Wavelet Feature Selection for Microarray Data

Yihui Liu
School of Computer Science and Information Technology,
Shandong Institute of Light Industry, Jinan, Shandong,
China, 250353
Yihui_liu_2005@yahoo.co.uk

***Abstract*—A hybrid method of feature selection based on wavelet analysis and genetic algorithm (GA) is proposed in this study for high dimensional microarray data. A set of orthogonal wavelet approximation coefficients based on wavelet decomposition are extracted to compress the gene profiles and reduce the dimensionality of microarray data. Then genetic algorithm is performed to select the optimized features from approximation coefficients. Linear discriminant analysis (LDA) is employed to evaluate the classification performance. Experiments are performed on four datasets. Our results show that this hybrid method is efficient and robust.**

I. INTRODUCTION

Recently, huge advances in DNA microarray have allowed the scientist to test thousands of genes in normal or tumor tissues on a single array and check whether those genes are active, hyperactive or silent. Therefore, there is an increasing interest in changing the criterion of tumor classification from morphologic to molecular. Generally speaking, approaches usually use a criterion relating to the correlation degree to rank and select key genes, such as signal-to-noise ratio (SNR) method [1], the partial least squares method [2], Pearson correlation coefficient method [3] and t -test statistic method [4]. Independent component analysis [5] also is used in the analysis of DNA microarray data. To equip the system with the optimum combination of classifier, gene selection and cross-validation methods, researchers perform a systematic and comprehensive evaluation of several major algorithms [6]. A very promising solution to combine the two ensemble schemes bagging and boosting, called BagBoosting, is proposed in the paper [7]. The predictive potential is confirmed by comparing BagBoosting to several established class prediction tools for microarray data. The researchers [8] discover many diversified and significant rules from high dimensional profiling data and propose to aggregate the discriminating power of these rules for reliable predictions. The discovered rules are found to contain low-ranked features; these features are found to be sometimes necessary for classifiers to achieve perfect accuracy. The researcher [9] focus on three different supervised machine learning techniques in cancer classification, namely C4.5 decision tree, and bagged and boosted decision trees. They have performed classification tasks on seven publicly available cancerous microarray data and compared the classification/prediction performance of these methods. They have observed that ensemble learning (bagged and boosted decision trees) often performs better than single decision trees in this classification task. Researchers [10] propose using a mutual information-based feature selection method where features are wavelet-based. They select Daubechies bases 4 which has four non-zero coefficients of the compact support wavelet orthogonal basis. They use approximation coefficients and wavelet coefficients to perform mutual information-based feature selection. In this research we only use wavelet approximation coefficients to compress gene profiles and reduce dimensionality. Approximation coefficients act as “fingerprint” of gene profiles and characterize the essential information contained in gene profiles. We perform wavelet decomposition at 2nd level on gene profiles, in order to reduce dimensionality and not to lose too much information. Genetic algorithm is performed to select the optimized features from approximation coefficients. Experiments are carried out on four datasets and experimental results show that our method is efficient and robust.

II. WAVELET ANALYSIS

For wavelet analysis for gene expression data, a gene expression profile can be represented as a sum of wavelets at different time shifts and scales using discrete wavelet analysis (DWT). The DWT is capable of extracting the local features by separating the components of gene expression profiles in both time and scale. According to DWT, a time-varying function $f(t) \in L^2(R)$ can be expressed in terms of $\phi(t)$ and $\psi(t)$ as follows:

$$f(t) = \sum_k c_0(k)\phi(t-k) + \sum_k \sum_{j=1} d_j(k) 2^{\frac{-j}{2}} \psi(2^{-j}t-k)$$

$$= \sum_k c_{j0}(k) 2^{\frac{-j0}{2}} \phi(2^{-j0}t-k) + \sum_k \sum_{j=j0} d_j(k) 2^{\frac{-j}{2}} \psi(2^{-j}t-k)$$

where $\phi(t), \psi(t), c_0$, and d_j represent the scaling function, wavelet function, scaling coefficients (approximation coefficients) at scale 0, and detail coefficients at scale j, respectively. The variable k is the translation coefficient for the localization of gene expression data. The scales denote the

* This study was supported by Shandong Institute of Light Industry under Grant12041653.

978-1-4244-1813-8/08/$25.00 © 2008 IEEE

different (low to high) scale bands.

The wavelet transform breaks the microarray vector into approximations and details at different levels. Figure 1 shows wavelet decomposition tree. Approximation coefficients compress microarray data; and detail coefficients characterize the changes of microarray data based on wavelet basis. Normally noise hidden in microarray data is contained in details at first levels of decomposition. When the decomposition level is getting higher, approximations lose more high frequency information, which is contained in details.

