Front. Electr. Electron. Eng. China 2009, 4(3): 243–250
DOI 10.1007/s11460-009-0041-y

REVIEW ARTICLE

Xuesong LU, Xuegong ZHANG

Pattern recognition methods in microarray based oncology study

Abstract With the development of microarray technology, more and more microarray-based oncology studies have been carried out. Huge amounts of data and the complexity of cancer mechanisms make data analysis methods a much more important part of these studies. In this article, we will mainly focus on the pattern recognition methods used in oncology studies. According to the availability of sample information, the unsupervised methods and supervised methods are reviewed separately. Finally, some possible future directions are proposed.

Keywords pattern recognition methods, microarray, oncology

1 Introduction

Microarray technology was introduced in the 1990s, and enables researchers to measure the expression levels of thousands of genes simultaneously. In 1999, Golub et al. [1] successfully applied microarray technology in a leukemia study. In the same year, Alon et al. [2] used microarray technology in a study of colon. Since then microarray technology has become a standard tool for cancer study and almost all of common cancers have been profiled [1–7]. A typical microarray experiment contains several dozen to hundreds of samples and each sample has the expression values of thousands of genes, so researchers face the problem that the feature number is far larger than the sample size. The characteristic of the problem increases the difficulties in microarray data analysis. In the past decade, pattern recognition methods have been applied or developed, to address these problems. An overview of these methods will not only help us understand the state of arts in this area, but also may inspire us with better ideas to solve these problems.

In this review, we will focus on the microarray data analysis methods used in oncology studies. To interpret the microarray data analysis problem more easily, we formulate the problem as follows.

A microarray dataset can be described as a data matrix $\boldsymbol{X}_{N\times M}$, where M is the number of samples and N is the number of features/genes. Each column represents a sample S_i, $i\in\{1,2,\ldots,M\}$ and each row represents a gene G_i, $i\in\{1,2,\ldots,N\}$. For most problems there will be a vector $\boldsymbol{Y}$, which includes the clinical information of each sample. The clinical information can either be discrete (e.g., normal vs cancer or different subtypes of cancer) or be continuous (e.g., survival time).

To avoid the effect of noise and shorten the computation time, researchers usually perform preliminary gene filtering before the analysis to remove the non-informative genes. There are four commonly used criteria for defining non-informative genes. First, the fold change (defined as the ratio between the expression value of treated and that of control samples) is less than a pre-determined cutoff value. In this case, the treated samples can be a group treated with a particular drug or a group of cancer patients and the control samples are a different group of untreated or non-cancerous patients. Second, the percentage of missing data is higher than a pre-determined cutoff value. Third, the percentage of present call is less than a pre-determined cutoff value. Last, the standard deviation is less than a pre-determined cutoff value. The choice of the cutoff value is usually arbitrary and different studies may use a different combination of criteria. For example, van't Veer et al. filtered the genes by requiring that the informative gene must have a two-fold difference and a p-value larger than 0.01 in more than 5 tumors [3]. Kapp et al. removed the genes which had a standard deviation of less than 1.5 and had less than 90% of data present [8].

Received June 23, 2008; accepted July 2, 2008

Xuesong LU
Merck Sharp & Dohme (China) Ltd., Shanghai Office, Shanghai 200040, China

Xuegong ZHANG (✉)
Bioinformatics Division, TNLIST/Department of Automation, Tsinghua University, Beijing 100084, China
E-mail: zhangxg@tsinghua.edu.cn

Based on whether clinical information is incorporated in the data analysis scheme, the data analysis methods can be separated into two categories — unsupervised analyses which only focus on expression data, and supervised analyses which utilize information other than expression data. We will discuss these two categories separately.

2 Unsupervised analysis methods

The main purpose of this type of method is to aid the class discovery, and here we not only mean the class for samples, but also class for genes. One of the most commonly used unsupervised methods in microarray data analysis is hierarchical clustering [9–11]. Because this type of analysis clusters genes and samples step by step and generates a clustering tree to indicate the distance between genes or samples, it is often used as the first step when analyzing data and it gives researchers an overall view of the data. For clustering methods, there are three key issues, distance measurement, optimization function, and evaluation of clusters. Different choices may need different solutions and may lead to different discoveries.

