Malaysian Journal of Mathematical Sciences 2(2): 73-81 (2008)

Evaluation Method in Random Forest as Applied to Microarray Data

Ng Ee Ling and Yahya Abu Hasan
School of Mathematical Sciences,
Universiti Sains Malaysia,
11800 Minden, Pulau Pinang
E-mail: eeneling@gmail.com

ABSTRACT

Unlike other decision tree classifiers, Random Forest grows multiple trees which create a forest-like classification. Thus, Random Forest produces better performance as compared to that of a single tree classifier. We consider several evaluation methods which include the 10-fold cross validation, leave-one-out cross validation and bootstrap estimation. These evaluation methods are to assess the performance of the Random Forest classifier. The usage of different evaluation methods certainly shows the durability of Random Forest. To help illustrate the problem better, the four microarray datasets of binary-class and multi-class are used as experimental datasets. The evaluation method is a subjective issue and it is bound to the researcher and his study scope when selecting an evaluation method. However, we have shown that Random Forest is best evaluated using 10-fold cross validation and bootstrap estimation.

Keywords: Microarray; Random Forest; Classification; Evaluation Methods

INTRODUCTION

A DNA microarray is a collection of microscopic DNA spots attached to a solid surface, such as glass, plastic or silicon chip to form an array. Using principles of base-pairing or hybridization, DNA microarray chips can measure the expression levels of up to tens of thousands of genes simultaneously. In other words, the technology of microarray has enabled the capability to monitor the whole genome (a complete set of chromosomes with its associated genes) in a single chip [2]. There are several different types of microarrays, including short oligonucleotide arrays or better known as Affymetrix arrays (made by Affymetrix), DNA or spotted arrays (originated by Pat Brown lab at Stanford), long oligonucleotide arrays (Agilent Inkjet) and fiber-optic arrays.

The development of microarray technology has indeed help in the areas of biology and medicine especially in cancer classification. DNA

microarray data mining is very important as it can help in early detection of genes mutation, diagnosis of disease of which, if diagnosed early can help prevent death.

Cancer classification has been a popular study over the past few years and was previously done using the most traditional method which is based on the combinations of a few clinical techniques. These techniques include looking at the differences of the cell shapes and detecting enzymes that are not normally produced by certain cells. However, studies show that not one of those tests are 100% accurate and are always inconclusive [8]. The DNA microarray is a new diagnostic tool in cancer classification [8]. With the combination of mathematical modeling and biological technology, this is certainly a comprehensive way not only to classify disease but also to examine disease outcome and discover new cancer subtypes contributing to the bioinformatics field.

The contributions of good cancer classifications are certainly important to the future of medical fields as it might brighten the medical perspective in cancer treatment. Good classified models can be used to test on unseen data and in future, doctors can save time in disease diagnosis. When disease is to be diagnosed earlier, it also means that the patients' chances of survival are higher.

A DNA microarray experiment consists of the measurement of the relative representation of each mRNA in a set of biological samples. This is done using principles of base-pairing or hybridization. The collection of DNA spots is usually done on a solid surface, such as glass, plastic or silicon chip to form an array [7]. The result is the ratio of the relative abundance of genes in the experimental sample and the common reference sample. Therefore, the ratio obtained can consist of positive and negative values. The positive values indicate a higher expression in the target sample versus the reference sample and vice versa for the negative values obtained [1].

The objective of this paper is to present the durability of Random Forest in the classification of microarray data. We all know that evaluation method is used to evaluate the performance of a certain classifier. Thus, our main aim is to choose the best evaluation method for evaluating the performance of Random Forest.

A brief comparison is first made to show that creating a forest indeed improves the classification accuracy as compared to when creating just a single tree in classification. Random Forest classification offers a wide

variety of components that can be tuned to obtain optimized results. However, our main aim is to choose the best evaluation method for evaluating the performance of Random Forest.

Three evaluation methods which include the 10-fold cross validation, leave-one-out cross validation or sometimes known as LOOCV and bootstrap estimation were used to estimate the performance of Random Forest classification. Best evaluation methods are expected to give lowest error rate which is the percentage of misclassified instances or samples in the particular dataset.

Four microarray datasets with different number of classes are used to illustrate to performance of the Random Forest classification method.

In short, section 2 of this paper describes the methodology used in this study which includes the Random Forest classification method and the evaluation method. In section three, we describe the datasets used and the experiments carried out while in section 4 we publish and discuss our results. We conclude our study by re-summarizing the whole purpose of this study together with our result in section five.

