A Maximally Diversified Multiple Decision Tree Algorithm for Microarray Data Classification

Hong Hu[1] **Jiuyong Li**[1] **Hua Wang**[1] **Grant Daggard**[2] **Mingren Shi**[1]

[1]Department of Mathematics and Computing
[2]Department of Biological and Physical Sciences
University of Southern Queensland,
Toowoomba, QLD 4350, Australia
Email: huhong@usq.edu.au

Abstract

We investigate the idea of using diversified multiple trees for Microarray data classification. We propose an algorithm of Maximally Diversified Multiple Trees (MDMT), which makes use of a set of unique trees in the decision committee. We compare MDMT with some well-known ensemble methods, namely AdaBoost, Bagging, and Random Forests. We also compare MDMT with a diversified decision tree algorithm, Cascading and Sharing trees (CS4), which forms the decision committee by using a set of trees with distinct roots. Based on seven Microarray data sets, both MDMT and CS4 are more accurate on average than AdaBoost, Bagging, and Random Forests. Based on a sign test of 95% confidence, both MDMT and CS4 perform better than majority traditional ensemble methods tested. We discuss differences between MDMT and CS4.

Keywords: ensemble classifier, diversified classifiers, decision tree, Microarray data.

1 Introduction

DNA Microarray technology provides capability to monitor the expression levels of thousands of genes at one time. Microarray data analysis offers the potential for discovering the causes of diseases, and identifying the marker genes which might be the signature of certain diseases.

In response to this potential, many Microarray classification algorithms have been proposed in the past ten years. Most of them have been adapted from data mining and machine learning methods, such as support vector machines (SVMs) (Brown, Grundy, Lin, Cristianini, Sugnet, Furey, Jr & Haussler 2000, Guyon, Weston, Barnhill & Vapnik 2002), k-nearest neighbor classifier (Yeang, Ramaswamy, Tamayo & et al. 2001), ensemble methods including Bagging and Boosting (Tan & Gibert 2003, Dietterich 2000), etc. Many researchers have focused their efforts to the study of ensemble decision tree methods (Li & Liu 2003, Tan & Gibert 2003, Dettling 2004, Zhang, Yu & Singer 2003) since they have shown promise to achieve high classification accuracy and its results are very easy to be interpreted.

Ensemble methods combine multiple classifiers (models) built on a set of re-sampled training data sets, or generated from various classification methods on a training data set. This set of classifiers form a decision committee, which classifies future coming samples. The classification of the committee can be simple vote or weighted vote of individual classifiers in the committee. We focuss on ensemble methods of combining multiple classifiers built on a set of re-sampled training data sets. 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.

A quick way to create diversity in the decision committee is to include a set of unique trees. This is a motivation of our proposed algorithm. A concern for such a split is that it might break down some attribute combinations or remove some informative genes that are good for classification. However, it is workable for Microarray data. Firstly, a Microarray data set contains a large number of genes, thousands to tens thousands, and this large number of genes can afford for the removal of small number of genes in subsequent trees. Secondly, Microarray data normally contains many noise values. It is very likely that expression levels of some genes are falsely correlated to outcomes (cancer or normal) due to noises. If those genes are repeatedly used in a decision committee, they will cause unreliable predictions in new cases. The diversified trees can avoid such problem. Thirdly, biologists are interested in gene interactions, the use of top genes by information gain ratio may lead to the discovery of trees of few genes. By removing these top genes, more gene combinations may be discovered.

CS4–cascading-and-sharing trees (Li & Liu 2003) is a diversified decision tree ensemble. CS4 selects n top genes and then builds n trees from the roots of n top genes. Apart from the root of the tree is fixed, other level of trees are constructed by using a normal tree construction method. CS4 has been shown achieving higher classification accuracy than Bagging and Boosting. It was reported that CS4 is better than other ensemble decision tree methods for Microarray data analysis. However, apart from the top level genes, other genes in the tree are shared. A number of trees may use some genes repeatedly. Thus, noise from one gene may affect most trees. Also, the performance of CS4 largely replies on the selection of top genes.

A distinction between CS4 and our proposed algorithm is that there are no common genes in our

This project was partially supported by Australian Research Council Discovery Grant DP0559090.

trees in the decision committee whereas genes in trees of CS4 are overlapping except the root genes. We will compare these two diversified decision tree approaches in this paper, and compare them with other traditional ensemble methods.

Complete-random classifiers (Liu, Ting & Fan 2005) also maximize the diversity of ensemble classifiers. Randomly generated trees may overlap, but a large number of trees, for example, thousands to ten thousands, diminish the effect of the overlaps. The results of complete random decision trees are promising too. We do not consider this diversifying approach in this paper based on efficiency consideration.

