Expert Systems with Applications 36 (2009) 9919–9923

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

# A sequential feature extraction approach for naïve bayes classification of microarray data

Liwei Fan [a,*], Kim-Leng Poh [a], Peng Zhou [b]

[a] Department of Industrial and Systems Engineering, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260, Singapore
[b] College of Economics and Management, Nanjing University of Aeronautics and Astronautics, 29 Yudao Street, Nanjing 210016, PR China

ARTICLE INFO

*Keywords:*
Microarray data
Naïve Bayes
Feature extraction
Independent component analysis (ICA)
Stepwise regression

ABSTRACT

Accurate classification of microarray data plays a vital role in cancer prediction and diagnosis. Previous studies have demonstrated the usefulness of naïve Bayes classifier in solving various classification problems. In microarray data analysis, however, the conditional independence assumption embedded in the classifier itself and the characteristics of microarray data, e.g. the extremely high dimensionality, may severely affect the classification performance of naïve Bayes classifier. This paper presents a sequential feature extraction approach for naïve Bayes classification of microarray data. The proposed approach consists of feature selection by stepwise regression and feature transformation by class-conditional independent component analysis. Experimental results on five microarray datasets demonstrate the effectiveness of the proposed approach in improving the performance of naïve Bayes classifier in microarray data analysis.



## 1. Introduction

Recent advancements in DNA microarray technology have enabled us to monitor and measure the expression levels of hundreds of thousands of genes simultaneously, which allowed a great deal of microarray data to be generated. Accordingly, microarray data analysis, which can provide useful information for cancer prediction and diagnosis, has also attracted many researchers from diverse areas. Different methods, from simple statistical techniques such as linear regression to complex machine learning algorithms such as support vector machines, have been employed to select informative genes and do classification of microarray data (Guyon, Weston, & Barnhill, 2002; Huang & Pan, 2003; Kim & Cho, 2004; Lin & Chien, in press; Park, Yoo, & Cho, 2007; Wong & Hsu, 2008; Zheng, Huang, & Shang, 2006).

Naïve Bayes classifier is a simple Bayesian network classifier built upon the strong assumption that different attributes are independent with each other given the class (Friedman, Geiger, & Goldszmidt, 1997; Langley, Iba, & Thompson, 1992). Despite its simplicity, naïve Bayes classifier has been found to be surprisingly effective compared with other more sophisticated classifiers (Hall, 2007). It is therefore not surprising that naïve Bayes classifier has gained popularity in solving various classification problems including microarray data analysis, e.g. Sandberg et al. (2001), Kelemen, Zhou, Lawhead, and Liang (2003) Chen, Huang, Tian, and Qu (2009). In addition to its application, several recent studies also focus on the theoretical aspect of naïve Bayes classifier. For instance, Hsu, Huang, and Chang (2008) extend the traditional naïve Bayes classifier to deal with the mix data with both discrete and continuous attribute values.

Nevertheless, there exist two major limitations that may severely affect the successful application of naïve Bayes classifier to microarray data analysis (Fan & Poh, 2008). The first is the conditional independence assumption embedded in the classifier itself, which is hardly satisfied by the microarray data. This limitation could be, at least theoretically, overcome by the class-conditional independent component analysis (CC-ICA) technique proposed in Vitria, Bressan, and Radeva (2007). Previous experimental results have shown that CC-ICA could effectively improve the performance of naïve Bayes classifier in some application domains (Fan & Poh, 2007).

Another limitation comes from the intrinsic characteristics of microarray dataset, which usually consists of thousands of genes with only tens of samples due to the expensive experiment. The extremely high dimensionality of microarray data may greatly increase the computational costs of naïve Bayes classifier. In addition, since the sample size is far less than the gene size, the use of CC-ICA can hardly enhance the independence among genes as well as improve the performance of naïve Bayes classifier. When the sample size in some classes is not large enough to do ICA, the implementation of CC-ICA even becomes infeasible (Fan & Poh, 2007). It is therefore necessary to do feature selection to reduce the dimensionality of genes before applying CC-ICA for naïve Bayes classification of microarray data.

