Comprehensibility of Classifiers for Future Microarray Analysis Datasets

Gregor Stiglic[1] , Simon Kocbek[1] and Peter Kokol[1,2]

[1] Faculty of Health Sciences, University of Maribor, Zitna ulica 15, 2000 Maribor, Slovenia
[2] Faculty of Electrical Engineering and Computer Science, University of Maribor, Smetanova ulica 17, 2000 Maribor, Slovenia

{gregor.stiglic, kokol}@uni-mb.si

Abstract. DNA Microarrays represent one of the widely studied fields in bioinformatics from application of machine learning models perspective. Along with extremely high number of features one is confronted with additional problem of extremely small number of samples. Usual sample size in microarray experiments rarely exceeds hundred samples. This problem severely limits performance of some machine learning techniques such as decision trees or rule based classifiers. This study tries to measure improvement in performance of decision tree models as a function of increasing number of samples in microarray analysis. Results show that it is possible to significantly reduce the gap in performance between the best available classifiers and decision trees by increasing number of samples. Our results hint at a possibility that by increasing sample size in microarray results we will be able to use much more user friendlier and comprehensible classifiers in the future than we do today.

Keywords: gene expression analysis, decision trees, classification, classifier comprehensibility.

1 Introduction

Datasets from microarray analysis, that enables the measurement of molecular signatures of different cells, have become an important part of introducing artificial intelligence to bioinformatics. Supervised machine learning techniques are mainly used to build classifiers, which can be used to improve the prediction of diagnosis. Although their results can be easily interpreted there were very few decision trees used in the early days of machine learning introduction to microarray based gene expression [1,2]. Due to their weak classification accuracy they were soon replaced by more robust ensembles of classifiers that were often based on decision trees, but did not offer as much comprehensibility as basic decision trees. Especially bagging, boosting and Random Forests were on their rise at that time [3], which meant decision trees were usually significantly outperformed by ensemble based methods. But it was not only greater robustness and higher classification accuracy that made such a

difference. One should notice that most of the microarray studies available at that time contained less than 100 samples. Although the prices for microarray experiments dropped significantly in the past ten years we are still facing microarray analysis studies with very low number of samples. Recent review paper on evaluation of microarray based classifiers by Dupuy and Simon [4] reveals that out of 90 reviewed microarray studies in oncology, only 5 of them included more than 100 patients.
Our paper empirically compares performance of four classical decision tree classifiers that are suitable for direct interpretability of their results, to SVM based classification algorithm. Section 2 describes the problem of dimensionality and presents currently available microarray datasets. Experimental setup and evaluation process are presented in section 3. Results are presented in section 4 and are followed by Conclusions that are summarized in the last section.

2 Decision Trees and Gene Expression Classification

One of the main advantages of decision trees is their comprehensibility. Additional to their possible use as classifiers, they represent a generalization of knowledge that is needed to differentiate between clinical outcomes based on gene expression of multiple genes. Decision trees along with rule based classifiers are the only group of classifiers that perform classification by a sequence of simple, easy-to-understand tests whose semantics are intuitively clear to domain experts [5].
There are two disadvantages that also represented a major drawback in using decision tree for microarray analysis problems in the past. The first one is their instability that is tightly connected with the second disadvantage – i.e. difficulties at branching the trees when the number of samples is too low. Instability of decision trees was successfully solved by ensembles of decision trees where multiple trees built from different subsets of the initial dataset were built to improve the robustness of the final classifier. Unfortunately, ensembles of classifiers posses very low level of their knowledge comprehensibility and are not appropriate for interpretation of the acquired knowledge. There were some studies that approached the problem of knowledge extraction from ensembles of classifiers [6,7,8] but all of them are too limited to be useful for practical use.
Quality of branching in decision tree is of vital importance to the final success of classifier. Unfortunately the nature of building a decision tree causes a significant reduction of samples in each subsequent node where two or more branches divide samples in the same number of subgroups. When the number of samples is very low, one will run out of samples in a few subsequent branching steps when building a decision trees no matter what kind of splitting criterion was used. Therefore it is of highest importance to assure a sufficient number of samples are available to build accurate decision tree.
Due to high cost per patient in microarray studies it is nowadays still acceptable for studies with 100 or even less samples to represent benchmarking datasets for evaluation of the most complex classifiers [4]. On the other hand a recent paper by Ein-Dor et al. [9] demonstrates that thousands of genes would be needed to find a reliable set of genes and subsequently a reliable classifier. Most of the microarray

