Variable selection using linear sparse Bayesian models for medical classification problems

Chuan Lu, Johan A. K. Suykens, Sabine Van Huffel

Abstract— **In this work, we consider variable selection by using Tipping's fast sparse Bayesian learning method with linear basis functions. The selected variables were used in different types of probabilistic linear classifiers. We applied this method to real-life medical classification problems including two cancer diagnosis problems based on micro-array data and a brain tumor classification problem based on MRS spectra. The generalization performance of the compared models have been improved via the proposed variable selection procedure. Moreover, this algorithm appeared to be fast and efficient in dealing with very high-dimensional datasets as used in our experiments. This implies that the linear sparse Bayesian models should have some role in variable selection for biomedical classification tasks.**

Keywords— **Variable selection, Sparse Bayesian modeling, probabilistic classifiers, Micro-array, Brain tumors.**

I. Introduction

In medical classification problems, variable selection can have an impact on the economics of data acquisition and the accuracy and complexity of the classifiers, and is helpful in understanding the underlying mechanism that generated the data. In this paper, we investigate the use of sparse Bayesian learning method with linear basis functions in variable selection. The selected variables were then used in different types of probabilistic linear classifiers, including linear discriminant analysis (LDA) models, logistic regression (LR) models, relevance vector machines (RVMs) with linear kernels [1] and the Bayesian least squares support vector machines (LS-SVM) with linear kernels [3].

II. Methods

A. Sparse Bayesian modelling

In a supervised learning problem, given a set of input data $X=\{\mathbf{x_n}\}_{\mathbf{n=1}}^{\mathbf{N}}$ together with corresponding target values $T=\{t_n\}_{n=1}^{N}$, the goal is to make predictions of t for new values of $\mathbf{x}$ by using the training data. Sparse Bayesian learning is the application of Bayesian automatic relevance determination (ARD) to models linear in their parameters, by which the sparse solutions to the regression or classification tasks can be obtained [1]. The predictions are based upon some functions $y(\mathbf{x})$ defined in the input space $\mathbf{x}$: $y(\mathbf{x};\mathbf{w})=\sum_{m=0}^{M} w_m\phi_m(\mathbf{x})=\mathbf{w^T}\phi(\mathbf{x})$.

We consider two forms for the basis functions $\phi_m(\mathbf{x})$, namely $\phi_m=x_m$ (the original input variable), and $\phi_m=K(\mathbf{x},\mathbf{x_m})$, where $K(.,.)$ denotes some symmetric kernel functions. Support vector machines (SVMs) and RVMs generally adopt the kernel representation.

C. Lu, J.A.K. Suykens and S. Van Huffel are with SCD-SISTA, ESAT, Dept. of Electrical Engineering, Katholieke Universiteit Leuven (KUL), Leuven, Belgium. E-mail: {chuan.lu, johan.suykens, sabine.vanhuffel} @esat.kuleuven.ac.be .

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

$$p(T|\mathbf{w},\sigma^2)=\prod_{n=1}^{N} p(t_n|\mathbf{w},\boldsymbol{\sigma^2}), \tag{1}$$

where σ^2 is the variance of the i.i.d. noise. The parameters $\mathbf{w}$ are given a Gaussian prior $p(\mathbf{w}|\alpha)=\prod_{m=0}^{M}\mathcal{N}(w_m|0,\alpha_m^{-1})$, where $\alpha=\{\alpha_m\}$ is a vector of hyperparameters, with one α_m assigned to each model parameter w_m. This is equivalent to using a penalty function $\sum_m \log|w_m|$ in terms of regularization, with preference to a smoother model. These hyperparameters can be estimated using the framework of type II maximum likelihood in which the marginal likelihood $p(T|\mathbf{w},\boldsymbol{\sigma^2})$ is maximized with respect to α and σ^2. This optimization process can be performed efficiently using an iterative re-estimation procedure. A fast sequential learning algorithm is also available [2]. It starts from a zero basis, at each iteration step adds or deletes a basis function, or updates a hyperparameter α_m until convergence. This greedy selection procedure enables us to process the data of high dimensionality efficiently.

B. Sparse Bayesian logit model for variable selection

For binary classification problems, one can utilize the logistic function $g(y)=1/(1+e^{-y})$ [1]. The marginal likelihood is binomial $p(T|\mathbf{w},\sigma^2)=\prod_{n=1}^{N} g(y_n)^{t_n}[1-g(y_n)]^{1-t_n}$, where $t\in\{0,1\}$. There is no noise variance in this case, and a local Gaussian approximation is used to compute the posterior distribution of the weights.

The most relevant variables for the classifier can be obtained from the resulting sparse solutions, if the original variables are taken as the basis function in the linear sparse Bayesian classifier. However, we should also be aware of the uncertainty involved in the variable 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, model averaging and committee machines. Here we focus only on the selection of a single subset of variables.

III. Experiments

A. Data

Two binary cancer classification problems based on gene expression data have been considered. The first one is to distinguish two types of Leukemia: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML), based on 72 samples with 7129 gene expression values [5]. The second one is aimed to discriminate tumors from normal

tissues in the colon cancer data, which contains information of 62 tissues, each tissue associated with 2000 gene expression values [5]. All the micro-array data have been normalized to have mean zero and variance one.