The rest of this paper is organized as follows. In section 2, we describe the related work on ensemble decision tree classification. In section 3, we introduce our maximally diversified multiple decision tree algorithm (MDMT). In section 4, we show experimental results. In Section 5, we present discussions. In section 6, we conclude the paper.

2 Related work

Bagging, Boosting and Random forests are some well-known ensemble methods in the machine learning field.

Bagging was proposed by Leo Breiman (Breiman 1996) in 1996. Bagging uses a bootstrap technique to re-sample the training data sets. Some samples may appear more than once in a data set whereas some samples do not appear. A set of alternative classifiers are generated from a set of re-sampled data sets. Each classifier will in turn assign a predicted class to an coming test sample. The final predicted class for the sample is determined by the majority vote. All classifiers have equal weights in voting.

Boosting was first developed by Freund and Schapire (Freund & Schapire 1996) in 1996. Boosting uses a re-sampling technique different from Bagging. A new training data set is generated according to its sample distribution. The first classifier is constructed from the original data set where every sample has an equal distribution ratio of 1. In the following training data sets, the distribution ratios are made different among samples. A sample distribution ratio is reduced if the sample has been correctly classified; Otherwise the ratio is kept unchanged. Samples which are misclassified often get duplicates in a re-sampled training data set. In contrast, samples which are correctly classified often do not appear in a re-sampled training data set. A weighted voting method is used in the committee decision. A higher accuracy classifier has larger weight than a lower accuracy classifier. The final verdict goes along with the largest weighted votes.

Tan and Gilbert (Tan & Gibert 2003) used Bagging and Boosting C4.5 decision trees. For Microarray data classification, the results showed that both methods outperform C4.5 single tree on some Microarray cancer data sets. Statistik and Surich developed a new BagBoosting method (Dettling 2004). Their experiments showed that BagBoosting outperforms constantly over Boosting and Bagging methods and achieved a better accuracy result on some Microarray data sets compared with some well-known single classification algorithms such as SVM and kNN.

Zhang and et al. (Zhang et al. 2003) proposed a new ensemble decision tree method called deterministic forest which was a modified version of random forests. Instead of re-sampling the training data set, this method selects a specified number of the top splits of the root node and then generates a number of alternative trees. The accuracy of results from deterministic forests are comparable to random forests.

CS4–cascading-and-sharing proposed by Jinyan Li and Huiqing Liu (Li & Liu 2003) makes use of both in their ensemble C4.5 algorithm for Microarray data classification. CS4 first uses the information gain ratio to select top n genes from the original data set. Then each of n genes in turn is used as the root node of an alternative tree of ensemble trees. Root nodes of ensemble trees are not determined by C4.5, but the remaining parts of trees are constructed by C4.5. CS4 diversifies roots of ensemble decision trees, but does not diversify all trees in the committee as our proposed algorithm.

3 Maximally diversified multiple decision tree algorithm (MDMT)

To improve the accuracy and reliability of ensemble decision tree methods for Microarray classification, we propose a new maximally diversified multiple decision tree (MDMT) method. We avoid the overlapping genes among alternative trees during the tree construction stage. MDMT guarantees that constructed trees are truly unique and maximizes the diversity of the final classifiers. By doing this, MDMT will reduce the instability caused by overlapping genes in current ensemble methods. For example, if the expression level of one gene is read wrongly, it only affects one tree and all other trees are unaffected.

MDMT algorithm consists of the following two steps:

1. Tree construction

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

Algorithm 1 Maximally diversified multiple decision tree (MDMT)

```
train($D, \mathcal{T}, n$)
  INPUT: A Microarray data set $D$, and the number of trees $n$.
  OUTPUT: A set of disjointed trees $\mathcal{T}$
  let $\mathcal{T} = \emptyset$
  for $i = 0$ to $n - 1$ do
    call c4.5 to build tree $T_i$ on $D$;
    remove genes used in $T_i$ from $D$;
    $\mathcal{T} = \mathcal{T} \cup T_i$.
  end for
  Output $\mathcal{T}$;
CLASSIFY($\mathcal{T}, x, n$)
  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($c$) + accuracy($T_j$);
  end for
  Output $c$ that maximizes vote($c$);
```

2. Classification

   Since the k-th tree can only use the genes that have not been selected by the previously created k-1 trees, the quality of k-th tree might be decreased. To avoid this problem, The final predicted class of a coming unseen sample is determined by the weighted votes from all trees. Each tree is given the weight of its training classification accuracy rate. The value of each vote is weighted by accuracy of tree making prediction. The majority vote is endorsed as the final predicted class. When the vote is tie, the class predicted by the first tree is advantaged. Since all trees are built on the original data set, all trees are accountable on all samples. This avoids unreliability of voting caused by sampling a small data set. Since all trees make use of different sets of genes, trees are independent. This brings another merit to this diversified committee. One gene containing noise or missing values only affects one tree but not multiple trees. Therefore, it is expected to be reliable in Microarray data classification where noise and missing values prevail.

