Knowledge Extraction from Microarray Datasets Using Combined Multiple Models to Predict Leukemia Types

Gregor Stiglic[1], Nawaz Khan[2], and Peter Kokol[1]

[1] Faculty of Electrical Engineering and Computer Science, University of Maribor, 2000 Maribor, Slovenia
gregor.stiglic@uni-mb.si

[2] School of Computing Science, Middlesex University, The Burrough, Hendon, London NW4 4BT, UK
n.x.khan@mdx.ac.uk

Summary. Recent advances in microarray technology offer the ability to measure expression levels of thousands of genes simultaneously. Analysis of such data helps us identifying different clinical outcomes that are caused by expression of a few predictive genes. This chapter not only aims to select key predictive features for leukemia expression, but also demonstrates the rules that classify differentially expressed leukemia genes. The feature extraction and classification are carried out with combination of the high accuracy of ensemble based algorithms, and comprehensibility of a single decision tree. These allow deriving exact rules by describing gene expression differences among significantly expressed genes in leukemia. It is evident from our results that it is possible to achieve better accuracy in classifying leukemia without sacrificing the level of comprehensibility.

1 Introduction

Clinical diagnosis for disease prediction is one of the most important emerging applications of microarray gene expression study. In the last decade, a new technology, DNA microarrays, has allowed screening of biological samples for a huge number of genes by measuring expression patterns. This technology enables the monitoring of the expression levels of a large portion of a genome on a single slide or “chip”, thus allowing the study of interactions among thousands of genes simultaneously [1]. Usually microarray datasets are used for identification of differentially expressed genes, which from data mining point of view represents a feature selection problem. The objectives of this research are to select important features from leukemia predictive genes and to derive a set of rules that classify differentially expressed genes. The study

G. Stiglic et al.: *Knowledge Extraction from Microarray Datasets Using Combined Multiple Models to Predict Leukemia Types*, Studies in Computational Intelligence (SCI) **118**, 339–352 (2008)
www.springerlink.com

follows the comprehensibility of a single decision tree. Although, there are many research that have demonstrated a higher level of accuracy in classifying cancer cells, for example [2, 3], the comprehensibility issue of decision trees to gain best accuracy in the domain of microarray data analysis has been ignored [4–7].

In this study, we attempt to combine the high accuracy of ensembles and the interpretability of the single tree in order to derive exact rules that describe differences between significantly expressed genes that are responsible for leukemia. To achieve this, Combined Multiple Models (CMM) method has been applied, which was proposed originally by Domingos in [8]. In our study the method is adapted for multidimensional and real valued microarray datasets to eliminate the colinearity and multivariate problems. All datasets from our experiment are publicly available from the Kent Ridge Repository described in [20]. These microarray samples are the examples of human tissue extracts that are related to a specific disease and have been used for comprehensible interpretation in this study. The following sections explore the datasets, methods of CMM adaptation and testing. It also presents the results that are obtained by applying the adapted method on four publicly available databases. Finally the chapter presents a validation study by providing an interpretation of the results in the context of rule sets and then by comparing the proposed adaptations with the combined and simple decision trees for leukemia grouping.

2 Combined Multiple Models for Gene Expression Analysis

Data mining is the process of autonomously extracting useful information or knowledge from large datasets. Many different models can be used in data mining process. However, it is required for many applications not only to involve model that produce accurate predictions, but also to incorporate comprehensible model. In many applications it is not enough to have accurate model, but we also want comprehensible model that can be easily interpreted to the people not familiar with data mining. For example, Tibshirani and Knight [9] proposed a method called Bumping that tries to use bagging and produce a single classifier that best describes the decisions of bagged ensemble. It builds models from bootstrapped samples and keeps the one with the lowest error rate on the original data. Typically this is enough to get good results also on test set. We should also mention papers that suggest different techniques of extracting decision trees from neural networks or ensembles of neural networks that can all be seen as a “black-box” method [10–12].

