Feature Selection and Classification for Small Gene Sets

Gregor Stiglic[1,2], Juan J. Rodriguez[3], and Peter Kokol[1,2]

[1] University of Maribor, Faculty of Health Sciences, Zitna ulica 15, 2000 Maribor, Slovenia
[2] University of Maribor, Faculty of Electrical Engineering and Computer Science, Smetanova 17, 2000 Maribor, Slovenia
{gregor.stiglic,kokol}@uni-mb.si
[3] University of Burgos, c/ Francisco de Vitoria s/n, 09006 Burgos, Spain
jjrodriguez@ubu.es

Abstract. Random Forests, Support Vector Machines and k-Nearest Neighbors are successful and proven classification techniques that are widely used for different kinds of classification problems. One of them is classification of genomic and proteomic data that is known as a problem with extremely high dimensionality and therefore demands suited classification techniques. In this domain they are usually combined with gene selection techniques to provide optimal classification accuracy rates. Another reason for reducing the dimensionality of such datasets is their interpretability. It is much easier to interpret a small set of ranked genes than 20 or 30 thousands of unordered genes. In this paper we present a classification ensemble of decision trees called Rotation Forest and evaluate its classification performance on small subsets of ranked genes for 14 genomic and proteomic classification problems. An important feature of Rotation Forest is demonstrated – i.e. robustness and high classification accuracy using small sets of genes.

Keywords: Gene expression analysis, machine learning, feature selection, ensemble of classifiers.

1 Introduction

There are many new classification methods and variants of existing techniques for classification problems. One of them is Random Forests classifier that was presented in [1] by Breiman and Cutler. It has proven to be fast, robust and very accurate technique that can be compared with the best classifiers (e.g. Support Vector Machines [2] or some of the most efficient ensemble based classification techniques) [3]. Most of these techniques are also used in genomic and proteomic classification problems where classifiers need to be specialized for high dimensional problems. The other option is integration of feature pre-selection into classification process where initial feature set is reduced before the classification is done. Most of the early experiments using microarray gene expression datasets used simple statistical methods of gene ranking to reduce the initial set of attributes. Recently more advanced feature selection methods from the machine learning field are applied to pre-selection step in genomic and proteomic classification problems. Although a small number of genes is

M. Chetty, A. Ngom, and S. Ahmad (Eds.): PRIB 2008, LNBI 5265, pp. 121–131, 2008.

preferred, we try to avoid extremely small subsets of genes, like Wang et al. [4], where subsets with only two or three genes were used for classification.

This paper attempts to evaluate two widely used feature selection techniques to determine the most appropriate number of features that should be retained in preselection step to achieve the best classification performance. Additionally, this paper introduces one of the most recent classification techniques called Rotation Forest [5] to genomic and proteomic classification using small sets of genes.

Section 2 of this paper presents a novel ensemble based classification model called Rotation Forests. The rest of the paper is organized as follows: in section 3 we review the feature selection and classification methods used in this paper, in section 4 we present results of our experiments comparing classification accuracy of Rotation Forests to three classification methods. Section 5 concludes the paper and gives some future research directions on usage of Rotation Forests in genomic and proteomic classification problems.

2 Rotation Forest

Rotation Forest is a novel classification technique that was initially presented by Rodriguez et al. [4] and applied to several machine learning problems. In order to obtain successful ensembles, the member classifiers have to be accurate and diverse. Because of sampling process in Bagging and Random Forests it is necessary to obtain diverse classifiers, but using a subset of the examples to train the classifiers can degrade the accuracy of the member classifiers. Hence, a natural question is if it is possible to obtain diverse classifiers without discarding any information in the dataset.

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 the trees are used as members of an ensemble. The 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. In Rotation Forests, the transformation of the dataset consists of the following steps:

- Features are randomly grouped in k groups.
- For each group of features:
  - A new dataset consisting of all examples using sets of features from step one is created.
  - All examples of randomly selected classes are removed from this new dataset.
  - A subset of randomly chosen examples is eliminated from the new dataset (by default 25% of samples are removed)

    - PCA (Principal Component Analysis) is applied to the remaining samples in a dataset.
    - PCA components are considered as a new set of features. None of the components is discarded.
- All training samples are transformed using new variables selected by PCA for each group.
- A classifier is built from transformed training set.
- Another classifier is build by returning to the first step in case final number of classifiers in ensemble is not reached.

This transformation produces a rotation of the axis. The transformed dataset has as many examples as the original dataset and 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 the ensemble methods.

