Chapter 21
Rotation of Random Forests for Genomic and Proteomic Classification Problems

Gregor Stiglic, Juan J. Rodriguez, and Peter Kokol

Abstract Random Forests have been recently widely used for different kinds of classification problems. One of them is classification of gene expression samples that is known as a problem with extremely high dimensionality, and therefore demands suited classification techniques. Due to its strong robustness with respect to large feature sets, Random Forests show significant increase of accuracy in comparison to other ensemble-based classifiers that were widely used before its introduction. In this chapter, we present another ensemble of decision trees called Rotation Forest and evaluate its classification performance on different microarray datasets. Rotation Forest can also be applied to different already existing ensembles of classifiers like Random Forest to improve their accuracy and robustness. This study presents evaluation of Rotation Forest classification technique based on decision trees as base classifiers and was evaluated on 14 different datasets with genomic and proteomic data. It is evident that Rotation Forest as well as the proposed rotation of Random Forests outperform most widely used ensembles of classifiers including Random Forests on majority of datasets.

1 Introduction

There have been many supervised classification techniques that were applied to the analysis of microarray and mass spectrometry-based data in recent years. Diaz-Uriarte and Alvarez [1] have shown that ensembles of classifiers can perform very good even in very high-dimensional and noisy gene expression domain, where they can achieve at least as good results as some other advanced machine learning techniques such as support vector machines (SVM) [2, 3], nearest neighbours [4, 5]

G. Stiglic (✉)
Faculty of Health Sciences, University of Maribor, Zitna ulica 15, 2000 Maribor, Slovenia
and
Faculty of Electrical Engineering and Computer Science, University of Maribor,
Smetanova 17, 2000 Maribor, Slovenia
e-mail: gregor.stiglic@uni-mb.si

H.R. Arabnia and Q.-N. Tran (eds.), *Software Tools and Algorithms for Biological Systems*, Advances in Experimental Medicine and Biology 696,
DOI 10.1007/978-1-4419-7046-6_21, © Springer Science+Business Media, LLC 2011

or neural networks [6]. The key idea of building multiple classification models and assembling them in committees of classifiers is the ability of ensembles to increase stability and accuracy compared to a single classifier [7]. This is especially true when classifiers like decision trees or neural networks that are very sensitive to changes in the underlying training set are used. Diversity among the members of an ensemble of classifiers is deemed to be a key issue in classifier combination as it ensures that independent members of an ensemble are built from the same initial dataset.

2 Methods

2.1 Basic Ensemble Building Techniques

One of the first ensemble building techniques was called bagging and was introduced by Breiman in [8]. It is based on random sampling of examples from the training set which is also the basis of bootstrapping [9]. An ensemble of classifiers that was built using bootstrapping is then used to classify an example using the majority vote of the ensemble.

Another basic and at the same time widely used method for building ensembles is called boosting [10]. In our research, boosting is represented by AdaBoost.M1 variant, which is the most commonly used algorithm from boosting family of ensemble building techniques. The main idea of boosting is re-weighting of examples, and it is therefore assumed that each base classifier can handle weighted examples. In cases where this is not possible, a dataset is obtained from a random sample, taking into account the weights distribution. In comparison to bagging, classifiers are built in a sequential process that cannot be parallelized. Although the idea of boosting is very promising and also achieves good accuracy results in practice, there are some drawbacks we should consider when using boosting. One of them is overfitting to the training set examples; although early literature mentions that boosting would not overfit even when running for a large number of iterations. Recent research clearly shows negative overfitting effects when boosting is used on datasets with higher noise content [11].

3 Random Forests

To increase the diversity of classifiers in bagging, Breiman upgraded the basic idea of bagging by combining bootstrapping with random feature selection for decision tree building. It has to be noted that feature randomization that represents an integral part of Random Forests method was introduced earlier by Ho [12] and Amit et al. [13].

Random decision trees created this way are grown by selecting the feature to split on at each node from randomly selected set of features. The number of chosen features is a parameter of the method. In this work, the number of chosen features is set to $\log_2(k+1)$ as in [14], where k is the total number of features.

Random Forests is an ensemble building technique that works well even with noisy content in training dataset and is considered as one of the most competitive and robust methods that can be compared to bagging or boosting [15].

4 Decorate

One of the recently proposed ensemble building techniques that could also be seen as a somehow alternative approach as it significantly differs from the above described techniques is called DECORATE (Diverse Ensemble Creation by Oppositional Relabeling of Artificial Training Examples) [16]. Base classifiers are built using additional artificially constructed training examples. These examples are given outcome labels that disagree with the current decision of the committee, thereby directly increasing diversity of classifiers within the committee.