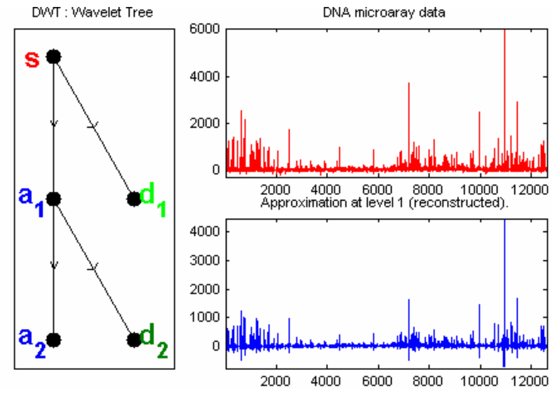

Fig. 1. Wavelet decomposition tree.This figure shows wavelet decomposition tree and approximation at first level.

III. Genetic Algorithm

The Genetic Algorithm (GA) is an evolutionary computing technique that can be used to solve problems efficiently for which there are many possible solutions [11]. In our research we perform GA on wavelet features to select the best discriminant features and reduce dimensionality of wavelet feature space.

The binary coding is designed in genetic algorithm. The Genome length is set to the number of wavelet features. We select different number of features in our study respectively to evaluate the performance of classification. Firstly the algorithm creates initial population by ranking key features based on a two-way T-test with pooled variance estimate. The algorithm then creates a sequence of new populations. At each step, the algorithm uses the individuals in the current generation to create the next population. To create the new population, the algorithm performs the following steps:

a. Each member of the current population is scored by computing its fitness value. The algorithm usually selects individuals that have better fitness values as parents. A fitness function acts as selective pressure on all of the data points. This function determines which data points get passed on to or removed from each subsequent generation. To apply a genetic algorithm on the microarray data, we use a linear discriminant classifier as fitness function to evaluate how well the data gets classified. Linear discriminate analysis (LDA) is a popular discriminant criterion, which is used to find a linear projection of the original vectors from a high-dimensional space to an optimal low-dimensional subspace in which the ratio of the between-class scatter and the within-class scatter is maximized [12]

b. The genetic algorithm creates three types of children for the next generation:

Elite children, that are the individuals in the current generation with the best fitness values, automatically survive to the next generation. In this research two elite children are selected.

Crossover children are created by combining the vectors of a pair of parents. The scattered crossover function is used in this study, which randomly selects a gene at the same coordinate from one of the two parents and assigns it to the child. The crossover fraction, which specifies the fraction of each population other than elite children, is set to 0.8.

The mutation algorithm creates mutation children by randomly changing the genes of individual parents. In this study the algorithm adds a random vector from a Gaussian distribution to the parent.

c. The algorithm stops when one of the stopping criteria is met. GA uses four different criteria to determine when to stop the solver. GA stops when the maximum number of generations is reached; the maximum number of generations is set to 70 in this research. Fitness limit is considered and the algorithm stops if the best fitness value is less than or equal to the value of fitness limit. GA also detects if there is no change in the best fitness value for some time given in seconds (stall time limit=20), or for some number of generations (stall generation limit=50).

IV. Results

In this study we use correct rate, sensitivity, specificity, PPV and NPV to evaluate the performance. Sensitivity is defined as $\frac{TP}{TP+FN}$; Specificity is defined as $\frac{TN}{TN+FP}$; PPV (Positive Predictive Value) is defined as $\frac{TP}{TP+FP}$; NPV (Negative Predictive Value) is defined as $\frac{TN}{TN+FN}$; Correct rate is defined as $\frac{TP+TN}{TP+TN+FP+FN}$. Where TP, TN, FP and FN stand for the number of true positive (cancer), true negative (control), false positive and false negative samples. Firstly, we do the preprocessing on microarray profiles by filtering out genes with small variance over time. The microarray data has high dimensionality and a lot of information corresponds to genes that do not show any key changes during the experiment. To make it easier to find the significant genes, we reduce the size of the data set by removing genes with 0 profile variance. After filtering, approximation coefficients at second level is selected to characterize the features of gene profiles, reducing the dimensionality and not losing too much small information contained in gene profiles. Here we use Daubechies basis 7(db7)[13], which has seven non-zero coefficients of the compact support wavelet orthogonal basis, for wavelet analysis of DNA microarray data.

Leukemia (ALL v.s. AML)

Training dataset consists of 38 bone marrow samples (27 ALL and 11 AML) with 7129 attributes from 6817 human genes, and 34 testing samples include 20 ALL and 14 AML[1].

