Classification and gene selection of microarray data using AdaBoost and Neural Network

Anna Lu [1] Yang Wan [1]

[1] School of Biomedical Engineering, Drexel University, USA

Course : Bmes543
Instructor : Ahmet Sacan
Date : 2017-06-08

ABSTRACT

Gene classification is a common task in expression studies. However, univariate gene selection by ranking is inaccurate, insufficient for multiclass microarray data. AdaBoost and Neural Networks are proposed as more robust methods for microarray data classification appropriate for multi-class problems. MATLAB's built-in neural network library and AdaBoost in MATLAB's Classification Learning App are used to assess the error rate in classification of three microarray datasets: leukaemia, brain, and NCI 60 human tumor cell lines. These three microarray data sets are selected for their variable number of classes. Findings show that AdaBoost provides greater accuracy than neural network for the three selected microarray datasets. Both AdaBoost and neural network are robust to variable number of classes: two, five, and eight for leukemia, brain, and NCI 60 respectively.

1 INTRODUCTION

Microarray are a collection of DNA fragments attached to a chip. Microarray data are useful for analysis of gene expression studies. Gene selection is a common task in expression studies for identifying a small subset of genes useful for diagnostic and clinical purposes.

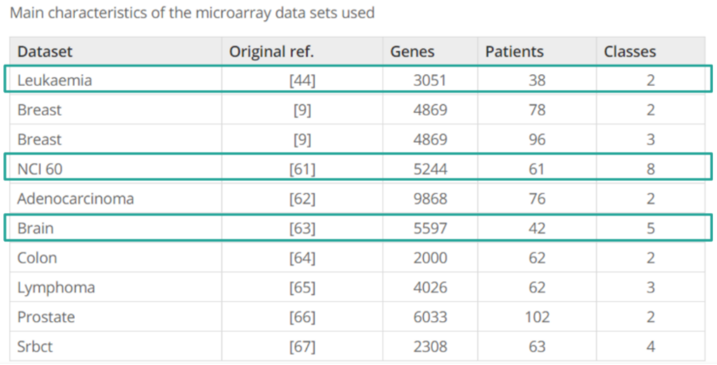
The goal of this project is to contribute comparative evaluations of two classification methods for microarray data: AdaBoost and neural networks, which were not assessed in the original publication. [1] Findings from a comprehensive evaluation across many classification methods for microarray data aids medical researchers in identifying the most accurate classification methodology for microarray data toward gene selection. Improved accuracy in microarray classification for gene selection corresponds to better diagnostic and clinical outcomes and increased efficiency of research efforts on accurately selected genes.

Standard methods for gene classification include ranking genes, principal component analysis, linear discriminant analysis, K-nearest neighbors, and support vector machines. [1, 2]
Other classification methods have been proposed for gene selection: Random forest [1] and Bayesian models [3, 4].

2 DATASET

Real microarray datasets from cancer studies were used for classification analysis. The three microarray datasets selected for classification are Leukaemia, Brain, and NCI 60 for their variable number of classes, two, five, and eight classes respectively. Providing data sets with a variable number of classes allows assessment of classifier robustness for both two class and multiclass data sets.

Table 1 Three microarray data sets used boxed in green out of the ten original publication data sets.



Main characteristics of the microarray data sets used

Dataset	Original ref.	Genes	Patients	Classes
Leukaemia	[44]	3051	38	2
Breast	[9]	4869	78	2
Breast	[9]	4869	96	3
NCI 60	[61]	5244	61	8
Adenocarcinoma	[62]	9868	76	2
Brain	[63]	5597	42	5
Colon	[64]	2000	62	2
Lymphoma	[65]	4026	62	3
Prostate	[66]	6033	102	2
Srbct	[67]	2308	63	4

3 METHODS

Method 1 AdaBoost

MATLAB's Boosted Trees package in Ensemble Learning of the Classification Learner App includes a classic AdaBoost algorithm. Each entire data set is used for AdaBoost. Workspace data was loaded into the App using rows as variables, row_1 (class labels) as the response (target), and the remaining rows as predictors (features). No cross validation was used to protect against overfitting. The following AdaBoost default parameters are used: 20 maximum splits, 30 learners, and 0.1 learning rate.

