Available online at www.sciencedirect.com

ScienceDirect

Progress in Natural Science 19 (2009) 1365–1371

Progress in Natural Science

www.elsevier.com/locate/pnsc

Prominent feature selection of microarray data

Yihui Liu

School of Computer Science and Information Technology, Shandong Institute of Light Industry, Jinan 250353, China

Received 14 June 2008; received in revised form 15 January 2009; accepted 16 January 2009

Abstract

For wavelet transform, a set of orthogonal wavelet basis aims to detect the localized changing features contained in microarray data. In this research, we investigate the performance of the selected wavelet features based on wavelet detail coefficients at the second level and the third level. The genetic algorithm is performed to optimize wavelet detail coefficients to select the best discriminant features. Experiments are carried out on four microarray datasets to evaluate the performance of classification. Experimental results prove that wavelet features optimized from detail coefficients efficiently characterize the differences between normal tissues and cancer tissues.

Keywords: Microarray data; Wavelet detail coefficients; Feature extraction; Feature selection

1. Introduction

The recent development of the microarray technique is rapidly accelerating progress in many areas of biomedical research. Statistical scoring of features has led to the discovery of new genes associated with cancer and other diseases. Explicitly, the advent of these technologies will revolutionize biology and medicine, but their full utilization will depend heavily on accurate data processing and analysis techniques, and the central role of data analysis will become even more critical in the future [1].

Approaches to classify the microarray data usually use the feature selection method and the feature extraction method. Feature selection methods concern a criterion relating to the correlation degree to rank and select key genes, such as the signal-to-noise ratio (SNR) method [2], the partial least squares method [3], the Pearson correlation coefficient method [4], the *t*-test statistic method [5], and the decision tree combined with the Bag-Boosting scheme [6–8]. Feature extraction methods are based on a new transform space of DNA microarray data, such as independent component analysis [9] and wavelet feature-based methods [10–12]. For transformations, a set of new basis is normally chosen for the data. The selection of the new basis determines the properties that will be held by the transformed data. Principal component analysis (PCA) maximizes the total scatter across all classes. So not only the between-class scatter but also the within-class scatter is maximized. However, maximizing within-class scatter is unwanted information. Linear discriminant analysis (LDA) is used to extract discriminant information from microarray data by maximizing between-class variations and minimizing within-class variations. Instead of transforming uncorrelated components, like PCA and LDA, independent component analysis (ICA) attempts to achieve statistically independent components in microarray data; it is sensitive to high-order statistics in the data, not just the covariance matrix [9]. PCA, LDA and ICA do not detect the localized features of microarray data. For wavelet transform, a set of orthogonal wavelet basis aims to detect the localized features contained in microarray data. In our papers [11–13], we investigated the performance of approximation coefficients. Approximation coefficients reduce the dimensionality of microarray data and are “compress version” of the microarray vector. The key gene information is

E-mail address: yihui_liu_2005@yahoo.co.uk

1002-0071/$ - see front matter © 2009 National Natural Science Foundation of China and Chinese Academy of Sciences. Published by Elsevier Limited and Science in China Press. All rights reserved.
doi:10.1016/j.pnsc.2009.01.014

investigated based on optimized wavelet features from reconstructed approximation using the genetic algorithm (GA) [14]. Though approximation coefficients of low frequency characterize the major part of microarray vectors, detail coefficients of high frequency represent the changing information of microarray data, which is ignored by approximation coefficients of low frequency. Detail coefficients measure the differences between cancer tissues and normal tissues using the orthogonal wavelet coefficient basis. In our research paper [15], we investigated the classification performance of detail coefficients at different levels. Experimental results showed that detail coefficients at the second and the third levels achieve good and robust performance. In this study, we carried out our research by using the genetic algorithm to select the best discriminant features from detail coefficients.

First, multilevel wavelet decomposition is performed to break the gene profile into approximations and details. Approximation coefficients compress gene profiles and detail coefficients detect the change points of gene profiles. We extracted detail coefficients at the second and the third level to measure the differences between normal tissues and cancer tissues and reduced dimensionality. Then, the genetic algorithm is performed to select the best features from detail coefficients at the second and the third level, respectively. The optimized features further reduce the dimensionality of the microarray vector.

