A Robust Ensemble Classification Method for Microarray Data Analysis

Zhongwei Zhang[1], Jiuyong Li[2], Hong Hu[3] and Hong Zhou[4]

Abstract Apart from the dimensionality problem, the uncertainty of Microarray data quality is another major challenge of Microarray classification. Microarray data contains various levels of noise and quite often are high levels of noise, and these data lead to unreliable and low accuracy analysis as well as the high dimensionality problem. In this paper, we propose a new Microarray data classification method, based on diversified multiple trees. The new method contains features that, (1) make most use of the information from the abundant genes in the Microarray data, and (2) use a unique diversity measurement in the ensemble decision committee. The experimental results show that the proposed classification method (DMDT) and the well known method (CS4), which diversifies trees by using distinct tree roots, are more accurate on average than other well-known ensemble methods, including Bagging, Boosting and Random Forests. The experiments also indicate that using diversity measurement of DMDT improves the classification accuracy of ensemble classification on Microarray data.

1. Introduction

The primary purpose of Microarray data classification is to build a classifier from the classified historical Microarray data, and then use the classifier to classify future incoming data or predict the future trend of data. Due to the advent of DNA Microarrays technology, the vast amount of DNA microarray datasets have

[1] Zhongwei Zhang
Department of Mathematics and Computing, University of Southern Queensland, Australia, e-mail: zhongwei@usq.edu.au

[2] Jiuyong Li
School of Computer and Information Science, University of South Australia, Mawson Lakes, Adelaide, Australia, e-mail: Jiuyong.Li@unisa.edu.au

[3] Hong Hu
Office of Research and Higher Degrees, University of Southern Queensland, Australia, e-mail: huhong@usq.edu.au

[4] Hong Zhou
Faculty of Engineering and Surveying, University of Southern Queensland, Australia, e-mail: hong.zhou@usq.edu.au

been widely available for Microarray data classification. However, as a new technology, Microarrays present new statistical problems to the Microarray data classification

1. *Curse of dimensionality problem*. Microarray data contains a huge number of genes with small number of samples and his problem has prevented many existing classification systems from direct dealing with this type of databases.
2. *Robustness problem*. In addition, DNA Microarray database contains high level of noise, irrelevant and redundant data, and those data will lead unreliable and low accuracy of analysis. Most current system can not robust enough to handle these types of data properly.

Ensemble decision tree classification methods [6, 4] have shown promise for achieving higher classification accuracy than single classifier classification method, such as C4.5 [8]. The essence of ensemble methods is to create diversified classifiers in the decision committee. Aggregating decisions from diversified classifiers is an effective way to reduce bias existing in individual trees. However, if classifiers in the committee are not unique, the committee has to be very large to create certain diversity in the committee.

Up to date, all ensemble decision trees methods have kept diversity in mind [1, 9]. However, among those methods, most of them, such as Boosting and Bagging, do not guarantee that each ensemble decision tree in the committee is different from outputs, namely identical trees and overlapping genes are not prohibited from an ensemble committee. Identical trees decrease the diversity of an ensemble committee, and noise in one gene may affect a number of ensemble decision trees; the noise will ultimately affect the reliability of Microarray classification. Therefore, committees built on those methods may not be as effective as a committee that contains no identical trees and overlapping genes.

A quick fix to improve diversity in the ensemble decision tree committee is to include a set of diversified decision trees with no overlapping genes. If classifiers in the ensemble decision tree committee are not guaranteed to be different to each other, the committee must be very large, in order to create certain diversity in the committee. This behooves us to pay special attention while designing our algorithm. One concern for such a split is that it might break down some attribute combinations that are good for classification. However, an apparent benefit of such trees is that a noise attribute cannot affect more than one tree in the committee. Considering that Microarray data normally contains much noise and many missing values, the idea of using diversified trees with no overlapping genes may provide a better solution.

The rest of this chapter is organized as follows. In Section 2, we discuss the measurement of diversity. In Section 3, we introduce our diversified multiple decision tree algorithm (DMDT). In Section 4, we show our experimental results. In Section 5, we conclude the paper.