The number of features in each group (or the 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 3 features. The selection of the optimal value of this parameter would increase notably the time necessary for the training of the classifiers and would give an advantage of Rotation Forests with respect to other ensemble methods that do not optimize the value of any parameters.

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; all the examples are used for training the classifiers in the ensemble.

3 Feature Selection and Classification Techniques

The main idea of feature selection is to choose a subset of variables that can significantly improve the time complexity and accuracy of a classification model. This is even more important in microarray based classification problems where initial set of features consists of thousands of gene expression values. With such a large amount of features it is of special interest to search for a dependency between optimal number of selected features and accuracy of classification model. There are two groups of feature selection techniques – filter and wrapper based methods [5]. Filter based methods rely on information content of features. Different metrics like distance metrics, information measures, correlation and consistency metrics can be used to get useful subsets when filter based feature selection is used. In wrapper approach subsets of features are selected based on how well those features classify training samples. The selection is done using the induction algorithm as a black box. Usually a search for a quality subset is done using the induction algorithm itself as a part of the evaluation function.

Symons and Nieselt [6] showed that in most microarray gene expression classification problems, filter based approaches outperform wrapper based approaches. In our experiments the following filter based approaches were used:

- ReliefF
- Support Vector Machine Recursive Feature Elimination (SVM-RFE)

Additional to two feature selection methods a set of four classification techniques was used in experiments presented in this paper:

- Random Forests
- Rotation Forests
- Support Vector Machines (SVM)
- k-Nearest Neighbors (k-NN)

A machine learning software framework named Weka [7] was used for all experiments described in this paper. Each of the above mentioned methods, except self-developed Rotation Forest algorithm is already implemented in Weka.

All above mentioned methods except Rotation Forest, that were explained earlier, are briefly described in the remainder of this section.

3.1 ReliefF

ReliefF feature selection algorithm is based on original Relief algorithm [8] that could only be used for classification problems with two class values. Basic idea of Relief algorithm is ranking of features based on their ability to distinguish between instances that are near to each other. Original algorithm was extended by Kononenko [9] so that it can deal with multi-class problems and missing values. Later it was further improved by Robnik-Sikonja and Kononenko [10] so that it is suitable for noisy data and can also be used for regression problems. Default settings for Weka implementation of ReliefF that also supports feature selection for regression problems were used in our study.

3.2 Support Vector Machines - Recursive Feature Elimination (SVM-RFE)

SVM in combination with Recursive Feature Elimination (SVM-RFE) were introduced to gene selection in bioinformatics by Guyon et al. [11]. SVM-RFE feature selection method is based on linear SVM used as the learning algorithm in recursive selection of nested subsets of features. In the final step of each cycle, all feature variables are ranked and a pre-selected number of the worst ranked features are eliminated. By default a single feature is eliminated in each round, it is also possible to remove more than one feature per round. In our experiment a setting where 50% of the remaining features are removed in each step was used.

3.3 Random Forests

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 then grown by selecting the feature to split on at each node from randomly selected

number of features. Number of chosen features is set to $log_2(k+1)$ as in [12], where k is the total number of features.

Random Forests is an ensemble building method that works well even with noisy content in the training dataset and is considered as one of the most competitive methods that can be compared to boosting [13].

3.4 Support Vector Machines (SVM)

SVM are increasingly popular classifiers in many areas, including bioinformatics [2]. The most basic variant of SVM use linear kernel and try to find an optimal hyperplane that separates samples of different classes. When classes can be linearly separated, the hyperplane is located so that there is maximal distance between the hyperplane and the nearest sample of any of the classes. In cases when samples cannot be linearly separated, there is no optimal separating hyperplane; in such cases, we try to maximize the margin but allow some classification errors. For all experiments in this study an advanced version of SVM called Sequential Minimal Optimization (SMO) proposed by Platt [14, 15] is used. It offers very quick and reliable learning of the decision models based on SVM.

3.5 k-Nearest Neighbors (k-NN)

Nearest Neighbors classifier is a typical representative of case based classifiers where all samples are stored for later use in the classification process [16]. It aims to classify samples according to similarities or distance between them. A class value is defined using class values of k nearest samples. Similarity to neighboring samples is calculated using distance between samples that is usually measured using Euclidean distance metric.