data today is collected in centralized repositories containing large numbers of samples like Gene Expression Omnibus [10] by National Center for Biotechnology Information or ArrayExpress [11] by European Bioinformatics Institute. Unfortunately such repositories are too large and contain data coming from various sources using different protocols to serve as a benchmarking collection of datasets.
Our study takes advantage of one of the largest publicly available repositories of gene expression measurements that were collected by International Genomics Consortium. Containing more than 2000 samples, expO repository [12] is currently one of the most appropriate collections of gene expression samples for evaluation of classification methods. Our study uses datasets from systematically organized expO repository samples that are collected in Gene Expression Machine Learning Repository (GEMLeR) [13]. Four datasets containing more than 500 samples were used for a first empirical evaluation and comparison of decision trees to current state-of-the-art classification method. Ultimate goal of this comparison was to find out how far decision trees can go in terms of classification accuracy if the number of samples in microarray study rises above usual numbers for such kind of experiments. Based on study by Statnikov et al. [14] Support Vector Machines (SVM) classifier was selected as the current "state-of-the-art" classification technique.

3 Experimental Setup

All experiments described in this paper were performed using libraries from Weka machine learning environment [15]. Four classical decision tree classifiers along with Sequential Minimal Optimization (SMO) [16] method that represents optimized SVM implementation in Weka environment were used for classification. Although there are some more sophisticated decision tree implementations available in Weka, we selected four decision tree methods according to comprehensibility of the final decision tree.

The four selected methods from Weka framework are: J48, REPTree, SimpleCart and BFTree. All of the methods are decision trees techniques which represent supervised machine learning approach. J48 is an implementation of a decision tree technique that is based on C4.5 algorithm which was originally proposed by Quinlan [17]. C4.5 algorithm is an extension of Quinlan's previous ID3 (Iterative Dichotomiser 3) method. REPtree method is also based on C4.5 algorithm and can produce classification (discrete outcome) or regression trees (continuous outcome). It sorts numeric attributes only once. SimpleCart method is CART (Classification And Regression Tree) analysis which is based on the paper by Breiman et al. [18]. CART approach can also produce classification or regression trees, which depends on the type of the dependent variable (categorical or numerical). BFTree is a best-first decision tree learner and it is a learning algorithm for supervised classification learning [19]. Best-first decision trees represent an alternative approach to standard decision tree techniques such as C4.5 algorithm since they expand nodes in best-first order instead of a fixed depth-first order.

Each classification method was used "as it is" in Weka environment which means that no additional parameter tuning was performed before or during classification performance comparison.

InfoGain [20] and SVM Recursive Feature Elimination (SVM-RFE) [21] feature selection methods were used to reduce the initial set of available genes. The first feature selection method based on information gain is representative of methods that evaluate one gene at a time, while the second one represents a family of feature selection methods that are able to detect complex patterns through evaluation of multiple genes simultaneously. Each feature selection method was used in combination with all five classification models included in comparison.

To observe the dynamics of classification performance when the number of samples rises, we designed an experimental setup where *n* samples from initial dataset are randomly sampled from original dataset and used in 10-fold cross-validation evaluation. This step is repeated 10 times for each n that is initially set to 50 and increases in steps of 50. Two measures of accuracy were used to evaluate classification performance of four decision tree and a linear kernel SMO classifier. Additionally to the classical accuracy estimation (denoted as ACC), that can be calculated as number of correctly classified samples divided by number of all samples, area under ROC curve or shortly area under curve (AUC) was used. Classification performance estimation was done using 10-fold cross validation cycle that was repeated 10 times to ensure higher reliability of results, especially in cases where small number of samples was used. Number of selected genes was equal to 100. To avoid so called selection bias, which was previously exposed in papers by Ambroise and McLachlan [22] and Simon [23], gene selection along with classification model building was done exclusively on training set inside cross-validation process.

Most evaluations for this study were conducted using four datasets from GEMLeR that were chosen by the highest number of available samples. Therefore a collection of four datasets with more than 500 samples (Table 1) were selected for experimental work. It would be possible to use One-Versus-All (OVA) datasets from GEMLeR containing 1545 samples, but due to high degree of imbalance between the two classes, four largest datasets from All-Paired group of datasets were used. Table 1 shows detailed information on all tissue comparison datasets from GEMLeR.

