Bagging linear sparse Bayesian learning models for variable selection in cancer diagnosis

Chuan Lu, Andy Devos, Johan A. K. Suykens, *Member, IEEE*, Carles Arús, and Sabine Van Huffel, *Senior Member, IEEE*

Abstract

In this work, we investigated variable selection and classification for biomedical datasets with a small sample size and a very high input dimension. The sequential sparse Bayesian learning methods with linear bases were used as the basic variable selection algorithm. The selected variables were then fed to the kernel based probabilistic classifiers: Bayesian least squares support vector machines (LS-SVMs) and relevance vector machines (RVMs). We employed the bagging techniques both for variable selection and model building in order to improve the reliability of the selected variables and the predictive performance. This modelling strategy has been applied to real-life medical classification problems including two binary cancer diagnosis problems based on microarray data and a brain tumor multiclass classification problem using spectra acquired via magnetic resonance spectroscopy (MRS). Other variable selection methods exploiting support vector machine (SVM) based recursive feature elimination (RFE) variable ranking or Fisher's criterion have been compared as well. It is shown that the use of bagging can improve the reliability and stability of the variable selection and model prediction.

Index Terms

Variable selection, sparse Bayesian learning, bagging, kernel based probabilistic classifiers, microarray, magnetic resonance spectroscopy (MRS).

I. Introduction

Recent advances in technologies such as microarrays and magnetic resonance (MR) have facilitated the collection of genomic, proteomic and metabolic data that can be used for medical decision support. For example, DNA microarrays enable us to simultaneously monitor the expression of thousands of genes [1][2]. It is then possible to compare the overall differences in gene expression between normal and diseased cells. Magnetic resonance

This work was supported by the projects of Belgian Federal Government IUAP IV-02 and IUAP V-22, of the Research Council KUL MEFISTO-666 and IDO/99/03, the FWO projects G.0407.02 and G.0269.02, and the EU fp6 integrated project eTUMOUR and the EU fp6 network-of-excellence BIOPATTERN. C. Lu is supported by a doctoral grant of K.U.Leuven. A. Devos is supported by an IWT grant (IWT-Vlaanderen).

C. Lu, A. Devos, J.A.K. Suykens and S. Van Huffel are with SCD-SISTA, ESAT, Dept. of Electrical Engineering, Katholieke Universiteit Leuven (K.U.Leuven), Leuven, Belgium. E-mail: {chuan.lu, andy.devos, johan.suykens, sabine.vanhuffel} @esat.kuleuven.ac.be . C. Arús is with Departament de Bioquímica i Biologia Molecular, Universitat Autonòma de Barcelona, 08193 Cerdanyola del Vallès, Spain.

spectroscopy (MRS) [11] is able to provide detailed chemical information about the metabolites presented in living tissue. In particular, *in vivo* proton MRS offers considerable potential for clinical applications, e.g. for brain tumor diagnosis [11][12].

A great deal of attention has been paid to class prediction in the context of such new diagnostic tools, particularly for cancer diagnosis. The challenge of classification using microarrays and MR spectra lies in: (1) the large number of input variables and a relatively small number of samples, (2) the presence of noise and artefacts.

Kernel based methods are of particular interest for this task since they can deal with high dimensional data in nature and have been supported by the statistical learning theory and good empirical results [3][4][5]. Despite the early success, the presence of a significant amount of irrelevant variables or measurement noise might hamper the performance and interpretation of predictive models. Variable selection (VS) is used to identify the variables (features) most relevant for classification. This is important for medical classification, it can have an impact not only on the accuracy and complexity of the classifiers, but also on the economics of data acquisition. Moreover it is helpful for understanding the underlying mechanisms of the disease.

Several statistical and computational approaches to variable selection exist for classifying such data. The first approach is variables ranking followed by a selection (or filtering) step, usually accompanied with cross-validation (CV), so as to determine the number of variables to use in the classifier. The ranking criteria could be based on some univariate methods such as correlation, t-statistics, and some goodness measures of the corresponding univariate models. Variables could also be ranked in a multivariate manner taking into account the interactions among variables which are not accessed by the univariate methods. One popular method of this type is support vector machine (SVM) based recursive feature elimination (RFE) [3][5]. The second approach is the so-called wrapper approach, which searches for the optimal combination of variables according to some performance measures of the models [7][14][13]. The third approach is the embedded approach. It combines the two tasks of variable selection and model fitting into one optimization procedure. This approach is considered to be computationally more efficient than ranking methods and wrapper methods. The embedded SVM based algorithms typically reformulate the standard SVM optimization problem in order to select only a fixed number of variables. This can be done via imposing additional constraints and adopting another objective function such as a generalization bound [4]. Nevertheless, these methods usually require an additional step (typically of cross-validation) for choosing the predefined number of variables. In [6] the Bayesian automatic relevance determination (ARD) algorithms were exploited. This is another type of embedded methods, in which the number of selected variables can be determined automatically. However, these algorithms seem to be sensitive to a small permutation of the training set, rendering their results less reliable from a biological point of view. In this paper, we explore a similar type of Bayesian ARD methods, and show how the reliability of the selected variables and classification performance can be improved by using bagging and feeding various bootstrap variables to various types of probabilistic models. Furthermore, by utilizing the sparse Bayesian learning with logistic functions, we do not need to tune nuisance hyperparameters such as the regularization parameter used in SVMs, or a priori known noise variance in regression models as used in [6].

The rest of the paper is organized as follows. Section II starts with an introduction to sparse Bayesian learning

followed by a detailed description of a fast sequential learning algorithm. Then a brief review of the two kernel based probabilistic classification algorithms is given, including Bayesian least squares support vector machines (LS-SVM) and relevance vector machines (RVM). The bagging strategy for variable selection and modelling are proposed afterwards. In section VI we apply this method to three cancer classification problems, present the results and provide some biological interpretation of the selected variables. A discussion and conclusions are given at the end of the paper.

II. Basic algorithm for variable selection

A. Sparse Bayesian learning

Supervised learning infers a functional relation $y \leftrightarrow f(\mathbf{x})$ from a training set $\mathbf{D} = \{\mathbf{x}_n, y_n\}_{n=1}^N$, with $\mathbf{x}_n \in \mathbb{R}^d$ and $y \in \mathbb{R}$. Sparse Bayesian learning (SBL) applies Bayesian ARD to models linear in their parameters so that sparse solutions (i.e. with many parameters equal to zero) can be obtained [18]. Its prediction on y given $\mathbf{x}$ can be based upon:

$$f(\mathbf{x}; \mathbf{w}) = \sum_{m=0}^{M} w_m \phi_m(\mathbf{x}) = \mathbf{w}^{\mathbf{T}} \boldsymbol{\phi}(\mathbf{x}). \tag{1}$$

Two forms of basis functions $\phi_m(\mathbf{z})$ are considered here, namely $\phi_m(\mathbf{z}) = z_m, m = 1, \cdots, d$, (i.e. $\phi_m(\mathbf{z})$ as the original input variable), or $\phi_m(\mathbf{z}) = K(\mathbf{z}, \mathbf{x}_m), m = 1, \cdots, N$, where $K(.,.)$ denotes some symmetric kernel function. $\phi_0(\mathbf{z})$ is always set to 1 in order to include an intercept (bias) term in the model. The basic variable selection algorithm relies on the sparse Bayesian learning model using the first form of basis functions, which are referred to as the linear basis functions. In contrast, the relevance vector machines (RVMs) take the second form, i.e. the kernel representation, for the basis function. The RVM will be revisited in section III-B as a probabilistic classifier.