2.1 Distance measurement

Different genes have very different scales of expression values, so it does not have too much biological meaning to measure the distance in Euclidean space. In biological studies, it is more meaningful if two genes or samples have the same expression pattern (i.e., two genes that always have high or low expression values in the same sample may imply that these two genes have a similar function). Because of this, the Pearson correlation coefficient is always used as a distance measurement [9,12]. Though direct measurement of the distance in Euclidean space may not be feasible, standardizing the data into the same scale or using ratio values makes the measurements more reasonable. Instead of measuring the Euclidean distance with microarray data, Nilsson et al. used an Isomap algorithm to measure the geodesic distance between two points [13], which is to measure the length of the shortest path between these two points. Recently, with more and more gene annotation information available, Boratyn et al. proposed to measure the distance based on both expression data and biological data, which include the biological functions of each gene [14]. The distance by the new measurement is the sum of distance measured by expression data and distance measured by biological data. Here the distance measured by biological data is defined based on three criteria, the distance between two genes with similar function is smaller than two genes with different functions, the distance between annotated and un-annotated genes is smaller than that between two un-annotated genes, and the distance between two un-annotated genes is smaller than that between two genes with different functions. Though we do not know the correct answer in most oncology studies, these improvements in the distance measurement provide interesting applications.

2.2 Optimization function

Almost all clustering problems are due to optimization problems. The aim of optimization is to maximize the distance between different clusters and to minimize the distance within the same cluster at the same time. To solve this optimization problem, different optimization functions are proposed and solutions are provided. Bagirov et al. designed the optimization function to find the cluster centroids by minimizing the sum of the deviations of all tissue samples [15]:

$$f(C_1,C_2,...,C_q) = \sum_{i=1}^{n} \min_{s=1,2,...,q} \|S_i - C_s\|,$$

where q is the total number of clusters and $C_1,C_2,...,C_q$ are the centroids of clusters. This optimization problem was solved by iteratively adding one centroid at a time. Gao et al. proposed to use the sparse non-negative matrix factorization method to do the clustering [16]. In this method, the data matrix $\boldsymbol{X}$ is factored into the product of two non-negative matrices $\boldsymbol{W}_{N\times k}$ and $\boldsymbol{H}_{k\times M}$, where k is the number of clusters. Then given the sparseness control parameter λ, the optimization function becomes

$$\min\left\{\frac{1}{2}\|\boldsymbol{X}-\boldsymbol{W}\boldsymbol{H}\|^2 + \frac{1}{2}\lambda\|\boldsymbol{H}\|^2\right\},$$

which can be solved iteratively. After calculating the $\boldsymbol{H}_{k\times M}$, sample j is placed in cluster i if h_{ij} is the largest entry in column j.

Sometimes there are other background information available. Though the information is not enough for supervised analysis, it will help the clustering process if we have a good way to utilize it. Sese et al. added additional features to each sample. The optimization function becomes maximizing for the interclass variance with the restriction that only allow those splits that can be explained by a common feature [17]. Dotan-Cohen et al. proposed a hierarchical tree snipping method which incorporates the gene annotation information from Gene Ontology (GO) to obtain clusters that are substantially enriched in genes participating in the same biological process [18].

Since a gene can function in several pathways and a sample can have multiple clinical statuses, limiting each sample or gene to be present in only one cluster may lead to loss of some information. Fuzzy clustering methods have been applied in this area (e.g., Belacel et al. successfully applied fuzzy K-means and fuzzy J-means clustering to breast cancer data [19]).

The optimization functions discussed above only focus on a gene or sample at a time, but sometimes we need to consider both the gene and sample simultaneously. The coupled two-way clustering (CTWC) method introduced by Getz et al. tries to address this issue [20,21]. The method does the clustering in an iterative way, which is first to cluster genes and samples separately, then iteratively perform clustering to the sub-matrix defined by the pair of clusters generated in gene and sample dimension. Kluger et al. applied spectral bi-clustering methods to tackle this problem [22]. They model the expression of each gene as $x_{ij} = e_{ij} + \rho_i + \chi_j$, where e_{ij} is the base expression level for gene i in sample j, ρ_i is the tendency of gene i to be expressed in all samples, χ_j is the tendency of all genes to be expressed in sample j. So the objective in the simultaneous clustering of genes and samples is, given data matrix, to find the underlying block structure of E.

2.3 Evaluation of clusters

Since there is no golden standard for the unsupervised problem, it is hard to evaluate the clustering results especially when the number of clusters is unknown. Usually, researchers choose the number of clusters based on their knowledge of the data. Or they try several possible numbers to select those with obvious patterns. Such choices are always used for hierarchical clustering, K-means clustering, self-organizing map (SOM), etc. [9–12, 23–25]. Currently, there are many methods that focus on estimating the number of clusters and evaluating the clusters. Some of them work in an iterative way. For example, Hsu et al. initially separated the samples into k classes, identified the differentially expressed genes (e.g., the expression values of the genes have different distribution in different classes) and built the classifier to predict the label of each sample. Finally, they iteratively refined the clustering results [26]. This method can both suggest the number of clusters and generate stable clusters. Li et al. also provided a program called SamCluster to iteratively discover stable clusters [27].