Another important parameter that has to be set is number of neighbors that will be used for calculation of class value. The most common settings for this parameter are 1, 3 or 5. In our experiments we always use 5 neighbors for class value estimation whose vote for final class is weighted according to their distance from the neighbor.

k-NN based classifiers are most useful in cases with continuous attribute values that also include genomic and proteomic datasets. It is also welcome if datasets contain low number of samples (e.g. gene expression datasets), because of high computational cost of k-NN classification process when number of samples rises.

4 Experiment Settings and Results

In our experiments two feature selection methods from section 3 were tested on 14 publicly available genomic and proteomic datasets presented in Table 1. No modification of original data in form of normalization or discretization was needed. All datasets are available at Kent Ridge Biomedical Data Set Repository [17] where additional information including references to original work for each of the datasets can be found. All tests were done using 10-fold cross-validation measuring the classification accuracy that can be calculated as a quotient between number of correctly classified and number of all samples in a testing set. To avoid feature selection bias, as discussed in Ambroise and McLachlan [18], a separate feature selection process was done for each training and test set during 10-fold cross validation.

Table 1. Details for genomic and proteomic datasets from Kent Ridge repository

Dataset	Original Work	Genes	Patients	Classes
ALL	Yeoh et al.	12558	327	7
ALLAML	Golub et al.	7129	72	2
Breast	Van't Veer et al.	24481	97	2
CNS	Mukherjee et al.	7129	60	2
Colon	Alon et al.	2000	62	2
DLBCL	Alizadeh et al.	4026	47	2
DLBCL-NIH	Rosenwald et al.	7399	240	2
DLBCL-Tumor	Shipp et al.	6817	77	2
Lung	Gordon et al.	12533	181	2
Lung-Harvard	Bhattacharjee et al.	12600	203	5
Lung-Michigan	Beer et al.	7129	96	2
MLL	Armstrong et al.	12582	72	3
Ovarian	Petricoin et al.	15154	253	2
Prostate	Singh et al.	12600	102	2

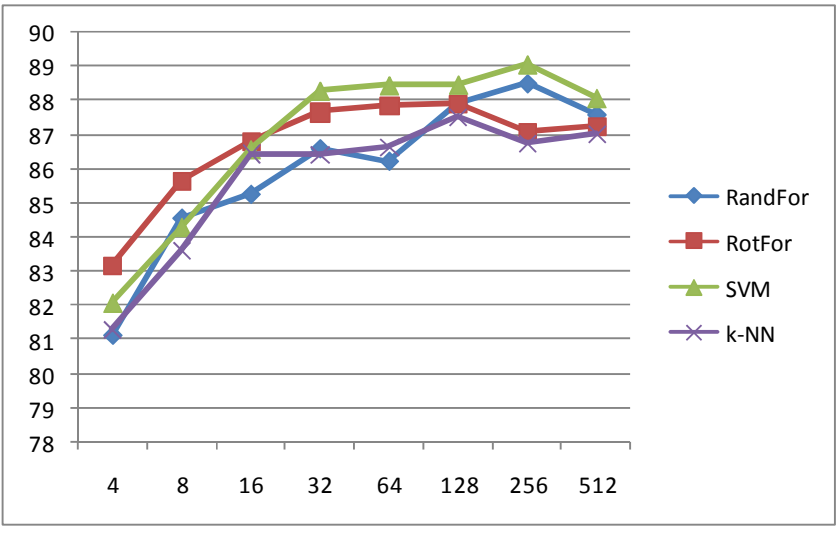

Fig. 1. Average accuracy on all datasets using four classification methods on reduced datasets with different number of genes (ReliefF feature selection)

Each ensemble (Random Forests and Rotation Forest) consisted of 100 decision trees, while number of features used for classification ranged from 4 to 512 and was defined as 2^i, where $i = 2,\ldots,9$.

In the first experiment a set of classification accuracy measurements was done based on reduced gene sets. ReliefF was used for feature selection using default settings of Weka environment. Averaged classification accuracy for specific feature

selection settings using k-top most ranked genes, where k ranges from 4 to 512, is presented in Figure 1. It can be observed that with number of selected features under 16, Rotation Forest outperforms all other methods, while SVM take over for higher numbers of selected genes. The highest classification accuracy was obtained using 256 most significant genes according to ReliefF, using SVM classifier (89.06%).

