Robustness analysis of diversified ensemble decision tree algorithms for Microarray data classification

Hong Hu[*], Jiuyong Li[†], Hua Wang[*], Grant Daggard[‡]

[*]Department of Mathematics and Computing
University of Southern Queensland
QLD 4350, Australia
Email: {huhong,hua}@usq.edu.au
[†]School of Computer and Information Science
University of South Australia,
Mawson Lakes, Adelaide, SA 5001, Australia
Email: Jiuyong.Li@unisa.edu.au
[‡]Department of Biological and Physical Sciences
University of Southern Queensland
QLD 4350, Australia
Email: grant@usq.edu.au

Abstract— **Ensemble classification methods have shown promise for achieving higher classification accuracy for Microarray data classification analysis. As noise values do exist in all Microarray data even after Microarray data preprocessing stage, robustness is therefore another very important criteria in addition to accuracy for evaluating reliable Microarray classification algorithms. In this paper, we conduct experimental comparison of our newly developed MDMT with C4.5, BaggingC4.5, AdaBoostingC4.5, Random Forest and CS4 on four Microarray cancer data sets. We test and evaluate how well a given single or ensemble classifier can tolerate noise data in unseen test data sets, particularly with increasing levels of noise. The experimental results show that MDMT tolerates the noise values in unseen test data sets better than other compared methods do, particularly with increasing levels of noise data. We observe that Random forests is comparable to MDMT in term of resistance to noise. The experimental results also show that ensemble decision tree methods tolerate the noise values better than single tree C4.5 does. We conclude that avoiding overlapping genes exist among the ensemble trees is an intuitive, simple and effective way to achieve higher degree of diversity for ensemble decision tree methods. The algorithm based on this principal is more reliable to deal with Microarray data sets with certain level of noise data.**

Keyword: Robustness, Microarry data, Ensemble decision trees, classification, diversity

I. Introduction

DNA microarray technology provides a broad snapshot of the state of a cell by measuring the expression levels of thousands of genes simultaneously. The primary purpose of Microarray data classification is to build a classifier using classified historical Microarray data, and then use the classifier to classify new incoming data or predict the future trend of data. Microarray data classification has great potential for identifying the gene signature of certain diseases, assisting medical diagnosis and many other applications associated with available Microarry data.

However, Microarray data commonly contains high level of noise. A DNA Microarray production involves several steps, such as, sample preparation, spotting samples on the chip, hybridization, results collection, etc. Unfortunately, every step can potentially bring in noise due to the quality of DNA samples, experimental set up, quality of chips, and finally the quality of reading equipment and statistical methods. Microarray data with high level of noise will mislead the Microarray data classification.

Robustness refers to the toleration of noise data and it is associated with predictions on data with noise values. A robust Microarray classification algorithm should performance accurately and reliably even with increasing levels of noise data. Hence, to increase the reliability of Microarray classification, we have to ensure that the algorithms we apply for are robust for tolerating the high level of noise. Otherwise, Microarray data classification based on Microarray data with high level of noise will lead to unreliable and low accuracy analysis.

Decision tree classification with a single classifier has been very successful in general classification problems [17], [2]. However, due to the nature of Microarray data's "curse of dimensionality" problem - huge number of genes with small number of samples, it is often difficult for single classification algorithm to predict Microarray data accurately. In addition, by using single decision tree algorithm, only a very small group of genes appears in the decision tree. This small tree will make the classification very unstable. Therefore, we need new classification algorithms which can be able to deal with noise data effectively.

Ensemble methods combine multiple classifiers 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. In the past decade, many researchers have devoted their efforts to the study of ensemble decision tree classification methods, such as Bagging [3], boosting [8], Random Forests [4] and CS4 [14], etc. Ensemble decision tree classification methods have shown promise for achieving higher classification accuracy than single classifier classification method, such as C4.5 [18].

Robustness is one of the most important criteria for judging Microarray classification algorithms due to the nature of Microarray data. In the past, Many comparisons between existing ensemble methods for Microarray classification have been carried out, but they mainly focused on the predictive accuracy [19], [4], [14]. Robustness comparisons between existing ensemble methods have been therefore ignored in most research literatures.

