Selekcja cech przy użyciu lasów losowych dla wieloklasowych danych mikromacierzowych

Gene Selection Using Random Forests Method For Multiclass Microarray Data

Sebastian Student, Institute of Automation, Silesian University of Technology, Gliwice

Abstract

The gene selection for classifier is a very important problem. Over the past few years many algorithms were proposed to resolve this problem. However, the optimal selection of informative genes for multiclass analysis is still an open problem. In this article the random forests method is used for gene selection. With help of variable importance returned from random forests it is possible to select significant genes. Tests on real 3-class colon cancer microarrays data show, that this method is promising in cancer research. Preliminary study of obtained gene list shows, that some of this genes are involved in neoplasia of other cancer types. The review of KEGG pathways shows, that selected genes are recognized as important in tumor grown process.

Parallel computation performed on computer cluster allowed us to work with big DNA microarray datasets and reduce the run time.

1. Introduction

Recent studies suggest that gene expression profiles may represent a promising alternative for clinical cancer classification. Molecular-based approaches has opened the possibility of investigating the activity of thousands of genes simultaneously and can be used to find genes involved in neoplasia. A big problem in applying microarrays in classification problem is dimension of this data [1]. Traditional statistical methodology for classification does not work well when there are more variables than samples. Thus, methods able to cope with the high dimensionality of the data are needed. In this paper we describe multiclass classification and dimension reduction which are intrinsically more difficult than binary ones [2]. The gene selection for classifier is a very important problem. Over the past few years many algorithms were proposed to resolve this problem. However, most of the studies are designed to binary dimension reduction problems and only a few involve multiclass cases. The optimal selection of informative genes for multiclass analysis is still an open problem. In this article the random forests [3] method is used for gene selection. There are two approaches to solve dimensionality problem of microarray data. We can obtain probably large set of genes to obtain the smallest error rate. In this case we choose genes even if they are highly correlated and perform similar functions. Second objective is identify the smallest set of genes that can still achieve good predictive performance. Finding small set of genes is very important in clinical practice as biomarkers for PCR method or to verify the biological function the selected genes and the corresponding proteins.

2. Random forests

Random forests method is an extension for bagging of decision trees. In bagging [4] successive trees do not depend on earlier trees, but each is independently constructed using a bootstrap sample of the data set. In the end, a simple majority vote is taken for prediction. Random forests add an additional layer of randomness to bagging. In standard trees, each node is split using the best split among all variables. In a random forest, each node is split using the best split among a subset of predictors randomly chosen at that node. This approach is robust against overfitting [3]. First we draw n_{tree} bootsrap samples from our data. For each of the bootstrap samples unpruned classification tree is build. Predictor in every node is choosing to obtain best split, from m_{try} randomly sample (choose from all predictors). Bagging and random variable selection results in low correlation of each tree. At the end new data are predicted by aggregating the predictions of the n_{tree} trees.

Standard way to obtain estimation of the error rate is to use data not included in the bootstrap sample at each bootstrap iteration. This data (out-of-bag data, OOB) is predicted using the tree grown with the bootstrap sample. The OBB estimate is a calculated error rate for the aggregated predictions. Random forests method can handle data with many more variables than observations. There are many advantage, but measurement of variable importance, incorporation interactions among predictor variables are the most important for extraction small sets of genes. Random forests method need to tune only three parameters. The m_{try} as the number of imput variables tried in each split, n_{tree} is the number of trees in forest and the minimum size of the terminal nodes: *nodesize*.

2. Methods

Random forest returns measures of variable importance. One of them is based on the decrease of classification accuracy when values of class labels are permuted randomly. The variable importance used here for discarding the worst genes is based as before on the decrease of classification accuracy, but in case when values of a variable in a node of tree are permuted randomly [5]. We discard those variables, that have the smallest difference between OOB error before and after randomly permutation variable values. Random forests are iteratively fitted, at each iteration eliminating the worst fraction *fraction.dropped* of variables. Gene importance ranking is generated only in first iteration. After fitting all forests, solution with the smallest number of genes and the OOB error rate within standard error of the minimum obtained error of all forests is chosen.

