Neural Comput & Applic (2007) 16:527–539
DOI 10.1007/s00521-007-0110-1

ICONIP2006

Classification consistency analysis for bootstrapping gene selection

Shaoning Pang · Ilkka Havukkala · Yingjie Hu · Nikola Kasabov

Received: 22 November 2006 / Accepted: 6 March 2007 / Published online: 30 March 2007

Abstract Consistency modelling for gene selection is a new topic emerging from recent cancer bioinformatics research. The result of operations such as classification, clustering, or gene selection on a training set is often found to be very different from the same operations on a testing set, presenting a serious consistency problem. In practice, the inconsistency of microarray datasets prevents many typical gene selection methods working properly for cancer diagnosis and prognosis. In an attempt to deal with this problem, this paper proposes a new concept of classification consistency and applies it for microarray gene selection problem using a bootstrapping approach, with encouraging results.

1 Introduction

The advent of microarray technology has made it possible to monitor the expression levels for thousands of genes simultaneously, which can help clinical decision making in complex disease diagnosis and prognosis, especially for cancer classification, and for predicting the clinical outcomes in response to cancer treatment. However, often only a small proportion of the genes contribute to classification, and the rest of genes are considered as noise. Gene selection is used to eliminate the influence of such noise genes, and to find out the informative genes related to disease.

1.1 Review of gene selection methods

Selecting informative genes, as a critical step for cancer classification, has been implemented using a diversity of techniques and algorithms. Simple gene selection methods come from statistics, such as t-statistics, Fisher's linear discriminate criterion and principal component analysis (PCA) [1–4]. Statistical methods select genes by evaluating and ranking their contribution or redundancy to classification [5], and are able to filter out informative genes very quickly. Margin-based filter methods have also been introduced recently [6]. However, the performance of these methods is not satisfactory, when applied on datasets with large numbers of genes and small numbers of samples.

More sophisticated algorithms are also available, such as noise sampling method [7], Bayesian model [8, 9], significance analysis of microarrays (SAM) [10], artificial neural networks [11], and neural fuzzy ensemble method [12]. These methods define a loss function, such as classification error, to evaluate the goodness of a candidate subset. Most are claimed to be capable of extracting out a set of highly relevant genes [13], however, their computational cost is much higher than in the simpler statistical methods.

A bootstrapping approach can also be used. This can select genes iteratively in a number of iterations, and can use a diversity of criterions simultaneously. For example, Huerta et al. [14] proposed a GA/SVM gene selection method that achieved a very high classification accuracy

S. Pang (✉) · I. Havukkala · Y. Hu · N. Kasabov
Knowledge Engineering and Discovery Research Institute, Auckland University of Technology, Private Bag 92006, Auckland 1020, New Zealand
e-mail: spang@aut.ac.nz

I. Havukkala
e-mail: ihavukka@aut.ac.nz

Y. Hu
e-mail: creekhu@yahoo.com

N. Kasabov
e-mail: nkasabov@aut.ac.nz

(99.41%) on Colon Cancer data [15]. Li et al. [16] introduced a GA/KNN gene selection method that is capable of finding a set of informative genes, and the selected genes were highly repeatable. Wahde and Szallasi [17] used an evolutionary algorithm based on a gene relevance ranking, and surveyed such methods [18]. The main drawbacks of the bootstrapping methods are in the difficulties of developing a suitable post-selection fitness function and in determining the stopping criterion.

1.2 Motivation of consistency for gene selection

For a disease microarray dataset, we do not know initially which genes are truly differentially expressed for the disease. All gene selection methods seek a statistic to find out a set of genes with an expected loss of information minimized. Most previous methods work by estimating directly a 'class-separability' criterion (i.e. rank of contribution to classification, or loss of classification) for a better gene selection. In a different vein, reproducibility is addressed by Mukherjee [19] as the number of common genes obtained from the statistic over a pair of subsets randomly drawn from a microarray dataset under the same distribution.

Class-separability criteria approximate the 'ground-truth' as the class-separation status of the training set (one part of a whole dataset). However, this whole dataset normally is just a subset of a complete dataset of disease (a dataset includes all possible microarray distributions of a disease). This leads to bad reproducibility, i.e. the classification system works well on the dataset that it was built on, but fails on future data. Reproducibility criteria take advantage of certain properties of microarray data, thus they do not approximate the 'ground truth', but indirectly minimize the expected loss under true data generating distribution.

However, it is not clear to what extent the selected highly differentially expressed genes using common-gene reproducibility criterion are correlated to a substantially good performance on the classification of microarray data. In other words, an erroneous cancer classification may also occur using a set of genes which are selected under the criterion of common-gene reproducibility.

Consistency in terms of classification performance is addressed in this paper to derive a gene selection model with both good class-separability and good reproducibility. A bootstrapping consistency method was developed by us with the purpose of identifying a set of informative genes for achieving replicably good results in microarray data analysis.

This rest of the paper is organized as follows. Section 2 gives the definition of classification consistency. Section 3 derives the novel bootstrapping gene selection method based on the classification consistency. Section 4 describes cancer diagnosis experiments on six well-known benchmark cancer microarray datasets and one proteomics dataset. Finally, we present conclusions and directions for further research in Sect. 5.

2 Consistency concepts for gene selection

A microarray can contain more than ten thousands genes, but only a few samples involving different types of disease. Gene selection, similar to feature selection in the traditional pattern recognition, works for selecting a set of genes/features to achieve better patient disease diagnosis. (i.e. classification of disease). Biologists are also interested to find out those genes informative to the disease for further disease research.

Given a microarry dataset D pertaining to a bioinformatics classification task, consisting of n samples with m genes, we define D_a and D_b as two subsets of D obtained by random subsampling; these two subsets serve as training and testing data, respectively.

$$D = D_a \cup D_b \;\&\; D_a \cap D_b = \varnothing \tag{1}$$

Provided an operation F over D such as a classification or a clustering, and a gene selection function f_s for selecting a subset of genes/features over training data D_a to achieve better disease-diagnosis/classification on testing data D_b, the fundamental consistency concept C on gene selection f_s can be modelled on a pair of subsets (D_a and D_b) drawn from the whole microarray dataset D under the same distribution,

$$C(F, f_s, D) = |P_a - P_b| \tag{2}$$

where P_a and P_b are the outcomes of function F on D and D_b, respectively. Outcome of operation F on a subset D_i can be formulated as P_i,

$$P_i = F(f_s(D_i), D_i) | i = a, b. \tag{3}$$

where i is the index of a subset, because consistency in Eq. (2) represents an comparison between P_a and P_b, i here represents a or b, indicating a subset for training or testing, respectively.

Note that F basically can be any of data processing models, such as a common-gene computation, clustering function, feature extraction function, or a classification function, etc. F determines the feature space on which the consistency is based on.