Some of the evaluation methods are based on re-sampling [28–32]. The basic idea of these methods is to re-sample several datasets from original data and get the clustering results, then evaluate the stability of the cluster result based on some statistics. Instead of using a re-sampling technique, Swift et al. used a set of clustering methods to generate clusters and evaluate their stability [33].

Besides clustering methods, there are also some unsupervised methods used in oncology studies for reducing the dimension, such as principal components analysis (PCA), singular value decompositions (SVD), and independent components analysis (ICA) [12,34–37].

3 Supervised analysis methods

Like microarray studies, the feature/gene number is always far larger than the sample size. Reducing the feature dimension is usually the first step in a supervised analysis. In this section, we will review the dimension reduction methods and subsequent analysis methods separately.

3.1 Dimension reduction

Reducing the feature dimension not only avoids over-fitting in model building, but also saves computation time and provides better interpretation of the data. Among these methods, feature selection is a commonly used method. The purpose of feature selection is to remove the less informative features and keep the more informative ones. The most common way to measure whether a gene is informative in microarray studies is to test whether the gene is differentially expressed in different classes. Usually, feature selection methods can be separated into two categories — filter methods and wrapper methods.

Filter methods evaluate each feature only based on the data. In most cases, the features are ranked by some statistics and given a cutoff value, the top ranked features are selected. If we assume that gene expression values in different classes belong to two Gaussian distributions with different mean values. The t statistic is often used to evaluate the performance of each gene for discriminating two classes [38–42]. Some other statistical methods are used for selecting differentially expressed genes, such as Wilcoxon rank test [41–43] or F-test [44]. Besides these statistical methods, signal-to-noise calculation which is defined as

$$\mathrm{S2N}_i = \frac{\mu_{1i} - \mu_{2i}}{\sigma_{1i} + \sigma_{2i}},$$

where μ_{ki} is mean value of gene i in class k, σ_{ki} is standard deviation of gene i in class k [1,45,46], ratio between class sum of squares to within class sum of squares [46], correlation between gene expression G_i and class label Y [42,47], entropy-based method [48] are applied in feature selection process. Because most of the filter methods only consider one gene at a time, they will fail to identify the combination effect of several genes. Bø et al. proposed a method to evaluate the discriminating ability of a pair of genes [49].

Wrapper methods embed the feature selection into the model optimization process. Usually, the performance of the features selected is evaluated by the classification error rate. Since feature selection is an NP-hard problem and the feature number in microarray studies is very large, many optimization methods have been applied to this problem. Support vector machine (SVM), recursive feature elimination (RFE) [50], recursive-SVM [51], and entropy-based recursive feature elimination (E-RFE) [52] all use similar

search methods to get optimal results. For these methods, in each iteration the classifier is trained with the features left. Features are ranked based on different criteria, and then one or more features with the worst performance are eliminated until a stop criterion is reached. Heuristic algorithms sequential forward selection (SFS), sequential forward floating search (SFFS), genetic algorithm (GA) and evolutionary algorithm (EA) are all incorporated with different classifiers to select features. For example, SFS with Fisher's linear discriminant analysis (LDA) [53], SFS with IB1, Naïve Bayes, C4.5 and CN2 [54], SFS, SFFS with LDA, logistic regression and SVM [55], GA with SVM [56,57], GA with *k*-nearest neighbor (*k*-NN) [58], GA with maximum likelihood classification method [59], EA with *k*-NN [60,61], EA with LDA [61].

Other than the filter and wrapper methods, there are some feature selection methods that can rank the features in the model training process and select important features. Saeys et al. called them embedded methods [62]. Krishnapuram et al. used a sparse Bayesian approach to model the problem, and then a set of important features were selected based on the weight of each feature [63]. Also, BLogReg [64] adopted a similar approach to this problem.

Similar to the PCA approach, partial least squares (PLS) has also been used for dimension reduction when sample labels are available [65,66].

3.2 Class prediction

Usually microarray data class prediction problems are similar to class prediction problems in other areas, so most of the classic class prediction methods have been applied to microarray data analysis — LDA projects the samples into one dimension space to maximize the distance between classes and minimize the distance in each class at the same time [46]; *k*-NN predicts the test sample based on the class labels of the *k* samples near the test sample [37,45,46,58, 60]; decision tree is a classifier in the form of a tree structure and each node either indicates the label of the test sample or specifies some test to select which sub-tree to go [67]; SVM aims to find an optimal hyper-plane to separate the two classes and maximize the distance between the hyper-plane and the closest data pointing to the hyper-plane [24,47,50–52,56,57]; and artificial neural networks take genes as input nodes and class label as the output node to learn the parameters connecting to nodes of different layers and predict the unknown samples [68–70]. Some modifications to the classic methods are also applied to this problem, such as Linder et al. who proposed the subsequent artificial neural network (ANN) method, which has two levels of ANN. The first level functions as a pre-selection which will narrow the possible predicted label of test sample and let the ANN of second level to give a final prediction [71].