* Corresponding author. Tel.: +65 6516 4573; fax: +65 6777 1434.
*E-mail address:* g0600308@nus.edu.sg (L. Fan).

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.01.075

In this paper, we propose a sequential feature extraction approach for naïve Bayes classification of microarray data. The proposed feature extraction approach consists of two main steps, namely feature selection by stepwise regression and feature transformation by CC-ICA, which will be introduced in Section 2. In Section 3, we present the experimental results on five microarray datasets, which show that the proposed approach can not only improve the average classification accuracy rates but also reduce the variation of classification performance. Section 4 concludes this paper.

## 2. Proposed approach

Assume that there are $K$ samples and $M$ genes (usually $K \ll M$), and the expression level of gene $i$ for sample $k$ is $x_{ki}$. Let $x_k = (x_{k1}, x_{k2}, \ldots, x_{kM})$ and $X = (x_{ki})_{K\times M}$ respectively denote the gene expression profile of sample $k$ and the summarized microarray data matrix. Let $g_i$ ($i = 1,2,\ldots, M$) denote the variable representing gene $i$. Further assume that the class label of sample $k$ is $c_k$ where $c_k \in \Omega = \{1,2,\ldots, L\}$. Let $c$ and $C = (c_1, c_2, \ldots, c_K)^T$ respectively denote the class variable and the column vector of class labels for the $K$ samples. The purpose is to train a naïve Bayes classifier based on $X$ and $C$, which may be used to accurately classify a given test sample with unknown class labels.

### 2.1. Stepwise regression-based feature selection

One specific characteristic of microarray data is that its feature (gene) size is far larger than sample size, which is known as "the curse of dimensionality problem". It is therefore necessary to do feature selection on the original dataset. Effective feature selection can reduce the complexity in computation, increase the classification accuracy and enhance the generalization property of classifiers (Ding & Peng, 2005).

A number of methods have been developed and applied to do feature selection. A relatively comprehensive overview on alternative feature selection methods can be found in (Guyon & Elisseeff, 2003). In general, feature selection methods can be divided into two big categories, namely filtering approach and wrapper approach. In microarray data analysis, filtering approach seems to be more popular. Although many filtering methods focus on the rankings of individual genes in terms of their relevance with class variable, recent studies have shown that the methods following the "minimum redundancy – maximum relevance" principle may select more representative genes (Ding & Peng, 2005; Park et al., 2007).

Stepwise regression is an automatic statistical procedure for selecting the representative predictive variables, e.g., genes in microarray data, to build good regression models. At each step, forward selection adds the most statistically significant variable and backward selection deletes the least significant variable provided that the $p$-values for the two variables are respectively less than $p_{in}$ and larger than $p_{out}$, where $p_{in}$ and $p_{out}$ are the probabilities of Type I error related to entering and deleting a variable. Conceptually, stepwise regression also follows the "minimum redundancy – maximum relevance" principle as adopted by several recently proposed feature selection methods. Meanwhile, it is simple and easy to implement but has still good performance (Park et al., 2007). Therefore, we propose the use of stepwise regression rather than other methods for gene selection in our study. In terms of the determination of $p_{in}$ and $p_{out}$, a rule of thumb is to let them small enough so that the number of genes selected is less than the number of samples (in order to do CC-ICA effectively). Without loss of generality, we assume that only the first $N$ ($N < K$) genes are retained after stepwise regression-based feature selection. The microarray data matrix after feature selection is denoted by $Y$ where $Y = (g_1, g_2, \ldots, g_N) = (x_{ki})_{K\times N}$.

### 2.2. ICA-based feature transformation

ICA is a relatively new multivariate statistical technique for finding the hidden factors that underlie a set of random variables (Hyvärinen, Karhunen, & Oja, 2001). Compared to principal component analysis that attempts to transform these variables into a set of uncorrelated variables, ICA attempts to transform them into new variables that are mutually independent or as independent as possible with each other (Hyvärinen et al., 2001). It is therefore a more powerful technique that has been widely used in solving various classification problems, e.g. microarray data analysis (Zheng et al., 2006) and ECG beat classification (Yu & Chou, 2007, 2008b, 2008a).