After wavelet decomposition at 2nd level is performed on gene profile, we obtain 1791 approximation coefficients. The original training matrix is 38x7129, now it is 38x1791 using wavelet features at 2nd level. After genetic algorithm is used to select the optimized features from approximation coefficients, 8 GA features and 15 GA features are selected based on 38x1791 approximation coefficients. For 8 GA feature, the training matrix is 38x8 and the testing matrix is 34x8. 97.06% correct rate is achieved based on linear discriminate analysis (LDA). Table 1 shows the performance of selected GA features. For 8 GA features, the population size is set to 223 in order to search full range of wavelet features. For 15 GA features, the population size is set to 119. Figure 2 shows wavelet features and selected GA features of the performance, which is shown in Table 1. Our result is the same as the Bayesian variable method [14]. It is better than 82.3% of the PCA disjoint models [15] and 88.2% of the between-group analysis [16]. Also Table 2 shows SingleC4.5, BaggingC4.5 and AdaBoostC4.5 methods achieve 91.18% accuracy, which is inferior to our method.

TABLE I
PERFORMANCE OF SELECTED GA FEATURES

GA FN	CR	SE	SP	PPV	NPV
8	0.9706	0.9286	1.0000	1.0000	0.9524
15	0.9706	0.9286	1.0000	1.0000	0.9524

This Table shows performance of Leukemia (ALL v.s. AML) data set. PPV stands for Positive Predictive Value; NPV stands for Negative Predictive Value. PPV stands for Positive Predictive Value; NPV stands for Negative Predictive Value. FN represents Feature Number. SE and SP represent Sensitivity and Specificity.

TABLE II
PREDICTIVE ACCURACY OF THE CLASSIFIERS [9]

Dataset	Accuracy		
	Single C4.5	Baggin g C4.5	AdaBoost C4.5
Leukemia	91.18	91.18	91.18
Lung cancer	92.62	93.29	92.62
Prostate cancer	67.65	73.53	67.65

MLL-Leukemia (ALL v.s. MLL v.s. AML)

Leukemia data [17] contains 57 training leukemia samples (20 ALL, 17 MLL and 20 AML). Testing data contains 4 ALL, 3 MLL and 8 AML samples. The number of attributes is 12582.

Approximation coefficient vector of 3155 dimensions is obtained based on decomposition at 2nd level. When 13 GA features and 18 GA features are selected, 100% correct rate is achieved. For 15 GA features, the population size is set to 210 in order to search full range of wavelet features. For 18 GA features, the population size is set to 175. We have the same performance with Li's method [8], boosting method, and better than C4.5, Bagging methods, which are shown in Table 4.

TABLE III
PERFORMANCE OF SELECTED GA FEATURES

GA FN	CR	SE	SP	PPV	NPV
13	1.0000	1.0000	1.0000	1.0000	1.0000
18	1.0000	1.0000	1.0000	1.0000	1.0000

Performance for MLL-Leukemia (ALL v.s. MLL v.s. AML) dataset.

TABLE IV
THE TEST ERROR NUMBERS BY FOUR MODELS [8]

		Test error numbers		
Dataset	Li's method	C4.5	Bagging	Boosting
MLL Leukemia	0	4(2:2:0)	2(1:1:0)	0

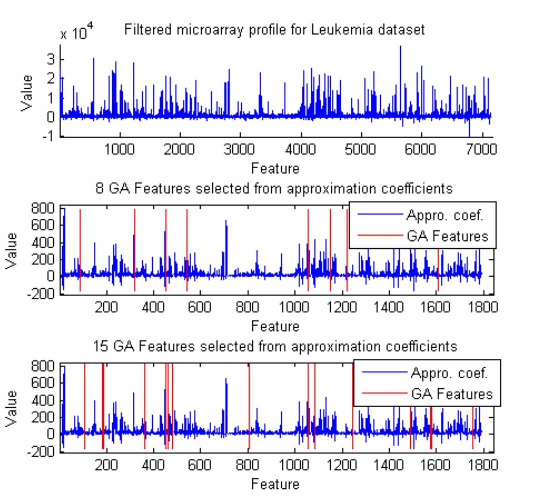

Fig. 2. Wavelet features and selected GA features. This Figure shows 8 GA features selected from approximation coefficients at 2nd level and 15 GA features selected from approximation coefficients at 2nd level for Leukemia dataset.

Prostate cancer

Prostate cancer data [18] contains training set of 52 prostate tumor samples and 50 non-tumor (labelled as "Normal") prostate samples with 12600 genes. An independent set of testing samples is also prepared, which is from a different experiment. The testing set has 25 tumor and 9 normal samples.

After wavelet decomposition at 2nd level, we extract 3159 approximation coefficients. We obtain 97.06% correct rate when 3 GA and 6 GA optimized features are selected from approximation coefficients. For 3 GA features, the population size is set to 1053; For 6 GA features, the population size is set to 526. Table 5 shows the performance of selected GA features. Table 2 shows that SingleC4.5, BaggingC4.5 and AdaBoostC4.5 methods achieve 67.65%, 75.53% and 67.65% accuracy, which are inferior to our method.