To obtain a better picture of dominance between compared methods and to avoid unreliable averaging of results, we did a comparison using statistical test. Non-parametric Friedman's statistical test [19] was used to compute average ranks of compared methods. Figure 2 presents Friedman's average ranks for all compared classifiers and different feature selection settings using ReliefF. It can be seen that Rotation Forest strongly dominates all other methods in the first three points, while SVM strongly dominate Rotation Forest in the last two settings. Average ranks shown in Figure 2 were calculated for results from all 14 datasets using SPSS statistical tools. Average rank, in our case of four compared methods, can have a value from 1 to 4. If a method hypothetically wins all comparison tests based on average accuracy it would be assigned an average rank of 4, while method losing all pairwise comparisons would score an average rank of 1.

The same settings as in the first experiment were used for the second experiment where SVM-RFE was used for feature selection tasks. Figure 3 presents results of average accuracy levels across all 14 datasets. It can be observed that Rotation Forest classifier is in front all the way up to the point of the highest classification accuracy at 128 selected genes (89,51% accuracy). Similar to the previous experiment the performance of Rotation Forest deteriorates with high numbers of selected features, where SVM perform better again.

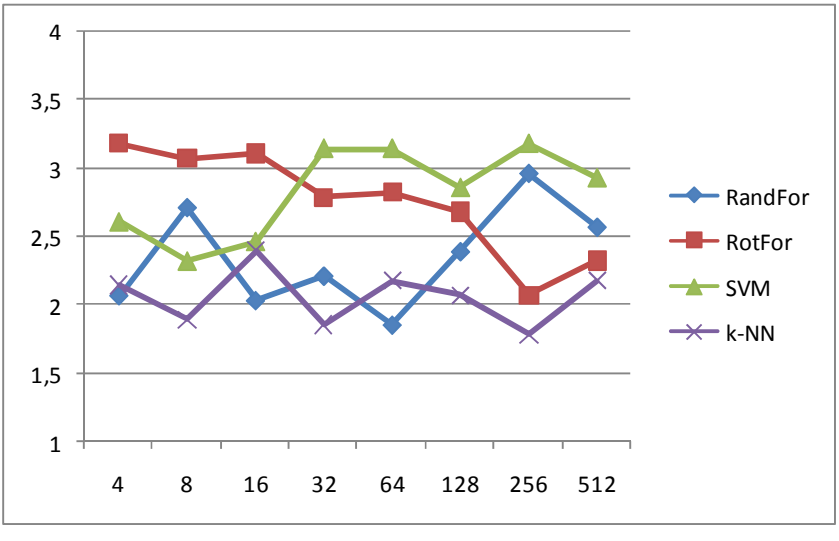

Fig. 2. Average rank for all four classification methods on reduced datasets with different number of genes using ReliefF feature selection