**Table 2**
Classification accuracy rates (%) of three classification rules on five datasets.

| Dataset | Leukemia-ALLAML | Leukemia-MLL | Colon Tumor | Lung Cancer I | Lung Cancer II |
|---|---|---|---|---|---|
| FS + NB | 66.7 | 43.1 | 74.2 | 78.8 | 87.8 |
| FS + ICA + NB | 88.9 | 77.8 | 80.6 | 80.8 | 92.8 |
| FS + CCICA + NB | 95.8 | 83.3 | 82.3 | 82.3 | 98.3 |

**Table 1**
Summary of five microarray datasets.

| Dataset | Leukemia-ALLAML | Leukemia-MLL | Colon Tumor | Lung Cancer I | Lung Cancer II |
|---|---|---|---|---|---|
| Data source | Golub et al. (1999) | Armstrong et al. (2002) | Alon et al. (1999) | Bhattacharjee et al. (2001) | Gordon et al. (2002) |
| Number of attributes | 7129 | 12528 | 2000 | 12600 | 12533 |
| Number of classes | 2 | 3 | 2 | 5 | 2 |
| Number of instances | 62 | 72 | 62 | 203 | 181 |

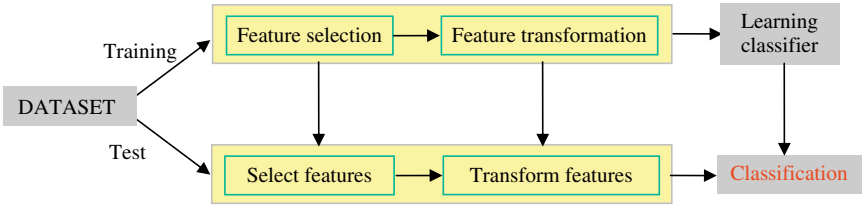


**Fig. 1.** Flow chart for implementing FS + ICA/CCICA + NB.

Given the microarray data matrix $Y$, the basic ICA model for feature transformation can be written as

$$Z^T = W \cdot Y^T \tag{1}$$

where $W$ is a $N$ by $N$ mixing matrix and $Z$ is a $K$ by $N$ source matrix. Every column of $Z$ represents one "independent component" and all the columns consist of the new features for classification purpose. The task is to estimate $W$ and $Z$. In ICA, there are many principles and algorithms for performing the task (Hyvärinen et al., 2001). We here suggest the use of FastICA algorithm, proposed by Hyvarinen and Oja (1997) and widely accepted as a computationally highly efficient method, to get $W$ and $Z$.

Class-conditional ICA (CC-ICA) is built upon the idea that ICA is done within each class so that one mixing matrix can be obtained for each class (Vitria et al., 2007). In this way, the new attributes after transformation may satisfy the class-conditional independence assumption of the naïve Bayes classifier well. If we split the microarray data matrix $Y$ into a set of sub-matrices $Y_l (l = 1,2,\ldots, L)$ according to the class label, the set of models for doing CC-ICA can be written as

$$Z_l^T = W_l \cdot Y_l^T \tag{2}$$

where $W_l$ is a $N$ by $N$ mixing matrix and $Z_l$ is a $K_l$ by $N$ source matrix for class $l$. Similarly, we can still use FastICA algorithm to estimate $W_l$ and $Z_l$ for each class.

### 2.3. Naïve bayes classifier

We shall use the data after feature extraction, i.e. $Z_l (l = 1,2,\ldots, L)$ to build a naïve Bayes classifier, which is used to classify a new test sample with attribute (gene) values $z_1, z_2, \ldots, z_N$ (after ICA-based feature transformation). In general, Bayesian network classifier computes the posterior probability that the sample belongs to class $c$ by using the Bayes rule as follows:

$$p(c|z_1, z_2, \ldots, z_N) = \frac{p(z_1, z_2, \ldots, z_N|c)p(c)}{p(z_1, z_2, \ldots, z_n)} \tag{3}$$