2. Measurement of diversity

All ensemble decision tree classification methods generate a set of decision trees to form a committee. Due to the different approaches applied to generate the committee, the decision trees in the final ensemble committee could be diverse from each other in certain ways. In the past decades, measuring diversity has become a very important issue in the research of Microarray ensemble classification methods [1, 5, 9].

Measuring outputs is a most natural way to measure the diversity of ensemble classifiers [1]. The output from measuring the classifiers in a committee may give a result of total different, partially different or identical with each other. If the classifiers in a committee are all identical in a committee, we can say these classifiers are not diversified; if the classifiers are partially different, we can say they are diversified. When the classifiers are totally different or unique to each other, we say the classifiers are maximally diversified.

There are also many statistical diversity measures available, such as diversity of errors [1, 2, 7], and pairwise and non-pairwise diversity measures [1, 9, 5]. It is desirable if every classifier in an ensemble committee can agree on most samples which are predicted correctly. At the same time, we also expect that they do not make same incorrect predictions on testing samples. Those methods are also very important measurements of diversity, because if their errors were correlated, classification prediction would not lead to any performance gain by combining them.

The approach of measuring diversity based on statistical methods has drawbacks. There is a lack of robustness consideration in Microarray classification in terms of incorrect and missing data values. Identical trees are excluded from the ensemble committee since they are not helpful in improving the prediction accuracy of classification. However, this measurement allows overlapping genes among diversified trees. Overlapping genes are a problem for reliable Microarray data classification.

In our proposed method, diversity is measured by the difference of outputs for Microarray data classification problems. The degree of diversity is dependent on how many overlapping genes are included between the decision trees of an ensemble committee.

Definition 1(Degree of diversity) *Given a data set* D *with* n *attributes,* $A = \{att_1, \cdots, att_n\}$*;* C *is an ensemble decision tree committee with* $k(k > 1)$ *individual decision trees generated from* D*,* $C = \{c_1, \cdots, c_k\}$*;* $c_i \in C$ *and* $c_j \in C$ *are any single decision trees;* c_i *contains a set of attributes* $A_{c_i} = \{att \mid att \in A\}$ *and* $A_{c_j} = \{att \mid att \in A\}$*;*

Let $| A_{c_i} \cap A_{c_j} |$ *= the number of elements contained in* $A_{c_i} \cap A_{c_j}$*, and* $| A_{c_i} \cup A_{c_j} |$ *= the number of elements contained in* $A_{c_i} \cup A_{c_j}$*. Then the degree of diversity between* c_i *and* c_j *is*

$$DD = 1 - \frac{|A_{c_i} \cap A_{c_j}|}{|A_{c_i} \cup A_{c_j}|} \quad (0 \le DD \le 1)$$

When an ensemble committee contains only decision trees which have totally different outputs, or unique trees with no overlapping genes, we say that the ensemble committee is maximally diversified. According to Definition 1, $DD = 1$ for the unique decision trees.

Definition 2 (Unique decision trees). *c_i and c_j are called unique decision trees, if $A_{c_i} \cap A_{c_j} = \phi$.*

We say an ensemble decision tree classification method has greater diversity when its decision trees have a higher degree of different outputs with less overlapping genes. It is clear that diversified decision trees have a DD value which is between 0 and 1.

Definition 3 (Diversified decision trees). *If $A_{c_i} \ne A_{c_j}$ and $A_{c_i} \cap A_{c_j} \ne \phi$, then c_i and c_j are called diversified decision trees.*

Similarly, if all decision trees in an ensemble decision tree committee are identical, the degree of its diversity would be 0 .