2.1 Common-gene consistency

Mukherjee et al. [20] set F as a common-gene computation; their approach is as follows. Suppose f_s is a ranking function for gene selection generating two lists of sorted genes from the two datasets. Let top-ranked genes in each case be selected and denoted by S_a and S_b. Then, the consistency in terms of common-gene C_g is defined as

$$C_g(f_s, D_a, D_b) = |S_a \cap S_b| \tag{4}$$

Consistency C_g in Eq. (4) depends on the ranking function, data and number of selected genes. Hence, a greater C_g value represents a more consistent gene selection.

2.2 Classification consistency

As F is assigned as a classification function, the above consistency C in Eq. (2) is called a classification consistency C_c, where Eq. (3) can be implemented as

$$P_a = F(f_s(D), D_a, D_b), \text{ and } P_b = F(f_s(D), D_b, D_a) \tag{5}$$

Substituting Eq. (5) into Eq. (2), we have

$$C_c(F, f_s, D) = |F(f_s(D), D_a, D_b) - F(f_s(D), D_b, D_a)| \tag{6}$$

where $f_s(D)$ specifies D as the dataset for gene selection. D_a in the first term of Eq. (5) is assigned for classifier training, and D_b is for testing. The second term of Eq. (5) specifies a reversed training and testing position for D_a and D_b, respectively. Note that a smaller C_c value here represents a more consistent gene selection. Figure 1 illustrates the procedure of computing Eq. (5). First, the performance P_a is computed by one classification on training subset D_a, then, P_b is obtained by another classification on testing subset D_b.

Alternatively, Eq. (7) gives another form of the classification consistency definition, which is obtained by switching the training and testing sets of Eq. (5),

$$C_c(F, f_s, D) = |F(f_s(D), D_a, D_a) - F(f_s(D), D_a, D_b)|. \tag{7}$$

Figure 2 shows the procedure of computing Eq. (7). Here, the classifier is trained on D_b, and then the performance is computed on the other subset D_a. The important difference to the procedure in Fig. 1 is that the testing and training subsets are switched. Ideally, when doing the analysis, one should use both procedures, to check which dataset gives better consistency when used for training. This is to safeguard for the training dataset having some kind of bias resulting in suboptimal results, when the training dataset is not a truly random sample of all data.

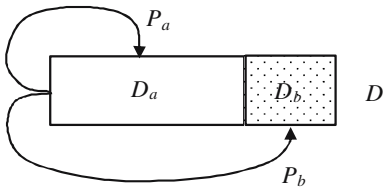

Fig. 1 Procedure of computing consistency (Form 1)

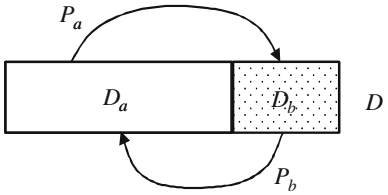

Fig. 2 Procedure of computing consistency (Form 2)

3 Gene selection based on classification consistency

Unlike common-gene based consistency [20], classification consistency needs a testing classification function F to estimate the contribution of selected genes, so that gene selection seeks an optimal f_s* with the following consistency C_c minimized,

$$f_s^* = \arg\min_{f_s \in \mathcal{F}} C_c(F, S, D) \tag{8}$$

where S is a set of currently selected genes. $\mathcal{F}$ refers to a family of gene selection functions, $C_c()$ represents a classification consistency computation that has F and f_s as classification function and gene selection function, respectively.

In practice, the evaluation of consistency eventually is a multi-objective optimizing problem, because there is a possibility that the improvement of consistency might be coupled with the deterioration of performance. This means that even if the consistency of one set of genes is better than that of another set, the performance of classification P on microarray data may not be as good as we expect. In other words, it might be a case of consistently bad classification (with low classification accuracy). Therefore, a ratio of consistency to performance R is used for the purpose of optimizing these two variables simultaneously,

$$R = \frac{C_c}{w \cdot P} \tag{9}$$

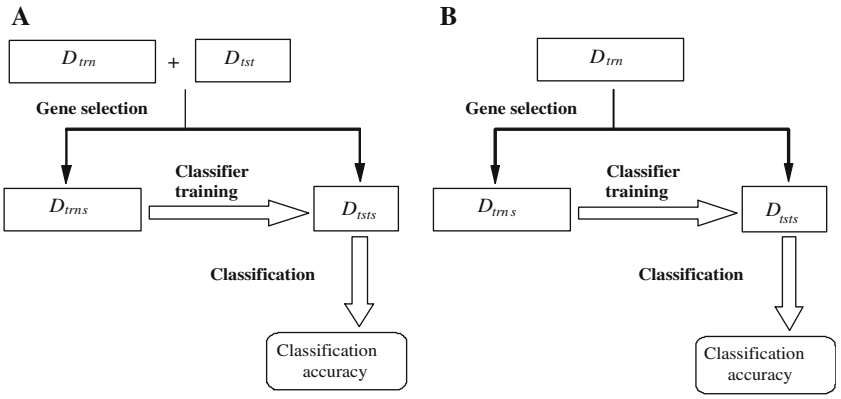

Fig. 3 Comparison between a biased and a totally unbiased verification scheme, where D_{trn} and D_{tst} are training and testing sets, and D_{trns} and D_{tsts} are training and testing set with genes selected, respectively. In case **a** (biased verification scheme), the testing set is used twice in gene selection and classifier procedures, which creates a bias error in the final classification results. In case **b** (the totally unbiased scheme), the testing set is independent of the gene selection and classifier training procedures

where w is a pre-defined weight (relative importance of C_c and P) for adjusting the ratio in experiment, and P is the performance evaluation on dataset D.

In this sense, Eq. (8) can be rewritten as,

$$f_s^* = \arg\min_{f_s \in \mathcal{F}} R(F, S, D) \tag{10}$$

where S is a set of currently selected genes. Function $C_c(\ )$ in Eq. (8) is replaced by a desired R to ensure a good balance between consistency and classification performance.

3.1 Bootstrapping gene-selection algorithm based on classification consistency

The algorithm can be simply summarized into the following steps:

1. Split all genes of dataset D into N segments based on their mean value.
2. Randomly select one gene from each of N segments. The initial candidate gene set contains N genes and is denoted by S.
3. Apply the operation function F (i.e. classifier) to the data containing those genes listed in S, and compute the consistency C_c by Eq. (5) or Eq. (6).
4. Perform gene selection function f_s on S to get a new generation of genes S', and recompute the consistency C'_c.
5. If $C_c^{'} > C_c$, then $C_c = C_c^{'}$ and $S = S'$.
6. Repeat Steps 3–5 for until C_c becomes smaller than a predefined threshold value.
7. Output the finally selected genes.