3. Prediction error estimation

To estimate prediction error the .632+ estimator [6] is used. It is shown that bootstrap methodology [7] gives better performance than cross-validation and resubstitution for relatively small sample microarray classification [8]. Suppose we have dataset of size l: $Z = (z_1, z_2, .., z_l)$ where $z_i = (x_i, y_i)$ is the observation. $\mathbf{X} = (x_1, x_2, .., x_l)$ is the inputs matrix and $Y = (y_1, y_2, .., y_l)$ is the response (class labels). For multiclass problem $y_i \in \{1, ., K\}$, where K is the number of classes. To divide our samples into training and test datasets bootstrap method is used. Bootstrap sample is a random sample with replacement of the observations and has the same size as our original dataset. The probes that appear in bootstrap sample compose a training dataset and the rest of observations are used as a test dataset. This is done B times to produce B bootstrap samples.

Prediction model $\hat{f}(\mathbf{X})$ has been estimated from a training sample. First we must introduce the loss function for measuring errors between Y and $\hat{f}(\mathbf{X})$ as $L(Y, \hat{f}(\mathbf{X}))$. This function returns 0 if response Y equals predicted value $\hat{f}(\mathbf{X})$ and 1 otherwise.

Now we can define the resubstitution error

$$\overline{Err}_{boot} = \frac{1}{Bl}\sum_{b=1}^{B}\sum_{i=1}^{l} L(y_i, \hat{f}^{*b}(x_i)) \qquad (1)$$

where $\hat{f}^{*b}(x_i)$ is the predicted value at x_i of the b-th bootstrap sample. This predictor can make overfitted predictions and the estimated error rate will be downward biased. That's why we obtain error estimator for test data sets

$$\overline{Err}_{test} = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{|C_b|}\sum_{i\in|C_b|} L(y_i, \hat{f}^{*b}(x_i))$$

The model trained on training set will be tested on the other samples, not used to fit the model. This provides protection against overfitting.

As we said before, we compute the error rate for B sets C_b containing samples that don't appear in b-th bootstrap sample and $|C_b|$ is number such samples. This estimator will overestimate the true prediction error, and when the test set will be small it can have high variance [8]. To resolve this problem .632+ estimator is used. This is a modified version of .632 estimator to avoid downward bias in overfitting case of our classifier. Let's define γ to be the error rate of our prediction rule if the inputs and class labels are independent. Let p_k be the observed proportion of responses y_i which equal k and q_k be the proportion of predictions $\hat{f}(x_i)$ which equal k, where k is the class label of class K. Then

$$\hat{\gamma} = \sum_{k=1}^{K} p_k(1 - q_k) \qquad (2)$$

The relative overfitting rate

$$\hat{R} = \frac{\overline{Err}_{test} - \overline{Err}_{boot}}{\hat{\gamma} - \overline{Err}_{boot}} \qquad (3)$$

Now we can define the .632+ estimator by

$$\overline{Err}_{(632+)} = (1 - \hat{w})\overline{Err}_{boot} + \hat{w}\overline{Err}_{test} \qquad (4)$$

$$\hat{w} = \frac{0.632}{1 - 0.368\hat{R}} \qquad (5)$$

When there is no overfitting problem the .632+ estimator is equal to .632 estimator

$$\hat{Err}_{(632+)} = 0368\hat{Err}_{boot} + 0632\hat{Err}_{test} \quad (6)$$

The bootstrap resampling is very computationally costly. We use the computer cluster (12 processors with HT technology) and PVM linux application.

4. Results

For test we used 3-class microarray cancer dataset with microarrays (HG-U133 Plus 2.0 oligonucleotide arrays). This dataset was obtained from MSC Cancer Center and Institute of Oncology, Warszawa. The specimens included collection of 34 normal colon tissues (class 1), 61 colon polyps (class 2), 48 colon cancers (class 3). The data was normalized with in Bioconductor and "Ferrari" reannotation procedure (environment containing the location probe set membership mapping) to remove probes matching transcripts from more than one gene and probes which do not match any transcribed sequence [9]. All computations are make with R and Bioconductor. There was only *Fraction.drooped* parameter tested in this article, because this parameter is important for gene selection. We compute 100 bootstrap iteration. In each iteration the numbers of trees used in first forest was 5000 and 2000 for all additional forests.

