Additional predictive value of microarray data compared to clinical variables

Jana Šilhavá and Pavel Smrž

Faculty of Information Technology, Brno University of Technology,
Božetěchova 2, 612 66 Brno, Czech Republic
{silhava,smrz}@fit.vutbr.cz

Abstract. Microarrays as a promising technology for prediction of cancer diagnosis have attracted attention of many researchers in recent years. Researchers often neglect clinical data used for prediction of diagnosis compared to pre-microarray era. An important problem is determination of an additional predictive value of microarray data in relation to clinical variables. We propose a new two-step method (LOG/Z+BB/X) combining logistic regression and BinomialBoosting models, to determine the additional predictive value of microarray data. LOG/Z+BB/X method is evaluated on two benchmark breast cancer datasets together with PLS-PV+RF/XZ method. The new method can combine clinical and microarray data more effectively than PLS-PV+RF/XZ method and enables simple addition of various types of data into the combined prediction.

Key words: logistic regression, boosting, generalized linear models, prediction, microarray data, clinical data, breast cancer

1 Introduction

Microarray technology [1] has attracted attention of many scientists in recent years. This technology offers new insights into biological processes. Microarray experiments are expected to contribute significantly to progress in cancer treatment by enabling precise and early diagnosis. However, the researchers do not pay attention to given clinical data used for the prediction of the diagnosis in the same manner as in pre-microarray era, often leaving clinical data unused. Clinical data typically include patient diagnoses, laboratory results, procedures and medications.

This paper relates to classification problem where the algorithm learns from samples with known class membership (training set) and establishes a prediction rule to classify new samples (test set). It is also called class prediction in literature. We focus on a binary class prediction here. We define the disease outcome as a variable that can have two values: poor prognosis or good prognosis.

An important problem is determination of an additional predictive value of microarray data (APVMD) in relation to clinical variables and a comparison of predictors. There are papers that demonstrate that microarray predictor looks

very strong, but it is not surprising as it was derived using the same set of cases that was used in model [2]. Thus, it may lead to results strongly biased in favor of the microarray predictor. There are many examples in literature that do not evaluate microarray experiments correctly [3]. There are some papers related to APVMD [2, 4] and some papers just combining microarray data with clinical data to construct a joint predictor [5, 6, 8].

The approach described in this paper deals with two problems. The first one looks for the way of determining APVMD. The second problem determines whether the microarray predictor can add some APVMD to the clinical predictor-thus if microarray data has some additional information in comparison to clinical variables.

The proposed method can construct a classifier combining both types of data. This method combines logistic regression (LOG) and BinomialBoosting (BB) [9]. LOG has been widely used with clinical data in clinical studies. The use of LOG with high-dimensional data without dimension reduction step is not suitable, because it can produce numerically unstable estimates and can not be easily generalized [10]. An approach involving LOG with highdimensional data offers BB [9]. BB algorithm is similar to LogitBoost [11], which has been successfully applied to microarray data in [12]. The characters of both aforementioned machine learning algorithms allow for their combination. The method LOG/Z+BB/X proposed in Section 2 is compared with PLS-PV+RF/XZ method [4] and evaluated on two publicly available breast cancer datasets [7, 8] in Section 3. Section 4 concludes this paper and gives directions of our future work.

2 Methods

2.1 Microarray and clinical data integration

Notation: Let Z be the $n \times q$ matrix with n samples of q clinical data. The response variable is a n-vector Y. Then let X be other matrix with microarray data. X is the $n \times p$ matrix containing n samples of expression values of p genes.