Definition 4 (Identical decision trees). *We call c_i and c_j are identical decision trees, if $A_{c_i} = A_{c_j}$*

3. Diversified multiple decision trees algorithm

We design a diversified multiple decision tree (DMDT) algorithm to deal with the problems of Microarray classification, namely small samples versus high dimensions and noisy data. DMDT aims to improve the accuracy and robustness of ensemble decision tree methods. In our proposed algorithm, we avoid the overlapping genes among alternative trees during the tree construction stage. DMDT guarantees that constructed trees are truly unique and maximizes the diversity of the final classifiers. Our DMDT algorithm is presented in Algorithm 1. The DMDT algorithm consists of the following two steps:

1. Tree construction

The main idea is to construct multiple decision trees by re-sampling genes. All trees are built on all of the samples but with different sets of genes. We conduct re-sampling data in a systematic way. First, all samples with all genes are used to build the first decision tree. The decision tree is built using the C4.5 algorithm. After the decision tree is built, the used genes appearing in the decision tree are removed from the data. All samples with the remaining genes are used to build the second decision tree. Then the used genes are removed and so on. This process repeats until the number of trees reaches a preset number. As a result, all trees are unique and do not share common genes.

Algorithm 1: Diversified multiple decision trees algorithm (DMDT)

```
1. TREECONSTRUCTION ( $D$ , $\mathcal{T}$, $DD$ , $n$ )
INPUT: A Microarray dataset $D$ , the degree of diversity $DD$ and the number of trees $n$.
OUTPUT: A set of disjointed trees $\mathcal{T}$
let $\mathcal{F} = \phi$
let $DD = 1$
for $i = 0$ to $n - 1$ do
    call C4.5 to build tree $T_i$ on $D$ ;
    remove genes used in $T_i$ from $D$ ;
    $\mathcal{T} = \mathcal{T} \cup T_i$ .
endfor
Output $\mathcal{T}$;
2. CLASSIFICATION( $\mathcal{T}$, $x$ , $n$ )
INPUT: A set of trained trees $\mathcal{T}$, a test sample $x$ , and the number of trees $n$ .
OUTPUT: A class label of $x$
let $vote(i) = 0$, where $i = 1$ to $c$ = the number of classes.
for $j = 1$ to $n$ do
    let $c$ be the class outputted by $T_j$ ;
    $vote(c) = vote(c) \times accuracy(T_j)$ ;
endfor
Output $c$ that maximizes $vote(c)$ ;
```

2. Classification

Since the k th tree has only used the genes that have not been selected by the previously created $k-1$ trees, the quality of the k th tree might be decreased. To fix this problem, we take a vote approach; that is to say, the final predicted class of an unseen sample is determined by the weighted votes from all constructed trees. Each tree is given the weight of its training classification accuracy rate. When the vote is a tie, the class predicted by the first tree is preferred. Since all trees are built on the original data set, all trees are accountable on all samples. This avoids the unreliability of voting caused by sampling a small data set. Since all trees make use of different sets of genes, trees are independent. This adds another merit to this diversified committee. One gene containing noise or missing values affects only one tree, and not multiple trees. Therefore, it is expected to be more reliable in Microarray data classification where noise and missing values prevail.

4. Experimental results and discussion

In this section, we first present the accuracy of individual methods and the average prediction accuracy of the six methods [3, 4], which are all based on the ten-fold validation technique [6, 10, 11]. Table 1 shows the individual and average accuracy results of the six methods based on tenfold cross-validation method.

Table 1: Average accuracy of five datasets with six classification algorithms

Data set	C4.5	Random Forests	AdaBoost C4.5	Bagging C4.5	CS4	DMDT
Breast Cancer	62.9	61.9	61.9	66.0	68.0	64.3
Lung Cancer	95.0	98.3	96.1	97.2	98.9	98.9
Lymphoma	78.7	80.9	85.1	85.1	91.5	94.1
Leukemia	79.2	86.1	87.5	86.1	98.6	97.5
Colon	82.3	75.8	77.4	82.3	82.3	85.8
Average	79.62	80.6	81.6	83.3	87.9	88

Our DMDT outperforms other ensemble methods. For instance, compared to the single decision tree, DMDT is a more favorable ensemble method and outperforms C4.5 by *10.0%* on average.