Algorithm 1 presents the Bootstrapping consistency method for gene selection in pseudo-code, where consistency and performance are optimized simultaneously.

Algorithm 1: Bootstrapping Consistency Gene selection

```
/* Initial gene selection */
Initialize the number of initial genes N;
Build gene spectrum by sorting genes in a increasing
 order of mean value;
Divide the above gene spectrum into N segments;
S ← ∅;
for each segment
    Randomly select one representative gene g;
    S = S ∪ g;
end

 /* Consistency computing */
for  j = 1 to B      /* B is predefined resampling times */
    Partition data D into D_a, D_b;
    Calculate P_a, P_b on S ;
    Calculate consistency score C_cj by Eq. (5) or Eq. (6);
end
Calculate consistency C = Σ_{j=1}^{B} C_cj;
Calculate classification accuracy P on D and S;
Calculate ratio R;

/* Bootstrapping gene selection */
while R > ξ;          /* ξ is predefined stop criterion */
    Update S to S' by mutation or crossover operation;
    Update consistency C'_c on S' ;
    Update R' ;
    If  R' ¿ R
        S ← S'; R ← R';
    end
end
Output S     /* final selected informative genes */
```

The optimized gene selection is obtained by generations of optimization on consistency and classification performance. In each generation, D_a and D_b are resampled B times depending on the size of samples. For example, if the sample size of dataset is larger then 30, B is set to 50,

Table 1 Cancer data sets used for testing the algorithm. Columns for training and validation data show the total number of patients

The numbers in brackets are the ratios of the patients in the two classes

Cancer	Class 1 vs. Class 2	Genes	Train data	Test data	Ref.
Lymphoma(1)	DLBCL vs. FL	7,129	(58/19)77	–	[21]
Leukaemia	ALL vs. AML	7,129	(27/11)38	34	[22]
CNS cancer	Survivor vs. Failure	7,129	(21/39)60	–	[23]
Colon cancer	Normal vs. Tumour	2,000	(22/40)62	–	[15]
Ovarian cancer	Cancer vs. Normal	15,154	(91/162)253	–	[24]
Breast cancer	Relapse vs. Non-Relapse	24,482	(34/44)78	19	[25]
Lung cancer	MPM vs. ADCA	12,533	(16/16)32	149	[26]

otherwise 30. Consequently, C is the mean value of the consistency scores for B rounds of computation.

4 Cancer diagnosis experiments

4.1 Datasets

The proposed concept for gene selection is applied to six well-known benchmark cancer microarray datasets and one proteomics dataset. Table 1 summarizes the seven datasets used for gene selection in the experiment.

4.2 Experimental setup

As suggested in literature for estimating generalization error [27, 28], a fivefold cross-validation schema is used in all the experiments, except for those datasets which had available originally separated training and testing sets. For each cross validation, a totally unbiased verification scheme shown in Fig. 3b is used, where both gene selection and classification are working only on the training set, so that no testing information is included in any part of the cancer diagnosis modelling.

For consistency evaluation, the dataset is randomly partitioned into two subsets. One subset contains one-third of all samples, and the other subset has the rest two-thirds of samples. Using a classifier such as KNN or SVM, two classification accuracies can be computed on two subsets, respectively, the absolute difference between these two accuracies is defined as the consistency (C) in terms of classification performance (refer to Eq. 5 and Eq. 6). After several hundred iterations, the mean value of the computed consistencies is taken as the final result.

In our example we use the above bootstrapping consistency gene selection method and Eq. (6) for consistency evaluation. All genes of a given microarray dataset (the search space) are first segmented into N segments, and N is set as 20. For each fold of the given dataset, the dataset is initially partitioned into training and testing set, on which the bootstrapping runs generation optimization until R becomes less than a predefined threshold ξ, and the selected informative genes are output. There are two setting choices for resampling times B. Depending on the size of dataset, B is set as 50 for those datasets with more than 30 samples, and B is set as 30 for the datasets with smaller sample sizes. w in Eq. (9) is set as 5.0 indicating that consistency is made more important than performance in the optimization. ξ is set as 0.1.

4.3 Results and discussion

Experiments are presented in this section to verify the classification consistency and bootstrapping gene selection method. The experiments use seven benchmark datasets, six cancer microarray datasets and one proteomics dataset, and then compare them with the experimental results of these datasets reported in the original publications, in terms of the cancer diagnosis prediction accuracy (refer to the cited papers in Table 1).

4.3.1 Lymphoma data

Table 2 shows the bootstrapping classification results for Lymphoma data, and Fig. 4 illustrates the optimizing

Table 2 The classification validation results on Lymphoma data

Note that fivefold cross-validation is used for calculating classification accuracy

TP True positive, *TN* true negative, *FP* false positive, and *FN* false negative

Lymphoma	Number of selected genes	TP	TN	FP	FN	Classification accuracy (%)
Fold1	36	8	10	0	1	94.74
Fold2	25	12	6	0	1	94.74
Fold3	34	11	7	1	0	94.74
Fold4	36	10	9	0	0	100
Fold5	32	10	8	0	1	94.74
Overall accuracy: 95.84%						