The proposed method, denoted as LOG/Z+BB/X [1], consists of the two models: logistic regression (LOG) and BinomialBoosting (BB). The integration of microarray and clinical data is at the level of predictions, see Fig. 1 (Step 1). There are some similar properties of LOG [10] and BB [9], that allow us to combine linear outputs of these models:

- generalized linear models: $Y_i = g(\eta_i)$ where g is a link function. η_i is the linear model as follows:

$$\eta_i = \beta_0 + \sum_{j=1}^{k} \beta_j Q_{i,j} \qquad \text{for} \quad i = 1, \ldots, n \quad , \tag{1}$$

[1] In the rest of this paper, a slash in a title of a method separates a model and type of data.

where β denotes coefficients, k and Q can be specified as: p and X for microarray data; q and Z for clinical data.
- response variable Y_i is considered as a binomial (Bernoulli) random variable p_i: $Y_i \sim \text{binomial}(p_i, n)$. Binomial response variables relate to logit function: $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$. Inverse logit is the link function g in LOG. In BB, logit function is included in binomial loss function as a population minimizer. BB with the componentwise linear least squares as a base procedure yields a fit of a linear logistic regression model [9].

LOG/Z+BB/X method can be described as follows, see Fig. 1. Microarray and clinical data are repeatedly split into training and test sets via Monte Carlo cross-validation (MCCV) procedure, see [14]. Each clinical training set is fitted to LOG model. Then the linear prediction of each clinical test set gives predictions η_i^Z of the linear model (1) denoted for clinical data with the upper index Z. Each microarray training set is fitted to the model using BB. The optimal number of boosting iteration is the main tuning parameter which is determined with Akaike information criterion (AIC) [15]. Then the linear prediction of each microarray test set gives predictions η_i^X of the linear model (1) denoted for microarray data with the upper index X. According to the additivity rule that is valid for linear models, we can sum up the linear predictions:

$$\eta_i = \eta_i^Z + \eta_i^X \quad . \tag{2}$$

Then the logit inversion of η_i gives a response:

$$Y_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}} \quad . \tag{3}$$

2.2 Additional predictive value of microarray data determination

The determination of APVMD involves two steps, see Fig. 1. Step 1 is a prediction of a response with LOG/Z+BB/X method that constructs a classifier combining microarray and clinical data. Step 2 is a prediction of a response with clinical data only (LOG/Z). Prediction errors are evaluated through mean error rates and the use of MCCV [14]. There is APVMD if mean error rate of Step 1 is lower than mean error rate of Step 2.

Otherwise APVMD is inadequate and it is better not to include the microarray data in the prediction of the response and use the clinical data alone.

3 Results and discussion

We performed experiments in R environment using packages 'stats' and 'mboost'. We compared Step 1 (LOG/Z+BB/X) with PLS-PV+RF/XZ [2] [4]. Both meth-

[2] The name is taken over from Boulesteix et al. [4]. The method combines Partial Least Squares (PLS) dimension reduction and the principle of pre-validation (PV) for avoiding overfitting. Random forests (RF) are then applied with both the new components and the clinical variables as predictors.

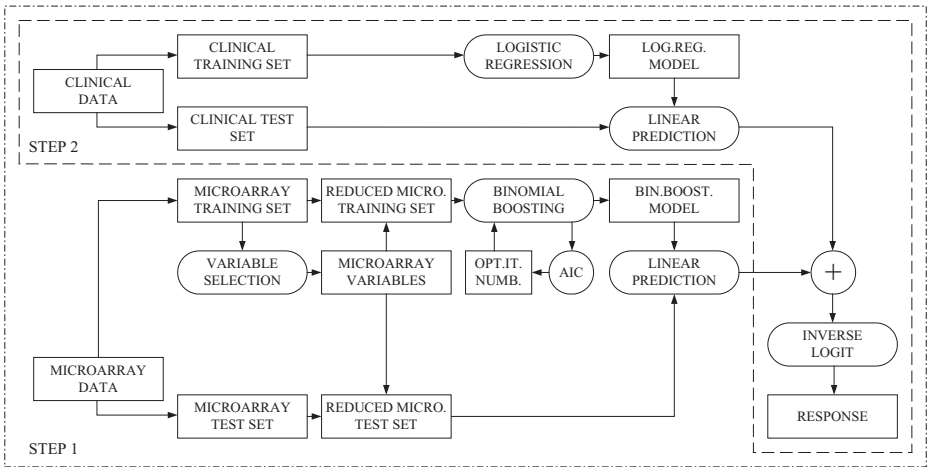

Fig. 1. Two step method for the determination of APVMD. The dot-and-dashed line denotes Step 1 (LOG/Z+BB/X). The dashed line denotes Step 2 (LOG/Z). Figure includes an evaluation schema and a variable selection part used in evaluation.

ods carry out an internal variable selection. We perform tests for different numbers of variables in order to inspect efficiency of both methods. Variables are selected on the basis of the absolute value of the t-statistic using R package 'st'.