For a regression problem, the likelihood of the data for a sparse Bayesian learning model can be expressed as:

$$p(\mathbf{y}|\mathbf{w}, \sigma^2) = (2\pi\sigma^2)^{-N/2} \exp\{-\frac{1}{2\sigma^2} \|\mathbf{y} - \boldsymbol{\Phi}\mathbf{w}\|^2\}, \tag{2}$$

where σ^2 is the variance of the i.i.d. noise, the $N \times M$ design matrix $\boldsymbol{\Phi} = [\boldsymbol{\phi}(\mathbf{x}_1), \boldsymbol{\phi}(\mathbf{x}_2), \ldots, \boldsymbol{\phi}(\mathbf{x}_N)]^T$. The parameters $\mathbf{w}$ are given a Gaussian prior

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{m=0}^{M} (\frac{\alpha_m}{2\pi})^{1/2} \exp(-\frac{\alpha_m w_m^2}{2}) \tag{3}$$

where $\boldsymbol{\alpha} = \{\alpha_m\}$ is a vector of hyperparameters, the Gaussian prior of $w_m \sim \mathcal{N}(0, \alpha_m^{-1})$ has mean zero and an individual variance α_m^{-1}. As illustrated in [18][20], this is equivalent to using a regularized cost function with a penalty term of $\sum_m \log |w_m|$, which encourages sparsity. The hyperparameters $\boldsymbol{\alpha}$ can be estimated using the framework of type II maximum likelihood, in which the marginal likelihood $p(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2) = \int_{-\infty}^{\infty} p(\mathbf{y}|\mathbf{w}, \sigma^2) p(\mathbf{w}|\boldsymbol{\alpha}) \mathbf{dw}$ is maximized. The marginal likelihood with respect to $\boldsymbol{\alpha}$ and σ^2 can be computed by:

$$p(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2) = (2\pi)^{-N/2} |\mathbf{C}|^{-1/2} \exp(-\frac{1}{2}\mathbf{y}^T \mathbf{C}^{-1} \mathbf{y}), \tag{4}$$

where $\mathbf{C} = \mathbf{B^{-1}} + \mathbf{\Phi A}^{-1}\mathbf{\Phi}^T$, with $\mathbf{B} = \sigma^{-2}\mathbf{I}$ and $\mathbf{A} = \mathrm{diag}(\alpha_1, \ldots, \alpha_M)$.

For binary classification problems, one can utilize the logistic function $g(a) = 1/(1+e^{-a})$ [18]. The computation of the likelihood is based on the Bernoulli distribution:

$$p(\mathbf{y}|\mathbf{w}) = \prod_{n=1}^{N} g(f(\mathbf{x}_n; \mathbf{w}))^{y_n}[1 - g(f(\mathbf{x}_n; \mathbf{w}))]^{1-y_n}, \tag{5}$$

where $y_n \in \{0, 1\}$. There is no noise variance in this case, and a local Gaussian approximation is used to compute the posterior distribution of the weights and the marginal likelihood $p(\mathbf{y}|\boldsymbol{\alpha})$.

This optimization process can be performed efficiently using an iterative re-estimation procedure and utilizing the conditions on the maximum of $p(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2)$ [18][20]. A fast sequential learning algorithm has also been introduced in [19], which enables us to efficiently process data of high dimensionality. We have adapted this algorithm to our applications, which will be detailed in the next subsection.

The most relevant variables for the classifier can be obtained from the resulting sparse solutions, if the original variables are taken as basis functions in the SBL model. This type of models is referred to as linear SBL models in this paper.

B. Sequential sparse Bayesian learning algorithm

The sequential SBL algorithm [19] starts from a zero basis, adds or deletes a basis function at each iteration step, or updates a hyperparameter α_m until convergence.

For optimization of the hyperparameters $\boldsymbol{\alpha}$, the objective function uses the logarithm of the marginal likelihood $\mathcal{L}(\boldsymbol{\alpha}) = \log p(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2)$. It is shown in [19][20] that we could analyze the properties of $\mathcal{L}(\boldsymbol{\alpha})$ by decomposing it into the marginal likelihood $\mathcal{L}(\boldsymbol{\alpha}_{-i})$ with $\boldsymbol{\phi}_i$ (the ith column of $\mathbf{\Phi}$) excluded, and the marginal likelihood $\ell(\alpha_i)$ including only $\boldsymbol{\phi}_i$. That is $\mathcal{L}(\boldsymbol{\alpha}) = \mathcal{L}(\boldsymbol{\alpha}_{-i}) + \ell(\alpha_i)$, where $\ell(\alpha_i) = \frac{1}{2}[\log\alpha_i - \log(\alpha_i + s_i) + \frac{q_i^2}{\alpha_i + s_i}]$, $s_i = \boldsymbol{\phi}_i^T\mathbf{C}_{-i}^{-1}\boldsymbol{\phi}_i$ and $q_i = \boldsymbol{\phi}_i^T\mathbf{C}_{-i}^{-1}\mathbf{y}$, $\mathbf{C}_{-i}$ is $\mathbf{C}$ with the contribution of $\boldsymbol{\phi}_i$ removed. Since s_i and q_i are independent of α_i, one can obtain a unique maximum of $\mathcal{L}(\boldsymbol{\alpha})$ with respect to α_i by setting the first derivative of $\ell(\alpha_i)$ to zero. The optimal values for α_i are:

$$\tilde{\alpha}_i = \frac{s_i^2}{q_i^2 - s_i}, \text{ if } q_i^2 > s_i, \tag{6}$$

$$\tilde{\alpha}_i = \infty, \text{ if } q_i^2 \leq s_i. \tag{7}$$

One convenient way to derive s_i and q_i is to utilize these expressions: $s_i = \frac{\alpha_i S_i}{\alpha_i - S_i}$, $q_i = \frac{\alpha_i Q_i}{\alpha_i - S_i}$, and when $\alpha_i = \infty$, $s_i = S_i$ and $q_i = Q_i$. In practice the two quantities S_i and Q_i are computed using the following equations:

$$S_i = \boldsymbol{\phi}_i^T\mathbf{C}^{-1}\boldsymbol{\phi}_i = \boldsymbol{\phi}_i^T\mathbf{B}\boldsymbol{\phi}_i - \boldsymbol{\phi}_i^T\mathbf{B}\mathbf{\Phi}\hat{\mathbf{\Sigma}}\mathbf{\Phi}^T\mathbf{B}\boldsymbol{\phi}_i \tag{8}$$

$$Q_i = \boldsymbol{\phi}_i^T\mathbf{C}^{-1}\hat{\mathbf{y}} = \boldsymbol{\phi}_i^T\mathbf{B}\hat{\mathbf{y}} - \boldsymbol{\phi}_i^T\mathbf{B}\mathbf{\Phi}\hat{\boldsymbol{\mu}}, \tag{9}$$

where $\mathbf{\Phi}$, $\mathbf{\Sigma}$ and $\hat{\boldsymbol{\mu}}$ contain only the parts corresponding to the basis functions included in the model (with $\alpha_m < \infty$). In the regression framework, $\mathbf{B} \equiv \sigma^{-2}\mathbf{I}$, $\hat{\mathbf{\Sigma}} \equiv \mathbf{\Sigma} = (\sigma^2\mathbf{\Phi}^T\mathbf{\Phi} + \mathbf{A})^{-1}$, $\hat{\boldsymbol{\mu}} \equiv \boldsymbol{\mu} = \sigma^{-2}\mathbf{\Sigma}\mathbf{\Phi}^T\mathbf{y}$, and

$\hat{\mathbf{y}} \equiv \mathbf{y}$. Here $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ correspond to the mean and variance of the posterior distribution over $\mathbf{w}$, which is also Gaussian. In the case of classification utilizing a logistic link function, the weight posterior $p(\mathbf{w}|\mathbf{y}, \boldsymbol{\alpha})$ is however not analytically available, but can be computed using a Gaussian approximation. For the current value of $\boldsymbol{\alpha}$, we can estimate the quantities of interest through an iteratively reweighted least squares (IRLS) algorithm such as the Newton-Raphson method. The following expressions can be exploited [18]:

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}} &= (\boldsymbol{\Phi}^T \mathbf{B} \boldsymbol{\Phi} + \mathbf{A})^{-1}, \\ \widehat{\boldsymbol{\mu}} &= \widehat{\boldsymbol{\Sigma}} \boldsymbol{\Phi}^T \mathbf{B} \widehat{\mathbf{y}}, \\ \text{and } \widehat{\mathbf{y}} &= \boldsymbol{\Phi} \widehat{\boldsymbol{\mu}} + \mathbf{B}^{-1}(\mathbf{y} - g(\boldsymbol{\Phi} \widehat{\boldsymbol{\mu}})), \end{aligned}$$

where $\mathbf{B} = \mathrm{diag}(\beta_1, \ldots, \beta_N)$, with $\beta_n = g(f(\mathbf{x}_n; \widehat{\boldsymbol{\mu}}))[1 - g(f(\mathbf{x}_n; \widehat{\boldsymbol{\mu}}))]$.