Besides these classic methods, there are other methods used in microarray class prediction, such as the weighted voting scheme introduced by Golub et al. [1]. The basic idea of this method is that each selected informative gene predicts the class label based on which class the test sample is near and the sample is predicted to the class with more votes. Van't Veer et al. successfully used a correlation-based classification method on a breast cancer dataset [3]. West et al. applied a Bayesian regression model to predict the clinical status of breast cancer patients [35]. Zhang et al. introduced a vector $\alpha=\{\alpha_1,\alpha_2,\ldots,\alpha_n\}$ to indicate whether the class label is switched, e.g., $\alpha_i=1$ indicates the label of sample i is switched. With this new variable and regression method, Zhang et al. tried to find the mislabeled samples in an iterative way [72]. Gevaert et al. proposed three ways to integrate clinical data into the Bayesian network learning — full integration, which means to put clinical data and expression data together to train the Bayesian network; decision integration, which means to train two Bayesian networks separately and integrate the prediction together; and partial integration, which means to learn the structure of the Bayesian network separately then integrate the structures together for parameter learning and prediction. The results show that only partial integration and decision integration perform significantly better than each data source separately [73].

Survival analysis is a specific class of class prediction problem. Unlike the problems we have discussed above, the class label here is a continuous value. Several regression-based methods have been applied to this kind of data analysis [74–78]. All these methods are derived from the same basic model

$$\lambda(t) = \lambda_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)$$

$$= \lambda_0(t)\exp(\boldsymbol{\beta}\boldsymbol{X}^{\mathrm{T}}),$$

where $\lambda_0(t)$ is a baseline function, $\boldsymbol{\beta}=\{\beta_1,\beta_2,\ldots,\beta_n\}$ is the regression coefficients and $\boldsymbol{X}=\{x_1,x_2,\ldots,x_n\}$ is the expression level of the sample studied. According to different focuses, these methods also have different modifications to that model. Goeman et al. put an extra assumption on the regression coefficients that the regression coefficients of the genes were random and independent with mean zero and common variance τ^2 [74]. Kaderali et al. proposed to solve the regression problem with a hierarchical Bayesian approach which introduced a prior distribution for regression coefficients [77]. One advantage of regression analysis is that it can handle different sources of data simultaneously (i.e., Fernandez-Teijeiro et al. incorporated both gene expression and clinical information in the survival analysis [79]).

Though feature selection and sophisticated analysis methods can somewhat avoid over-fitting and improve classification results, we cannot be sure whether improved results are the consequence of improved methods or

chance. There are several methods aimed at evaluating the analysis results or increasing the stability of the classifier. Testing the classifier with an independent dataset is usually the unbiased way to measure the performance of the classifier. If the independent test set is unavailable, cross validation is also a common way to estimate the performance of the classifier, *n*-fold cross validation means separating the original dataset into *n* subsets and each time choosing one subset as test set and others as training set, leaving one out cross validation means each time leaving one sample as test set and others as training set to evaluate the performance of classifier. For microarray data analysis, feature selection steps are external to the cross validation process; there are two schemes to evaluate the performance when feature selection is used. As defined by Zhang et al. [51], CV1 is the scheme where feature selection is done with all the samples and cross validation is only for the classification step, CV2 is the scheme where the cross validation is used both for feature selection and classification which means leaving the test set first then performing feature selection and classification with the other samples. Since CV1 scheme uses all the samples to do the feature selection which provides information for the classification step, the cross validation result of CV1 scheme will be optimistic and that of CV2 is more faithful. Though cross validation provides a way to evaluate the performance of the data analysis methods, there is still a large variation in cross validation results. Sometimes permutation method is often used to permute the label of samples to estimate the null distribution of the statistics or classification accuracy. It is a non-parametric way to test the significance of the analysis result [45,80,81]. Boosting is used to increase the accuracy and stability by increasing the weight of misclassified samples [43,82] and bagging is used for the same purpose by combining the several classifiers trained on bootstrapping datasets [82]. There are also some methods to detect possible mislabeled samples in an iterative way [72,83].

4 Biology behind data

Simply clustering the data, selecting some differentially expressed genes, or building a classifier for microarray analysis is not always enough. Sometimes we need to know more about the biology behind the data, such as which pathway or biological function is associated with some specific type of cancer. As we learn more about gene function, methods have been proposed to analyze the expression data with known gene annotations.