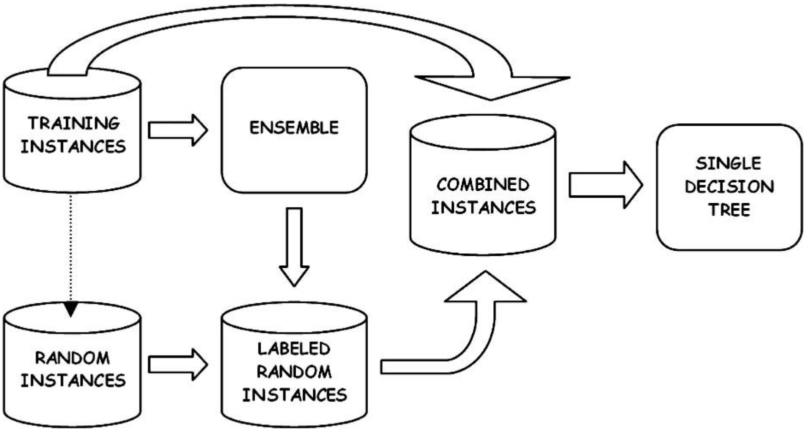

Fig. 1. Building decision tree from ensemble using CMM

2.1 CMM Approach

One of the methods that are able to build comprehensible model from an ensemble of models is called Combined Multiple Models (CMM) and was presented in [8]. CMM was later studied and improved by Estruch et al. in [13]. Basic idea of CMM is to build a single classifier that would retain most of the accuracy gains of the ensemble models. This is done by adding additional "artificial data points" to the learning dataset. Those additional data points are then classified (i.e. labeled) by applying the ensemble of classifiers that was trained on the learning dataset. The next step is joining the original training dataset with the new "artificial" dataset. This final dataset is used to build a single comprehensible classifier. The whole process is shown in Fig. 1. The idea of generating the "artificial data points" when building classifiers, was already used in several papers. One of the first such methods is the active learning method proposed by Cohn et al. in [14]. Another application of artificial examples was presented by Craven and Shavlik in [15] where they describe the learning of decision trees from neural networks. This approach was later used in several papers on neural networks knowledge extraction.

2.2 Proposed Modifications

This chapter presents the CMM based method that is specialized for microarray dataset classification problems. Optimization of the original method was done on artificial data points creation due to specific structure of the microarray datasets. Opposite to the original research [8], where most of the best results were achieved on the datasets containing nominal values, microarray analysis presents pure continuous-valued data-sets. Therefore a new method called Combined Multiple Models for Continuous values (CMMC) is proposed. In this chapter three new artificial data points creation techniques for decision tree building are examined which will be referred to as CMMC-1, CMMC-2 and CMMC-3. Both methods are based on multiplication of the original training set instances. Data points are generated from original training set by creating copies of original training set instances by slightly changing the values of attributes. First method is based on the variance of the gene expression values and each attribute can be changed by adding the random value from one of the intervals $\left\{-\sigma, -\frac{1}{3}\sigma\right\}$ and $\left\{\frac{1}{3}\sigma, \sigma\right\}$ to the original gene expression value. Because of the large number of attributes we change only 50% randomly selected attributes. Result of such data point multiplication is a wide dispersion of the points around their base data point, but original training set distribution of the samples is still preserved. Second method tries to maintain the original distribution on even tighter area than the first one, especially when data points lie tightly together. This is done by generating the random points in the interval $x \quad d$, where x is the value of the attribute and d is the distance to the nearest neighbour value of this attribute. Again only 50% of attributes are randomly selected for modification. Another modification of the original approach was done in application of different ensemble building method. Based on our own tests and also reports in some papers [16], we decided to use Random Forest ensemble building method that is based on one of the first ensemble building methods called bagging [17]. To compose ensemble from base classifiers using bagging, each classifier is trained on a set of n training examples, drawn randomly with replacement from the original training set of size m. Such subset of examples is also called a bootstrap replicate of the original set. Breiman upgraded the idea of bagging by combining it with the random feature selection for decision trees. This way he created Random Forests, where each member of the ensemble is trained on a bootstrap replicate as in bagging. Decision trees are than grown by selecting the feature to split on at each node from randomly selected number of nodes. We set number of chosen features to $log_2(k+1)$ as in [18], where k is the total number of features. Random Forests are the ensemble method that works well even with noisy content in the training dataset and are considered as one of the most competitive methods that can be compared to boosting [19]. To get the most out of the proposed multiplication of data points another version of CMMC algorithm was derived. This version (CMMC-3) is based upon multiplication of data points in each of the leafs which are later extended by additional subtree. Since our decision trees are pruned they achieve good generalization and are less complex than unpruned decision trees. Therefore we try to "upgrade" each leaf by attaching another subtree under the leaf. These subtrees are built using CMMC technique described above (basically the same as CMMC-2). Thereby existing data points that got to the leaf are multiplied (again by adding 1,000 artificial data points labelled by Random Forest) and the problem of a small number of samples in lower nodes of trees is reduced but not solved as we cannot be certain about the correct labelling of the artificial samples.