The complete list of MMDT algorithm is given in Algorithm 1.

We give some explanations of the algorithms in the following.

C4.5 is itself a gene selection algorithm based on information gain ratio. Therefore, no gene selection algorithm is required. In addition, C4.5 discretizes continuous values by information gain ratio. No discretization pre-process is required for this algorithm. The algorithm works on the set of the original data set.

The input is a Microarray data set and a preset number of trees. The first tree (T_1) is constructed based on the original training data set. The second tree (T_2) is based on a re-sampled training data set where genes used in T_1 are removed. As a result, T_1 and T_2 share no common genes and hence are unique. The process repeats until the required number of trees k is generated.

4 Experimental results

To evaluate the performance of ensemble decision tree methods, Seven data sets from Kent Ridge Biological Data Set Repository (Li & Liu 2002) are selected. Table 1 shows the summary of the characters of the seven data sets. We conduct our experiments by using tenfold cross-validation on the merged original training and test data sets.

Table 1: Experimental data set details

Data set	Genes	Class	Record
Breast Cancer	24481	2	97
Lung Cancer	12533	2	181
Lymphoma	4026	2	47
Leukemia	7129	2	72
Colon	2000	2	62
Ovarian	15154	2	253
Prostate	12600	2	21

Our developed MDMT algorithm is compared with five well known single and ensemble decision tree algorithms, namely C4.5, Random Forests, AdaBoostC4.5, Baggingc4.5 and CS4. We have done our experiments with all four algorithms apart from CS4 using the Weka-3-5-2 package which is available online (`http://www.cs.waikato.ac.nz/ml/weka/`). We have done the experiments with CS4 using the software tool provided by Dr Jinyan Li and Huiqing Liu. Default settings are used for all compared ensemble methods. We were aware that the accuracy of some methods on some data sets can be improved when parameters were changed. However, it was difficult to find another uniform settings good for all data sets. Therefore, we did not change default settings since the default produced higher accuracy on average. From our experiments, we found that a large number of ensemble trees does not necessarily improve the prediction accuracy. We use C4.5 default settings for our MDMT algorithm and set the number of trees as 25 for the tenfold cross-validation test since further increasing the number of ensemble trees does not help to improve the prediction accuracy of classification.

Table 2 shows the individual and average accuracy results of the six methods based on tenfold cross-validation method.

Based on tenfold cross-validation test, our MDMT outperforms other ensemble methods. Compared to the single decision tree, MDMT is the best ensemble method and outperforms C4.5 by 10.0% on average. CS4 also performs very well and improve the accuracy on average by 8.4%. Random Forests, Adaboostc4.5 and BaggingC4.5 improves the accuracy on average by up to 4.3%. Among the five ensemble methods, MDMT is the most accurate classification algorithm and improves the accuracy of classification on all cancer data sets by up to 26.7%. CS4 is comparable to MDMT in the test and improves the accuracy of classification on all data sets by up to 17.4%. Baggingc4.5 also outperforms C4.5 on all data sets by up to 9.6%. 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 only improves the accuracy on Lung Cancer,Lymphoma and Leukemia and decreases the accuracy performance on Breast Cancer and Colon data sets.

To determine whether MDMT and CS4 significantly outperform ensemble traditional methods, we also conducted a sign test. The results are shown in Table 3. Based on a sign test of 95% confidence level, MDMT performs better than C4.5, Random Forests, AdaBoostC4.5 and BaggingC4.5. CS4 performs better than Random Forests and AdaBoostC4.5. Not enough evidence supports that CS4 is better than C4.5 and BaggingC4.5. Both MDMT and CS4 do not perform differently based on this test.

5 Discussions

Our experiments show that diversified ensemble classifiers outperform majority traditional ensemble classifiers tested. This suggests that diversity improves classification accuracy of ensemble classification. However, no evidence shows which diversified decision tree method is better between CS4 and MDMT. In this section, we discuss their relative strengths and weaknesses.

CS4 includes a set of decision trees in the decision committee with a set of distinct top genes at roots. The top genes are identified using information gain ratio in current CS4 algorithm. Apparently, other criteria can be used to find top genes too. If top genes are biologically meaningful, this algorithm is very useful for biologists. It groups genes by some informative genes and builds classifier based on meaningful gene groups. However, if the top genes are misidentified due to noise, the classifier committee is misleading. In addition, apart from the top genes, other genes in trees overlap. One noise gene may affect a number of trees.