The marginal likelihood maximization algorithm jointly optimizes the weights and the hyperparameters $\{\alpha_m\}_{m=0}^{M_{\mathrm{all}}}$, with M_{all} the maximum index for the basis functions. In case of linear basis functions, $M_{\mathrm{all}} = d$; in case of linear kernel basis functions $M_{\mathrm{all}} = N$. Define the complete set of possible indices for the basis functions as $\mathcal{I}_{\mathrm{all}}$, containing the integer numbers from 0 to M_{all}. Our modified algorithm for classification utilizing the logistic function is as follows.

1) Initialize the model with only an intercept: $\alpha_0 < \infty$ (e.g. $\alpha_0 = (\mathbf{y}^T \mathbf{y}/N)^{-2}$), and $\forall m > 0, \alpha_m = \infty$. Initialize the index set of the bases in the model $\mathcal{I}_{\mathrm{sel}} \leftarrow \{0\}$.
2) Given current $\boldsymbol{\alpha}$, estimate $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\mu}}$ using the IRLS algorithm for the logit model. Note that $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\mu}}$ are only related to the basis functions included in the current model, initially with only one scalar element. And $\boldsymbol{\Phi}$ starts with only one column vector $\boldsymbol{\phi}_0 = [1, \cdots, 1]_{1\times N}^T$.
3) Randomly select M_{out} bases with the indices of $\mathcal{I}_{\mathrm{out}} \subseteq (\mathcal{I}_{\mathrm{all}} \setminus \mathcal{I}_{\mathrm{sel}})$. Set $\mathcal{I}_{\mathrm{can}} \leftarrow \mathcal{I}_{\mathrm{out}} \cup \mathcal{I}_{\mathrm{sel}}$.
4) For each basis vector in the candidate sets $\boldsymbol{\phi}_m, m \in \mathcal{I}_{\mathrm{can}}$, compute the value of s_m and q_m, find out the optimal action with respect to each α_m, then calculate ∇_m, the corresponding change in marginal likelihood $\mathcal{L}(\boldsymbol{\alpha})$ after taking that action. The following rules are used:
   - If $q_m^2 > s_m$ and $\alpha_m < \infty$ (i.e. $\boldsymbol{\phi}_i$ is in the model), estimate $\tilde{\alpha}_m$ using (6), $\nabla_m = \frac{Q_m^2}{S_m + [\tilde{\alpha}_m^{-1} - \alpha_m^{-1}]^{-1}} - \log\{1 + S_m[\tilde{\alpha}_m^{-1} - \alpha_m^{-1}]\}$.
   - If $q_m^2 > s_m$ and $\alpha_m = \infty$, add $\boldsymbol{\phi}_m$ to the model, compute $\tilde{\alpha}_m$ using (6), $\nabla_m = \frac{Q_m^2 - S_m}{S_m} + \log \frac{S_m}{Q_m^2}$.
   - If $q_m^2 \leq s_m$ and $\alpha_m < \infty$, then delete $\boldsymbol{\phi}_m$, setting $\tilde{\alpha}_m = \infty$, $\nabla_m = \frac{Q_m^2}{S_m - \alpha_m} - \log(1 - \frac{S_m}{\alpha_m})$.
5) Select one basis $m* = \arg\max \nabla_m \mathcal{L}$, and take the corresponding action, i.e. $\alpha_{m*} \leftarrow \tilde{\alpha}_{m*}$ and update $\boldsymbol{\Phi}$.
6) If convergence is reached then stop, otherwise goto step 2.

The number of bases to be screened for updating α_m is the number of bases in the model plus M_{out}, the predefined number of randomly selected bases from those not used by the model. In our experiments, M_{out} has been fixed to 100. And in this paper, the optimization procedure is considered to be converged when the maximum value of $|\log(\tilde{\alpha}_m/\alpha_m)|_{m\in\mathcal{I}_{\mathrm{can}}}$ in step 4 is lower than 10^{-6}.

However, we should also be aware of the uncertainty involved in the basis function selection, which might result from the existence of multiple solutions, and the sensitivity of the algorithm to small perturbations of experimental conditions. Attempts to tackle this problem are for example bagging and committee machines. Here we will focus

on the very simple bagging approach, which will be described in Section IV.

III. KERNEL BASED PROBABILISTIC CLASSIFIERS

Support Vector Machines (SVM) are now a state-of-the-art technique for pattern recognition [21]. A standard SVM classifier takes the form $y(\mathbf{x}) = \text{sign}[\mathbf{w}_\text{f}^T \boldsymbol{\varphi}(\mathbf{x}) + b]$ in the feature (primal) space with $\boldsymbol{\varphi}(.) : \mathbb{R}^d \rightarrow \mathbb{R}^{d_\text{f}}$, where d_f is the dimension of the (potentially infinite dimensional) feature space. It is inferred from data with binary targets $y_i \in \{\pm 1\}$ by solving the following optimization problem:

$$\min_{\mathbf{w}_\text{f}, b, \boldsymbol{\xi}} \mathcal{J}(\mathbf{w}_\text{f}, b, \boldsymbol{\xi}) = \frac{1}{2}\mathbf{w}_\text{f}^T \mathbf{w}_\text{f} + C \sum_{i=1}^{N} \xi_i, \tag{10}$$

subject to $y_i(\mathbf{w}_\text{f} \boldsymbol{\varphi}(\mathbf{x}_i) + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \ldots, N$.

This can be conveniently solved in its dual formulation. It turns out that $f(\mathbf{x}; \mathbf{w}_\text{f}, b) = \mathbf{w}_\text{f}^T \boldsymbol{\varphi}(\mathbf{x}) + b = \sum_{i=1}^{N} a_i \, y_i \, K(\mathbf{x}, \mathbf{x}_i) + b$, where a_i is called a support value, and $K(\cdot, \cdot)$ is a chosen positive definite kernel satisfying Mercer's condition. The most common kernels include linear kernels and radial basis function (RBF) kernels. Here we only considered models with linear kernels, defined as $K(\mathbf{x}, \mathbf{z}) = \mathbf{x}^T \mathbf{z}$.

One can see that SVMs share actually the same functional form with the sparse Bayesian learning models. However, the basis function used in SVMs must be a positive definite kernel.

A. Bayesian LS-SVM classifier

The LS-SVM is a least squares version of SVM, and is closely related to Gaussian processes and kernel Fisher discriminant analysis [22][23]. The training procedure for LS-SVM is reformulated as

$$\min_{\mathbf{w}_\text{f}, b, \mathbf{e}} \mathcal{J}(\mathbf{w}_\text{f}, b, \mathbf{e}) = \frac{1}{2}\mathbf{w}_\text{f}^T \mathbf{w}_\text{f} + \frac{\lambda}{2} \sum_{i=1}^{N} e_i^2, \tag{11}$$

subject to $y_i[\mathbf{w}_\text{f}^T \boldsymbol{\varphi}(\mathbf{x})_i + b] = 1 - e_i, i = 1, \ldots, N$.

This optimization problem can be transformed and solved through a linear system in the dual space instead of a quadratic programming problem as for the standard SVM case [23]:

$$\left[\begin{array}{c|c} 0 & \mathbf{y}^T \\ \hline \mathbf{y} & \boldsymbol{\Omega} + \lambda^{-1}\mathbf{I} \end{array}\right] \left[\begin{array}{c} b \\ \hline \mathbf{a} \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline \mathbf{1} \end{array}\right] \tag{12}$$

with $\mathbf{e} = [e_1, \cdots, e_N]^T$, $\mathbf{a} = [a_1, \cdots, a_N]^T$, and $\mathbf{1} = [1 \cdots 1]^T$. The matrix $\boldsymbol{\Omega}$ is defined as follows: $\boldsymbol{\Omega}_{ij} = y_i y_j \boldsymbol{\varphi}(\mathbf{x}_i)^T \boldsymbol{\varphi}(\mathbf{x}_j) = y_i y_j \, K(\mathbf{x}_i, \mathbf{x}_j)$.