5 Rotation Forest

One of the most recent classification techniques not only in bioinformatics but also in the machine learning field is called Rotation Forest and was developed by Rodriguez et al. [17]. Rotation Forest classifier was introduced to gene expression classification problems by Stiglic et al. in [18], where it was used as a meta-classifier for meta-classification scheme and applied to 14 different gene expression classification problems. This chapter presents the potential of Rotation Forest in the field of gene expression classification on an even wider scale and additionally presents a novel way to use the rotation of datasets using PCA transformation for building new ensembles of classifiers.

Most ensemble methods can be used with any classification method, but decision trees are one of the most commonly used. There are ensemble methods designed specifically for decision trees, such as Random and Rotation Forests. The latter is based on the sensibility of decision trees to axis rotations; the classifiers obtained with different rotations of a dataset can be very different. This sensibility is usually considered as a disadvantage, but it can be very beneficial when decision trees are used as members of an ensemble. Decision trees obtained from a rotated dataset can still be accurate, because they use all the information available in the dataset, but simultaneously they can be very diverse.

As in Bagging and Random Forests, each member of the ensemble is trained with a different dataset. These datasets are obtained from a random transformation of the original training data. This transformation produces a rotation of the axis.

The transformed dataset has as many examples as the original dataset. All the information that was in the original dataset remains in the transformed dataset, because none of the components is discarded and all the training examples are used for training all classifiers in an ensemble.

Number of features in each group (or number of groups) is a parameter of the method. The optimal value for this parameter depends on the dataset and it could be selected with an internal cross validation. Nevertheless, in this work the default value was used, and groups were formed using three features.

The elimination of classes and examples of the dataset is done because PCA is a deterministic method, and it would not be difficult (especially for big ensembles) that some members of the ensemble had the same (or very similar) grouping of variables. Hence, an additional source of diversity was needed. This elimination is only done for the dataset used to do PCA, while training of ensemble classifiers is done using all examples.

5.1 Rotation of Random Forests

This chapter compares basic and most widely used ensemble building techniques with a novel technique called Rotation of Random Forests (RRF).

Random Forest is based on bagging, using Random Trees as base classifiers. It is also possible to use the Rotation Forest method using Random Trees as base classifiers. We call this method Rotation of Random Forests. The same relationship could be expected between Rotation of Random Forest and Rotation Forest and then between Random Forest and Bagging, that is, the base classifiers will be less accurate but more diverse and this could be beneficial for the ensemble.

On the other hand, one of the advantages of Random Forests over Rotation Forest is that the former are faster. Using Rotation Forest with Random Trees could reduce this time difference, because it is also possible to construct several Random Trees in the same rotated space. This is the equivalent to a Rotation Forest ensemble using Random Forest as base classifiers.

In this chapter, the following configuration is used: ten rotations, with a ten-tree Random Forest classifier in each rotation.

6 Rotation of a Single Decision Tree Example

This section shows an example to illustrate the procedure of single decision tree rotation that represents an integral part of Rotation Forest consisting of several such trees. Decision tree is built using DLBCL-Tumour dataset. Before constructing the forest, a set of features is selected. For this example, ReliefF [19] selection method was used. For simplicity reasons, only six features were selected. The selected features are M14328_s, X02152, X12447, L19686_rna1, J04988 and J03909 that represent gene identifiers from DLBCL-Tumour dataset.

The features are randomly grouped. In our example, for one of the trees in the forest, one of the groups is formed by M14328_s, X12447_at and L19686_rnal; the other group includes the remaining features. Now we have two datasets, one for each group. For each of these datasets, a random proper subset of the classes is selected, and the examples of the classes from the subset are removed from the dataset. For a two-class dataset, a proper subset has zero or one class, and at least the examples of one of the classes will remain in the dataset. From the remaining dataset, a subset of 25% randomly selected examples is removed. Then, PCA is applied to the resulting datasets. The result of PCA is a set of components. For the first group of features, in a particular run, these components were as follows:

$$
\begin{aligned}
f_{11} &= 0.580 \times M14328\_s + 0.579 \times X12447 + 0.574 \times L19686\_rnal \\
f_{12} &= 0.814 \times L19686\_rnal - 0.483 \times X12447 - 0.323 \times M14328\_s \\
f_{13} &= -0.748 \times M14328\_s + 0.657 \times X12447\_at + 0.093 \times L19686\_rnal.
\end{aligned}
$$