From Table 1, we notice that CS4 also performs very well and improves the accuracy by *8.4%* on average. Random Forests, AdaboostC4.5 and BaggingC4.5 improve the accuracy on average by up to *4.3%*. More specifically,

1. Among the five ensemble methods used in our experiments, DMDT turns to be the most favorable classification algorithm with the highest accuracy, which improves the accuracy of classification on all cancer data sets by up to *26.7%*.
2. CS4 is comparable to DMDT in the test which improves the accuracy of classification on all data sets by up to *17.4%*.
3. BaggingC4.5 also outperforms C4.5 on all data sets by up to *9.6%*.
4. Random Forests improves the accuracy on lung cancer, Lymphoma, Leukemia and Prostate data sets by up to *19.1%*, but fails to improve the accuracy on breast cancer, Colon and Ovarian data sets. AdaBoostC4.5 can only improve the accuracy on Lung Cancer, Lymphoma and Leukemia and decreases the accuracy performance on the Breast Cancer and Colon data sets.

It is interesting to see that traditional ensemble decision tree algorithms do not always outperform a single tree algorithm. This is because the traditional ensemble methods assume that a training data set has a large number of samples with small numbers of attributes. As a result, the re-sampled data set is only slightly different from the original data set. The trees constructed on those re-sampled data are still reliable. However, in Microarray data analysis, the problem that we are facing is completely the opposite: a small number of samples with large numbers of attributes(genes). As most Microarray data contains less than 200 samples, a slight change of samples may cause a dramatic structural change in the training data set. The trees constructed on such unreliable data sets are more likely to lead to higher risk of the problem of unreliability. This risk affects the performance of classification. In contrast, DMDT and CS4 are designed specially for Microarray data analysis. DMDT keeps the alternative trees using all available samples in order to minimize the impact of the unreliability problem.

5. Conclusions

In this chapter, we studied the concept of diversity measurement in ensemble classifiers. We then proposed an algorithm that diversifies trees in the ensemble decision tree committee. We conducted experiments on six Microarray cancer data sets. We conclude that the proposed DMDT performs the best among all algorithms used in the experiments. DMDT is more resistant to the noise data while keep the highest classification accuracy rate. From the robustness point of view, Random Forests is comparable to DMDT and outperform than other compared algorithms. Without increase the noise data level, CS4 is comparable to DMDT. However, its performance decreases while comparing with DMDT when the noise level increases in the training and test data.

References

1. Matti Aksela and Jorma Laaksonen. Using diversity of errors for selecting members of a committee classifier. Pattern Recognition, 39(4):608–623, 2006.
2. M. Brown, W. Grundy, D. Lin, N. Cristianini, C. Sugnet, T. Furey, M. Jr, and D. Haussler. Knowledge-based analysis of Microarray gene expression data by using support vector machines. In Proc. Natl. Acad. Sci., volume 97, pages 262–267, 2000.
3. T. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, Boosting, and randomization. Machine Learning, 40(2):139–158, 1998.
4. Mordechai Gal-Or, Jerrold H. May, and William E. Spangler. Using decision tree models and diversity measures in the selection of ensemble classification models. In Nikunj C. Oza, RobiPolikar, Josef Kittler, and Fabio Roli, editors, Multiple Classifier Systems, volume 3541 of Lecture Notes in Computer Science, pages 186–195. Springer, 2005.
5. Ludmila I. Kuncheva and Christopher J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine Learning, 51(2):181–207, 2003.
6. Jinyan Li and Huiqing Liu. Ensembles of cascading trees. In ICDM, pages 585–588, 2003.
7. Derek Partridge and Wojtek Krzanowski. Distinct failure diversity in multiversion software. Technical report, Dept. Computer Science, University of Exeter, sec@dcs.exeter.ac.uk, 1999.
8. J. R. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, California, 1993.
9. Geoffrey I. Webb and Zijian Zheng. Multistrategy ensemble learning: Reducing error by combining ensemble learning techniques. IEEE Trans. Knowl. Data Eng, 16(8):980–991, 2004.
10. C. Yeang, S. Ramaswamy, P Tamayo, and et.al. Molecular classification of multiple tumor types. Bioinformatics, 17(Suppl 1):316–322, 2001.
11. Heping Zhang, Chang-Yung Yu, and Burton Singer. Cell and tumor classification using gene expression data: Construction of forests. Proceeding of the National Academy of Sciences, 100(7): 4168–4172, April 1 2003