After training the model with data, a percent accuracy is calculated and reported from the classifier model's properties. Error rate is calculated as 1 – accuracy. Scatter plots of the original data and the model's predicted results are compared side-by-side.

Method 2 Neural Network

MATLAB's built-in neural network library is used. Data sets are randomly permutated, column-wise to maintain class labels along the first row. A cutoff threshold of ⅔ was selected to divide training and test data. ⅔ of the original data are used for training the network and the remaining ⅓ is used for testing the network. Error rate is calculated as an arithmetic mean from the set of differences between the actual class labels and the outputs predicted from the neural network. Default parameters are used

Both Neural Network and AdaBoost algorithms were implemented and tested using MATLAB R2016b on a 64-bit Windows 8.1 operating system with 8 GB RAM and an i7-5500 CPU.

4 EXPERIMENTS AND RESULTS

Each of the three leukaemia, NCI 60, and brain datasets were trained using both neural network and AdaBoost for a total of six experiments. Resulting prediction error rates are summarized from these six experiments and compared to literature error rates as a metric for evaluation of the classifier.

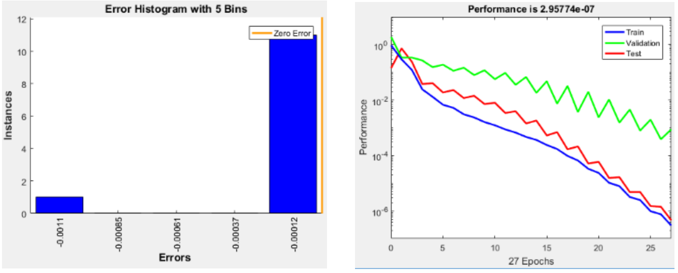


Figure 1 (Left) Leukaemia Neural Network Error Histogram with 5 Bins of tested data (n=12). Mean error rate = 0.0781. A negative skew is visible to the left of the yellow zero error marker. **(Right) Leukaemia Neural Network Performance Plot** shows that training, test, and validation performance decrease over 20 epochs.

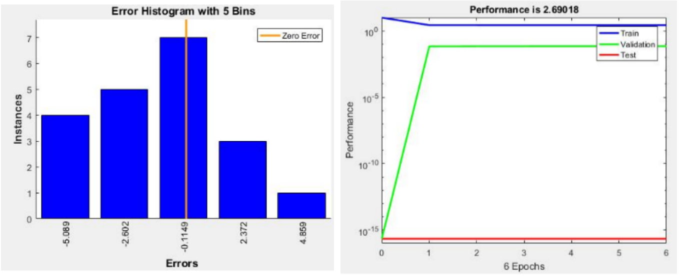


Figure 2 (Left) NCI 60 Neural Network Error Histogram with 5 Bins of tested data (n=20) centered at the zero error market. Mean error rate = 0.3258**. (Right) NCI 60 Neural Network Performance Plot** with score of 2.69018. The performance plot shows that training performance decreases, test performance increases gradually, and validation performance increases greatly at after the 20th Epoch.

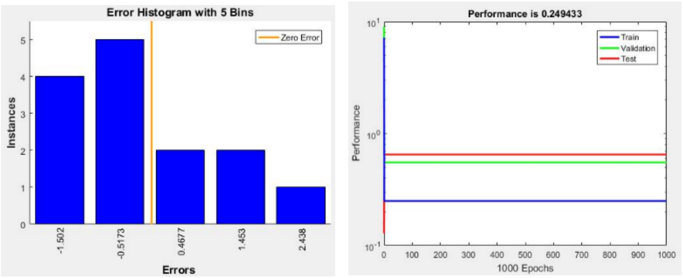


Figure 3 (Left) Brain Neural Network Error Histogram with 5 Bins of tested data (n=14). Mean error rate = 0.6331. **(Right) Brain Neural Network Performance Plot** with score of 2.71811. The performance plot shows that training and test performance decrease, and validation performance increases solely for the first of six Epochs.