If the assumption of class-conditional independence among attributes is imposed, the following Naïve Bayes classifier can be obtained

$$p(c|z_1, z_2, \ldots, z_n) = \frac{p(c)\prod_{i=1}^{n} p(z_i|c)}{p(z_1, z_2, \ldots, z_n)} \tag{4}$$

Since $p(z_1, z_2, \ldots, z_n)$ is a common factor for a certain sample, it can be ignored in classification process. In addition, since the attribute variables are continuous in microarray data analysis, we can use the probability density value $f(z_i|c)$ to replace the probability value $p(z_i|c)$. The class-conditional probability density $f(\cdot|c)$ for each attribute and the prior $p(c)$ can be obtained from the learning process. We here suggest the use of nonparametric kernel density estimation method to estimate $f(\cdot|c)$. As a result, the final naïve Bayes classification model can be written as

$$c^* = \arg\max_{c \in \Omega} p(c) \prod_{i=1}^{n} f(z_i|c) \tag{5}$$

## 3. Experimental results

We evaluate the performance of the proposed feature extraction approach for naïve Bayes classifier based on five well-known gene expression datasets, namely Leukemia-ALLAML, Leukemia-MLL, Colon Tumor, Lung Cancer I and Lung Cancer II. Table 1 shows the five datasets with their properties. In addition to feature selection integrated with CC-ICA plus naïve Bayes classifier (FS + CCICA + NB),