In order to achieve a high performance by means of LS-SVM models, one still needs to tune the regularization and possible kernel parameters. Bayesian LS-SVM (BayLSSVM), which was introduced in [25][23], applied the evidence framework [24] to LS-SVMs with three levels of inferences. The regularization parameter λ is optimized automatically by maximizing the posterior probability of the model. The inference in Bayesian LS-SVM modelling follows intrinsically the linear regression framework, assumes a Gaussian noise model over the target variable, and the prior distribution for $\mathbf{w}_\text{f}$ is a Gaussian with a fixed variance α^{-1} for all weights, which is related to

the regularization parameter λ in the standard LS-SVM. The most probable model parameters $\mathbf{w}_{\mathrm{MP}}$ and b_{MP} are estimated implicitly using conventional LS-SVM training methods, i.e. by solving (12) in the dual space. The class probabilities for the classification can be calculated in the following way. For a given test case, compute the conditional class probabilities $p(\mathbf{x}_*|y_* = \pm 1, \mathbf{y}, \alpha, \sigma^2)$ using the two Gaussian probability densities for two classes at the most probable values: $f(\mathbf{x}_*; \mathbf{w}_{\mathrm{f}})|_{\mathbf{w}_{\mathrm{MP}},\pm} \sim \mathcal{N}(\mu_\pm, \sigma^2_{*,\pm})$. The mean of each distribution $\mu_\pm$ is defined as the class center of the latent output $f(\mathbf{x}; \mathbf{w}_{\mathrm{MP}}, b_{\mathrm{MP}})$ in the training set. The variance $\sigma^2_{*,\pm}$ comes from both the target noise and the uncertainty in the parameter $\mathbf{w}_{\mathrm{f}}$. For more computational details, the interested readers are referred to [23][25]. By applying Bayes' rule the posterior class probabilities of the LS-SVM classifier can be obtained:

$$p(y_*|\mathbf{x}_*, \mathbf{y}, \alpha, \sigma^2) = \frac{p(y)p(\mathbf{x}_*|y_*, \mathbf{y}, \alpha, \sigma^2)}{\sum_{y'=\pm 1} p(y')p(\mathbf{x}_*|y', \mathbf{y}, \alpha, \sigma^2)}, \tag{13}$$

$p(y)$ corresponds to the prior class probability. Its default value is set to the proportion of a certain class among the training data.

B. Relevance vector machines for classification

As mentioned previously, the relevance vector machine is a special case of sparse Bayesian learning models, in which the basis functions are given by kernel functions of the same type as for SVMs. The sequential learning algorithm introduced in Section II-B was again applied to the optimization of RVMs. The predicted probability of being positive for a given input $\mathbf{x}_*$ can be computed by using the logistic function:

$$p(y_* = 1|\mathbf{x}_*, \mathbf{y}, \boldsymbol{\alpha}) = \frac{1}{1 + e^{-\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_*)}} \tag{14}$$

Prior class probabilities can be incorporated into the model by adjusting the bias term w_0 to $w_0 - \log(N_+/N_-) + \log(P_+/P_-)$, where $N+$ and N_- denote the number of positive and negative cases in the training set, P_+ and P_- the prior class probability for the positive and negative class, respectively.

No simulation results of the models with nonlinear kernels (such as RBF kernels) are reported in this paper, as linear classifiers perform sufficiently well for our problems at hand, and nonlinear models have shown no improvement over the simple linear classifiers.

IV. BAGGING STRATEGY

A. Bagging the selected variables and models

Bagging is a "bootstrap" ensemble method that generates individuals for its ensemble by training each classifier on a random redistribution of the training set [27]. Each classifier's training set is generated by randomly drawing, with replacement, the same number of examples as in the original training set. It is shown that the bootstrap mean is approximately a posterior average of a quantity of interest [28]. Suppose a model is fitted to our training set $\mathbf{D}$, obtaining the prediction $f(\mathbf{x})$ at input $\mathbf{x}$. This prediction could be the latent outcome of e.g. a standard

SVM model, or the predicted class probability of a probabilistic model. Bootstrap aggregation or bagging averages this prediction over a collection of bootstrap samples, thereby reducing its variance. For each bootstrap sample $\mathbf{D}^{*b}, b = 1, 2, \cdots, B$, we fit the model, giving prediction $f^{*b}(\mathbf{x})$. The bagging estimate is defined by

$$f_{\text{bag}}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} f^{*b}(\mathbf{x}). \tag{15}$$

The final class label will be decided by thresholding the bootstrap estimate of the class probability or the latent outcome. Bagging can push a good but unstable procedure a significant step towards optimality, which has been evidenced both experimentally and theoretically [27][28].

Note that, the class proportion in the bootstrap samples was not controlled in the experiments. However, the prior class probabilities have been fixed for each bagged model based on the training set class distribution.

An alternative bagging strategy is to bag only the predicted class labels and the final prediction will be given by voting. However, a reliable estimate of the class probability is essential for medical diagnosis. The prediction averaging strategy tends to produce bagged estimates with lower variance, especially for small B. Therefore, the prediction averaging strategy is preferred and advocated here.

We now sketch our bagging strategy for variable selection and modelling. For a given training set, B bootstrap data are randomly generated with replacement. For each bootstrap training set, one subset of variables is selected via the SBL logit models with linear basis functions, followed by feeding these variables to a model of interest such as Bayesian LS-SVM. Then B subsets of variables are selected and B corresponding models are built based upon the B bootstrap training data. Given input data $\mathbf{x}$, the class probability or the latent outcome for the bagged binary classifier will be the average of the B class probabilities or latent outcomes. We took $B = 30$ in our experiments.

B. Strategy for the multiclass classification problems

For multiclass classification problems, we reduce the k-class classification problem into $k(k-1)/2$ pairwise binary classification problems, which yield the conditional pairwise probability estimates for a given input $\mathbf{x}$. The conditional probabilities are then coupled to obtain the joint posterior probability for each class by using Hastie's method [26]. The final prediction of the class will be the one for which the highest joint probability is achieved. Accordingly the variables used should be the union of the $k(k-1)/2$ sets of variables that are used in the same number of binary classifiers. Bagging is applied to each binary classification individually. Only the mean predicted probabilities from the bagged binary classifiers are coupled in order to get the final joint posterior probability for the multiclass classification problems.

V. COMPARED METHODS

In order to see the potential performance gain of using our proposed methods, we also assessed the performance of some reference methods. We denote the proposed variable selection approach as "LinSBL+Bag", which bags the variables selected from the linear SBL logit models. Accordingly the model fitting and prediction will be "bagged"

as well. Its counterpart method "LinSBL" forms a classifier using only one subset of variables selected from a single linear SBL logit model, which is based on the whole training set without bootstrap repetition. Other variable selection methods used for comparison include two typical variable ranking methods, in which the models are built on the N_{v} variables with the highest ranks.

One of these is the popular SVM based recursive feature elimination (RFE) method [3]. The idea of this method is to recursively eliminate the variable which contributes the least in the SVM model, and then rank the variables based on the reverse order of their elimination. The contribution of the mth variable is evaluated by means of the change in the cost function $\nabla\mathcal{J}_m$ caused by removing the mth variable. When a linear kernel is used, $\nabla\mathcal{J}_m = w_m^2$ with w_m the corresponding weight in the linear SVM model: $w_m = \Sigma_{i=1}^{N} a_i x_{im} y_i$.