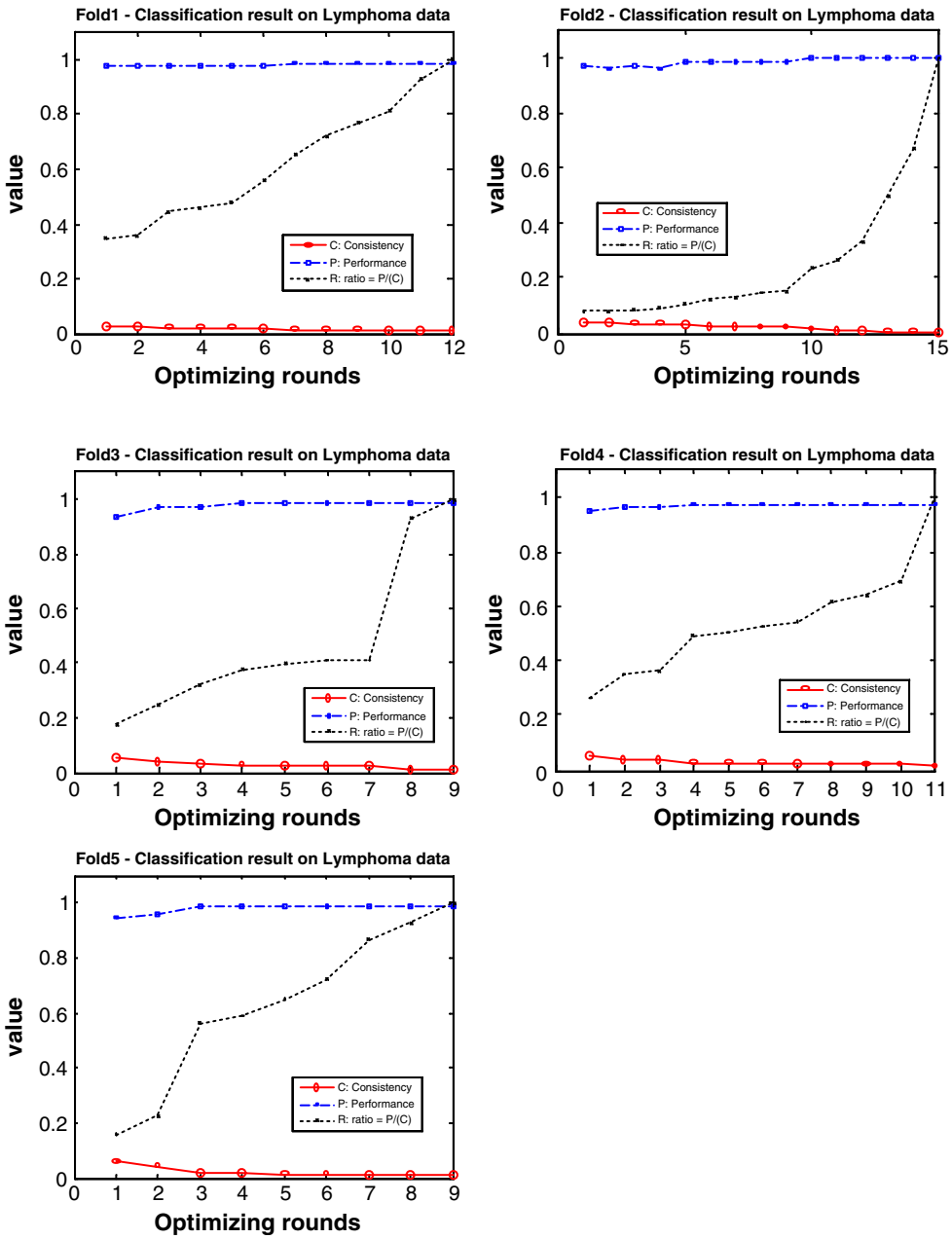

Fig. 4 The optimization results on lymphoma data, where horizontal axis represents the optimizing rounds, and vertical axis shows the results of consistency (*C*), classification performance (*P*) and the ratio (*R*) of consistency and performance calculated in the optimizing process. Note that accuracy P is the training classification accuracy obtained in the classifier optimizing process

procedure in fivefold cross-validation, where consistency and classification accuracies are recorded at every optimizing step.

As shown in Table 2, the overall classification accuracy on the testing set of Lymphoma dataset is fairly high (greater than 95%). The number of selected informative genes is around 30, and the final calculated classification accuracy is stable (94.74– 100%). Moreover, the results of confusion matrix (TP, TN, FP and FN) show that the proposed method is very effective on Lymphoma dataset in terms of both classification accuracy (TP and TN) and misclassification rate (FP and FN).

Figure 4 presents the optimizing procedure of bootstrapping consistency gene selection method. The optimized consistency is seen to decrease to below 0.1, while the training classification accuracy increases to above 90%. It shows that the proposed method is capable of improving

Table 3 The classification validation result on Leukaemia data. An independent test dataset was used for validation

dataset	Number of selected genes	TP	TN	FP	FN	Classification accuracy
Leukaemia	35	12	20	0	2	94.12%

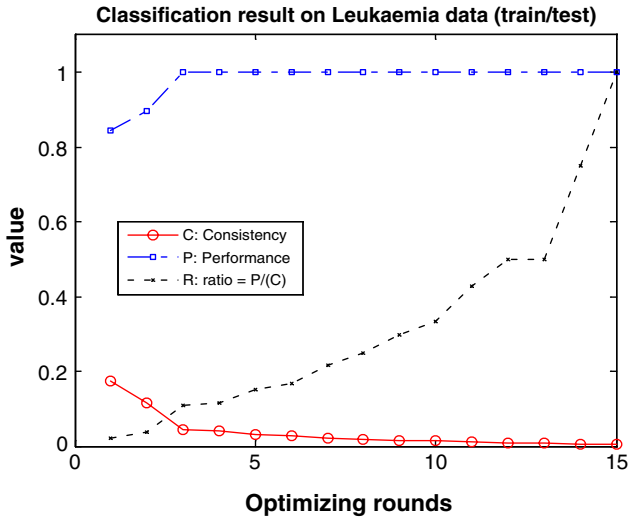

Fig. 5 The results of optimization on leukaemia data

Table 4 The classification validation results of bootstrapping consistency method on CNS Cancer data

CNS Cancer	Number of selected genes	TP	TN	FP	FN	Classification accuracy (%)
Fold1	44	9	1	2	0	83.33
Fold2	56	4	3	0	5	58.33
Fold3	43	3	2	2	5	41.67
Fold4	44	7	2	3	0	75.00
Fold5	44	6	2	4	0	66.67
Overall accuracy: 65.00						

consistency simultaneously with classification performance. Note that a smaller consistency value indicates a better consistency characteristic of data.

4.3.2 *Leukaemia data*

Table 3 and Fig. 5 present the classification and consistency results obtained by the described bootstrapping consistency method on Leukaemia data. Table 3 shows that the achieved classification accuracy on the testing set is about 95%, when 35 genes are used for constructing the final optimized classifier. In Fig. 5, after 15 rounds optimization based on the improvement of ratio R to consistency and classification performance (refer to Eq. (9)), the classification accuracy on the training set improves to 1 and the consistency value is reduced to 0, indicating that the maximum possible consistency is obtained.

4.3.3 *CNS cancer data*

Table 4 and Fig. 6 present the experimental results obtained by the described bootstrapping consistency method on CNS cancer data. Table 4 shows that the classification results on 5 folds of CNS Cancer dataset have high variance: the highest accuracy is 83.33%, while the lowest is only 41.67%. The overall accuracy is only 65%, which is not acceptable for solving the real clinical problem of disease diagnosis. The confusion matrix clearly shows that one misclassification rate (FN) is high, i.e. number of false negatives (FN) obtained on fold2 and fold3 is 5 individuals; this is larger than the TN accuracy rate.

Figure 6 shows that the initial consistency value of CNS cancer dataset is quite high (around 0.4) and cannot be decreased in the optimizing process as much as in the previous datasets. The classification performance on training sets on folds 1–4 data rises approximately from 60 to 80%, while the consistency is decreased from 0.4 to 0.2. Although the accuracy on the fold 5 data is significantly improved, from 40 to 80%, the best consistency is still greater than 0.2, which means the consistency is not satisfactory, probably due to inherent problems with the dataset. Such a situation results in the bad overall classification accuracy (65.00%).

