Variable selection using linear sparse Bayesian models for medical classification problems

Chuan Lu, Johan A. K. Suykens, Sabine Van Huffel
Katholieke Universiteit Leuven - Dept. of Electrical Engineering (ESAT - SCD - SISTA)
Kasteelpark Arenberg 10, B-3001 Heverlee (Leuven), Belgium
{*chuan.lu, johan.suykens, sabine.vanhuffel*} *@esat.kuleuven.ac.be*

1 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 work, we investigate the use of Tipping's 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].

2 Methods

2.1 Sparse Bayesian modelling

- 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 $\mathbf{x}$:

$$y(\mathbf{x};\mathbf{w}) = \sum_{m=0}^{M} w_m \phi_m(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) \qquad (1)$$

- Two forms for the basis functions $\phi_m(\mathbf{x})$:
  Original input variables $\phi_m = x_m$
  Kernel basis function $\phi_m = K(\mathbf{x}, \mathbf{x}_m)$,
  where $K(.,.)$ denotes some symmetric kernel functions.
- For a regression problem,
  the likelihood of the data for a sparse Bayesian model can be expressed as:

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

  where σ^2 is the variance of the i.i.d. noise.
- The parameters $\mathbf{w}$ are given a Gaussian prior

$$p(\mathbf{w}|\alpha) = \prod_{m=0}^{M} \mathcal{N}(w_m|0,\alpha_m^{-1}), \qquad (3)$$

  where $\alpha = \{\alpha_m\}$ is a vector of hyperparameters, with a uniform prior on $\log(\alpha_\mathbf{m})$.
  $\Leftrightarrow$ using a penalty function $\sum_m \log|w_m|$ in terms of regularization, with preference to a smoother model.
- Estimate these hyperparameters using the framework of type II maximum likelihood: maximizing marginal likelihood $p(T|\mathbf{w},\sigma^2)$ 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]. The greedy selection procedure enables us to process the data of high dimensionality efficiently.

2.2 linear Sparse Bayesian logit model for variable selection

- For binary classification problems, utilize the logistic function $g(y) = 1/(1+e^{-y})$ [1]. The marginal likelihood is binomial

$$p(T|\mathbf{w},\sigma^2) = \prod_{n=1}^{N} g(y_n)^{t_n}[1-g(y_n)]^{1-t_n}, \qquad (4)$$

  where $t \in \{0,1\}$.
- 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 this 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.

- Be aware of the uncertainty involved resulting from
  – the existence of multiple solutions,
  – the sensitivity of the algorithm to small perturbations of experimental conditions.
- Attempts to tackle this problem are e.g. bagging, model averaging and committee machines.
- Here we focus only on the selection of a single subset of variables.

3 Experiments

3.1 Data

Binary cancer classification

- Based on Micro-array gene expression data
- normalized to have mean zero and variance one.

Cancer	#sample	#genes	2 classes
Leukemia	72	7129	lymphoblastic (ALL) / acute myeloid (AML)
Colon	62	2000	tumor/normal

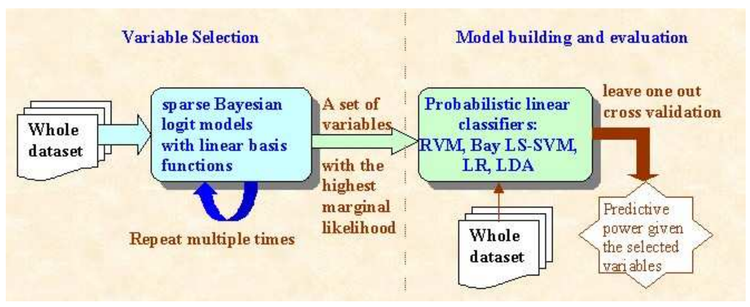
Multiclass classification of brain tumors

- Based on the ^{1}H short echo magnetic resonance spectroscopy (MRS) spectra data [6].
- Four major types of brain tumors:
  – benign (glioblastomas, metastases)
  – malignant (menigiomas, astrocytomas of grade II).
- 205 spectra $\times$ 138 L2 normalized magnitude values in the frequency domain

3.2 Experimental settings

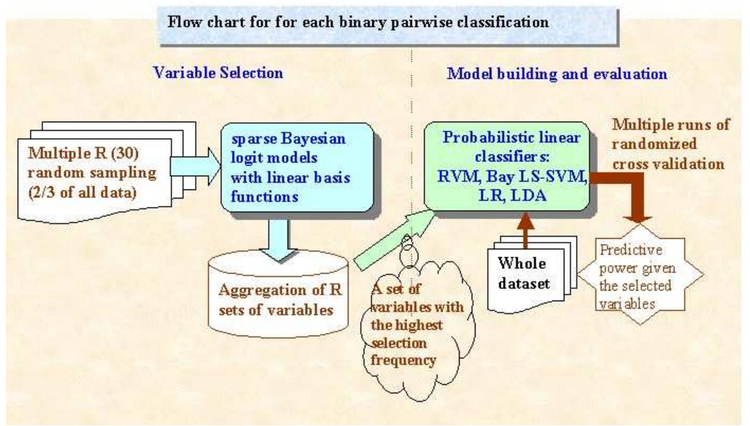
Since the number of samples is very small compared with the dimension of the variables, variable selection was not purely based on one single training set.

For the two binary classification problems



For the multiclass classification problem

- reduced the 4-class classification problem into 6 pairwise binary classification problems, which yielded the conditional pairwise probability estimates.
- coupled the conditional pairwise probability to obtain the joint posterior probability for each class by using Hastie's method [4].
- the variables used should be the union of the variables selected by the 6 binary sparse Bayesian logit models.



3.3 Results

LOO accuracy for binary classification problems.

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.

(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 available' due to numerical problems.

Test performance for 4-class brain tumor classification.

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.

#Var	RVM(%)	LS-SVM(%)	LR(%)	LDA(%)
all:138	69.95 ± 2.88	68.48 ± 3.03	NA	NA
sel: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 available' due to numerical problems.

4 Discussion and Conclusions

- ☞ Use of the proposed variable selection preprocessing can increase the generalization performance of the linear models.
- ☞ The algorithm appeared to be fast and efficient in dealing with datasets with very high dimentionality.
- ☞ The results from these experiments are somehow biased
- ☞ Future work requires more experiments in order to see
  - ◇ the characteristics of this variable selection procedure (esp. when combined with bagging)
  - ◇ 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.