3 Experiment and Results

3.1 Datasets

Four widely used publicly available gene expression datasets were used in our experimental evaluation of the proposed method. They were obtained from Kent Ridge Biomedical Data Set Repository which was described in [20].

Leukemia1 Dataset (amlall)

The original data comes from the research on acute leukemia by Golub et al. [21]. Dataset consists of 38 bone marrow samples from which 27 belong to acute lymphoblastic leukemia (ALL) and 11 to acute myeloid leukemia (AML). Each sample consists of probes for 6,817 human genes. Golub used this dataset for training. Another 34 samples of testing data were used consisting of 20 ALL and 14 AML samples. Because we used leave-one-out cross-validation, we were able to make tests on all samples together (72).

Breast Cancer Dataset (Breast)

This dataset was published in [22] and consists of extremely large number of scanned gene expressions. It includes data on 24,481 genes for 78 patients, 34 of which are from patients who had developed distance metastases within 5 years, the rest 44 samples are from patients who remained healthy from the disease after their initial diagnosis for interval of at least 5 years.

Lung Cancer Dataset (Lung)

Lung cancer dataset includes the largest number of samples in our experiment. It includes 12,533 gene expression measurements for each of 181 tissue samples. The initial research was done by Gordon et al. [23] where they try to classify malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung.

Leukemia2 Dataset (mll)

This Leukemia dataset tries to discern between three types of leukemia (ALL, MLL, AML). Dataset contains 72 patient samples, each of them containing 12,582 gene expression measurements. Data was collected by Armstrong et al. and results published in [24].

3.2 Gene Selection

It has been shown that selecting a small subset of informative genes can lead to improved classification accuracy and greatly improves execution time of data mining tools [25]. The most commonly used gene selection approaches are based on gene ranking. In these gene ranking approaches, each gene is evaluated individually and assigned a score representing its correlation with the class. Genes are then ranked by their scores and the top ranked ones are selected from the initial set of features (genes). To make our experiments less dependent of the filtering method, we use three different filtering methods. This way we get 12 different microarray datasets with a pre-defined number of most relevant gene expressions. All used filtering methods are part of WEKA toolkit [26] that we were using in our experiments. The following filtering methods were used:

GainRatio filter. This is the heuristic that was originally used by Quinlan in ID3 [27]. It is implemented in WEKA as a simple and fast feature selection method. The idea of using this feature selection technique for gene ranking was already presented by Ben-Dor et al. [28].

Relief-F filter. The basic idea of Relief-F algorithm [29] is to draw instances at random, compute their nearest neighbors, and adjust a feature weighting vector to give more weight to features that discriminate the instance from neighbors of different classes. A study comparing Relief-F to other similar methods in microarray classification domain was conducted by Wang and Makedon [30] where they conclude that the performance of Relief-F is comparable with other methods.

SVM filter. Ranking is done using Support Vector Machines (SVM) classifier. Similar approach using SVM classifier for gene selection was already used in papers by Guyon et al. [31] and Fujarewicz et al. [32].