Gamberoni et al. proposed a novel approach that studies the correlations between genes and their relation to Gene Ontology (GO) [84]. The method first selects a significantly correlated pair of genes, and then counts the number of gene pairs linked to the same GO term. Finally, the method uses a bootstrap analysis to evaluate whether the numbers of gene pairs linked to the same GO term is significantly higher than that in a simulation study. Subramanian et al. proposed a method called gene set enrichment analysis (GSEA) to evaluate whether a predefined group of genes have different expression patterns in two classes of samples [85]. Similar to the GSEA approach, Al-Shahrour et al. proposed a method to find groups of genes that have common functional labels which are significantly over- or under-expressed as a block [86].

5 Discussion

Microarray data analysis is a biology-driven problem; different biological purposes will need different analysis methods. Though the existing methods have met data analysis requests, there are still many directions that need more focus. For example, most of the existing classification methods cannot use clinical data and expression data simultaneously. This leads to the question of how to effectively remove the dominant factor in the dataset.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 60575014, 30625012 and 60721003), and National High-tech R&D Program (No. 2006AA02Z325).

References

1. Golub T R, Slonim D K, Tamayo P, Huard C, Gaasenbeek M, Mesirov J P, Coller H, Loh M L, Downing J R, Caligiuri M A, Bloomfield C D, Lander E S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 1999, 286(5439): 531–537
2. Alon U, Barkai N, Notterman D A, Gish K, Ybarra S, Mack D, Levine A J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences of the United States of America, 1999, 96(12): 6745–6750
3. van't Veer L J, Dai H, van de Vijver M J, He Y D, Hart A A, Mao M, Peterse H L, van der Kooy K, Marton M J, Witteveen A T, Schreiber G J, Kerkhoven R M, Roberts C, Linsley P S, Bernards R, Friend S H. Gene expression profiling predicts clinical outcome of breast cancer. Nature, 2002, 415(6871): 530–536
4. Alizadeh A A, Eisen M B, Davis R E, Ma C, Lossos I S, Rosenwald A, Boldrick J C, Sabet H, Tran T, Yu X, Powell J I, Yang L, Marti G E, Moore T, Hudson J Jr, Lu L, Lewis D B, Tibshirani R, Sherlock G, Chan W C, Greiner T C, Weisenburger D D, Armitage J O, Warnke R, Levy R, Wilson W, Grever M R, Byrd J C, Botstein D, Brown P O, Staudt L M. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature, 2000, 403(6769): 503–511
5. Beer D G, Kardia S L, Huang C C, Giordano T J, Levin A M, Misek D E, Lin L, Chen G, Gharib T G, Thomas D G, Lizyness M L, Kuick R, Hayasaka S, Taylor J M, Iannettoni M D, Orringer M B, Hanash S. Gene-expression profiles predict survival of patients with lung