And for the second group:

$$
\begin{aligned}
f_{21} &= 0.583 \times X02152 + 0.581 \times J03909 + 0.568 \times J04988 \\
f_{22} &= -0.822 \times J04988 + 0.438 \times J03909 + 0.365 \times X02152 \\
f_{23} &= 0.726 \times X02152 - 0.686 \times J03909 - 0.043 \times J04988.
\end{aligned}
$$

All these components define a new set of features. The original dataset is then transformed using these components and consists of this new set of features. The transformed dataset will be used to construct the tree. Note that the removal of classes and examples is done only for PCA transformation, while the transformed dataset contains all the training examples. Obtained decision tree is shown in Fig. 1. A Rotation Forest classifier is formed by several decision trees obtained following the procedure described above.

In the case of Rotation of Random Forest method, for each transformed dataset, a Random Forest is constructed.

7 Results

In this section, we extensively compare Rotation Forest and RRF with other methods in the literature on public gene expression datasets. Dimensionality of each dataset is reduced before classification using ReliefF feature selection method. RelieF was recently used in an extensive study by Symons and Nieselt [20], where it was selected as the most effective feature selection method for gene expression classification.

One of the most problematic limitations of supervised classification methods is overfitting to training set of examples. Especially in cases where learning is

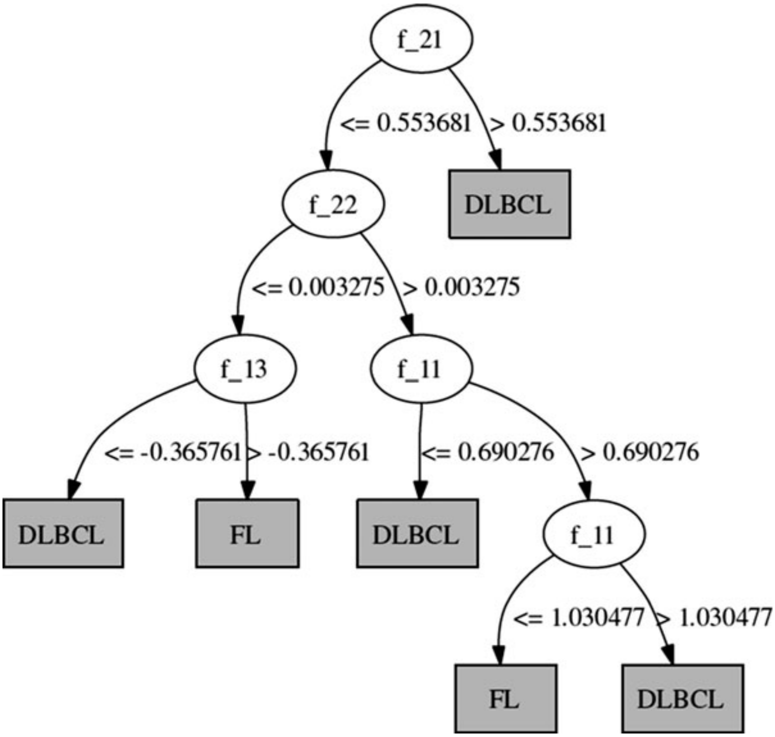

Fig. 1 An example of a single rotation forest decision tree

performed too long (e.g. neural networks) or where training examples are rare, the classifier may adjust to specific random variations of the training data that do not represent true relationships in test data. To solve this problem, we should try to get the most unbiased estimation of classifier accuracy. This is possible by using a so-called external validation, although there are still many papers that do not use that kind of validation and are presenting overly optimistic accuracy rates. The most appropriate methods, according to a study by Ambroise and McLachlan [21], are repeated k-fold cross-validation [22] or suitably defined bootstrap validation methods [23].

Following recommendations from Ambroise and McLachlan, a tenfold cross-validation was used in all experiments. As there were many classifiers that are using randomness in classifier building process, each tenfold cross-validation was repeated 20 times. Instances were randomly shuffled and stratified according to the class values of samples before they were divided into ten subsets, also called folds. The same procedure was done for all tests except the last experiment where 5 × 10-fold cross-validation was used because of very high computational complexity of testing procedure.