METHODOLOGY

The decision tree as we know is a powerful and popular tool for classification and prediction. Decision tree is derived from the simple divide-and-conquer algorithm. However, simple as it is, these are only constructions of single classification trees.

Now, with Random Forest, we are able to compute a collection of single classification trees. This creates a forest-like classification. The basic algorithm in Random Forest works in such a way that each tree is constructed using a different bootstrap sample built from the original data. The bootstrap data points are a random sample of size n drawn with replacement from the sample $(x_1, ..., x_n)$. This means that the bootstrap data set consists of members of the original data set, some appearing zero times, some appearing once twice, etc. [4].

The bootstrap sample usually consists of about two-thirds of the data. The other one-third will then be used as the 'test' set to get the classification result. Classification is done by getting the majority vote (particular class) of

each 'test' set in a certain collection [3]. 9:1 cross validation or better known as 10-fold cross validation of which 90% of the data will be used to train leaving the other 10% to be tested. The 9:1 cross validation is also a standard choice of measuring the performances of a classifier according to [9]. The advantage of this method is that it matters less how the data is divided because each data point gets to be tested exactly once and also gets to be in a training set (*k-1*) times.

Leave-one-out cross-validation is a simple *n*-fold cross-validation, where *n* is the number of samples or instances in a dataset. Each sample in turn is left out and the learning scheme is trained on all the remaining samples. This method enables the greatest possible amount of data used for training in each case and thus ought to increase the accuracy for some classifiers. In addition, as the procedure is deterministic, no sampling is involved and thus there is no need to conduct any repetition [9].

Bootstrap estimation is another estimation method that is based on the statistical procedure of sampling with replacement. The data set is sampled *n* times to build a training set of *n* instances. However, the disadvantage is that some instances will be picked more than one time and the instances that are never picked are used for testing [9]. The whole bootstrap procedure is repeated several times, with different replacement samples for the training set and the result is averaged. This method of estimation is also the original estimation used in the Random Forest algorithm by [3].

EXPERIMENT

Datasets

The four datasets used in this study are described below.

1. Brain tumor (BRAIN)

This dataset consists of 7070 genes obtained using the Affymetrix gene chip. It has five classes (MED, EPD, MGL, RHB, JPA) and 69 samples of which 39 are MED, 10 are EPD, 7 both for each MGL and RHB and 6 for JPA.

2. Diffuse large b-cell lymphoma (DLBCL)

The DLBCL dataset contains 58 DLBCL samples and 19 FL samples. Each sample is described by 6817 genes.

3. Leukemia (LEU)

The leukemia dataset has three classes namely ALL, MLL and AML and 12584 genes. The 57 samples in the dataset are divided as such 20 are ALL and AML respectively while the other 17 is MLL.

4. Lung cancer (LUNG)

This dataset has 103 samples consisting of 39 samples of lung adenocarcinomas (ADEN), 21 samples of squamous cell lung carcinomas (SQUA), 20 samples of pulmonary carcinoids (COID), 6 samples of small-cell lung carcinomas (SCLC) and 17 normal lung samples (NORMAL).
Each sample is described by 12600 genes.

* Dataset 1 is obtained from [6] while datasets 2, 3 and 4 are obtained from [5].

Data-preprocessing

Note that the expression values in most microarray data can vary very drastically. Therefore, thresholding is essential. A standard minimum value of 20 and maximum value of 16000 are used in this study [7]. Besides that, assessing gene variability is important. This can be done by calculating their fold difference. Fold difference is the maximum value across samples divided by minimum value. Fold difference is frequently used by biologists to assess the changes of genes. Usually, genes with values of fold difference less than 2 are excluded. Carrying out this process is vital as some genes are not well expressed and do not vary sufficiently to be useful [7].

Data-processing

As microarray data usually have more variables (genes) then samples, it is significant to select only important genes to be used when doing our classification. Nevertheless, in Random Forest, the important variables are automatically selected. Thus, there is no need to do feature selection.

The first section of our experiment is done by evaluating Random Forest using the three evaluation method mentioned. The second part of our experiment includes the tuning of class-weights to optimize our results. This is because, in microarray data, classes are often not well distributed. Thus,

class weights play an important role to balance the prediction error for individual class. We shall see the effect of class weight tuning to balance the proportional of the class population given in the original data. In other words, we tune the class weights so that they become 'virtually' equal in population wise.

RESULTS AND DISCUSSION

Results

The results were run as stated in the methodology section and results shown are the error rates or percentage of incorrectly classified samples.

