Computers in Biology and Medicine 40 (2010) 698–704

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

# Decision forest for classification of gene expression data

Jianping Huang [a,b], Hong Fang [c], Xiaohui Fan [a,*]

[a] Pharmaceutical Informatics Institute, College of Pharmaceutical Sciences, Zhejiang University, 388 YuHangTang Road, Hangzhou 310058, China
[b] National Center for Toxicological Research, U.S. Food and Drug Administration, 3900 NCTR Road, Jefferson, AR 72079, USA
[c] Z-Tech Corporation, An ICF International Company at NCTR/FDA, 3900 NCTR Road, Jefferson, AR 72079, USA

ARTICLE INFO

*Article history:*
Received 6 October 2009
Accepted 12 June 2010

*Keywords:*
Decision forest
Gene expression data
Classification
Microarray
Ensemble

ABSTRACT

This study attempts to propose an improved decision forest (IDF) with an integrated graphical user interface. Based on four gene expression data sets, the IDF not only outperforms the original decision forest, but also is superior or comparable to other state-of-the-art machine learning methods, especially in dealing with high dimensional data. With an integrated built-in feature selection (FS) mechanism and fewer parameters to tune, it can be trained more efficiently than methods such as support vector machine, and can be built with much fewer trees than other popular tree-based ensemble methods. Moreover, it suffers less from the curse of dimensionality.



## 1. Introduction

DNA microarray technology has been playing an important role in pharmaceutical and clinical research by allowing monitoring of genome-wide gene expression profiles simultaneously. It enables the possibility of cancer and other disease classification at the gene expression level. The improvement of classification capability of corresponding tools on the basis of gene expression data is thus necessitated by the advance of cancer diagnosis, prognosis and therapy. Although a number of classification methods have been proposed and widely applied to gene expression data, there is still no method can best fit all potential data sets derived from different experiments. Moreover, the diversity and instability of results produced by the available techniques, as have been shown on different data sets, are of great challenge for practitioners [1]. Irrespective of the sample size and the quality of the data set applied, classification capability can also be influenced remarkably by the adopted classification method and the corresponding parameters, as well as the feature selection method employed in the classifier construction [2,3]. In this study, we aimed at proposing an improved decision forest (IDF) that suffers less from the curse of dimensionality. IDF is able to achieve promising results even without additional feature selection methods, and usually contains much fewer trees when compared with other prevailing tree-based ensemble methods, e.g. random forest (RF) [4], and only three parameters need to be considered.

It is well-known for gene expression data that the number of genes (variables) far exceeds the number of samples. One of the major challenges for such classification problem, which precludes conventional statistical and machine learning methods from being widely accepted, is often termed as 'curse of dimensionality' [5]. Gene expression data acquired from different microarray experiments often comprise of thousands or ten thousands of genes, most of which are usually redundant or irrelevant to the biological/clinical objectives of interest. Classification tools that cannot identify or utilize informative genes appropriately in the classifier building process often fail with deterioration of generalization capability.

To make up this deficiency, feature selection methods were often employed. It is vital to couple them with classification methods for achieving good performance during classifier construction [5,6]. Hence, an accuracy/computation tradeoff has to be taken into account when a filter or wrapper-based feature selection approach is introduced. Furthermore, the combination of different classification and feature selection methods, as well as different combination manners used can both lead to distinct results.

Alternatively, application of machine learning methods with built-in feature selection as a part of the training process in classifier construction is also appreciable. They may be more efficient in that they can avoid splitting the original training set into a learning and validation set, which is often needed in evaluating each potential subset of features generated with a separate feature selection procedure [7]. Support vector machines based on recursive feature elimination (RFE-SVM) [8] is one of such successful examples in the classification of gene expression data. Another typical example is decision tree [9,10], one of the most popular methods applied in the

* Corresponding author. Tel.: +86 571 88208596; fax: +86 571 88208428.
*E-mail address:* fanxh@zju.edu.cn (X. Fan).

0010-4825/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2010.06.004

machine learning community. Its success is due largely to its efficiency and interpretability. Major limitations of decision tree are its underlying instability and low accuracy. However, owing to the underlying unstable nature of decision tree, accuracy can often be gained by ensemble diverse trees. Therefore, decision tree is often used as the basic algorithm in the ensemble methods such as bagging [11], boosting [12] and RF, which have been widely accepted in bioinformatics [13–17]. Another advantage of tree-base method, which is contributed by its intrinsic mechanism, is that it can be manifested by using a rule-based manner through translation [14].