adenocarcinoma. Nature Medicine, 2002, 8(8): 816–824

6. Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, Sampas N, Dougherty E, Wang E, Marincola F, Gooden C, Lueders J, Glatfelter A, Pollock P, Carpten J, Gillanders E, Leja D, Dietrich K, Beaudry C, Berens M, Alberts D, Sondak V. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature, 2000, 406(6795): 536–540
7. Dyrskjøt L, Thykjaer T, Kruhøffer M, Jensen J L, Marcussen N, Hamilton-Dutoit S, Wolf H, Orntoft T F. Identifying distinct classes of bladder carcinoma using microarrays. Nature Genetics, 2003, 33 (1): 90–96
8. Kapp A V, Jeffrey S S, Langerød A, Børresen-Dale A L, Han W, Noh D Y, Bukholm I R, Nicolau M, Brown P O, Tibshirani R. Discovery and validation of breast cancer subtypes. BMC Genomics, 2006, 7: 231
9. Ross D T, Scherf U, Eisen M B, Perou C M, Rees C, Spellman P, Iyer V, Jeffrey S S, van de Rijn M, Waltham M, Pergamenschikov A, Lee J C, Lashkari D, Shalon D, Myers T G, Weinstein J N, Botstein D, Brown P O. Systematic variation in gene expression patterns in human cancer cell lines. Nature Genetics, 2000, 24(3): 227–235
10. Huang Y, Prasad M, Lemon W J, Hampel H, Wright F A, Kornacker K, LiVolsi V, Frankel W, Kloos R T, Eng C, Pellegata N S, de la Chapelle A. Gene expression in papillary thyroid carcinoma reveals highly consistent profiles. Proceedings of the National Academy of Sciences of the United States of America, 2001, 98(26): 15044–15049
11. Hastie T, Tibshirani R, Botstein D, Brown P. Supervised harvesting of expression trees. Genome Biology, 2001, 2(1): research0003.1–research0003.12
12. Huang E, Cheng S H, Dressman H, Pittman J, Tsou M H, Horng C F, Bild A, Iversen E S, Liao M, Chen C M, West M, Nevins J R, Huang A T. Gene expression predictors of breast cancer outcomes. Lancet, 2003, 361(9369): 1590–1596
13. Nilsson J, Fioretos T, Höglund M, Fontes M. Approximate geodesic distances reveal biologically relevant structures in microarray data. Bioinformatics, 2004, 20(6): 874–880
14. Boratyn G M, Datta S, Datta S. Incorporation of biological knowledge into distance for clustering genes. Bioinformation, 2007, 1(10): 396–405
15. Bagirov A M, Ferguson B, Ivkovic S, Saunders G, Yearwood J. New algorithms for multi-class cancer diagnosis using tumor gene expression signatures. Bioinformatics, 2003, 19(14): 1800–1807
16. Gao Y, Church G. Improving molecular cancer class discovery through sparse non-negative matrix factorization. Bioinformatics, 2005, 21(21): 3970–3975
17. Sese J, Kurokawa Y, Monden M, Kato K, Morishita S. Constrained clusters of gene expression profiles with pathological features. Bioinformatics, 2004, 20(17): 3137–3145
18. Dotan-Cohen D, Melkman A A, Kasif S. Hierarchical tree snipping: clustering guided by prior knowledge. Bioinformatics, 2007, 23 (24): 3335–3342
19. Belacel N, Cuperlović-Culf M, Laflamme M, Ouellette R. Fuzzy J-means and VNS methods for clustering genes from microarray data. Bioinformatics, 2004, 20(11): 1690–1701
20. Getz G, Levine E, Domany E. Coupled two-way clustering analysis of gene microarray data. Proceedings of the National Academy of Sciences of the United States of America, 2000, 97(22): 12079–12084
21. Getz G, Gal H, Kela I, Notterman D A, Domany E. Coupled two-way clustering analysis of breast cancer and colon cancer gene expression data. Bioinformatics, 2003, 19(9): 1079–1089
22. Kluger Y, Basri R, Chang J T, Gerstein M. Spectral biclustering of microarray data: coclustering genes and conditions. Genome Research, 2003, 13(4): 703–716
23. Wang J, Delabie J, Aasheim H, Smeland E, Myklebost O. Clustering of the SOM easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study. BMC Bioinformatics, 2002, 3: 36
24. Hanczar B, Courtine M, Benis A, Hennegar C, Clément K, Zucker J D. Improving classification of microarray data using prototype-based feature selection. ACM SIGKDD Explorations Newsletter, 2003, 5(2): 23–30
25. Crescenzi M, Giuliani A. The main biological determinants of tumor line taxonomy elucidated by a principal component analysis of microarray data. FEBS Letters, 2001, 507(1): 114–118
26. Hsu A L, Tang S L, Halgamuge S K. An unsupervised hierarchical dynamic self-organizing approach to cancer class discovery and marker gene identification in microarray data. Bioinformatics, 2003, 19(16): 2131–2140
27. Li W, Fan M, Xiong M. SamCluster: an integrated scheme for automatic discovery of sample classes using gene expression profile. Bioinformatics, 2003, 19(7): 811–817
28. Dudoit S, Fridlyand J, Speed T. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. Technical Report 576. Berkeley, CA: Department of Statistics, University of California, 2000
29. Smolkin M, Ghosh D. Cluster stability scores for microarray data in cancer studies. BMC Bioinformatics, 2003, 4: 36
30. Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learning, 2003, 52(1–2): 91–118
31. Bhattacharjee A, Richards W G, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark E J, Lander E S, Wong W, Johnson B E, Golub T R, Sugarbaker D J, Meyerson M. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proceedings of the National Academy of Sciences of the United States of America, 2001, 98(24): 13790–13795
32. Dudoit S, Fridlyand J. Bagging to improve the accuracy of a clustering procedure. Bioinformatics, 2003, 19(9): 1090–1099
33. Swift S, Tucker A, Vinciotti V, Martin N, Orengo C, Liu X, Kellam P. Consensus clustering and functional interpretation of gene-expression data. Genome Biology, 2004, 5(11): R94
34. Martoglio A M, Miskin J W, Smith S K, MacKay D J. A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer. Bioinformatics, 2002, 18(12): 1617–1624
35. West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson J A Jr, Marks J R, Nevins J R. Predicting the clinical