Table 1 Kent ridge repository datasets overview

Dataset	Source	Genes	Patients	Classes
ALL	Yeoh et al. [25]	12,558	327	7
ALLAML	Golub et al. [26]	7,129	72	2
Breast	Van't Veer et al. [27]	24,481	97	2
CNS	Mukherjee et al. [28]	7,129	60	2
Colon	Alon et al. [29]	2,000	62	2
DLBCL	Alizadeh et al. [30]	4,026	47	2
DLBCL-NIH	Rosenwald et al. [31]	7,399	240	2
DLBCL-Tumor	Shipp et al. [32]	6,817	77	2
Lung	Gordon et al. [33]	12,533	181	2
Lung-Harvard	Bhattacharjee et al. [34]	12,600	203	5
Lung-Michigan	Beer et al. [35]	7,129	96	2
MLL	Armstrong et al. [36]	12,582	72	3
Ovarian	Petricoin et al. [37]	15,154	253	2
Prostate	Singh et al. [38]	12,600	102	2

7.1 Data

For effective evaluation of proposed classification methods, a set of 14 gene expression datasets was used. All of them can be downloaded from Kent Ridge Biomedical Data Set Repository [24] and represent different biomarker classification problems originating from microarray or mass spectrometry-based studies. Table 1 contains more details on datasets that can also be found at the above-mentioned repository. Most datasets from this repository are well known, which were used in many studies where classification accuracy was important, and they served for benchmarking the proposed classification methods.

7.2 Classification Accuracy

An important issue in classification performance of gene expression classifiers is the number of selected genes before the classification algorithm is applied. It is well known that feature selection can significantly improve the performance of classification and is an integral part of most gene expression classification schemes. Recent studies [20, 39] show that the most effective classifiers achieve the highest accuracy rates with number of features between 100 and 500 genes. Although feature selection does not play a significant role when comparing classifiers, two different feature selection settings were used in our experiments where 100 and 250 features were selected during cross-validations.

In this study, a set of widely used ensemble building classification methods including Random Forests is compared to two proposed methods – Rotation Forests and RRF. The first one is a novel classification technique in bioinformatics, while the second one proposes an improvement of already proved Random Forests. Compared ensembles of classifiers were built using 100 classifiers inside Weka [40] machine learning framework. Implementations of all methods can be found in Weka environment.

Initially, a comparison of both feature selection methods using 100 and 250 best ranked features was done using average ranks from Friedman test. Results for all six ensemble building methods from 20 × 10-fold cross-validation were used. Rotation Forest, RRF and Random Forests achieve the highest ranks. When comparing feature selection method settings, it can be noted that ReliefF with 250 selected features achieves slightly better results. More detailed results representing averaged 20 × 10-fold cross-validation accuracy rates for ReliefF-250 are presented in Table 2. To directly compare three most accurate ensembles of classifiers, a set of pair-wise Wilcoxon tests was done where Random Forests, Rotation Forest and RRF were compared.

There were no significant differences in accuracy of compared classification methods (using $p < 0.05$) when results from ReliefF-250-based feature selection were used. On the other hand, when compared to the rest of the methods a significant difference can be observed.

Even though there are no statistically significant differences when average accuracy is compared between the three most successful methods, one can observe the superiority of Rotation Forest and RRF by observing the average rank. It can also be observed that Rotation Forest outperforms Random Forest in 9 of 14 datasets when average accuracy is compared. RRF goes even further and can outperform Random Forest in 10 of 14 dataset.

Table 2 Classification accuracy for ReliefF feature extraction using 250 top ranked genes (lower is better)