3.3 Experiment Setting

The experiments are designed to test the accuracy gain of all three CMMC methods compared to accuracy of a single J48 tree (Java C4.5 tree implementation in WEKA toolkit). The study has followed n-fold cross-validation process for testing. The n-fold cross-validation is typically implemented by running the same learning system n times and each time on a different training set of size *(n-1)/n* times the size of the original data set. A specific variation of n-fold cross-validation, called leave-one-out cross-validation method (LOOCV), is used in the experiment. In this approach, one sample in the training set is withheld, the remaining samples of the training set are used to build a classifier to predict the class of withheld sample, and the cumulative error is then calculated. LOOCV was often criticized, because of higher error variance in comparison to five or tenfold cross-validation [33], but a recent study by Braga-Neto and Dougherty [34] shows that LOOCV can be considered very useful for microarray datasets, because they have not been able to verify the substantial differences in performance among the mentioned methods. Because of random nature in tested classifier building methods, this research attempts to repeat LOOCV ten times for all random based methods (both CMMC and Random Forests) and then computes average accuracy for all runs. As indicated in [8], 1,000 artificial data points are generated for CMMC-1 and CMMC-2 methods, while CMMC-3 uses the same number of artificial data points in every "upgraded" leaf.

3.4 Results

This section highlights the key findings that are obtained by applying the adapted CMM method on four microarray datasets available in public domain. Table 1 shows the accuracy comparison. The tests show that all three proposed methods gained some accuracy comparing to a simple decision tree, but they are still lacking a lot of accuracy compared to ensemble of classifiers.

To keep the complexity level low for built decision trees, we used pruning in all decision trees that are used in the experiment. Average complexity (i.e. number of rules) of decision trees is presented in Table 2. We do not present the complexity of Random Forest Method as it can be simply estimated as approximately 100 times larger than the simple decision tree and therefore completely unacceptable for interpretation. The most significant fact revealed from the Table 2 is low rule complexity of CMMC-2 generated decision trees, especially when compared to CMMC-1 trees. Trees from our second proposed method

Table 1. Comparison of accuracy for decision tree (C4.5), proposed decision tree building methods and Random Forests (RF)

Dataset	C4.5	CMMC1	CMMC2	CMMC3	RF
amlall1	80.56	90.40	89.20	87.58	97.92
amlall2	79.17	91.29	88.70	88.44	98.30
amlall3	79.17	90.85	87.94	88.30	98.80
amlallAvg	79.63	90.85	88.61	88.11	98.34
breast1	66.67	73.89	67.19	72.85	85.84
breast2	61.54	64.74	65.62	65.66	81.00
breast3	71.79	66.49	66.78	65.04	90.21
breastavg	66.67	68.37	66.53	67.97	85.68
lung1	96.13	96.37	97.55	96.64	99.45
lung2	97.79	96.61	97.95	97.06	98.97
lung3	98.90	96.53	98.58	97.28	99.45
lungavg	97.61	96.50	98.03	96.99	99.29
mll1	79.17	88.89	89.48	87.08	97.62
mll2	88.89	88.29	88.89	91.96	94.64
mll3	84.72	88.29	88.89	87.62	96.03
mllAvg	84.26	88.49	89.09	88.87	96.10
Average	82.04	86.05	85.56	85.49	94.85

Table 2. Comparison of tree complexity (number of leafs) for decision tree (C4.5) and proposed decision tree building methods (CMMC-1, CMMC-2 and CMMC-3)

Dataset	C4.5	CMMC1	CMMC2	CMMC3
amlall1	2.93	73.30	5.24	6.19
amlall2	2.93	75.57	5.38	5.45
amlall3	2.93	75.77	5.05	6.29
amlallAvg	2.93	74.88	5.22	5.98
breast1	6.26	46.35	18.78	11.22
breast2	7.12	11.37	15.65	14.53
breast3	6.76	11.92	14.68	23.05
breastAvg	6.71	23.21	16.37	16.27
lung1	3.99	75.32	5.05	4.35
lung2	4.00	64.05	9.21	4.31
lung3	4.00	72.71	5.52	4.21
lungAvg	4.00	70.69	6.59	4.29
mll1	3.00	115.94	6.39	4.88
mll2	3.00	118.31	8.68	5.24
mll3	3.92	121.76	7.60	5.00
mllAvg	3.31	118.67	7.56	5.04
Average	4.24	71.86	8.94	7.90