2. Wavelet analysis

The wavelet transform has nice features of space–frequency localization and multiresolutions. Wavelet technology is applied widely in many research areas, and the major reasons are its complete theoretical framework and low computational complexity. The wavelet-transform method, proposed by Grossmann and Morlet [16], analyzes a signal by transforming its input time domain into a time–frequency domain. The continuous wavelet transform of the one-dimensional signal is defined as follows:

$$W(a,b) = \int_R s(t)\frac{1}{\sqrt{a}}\Psi\left(\frac{t-b}{a}\right)\mathrm{d}t$$
$$a \in R^+ - \{0\}, b \in R$$

where *s, a, b* and Ψ denote signal, scale, position, and wavelet function parameters, respectively. The continuous wavelet transform (CWT) means continuous in terms of scaling and shifting. The analyzing wavelet at every scale is shifted smoothly over the full domain of the analyzed signal. The CWT is a redundant representation because it uses much more scale and position information than the non-redundant discrete scale-position representation, which has a set of orthonormal basis. The discrete wavelet transform (DWT) can be discretized by restraining a and b to a discrete lattice, i.e. $a = 2^n, b \in Z$. The wavelet filter-banks approach was developed by Mallat [17]. The wavelet analysis involves two compounds: approximations and details. For one-dimensional wavelet decomposition, starting from the original microarray vector s, the first step produces two sets of coefficients: approximation coefficients (scaling coefficients) c_1 and detail coefficients (wavelet coefficients) d_1. These coefficients are computed by convolving the original microarray vector with the low-pass filter for approximation and with the high-pass filter for detail. The convolved coefficients are downsampled by keeping the even indexed elements. Then, the approximation coefficients c_1 are split into two parts by using the same algorithm and replaced by c_2 and d_2, and so on. This decomposition process is repeated until the required level is reached. The coefficient vectors are produced by downsampling and are only half the length of the signal or the coefficient vector at the previous level.

Fig. 1 shows the wavelet decomposition tree at three levels. Fig. 2 shows wavelet detail coefficients of the sample vector of prostate cancer at four levels. We have 6306, 3159, 1586, and 799 detail coefficients for prostate cancer data from the first level to the fourth level, respectively. After wavelet decomposition is performed on each sample vector, wavelet vectors of 3159 dimensions at the second level or 1586 dimensions at the third level are extracted to represent the original microarray vector of 12,600 dimensions. For the prostate cancer dataset, we obtain 100 × 1586 matrix of wavelet feature vectors for training samples and 34 × 1586 matrix of wavelet feature vectors for test samples. The feature matrix dramatically reduces computation load. Then, the genetic algorithm is performed to further select the best discriminant features from extracted wavelet detail coefficients, combining with linear discriminant analysis [14]. Fig. 3 shows the classification process of microarray data based on optimized GA features.

3. Experiments and results

Four microarray datasets are used to evaluate the performance of the proposed method. We did the preprocessing on microarray profiles by filtering out genes with 0 profile variance over time. After filtering, we extracted wavelet detail coefficients from the filtered data to measure the differences between cancer tissues and normal tissues. When the decomposition level is higher, the wavelet coeffi-

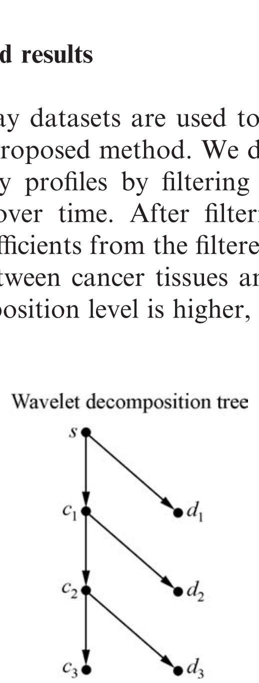

Fig. 1. Wavelet decomposition tree at three levels. Symbol s represents microarray profiles; c_1, c_2 and c_3 represent approximation coefficients at three levels; d_1, d_2 and d_3 represent detail coefficients at three levels.