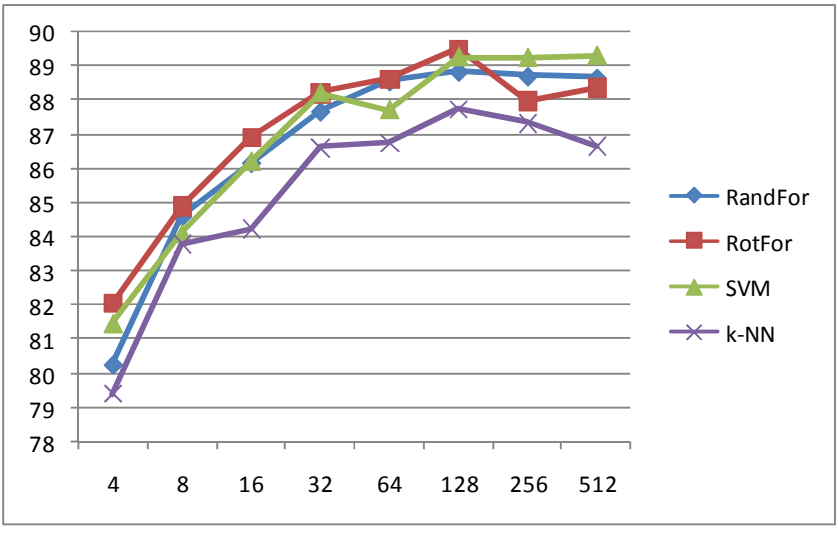

Fig. 3. Average accuracy using SVM-RFE based feature selection

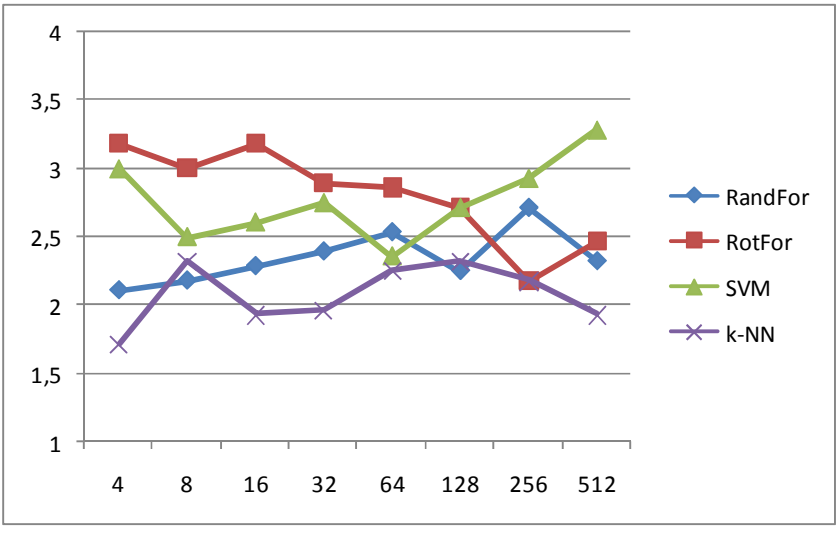

Fig. 4. Average ranks using SVM-RFE based feature selection

Friedman test shows significant differences among compared methods again. When Friedman test hypothesis is rejected, it is usually followed by a pairwise comparison of classification methods. This can be done by Wilcoxon signed-rank test [20] that represents non-parametric alternative to the paired Student t-test for two related measurements.

Wilcoxon test was done on all pairwise combinations in the first ReliefF based and the second SVM-RFE based experiments. In case of ReliefF results a significant dominance of Rotation Forest and SVM methods compared to Random Forests and k-NN is shown. However there is no significant difference between Rotation Forest and SVM ($p = 0.828$). There is also no significant difference between results of Random Forests and k-NN ($p = 0.167$). Results were almost the same for SVM-RFE feature selection with the only exception – Rotation Forest did not manage to significantly outperform Random Forests ($p = 0.066$) although it was performing better.

Figure 4 confirms results from Figure 3 where it can be seen that Rotation Forest dominates all other classification methods up to the and including a point where 128 most significant genes were selected.

Given the highest accuracy of 89.51% one would assume that a combination of Rotation Forest and SVM-RFE based feature selection using 128 most significant genes is the best combination. But in many cases biologists are interested in smaller sets of genes that can be more descriptive and give more information than large sets of genes that are difficult to interpret.

Figure 5 shows a combination of both best methods (Rotation Forest and SVM) using average accuracy levels for both feature selection techniques simultaneously. One should notice that although SVM-RFE achieves better average accuracy overall, it is evident that ReliefF should be preferred when a small number of selected genes should be obtained.

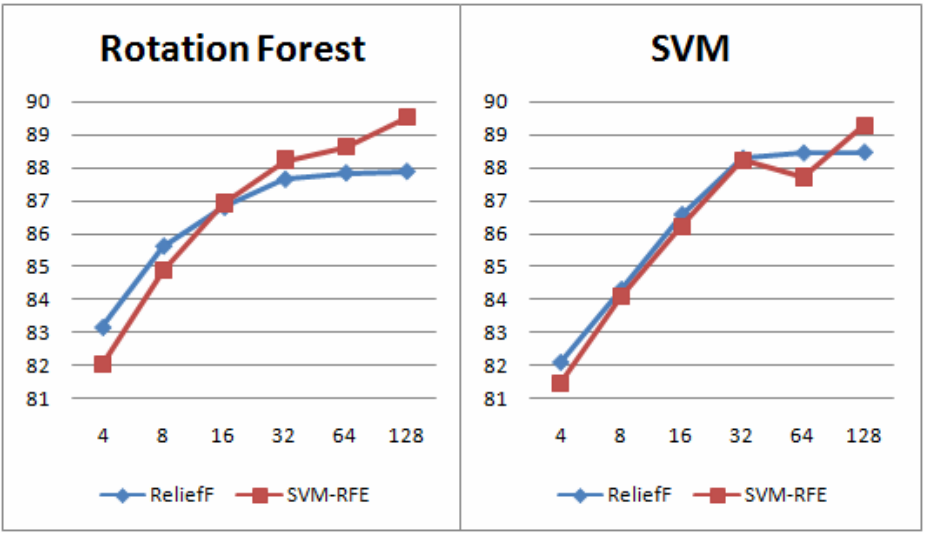

Fig. 5. Simultaneous comparison of ReliefF and SVM-RFE feature selection techniques