In this paper, we focus on the robustness comparison between existing single and ensemble decision tree methods including our newly developed Maximally diversified multiple decision tree algorithm (MDMT). We test and evaluate how well a given single or ensemble decision tree classification methods can tolerate noise values, particularly with increasing level of noise.

The rest of this paper is organized as follows. In section 2, we review the ensemble decision tree and a single decision tree classification methods. In section 3, we present the design methodology for comparing the robustness of selected single and ensemble decision tree classification methods. In section 4, we test the robustness of selected algorithms. The results are summarized into figures and tables. In section 5, we discuss the results. In section 6, we conclude the paper.

II. Ensemble classification algorithms

Many researchers in the past decade have devoted their efforts to the study of combining decision trees for Microarray classification in order to enhance the predictive power of decision tree in term of accuracy and robustness to the Microarray data analysis [3], [8], [4], [14]. The results show that the ensemble methods are more accurate than a single classification method.

Decision tree classification [17], [2] has been very popular and successful in machine learning and data mining fields in past few decades because of the accuracy and easy interpretability of decision tree classifier.

A decision tree [15] classifies records by trace them down the tree from the root to leaf nodes, which specify classes. For example, C4.5, a benchmark decision tree classification algorithms in machine learning and data mining, partitions a training data into some disjoint subsets simultaneously, based on the values of an attribute. At each step in the construction of the decision tree, C4.5 selects an attribute which separates data with the highest information gain ratio [18]. The same process is repeated on all subsets until each subset contains only one class. To simplify the decision tree, the induced decision tree is pruned using pessimistic error estimation [18].

Bagging was proposed by Leo Breiman [3] 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.

The boosting method was first developed by Freund and Schapire [8] 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 may 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.

Random decision forests ensemble decision tree methods have been researched extensively [11], [10], [4], [21]. Leo Breiman proposed a random decision forests method called Random Forests [4] in 1999. This method combines Bagging and random feature selection methods to generate multiple classifiers. First, boostrap is adapted to form a re-sampled training data set which a tree will be constructed from. During the tree construting stage, at each node, a fixed number of features is selected randomly for splitting on. Among the selected set of features, the one with higher information gain ratio is selected to split the training data set.

CS4–cascading-and-sharing proposed by Jinyan Li and Huiqing Liu [14]. 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. It was reported that CS4 is better than other ensemble decision tree methods for Microarray data analysis in term of accuracy.

We design a maximally diversified multiple decision tree (MDMT) algorithm [12] to deal with the problem of small samples versus high dimensions in Microarray data. MDMT aims to improve the accuracy and reliability of ensemble decision tree methods. In our proposed algorithm, 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 of

the 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.

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 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.

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.

III. Experimental design methodology

In addition to the accuracy of prediction , robustness is another important issue in Microarray classification. The objective of robustness analysis is to exam the reliability of a given algorithm based on noise Microarray data, particularly with increasing level of noise values.

A. Test data sets

Four data sets from the Kent Ridge Biological Data Set Repository [13] are selected. These data sets were collected from very well researched journal papers, namely ALL-AML Leukemia [20], Colon [7], Lymphoma [1] and Lung Cancer [9].

Table I shows the summary of the characteristics of the four data sets. We conduct our experiments by using tenfold cross-validation on the original and perturbed Microarray data sets.

TABLE I
Experimental data set details

Data set	Genes	Class	Record
Leukemia	7129	2	72
Colon	2000	2	62
Lymphoma	4026	2	47
Lung Cancer	12533	2	181

B. Test data preparation

Microarray data sets used for experiments may contain various amount of noise data. To be able to compare robustness between classifiers, we use a program to increase the degree of noise data both in training and test data.

A Microarray data set organizes data into columns and rows (samples). The columns contain a set of gene values and a category value. Each column contains the expression levels of a single gene for every sample. Each row in that table contains sample information about the expression levels of all genes with a consequent class. In our experiments, the Polar form of the Box-Muller transformation method [5] has been used to generate additional White Gaussian noise (n) independently to each gene in the original data set. For example, let g be a gene expression level value of gene G in the original data set. The perturbed value of g will be $g' = g + n$. n is generated for every gene value of gene G. The set of n has a mean of 0 and a variance of $d * \delta$ [16]. d represents the noise level while δ represents the variance of gene G in the original data set.