Table 1. Overview of four GEMLeR datasets used in this study.

Dataset	Num. of Samples	Class 1	Class 2	Num. of Genes
AP_Breast_Colon	630	344	286	10937
AP_Breast_Kidney	604	344	260	10937
AP_Breast_Ovary	542	344	198	10937
AP_Colon_Kidney	546	286	260	10937

Final experiment in this study included 36 different datasets from all cancer tissue pairs found in GEMLeR. To show the significant difference in classification performance in small versus large microarray datasets we repeated classification using 50 samples that were randomly sampled from original datasets and compared the results to classification using all available samples (from 130 in the smallest to 630 samples in the largest dataset). Using all 36 datasets allows additional statistical testing of significance that was done using SPSS software package.

4 Results

Initially, an evaluation of classification performance using SVM-RFE based feature selection was measured using cross-validation evaluation. Results of ACC and AUC for 10 x 10-fold cross-validation for all number of samples settings are presented in Figures 1 and 2. Accuracy and AUC estimation results do not differ significantly and also show the known fact that decision trees cannot compete with much more effective SMO classifier when the number of samples is low. One can observe very similar results for all four decision tree algorithms. Weak performance of REPTree can be observed in the initial stages with number of samples at 50, but this lag behind other decision tree methods is soon gone as the number of samples rises.

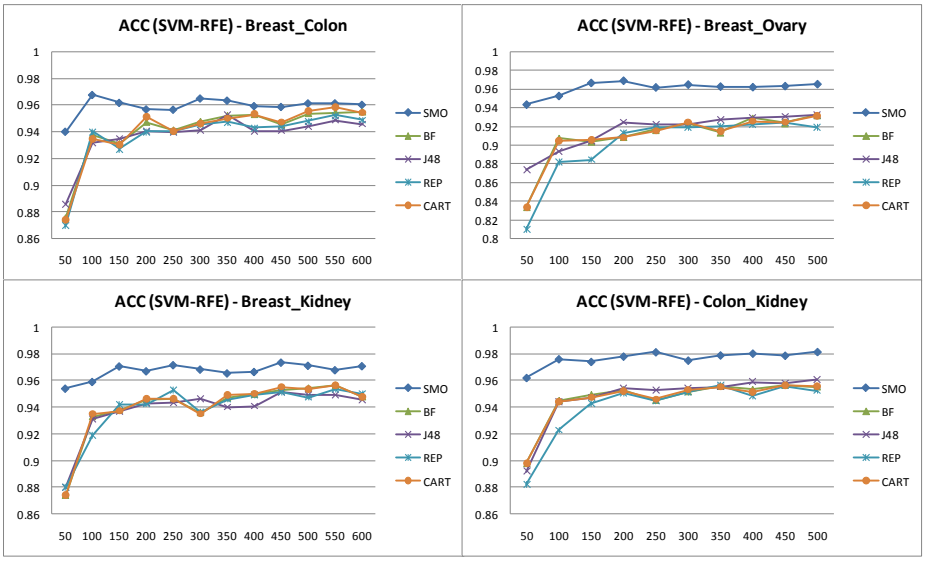

Fig. 1. Accuracy as function of increasing number of samples for SVM-RFE feature selection.

Observing AUC levels for decision tree classifiers at higher number of samples reveals that they managed to significantly narrow the gap to SVM. In the first dataset (Breast vs. Colon Cancer) the difference to SMO for 600 samples is extremely small for simpleCART (0.58%) and REPTree (0.49%) classifiers.