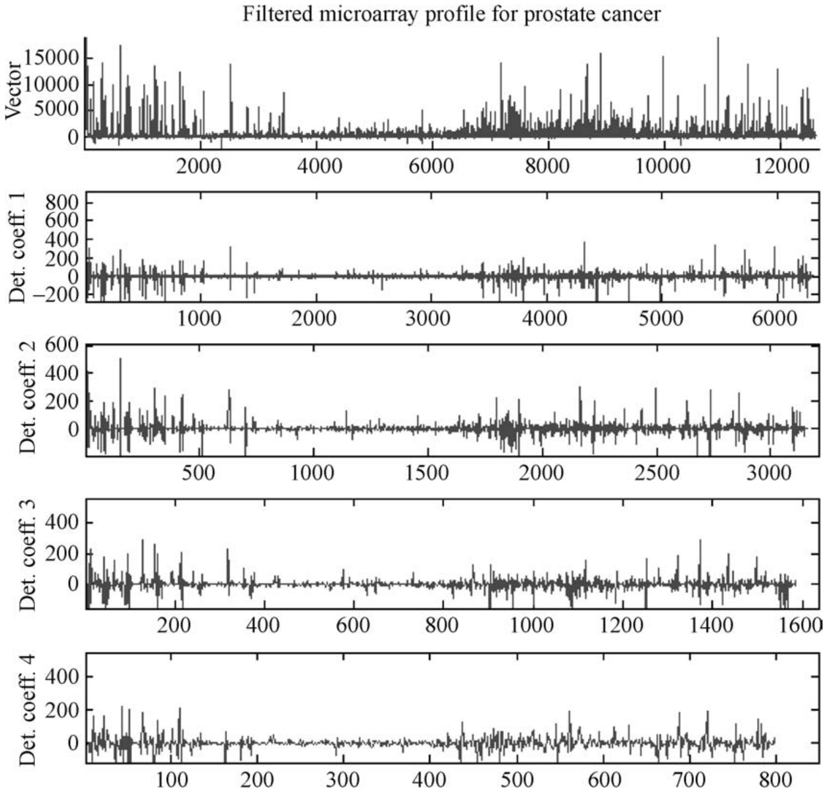

Fig. 2. Wavelet detail coefficients of sample vector of prostate cancer data at four levels.

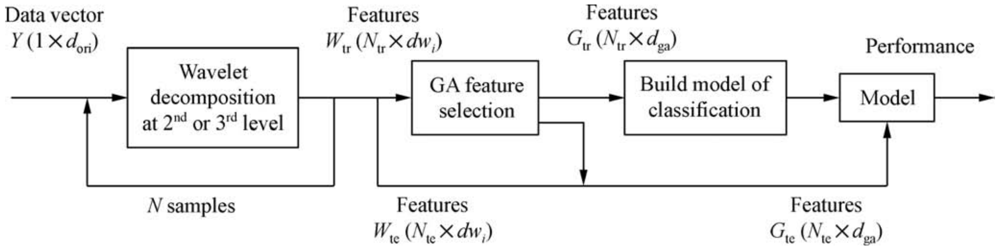
Data vector
Y ($1 \times d_{\text{ori}}$)
Wavelet decomposition at 2nd or 3rd level
Features
W_{tr} ($N_{\text{tr}} \times dw_i$)
GA feature selection
Features
G_{tr} ($N_{\text{tr}} \times d_{\text{ga}}$)
Build model of classification
Performance
Model
N samples
Features
W_{te} ($N_{\text{te}} \times dw_i$)
Features
G_{te} ($N_{\text{te}} \times d_{\text{ga}}$)

Fig. 3. The process of GA feature selection and classification. Y, W_{tr}, W_{te}, G_{tr}, and G_{te} represent the original microarray vector, wavelet feature vectors for training and test samples, and selected GA feature vectors for training and test samples. N, N_{tr}, N_{te}, d_{ori}, dw_i, and d_{ga} represent the number of whole samples, training samples and test samples, dimension number of the original data vector, dimension number of wavelet detail coefficients at the second or the third level, and the number of selected GA features, respectively.

cients represent larger changes of gene profiles based on higher derivatives of microarray data, and the small changes of gene profiles are ignored. First, detail coefficients at the second and the third level are selected to reduce dimensionality, respectively. Then, the genetic algorithm is implemented to optimize the wavelet detail coefficients. The optimized wavelet features are used to evaluate the performance of classification. Daubechies basis 7 [18], which has seven non-zero coefficients of the compact support wavelet orthogonal basis, is performed for wavelet analysis of DNA microarray data.