C. Softwares used for comparison

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 have migrated our MDMT in to the Weka-3-5-2 package. We 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 average prediction accuracy of classification significantly for most Microarray data sets we used. Figure 1 shows the individual and average accuracy results of the MDMT algorithm with different numbers of decision trees based on Leukemia, Colon, Lymphoma and Lung cancer data sets.

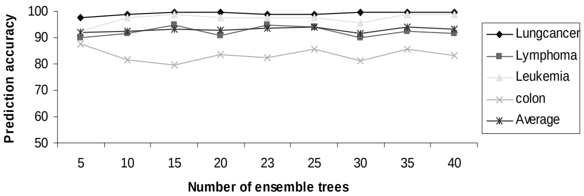

Fig. 1. Prediction accuracy vs number of ensemble trees using MDMT on Leukemia, Colon, Lymphoma and Lung cancer data sets

IV. Experimental results

Figure 2 shows the average accuracy results for the six selected algorithms over the four data sets with noise level of 0%, 20% and 60%. Table II, Table III, Table IV, Table V, Table VI and table VII show the details of the accuracy results for C4.5, Random Forests, AdaBoostC4.5, Baggingc4.5, CS4 and MDMT respectively.

From the experimental results, we have the following observations:

1) Based on the original data sets, compared to the single decision tree MDMT and CS4 are the best ensemble methods and outperforms C4.5 by up to 10.2% on average. Random Forests, Adaboostc4.5 and BaggingC4.5 improves the accuracy on average by up to 3.9%. Among the five ensemble methods, MDMT and CS4 are the most accurate classification algorithms and improve the accuracy of classification on all cancer data sets by up to 19.4%. CS4 is comparable to MDMT in the test. MDMT performs better than CS4 on Colon and Lymphoma by up to 3.2% while CS4 outperforms MDMT on Leukemia by 1.1%. And MDMT and CS4 perform equally on the Lung cancer data set. Baggingc4.5 also outperforms C4.5 on all data sets by up to 6.9%. Random Forests and AdaBoostc4.5 improve the accuracy on lung cancer, Lymphoma and Leukemia data sets by up to 8.3%, but fails to improve the accuracy on the Colon data set.

TABLE II
Prediction accuracy of C4.5 over four data sets with different level of noise values

Data set	original	20%	60%
Leukemia	79.2	70.4	66.7
Colon	82.3	72.5	47.9
Lymphoma	78.7	78.2	70.7
Lung Cancer	95.0	72.3	65.8
Average	83.8	73.3	62.8

2) With a lower noise level of 20%, MDMT and CS4 perform the best with a slightly change over the original data by up to 0.9%. BaggingC4.5 performs well with decreasing accuracy on average by 1.1%. AdaBoostC4.5 decreases the accuracy by 5.9% while single tree method C4.5 decreases the accuracy on average by 10.5%. For self comparison to the original results, Baggingc4.5 increases its accuracy on Leukemia data set by 2.2%, MDMT keeps the accuracy unchanged while all other algorithms decrease their accuracy by up to 8.8%; all algorithms increase their accuracy on Colon data set by up to 5.1% except C4.5 and CS4; Adaboostc4.5, Baggingc4.5 and MDMT increase their accuracy on Lymphoma despite C4.5, Random Forests and CS4 decreasing their performance; All algorithms decrease the accuracy on the Lung cancer data set while Adaboostc4.5 performs the worst among the ensemble

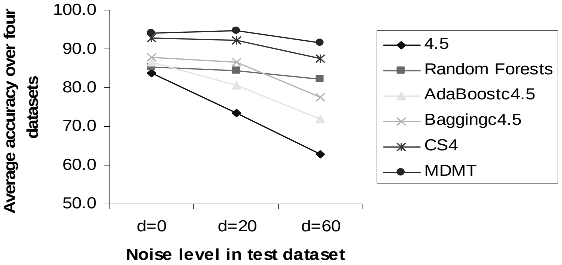

Fig. 2. Average accuracy of six algorithms over four data sets with different level of noise values

TABLE III
Prediction accuracy of Random Forests over four data sets with different level of noise values

Data set	original	20%	60%
Leukemia	86.1	80.8	83.3
Colon	75.8	80.0	76.3
Lymphoma	80.9	80.7	76.4
Lung Cancer	98.3	96.1	92.3
Average	85.3	84.4	82.0