AdaBoost

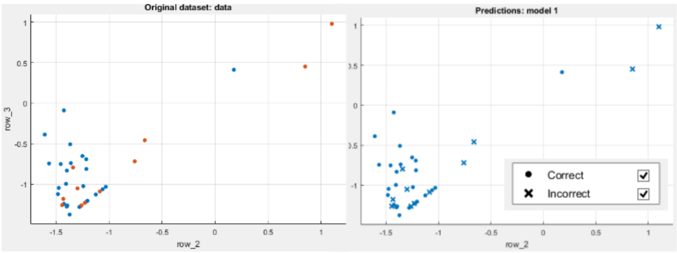


Figure 5 (Left) Leukaemia original data scatterplot shows two classes distinguished by color. **(Right) Leukaemia AdaBoost model** classifier reports an accuracy of 71.1% and an error rate of 0.289. Total training time was 12.32 seconds.

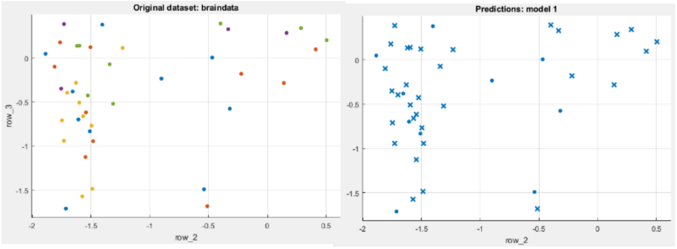


Figure 4 (Left) Brain original data scatterplot shows five classes distinguished by color. **(Right) Brain AdaBoost model** classifier reports an accuracy of 23.8% and an error rate of 0.762. Total training time was 10.113 seconds.

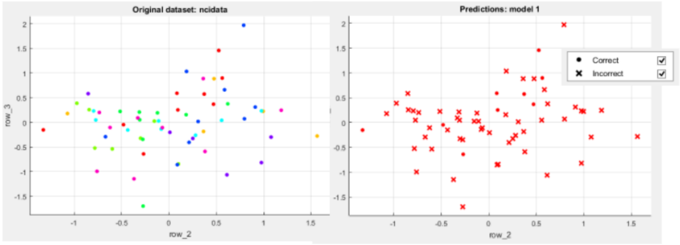


Figure 6 (Left) NCI 60 original data scatterplot shows eight classes distinguished by color. **(Right) NCI 60 AdaBoost model** classifier reports an accuracy of 14.8% and an error rate of 0.852. Total training time was 11.117 seconds.

5 DISCUSSION

Neural Network outperformed all other classifiers for two microarray datasets: leukaemia and brain with 2 and 5 classes respectively. Compared error rates to the original publication [1] on gene selection in random forest show that other methods SVM, KNN, SC, and random forest outperform AdaBoost consistently with less error rates in **Comparative error rates of AdaBoost and neural network to other classification methods. Support vector** machine (SVM), k-nearest neighbors (KNN), discrete linear discriminant analysis (DLDA), shrunken centroids (SC), nearest neighbor with variable selection (NN.vs), and random forest outperform AdaBoost and neural network with less error rates. The column 'no info' refers to minimal error, if no information is given from genes (i.e., we always bet on the most frequent class).Of interest, 'no info' matched AdaBoost error rates indicating inferior performance comparative to other classifiers.