4.3.4 *Colon data*

Table 5 and Fig. 7 show the experimental results obtained by the bootstrapping consistency method on Colon cancer data. As presented in Table 5, the highest classification accuracy (91.67%) is obtained on fold 1 and fold 4 data in the classifier optimizing process, while the lowest one (66.67%) appears on fold 3. The difference between these computed classification accuracies is large, which shows the Colon Cancer dataset has a relatively high variability of consistency. The final number of selected informative genes is on average 23, and the overall classification accuracy is about 84%.

Figure 7 shows that both the consistency and performance improve significantly. For example, in fold 1, the classification performance rises approximately 10% (from 80 to 90%) coupled with the improvement of consistency (from 0.2 to 0.1). The improvement of classification performance obtained on fivefolds data is quite different, though: the performance on folds 3–5 is improved much more than that on folds 1–2. Meanwhile, the optimizing rounds are also different. The classifier is optimized within 25 rounds in the cases of folds 3–5, but in folds 1–2, the classifier is optimized in less than 20 rounds.

4.3.5 *Ovarian data*

Table 6 and Fig. 8 give the experimental results obtained by iterative bootstrapping method on Ovarian Cancer dataset. Table 6 shows the classification results based on

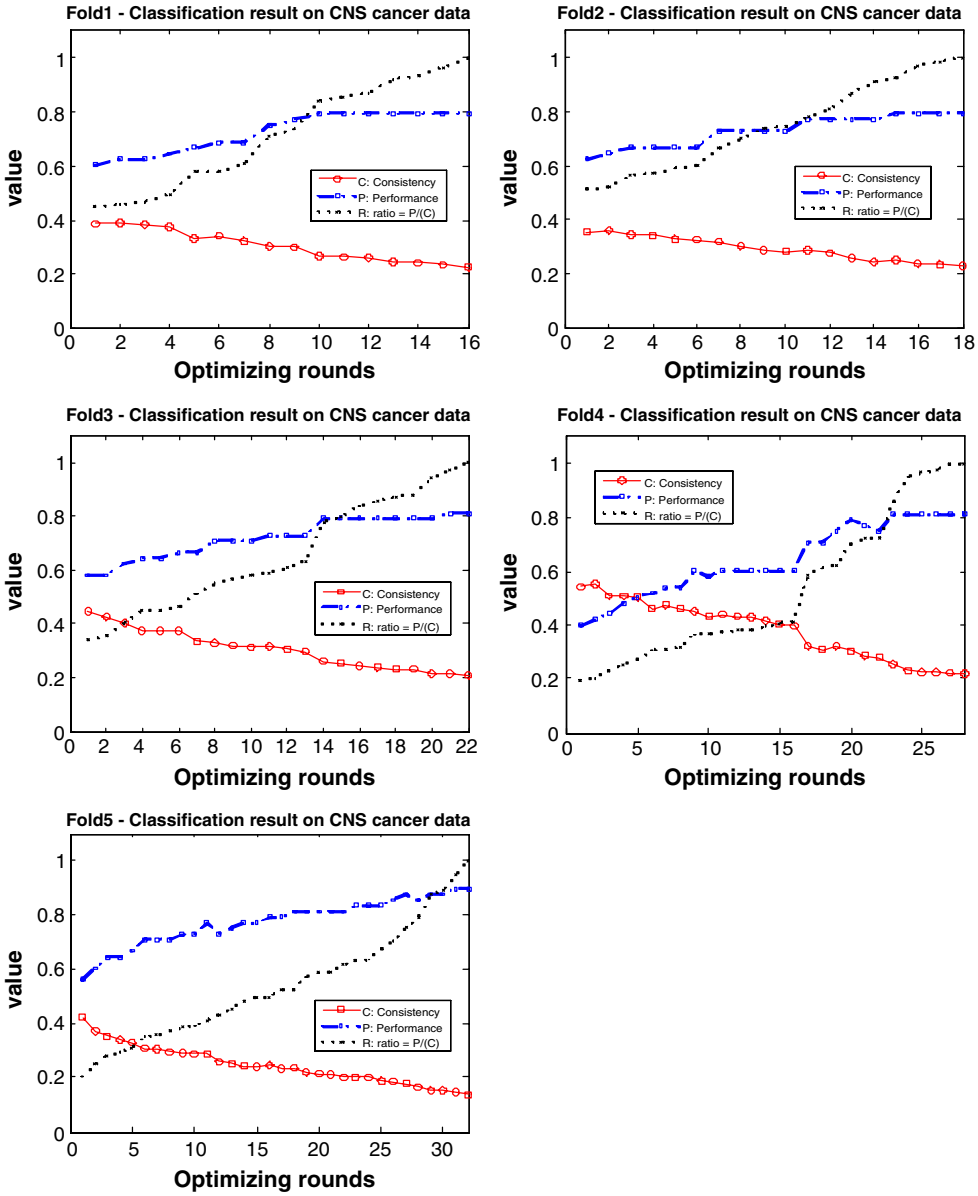

Fig. 6 The results of iterative bootstrapping optimization on CNS cancer data

Table 5 The classification validation results of iterative bootstrapping method on Colon Cancer data

Colon Cancer	Number of selected genes	TP	TN	FP	FN	Classification accuracy (%)
Fold1	22	4	7	0	1	91.67
Fold2	17	4	6	2	0	83.33
Fold3	21	2	6	1	3	66.67
Fold4	29	5	6	1	0	91.67
Fold5	28	1	11	0	2	85.71
Overall accuracy: 83.81%						

the informative genes selected. The proposed method produces an overall accuracy of 98.80%. The difference between the highest accuracy (100%) and the lowest accuracy (98%) is only 2%. Moreover, the confusion matrix shows both the classification accuracy rate and misclassification rate are very good, e.g. there are no misclassified samples in the cases of fold 4 and fold 5.

Figure 8 shows that both the classification performance and consistency is stable during the process of classifier optimization. It turns out that the ovarian dataset has a good and little varying consistency characteristic, which results in successful classification results on all cross-validation sets. Consequently, the improvement of consistency is less than 0.05 in all 5 folds.