3.1. Prostate cancer

Prostate cancer data [19] contain a training set of 52 prostate tumor samples and 50 non-tumor samples labeled as “Normal” with 12,600 genes. An independent set of test samples has 25 tumor and 9 normal samples from a different experiment.

After filtering the microarray vector with 0 profile variance, vectors of 12,599 dimensions are obtained. There are 3159 and 1586 detail coefficients obtained based on wavelet decomposition at the second and the third level, respectively. Fig. 4 shows six optimized features from 3159 detail coefficients at the second level. Fig. 5 shows six optimized detail coefficients for test samples of the prostate cancer dataset. A 97.06% performance is obtained based on six selected detail coefficient features. When 11 optimized features are selected from 1586 detail coefficients at the third level a 100% recognition rate is achieved. Fig. 6 shows the 11 optimized features are selected from detail coefficients at the third level. Fig. 7 shows 11 selected

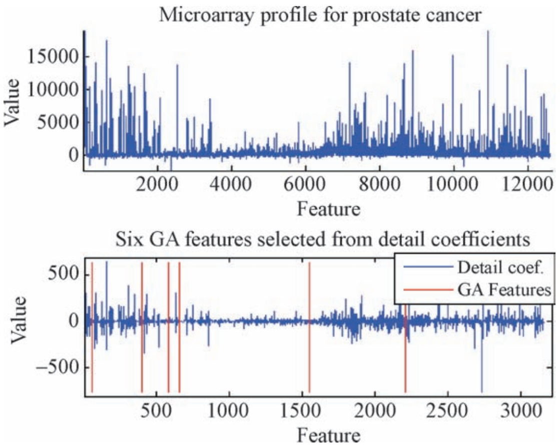

Fig. 4. Detail coefficients at the second level and six selected GA features for the prostate cancer dataset.

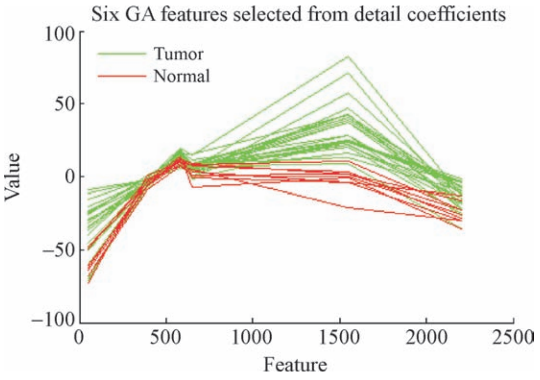

Fig. 5. A 97.06% performance of six GA features selected from detail coefficients at the second level for test samples of the prostate cancer dataset.

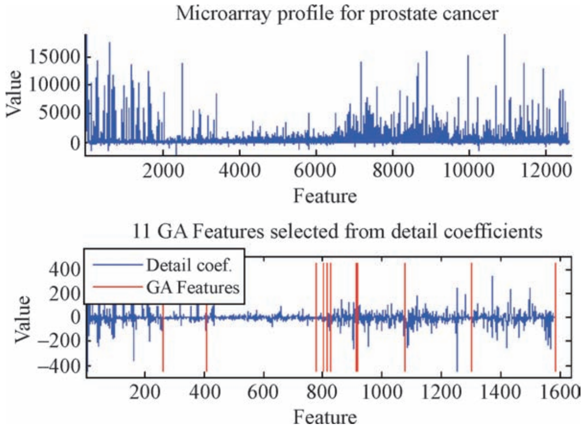

Fig. 6. Detail coefficients at the third level and 11 selected GA features.

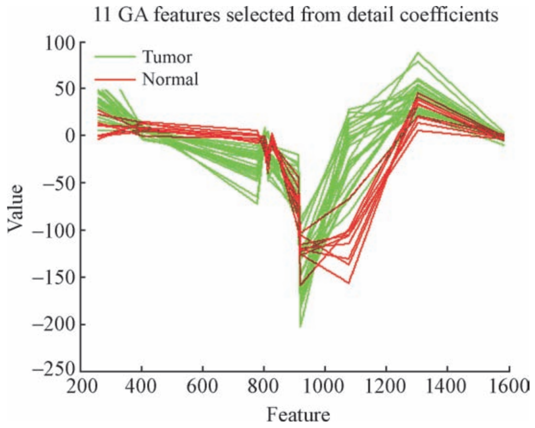

Fig. 7. A 100% performance of 11 GA features selected from detail coefficients at the third level for test samples of the prostate cancer dataset.