Dataset	Bagging	Boosting	Decorate	Random Forests	Rotation Forest	RRF
ALL	87.85 ± 5.2	90.86 ± 4.5	86.13 ± 5.2	89.67 ± 4.1	91.38 ± 4.2	88.97 ± 4.4
ALLAML	92.53 ± 9.9	89.67 ± 11.2	94.97 ± 7.8	97.07 ± 6.3	95.55 ± 6.9	97.63 ± 5.2
Breast	64.57 ± 14.4	64.45 ± 14.4	63.74 ± 13.9	68.02 ± 13.6	68.23 ± 13.3	68.42 ± 13.4
CNS	63.67 ± 15.5	60 ± 18.2	59.5 ± 17.9	64.42 ± 13.6	60.58 ± 15.5	65.33 ± 15.8
Colon	83.98 ± 12.3	77.79 ± 14.2	83.54 ± 12.3	83.27 ± 13.3	84.26 ± 12.6	84.24 ± 13.2
DLBCL	89.25 ± 13.4	86.33 ± 16	92.48 ± 12	94.15 ± 9.8	93.58 ± 11.3	94.33 ± 10.3
DLBCL-NIH	61.5 ± 7.8	60.02 ± 8.2	60.77 ± 9	62.21 ± 8.3	62.13 ± 7.9	61.83 ± 8.5
DLBCL-Tumor	89.24 ± 9.6	88.71 ± 10.4	90.08 ± 9.4	92.54 ± 8.8	95.66 ± 7	94.54 ± 7.8
Lung	97.49 ± 3.3	95.58 ± 4.9	99.18 ± 2	99.17 ± 2	99.23 ± 1.9	99.23 ± 1.9
Lung-Harvard	92.43 ± 5.5	93.89 ± 4.9	92.83 ± 5.1	92.79 ± 4.9	93.57 ± 5	92.77 ± 5.1
Lung-Michigan	98.92 ± 3.2	98.92 ± 3.2	97.96 ± 4.1	99.42 ± 1.6	100 ± 0	99.64 ± 1
MLL	92.3 ± 10.1	92.27 ± 10.6	94.92 ± 7.3	94.82 ± 7.8	93.56 ± 8.4	93.38 ± 8.9
Ovarian	98.08 ± 2.3	98.42 ± 2.2	98.68 ± 2	99.31 ± 1.6	99.72 ± 0.8	99.35 ± 1.4
Prostate	90.49 ± 9.8	90.98 ± 8.8	91.59 ± 8.9	93.26 ± 8.1	93.6 ± 7.8	93.46 ± 7.7
Friedman's Avg. Rank	4.44	4.54	4.08	2.88	2.45	2.66

8 Conclusions

This chapter presents a novel Rotation Forest-based classification method for genomic and proteomic classification problems. It also includes experiments comparing leading supervised data mining approaches on a wide range of gene expression classification problems. The results, based on average ranks and average accuracy, indicate that Rotated Random Forest and Rotation Forest can be considered as two of the most accurate ensembles of classifiers in gene expression classification. It is also evident that both methods can improve the performance of Random Forest classifier that was described by Diaz-Uriarte et al. [1] as a "part of the standard tool-box of methods for the analysis of microarray data".

The main motivation for the use of Rotations of Random Forest is that they are faster than Rotation Forests; in this chapter, the used configuration was ten rotations for ensemble, with a ten-tree Random Forest for each rotation. If time is not a cause of main concern, it is possible to construct each Random Tree in a different rotated space. This would improve the diversity among base classifiers without a penalty in their accuracy. It remains to be seen if this configuration would be better than Rotation Forests.

In the near future, when microarray datasets with larger number of samples will become available, Rotated Random Forest can become a rational solution to classification problems, because of low time complexity compared to more accurate Rotation Forest. It should be noted that both methods proved to return more reliable results in comparison to Random Forests in most cases.

References

1. Díaz-Uriarte R, Alvarez de Andrés S (2006) Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7:3
2. Vapnik V (1998) Statistical learning theory. John Wiley and Sons, New York
3. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16:906–914
4. Wu W, Xing E, Mian I, Bissell M (2005) Evaluation of normalization methods for cdna microarray data by k-nn classification. BMC Bioinformatics 6(191):1–21
5. Dudoit S, Fridlyand J (2003) Classification in microarray experimentse. In: Speed T (Ed.), Statistical analysis of gene expression microarray data. Interdisciplinary statistics. Chapman & Hall/CRC, Virginia Beach, 93–158
6. Seiffert U, Hammer B, Kaski S, Villmann T (2006) neural networks and machine learning in bioinformatics – theory and applications. In: Proceedings of the 14th European Symposium on Artificial Neural Networks ESANN 2006, 521–532
7. Cunningham, P. (2007) Ensemble Techniques. Technical Report UCD-CSI-2007–5
8. Breiman L (1996) Bagging predictors. Machine Learning 24:123–140
9. Efron B, Tibshirani R (1994) An introduction to the bootstrap. Chapman & Hall/CRC, Virginia Beach
10. Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. In: Proceedings of the 13th International Conference on Machine Learning, 148–156