Table 3. Comparison of accuracy by feature selection method

Feature selection	C4.5	CMMC1	CMMC2	CMMC3	RF
GainRatio	80.63	87.39	85.86	86.04	95.21
ReliefF	81.85	85.23	85.29	85.78	93.23
SVM-FS	83.65	85.54	85.55	84.56	96.12

generate only two times more rules than simple decision trees. Even better results were obtained using CMMC-3 method. Low complexity at CMMC-3 based trees is a consequence of the decision tree building technique in which trees are generated at the beginning using the initial training sets without artificial data points. The artificial data points are added in later stages.

Table 3 presents the results based on average results on each dataset for each gene selection method. The best results, with exception of CMMC-3 method, were achieved when SVM based feature selection method was used. From this table it can also be seen that the majority of accuracy gain of the CMMC-1 method compared to CMMC-2 was due to first method's better accuracy when used with the GainRatio based feature selection method.

3.5 Subgrouping Leukemia Type

To demonstrate the practical advantage of our proposed adaptation to CMM method, two sample trees were constructed from the Leukemia (AML-ALL)

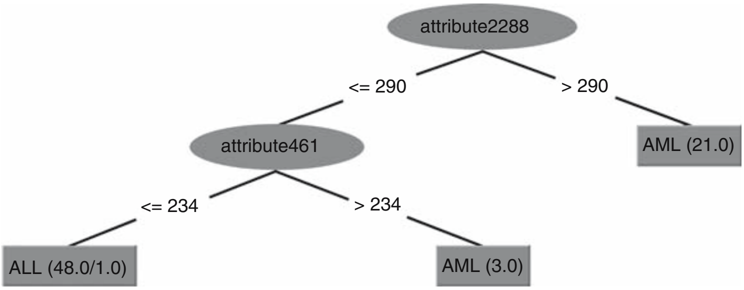

Fig. 2. J48 decision tree generated from amlall3 dataset (98.61% accuracy)

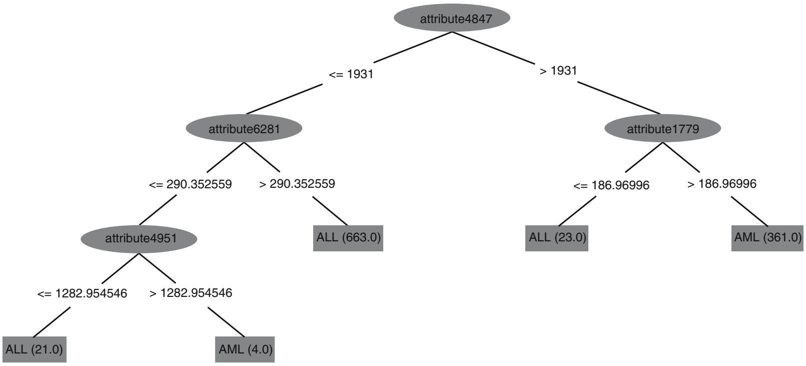

Fig. 3. CMMC-2 decision tree from amlall3 dataset (100% accuracy)

dataset. The first tree (Fig. 2) was constructed using J48 algorithm, while the second tree (Fig. 3) used CMMC-2 algorithm.

The first tree was built from all 72 samples of the amlall3 dataset, where SVM based filtering was used. J48 tree classified all but one sample correctly using only two genes. On the other hand CMMC-2 tree classified all 72 and additional 1,000 artificial samples correctly using only four genes. The decision tree yields five rules.

The goal of the decision tree presented in Fig. 3 is to derive short rules that explicitly clarify the types of leukemia and are convenient for expert interpretation. The rules are summarized in Table 5 that is accompanied by Table 4. The rule extraction process is based on the attribute to gene name mappings that are presented in Table 4.