Both methods are evaluated with two benchmark breast cancer datasets (van't Veer et al. [7]) and (Pittman et al. [8]). The first one gives the expression levels of 22483 genes for 78 breast cancer patients. According to distant metastases, 34 of these samples are classified into the poor prognosis group, the rest 44 samples belong to the the good prognosis group. The used dataset is prepared as described in [7] and is included in R package 'DENMARKLAB' [13]. This dataset includes 4348 resulting genes. Clinical variables are age, tumor grade, estrogen receptor status, progesterone receptor status, tumor size and angioinvasion. The second one gives the expression levels of 12625 genes for 158 breast cancer patients. According to recurrence of disease, 63 of these samples are classifed into the poor prognosis group, the rest 95 samples belong to the good prognosis group. The data was pre-processed using packages 'gcrma' and 'genefilter'. The genes that showed a low variablity across all samples were cleared out. The resulting dataset includes 8961 genes. Clinical variables are age, lymph node status, estrogen receptor status, family history, tumor grade and tumor size.

We compare prediction errors of the four methods in Table 1 and Table 2. According to the prediction errors in Table 1, van't Veer microarray data does not improve prediction accuracy yielded by clinical data alone. This finding coincides with conclusions in [4, 5]. The prediction errors of LOG/Z+BB/X and PLS-PV+RF/XZ methods remain more or less same for all cases of different numbers of variables. According to the prediction errors in Table 2, Pittman

Table 1. van't Veer dataset. The prediction errors (including mean error rates and standard deviations) of prediction methods outcomes evaluated over 100 MCCV iterations. p denotes a number of microarray variables.

Method	$p = 20$	$p = 50$	$p = 100$	$p = 200$	$p = all$
$LOG/Z + BB/X$	0.30 ± 0.11	0.30 ± 0.10	0.30 ± 0.11	0.29 ± 0.10	0.30 ± 0.11
$PLS - PV + RF/XZ$	0.33 ± 0.12	0.33 ± 0.12	0.32 ± 0.11	0.32 ± 0.12	0.31 ± 0.10
LOG/Z	0.28 ± 0.11	–	–	–	–
BB/X	0.37 ± 0.10	0.37 ± 0.11	0.38 ± 0.11	0.38 ± 0.10	0.40 ± 0.10

Table 2. Pittman dataset. The prediction errors (including mean error rates and standard deviations) of prediction methods outcomes evaluated over 100 MCCV iterations. p denotes a number of microarray variables.

Method	$p = 20$	$p = 50$	$p = 100$	$p = 200$	$p = all$
$LOG/Z + BB/X$	0.28 ± 0.07	0.27 ± 0.06	0.27 ± 0.06	0.26 ± 0.08	0.24 ± 0.07
$PLS - PV + RF/XZ$	0.30 ± 0.07	0.29 ± 0.08	0.30 ± 0.07	0.30 ± 0.08	0.29 ± 0.07
LOG/Z	0.35 ± 0.08	–	–	–	–
BB/X	0.29 ± 0.07	0.30 ± 0.07	0.29 ± 0.07	0.28 ± 0.07	0.27 ± 0.07

microarray data improves prediction accuracy. On the other hand, Pittman clinical data decreases prediction errors of both compared methods even if LOG/Z method produces the biggest prediction errors with this dataset. When we compare both methods (LOG/Z+BB/X and PLS-PV+RF/XZ) in Table 2, the prediction errors of LOG/Z+BB/X method have a decreasing tendency depending on the increasing number of microarray variables, while the prediction errors of PLS-PV+RF/XZ remain nearly same.