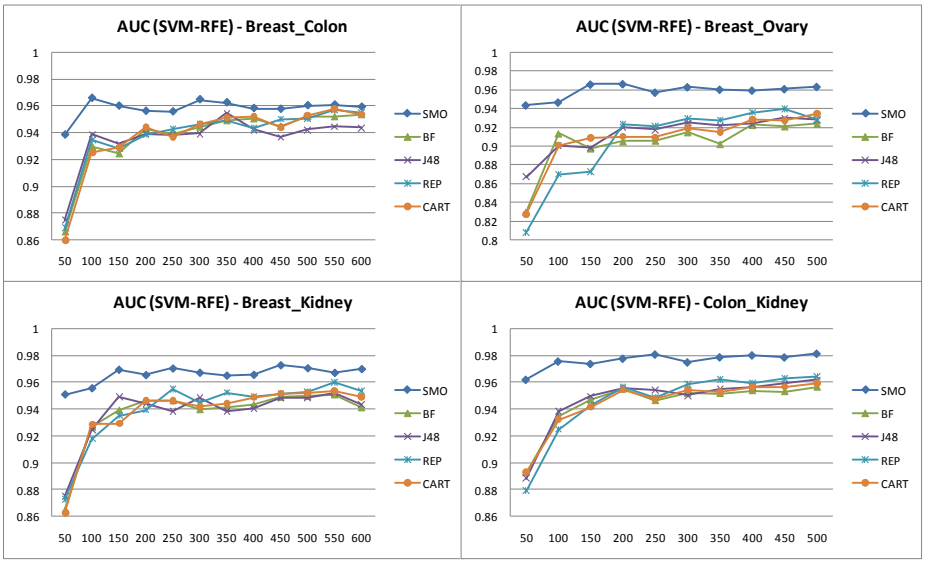

Fig. 2. Area Under Curve (AUC) as a function of increasing number of samples for SVM-RFE based feature selection.

Another experiment was conducted using InfoGain instead of SVM-RFE feature selection algorithm to ensure SMO results are not biased due to underlying SVM based feature selection. Figures 3 and 4 show accuracy and AUC for InfoGain based classification using 10 x 10 – fold cross validation.

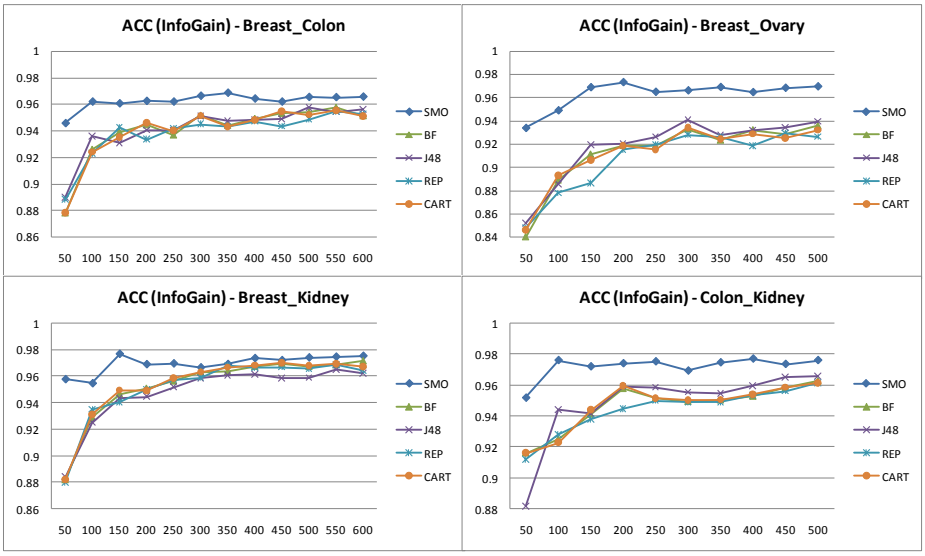

Fig. 3. Accuracy as a function of increasing number of samples for InfoGain based feature selection.

Observing results of accuracy in Figure 3 it is obvious that both SVM-RFE and InfoGain based feature selection methods return very similar results. Once again in two out of four cases, decision trees managed to completely narrow the gap between them and SMO classifier when enough samples were available. Except for minor changes in areas with lower number of samples, also AUC based on InfoGain feature selection shows very similar results to SVM-RFE based feature selection (Figure 4).