As we can see in Tab.1 the smallest error rate we obtain when, small groups of genes are discarded in each bootstrap iteration. The smallest mean genes number used in 100 bootstrap samples is obtained for *fraction.dropped*=0.4 (Tab.2).

Tab.1.

Classification error rate based on bootstrap resampling for different value of fraction.dropped parameter

Classification error		
Fr.dropped	**.0632**	**loo**
0.2	0.036	0.0557
0.3	0.04	0.0606
0.4	0.041	0.0633
0.5	0.0351	0.05379
0.6	0035	0,0538

Tab.2.

Number of genes used in bootstrap samples for different value of fraction.dropped

Number of genes			
Fr.dropped	**mean**	**min**	**max**
0.2	59.49	2	1637
0.3	85.1	2	3198
0.4	40.97	2	1048
0.5	95.72	2	7806
0.6	59.47	2	3198

As we said the OOB error rate is used to determine the number of genes used in each iteration. On Fig.1 we compare the OOB error for bootstrap samples and all samples. As we can see small set of genes is enough to obtain very good results. For large gene set we can observe that OOB error increase.

Fig.2 shows results of genes ranking for this case. There was relatively small number of genes that was at least one times chosen in bootstrap iteration. Only genes important for the classification were highly ranked. These genes we can choose for molecular markers to understand the basis of metastasis of various cancer. This figure can help decide on the number of genes for finding biomarkers.

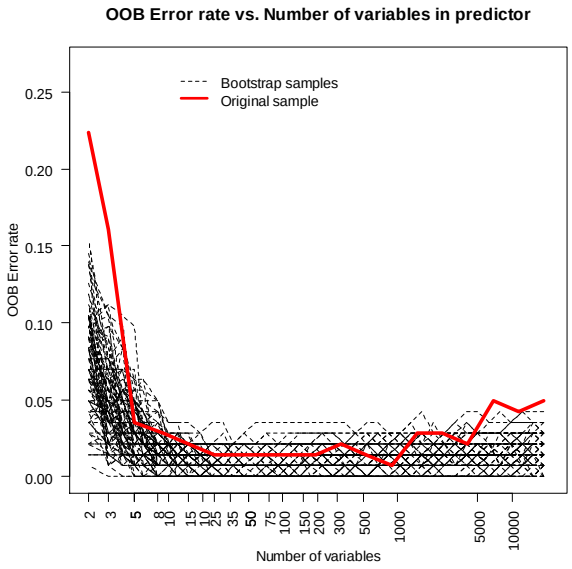


Fig.1. OOB error rate for different number of genes. (*fr.drooped*=0.4)

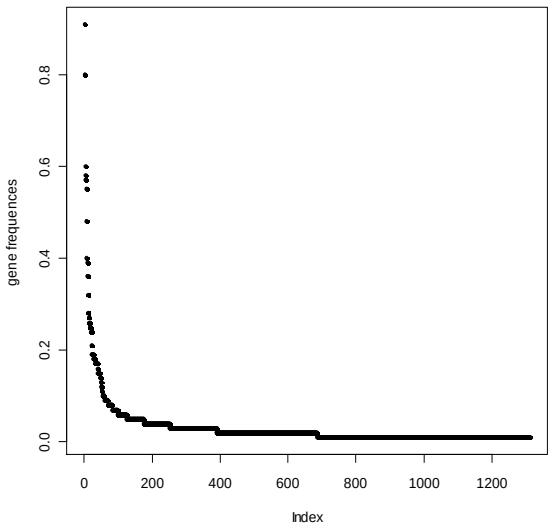


Fig.2. Results of bootstrap-based feature ranking (*fr.drooped*=0.4)

Very important factor is stability of selected genes. Comparing the probabilities value genes selected in bootstrap samples that appear in gen set

selected from whole samples set is a good measures of stability [10]. Fig.3 shows a big number of genes

form 60 genes selected form whole sample set that appears in more than 50% bootstrap iterations.