To demonstrate the difference between the two constructed trees (J48 and CMM), a set of the strongest rules were chosen that were revealed by the CMM tree but failed to be recognised by the J48 tree. The rules, presented in Table 4, are directly extracted according to the corresponding branches of the decision tree, presented in Fig. 3. An interpretation of the rules (Table 5) is provided in Sect. 4.

Table 4. Attributes for gene mapping used in both decision trees

Attribute no.	Gene description
2,288	DF D component of complement (adipsin)
461	Liver mRNA for interferon-gamma inducing factor (IGIF)
4,847	Zyxin X95735_at
6,281	MYL1 Myosin light chain (alkali)
1,779	MPO Myeloperoxidase
4,951	Nucleoside-diphosphate kinase

Table 5. Rules for Leukemia classification derived from CMM decision tree

1. IF (zyxin X95735 NOT EXPRESSED) AND (MYL1 Myosin Light Chain EXPRESSED) ⇒ ALL
2. IF (zyxin X95735 NOT EXPRESSED) AND (MYL1 Myosin Light Chain (Alkali) NOT EXPRESSED) AND (Nucleoside-diphosphate kinase EXPRESSED) ⇒ AML
3. IF (zyxin X95735 EXPRESSED) AND (MPO Myeloperoxidase EXPRESSED) ⇒ AML

4 Discussion and Conclusion

Acute leukemia which is of lymphoid origin is called Acute Lymphocytic Leukemia (ALL) and a malignant disorder where myeloid blast cells accumulate in the marrow and bloodstream is called Acute Myelocytic Leukemia (AML). A study conducted by Golub et al. [21] has revealed 50 predictive genes that differentiate between ALL and AML. The report has shown that over expression of myosin light chain (M31211) leads to the ALL. The report also has indicated that zyxin 2 X95735, an adhesion plaque protein a component of a signal transduction pathway that mediates adhesion stimulated changes in gene expression, plays a significant role in AML. In a recent study conducted by Umpai and Aitken [35] has demonstrated that the gene X95735 zyxin significantly determines AML whereas myosin light chain over expression frequency is higher in ALL patients. Some other studies also have demonstrated the similar result. For example, Aris and Rece [36] has demonstrated the differentiation technique of AML and ALL based on the fact that zyxin is significant in AML, on the other hand, myosin light chain expression is significant in ALL patients. The French–American–British (FAB) group [37] has standardized the nature of ALL and AML on the basis of myeloperoxidase (MPO) expression. According to the criteria, the AML group demonstrates greater than 3% MPO and/or Sudan Black B (SBB) blast. The study has also revealed that, 70–75% of AML cases show myeloid associated antigens positive, for example, CD13, CD33, MPO etc., thus, it is evident that zyxin

X95735 and myeloperoxidase overexpression determines the AML subgroup of acute leukemia. Expression of nm23-H1/Neocleoside diphosphate kinase (NDPK) correlates inversely with the metastasising potential of some human tumours. The nucleoside diphosphate kinase enzymatic activity possessed by several isomers, for example, Nm23 H1 and NM23 H2, is increased significantly in AML cells and a higher level of nm23-H1 expression is correlated with a poor prognosis in AML. A study, conducted by Yokoyama et al. revealed that 110 AML patients have demonstrated the increased nm23-H1 mRNA level which showed a resistance in response to initial chemotherapy. They also have demonstrated that nm23 H1 has an enormous prognostic affect in AML, especially in AML-M5 (acute monocytic leukemia) [38]. From the discussion above it is evident that proposed adaptation of CMM model gives accurate trees that carry additional knowledge compared to classical decision trees. The research shows that the proposed CMMC demonstrated 2% higher accuracy compared to the classical decision trees. In addition to that, it has been demonstrated that the best CMMC method even can manage to keep the complexity level of the tree very low. Therefore, it is evident that CMMC tree was only twice as large as an original tree. The proposed CMMC model used a well known C4.5 algorithm for building the final decision tree. One of the problems with classical decision tree algorithms is that splitting in the lower lying nodes is based on fewer samples than splitting in the nodes near the root of the tree. Therefore, splitting toward the leaves is less reliable. The idea of using less artificial data points in the upper nodes by introducing them at the lower nodes of a decision tree can be applied in future for better accuracy. Although a part of this idea was already proposed in this chapter, there are still a lot of possible variations to this idea left to be researched. This study has also opened another important direction for further research. A small ensembles comprising of only three or five classifiers that can still be interpreted in the form of rules or some novel visualization techniques can be introduced to demonstrate the comprehensibility of microarray data in order to predict diseases.