The variables can also be ranked based on Fisher's criterion [24], which is a measure of the correlation between the variables and the class labels. For a binary classification, the Fisher discriminant criterion for an individual variable is given by $(\mu_{m,+} - \mu_{m,-})^2/(\sigma_{m,-}^2 + \sigma_{m,+}^2)$, where $\mu_{m,+}$ and $\mu_{m,-}$ are the means of variable m within the positive and negative class, respectively, $\sigma_{m,+}$ and $\sigma_{m,-}$ the standard deviations of the variable within each class. The larger the Fisher's criterion, the higher the ranking of the variable.

N_{v} was tuned by means of 10-fold cross-validation using SVMs. A coarse-to-fine strategy was utilized to search for N_{v} within a range of possible values. The N_{v} with the lowest 10-fold CV error rate were selected, and the tie breaking rule is to choose the smallest number of variables. The ranking and the number of variables were determined by using the whole training set in each run of cross-validation. These two methods are denoted by "RFE+CV" and "Fisher+CV", respectively.

Note that no bagging was applied for these reference methods. Our preliminary experiments show that the effect in bagging the models is more prominent when the variables vary among different bootstrap data. However, it becomes too time consuming to bootstrap variable selection with the methods of "ranking + CV".

Concerning the modelling techniques, besides the advocated probabilistic models, we use the standard linear SVM classifier as baseline model. In this work, we fixed the regularization hyperparameter of SVM to 10^6, which is high enough to keep the training error low. Unlike the other probabilistic models, the final SVM classifiers do not generate naturally the probability output. Therefore, for the multiclass classification problem, the final predicted class labels were decided by voting using the pairwise binary SVM classification results.

VI. Experiments

A. Experimental settings

The generalization performance of the models in conjunction with variable selection was evaluated by 30 runs of randomized cross-validation. We applied a full cross-validation, where the variable selection was conducted prior to each model fitting process for each realization of the training data. An incomplete cross-validation, i.e. a cross-validation after variable selection might lead to a serious underestimation of the prediction error [8].

The data set was divided into two parts by stratified random splitting. The training set was used for model building including variable selection and model fitting, while the remaining data were used for test purposes. The

performance is measured by the mean accuracy (Acc) and the mean area under the ROC curve (AUC) [29] and the corresponding standard error (SE) of the mean. Our matlab programs used for these experiments were built upon several toolboxes, including the SparseBayes V1.0 [1] (with modifications) for the sequential sparse Bayesian learning, the Spider[2] for RFE and SVM modelling, and LS-SVMlab [3] for Bayesian LS-SVM modelling.

B. Binary cancer classification based on microarray data

Two benchmark binary cancer classification problems based on DNA microarray data have been considered. The first problem aims to discriminate between two types of leukemia: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). The dataset includes 72 samples (47 of AML and 25 of ALL) with 7129 gene expression values [4] [1][3]. The second problem aims to differentiate tumors from normal tissues using colon cancer data. The dataset contains information of 40 tumor and 22 normal colon tissues. Each tissue is associated with 2000 gene expression values [5] [2][3]. All microarray data have been normalized to zero mean and unit variance. Each realization of the training set contains 50 data points, the test set includes the rest of the data of size 22 and 12 in leukemia and colon cancer data, respectively.

The accuracy and the AUC for the leukemia and colon cancer classification problems are reported in Table I and Table II, respectively. The mean number of selected variables ($\overline{N_{\mathrm{v}}}$) for each variable selection (VS) method within one trial is also given.

Notice that the accuracy measure varies with the change of class priors or the probability cutoff values (which was fixed to 0.5 in our binary classifications). The most prominent changes are the following. When the priors P_+ and P_- in leukemia cancer prediction were both set to 0.5 instead of their default values 0.34 and 0.66, the accuracy for BayLSSVM using all variables increased to $95 \pm 0.78\%$. In the colon cancer diagnosis, bagged BayLSSVMs and RVMs reached a higher accuracy of $86.11 \pm 1.68\%$ and $82.78 \pm 1.92\%$, if P_+ and P_- were set to 0.2 and 0.8 (instead of their default values 0.36 and 0.64), respectively. On the contrary, the AUC measure is independent from the prior class distribution. Although AUCs for the bagged models might change slightly due to the bagging effect, e.g. the mean AUC for the bagged BayLSSVM changed from 89.06% to 89.38% for colon data.

Additionally, the LinSBL logit model is itself a probabilistic model with an automatic variable selection mechanism. It achieved a similar performance as the probabilistic kernel models. For prediction of leukemia cancer, a single LinSBL model and a bagged LinSBL model yielded a test performance in terms of mean AUC of $94.55 \pm 1.05\%$ and $98.51 \pm 0.46\%$, respectively. Using the colon data, the test performance for a single LinSBL and a bagged LinSBL reached a mean AUC of $82.08 \pm 1.94\%$ and $89.27 \pm 1.72\%$, respectively.

[1] http://research.microsoft.com/mlp/RVM/SparseBayesV1.00.tar.gz

[2] http://www.kyb.tuebingen.mpg.de/bs/people/spider/index.html

[3] http://www.esat.kuleuven.ac.be/sista/lssvmlab/

[4] available online at www.genome.wi.mit.edu/MPR/data_set_ALL_AML.html

[5] available online at microarray.princeton.edu/oncology/affydata/index.html

TABLE I

TEST RESULTS FOR LEUKEMIA CANCER CLASSIFICATION, REPORTED AS MEAN±SE OF THE ACCURACY AND AUC FROM 30 RUNS OF CROSS-VALIDATION. ALL CLASSIFIERS WERE TESTED WITH THE SAME SERIES OF VARIABLE SELECTION TECHNIQUES, AND UTILIZED ONLY LINEAR KERNELS. THE HIGHEST VALUE OF ACCURACY OR AUC FOR EACH TYPE OF CLASSIFIER (IN ROW) IS INDICATED IN BOLD.

VS method	All		Fisher+CV		RFE+CV		LinSBL		LinSBL+Bag	
$\overline{N}_{\mathrm{v}}$	7129		43.30		5.10		3.50		49.10	
Classifier	Acc(%)	AUC(%)	Acc(%)	AUC(%)	Acc(%)	AUC(%)	Acc(%)	AUC	Acc(%)	AUC(%)
SVM	**95.61**	**99.29**	92.42	97.29	90.30	94.88	90.45	94.43	92.42	98.01
	±0.73	±0.23	±0.89	±0.64	±1.19	±1.17	±0.92	±0.94	±0.92	±0.46
BayLSSVM	89.85	**98.78**	92.73	97.62	90.00	95.01	88.94	93.50	**93.79**	98.48
	±1.05	±0.34	±1.04	±0.59	±1.29	±1.11	±1.02	±1.40	±1.02	±0.41
RVM	90.15	96.22	93.03	97.26	90.45	94.93	89.85	93.72	**93.18**	**98.24**
	±1.34	±0.77	±0.85	±0.61	±1.24	±1.10	±1.17	±1.17	±1.00	±0.47

TABLE II

TEST RESULTS FOR COLON CANCER CLASSIFICATION, REPORTED AS MEAN±SE OF THE ACCURACY AND AUC FROM 30 RUNS OF CROSS-VALIDATION. ALL CLASSIFIERS WERE TESTED WITH THE SAME SERIES OF VARIABLE SELECTION TECHNIQUES, AND UTILIZED ONLY LINEAR KERNELS. THE HIGHEST VALUE OF ACCURACY OR AUC FOR EACH TYPE OF CLASSIFIER (IN ROW) IS INDICATED IN BOLD.

