBD algorithm [10] approximates $DJ(i)$ by expanding J in Taylor series to second order. At the optimum of J , the first order term can be neglected yielding:

$$DJ(i) = \left(\frac{1}{2}\right) \frac{\partial^2 J}{\partial w_i^2} (Dw_i)^2. \quad (5)$$

The change in weight $Dw_i = w_i$ corresponds to removing feature i . Since SVM minimize $J = \frac{1}{2} \|\mathbf{w}\|^2$ under certain constraints, then (5) reduces to w_i^2 and this justifies the use of w_i^2 as feature ranking criterion.

2.2.1 Recursive Feature Selection

A good feature-ranking criterion is not necessarily a good feature subset ranking criterion. The criteria $DJ(i)$ or w_i^2 estimate the effect of removing one feature at a time on the objective function. They become very suboptimal when it comes to removing several features at a time, which is necessary to obtain a small feature subset. This problem can be overcome by using the following iterative procedure that is called RFE:

- 1) Train the classifier (optimise the weights w_i respect to J);
- 2) Compute the feature with smallest ranking criterion;
- 3 Remove the feature with smallest ranking criterion.

For computational reasons it may be more efficient to remove several features at a time at the expense of possible classification performance degradation.

It's relevant to be noted that RFE has no effect on correlation methods since the ranking criterion is computed with information about a single feature; so as gene selection we utilize an SVM method based on RFE, that should take into account mutual information between genes in the gene selection process, and this could impact classification performance.

Since a gene selection utilising SVM method based on RFE was already used in [2] and they obtained good results compared to classical correlation methods, we also decide to use such a method.

3.1 Data Set

The data set we used in this work can be found at [11]. We used the gene expression data from 98 primary breast cancers: 34 from patients who developed distant metastasis within 5 years, 44 from patients who continued to be disease-free after a period of at least 5 years, 18 from patients with BRCA1 germline mutations and 2 from BRCA2. All "sporadic" patients were lymph nodes negative, and under 55 years of age at diagnosis. From each patient 5 μ g total RNA was isolated from snap frozen tumours material and used to derive complementary RNA (cRNA). A reference cRNA pool was made by pooling equal amounts of cRNA from each of sporadic carcinomas. Two hybridisation were carried out for each tumour using a fluorescent dye reversal technique on

microarray containing approximately 25000 genes. The 78 sporadic lymph node negative patients were selected specifically to search for a prognostic signature in their expression profiles [1].

3 Results

The data consisted of 78 training sample and 19 test samples. Each sample was a vector corresponding to ~ 25000 genes, but approximately 5000 genes (significantly regulated in more than 3 tumours out of 78, that is, at least a twofold difference and a p-value of less than 0.01 in more than 3 tumours) were selected from the 25000 genes on the microarray.

First we perform a simple pre-processing step: we normalize the data such that for each gene expression value we subtract its mean and divide the result by its standard deviation. Then we use the Recursive Feature Elimination method, as explained in Section 2.2.1. We eliminate chunks of genes at a time (for instance a way to proceed could be: reach the number of genes which is the closest power of 2, and at subsequent iterations eliminate half of the remaining genes). We thus obtain nested subsets of genes of increasing informative density. The quality of these subsets of genes is then assessed by training an SVM (look at the next subsection).

3.1 SVM Training and testing procedure

Given a number of feature, we begin by choosing a kernel and tune the tuneable parameters of SVM software in order to achieve the best performance on a leave-one-out cross-validation tests using the training dataset (the leave-one-out procedure consists of removing one example from the training set, constructing the decision function on the basis only of the remaining training data and then testing on the removed example). In this fashion one tests all examples of the training data and measures the fraction of errors over the total number of training examples). Then For the best parameters of the leave-one-out procedure, we test the SVM on the 19 test samples.

Unfortunately at the moment we are not able to publish our results because the work is still in progress

References

- [1] Laura J. van't Veer, Hongyue Dal, and Marc J. van de Vijver, Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer, *Nature*, VOL 415, 31 January 2002, 530-535.
- [2] Isabelle Guyon, Jason Weston, Stephen Barnhill and Vladimir Vapnik, Gene Selection for Cancer Classification using Support Vector Machine.
- [3] Vanessa Yen, Study of Genomic Alterations in Human Breast Tumours by RDA, *Biojournal*, VOL 3, 2000.
- [4] Michael Brown, William Grundy, David Lin, Nello Cristianini, Charles Sugnet, Terrence Furey, Manuel Ares and David Haussler, Knowledge-based analysis of microarray gene expression data by using support vector machines, *PNAS*, VOL 97, no.1, January 4, 2000, 262-267.
- [5] T. Furey, N. Cristianini, N. Duffy, D. Bednarski, M. Schummer and D. Haussler, Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data, *Bioinformatics*, 2000.
- [6] S. Mukherjee, P. Tamayo, D. Slonim, A. Verri, T. Golub, J.P. Mesirov and T. Poggio, Support Vector Machine Classification of Microarray Data, *A.I. Memo No.1677*, C.B.C.L. Paper No.182, 1998.

- [7] V. Vapnik, The nature of statistical learning theory, Springer Verlag, 1995.
- [8] T. R. Golub, D. K. Slonim, P. Tamajo, 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, VOL 286, 15 October 1999, 531-536.
- [9] Ron Kohavi and George John, Wrappers for feature subset selection, Artificial intelligence journal, VOL 97, Nos 1-2, 1997, 273-324
- [10] Y. Le Cun, J. S. Denker and S. A. Solla, Optimum Brain Damage, in D. Touretzky Ed. Advances in Neural Information Processing Systems 2, 1990, 598-605.
- [11] Microarray data are available at <http://www.rii.com/publications/default.htm>