TABLE IV
Prediction accuracy of AdaBoostc4.5 over four data sets with different level of noise values

Data set	original	20%	60%
Leukemia	87.5	80.0	76.7
Colon	77.4	82.5	64.6
Lymphoma	85.1	88.2	75.7
Lung Cancer	96.1	71.7	70.3
Average	86.5	80.6	71.8

TABLE V
Prediction accuracy of Baggingc4.5 over four data sets with different level of noise values

Data set	original	20%	60%
Leukemia	86.1	88.3	82.1
Colon	82.3	85.8	62.1
Lymphoma	85.1	85.7	84.1
Lung Cancer	97.2	86.7	81.9
Average	87.7	86.6	77.5

methods with the biggest decrease of 24.4%.

3) With a high noise level of 60%, all compared classification algorithms decrease the accuracy performance on average. Among the six methods, MDMT and Random Forests perform the best with a smallest decrease of 2.4% and 3.3% respectively. CS4, Baggingc4.5 and AdaBoostc4.5 decrease their accuracy by 5.2%, 10.2% and 14.7% respectively while C4.5 decreases the accuracy on average by 21%. For the individual results, only Random Forests increases slightly on Colon data by 0.5%.

TABLE VI

PREDICTION ACCURACY OF CS4 OVER FOUR DATA SETS WITH DIFFERENT LEVEL OF NOISE VALUES

Data set	original	20%	60%
Leukemia	98.6	95.0	89.6
Colon	82.3	81.3	70.4
Lymphoma	91.5	95.0	91.6
Lung Cancer	98.9	97.8	98.9
Average	92.8	92.3	87.6

TABLE VII

PREDICTION ACCURACY OF MDMT OVER FOUR DATA SETS WITH DIFFERENT LEVEL OF NOISE VALUES

Data set	original	20%	60%
Leukemia	97.5	97.5	92.5
Colon	85.5	87.5	82.1
Lymphoma	94.1	95.0	93.2
Lung Cancer	98.9	98.3	98.4
Average	94.0	94.6	91.6

V. DISCUSSIONS

1) Ensemble methods increase the robustness of decision tree classification. Experimental results show that ensemble decision tree methods tolerate the noise values better than single tree C4.5 does. As we know that Microarray data contains a huge number of noise values, it can be very difficult for a small tree to tolerate noise, hence it is not robust. For example, if the single tree is effected by noise, The whole classifier is effected and it leads to unreliable and lower accuracy result. In contrast, an ensemble decision tree methods contains multiple of trees, when one tree is effected by noise, other trees might not be effected at all. And the impact of noise is reduced due to the ensemble classifier voting process.

2) The robustness of Microarray classification is effected by the diversity of ensemble methods. The essence of ensemble methods is to generate diversified classifiers in the decision committee. Intuitively, if individual trees in an ensemble committee are all identical, the ensemble committee is little useful for improving the prediction performance over single decision tree algorithm.

   To increase the power of ensemble classification, ensemble decision tree algorithms must be able to generate a number of individual trees that are distinguish(diverse) to each other [6]. CS4 and MDMT are designed to guarantee diversified trees in an ensemble committee. CS4 guarantees the diversified trees by selecting distinguish 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. MDMT guarantees that constructed trees are truly unique by using disjointed genes among alternative genes. The results indicate both methods perform very good in dealing with noise data. In contrast, Bagging does not guarantee the diversity of ensemble tree. With bootstrap method, only about 2/3 of original training examples are used for constructing a individual decision tree. When re-sampled training data sets by bootstrap are identical, the decision generated from them are not diversified.

   Boosting uses the entire training data set for constructing the individual decision tree, therefore the prediction accuracy of each individual tree tends to be more accurate than Bagging. However, it still has its disadvantage for Microarray data classification. It has the same risk as Bagging in term of diversity. In addition, it is potentially not robust to noise data because its re-sample training data method. Boosting assigns more weight to the samples with higher prediction error rate. So it is the case when a sample with a higher weight contains high level of noise of genes or attributes. As we know that Microarray data contains high level of noise, as a result the re-sampled training data set contains increased noise data, and the decision tree based on such data set causes the overfit problem.