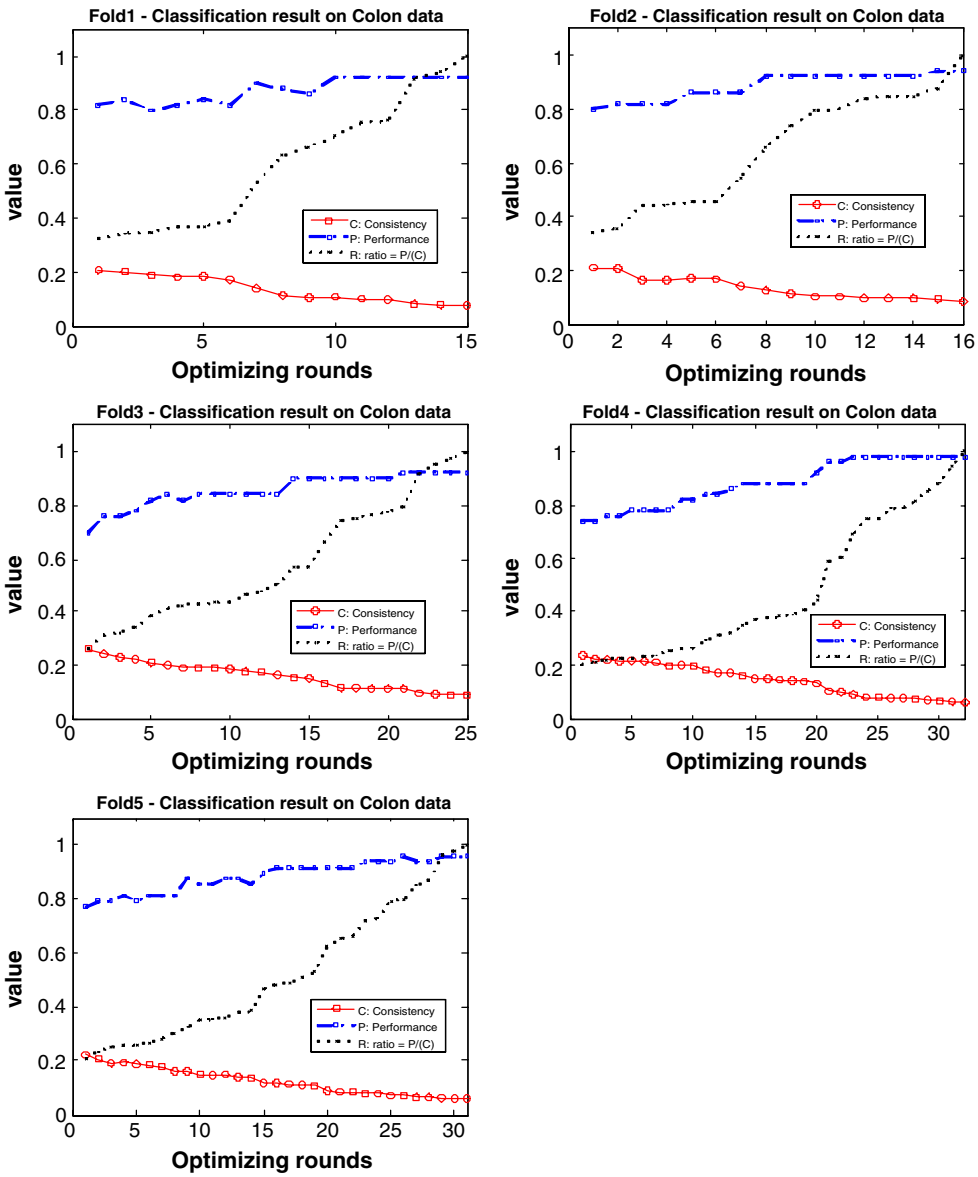

Fig. 7 The results of iterative bootstrapping optimization on colon cancer data

4.3.6 Breast cancer data

Table 7 and Fig. 9 show the experimental results obtained by the bootstrapping consistency method on Breast Cancer dataset. Table 7 shows that the low classification accuracy on the testing set is related to the high inconsistency characteristic of this breast cancer dataset. The classification accuracy obtained by iterative bootstrapping method with 50 selected informative genes is only 63.16%, which is not very useful for identifying disease patterns in real clinical practice.

Figure 9 presents the relatively poor consistency and classification accuracy obtained by iterative bootstrapping method in the optimizing process. The best classification performance on the training data in gene selection procedure is 80%, when the final optimized consistency (approximately 0.2) is achieved after nine iterations.

4.3.7 Lung cancer data

Table 8 and Fig. 10 show the results obtained by the bootstrapping consistency method on Lung Cancer data. As shown in Table 8, the experimental result of Lung Cancer data reaches a satisfactory level in which the classification accuracy on testing set is 91.28% with 34 selected genes identified by our method.

As shown in Fig. 10, the classifier is only optimized in only 9 rounds. Unlike in the relative poor consistency and classification performance in the Breast Cancer dataset, there was no difficulty in the optimizing process here, due