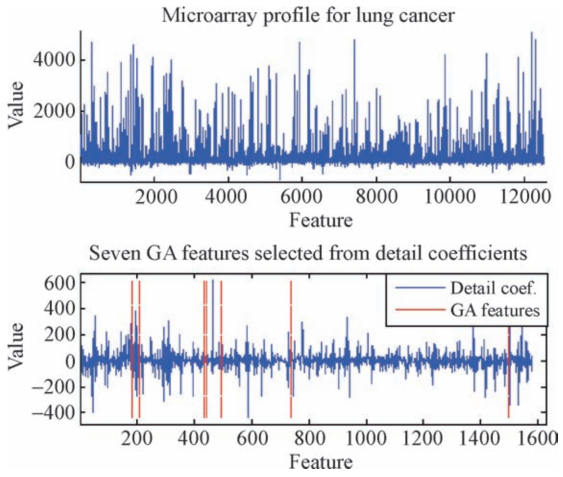

Fig. 8. Detail coefficients at the third level and seven selected GA features.

Table 1
The comparison results of different algorithms for the prostate cancer dataset.

Methods	Accuracy (%)
Eleven GA features (detail coefficients at the third level)	100
Six GA features (detail coefficients at the second level)	97.06
Detail coefficients (the second level and the third level) [15]	97.06
Seven GA features [14] (reconstructed approximation)	97.06
SingleC4.5, BaggingC4.5, AdaBoostC4.5 [8]	67.65, 75.53, 67.65

detail coefficients for test samples, and the differences between cancer tissues and normal tissues are illustrated clearly. In the paper of Tan et al. [8], SingleC4.5, BaggingC4.5 and AdaBoostC4.5 methods achieved 67.65%, 75.53%, and 67.65% accuracies, respectively, which are inferior to our method. A 97.06% performance [15] is achieved based on detail coefficients at the second and the third level, respectively. In the paper [14], seven key GA features optimized from reconstructed approximation at the second level obtain 97.06% performance. Table 1 shows the comparison results of different algorithms.

3.2. Lung cancer

Lung cancer data [20] contain two kinds of tissues, which are pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung. There are 32 training samples, including 16 MPM and 16 ADCA. There are 149 test samples, including 15 MPM and 134 ADCA samples. The number of genes of each sample is 12,533.

When filtering the microarray vector with 0 profile variance, the vector dimension changed to 12,532. After wavelet decomposition at the second and the third level is performed, 3142 and 1577 detail coefficients are extracted, respectively. After the genetic algorithm is implemented, when six optimization features are selected from detail coefficients at the second level, 97.99% accuracy is achieved. A 98.66% accuracy is achieved when seven optimized features are selected from detail coefficients at the third level. Fig. 8 shows seven selected GA features from detail coefficients at the third level. Fig. 9 shows seven selected features for test samples, and the difference between malignant pleural mesothelioma and adenocarcinoma samples is very clear. Table 2 shows the comparison results of different algorithms. Tan et al. [8] gave 92.62%, 93.29%, and 92.62% accuracies of SingleC4.5, BaggingC4.5 and AdaBoostC4.5 methods, which are lower than our 98.66% performance. The 98.66% performance is also better than 97.99% of Li's method and 97.99% of 20 GA features selected from reconstructed approximation. This is because the details at the third level reflect the changing information of microarray data and reveal the high-order information hidden in the microarray profile.

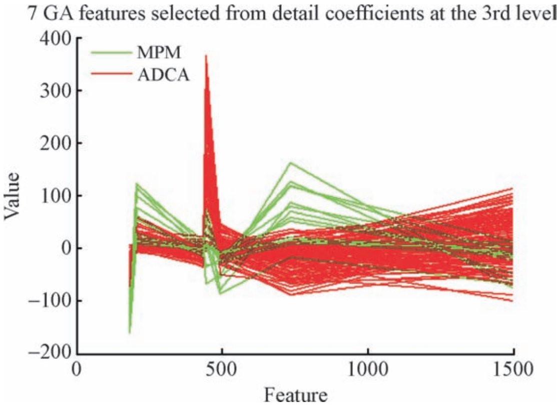

Fig. 9. A 98.66% performance of selected GA features from detail coefficients at the third level. It is for test samples of the lung cancer dataset.