VS method	All		Fisher+CV		RFE+CV		LinSBL		LinSBL+Bag	
$\overline{N}_{\mathrm{v}}$	2000		114.30		9.73		6.50		107.17	
Classifier	Acc(%)	AUC(%)	Acc(%)	AUC(%)	Acc(%)	AUC(%)	Acc(%)	AUC(%)	Acc(%)	AUC(%)
SVM	81.94	85.31	80.28	85.83	81.36	83.54	76.39	81.67	**86.11**	**87.71**
	±1.80	±2.22	±2.10	±2.26	±2.18	±3.10	±2.03	±2.23	±1.73	±1.93
BayLSSVM	85.00	88.65	**85.28**	**89.27**	81.11	86.15	76.67	84.48	84.44	89.06
	±1.73	±1.71	±1.56	±1.60	±2.11	±2.79	±2.08	±2.09	±1.65	±1.78
RVM	83.61	87.71	**83.61**	86.98	81.67	86.46	73.06	82.40	79.17	**87.71**
	±1.82	±2.16	±1.73	±2.17	±2.30	±2.75	±2.41	±2.23	±1.99	±1.88

C. Classification of brain tumors based on MRS data

The method has also been applied to a multiclass classification problem of brain tumors using short echo time ^{1}H MRS data. The dataset consists of 205 spectra in the frequency domain. The full spectrum (a row vector of magnitude values) has been normalized to unit norm. Only the frequency region of interest from 4.17 to 0 ppm (a measure of the chemical shift in a field independent frequency scale) was used in this study, corresponding to 138 input variables. The dataset contains the records from four major types of brain tumors: meningiomas (Class 1, 57 spectra), astrocytomas grade II (Class 2, 22 spectra), glioblastomas (87 spectra) and metastases (39 spectra) [15]. However, the last two types of tumors are very difficult to distinguish. Our experience on this dataset is that, the

TABLE III

TEST AUC (%) FOR PAIRWISE BINARY CLASSIFICATION OF BRAIN TUMORS, REPORTED AS MEAN±SE FROM 30 RUNS OF CROSS-VALIDATION. ALL CLASSIFIERS WERE TESTED WITH THE SAME SERIES OF VARIABLE SELECTION TECHNIQUES, AND UTILIZED ONLY LINEAR KERNELS. THE HIGHEST AUC OF EACH BINARY CLASSIFICATION FOR EACH TYPE OF MODEL (IN ROW) IS INDICATED IN BOLD.

VS method	All			Fisher+CV			RFE+CV			LinSBL			LinSBL+Bag		
Class pair	1vs.2	1vs.3	2vs.3	1vs.2	1vs.3	2vs.3	1vs.2	1vs.3	2vs.3	1vs.2	1vs.3	2vs.3	1vs.2	1vs.3	2vs.3
Classifier $\overline{N}_v$	138	138	138	39.17	114.57	9.03	6.10	23.10	14.07	4.30	9.50	6.20	31.17	59.67	55.40
SVM	**99.82**	97.52	92.14	98.75	96.17	95.06	98.18	96.56	92.23	96.99	96.51	90.78	98.17	**97.76**	**96.01**
	±0.08	±0.26	±0.77	±0.41	±0.89	±0.69	±0.56	±0.37	±0.95	±0.62	±0.37	±1.07	±0.88	±0.28	±0.44
BayLSSVM	99.62	97.33	95.15	98.67	96.85	95.43	97.44	97.04	94.23	96.94	96.88	93.38	**99.65**	**97.95**	**96.44**
	±0.16	±0.27	±0.56	±0.39	±0.39	±0.63	±0.57	±0.30	±0.70	±0.57	±0.34	±0.82	±0.15	±0.26	±0.40
RVM	98.47	97.55	**96.87**	98.60	96.82	95.80	97.52	97.18	95.57	96.99	97.02	94.52	**99.70**	**97.94**	96.19
	±0.37	±0.27	±0.37	±0.38	±0.47	±0.65	±0.74	±0.32	±0.67	±0.72	±0.35	±0.70	±0.13	±0.26	±0.42

trained models did not perform better than a majority classifier, which assigns the majority class in the training set to all the test cases. Therefore, we merged the two tumor types - glioblastomas and metastases - into one class of aggressive tumors (Class 3), and only dealt with the three-class classification problem. For details of the data acquisition and preprocessing procedure for this dataset, the readers are referred to [15].

Since the data are unbalanced, the model using the default priors will lead to a relatively low sensitivity for astrocytomas grade II. Therefore, we decided to use equal priors for all binary classifiers, which resulted into a "satisfactory" sensitivity and specificity for all three classes. The average test AUC for each pairwise binary classification is reported in Table III. Table IV presents both the training and test accuracy of the brain tumor classification problems using equal class priors.

Again the LinSBL model can be used to classify brain tumors. Using equal priors, the test performance for 3-class prediction reached a mean accuracy of $86.03 \pm 0.63\%$ by using a single LinSBL model, and $89.46 \pm 0.52\%$ by using a bagged LinSBL model.

Notice that, the mean number of variables selected by LinSBL+Bag for the 3-class classification problem is 98.73, which means that about 72% of the variables were involved in the classification. However, checking the number of variables selected for each pairwise binary classification provides more insight. For example the comparison of Class 1 vs. 2 used only an average of 31.17 variables.

D. Biological relevance of the selected variables

We examined the most frequently selected variables from the LinSBL+Bag and their biological relevance for each dataset. This was done by first calculating the number of occurrences o_{vm} of variable m being selected on the vth run of cross-validation, i.e. on the 30 bootstrap data of one training set realization. This number of occurrences

TABLE IV

TRAINING AND TEST ACCURACY FOR BRAIN TUMOR THREE-CLASS CLASSIFICATION, REPORTED AS MEAN±SE FROM 30 RUNS OF CROSS-VALIDATION. ALL CLASSIFIERS WERE TESTED WITH THE SAME SERIES OF VARIABLE SELECTION TECHNIQUES, AND UTILIZED ONLY LINEAR KERNELS. THE HIGHEST VALUE OF TRAINING ACCURACY OR TEST ACCURACY FOR EACH TYPE OF CLASSIFIER (IN ROW) IS INDICATED IN BOLD.

VS method	All		Fisher+CV		RFE+CV		LinSBL		LinSBL+Bag	
$\overline{N}_{\mathrm{v}}$	138		115.97		37.37		17.9		98.73	
Classifier	Train(%)	Test(%)	Train(%)	Test(%)	Train(%)	Test(%)	Train(%)	Test(%)	Train(%)	Test(%)
SVM	**100.00**	85.25	95.72	85.54	**100.00**	85.05	99.95	83.92	96.40	**86.91**
	±0.00	±0.69	±1.07	±1.10	±0.00	±0.70	±0.05	±0.43	±0.30	±0.71
BayLSSVM	**99.25**	86.37	94.57	86.47	98.54	85.15	96.20	86.37	96.84	**89.51**
	±0.12	±0.75	±0.52	±0.66	±0.18	±0.66	±0.25	±0.79	±0.20	±0.55
RVM	89.17	87.72	89.49	87.11	94.55	87.11	94.60	86.72	**94.67**	**89.95**
	±0.31	±0.76	±0.34	±0.61	±0.43	±0.70	±0.38	±0.76	±0.19	±0.56

of a variable can be used to rank the variables. The higher the number, the higher the rank for that variable. The variables that were selected only by chance, e.g. once or twice, should not be considered as important. For each binary classification, an occurrence matrix $\mathbf{O}$ was computed with element of $o_{vm}, v = 1, \cdots, 30$ and $m = 1, \cdots, d$. By plotting this matrix, we can have a visual assessment of the variables.

To facilitate the interpretation of the occurrence matrix, we ordered the variables and the occurrence matrix accordingly in descending order of the total number of selection occurrences ($\sum_{v=1}^{30} o_{mv}$) among the 30 trials. Then we plot the submatrix of $\mathbf{O}$ containing only the variables that had been selected at least 30 times in total. The results are shown in Fig. 1 for the leukemia data, and in Fig. 2 for the colon cancer data.

It is noteworthy that, the genes that were selected by the proposed bagged LinSBL logit models are mostly biologically interpretable. In the leukemia cancer classification, the three top ranked genes identified by our algorithm are all among the informative genes according to [1]. The highest ranked gene is zyxin (gene 4847 according to its order in the original dataset, accession no. X95735), which encodes a LIM domain protein localized at focal contacts in adherent erythroleukemia cells. CD33 (gene 1834, M23197) is the differentiation antigen encoding cell surface protein, for which monoclonal antibodies have been demonstrated to be useful in distinguishing lymphoid from myeloid lineage cells.