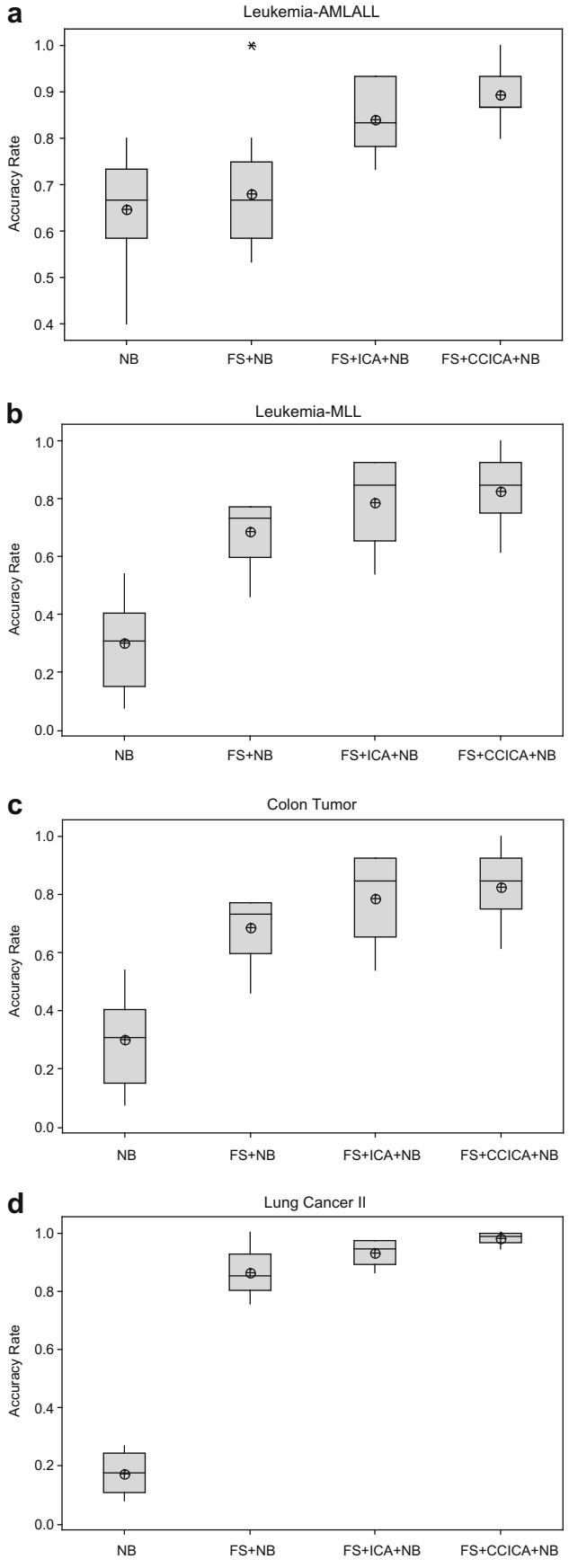


**Fig. 2.** Boxplots of four classification rules in terms of holdout classification accuracy rates.

we also implement three other classification rules, namely naïve Bayes classifier (NB), feature selection plus naïve Bayes classifier (FS + NB), and feature selection integrated with ICA plus naïve Bayes classifier (FS + ICA + NB) on the five datasets. Since the proposed sequential feature extraction approach aims to address the issues arising from naïve Bayes classification of microarray data, its integrations with other popular classifiers such as support vector machines are not considered in our experimental study.

Both leave-one-out and hold-out classification accuracy rates are used to give a relatively comprehensive comparison on the performances of alternative classification rules. Every dataset is partitioned into two parts, i.e. training and test datasets. The training dataset is used to do feature selection, carry out ICA/CC-ICA computation and train classifiers. The test dataset is used to evaluate the performances of alternative classifiers. Fig. 1 shows the flow chart for implementing FS + ICA + NB or FS + CCICA + NB on the five datasets.

For leave-one-out experiments, the pure naïve Bayes classifier was not included due to its extremely time-consuming computations. The classification accuracy rates for the other three classifiers are displayed in Table 2. It can be seen from Table 2 that both FS + CCICA + NB and FS + ICA + NB perform better than FS + NB in microarray data analysis, which demonstrates the effectiveness of the proposed approach. As for the comparison between the former two classification rules, FS + CCICA + NB performs obviously better than FS + ICA + NB in terms of classification accuracy.

Since leave-one-out classification accuracy rates cannot provide the information on the variation of classification performance, we applied holdout classification accuracy rates to further evaluate the performances of alternative classification rules. In our experiment, four fifth of the samples are randomly selected as the training data and the remaining one fifth of the samples are taken as the test data. Such a procedure is repeated ten times for each classification rule on the four datasets exclusive of the Lung Cancer I dataset, which is due to the fact that some classes in the training data for this dataset have not enough samples to implement CC-ICA.

Fig. 2 shows the box plots of the holdout classification accuracy rates for the four classification rules on the four datasets. It can be found that FS + CCICA + NB and FS + ICA + NB have better classification performances than FS + NB or NB, which is consistent with the leave-one-out classification results. Only feature selection by stepwise regression could improve the naïve Bayes classification accuracy rates, whereas the degree of performance improvement depends on the dataset. For instance, as shown in Fig. 2, the discrepancy between NB and FS + NB is not obvious for the first two datasets. However, for the last two datasets the classification performance of FS + NB is significantly better than that of NB. Although feature selection may not always be effective, its integration with ICA/CC-ICA transformation has been found to certainly improve the classification performance.

It can be seen from Fig. 2 that the FS + CCICA + NB is generally superior to the FS + ICA + NB in the sense that the former is more stable than the latter in terms of classification performance. The possible reason is that feature transformation by CC-ICA seems to be more reasonable for the data to satisfy the class-conditional independence assumption underlying naïve Bayes classifier. However, a limitation of FS + CCICA + NB is that it may not be implemented when the sample size in some classes is too small to do ICA for each class. In such cases, FS + ICA + NB is recommended for use since it still performs better than NB and FS + NB.

## 4. Conclusion

This paper presents a sequential feature extraction approach for naïve Bayes classification of microarray data. The proposed feature extraction approach starts from gene selection by stepwise regression, which is a simple but effective dimension reduction technique following the "minimum redundancy – maximum relevance" principle. The data on the genes selected are then transformed by CC-ICA, which makes the new features after transformation become as independent as possible. Our experimental results on five microarray datasets demonstrate the effectiveness of the proposed approach in improving the classification performance of naïve Bayes classifier in microarray data analysis. It is found that FS + CCICA + NB has not only the highest averages but also the lowest standard deviations in terms of classification accuracy rates.