Table 2
The comparison results of different algorithms for the lung cancer dataset.

Methods	Accuracy (%)
Seven GA features (detail coefficients at the third level)	98.66
Six GA features (detail coefficients at the second level)	97.99
Twenty GA features [14] (reconstructed approximation)	97.99
SingleC4.5, BaggingC4.5, AdaBoostC4.5 [8]	92.62, 93.29, 92.62
Li's method [7]	97.99

3.3. MLL (ALL vs MLL vs AML)

Leukemia data [21] contain 57 training leukemia samples, including 20 acute lymphoblastic leukemia (ALL), 17 mixed-lineage leukemia (MLL), and 20 acute myeloid leukemia (AML) samples. The test dataset contains four ALL, three MLL and eight AML samples. The number of attributes is 12,582.

When filtering the microarray vector with 0 profile variance, the dimensionality of the vector does not change and remains at 12,582. After wavelet decomposition at the second and the third level, 3155 and 1584 detail coefficients are extracted to measure the differences of different tissue samples, respectively. When five and six GA features are optimized based on the training matrix 57 × 3155 at the second level and 57 × 1584 at the third level, respectively, a 100% correct rate is achieved by using the property of changing points characterized by wavelet detail coefficients

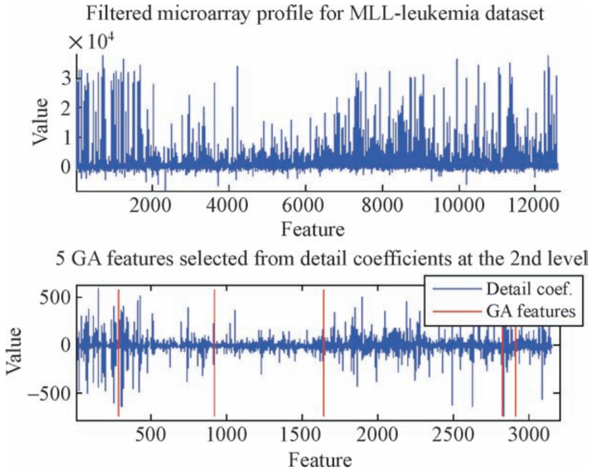

Fig. 10. Detail coefficients at the second level and five selected GA features.

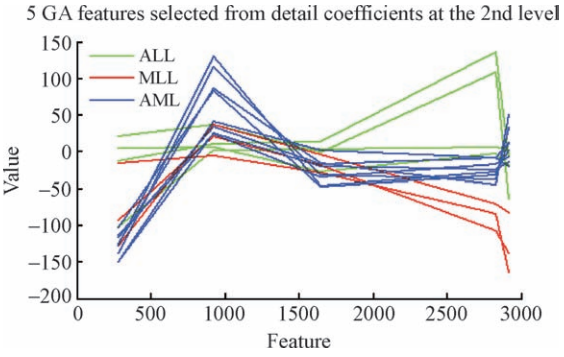

Fig. 11. A 100% performance of five selected GA features from detail coefficients at the second level for test samples of the MLL cancer dataset.

Table 3
The comparison results of different algorithms for the MLL dataset.

Methods	Accuracy (%)
Six GA features (detail coefficients at the third level)	100
Five GA features (detail coefficients at the second level)	100
Detail coefficients (the second level and the third level) [15]	100
Seven GA features [14] (reconstructed approximation)	100
Li's method, C4.5, Bagging, Boosting [7]	100, 73.33, 86.67, 100

of high frequency. Fig. 10 shows five GA features selected from detail coefficients at the second level. Fig. 11 shows five GA features selected from detail coefficients at the second level for test samples. Table 3 shows the comparison results of different algorithms for the MLL cancer dataset. We have the same performance as Li's method, the boosting method, and a better performance than C4.5, Bagging methods [7]. Detail coefficients at the second and the third level also achieve 100% performance [15]. Fourteen GA features optimized from reconstructed approximation of low frequency obtain 100% accuracy.