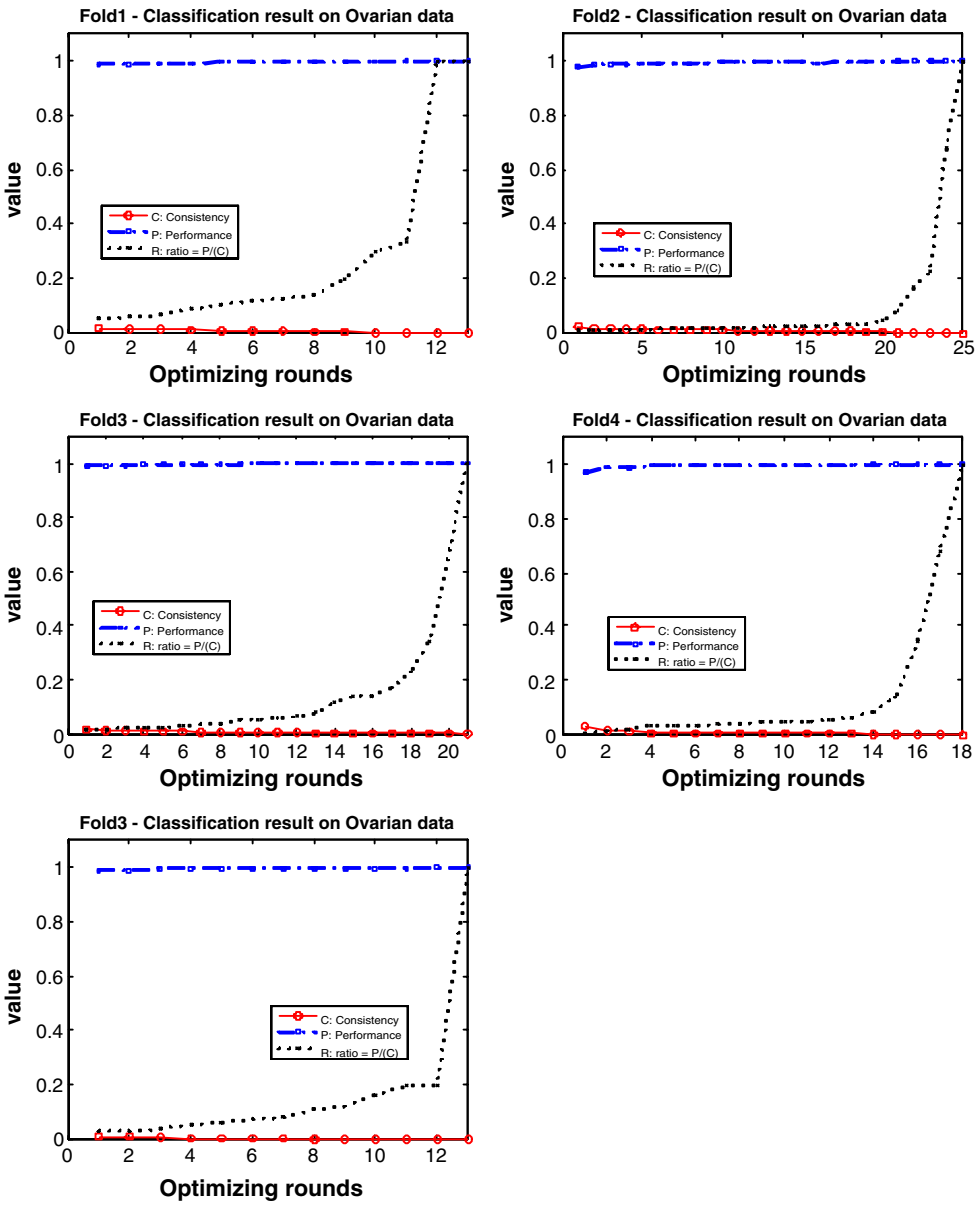

Fig. 8 The results of iterative bootstrapping optimization on ovarian data

to the already good initial consistency characteristic of Lung Cancer dataset. It can be seen that the initial classification accuracy is greater than 90%, and the consistency calculated in the first round is about 0.1, so that it only takes 9 optimizing rounds to achieve a high classification accuracy coupled with a good consistency in the training process.

4.4 Classification accuracy summary: consistency method versus publication

For clarity, the classification accuracies obtained by the presented bootstrapping consistency gene selection method are summarized and compared with the literature reported results in Table 9. Consistency method outperforms the published methods on four datasets, and the classification result on colon data is very close to the reported accuracy. However, the classification accuracies of two datasets (CNS, Breast) are much lower than the published ones. Many published classification results are based on a biased validation scheme as in Fig. 2, which results in the experiments being unreplicable and too optimistic.

However, the experimental results obtained by the consistency method can be easily reproduced, because of the totally unbiased validation scheme of Fig. 3b, as applied in this study. These results suggest that a reproducible prognosis is possible for only five of the seven used benchmark datasets.

Table 6 The classification validation results of iterative bootstrapping method on Ovarian Cancer data

Ovarian cancer	Number of selected genes	TP	TN	FP	FN	Classification accuracy (%)
Fold1	18	25	24	0	1	98.00
Fold2	28	31	18	1	0	98.00
Fold3	24	33	16	1	0	98.00
Fold4	24	34	16	0	0	100
Fold5	34	38	15	0	0	100
Overall accuracy 98.80						

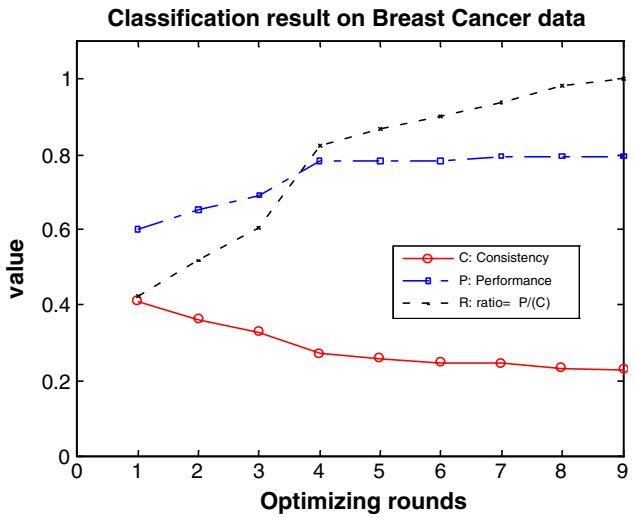

Fig. 9 The optimizing results of iterative bootstrapping method on breast cancer data

Table 7 The classification validation results of iterative bootstrapping method on Breast Cancer data. An independent test dataset was used for validation

Dataset	Number of selected genes	TP	TN	FP	FN	Classification accuracy
Breast cancer	50	5	7	5	2	63.16%

5 Summary

The results with the bootstrapping consistency gene selection method described in this paper have demonstrated that the consistency concept can be used for gene selection to solve the reproducibility problem in microarray data analysis. The main contribution of the consistency method is that it ensures the reliability and generalizability of microarray data analysis experiment, and improves the disease classification performance as well. In addition, because the method does not need previous knowledge about the given microarray data, it can be used as an effective tool in unknown disease diagnosis.

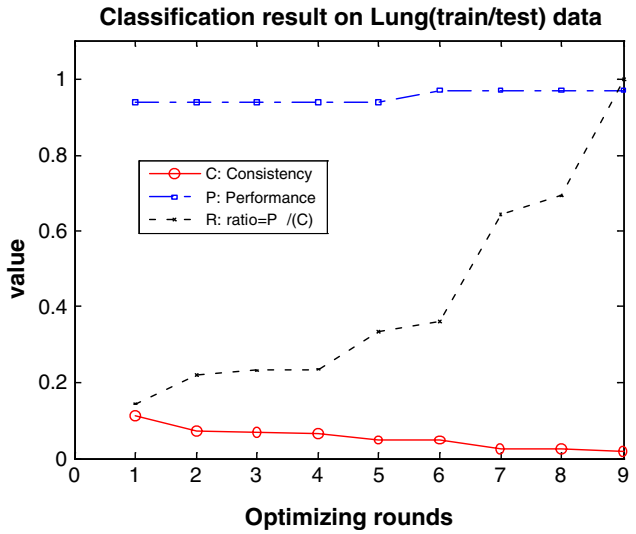

Fig. 10 The results of iterative bootstrapping optimization on Lung cancer data

Table 8 The classification results of iterative bootstrapping method on Lung Cancer data. An independent test dataset was used for validation

Dataset	Number of selected genes	TP	TN	FP	FN	Classification accuracy
Lung cancer	34	121	15	0	13	91.28%

From the perspective of generalization error, it should be pointed out that the experimental results can be seen as totally unbiased, because the data for validation is independent and never used in the training process, i.e. before the final informative genes are selected, the test data is isolated and has no correlation with these genes. Therefore, the selected informative genes are entirely fair to any given data for validation. Such a mechanism of gene selection might result in the bad performance in certain microarray datasets, which is due to the special characteristics of these datasets. This makes the reported good results in some published papers on these datasets suspect, as also discussed by [29]. Recently, many papers have reported on development of guidelines and procedures for more reliable microarray profiling [30–32], reviewed existing methods [33, 34] and suggested improvements in meta-analysis [35]. However, none of these works have tackled explicitly the problem of consistency in the gene selection step, as investigated by us.