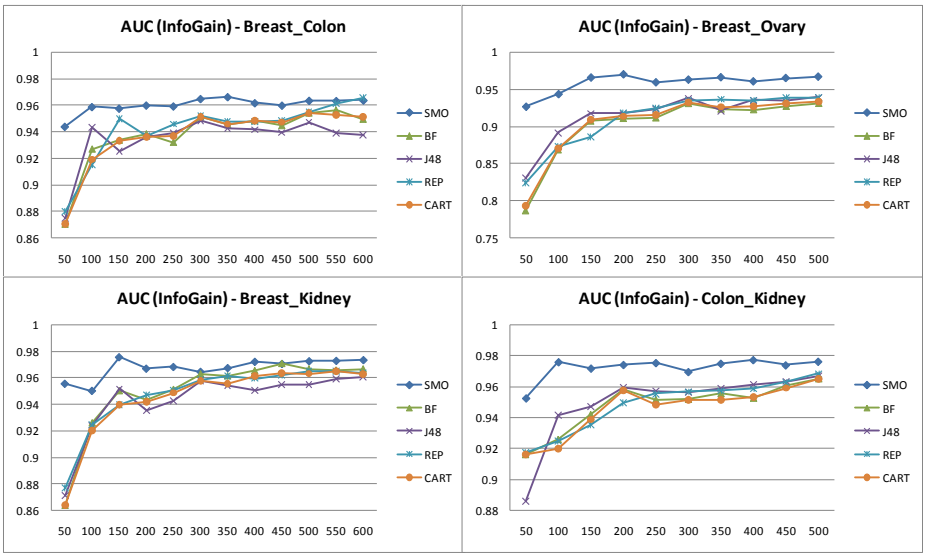

Fig. 4. Accuracy as a function of increasing number of samples for InfoGain based feature selection.

One of the parameters that were measured in this study was decision tree complexity that was measured as number of nodes in a decision tree. It is highly important that this number stays relatively low to (1) avoid overfitting and (2) to maintain simple comprehensibility of built decision trees. Average number of nodes used to build decision trees for different number of samples used is displayed in Figure 5. It can be observed that J48 classifier produces the most complex decision trees that can contain almost 20 nodes in cases where more than 500 samples were used to build the tree. On the other hand REPTree and SimpleCart achieved same or better AUC levels using significantly less nodes and represent the optimal solution in terms of comprehensibility-accuracy ratio. This could also depend on different default parameter setting in Weka environment that were not altered.

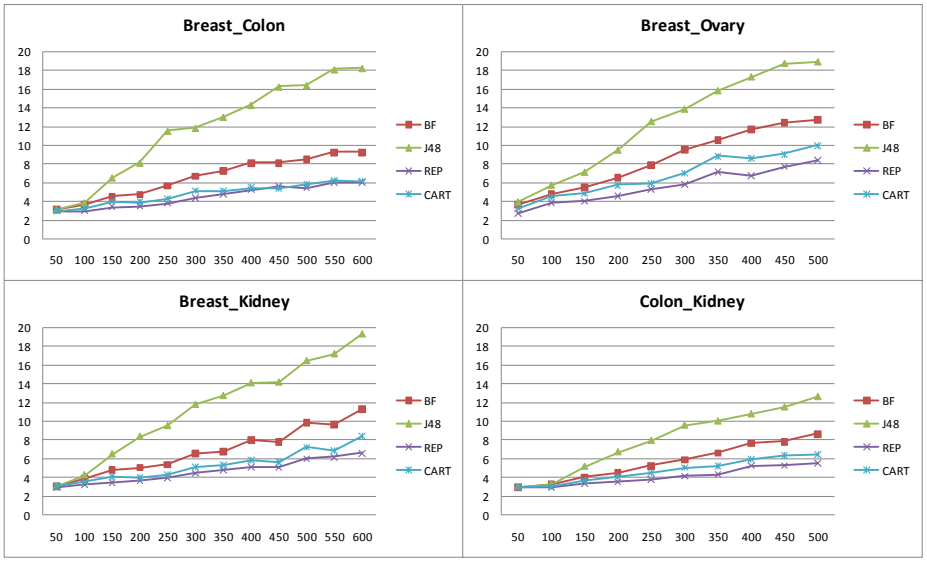

Fig. 5. Complexity (number of nodes) of decision trees as a function of increasing number of samples.

Taking into account accuracy and comprehensibility one can conclude that REPTree represents the most optimal classification solution. Therefore we built a new dataset using top 100 genes selected by SVM-RFE on complete Breast vs. Colon dataset. This exemplary decision tree is presented in Figure 6 which shows the practical usability of such a solution. It is very comprehensible and can be practically used even by non-experts at cost of sacrificing minimal amount of reliability. Probe numbers have been converted to gene names for easier interpretation of results.
Additional to coverage and error on training set (represented in brackets), each final decision tree node displays coverage and error on pruning dataset (squared brackets). Default setting for REPTree includes 3-fold pruning which means 1/3 of the data will be used for pruning purposes. In our case this means over 200 samples that will guarantee an effective pruning process that can also be seen in Figure 3.