Decision forest (DF) [18] is another tree-based ensemble method which was initially developed in the field of chemometrics and has been widely accepted in quantitative structure-activity relationship (QSAR) modeling and regulator application. However, DF has not been fully explored in the area of bioinformatics, except examples such as [19]. This may be partly due to its deficiency of performance in comparison with other popular methods. In present study, by simplifying the construction steps and reducing some parameters of the original DF, we proposed the more concise and friendly IDF. The IDF provides promising results in discriminating gene expression data, especially for data with high dimensionality, making it more suitable for application in the field of bioinformatics. Rather than other prevailing ensemble approaches that often combine up to hundreds or even thousands of weak classifiers (e.g. bagging, boosting and RF), it is designed to combine much fewer but strong trees without scarifying the accuracy and the diversity among the component trees. We demonstrated the superior performance of IDF by applying it to a large drug-induced hepatotoxicity data set and then verifying it with three well-known cancer data sets. Interesting results were yielded when compared with other prevailing methods, including support vector machines (SVMs) [20], RF, *k*-nearest neighbor (kNN), nearest centroids (NC), bagging and boosting. A graphical user interface (GUI) was also provided. Non-expert users and those who have limited access to user-friendly tools can avoid tedious and time-consuming programming merely by point and click (the IDF can be freely downloaded from the website: http://pharminfo.zju.edu.cn/computation/df.html).

## 2. Methods

### 2.1. Improved decision forest

Ensemble methods represented one of the main directions of machining learning in the past decade [21], and are still playing an important role in classification of gene expression data. The broad spreading of ensemble methods is owing to the accuracy gained by the ensemble which is often not achievable with a single classifier. With respect to the ensemble, it is widely accepted that the individual accuracy and the diversity in their predictions (making different errors) of the base classifier are two crucial elements.

Decision tree is the most popular basic classifier used in the ensemble methods due to its underlying unstable nature. A small disturbance on a node can lead to completely different descendent sub-trees. The IDF was accordingly attempting to be built with predictive and robust trees based on this idea. The diversity among the component trees can be achieved by adopting different variables in the root node during the tree construction. Therefore, in contrast to other extensively studied tree-based ensembles (e.g. bagging, boosting, RF) that usually combine a large number of weak component trees; IDF intends to combine less strong trees.

In the original DF [18], the development of a tree model consists of two steps, tree construction and tree pruning. The final decision of DF is made by averaging the probability of all trees. More trees will be built until a misclassification criterion is matched or the maximum number of trees is reached. The extent of pruning is determined by the misclassification criterion. The features (variables, descriptors) used in the previous model are removed from the feature pool, and only the remaining features are used for the development of next tree. In this original version, at least five parameters have to determined, namely, (1) minimum number of trees for the forest model, (2) minimum number of compounds in a node, (3) minimum reduction, (4) maximum number of trees for the forest and (5) the misclassification criterion.

In the IDF, these parameters are reduced into three, namely the minimum number of leaves in a node ($L$), the maximum number of trees ($T$) to create, and a new parameter ($R$) indicating the times of a feature that can be used in the forest. Unlike the DF that the use of every feature in the forest is limited to only once, the IDF allows a feature to present in the forest for more than once (not more than $R$ times) so that those most informative features may contribute more to the whole forest. However, repeatedly using the same features can run into the risk of the concurrence of identical trees. Thus, some measures should be taken into account in the construction process. As stated above, since decision tree is an unstable classifier which can be completely different when subjected to a slight disturbance, we can carefully pick up different features presented in the root node of each tree to make them different. Once a feature has been used in the root node of a tree, it would not be selected as root node splitter again in other trees. At the same time, features having high correlation coefficient (e.g. $>90\%$) cannot be chosen as root node splitter simultaneously. As a result, the IDF is greatly enhanced over the original one.

The construction of a forest can be described as follows:

1. Initialization of parameters $L$, $R$ and $T$.
2. Each feature is designated with a counter (each counter is initialized with zero), to ensure that it will not be used in the forest more than $R$ times.
3. Build the root node of the first tree.
4. Once a feature is selected and used as a node splitter, increase this feature's counter (FC) by 1. If FC $\geq R$, remove this feature from the feature pool.
5. Build (split) the children nodes after a parent node finished splitting, and GOTO 4. Stop splitting if a node contains no more than $L$ leaves or all leaves are in the same class. Repeat this step until no more nodes can be split.
6. Finish building a tree. If the number of trees built is less than $T$, GOTO 7, otherwise, GOTO 8.
7. Build the root node of another tree. Select the best feature as splitter from the feature pool. However, the following features will not be considered and will be skipped: A feature that has been used as a root node splitter in other trees, or a feature that has a high correlation coefficient (e.g. $>90\%$) with another root node splitter. GOTO 4.
8. End.

The development of a tree is similar to that of the classification and regression tree (CART) [9]. However, a variant of entropy function is employed as splitting criterion, and the tree is built without pruning.

The final decision of IDF is the same as DF that made by averaging the probability of all trees. A class with a mean