In the colon cancer data, the most important gene which is identified by our method corresponds to mRNA for uroguanylin precursor (gene 377, accession no. Z50753). Guanylin and uroguanylin have been recently found to be linked to colon cancer, and treatment with uroguanylin was found to have possible therapeutic significance [6][9]. The gene with the second highest rank (Gene 1772, H08393) is a collagen alpha2 (XI) chain which is involved in cell adhesion, and collagen degrading activity is part of the metastatic process for colon carcinoma cells [3][10].

For the class prediction of brain tumors, we examined the corresponding metabolites of which the magnitude

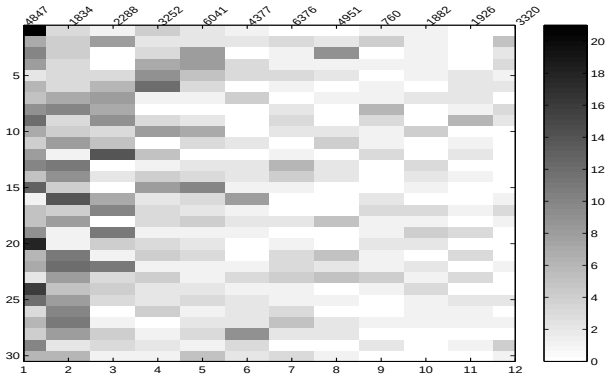


Fig. 1. Genes selected by LinSBL+Bag from the 30 realizations of the training sets for leukemia cancer microarray data. The x-axis labels in the bottom the rank of the gene, and on the top the index of the gene in the original microarray data matrix. The y-axis refers to the number of runs in the 30 randomized cross-validations. Only the genes that were selected more than 30 times in all the 30×30=900 linear SBL models are listed in the plot. The gray level in each cell corresponds to the number of occurrences that a gene was selected in bootstrapping for one realization of the training set.

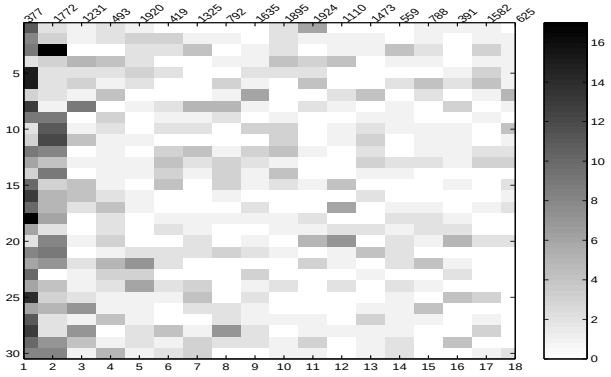


Fig. 2. Genes selected by LinSBL+Bag from the 30 realizations of the training set for colon cancer microarray data. See Fig. 1 for details.

values appeared to be significant in the pairwise binary classification problems. Fig. 3 depicts the mean spectra of the three classes, as well as the metabolites associated with their resonance frequencies [17]. The two horizontal bars, below each mean spectrum, represent the selection rate of each variable (corresponding to a frequency value) in pairwise discrimination with another tumor class. The selection rate for variable m was computed by dividing the total number of selection occurrences by $30 \times 30 = 900$. We can take the selection rate as a kind of importance measure for the variables. By examination of the figure and incorporating the domain knowledge, we were able to point out the metabolites that are important or useful for the classification. For example, in contrast to the other two classes, the astrocytomas grade II have a relatively high level[6] in the frequency regions of both total creatine (Cr) and myo-inositol (mI)/ glycine (Gly). These variables were also selected most frequently in all three pairwise binary

[6]The "level" here refers to relative intensity in the spectra as they are scaled to unit norm.

classification problems, particularly for differentiating Class 1 from 2. Indeed, in these regions the selection rate has the darkest color and reaches a value close to 0.5. To discriminate meningiomas from the aggressive class, more frequency regions are used: not only Cr and mI/Gly, but also Glutamate (Glu), Glutamine (Gln), Lipids, N-acetyl containing macromolecules (NAC). Interestingly, Cr does play a role in the maintenance of energy metabolism. While NAC resonances at the usual NAA (N-acetyl aspartate) chemical shift may appear in the solid or cystic areas of brain tumors. However one must be cautious when interpreting the selected variables for such MR spectra: the resonances at the same position may originate in different compounds depending on the tumor type. For example, the 2.03 ppm peak originates mostly from Lipids in Class 3 tumors, while it is safe to label it NAC for Class 1 and 2. It may have NAA contribution, but other N-acetyl compounds are contributing varying amounts [16]. The whole region at 2-2.6 ppm may have variable contribution from macromolecules (mostly proteins).

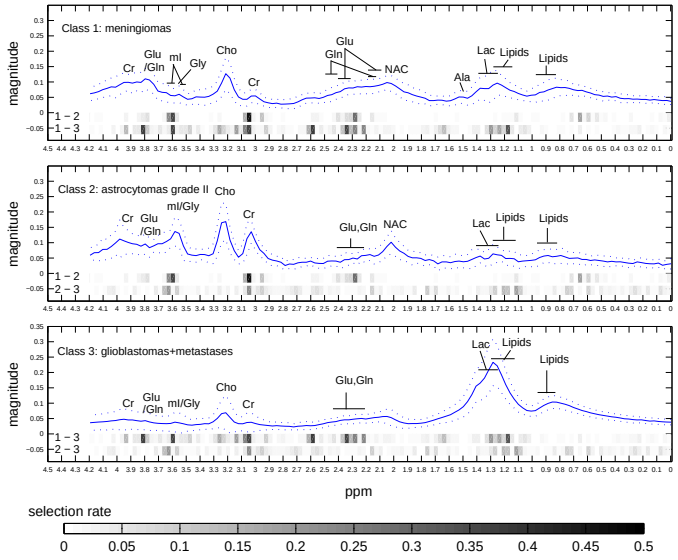


Fig. 3. Mean spectrum of the brain tumors and the selection rate of the variables using LinSBL+Bag from the 30 runs of cross-validation for the pairwise binary classification. The dotted line is the mean $\pm$ SD (standard deviation) spectrum.

VII. Discussion

From both the binary and multiclass examples, we can clearly see how bagging improves the performance of the single model generated by simply selecting one subset of variables. There is a significant gap in predictive performance between the methods of LinSBL and LinSBL+Bag in our experiments. These two sets of test performance were compared via paired t-tests for each classifier. The p-values of comparison on AUC are all $< 10^{-4}$ for the leukemia data, and all < 0.015 for the colon data. The t-tests on accuracy for the 3-class brain tumor diagnosis got a p-value $< 10^{-4}$ for each model type. This gap might result from: on the one hand the large uncertainty because of the small sample size of the training data, and on the other hand the sensitivity of SBL itself. Even for the

reference methods which are considered to be more stable than SBL, unreliability of the selected variables can still be observed from the big difference between the training and test performance.

As to the modelling techniques, the Bayesian probabilistic models performed somehow better than the standard SVMs. This might be partially due to the fact that the hyperparameters for SVMs (C, fixed to a high value of 10^6 to keep the training error low) were not optimized. Our main focus was on the models with probabilistic output which is important in biomedical diagnosis and without the burden of cross-validation for hyperparameter tuning. We have also tested the single model using the aggregated variables from the LinSBL+Bag selection, which led again to a worse performance compared to that of the bagged models with different variables.

The class priors or the cutoff values for probabilistic classification should be determined based upon the prevalence of the class in the target environment and the different misclassification costs. Practically, the uncertain cases with a posterior probability close to the cutoff values should be rejected by the classifier and referred to further examination.

To get an idea of the computational efficiency of our VS methods, we computed the average CPU time in a CV trial consumed by LinSBL+Bag on 30 bootstrap training samples. The simulations were conducted on cluster machines with Pentium processors (1GHz). Around 7 minutes and 3 minutes were needed for the leukemia data and the colon data, respectively. For prediction of brain tumors, in total 24 minutes were used for all 3 pairs of binary classification problems.