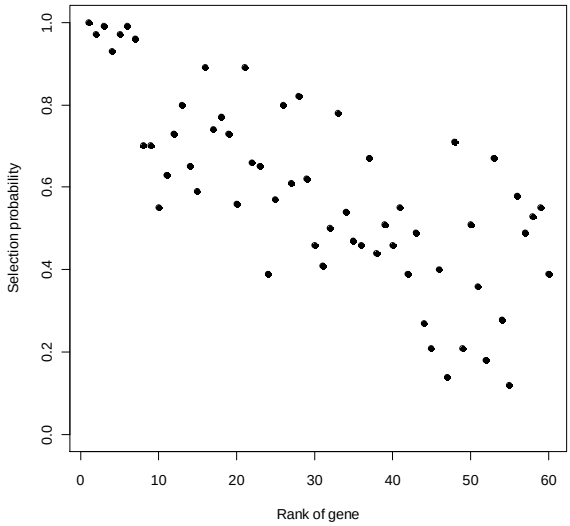


Fig.3. Selection probability in bootstrap iterations for genes selected from whole sample set (fr.drooped=0.4)

5. Conclusions

In this paper random forests method is used to find small subset of genes to differentiate different colon tissues. Our tests show, that we can obtain very good results for small number of selected genes. Preliminary review of obtained gene list shows, that some of this genes are involved in neoplasia of other cancer types. The study of KEGG pathways show, that selected genes are from pathways like TGF-beta signaling pathway, Cell adhesion molecules, Focal adhesion or Wnt signaling pathway. In the 30 mostly selected genes in bootstrap iterations are genes recognized as important in tumor grown process. Some of them are not yet recognized. This methods needs several improvements and is very promising in microarrays experiments.

The biggest problem is the computational cost of a random forests method and bootstrap resampling. Parallel computation performed on computer cluster (12 processors with HT technology) allowed us to work with big DNA microarray datasets. Presented approach makes possible to obtain more reliable classifier with less classifier error and can help to find new genes that take part in neoplasia.

5. Acknowledgments

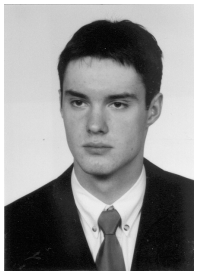
We thank prof. J. Ostrowski from Centre of Oncology, Warszawa for facilitate microarrays data.

This work was supported by Silesian University of Technology under grant BW.

References

1. Nguyen D. V., Rocke D. M.: *Tumor classification by partial least squares using microarray gene expression data*, Bioinformatics 2002, 18(1):39-50
2. Zhang T. at al.: *A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression*, Bioinformatics 2004, 20(15):2429-2437
3. Breiman L: *Random forests*, Machine Learning 2001, 45:5-32
4. Breiman L: *Random forests*, Machine Learning 1996, 24(2):123-140
5. Diaz-Uriarte Ramon, Alvarez de Andres.: Gene selection and classification of microarray data using random forest, BMC Bioinformatics 2006, 7 : 3
6. Efron B, Tibshirani R.: *Bootstrap methods: another look look at the jackknife*, Annals of Statistics, 1979, 7 : 1-26
7. Efron B.: *Improvements on cross-validation: the 632+ bootstrap method*, J. Amer. Statist. Assoc. 1997, 92 : 548-560
8. Braga-Neto U, Dougherty E. R.: *Is cross-validation valid for small-sample microarray classification?*, Bioinformatics 2004, 20(3) : 374-380
9. Ferrari F et al.:Novel definition files for human GeneChips based on GeneAnnot, BMC Bioinformatics 2007, 8 : 446
10. Pepe, M et al.: *Selecting differentially expressed genes from microarray experiments*, Biometrics 2003, 59, 133-142

Author address:



Mgr inż. Sebastian Student
Politechnika Śląska
Instytut Automatyki
Zakład Inżynierii Systemów
ul. Akademicka 16
44-100 Gliwice
tel. (032) 237 21 19
email:
sebastian.student@polsl.pl