4 Conclusion and future work

In this paper, we presented a two-step method (LOG/Z+BB/X) that can determinate APVMD and offers a possible solution of construction of a classifier combining the two different types of data. The method is evaluated on two benchmark breast cancer datasets [7, 8] together with PLS-PV+RF/XZ method [4]. According to findings in Section 3, van't Veer dataset does not have APVMD, while Pittman dataset has APVMD. Pittman clinical data decreases error rates of both compared methods, even if LOG/Z method produces the biggest prediction errors on this dataset. These findings demonstrate the fact that clinical data is still a valuable data source and it should be used if available. Advantages of LOG/Z+BB/X method are that it can combine clinical and microarray data more effectively than PLS-PV+RF/XZ method and enables simple addition of more than two different types of data for increased prediction accuracy.

An extension of LOG/Z+BB/X method into the future can be intelligent decision, whether and under what conditions to include a specific data source or data source combination into prediction. We plan to test other datasets and use more data sources for evaluation.

Acknowledgments. This work was partly supported by the Czech Ministry of Education research grants 2B06052 and MSM0021630528.

References

1. Lockhart, D.J., Winzeler, E.A.: Genomics, gene expression and DNA arrays. Nature 405: 827-836 (2000)
2. Tibshirani, R., Efron, B.: Pre-validation and inference in microarrays. Statistical applications in genetics and molecular biology, 1, ISSN 1544-6115 (2002)
3. Dupuy, A., Simon, R.M.: Critical Review of Published Microarray Studies for Cancer Outcome and Guidelines on Statistical Analysis and Reporting. Journal of the National Cancer Institute, 99, 147-157, ISSN 1460-2105 (2007)
4. Boulesteix, A.L., Porzelius, Ch., Daumer, M.: Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value. Bioinformatics, 24, 1698-1706, ISSN 1460-2059 (2008)
5. Eden, P., Ritz, C., Rose, C., Ferno, M., Peterson, C.: "Good Old" clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers. Eur. J. Cancer, 40, 1837-41, ISSN 1359-6349 (2004)
6. Gevaert, O., De Smet, F., Timmerman, D., Moreau, Y., De Moor, B.: Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. Bioinformatics, 22, 184-190, ISSN 1460-2059 (2006)
7. van't Veer, L.J., Dai, H., van de Vijver, M.J., et al.: Gene expression profiling predicts clinical outcome of breast cancer. Nature 415, 530-536 (2002)
8. Pittman, J., Huang, E., Dressman, H., et al.: Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. Proc.Natl.Acad.Sci.USA, Jun 1, 101(22):8431-6 (2004)
9. Bühlmann, P., Hothorn, T.: Boosting Algorithms: Regularization, Prediction and Model Fitting. Statist. Sci., 22, 477-505, ISSN 0883-4237 (2007)
10. Hosmer, D.W. and Lemeshow, S.: Applied Logistic Regression, Second Edition. New York: Wiley, ISBN-10: 0471356328 (2000)
11. Friedman, J.: Greedy function approximation: A gradient boosting machine. Ann. Statist., 29, 1189-1232, ISSN 00905364 (2001)
12. Dettling, M., Bühlmann, P.: Boosting for tumor classification with gene expression data. Bioinformatics, 19, 1061-1069, ISSN 1367-4803 (2003)
13. Fridlyand, J., Yang, J.Y.H.: DENMARKLAB R package. Advanced microarray data analysis: Class discovery and class prediction, http://genome.cbs.dtu.dk/courses/norfa2004/Extras/DENMARKLAB.zip.
14. Molinaro, A., Simon, R., Pfeiffer, R.M.: Prediction error estimation: a comparison of resampling methods. Bioinformatics 21, 3301-3307 (2005)
15. Hothorn, T. and Bühlmann, P.: mboost: Model-Based Boosting. R package version 0.5-8 (2007) Available at http://CRAN.R-project.org/