status of human breast cancer by using gene expression profiles. Proceedings of the National Academy of Sciences of the United States of America, 2001, 98(20): 11462–11467
36. Liebermeister W. Linear modes of gene expression determined by independent component analysis. Bioinformatics, 2002, 18(1): 51–60
37. Pomeroy S L, Tamayo P, Gaasenbeek M, Sturla L M, Angelo M, McLaughlin M E, Kim J Y, Goumnerova L C, Black P M, Lau C, Allen J C, Zagzag D, Olson J M, Curran T, Wetmore C, Biegel J A, Poggio T, Mukherjee S, Rifkin R, Califano A, Stolovitzky G, Louis D N, Mesirov J P, Lander E S, Golub T R. Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature, 2002, 415(6870): 436–442
38. Gordon G J, Richards W G, Sugarbaker D J, Jaklitsch M T, Bueno R. A prognostic test for adenocarcinoma of the lung from gene expression profiling data. Cancer Epidemiology, Biomarkers & Prevention, 2003, 12(9): 905–910
39. Gordon G J, Jensen R V, Hsiao L L, Gullans S R, Blumenstock J E, Ramaswamy S, Richards W G, Sugarbaker D J, Bueno R. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Research, 2002, 62(17): 4963–4967
40. Dabney A R. Classification of microarrays to nearest centroids. Bioinformatics, 2005, 21(22): 4148–4154
41. Thomas J G, Olson J M, Tapscott S J, Zhao L P. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. Genome Research, 2001, 11(7): 1227–1236
42. Troyanskaya O G, Garber M E, Brown P O, Botstein D, Altman R B. Nonparametric methods for identifying differentially expressed genes in microarray data. Bioinformatics, 2002, 18(11): 1454–1461
43. Dettling M, Bühlmann P. Boosting for tumor classification with gene expression data. Bioinformatics, 2003, 19(9): 1061–1069
44. Broët P, Lewin A, Richardson S, Dalmasso C, Magdelenat H. A mixture model-based strategy for selecting sets of genes in multiclass response microarray experiments. Bioinformatics, 2004, 20(16): 2562–2571
45. Nutt C L, Mani D R, Betensky R A, Tamayo P, Cairncross J G, Ladd C, Pohl U, Hartmann C, McLaughlin M E, Batchelor T T, Black P M, von Deimling A, Pomeroy S L, Golub T R, Louis D N. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. Cancer Research, 2003, 63(7): 1602–1607
46. Gormley M, Dampier W, Ertel A, Karacali B, Tozeren A. Prediction potential of candidate biomarker sets identified and validated on gene expression data from multiple datasets. BMC Bioinformatics, 2007, 8: 415
47. Furey T S, Cristianini N, Duffy N, Bednarski D W, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 2000, 16(10): 906–914
48. Li J, Wong L. Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns. Bioinformatics, 2002, 18(5): 725–734
49. Bø T H, Jonassen I. New feature subset selection procedures for classification of expression profiles. Genome Biology, 2002, 3(4): research0017.1–research0017.11
50. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning, 2002, 46(1–3): 389–422
51. Zhang X, Lu X, Shi Q, Xu X Q, Leung H C, Harris L N, Iglehart J D, Miron A, Liu J S, Wong W H. Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. BMC Bioinformatics, 2006, 7: 197
52. Furlanello C, Serafini M, Merler S, Jurman G. Entropy-based gene ranking without selection bias for the predictive classification of microarray data. BMC Bioinformatics, 2003, 4: 54
53. Li W, Xiong M. Tclass: tumor classification system based on gene expression profile. Bioinformatics, 2002, 18(2): 325–326
54. Inza I, Sierra B, Blanco R, Larrañaga P. Gene selection by sequential search wrapper approaches in microarray cancer class prediction. Journal of Intelligent and Fuzzy Systems, 2002, 12(1): 25–33
55. Xiong M, Fang X, Zhao J. Biomarker identification by feature wrappers. Genome Research, 2001, 11(11): 1878–1887
56. Liu J J, Cutler G, Li W, Pan Z, Peng S, Hoey T, Chen L, Ling X B. Multiclass cancer classification and biomarker discovery using GA-based algorithms. Bioinformatics, 2005, 21(11): 2691–2697
57. Peng S, Xu Q, Ling X B, Peng X, Du W, Chen L. Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines. FEBS Letters, 2003, 555(2): 358–362
58. Li L, Weinberg C R, Darden T A, Pedersen L G. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics, 2001, 17(12): 1131–1142
59. Ooi C H, Tan P. Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. Bioinformatics, 2003, 19 (1): 37–44
60. Deutsch J M. Evolutionary algorithms for finding optimal gene sets in microarray prediction. Bioinformatics, 2003, 19(1): 45–52
61. Jirapech-Umpai T, Aitken S. Feature selection and classification for microarray data analysis: evolutionary methods for identifying predictive genes. BMC Bioinformatics, 2005. 6: 148
62. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics, 2007, 23(19): 2507–2517
63. Krishnapuram B, Carin L, Hartemink A J. Joint classifier and feature optimization for comprehensive cancer diagnosis using gene expression data. Journal of Computational Biology, 2004, 11(2–3): 227–242
64. Cawley G C, Talbot N L C. Gene selection in cancer classification using sparse logistic regression with Bayesian regularization. Bioinformatics, 2006, 22(19): 2348–2355
65. Nguyen D V, Rocke D M. Tumor classification by partial least squares using microarray gene expression data. Bioinformatics, 2002, 18(1): 39–50
66. Nguyen D V, Rocke D M. Partial least squares proportional hazard regression for application to DNA microarray survival data. Bioinformatics, 2002, 18(12): 1625–1632
67. Chang H Y, Nuyten D S, Sneddon J B, Hastie T, Tibshirani R, Sørlie T, Dai H, He Y D, van't Veer L J, Bartelink H, van de Rijn M,