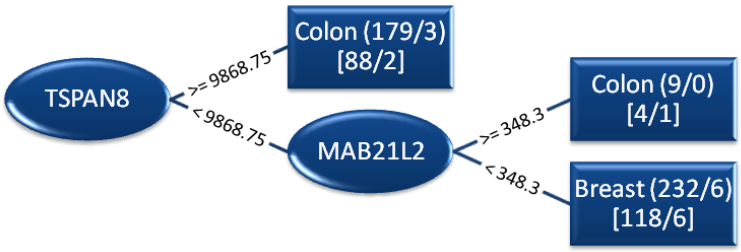

Fig. 6. Exemplary REPTree along with training and pruning set coverage/error results built on data from Breast vs. Colon dataset.

As already mentioned the final experiment included all 36 datasets from GEMLeR and measured classification accuracy on small datasets (random sampling from original datasets that was repeated 20 times) and large datasets (all available samples). SVM-RFE was used for gene selection as in previous experiment. Figure 7 shows gain in ACC and AUC for SMO and REP Tree comparing small to large datasets. Results demonstrate that there are 34 datasets where REP Tree gains more accuracy than SMO and there are 35 such datasets when AUC gain is observed.

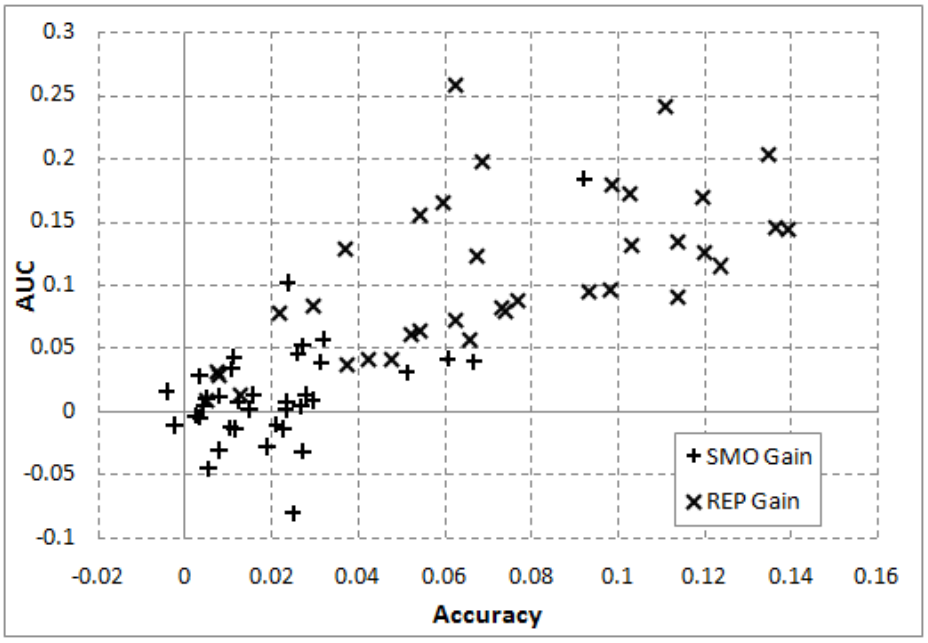

Fig. 7. Comparison of gain in accuracy and AUC for SMO and REP Tree when measured on small versus large datasets. Positive value indicates positive gain for large dataset.

However SMO still outperforms REP Tree classifier in both ACC and AUC, but the gap in performance between both methods is much lower for larger datasets. Wilcoxon signed ranks test was used to confirm this by setting a null-hypothesis that

both methods perform equally. Initially a comparison of ACC and AUC for both methods was conducted on reduced datasets using 50 samples. Both tests (for ACC and AUC) show that there is a significant difference in favour of SMO ($p < 0.0001$). In case of large datasets, statistical testing still points out a significant difference for ACC, but not for AUC. In case of AUC there are 18 datasets where REP Tree performed better, 4 datasets where methods performed equally and 14 datasets where SMO achieved higher values of AUC. Consequently the null-hypothesis cannot be rejected for AUC ($p = 0.108$) on large datasets.