The consistency concept was investigated on six benchmark microarray datasets and one proteomic dataset. The experimental results show that the different microarray datasets have different consistency characteristics, and that

Table 9 Classification accuracy comparison: Consistency method results vs. known results from literature

Dataset	Average classification accuracy	
	Consistency method (%)	Publication (%)
Lymphoma	95.84	72.50
Leukaemia	94.12	85.00
CNS cancer	65.00	83.00
Colon cancer	83.81	87.00
Ovarian cancer	98.80	97.00
Breast cancer	63.16	94.00
Lung cancer	91.28	90.00

better consistency can lead to an unbiased and reproducible outcome with good disease prediction accuracy.

The recommended protocol for using our method is as follows:

1. Use Eq. (6) with your training/test sets.
2. Run your classification algorithm of choice.
3. Use Eq. (7) with your training/test sets.
4. Run your classification algorithm of choice again with same settings.
5. Choose the results with the test/training set which gives better consistency in step 2 or 4.
6. Run the better model with total data or with new future datasets.

We believe our implementation of the classification consistency using iterative bootstrapping can provide a small set of informative genes which perform consistently with different data subsets. Compared with the traditional gene selection methods without using consistency measurement, bootstrapping consistency method can thus provide more accurate classification results. More importantly, results demonstrate that gene selection with the consistency measurement is able to enhance the reproducibility and consistency in microarray and proteomics based diagnosis decision systems. This is important when the classification models are used to analyze new future datasets.

Acknowledgments The research presented in the paper was partially funded by the New Zealand Foundation for Research, Science and Technology under the grant: NERF/AUTX02-01.

References

1. Ding C, Peng H (2003) Minimum Redundancy Feature Selection for Gene Expression Data. In: Paper presented at the Proc. IEEE Computer Society Bioinformatics Conference (CSB 2003), Stanford
2. Furey T, Cristianini N et al (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16(10):906–914
3. Jaeger J, Sengupta R et al (2003) Improved gene selection for classification of microarrays. In: Paper presented at the Pacific Symposium on Biocomputing
4. Tusher V, Tibshirani R et al (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 98(9):5116–5121
5. Zhang C, Lu X, Zhang X (2006) Significance of gene ranking for classification of microarray samples. IEEE/ACM Trans Comput Biol Bioinform 3(3):312–320
6. Duch W, Biesiada J (2006) Margin based feature selection filters for microarray gene expression data. Int J Inform Technol Intell Comput 1:9–33
7. Draghici S, Kulaeva O et al (2003) Noise sampling method: an ANOVA approach allowing robust selection of differentially regulated genes measured by DNA microarrays. Bioinformatics 19(11):1348–1359
8. Efron B, Tibshirani R et al (2001) Empirical bayes analysis of a microarray experiment. J Am Stat Assoc 96:1151–1160
9. Lee KE, Sha N et al (2003) Gene selection: a Bayesian variable selection approach. Bioinformatics 19(1):90–97
10. Tibshirani RJ (2006) A simple method for assessing sample sizes in microarray experiments. BMC Bioinform 7:106
11. Kauai H, Kasabov N, Middlemiss M et al (2003) A generic connectionist-based method for on-line feature selection and modelling with a case study of gene expression data analysis. In: Paper presented at the Conferences in Research and Practice in Information Technology Series: proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics 2003, vol 19, Adelaide, Australia
12. Wang Z, Palade V, Xu Y (2006) Neuro-Fuzzy ensemble approach for microarray cancer gene expression data analysis. In: Proceedings of 2006 international symposium on evolving fuzzy systems, pp 241–246
13. Wolf L, Shashua A et al (2004) Selecting relevant genes with a spectral approach (No. CBCL Paper No.238). Massachusetts Institute of Technology, Cambridge
14. Huerta EB, Duval B et al (2006) A hybrid GA/SVM approach for gene selection and classification of microarray data. Lect Notes Comput Sci 3907:34–44
15. Alon U, Barkai N et al (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 96(12):6745–6750
16. Li L, Weinberg CR et al (2001) Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics 17(12):1131–1142
17. Wahde M, Szallasi Z (2006) Improving the prediction of the clinical outcome of breast cancer using evolutionary algorithms. Soft Comput 10(4):338–345
18. Wahde M, Szallasi Z (2006) A Survey of methods for classification of gene expression data using evolutionary algorithms. Expert Rev Mol Diagn 6(1):101–110
19. Mukherjee S, Roberts SJ (2004) Probabilistic consistency analysis for gene selection. Paper presented at the CSB, Stanford
20. Mukherjee S, Roberts SJ et al (2005) Data-adaptive test statistics for microarray data. Bioinformatics 21(Suppl 2):ii108–ii114
21. Shipp MA, Ross KN et al (2002) Supplementary information for diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med 8(1):68–74
22. Golub TR (2004) Toward a functional taxonomy of cancer. Cancer Cell 6(2):107–108
23. Pomeroy S, Tamayo P et al (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 415(6870):436–442

24. Petricoin EF, Ardekani AM et al (2002) Use of proteomic patterns in serum to identify ovarian cancer. Lancet 359:572–577
25. Van 't Veer LJ, et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415:530–536
26. Gordon GJ, Jensen R et al (2002) Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Research 62:4963–4967
27. Breiman L, Spector P (1992) Submodel selection and evaluation in regression: the Xrandom case60. Int Stat Rev 60:291–319
28. Kohavi R (1995) A study of crossvalidation and bootstrap for accuracy estimation and model selection. In: Paper presented at the international joint conference on artificial intelligence (IJCAI), Montreal
29. Ransohoff DF (2005) Bias as a threat to the validity of cancer molecular marker research. Nat Rev Cancer 5(2):142149
30. Staal FJT, Cario G et al (2006) Consensus guidelines for microarray gene expression analyses in leukemia from three European leukemia networks. Leukemia 20:1385–1392
31. Allison DB, Cui X et al (2006) Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet 7:55–65
32. Kawasaki ES (2006) The end of the microarray tower of babel: will universal standards lead the way? J Biomol Tech 17:200–206
33. Pham TD, Wells C et al (2006) Analysis of microarray gene expression data. Curr Bioinform 1:37–53
34. Asyali MH, Colak D et al (2006) Gene expression profile classification: a review. Curr Bioinform 1:55–73
35. Sauerbrei W, Hollander N et al (2006) Evidence-based assessment and application of prognostic markers: the long way from single studies to meta-analysis. Commun Stat Theory Methods 35:1333–1342