Brown P O, van de Vijver M J. Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. Proceedings of the National Academy of Sciences of the United States of America, 2005, 102 (10): 3738–3743

68. Khan J, Wei J S, Ringnér M, Saal L H, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu C R, Peterson C, Meltzer P S. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine, 2001, 7(6): 673–679
69. O'Neill M, Song L. Neural network analysis of lymphoma microarray data: prognosis and diagnosis near-perfect. BMC Bioinformatics, 2003, 4: 13
70. Liu B, Cui Q, Jiang T, Ma S. A combinational feature selection and ensemble neural network method for classification of gene expression data. BMC Bioinformatics, 2004, 5: 136
71. Linder R, Dew D, Sudhoff H, Theegarten D, Remberger K, Pöppl S J, Wagner M. The 'subsequent artificial neural network' (SANN) approach might bring more classificatory power to ANN-based DNA microarray analyses. Bioinformatics, 2004, 20(18): 3544–3552
72. Zhang W, Rekaya R, Bertrand K. A method for predicting disease subtypes in presence of misclassification among training samples using gene expression: application to human breast cancer. Bioinformatics, 2006, 22(3): 317–325
73. Gevaert O, De Smet F, Timmerman D, Moreau Y, De Moor B. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. Bioinformatics, 2006, 22 (14): e184–e190
74. Goeman J J, Oosting J, Cleton-Jansen A M, Anninga J K, van Houwelingen H C. Testing association of a pathway with survival using gene expression data. Bioinformatics, 2005, 21(9): 1950–1957
75. Gui J, Li H. Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. Bioinformatics, 2005, 21(13): 3001–3008
76. Schumacher M, Binder H, Gerds T. Assessment of survival prediction models based on microarray data. Bioinformatics, 2007, 23(14): 1768–1774
77. Kaderali L, Zander T, Faigle U, Wolf J, Schultze J L, Schrader R. CASPAR: a hierarchical bayesian approach to predict survival times in cancer from gene expression data. Bioinformatics, 2006, 22(12): 1495–1502
78. Parmigiani G, Garrett-Mayer E S, Anbazhagan R, Gabrielson E. A cross-study comparison of gene expression studies for the molecular classification of lung cancer. Clinical Cancer Research, 2004, 10(9): 2922–2927
79. Fernandez-Teijeiro A, Betensky R A, Sturla L M, Kim J Y, Tamayo P, Pomeroy S L. Combining gene expression profiles and clinical parameters for risk stratification in medulloblastomas. Journal of Clinical Oncology, 2004, 22(6): 994–998
80. Barry W T, Nobel A B, Wright F A. Significance analysis of functional categories in gene expression studies: a structured permutation approach. Bioinformatics, 2005, 21(9): 1943–1949
81. Zhang C, Lu X, Zhang X. Significance of gene ranking for classification of microarray samples. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2006, 3(3): 312–320
82. Dettling M. BagBoosting for tumor classification with gene expression data. Bioinformatics, 2004, 20(18): 3583–3593
83. Lu X, Li Y, Zhang X. A simple strategy for detecting outlier samples in microarray data. In: Proceedings of the Eighth International Conference on Control, Automation, Robotics and Vision. Kunming: IEEE, 2004, 2: 1331–1335
84. Gamberoni G, Storari S, Volinia S. Finding biological process modifications in cancer tissues by mining gene expression correlations. BMC Bioinformatics, 2006, 7: 6
85. Subramanian A, Tamayo P, Mootha V K, Mukherjee S, Ebert B L, Gillette M A, Paulovich A, Pomeroy S L, Golub T R, Lander E S, Mesirov J P. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America, 2005, 102(43): 15545–15550
86. Al-Shahrour F, Diaz-Uriarte R, Dopazo J. Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. Bioinformatics, 2005, 21(13): 2988–2993