11. Rätsch G, Onoda T, Müller KR (2001) Soft margins for AdaBoost. Machine Learning 42(3):287–320
12. Ho, TK (1995) Random decision forest. In: Proceedings of the 3rd Int'l Conf on Document Analysis and Recognition, Montreal, Canada, August 14–18, 1995, 278–282
13. Amit Y, Geman D (1997) Shape quantization and recognition with randomized trees. Neural Computation 9:1545–1588
14. Breiman L (2001) Random forests. Machine Learning 45:5–32
15. Dietterich TG (2002) Ensemble learning. In: Arbib MA (Ed.) The handbook of brain theory and neural networks, 2nd ed. The MIT Press, Cambridge, MA, 405–408
16. Melville P, Mooney RJ (2003) Constructing diverse classifier ensembles using artificial training examples. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI 2003), 505–510
17. Rodríguez JJ, Kuncheva LI, Alonso CJ (2006) Rotation forest: A new classifier ensemble method. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(10):1619–1630
18. Stiglic G, Kokol P (2007) Effectiveness of rotation forest in meta-learning based gene expression classification. In: Proceedings of the 20th IEEE International Symposium on Computer-Based Medical Systems (CBMS 2007), 243–250
19. Robnik-Sikonja M, Kononenko I (1997) An adaptation of relief for attribute estimation in regression. In: Machine Learning: Proceedings of the Fourteenth International Conference (ICML'97), 296–304
20. Symons S, Nieselt K (2006) Data mining microarray data – comprehensive benchmarking of feature selection and classification methods (available at: www.zbit.unituebingen.de/pas/preprints/GCB2006/SymonsNieselt.pdf)
21. Ambroise C, McLachlan GJ (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. Proceedings of the National Academy of Sciences of the United States of America 99:6562–6566
22. Burman P (1989) A comparative study of ordinary cross-validation, v-fold cross-validation and repeated learning-testing methods. Biometrika 76:503–514
23. Efron B, Tibshirani R (1997) Improvements on cross-validation: the. 632 + bootstrap method. Journal of the American Statistical Association 92:548–560
24. Li J, Liu H (2003) Ensembles of cascading trees. In: Proceedings of the IEEE ICDM 2003 Conference 585
25. Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz R, Behm FG, Raimondi SC, Relling MV, Patel A, Cheng C, Campana D, Wilkins D, Zhou X, Li J, Liu H, Pui CH, Evans WE, Naeve C, Wong L, Downing JR (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell 1:133–143
26. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286:531–537
27. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH (2002) Gene expression profiling predicts clinical outcome of breast cancer. Letters to Nature, Nature 415:530–536
28. Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, Kim JY, Goumnerova LC, Black PM, Lau C, Allen JC, Zagzag D, Olson JM, Curran T, Wetmore C, Biegel JA, Poggio T, Mukherjee S, Rifkin R, Califano A, Stolovitzky G, Louis DN, Mesirov JP, Lander ES, Golub TR (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. Letters to Nature, Nature 415:436–442
29. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of National Academy of Sciences of the United States of America 96:6745–6750

30. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403:503–511
31. Rosenwald A, Wright G, Chan W, Connors JM, Campo E, Fisher R, Gascoyne RD, Muller-Hermelink K, Smeland EB, Staut LM (2002) The use of molecular profiling to predict survival after themotheropy for diffuse large-B-cell lymphoma. The New England Journal of Medicine 346:1937–1947
32. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, Ray TS, Koval MA, Last KW, Norton A, Lister TA, Mesirov J, Neuberg DS, Lander ES, Aster JC, Golub TR (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature Medicine 8:68–74
33. Gordon GJ, Jensen RV, Hsiao LL, Gullans SR, Blumenstock JE, Ramaswamy S, Richards WG, Sugarbaker DJ, Bueno R (2002) Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Research 62(17):4963–4967
34. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M (2002) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinomas subclasses. Proceedings of National Academy of Sciences of the United States of America 98:13790–13795
35. Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, Taylor JM, Iannettoni MD, Orringer MB, Hanash S (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nature Medicine 18(8):816–824
36. Armstrong SA, Staunton JE, Silverman LB, Pieters R, den Boer ML, Minden MD, Sallan SE, Lander ES, Golub TR, Korsmeyer SJ (2002) MLL Translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nature Genetics 30:41–47
37. Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC, Liotta LA (2002) Use of proteomic patterns in serum to identify ovarian cancer. The Lancet 359:572–577
38. Singh D, Febbo P, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Loda M, Kantoff PW, Golub TR, Sellers WR (2002) Gene expression correlates of clinical prostate cancer behaviour. Cancer Cell 1(2):203–209
39. Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S (2005) A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics 21:631–643
40. Witten IH, Frank E (2005) Data mining: practical machine learning tools with Java implementations. Morgan Kaufmann, Massachusetts