3.4. Leukemia (ALL vs AML)

The training dataset consists of 38 bone marrow samples, including 27 ALL and 11 AML samples with 7129 attributes from 6817 human genes. The test dataset has 34 samples including 20 ALL and 14 AML [2].

After filtering, 7129 dimensions of the vector change to 7128. Only one dimension is reduced. There are 1791 and 902 detail coefficients extracted after wavelet decomposition at the second and the third level, respectively. When two GA features are selected from detail coefficients at the second level, 97.06% accuracy is achieved. Fig. 12 shows two GA features are selected from detail coefficients at the second level for test samples of the leukemia dataset. Five GA features optimized from detail coefficients achieve 97.06% performance. Fig. 13 shows five GA features selected from detail coefficients at the third level for test samples. Table 4 shows the comparison results of different algorithms for the leukemia dataset. Our performance is the same as 97.06% of the Bayesian variable method [22], better than 82.3% of the PCA disjoint models [23], 88.2% of between-group analysis [24], 91.18% of SingleC4.5, BaggingC4.5 and AdaBoostC4.5 methods, and 88.24% of Li's method [7]. Four GA features selected from reconstructed wavelet approximation of low frequency achieve 97.06% performance [14]. Detail coefficients at the second and the third level achieve 100% performance using the support vector machine classifier [15] .

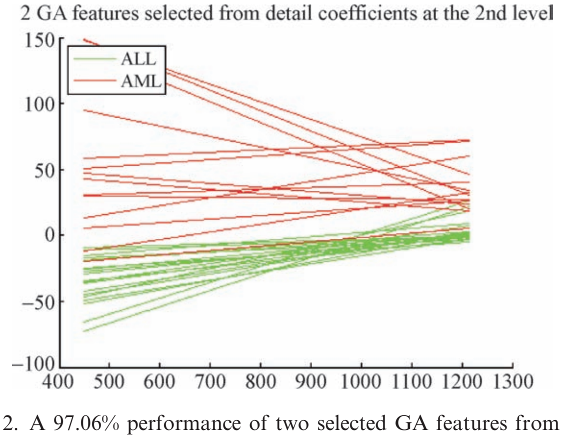

Fig. 12. A 97.06% performance of two selected GA features from detail coefficients at the second level for test samples of the leukemia cancer dataset.

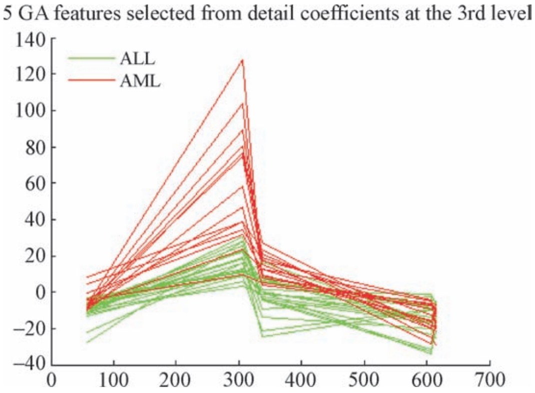

Fig. 13. A 97.06% performance of five selected GA features from detail coefficients at the third level for test samples of the leukemia cancer dataset.

4. Discussion and conclusions

In this paper, we propose a hybrid method combining features extract and feature selection methods for microarray data analysis. First, wavelet decomposition at the second and the third level is performed on the microarray vector to extract the high frequency features of wavelet detail coefficients, which takes full advantage of different order statistical information of the microarray vector. High-frequency orthogonal detail coefficients at the second level and the third level reduce the dimensionality of the

Table 4
The comparison results of different algorithms for the leukemia dataset.

Methods	Accuracy (%)
Five GA features (detail coefficients at the third level)	97.06
Two GA features (detail coefficients at the second level)	97.06
Detail coefficients (the second level and the third level) [15]	100
Four GA features [14] (reconstructed approximation)	97.06
Li's method, C4.5, Bagging, Boosting [7]	88.24, 91.18, 91.18, 91.18
Bayesian variable method [22]	97.06
PCA disjoint models [23]	82.3
Between-group analysis [24]	88.2

microarray vector and characterize major changing information and ignore "small change" of the microarray vector. The dimension number of 12,600, 12,533, 12,582, and 7129 for the prostate cancer dataset, lung cancer dataset, MLL dataset, and leukemia dataset has been reduced to 1586, 1577, 1584, and 902 detail coefficients at the third level. This dramatically reduces the computation load for the next calculation of feature selection based on the genetic algorithm. In order to select the best discriminant features of detail coefficients, the genetic algorithm is performed on the detail coefficients to select optimized features. Experiments were carried out on four independent datasets, and optimized GA features achieve competitive performance compared to other feature extraction and features selection methods.