TABLE 1: Comparison results for singles classification trees, Random Forest evaluated using 10-fold cross validation, LOOCV and bootstrap estimation

Datasets	Single classification tree	Random Forest- 10-fold CV	Random Forest- LOOCV	Random Forest- boostrap estimation
BRAIN	15.942	2.8986	8.6957	5.797101
DLBCL	19.4805	11.6883	14.2857	14.2857
LEU	22.807	14.0351	8.7719	14.03509
LUNG	23.301	20.3883	17.4757	13.59223

The above experiment has been done without considering the class weights properties. In the second experiment, we allow the class weights to be allocated accordingly.

TABLE 2: Results after reallocating class weights

Datasets	Results
BRAIN	5.797101
DLBCL	9.0909
LEU	14.03509
LUNG	26.31259

Further explaination of results

From Table 1, we have proved that the classification results using just one single classification tree is indeed poor. Accuracies immediately improved when using the Random Forest classifier. This is because the classification is

done by taking the majority votes in the forest that are created by several trees.

The best evaluation method is selected according to that which gives least error rate. From Table 1, the 10-fold cross validation and bootstrap estimation perform very well.

Leave-one-out cross-validation (LOOCV) gives poorer results when dealing with a small number of variables (genes) as in when dealing with the BRAIN and DLBCL data. Results obtained using the bootstrap estimation is also acceptable.

Results obtained in Table 2 shows the accuracies of classification after taking into consideration the weights of the classes. This is because microarray data usually has imbalanced class distribution that is one class might be larger than the other. Therefore, weight-tuning might minimize the overall error rate by keeping a lower error rate for the larger class. Nevertheless, while adjusting the weights of each class can decrease the percentage of misclassified samples for the individual class, the overall error rate might increase due to this adjustment. This is because the overall error rate must be increased to get a better balance of the whole model [3].

As can be seen in Table 2, tuning the class weights worsen the result obtained when training the LUNG data but improves the overall results for the other three datasets. Thus, weight-tuning does not show vast improvement in our experiment.

We also conclude that the number of samples play an important role. When the number of samples increases, the sensitivity of the evaluation method towards the classifier also increases.

CONCLUSION

This main purpose of this study is to come out with the best evaluation method for the Random Forest classifier. Three evaluation methods proposed in this study include the 10-fold cross validation, *n*-fold cross validation and bootstrap estimation. Four microarray datasets are used to illustrate the performance of Random Forest that is being evaluated using the three different evaluation methods.

Most recommendable evaluation method (in descending order):

1) 10-fold cross validation
2) Bootstrap estimation
3) Leave-one-out cross validation

We also suggested that the sensitivity of the evaluation method changes when dealing with a different number of samples. Further research can involve more microarray datasets of various number of samples and using the best evaluation method chosen from this study which is the 10-fold cross validation, we can now look at the Random Forest classifier in more detail and see tune some of its other parameter to increase classification accuracies in microarray datasets.

REFERENCES

[1] Aas, K., *Microarray Data Mining: A Survey*, SAMBA/02/01, January 2001, http://www2.nr.no/documents/samba/research_areas/SIP/microarraysurvey.pdf (Last accessed: Aug 2006)

[2] Albert G.Most, John W.Foster, Michael P.Spector, *Microbial Physiology 4th Edition*, WileyLiss, Inc., New York, 2002

[3] Breiman, L. and Cutler, A., *Random Forests*, http://www.stat.berkeley.edu/users/breiman/RandomForests/cc_home.htm (Last accessed: Aug 2006)

[4] Efron and Tibshirani, *Bootstrap,* http://www.fon.hum.uva.nl/praat/manual/Bootstrap.html (Last accessed: August 2006)

[5] Li, J. and Liu, H., Institute for Infocomm Research, Singapore, *Kent Ridge Biomedical Data Set Repository,* http://research.i2r.a-star.edu.sg/rp/ (Last accessed: Aug 2006)

[6] Piatetsky-Shapiro, G., Parker, G., http://www.kdnuggets.com/dmcourse/data_mining_course/data/index.html (Last accessed: Aug 2006)

[7] Piatetsky-Shapiro, G., Ramaswamy, S. and Khabaza, T., *Capturing Best Practice for Microarray Gene Expression Data Analysis*, http://www.kdnuggets.com/dmcourse/data_mining_course/microarray-best-practice.pdf (Last accessed: Aug 2006)

[8] Twyman, R. (RT), Dr, Welcome Trust, *DNA arrays and cancer classification*, http://genome.wellcome.ac.uk/doc%5Fwtd020927. html (Last accessed: Aug 2006)

[9] Witten, I. H. and Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Academic Press, 2000