## References


Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., et al. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy Sciences of the United States of America, 96*, 6745–6750.

Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D., et al. (2002). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics, 30*, 41–47.

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., et al. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy Sciences of the United States of America, 98*, 13790–13795.

Chen, J., Huang, H., Tian, F., & Qu, Y. (2009). Feature selection for text classification with naïve Bayes. *Expert Systems with Applications 36*, 5432–5435.

Ding, C., & Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology, 3*, 185–205.

Fan, L., & Poh, K. L. (2007). A comparative study of PCA, ICA and class-conditional ICA for naïve Bayes classifier. *Lecture notes in computer science (LNCS),* (Vol. 4507, pp. 16–22). Berlin: Springer.

Fan, L., & Poh, K. L. (2008). Improving the naïve Bayes classifier. In J. R. R. Dopico, J. Dorado, & A. Pazos (Eds.). *Encyclopedia of artificial intelligence,* (Vol. 3, pp. 879–883). IGI Publishing.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning, 29*, 131–163.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science, 286*, 531–537.

Gordon, G. J., Jensen, R. V., Hsiao, L. L., Gullans, S. R., Blumenstock, J. E., Ramaswamy, S., 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.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182.

Guyon, I., Weston, J., & Barnhill, S. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning, 46*, 389–422.

Hall, M. (2007). A decision tree-based attribute weighting filter for naïve Bayes. *Knowledge-Based Systems, 20*, 120–126.

Hsu, C. C., Huang, Y. P., & Chang, K. W. (2008). Extended naïve Bayes classifier for mixed data. *Expert Systems with Applications, 35*, 1080–1083.

Huang, X., & Pan, W. (2003). Linear regression and two-class classification with gene expression data. *Bioinformatics, 19*, 2072–2078.

Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis*. New York: John Wiley & Sons.

Hyvarinen, A., & Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural Computation, 9*, 1483–1492.

Kelemen, A., Zhou, H., Lawhead, P., & Liang, Y. (2003). Naïve Bayesian classifier for microarray data. In *Proceedings of the international joint conference on neural networks* (pp. 1769–1773).

Kim, K. J., & Cho, S. B. (2004). Prediction of colon cancer using an evolutionary neural network. *Neurocomputing, 61*, 361–379.

Langley, P., Iba, W., & Thompson, K. (1992). An analysis of Bayesian classifiers. In *Proceedings of the international conference on artificial intelligence*.

Lin, K. S., & Chien, C. F. (in press). Cluster analysis of genome-wide expression data for feature extraction. *Expert systems with applications*. doi:10.1016/j.eswa.2008.01.068.

Park, H. S., Yoo, S. H., & Cho, S. B. (2007). Forward selection method with regression analysis for optimal gene selection in cancer classification. *International Journal of Computer Mathematics, 84*, 653–668.

Sandberg, R., Winberg, G., Bränden, C., Kaske, A., Ernberg, I., & Cöster, J. (2001). Capturing whole-genome characteristics in short sequences using a naïve Bayesian classifier. *Genome Research, 11*, 1404–1409.

Vitria, J., Bressan, M., & Radeva, P. (2007). Bayesian classification of cork stoppers using class-conditional independent component analysis. *IEE Transactions on Systems, Man and Cybernetics, C37*, 32–38.

Wong, T. T., & Hsu, C. H. (2008). Two-stage classification methods for microarray data. *Expert Systems with Applications, 34*, 375–383.

Yu, S. N., & Chou, K. T. (2008a). Selection of significant independent components for ECG beat classification. *Expert Systems with Applications*, *36*, 2088–2096.

Yu, S. N., & Chou, K. T. (2007). A switchable scheme for ECG beat classification based on independent component analysis. *Expert Systems with Applications, 33*, 824–829.

Yu, S. N., & Chou, K. T. (2008b). Integration of independent component analysis and neural networks for ECG beat classification. *Expert Systems with Applications, 34*, 2841–2846.

Zheng, C. H., Huang, D. S., & Shang, L. (2006). Feature selection in independent component subspace for microarray data classification. *Neurocomputing, 69*, 2407–2410.