   Random Forests method combines Bagging and random feature selection methods to generate alternative classifiers. Decision trees generated by this way increase the diversity among alternative trees. It still does not guarantee that every decision tree in the committee is unique. However, due to the enormous number of genes existing in Microarray data set, Random forests gets good chance to generate higher degree of diversified trees with little or no overlapping genes among them. So Random decision forests algorithm should be more robust or more resistance to noise data than Bagging does. The results prove that MDMT, CS4 and Random Forests outperform BaggingC4.5 and BoostingC4.5.

3) Regarding the degree of diversity of ensemble decision trees, we can see from the results that MDMT and CS4 perform similar on the original test data. However, when test data contains more noise values, MDMT performs better than CS4 and other ensemble methods. In CS4 ensemble trees, apart from the top genes, other genes in trees might overlap. One noise gene may affect a number of trees. In contrast, In the MDMT algorithm, a noise gene affects only one tree, and hence MDMT should tolerate more noise than CS4 does. The results indicate that avoiding overlapping genes among the ensemble trees is an intuitive, simple and effective way to achieve a higher degree of diversity for ensemble decision tree

methods.

4) From the results, we observe that Random forests performs similar with MDMT regarding the robustness perspective. One of the possible reasons is that it is beneficial in the way it constructs the alternative trees. Unlike Bagging, Random forests constructs a tree by using random selected genes at each node. It therefore greatly increases the chance of get unique trees without overlapping genes.

VI. Conclusions

In this paper, we explored the robustness of ensemble decision tree methods. Perturbed data sets with increased noise data level were used to test the robustness of the ensemble decision trees generated from C4.5, Random Forests, AdaBoostC4.5, Baggingc4.5, CS4 and MDMT. We observed that MDMT, CS4 and Random Forests tolerate the noise values better than Baggingc4.5 and Boostingc4.5 methods do, particularly with increasing levels of noise data. Experimental results indicate that Random Forests is comparable to MDMT regarding the robustness issue and performs better than CS4 AdaBoostC4.5 and BaggingC4.5 on noise data, while CS4 is comparable to MDMT on original data sets. However, when the noise level increases in the training and test data, MDMT performs better than CS4. Experimental results also show that ensemble decision tree methods tolerate the noise values better than single tree C4.5 does.

References

[1] A. Alizadeh, M.B. Eishen, E. Davis, and C. Ma et. al. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.

[2] L. J. Breiman, R. A. Olshen, and C. J. Stone. *classification and regression trees*. Chapman and Hall, New York, 1984.

[3] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[4] Leo Breiman. Random forests–random features. Technical Report 567, University of California, Berkley, 1999.

[5] Luc Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag, New York, 1986.

[6] 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.

[7] U. Alon et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS*, 96:6745–6750, 1999.

[8] Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm. In *International Conference on Machine Learning*, pages 148–156, 1996.

[9] Gavin Gordon, Roderick Jensen, Li-Li Hsiao, and Steven Gullans et. al. Translation of microarray data into clinically relevant cancer diagnostic tests using gege expression ratios in lung cancer and mesothelioma. *Cancer Research*, 62:4963–4967, 2002.

[10] Tin Kam Ho. Random decision forests. In *ICDAR*, page 278, 1995.

[11] Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell*, 20(8):832–844, 1998.

[12] Hong Hu, Jiuyong Li, Hua Wang, Grant Daggard, and Mingren Shi. A maximally diversified multiple decision tree algorithm for microarray data classification(under review).

[13] Jingyan Li and Huiqing Liu. Kent ridge bio-medical data set repository. http://sdmc.lit.org.sg/gedatasets/datasets.html, 2002.

[14] Jinyan Li and Huiqing Liu. Ensembles of cascading trees. In *ICDM*, pages 585–588, 2003.

[15] T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[16] Krishnamurty Muralidhar and Rathindra Sarathy. Security of random data perturbation methods. *ACM Trans. Database Syst.*, 24(4):487–493, 1999.

[17] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1,1:81–106, 1986.

[18] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, California, 1993.

[19] Aik Choon Tan and David Gibert. Ensemble machine learning on gene expression data for cancer classification. *Applied Bioinformatics*, 2(3):s75–s83, 2003.

[20] T.R.Golub, D.K.Slonim, P. Tamayo, and et.al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[21] 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.