One limitation of the bagging strategy is that a single structure in the model is lost. To deal with this problem, one can adapt a similar approach as described in [13], in which the linear discriminant analysis (LDA) was bootstrapped, to generate an "average" classifier using the weighted average of the B sets of model parameters. Note that, the parameters of kernel models such as BayLSSVM and RVM are not directly ready for averaging. Whereas for linear kernel models, which is the case in this work, we could transform the model representation from kernel based to variable based, for which the parameters can be averaged and become easy to interpret. Also be aware that, unlike in [13] our bagging strategy uses different variables on different bootstrap samples, the weights for the less selected variables should be lower as well. However, the output class probability of the single "average" classifier will not be the same as the average class probability from the bagged models, which is usually more reliable and accurate than the outcome of a single model. Nevertheless, evaluation of such average models is still needed. One direction for future investigation could also be in finding a good way to transform the model aggregation into a single structure model, which is easy to explain clinically.

VIII. Conclusions

The largest potential problem for classification described here lies in the use of datasets with a small sample size and a huge dimensionality. The populations are usually underrepresented in this situation, which might result into a serious bias towards the training set, i.e. with a high training performance for a single model after variable selection and possibly a much lower generalization performance on the unseen data. This motivates the use of a bagging strategy in order to improve the reliability and lower the uncertainty in both variable selection and modelling. Experimental results confirm the advantages of the bagging strategy.

We have demonstrated that the use of the proposed variable selection can enhance the reliability of the models and thus increase the generalization performance of the models in these experiments. Furthermore, unlike the popular variable ranking methods such as RFE and the Fisher's criterion methods, the proposed method requires no additional step in order to decide on the number of variables to be used in the models. The variables are selected within a Bayesian framework, and the procedure is shown to be computationally efficient if the sample size is small. The number of occurrences of a variable being selected can serve as an importance measure for the variable. Our results imply that the linear sparse Bayesian learning plus bagging should play some role in variable selection for biomedical classification tasks.

ACKNOWLEDGMENT

The authors would like to thank Mike Tipping for providing useful comments on implementation of the sequential SBL algorithm. Use of the brain tumor data provided by the EU funded INTERPRET project (IST-1999-10310, http://carbon.uab.es/INTERPRET) is gratefully acknowledged.

REFERENCES

[1] 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. Caliguiri, C. D. Bloomeld, and E.S. Lander. Molecular classication of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, vol. 286, pp. 531-537, 1999.

[2] 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*, vol. 96, pp. 6745-6750, 1999.

[3] I. Guyon , J. Weston , S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, *Machine learning*, vol. 46, pp. 389-422, 2002.

[4] J. Weston, A. Elisseeff, M. Tipping, B. Schölkopf, Use of the zero norm with linear models and kernel methods. *Journal of Machine Learning Research*, vol. 3, pp. 1439-1461, 2002.

[5] L.M. Fu, E.S. Youn, Improving reliability of gene selection from microarray functional genomics data, *IEEE Transations on Information Technology in Biomedicine*, vol. 7, no. 3, pp. 191-196, 2003.

[6] Y. Li, C. Campbell, M. Tipping, Bayesian automatic relevance determination algorithms for classifying gene expression data. *Bioinformatics*, vol. 18, pp 1332-1339, 2002.

[7] M. Xiong, X. Fang, J. Zhao, Biomarker identification by feature wrappers. *Genome Res.*, vol. 11, pp. 1878-1887, 2001.

[8] R. Simon, M. Radmacher, K. Dobbin, L. McShane, Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst*, vol. 95, pp. 14-18, 2003.

[9] K. Shailubhai, H.H. Yu, K. Karunanandaa, J.Y. Wang, S.L. Eber, Y. Wang, N.S. Joo, H.D. Kim, B.W. Miedema and S.Z. Abbas *et al.*, Uroguanylin treatment suppress polyp formation in the $\mathrm{Apc}^{Min/+}$ mouse and induces apoptosis in human colon adenocarcinoma cells via cyclic GMP. *Cancer Res.*, vol. 60, pp. 5151-5163, 2000.

[10] G. Karakiulakis, C. Papanikolaou, S.M. Jankovic, A. Aletras, E. Papakonstantinou, E. Vretou, V. Mirtsou-Fidani, Increased type IV collagen-degrading activity in metastases originaing from priorary tumors of human colon *Invasion and metastasis*, vol. 17, no. 3, pp. 158-168, 1997.

[11] S.K. Mukherji (ed.), *Clinical Applications of Magnetic Resonance Spectroscopy*. Wiley-Liss, 1998.

[12] S.J. Nelson, Multivoxel magnetic resonance spectroscopy of brain tumors, *Molecular Cancer Therapeutics*, vol. 2, pp. 497-507, 2003.

[13] R.L. Somorjai, B. Dolenko, A. Nikulina, P. Nickersonb, D. Rushb, A. Shawa, M. Glogowskia, J. Rendella, R. Deslauriers, Distinguishing normal from rejecting renal allografts: application of a threestage classification strategy, *Vibr. Spectrosc*, vol. 28, pp. 97-102, 2002.

[14] A.E. Nikulin, B. Dolenko, T. Bezabeh, R.L. Somorjai, Near-optimal region selection for feature space reduction: novel preprocessing methods for classifying MR spectra, *NMR Biomed*, vol. 11, pp. 209-216, 1998.

[15] A. Devos, L. Lukas, J.A.K. Suykens, L. Vanhamme, A.R. Tate, F.A. Howe, C. Majós, A. Moreno-Torres, M. Van der Graaf, C. Arús, S. Van Huffel, Classification of brain tumours using short echo time ^{1}H MR spectra, *Journal of Magnetic Resonance*, 2004, in press.

[16] A.P. Candiota, C. Majós, A. bassols, M.E. Cabañas, J.J. Acebes, M.R. Quintero, C. Ar'uS, Assignment of the 2.03 ppm resonance in in vivo ^{1}H MRS of human brain tumour cystic fluid: contribution of macromolecules, *MAGMA 2004*, in press.

[17] V. Govindaraju, K. Young, and A. A. Maudsley, Proton NMR chemical shifts and coupling constants for brain metabolites, *NMR in Biomedicine*, vol. 13, pp. 129-153, 2003.

[18] M.E. Tipping, Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, vol. 1, pp. 211-244, 2001.

[19] M.E. Tipping and A. Faul, Fast marginal likelihood maximisation for sparse Bayesian models. *Proc. Artificial Intelligence and Statistics*, 2003.

[20] C.M. Bishop, and M.E. Tipping, Bayesian regression and classification. In J.A.K. Suykens *et al.* (Eds.), *Advances in Learning Theory: Methods, Models and Applications*, vol. 190, NATO Science Series III: Computer and Systems Sciences, IOS Press, pp. 267-288, 2003.

[21] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[22] J.A.K. Suykens, J. Vandewalle, Least squares support vector machine classifiers, *Neural Processing Letters*, vol. 9, no. 3, pp. 293-300, 1999.

[23] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, *Least Squares Support Vector Machines*. Singapore: World Scientific, 2002.

[24] C.M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.

[25] T. Van Gestel, J.A.K. Suykens, G. Lanckriet, A. Lambrechts, B. De Moor, J. Vandewalle, A Bayesian framework for Least Squares Support Vector Machine classifiers, Gaussian processes and kernel Fisher discriminant analysis," *Neural Computation*, vol. 14, pp. 1115-1148, 2002.

[26] T. Hastie, R. Tibshirani, Classification by pairwise coupling. In M.I. Jordan, M.J. Kearns, and S.A. Solla, editors, *Advances in Neural Information Processing Systems*, vol. 10, MIT Press, 1998.

[27] L. Breiman, Bagging predictors, *Machine Learning*, vol. 24, pp. 123-140, 1996.

[28] T. Hastie, R. Tibshirani, J. Friedman, *The elements of statistical learning - data mining, inference, and prediction*, Springer, New York, 2001.

[29] J.A. Hanley, B. McNeil, "The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve," *Radiology*, vol. 143, pp. 29-36, 1982.