References

1. L.-H. Loo, Identifying Differentially Expressed Genes in DNA Microarray Data, PhD Thesis, Drexel University, 2004
2. Z. Guo, T. Zhang, X. Li, Q. Wang, J. Xu, H. Yu, J. Zhu, H. Wang, C. Wang, E. J. Topol, Q. Wang and S. Rao, Towards precise classification of cancers based on robust gene functional expression profiles, BMC Bioinformatics, vol. 6, no. 1, p. 58, 2005
3. J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson and P. S. Meltzer, Classification and diagnostic pre-diction of cancers using gene expression profiling and artificial neural networks, Nature Medicine, vol. 7, no. 6, pp. 673–679, 2001

4. B. Brors, A. Kohlmann, S. Schnittger, C. Schoch, T. Haferlach and R. Eils, Classification of Cytogenetically Defined AML Patients by Decision Tree Analysis of Statistically Selected Gene Expression Data, in Proceedings of 43rd Annual Meeting of the American Society of Hematology (ASH01), Orlando, FL (USA), December 7–12, 2001
5. J. Li and K. Ramamohanarao, A Tree-based Approach to the Discovery of Diagnostic Biomarkers for Ovarian Cancer, in Proceedings of the PAKDD 2004, pp. 682–691, Sydney, Australia, February 2004
6. M. Dettling, BagBoosting for tumor classification with gene expression data, Bioinformatics, vol. 20, no. 18, pp. 3583–3593, 2004
7. D. P. Berrar, B. Sturgeon, I. Bradbury, C. S. Downes and W. Dubitzky, Microarray Data Integration and Machine Learning Techniques For Lung Cancer Survival Prediction, in Proceedings of Critical Assessment of Microarray Data Analysis (CAMDA 2003), Durham, North Carolina, USA, pp. 43–54, November 2003
8. P. Domingos, Knowledge discovery via multiple models, Intelligent Data Analysis, vol. 2 no. 1–4, pp. 187–202, 1998
9. R. Tibshirani and K. Knight, Model search and inference by bootstrap bumping, Journal of Computational and Graphical Statistics, vol. 8, pp. 671–686, 1999
10. O. Boz, Converting a Trained Neural Network To a Decision Tree DecText – Decision Tree Etxractor, PhD thesis, Computer Science and Engineering, Lehigh University, 2000
11. M. W. Craven, Extracting Comprehensible Models from Trained Neural Networks, PhD thesis, University of Wisconsin – Madison, 1996
12. Z.-H. Zhou and Y. Jiang, NeC4.5: neural ensemble based C4.5, IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 6, pp. 770–773, 2004
13. V. Estruch, C. Ferri, J. Hernndez-Orallo and M. J. Ramrez-Quintana, Simple Mimetic Classifiers, in Proceedings of IAPR International Conference on Machine Learning and Data Mining (MLDM2003), pp. 156–171, 2003
14. D. Cohn, L. Atlas and R. Ladner, Improving generalization with active learning, Machine Learning, vol. 15, pp. 201–221, 1994
15. M. W. Craven and J. W. Shavlik, Extracting comprehensible concept representations from trained neural networks, in Working Notes on the IJCAI'95 Workshop on Comprehensibility in Machine Learning, Montreal, Canada, pp. 61–75, 1995
16. H. Zhang, C. Y. Yu and B. Singer, Cell and Tumor Classification Using Gene Expression Data: Construction of Forests, in Proceedings of National Academy of Sciences U S A, vol. 100, no. 7, pp. 4168–4172, 2003
17. L. Breiman, Bagging predictors, Machine Learning, Vol. 24, no. 2, pp. 123–140, 1996
18. L. Breiman, Random forests, Machine Learning, Vol. 45, no. 1, pp. 5–31, 2001
19. T. G. Dietterich, Ensemble Learning, in The Handbook of Brain Theory and Neural Networks, 2nd ed., M. A. Arbib, Ed. MIT, Cambridge, MA, pp. 405–408, 2002
20. J. Li and H. Liu, Ensembles of Cascading Trees, in Proceedings of IEEE International Conference on Data Mining (ICDM 2003), IEEE Computer Society, Melbourne, p. 585

21. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield and E. S. Lander, Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, Science, vol. 286, no. 5439, pp. 531–537, 1999
22. L. J. van 't Veer, H. Dai, M. J. van De Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. Der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards and S. H. Friend, Gene expression profiling predicts clinical outcome of breast cancer, Nature, vol. 415, pp. 530–536, 2002
23. G. J. Gordon, R. V. Jensen, L.-L. Hsiao, S. R. Gullans, J. E. Blumenstock, S. Ramaswami, W. G. Richards, D. J. Sugarbaker and R. Bueno, Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma, Cancer Research, vol. 62, no. 17, pp. 4963–4967, 2002
24. S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Min-den, S. E. Sallan, E. S. Lander, T. R. Golub and S. J. Korsmeyer, MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia, Nature Genetics, vol. 30, no. 1, pp. 41–47, 2002
25. Y. Lu and J. Han, Cancer classification using gene expression data, Information Systems, vol. 28, no. 4, pp. 243–268, 2003
26. I. Witten and E. Frank, Data Mining: Practical Machine Learning Tools with Java Implementations, Morgan Kaufmann, San Francisco, 2000
27. J. R. Quinlan, Induction of decision trees, Machine Learning, vol. 1, pp. 81–106, 1986
28. A. Ben-Dor, N. Friedman and Z. Yakhini, Scoring genes for relevance, Agilent Technologies Technical Report AGL-2000-13
29. I. Kononenko, Estimating Attributes: Analysis and Extensions of Relief, in Proceedings of ECML'94, pp. 171–182, Springer, Berlin Heidelberg New York, 1994
30. Y. Wang and F. Makedon, Application of Relief-F Feature Filtering Algorithm to Selecting Informative Genes for Cancer Classification Using Microarray Data, in Proceedings of IEEE Computational Systems Bioinformatics Conference, pp. 497–498, Stanford, California, 2004
31. I. Guyon, J. Weston, S. Barnhill and V. Vapnik, Gene selection for cancer classification using support vector machines, Machine Learning, vol. 46, no. 1–3, pp. 389–422, 2002
32. K. Fujarewicz, M. Kimmel, J. Rzeszowska-Wolny and A. Swierniak, A note on classification of gene expression data using support vector machines, Journal of Biological Systems, vol. 11, no. 1, pp. 43–56, 2003
33. T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning, Springer, Berlin Heidelberg New York, 2001
34. M. Braga-Neto and E.R. Dougherty, Is cross-validation valid for small-sample microarray classification?, Bioinformatics, vol. 20, no. 3, pp. 374–380, 2004
35. T. Umpai and S. Aitken, Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes, BMC Bioinformatics, vol. 6, no. 148, 2005
36. V. Aris and M. Rece, A Method to Improve Detection of Disease Using Selectively Expressed Genes in Microarray Data, Methods of Microarray Data Analysis, Kluwer, Dordecht, 2002

37. A. Venditti, G.D. Peeta, F. Buccisano, A. Tambarini, et. al., Minimally differentiated acute myleoid leukemia (AML-MO): Comparisson of 25 cases with other French–American–British subtypes, Blood, vol. 89, no. 2, pp. 621–629, 1997

38. A. Yokoyama, J. Okabe-Kado, et. al., Evaluation by multivariate analysis of the differentiation inhibitory factor nm23 as a prognostic factor in acute myelogenous leukemia and application to other hematologic malignancies, Blood, vol. 91, no. 6, pp. 1845–1851, 1998