5 Conclusions

This study presents an initial empirical comparison of comprehensible decision tree methods to current state-of-the-art classifier. Four large binary datasets from microarray analysis, containing more than 500 samples each, were initially used to show that it is possible to significantly improve classification performance of decision trees by providing sufficient number of samples. In the final experiment we compared performance of decision tree algorithms to SVM on a large collection of 36 microarray datasets. However it has to be noted that all datasets originate from the same problem area (tumor vs. tumor comparison). Therefore one has to be aware that a lot more computational power and additional large datasets will be needed in the future to adequately compare classification performance of decision trees to state-of-the-art classifiers using statistical hypothesis tests. Another important addition to the current experimental design could be cross-validation based parameter tuning before the application of the classifier to the final testing dataset that could additionally narrow the gap in performance to the best available classification algorithms. Although they have almost been written off, experimental results show that decision tree classifiers might play an important role in microarray analysis in the future.

References

1. Boulesteix, A.L., Tutz, G., Strimmer, K.: A CART-based approach to discover emerging patterns in microarray data, Bioinformatics 18, 2465–2472 (2003)
2. Xiaojing Yuan, Xiaohui Yuan, Fan Yang, Jing Peng, Buckles, B.P.: Gene Expression Classification: Decision Trees vs. SVMs. FLAIRS Conference 2003: 92-97 (2003)
3. Tan, A.C. & Gilbert, D.: Ensemble machine learning on gene expression data for cancer classification. Applied bioinformatics 2 (2003)
4. Dupuy, A., Simon, R. M.: Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. J Natl Cancer Inst 99, 147-157 (2007)
5. Murthy, S. K. Automatic construction of decision trees from data: A multi-disciplinary survey. Data Mining and Knowledge Discovery 2, 345-389 (1998)
6. Craven, M. and Shavlik, J.: Rule Extraction: Where Do We Go from Here?, University of Wisconsin Machine Learning Research Group Working Paper 99-1 (1999)

7. Robert, W., Cunningham, P., Walsh, P., Byrne, S.: Explaining the output of ensembles in medical decision support on a case by case basis. Artificial Intelligence in Medicine, 28(2), 191-206 (2003)
8. Stiglic, G., Mertik, M., Podgorelec, V., Kokol, P.: Using Visual Interpretation of Small Ensembles in Microarray Analysis. Proceedings of Computer Based Medical Systems (CBMS 2006) Conference, Salt Lake City, UT, USA, 691-695 (2006)
9. Ein-Dor, L., Zuk, O. and Domany, E.: Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc Natl Acad Sci U S A 103, 5923-5928 (2006)
10. Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M., Edgar, R.: NCBI GEO: mining tens of millions of expression profiles-database and tools update. Nucleic Acids Res., 35(Database issue), D760-5 (2007)
11. Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., Holloway, E., Kolesnykov, N., Lilja, P., Lukk, M., Mani, R., Rayner, T., Sharma, A., William, E., Sarkans, U., Brazma, A.: ArrayExpress - a public database of microarray experiments and gene expression profiles. Nucleic Acids Res, 35(Database issue):D747-50 (2007)
12. Expo Public Geo Data, http://expo.intgen.org/geo/home.do
13. Gene Expression Machine Learning Repository, http://gemler.fzv.uni-mb.si
14. Statnikov, A., Wang, L. and Aliferis, C. F.: A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics 9, 319+ (2008)
15. Witten, I.H. and Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann (2005)
16. Platt, J.: Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. In: B. Scholkopf, J.C. Burges and A.J. Smola (Eds.), Advances in Kernel Methods - Support Vector Learning, Cambridge MA: MIT Press, pp. 185-208 (1999)
17. Quinlan, J.R.: Induction of decision trees. Machine learning, 1 (1986)
18. Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J.: Classification and regression trees. Monterey, CA, Wadsworth, Inc. (1984)
19. Shi, H.: Best-first decision tree learning, MSc Thesis, Hamilton, NZ (2007)
20. Hunt, E.B., Marin, J. and Stone, P.: Experiments in Induction. The American Journal of Psychology, 80(4), 651-653 (1967)
21. Guyon, I., Weston, J., Barnhill, S. and Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learning, 46, 389–422 (2002)
22. Ambroise, C. and McLachlan, G.J.: Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci U S A 99, 6562-6566 (2002)
23. Simon, R., Radmacher, M.D., Dobbin, K. and McShane, L.M.: Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J. Natl Cancer Inst., 95, 14–18 (2003)