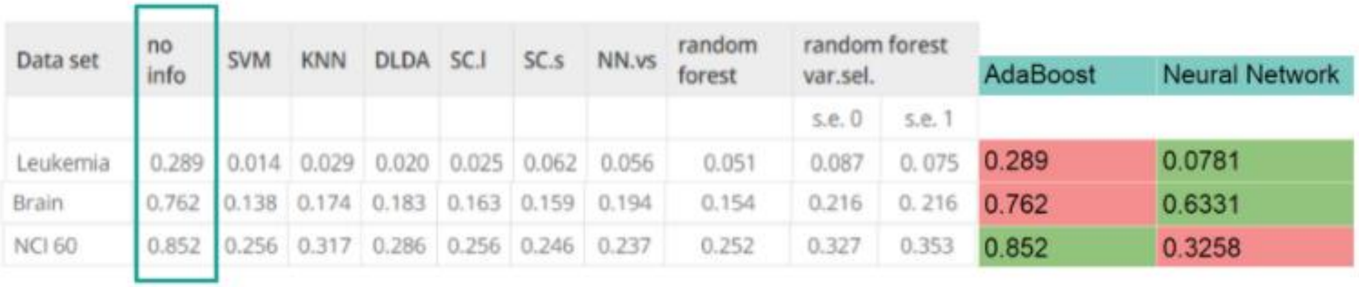
Conclusively AdaBoost is not recommended as a classifier of microarray data for gene selection. Neural network may be a candidate accurate classifier for data sets of fewer (than five) classes.

One limitation of the study is the relatively small sizes of the microarray data sets available. Patient sizes are n=38, 42, and 61 for leukaemia, brain, and NCI 60 data sets respectively. Default parameters and no cross validation are used in training both AdaBoost and neural network models.

Another limitation of both models is that a single output layer was used for multi-class data. More appropriately, accuracy may be improved by modeling AdaBoost and neural networks with as many output layers as classes. Lastly configuration of parameters such as number of hidden layers, maximum split, and learning rates enhance accuracy and multi-fold cross validation would better protect against overfitting the data.

Larger cohort cancer studies (n > 1000) can be performed by replicating the methods with increased confidence and reduced variability due to larger sample size. Additionally, feature selection and effects of parameter configurations on classification methods can be applied to both AdaBoost and neural networks. Classification robustness to demographic variety can be compared for a more comprehensive comparison across diverse cancer types, age groups, and other lifestyle factors with evidenced correlations to specific genes.

Table 2 **Comparative error rates of AdaBoost and neural network to other classification methods.** Support vector machine (SVM), k-nearest neighbors (KNN), discrete linear discriminant analysis (DLDA), shrunken centroids (SC), nearest neighbor with variable selection (NN.vs), and random forest outperform AdaBoost and neural network with less error rates. The column ‘no info’ refers to minimal error, if no information is given from genes (i.e., we always bet on the most frequent class).



Data set	no info	SVM	KNN	DLDA	SC.l	SC.s	NN.vs	random forest	random forest var.sel.		AdaBoost	Neural Network
									s.e. 0	s.e. 1		
Leukemia	0.289	0.014	0.029	0.020	0.025	0.062	0.056	0.051	0.087	0. 075	🟥 0.289	🟩 0.0781
Brain	0.762	0.138	0.174	0.183	0.163	0.159	0.194	0.154	0.216	0. 216	🟥 0.762	🟩 0.6331
NCI 60	0.852	0.256	0.317	0.286	0.256	0.246	0.237	0.252	0.327	0.353	🟩 0.852	🟥 0.3258

6 REFERENCES

1. Díaz-Uriarte R, Alvarez S: Gene Selection and classification of microarray data using random forest. *BMC Bioinformatics*, 2006. 7:3.

2. Lee JW, Lee JB, Park M, Song SH: An extensive evaluation of recent classification tools applied to microarray data. *Computation Statistics and Data Analysis* 2005, 48: 869–885.

3. Yeung KY, Bumgarner RE, Raftery AE: Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics* 2005, 21: 2394–2402.

4. Li Y, Campbell C, Tipping M: Bayesian automatic relevance determination algorithms for classifying gene expression data. *Bioinformatics* 2002, 18:1332-1339.