The method has also been applied to a multiclass classification problem on brain tumors using the ^{1}H short echo magnetic resonance spectroscopy (MRS) spectra data. The data set consists of 205 spectra $\times$ 138 magnitude values in the frequency domain, and includes information from the four major types of brain tumors: two types of benign tumors, and two types of malignant tumors [6].

B. Experimental settings

Since the number of samples is very small compared with the dimension of the variables, we decided not to select the variables based only on one single training set. For the two binary classification problems, we first ran the variable selection algorithm 5 times on the whole dataset, resulting into 5 sets of variables. The final set of selected variables is the one associated with the highest marginal likelihood, which was then fed to various linear classifiers, including LDA, LR, Bayesian LS-SVM with linear kernels and RVM with linear kernels. The generalization performance of the models using the selected variables were evaluated by the leave-one-out (LOO) errors.

For the multiclass classification problem, we reduced the 4-class classification problem into six pairwise binary classification problems, which yielded the conditional pairwise probability estimates. The conditional probability were then coupled to obtain the joint posterior probability for each class by using Hastie's method [4]. Hence the variables used should be the union of the variables selected by the 6 binary sparse Bayesian logit models. The data set was randomly divided into two parts. The training set contains two thirds of the data and was used to build the models, while the remaining data were used for test purpose. The splitting was stratified and was repeated 30 times. For each pairwise classification, the following procedure was used. For each realization of the training data, we selected one set of variables from one sparse Bayesian logit model. The final selected variables were the most frequently selected variables from the thirty runs of selection. The number of variables was chosen to be the same as the (rounded) average number of selected variables from all the runs. Then the final selected variables were fed to all the considered linear classifiers, and the predictive power of the models given the selected variables were estimated by the averaged test accuracy from the 30 runs of randomization.

C. Results

We obtained zero LOO errors by using only 4 and 5 selected genes on 3 out of the 4 linear classifiers, for the Leukemia and colon cancer data respectively (see Table I). As to the four-class classification of brain tumors, the averaged test performance, from 30 random cross-validation (CV) trials, increases from accuracy of 68.48% to 75.34% by using variable selection for the linear LS-SVM classifier that performs best in this experiment (see Table II).

TABLE I

LOO accuracy for binary classification problems.

(a) on Leukemia cancer data

#Var	RVM	LS-SVM	LR	LDA
all:7129	0.931	0.958	NA	NA
sel:4	1	1	1	0.986

(b) on colon cancer data

#Var	RVM	LS-SVM	LR	LDA
all:2000	0.823	0.871	NA	NA
sel:5	0.984	1	1	1

Note: 'NA' stands for 'not applicable' due to numerical problems.

TABLE II

Test performance for brain tumor 4-class classification.

#Var	RVM(%)	LS-SVM(%)	LR(%)	LDA(%)
138	69.95±2.88	68.48±3.03	NA	NA
27	74.07±2.82	75.34±3.55	74.61±3.64	75.05±3.47

Note: the mean test accuracy and the standard deviation over 30 splits is given with and without variable selection. 'NA' stands for 'not applicable' due to numerical problems.

IV. Discussion and Conclusions

We have demonstrated that use of the proposed variable selection preprocessing can significantly increase the generalization performance of the models in these experiments. Remarkably, this algorithm has exhibited its efficiency in dealing with such high-dimensional datasets as used in our experiments.

Due to the small size of available samples in these problems, selection of the variable were not purely based on the training set. In this way we haven't evaluated the generalization performance of the variable selection method itself. Performance measured by using the same subset of variables for all runs of CV is somehow biased.

As future work, more experiments are to be done in order to see the characteristics of this variable selection procedure (especially when combined with bagging), and the performance when compared with the other variable selection methods.

Acknowledgments

Use of the brain tumor data provided by the EU funded INTERPRET project (IST-19999-10310, http://carbon.uab.es/INTERPRET) is gratefully acknowledged.

This research was supported by the projects of IUAP IV-02 and IUAP V-22, KUL GOA-MEFISTO-666, IDO/99/03, FWO G.0407.02 and G.0269.02.

References

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

[2] M.E. Tipping and A. Faul, Fast marginal likelihood maximisation for sparse Bayesian models. In *Proceedings of Artificial Intelligence and Statistics '03*, 2003.

[3] J.A.K. Suykens, T. Van Gestel et al., *Least Squares Support Vector Machines*. Singapore: World Scientific, 2002.

[4] T. Hastie, R. Tibshirani, Classification by pairwise coupling. In *Advances in Neural Information Processing Systems*, vol. 10, MIT Press, 1998.

[5] I. Guyon , J. Weston et al., Gene selection for cancer classification using support vector machines, *Machine learning*, 2002.

[6] L. Lukas, A. Devos et al., Classification of brain tumours using ^{1}H MRS spectra, internal report, ESAT-SISTA, K.U.Leuven, 2003.