The proposed method uses the orthogonal wavelet basis to represent the microarray vector and is not dependent on the training dataset, not involved in large matrix computation, such as other feature extraction methods of PCA, LDA, and ICA. When we use Matlab software to run wavelet decomposition on 136 samples of prostate cancer with 12,600 dimensions and use the WINDOWS operate system, Intel 2.1 GHz CPU, and 1G RAM, it only takes about 3 s. Experimental results prove that the wavelet feature extraction method is robust and feasible. In our future work, we will focus on the gene information selection in original data space based on wavelet analysis.

Acknowledgements

This work was supported by SRF for ROCS, SEM, and the Natural Science Foundation of Shandong Province (Y2008G30), China.

References

[1] Campbell C. New analytical techniques for the interpretation of microarray data. Bioinformatics 2003;19:1045.

[2] Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531–7.

[3] Danh VN, David MR. Tumor classification by partial least squares using microarray gene expression data. Bioinformatics 2002;18: 39–50.

[4] Xiong MJ, Li MW, Boerwinkle E. Computational methods for gene expression-based tumor classification. Biotechniques 2000;29: 1264–70.

[5] Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. Bionformatics 2001;17:509–19.

[6] Dettling M. BagBoosting for tumor classification with gene expression data. Bioinformatics 2004;20:3583–93.

[7] Li J, Liu H, Ng SK, et al. Discovery of significant rules for classifying cancer diagnosis data. Bioinformatics 2003;19:93–102.

[8] Tan AC, Gilbert D. Ensemble machine learning on gene expression data for cancer classification. Appl Bioinform 2003;2:S75–83.

[9] Huang DS, Zheng CH. Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. Bioinformatics 2006;22:1855–62.

[10] Zhou X, Wang X, Dougherty ER. Nonlinear-probit gene classification using mutual-information and wavelet based feature selection. Biol Syst 2004;12:371–86.

[11] Liu Y. Wavelet feature selection for microarray data. IEEE/NIH on life science systems and applications workshop, 2007. LISA; 2007. pp. 205–8.

[12] Liu Y. Feature extraction for DNA microarray data. In: Twentieth IEEE international symposium on computer-based medical systems. CBMS; 2007. pp. 371–6.

[13] Liu Y, Shen J, Cheng J. Cancer classification based on the "Fingerprint" of microarray data. In: Proceedings of the first international conference on bioinformatics and biomedical engineering; 2007. pp. 180–3.

[14] Liu Y. Detect key genes information in classification of microarray data. EURASIP J Adv Signal Process 2008. doi:10.1155/2008/612397.

[15] Liu Y. Wavelet feature extraction for high-dimensional microarray data. Neurocomputing 2009;72:985–90.

[16] Grossmann A, Morlet J. Decomposition of Hardy functions into square integrable wavelets of constant shape SIAM. J Math Anal 1984;15:723–36.

[17] Mallat S. A theory for multiresolution signal decomposition: The wavelet representation. IEEE Trans Pattern Anal Mach Intell 1989;11:674–93.

[18] Daubechies I. Orthonormal bases of compactly supported wavelets. Commun Pure Appl Math 1988;41:909–96.

[19] Singh D, Febbol PG, Ross K, et al. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 2002;1:203–9.

[20] Gordon GJ, Jensen RV, Hsiao LL, et al. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Res 2002;62:4963–7.

[21] Armstrong SA, Staunton JE, Silverman LB, et al. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat Genet 2002;30:41–7.

[22] Kyeong EL, Sha N, Dougherty ER, et al. Gene selection: a Bayesian variable selection approach. Bioinformatics 2003;19:90–7.

[23] Bicciato S, Luchini A, Di Bello C. PCA disjoint models for multiclass cancer analysis using gene expression data. Bioinformatics 2003;19: 571–8.

[24] Aedin CC, Guy P, Elizabeth CC, et al. Between group analysis of microarray data. Bioinformatics 2002;18:1600–8.