In MDMT algorithm, a noise gene only affects one tree, and hence the MDMT should tolerate more noise than CS4 does. One concern of MDMT is that the enforcement of unique trees breaks up some gene combinations that are good for classification. However, the experimental results do not indicate that this is

Data set	C4.5	Random Forests	AdaBoostC4.5	BaggingC4.5	CS4	MDMT
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
Ovarian	95.7	94.1	95.7	97.6	99.2	96.4
Prostate	33.3	52.4	33.3	42.9	47.6	60
Average	75.3	78.5	76.7	79.6	83.7	85.3

Table 2: Average accuracy of seven data sets with six classification algorithms based on tenfold cross-validation

	C4.5	Random Forests	AdaBoostc4.5	Baggingc4.5	CS4	MDMT
MDMT	(7,0,0)	(7,0,0)	(7,0,0)	(5,2,0)	(3,3,1)	–
P-value	0.008	0.008	0.008	0.031	0.313	–
CS4	(6,0,1)	(6,1,0)	(7,0,0)	(6,0,1)	–	(3,3,1)
P-value	0.063	0.016	0.008	0.063	–	0.313

Table 3: Summary of sign test between MDMT and other classification methods. The second row summaries the pairwise comparison (higher, lower, tie) between MDMT and another classification method based on Table 2. The third rows show the P-values of the test. The same test for CS4 is listed in the next two rows.

a case. This does affect finding some combinations of highly informative genes with less informative genes. This is a minus. However, it finds some combinations of less informative genes that are missed by CS4. This is a plus. Keep in mind that many biologists believe that many "uninformative genes" play an important role in diseases. MDMT has potential for finding such genes combinations missed by CS4.

In short, CS4 is capable of finding informative genes and the combinations of informative genes with informative genes, and of informative genes with less informative genes. MDMT is capable of discovering combinations of informative genes with informative genes, and of less informative genes with less informative genes. In addition, MDMT has potential of being less sensitive to noise data than CS4. Note that informative or less informative genes may only make sense to data analyzers. For biologists, two methods use different gene sets and different combinations to equally explain a Microarray data. Both have potential to offer biologists some interesting discovery.

6 Conclusion

In this paper, we studied using diversified multiple decision trees to classify Microarray data. We proposed an algorithm that maximally diversifies trees in the ensemble decision tree committee. Trees in the committee share no common genes. Genes in trees are not randomly selected, but are chosen by C4.5 in a covering-algorithm manner. We conducted experiments on seven Microarray cancer data sets. The experimental results show that the proposed method and another existing diversified decision tree method, which diversifies trees by using distinct tree roots, are more accurate on average than other well-known ensemble methods, such as Bagging, Boosting and Random Forests. A sign test with 95% confidence shows that both diversified algorithms perform better than majority ensemble methods tested. The experiments indicate that diversity improves classification accuracy of ensemble classification on Microarray data. We discussed the relative strengths and weaknesses of both diversified ensemble classification methods.

References

Breiman, L. (1996), 'Bagging predictors', *Machine Learning* **24**(2), 123–140.

Brown, M., Grundy, W., Lin, D., Cristianini, N., Sugnet, C., Furey, T., Jr, M. & Haussler, D. (2000), Knowledge-based analysis of microarray gene expression data by using suport vector machines, *in* 'Proc. Natl. Acad. Sci.', Vol. 97, pp. 262–267.

Dettling, M. (2004), 'Bagboosting for tumor classification with gene expression data', *Bioinformatics* **20**(18), 3583–3593.

Dietterich, T. G. (2000), 'An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization', *Machine learning* **40**, 139–157.

Freund, Y. & Schapire, R. E. (1996), Experiments with a new boosting algorithm, *in* 'International Conference on Machine Learning', pp. 148–156.

Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. (2002), 'Gene selection for cancer classification using support vector machines', *Machine Learning* **46**(1-3), 389–422.

Li, J. & Liu, H. (2002), 'Kent ridge bio-medical data set repository. http://citeseer.ist.psu.edu/liu95chi.html'.

Li, J. & Liu, H. (2003), Ensembles of cascading trees, *in* 'ICDM', pp. 585–588.

Liu, F. T., Ting, K. M. & Fan, W. (2005), Maximizing tree diversity by building complete-random decision trees., *in* 'PAKDD', pp. 605–610.

Tan, A. C. & Gibert, D. (2003), 'Ensemble machine learning on gene expression data for cancer classification', *Applied Bioinformatics* **2**(3), s75–s83.

Yeang, C., Ramaswamy, S., Tamayo, P. & et al. (2001), 'Molecular classification of multiple tumor types', *Bioinformatics* **17**(Suppl 1), 316–322.

Zhang, H., Yu, C.-Y. & Singer, B. (2003), 'Cell and tumor classification using gene expression data: Construction of forests', *Proceeding of the National Academy of Sciences* **100**(7), 4168–4172.