5 Conclusions

This paper presents a novel classification method for genomic and proteomic data classification. The results indicate that Rotation Forests can be considered as one of

the most useful classification techniques on small gene sets. Another important issue that was researched in this paper is a problem of finding the optimal number of genes to get the most out of the classifier. It was shown that there is no optimal solution to this problem. One can get significantly different results when comparing classification accuracy when an extremely low number of genes is used to classification accuracy in higher dimensional problems using different classifiers. It is however practically impossible to define a fixed number of features that should be selected for optimal classification performance. On the other hand it was obvious that there are some classification techniques that should be used when a low number of genes is preferred (Rotation Forest) and some methods that demand higher number of genes (SVM) for optimal classification accuracy. It was shown that ReliefF should be used for extremely small sets of selected features, while SVM-RFE performs better in higher dimensions. It should also be noticed that SVM-RFE cannot be used for regression problems, where ReliefF will be the only available solution out of the two presented feature selection methods.

One of the issues for the future is evaluation of Rotation Forests on even more datasets. Unfortunately it is not possible to directly use Rotation Forests for feature selection, but there are other ways of using the power of Rotation Forests. One of such is their ability to very accurately estimate the similarity of cases and could therefore be used for implementation of clustering algorithms. As we know clustering is also one of the most widely used methods in unsupervised analysis that is being used in bioinformatics and therefore opens a lot of new areas where Rotation Forests could be used.

References

1. Breiman, L.: Random forests. Machine Learning 45, 5–32 (2001)
2. Vapnik, V.: Statistical learning theory. John Wiley and Sons, New York (1998)
3. Caruana, R., Niculescu-Mizil, A.: An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd international Conference on Machine Learning (ICML 2006), vol. 148, pp. 161–168 (2006)
4. Wang, L., Chu, F., Xie, W.: Accurate Cancer Classification Using Expressions of Very Few Genes. IEEE/ACM Transactions on Computational Biology and Bioinformatics 4(1), 40–53 (2007)
5. Rodríguez, J.J., Kuncheva, L.I., Alonso, C.J.: Rotation forest: A new classifier ensemble method. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(10), 1619–1630 (2006)
6. John, G.H., Kohavi, R., Pfleger, K.: Irrelevant Features and the Subset Selection Problem. In: Proceedings of the Eleventh International Conference on Machine Learning, pp. 121–129 (1994)
7. Symons, S., Nieselt, K.: Data Mining Microarray Data – Comprehensive Benchmarking of Feature Selection and Classification Methods. Pre-print, http://www.zbit.uni-tuebingen.de/pas/preprints/GCB2006/SymonsNieselt.pdf
8. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools with Java implementations. Morgan Kaufmann, San Francisco (2005)
9. Kira, K., Rendell, L.A.: A practical approach to feature selection. In: Proceedings of International Conference on Machine Learning (ICML1992), pp. 249–256 (1992)

10. Kononenko, I.: Estimating attributes: analysis and extension of relief. In: Proceedings of European Conference on Machine Learning (ICML1994), pp. 171–182 (1994)
11. Robnik-Sikonja, M., Kononenko, I.: An adaptation of Relief for attribute estimation in regression. In: Machine Learning: Proceedings of the Fourteenth International Conference (ICML 1997), pp. 296–304 (1997)
12. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learning 46, 389–422 (2002)
13. Díaz-Uriarte, R., Alvarez de Andrés, S.: Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7, 3 (2006)
14. Dietterich, T.G.: Ensemble Learning. In: Arbib, M.A. (ed.) The Handbook of Brain Theory and Neural Networks, pp. 405–408. The MIT Press, Cambridge (2002)
15. Platt, J.: Machines using Sequential Minimal Optimization. In: Schoelkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods - Support Vector Learning (1998)
16. Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K.: Improvements to Platt's SMO Algorithm for SVM Classifier Design. Neural Computation 13(3), 637–649 (2001)
17. Mitchell, T.: Machine Learning. McGraw Hill, New York (1997)
18. Kent Ridge Biomedical Data Set Repository: sdmc.i2r.a-star.edu.sg/rp/
19. Ambroise, C., McLachlan, G.J.: Selection bias in gene extraction on the basis of microarray gene-expression data. Proc. Natl. Acad. Sci. USA 2002 99, 6562–6566 (2002)
20. Friedman, M.: A comparison of alternative tests of significance for the problem of m rankings. The Annals of Mathematical Statistics 11(1), 86–92 (1940)
21. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bulletin 1, 80–83 (1945)