TABLE V
PERFORMANCE OF SELECTED GA FEATURES.

GA FN	CR	SE	SP	PPV	NPV
3	0.9706	1.0000	0.9600	0.9000	1.0000
6	0.9706	1.0000	0.9600	0.9000	1.0000

Performance for Prostate cancer dataset.

Lung cancer

Lung cancer data [19] contains two kinds of tissue including malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung. There are 181 tissue samples (31 MPM and 150 ADCA) including 32 training samples (16 MPM and 16 ADCA) and 149 testing samples (15 MPM and 134 ADCA). The number of genes of each sample is 12533.

Approximation coefficients at 2nd level have 3142 dimensionality. After genetic algorithm is performed on these approximation coefficients, 4 GA and 7GA features are obtained. Table 6 shows the performance of selected GA features. We obtain 97.32% correct rate for 4 GA features and 97.99% correct accuracy for 7 GA features. For 4 andGA features, the population size is set to 785 and the population size is set to 448 for 7 GA features, in order to search full range of wavelet features. Table 2 and Table 7 show the performance of other methods. Table 2 shows that SingleC4.5, BaggingC4.5 and AdaBoostC4.5 methods achieve 92.62%, 93.29% and 92.62% accuracy, which are inferior to our method. Our best performance 97.99% is as same as Li's method.

TABLE VI
PERFORMANCE OF SELECTED GA FEATURES

GA FN	CR	SE	SP	PPV	NPV
4	0.9732	0.9701	1.0000	1.0000	0.7895
7	0.9799	0.9776	1.0000	1.0000	0.8333

Performance for Lung cancer dataset.

TABLE VII
TEST ERROR NUMBERS OF FOUR MODELS [8]

	Test error numbers(MPM:ADCA)			
Dataset	Li's method	Single C4.5	Bagging C4.5	Boosting C4.5
Lung cancer	3(1:2)	27(4:23)	4(0:4)	27(4:23)

V. CONCLUSIONS

In this paper we combine wavelet analysis and genetic algorithm. Wavelet approximation coefficients at 2nd level are used to characterize the essential information contained in microarray data and reduce dimensionality of gene profiles. Genetic algorithm is further performed to select the optimized features from wavelet coefficients. Experiments are carried out on four independent datasets based on selected GA features. For the Leukemia dataset, a 97.06% correct classification rate is achieved only using 8 GA features; 100% correct rate is obtained using 13 GA features for MLL-Leukemia dataset; 3 GA features are used to get 97.06% correct accuracy for Prostate cancer dataset; 97.99% accuracy for Lung cancer dataset is achieved using 7 GA features. Experimental results prove that this hybrid method is feasible and robust.

REFERENCES

[1] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeeck, J,P. Mesirov, H. Coller, M.L. Loh, J.R. Downing and M.A. Caligiuri, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," Science, vol. 286, 1999, pp. 531–537.

[2] V.N. Danh and M.R. David, "Tumor classification by partial least squares using microarray gene expression data," Bioinformatics, vol. 18, 2002, pp. 39–50.

[3] M. Xiong, L. Jin, W. Li and E. Boerwinkle, "Computational methods for gene expression-based tumor classification," Biotechniques, vol. 29, 2000, pp. 1264–1270.

[4] P. Baldi and A.D. Long, "A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes," Bionformatics, vol. 17, 2001, pp. 509–519.

[5] D.S. Huang and C.H. Zheng, "Independent component analysis-based penalized discriminant method for tumor classification using gene expression data," Bioinformatics, vol. 22, 2006, pp. 1855–1862.

[6] A. Statnikov, C.F. Aliferis, I. Tsamardinos, D. Hardin and S. Levy, "A comprehensive valuation of multicategory classification methods for microarray gene expression cancer diagnosis," Bioinformatics, vol. 21, 2005, pp. 631-643.

[7] M. Dettling, "BagBoosting for tumor classification with gene expression data," Bioinformatics, vol. 20, 2004, pp. 3583-3593.

[8] J. Li, H. Liu, S.K. Ng and L. Wong, "Discovery of significant rules for classifying ancer diagnosis data," Bioinformatics, vol. 19, 2003, pp. 93-102.

[9] A.C. Tan and D. Gilbert, "Ensemble machine learning on gene expression data for cancer classification," Appl Bioinformatics, vol. 2, 2003, pp. S75-83.

[10] X. Zhou, X. Wang, and E.R. Dougherty, "Nonlinear-probit gene classification using mutual-information and wavelet based feature selection," Biological Systems, vol. 12, 2004, pp. 371–386.

[11] J.H. Holland, Adaptation in Natural and Artificial Systems, MA: MIT Press, Cambridge, 1992.

[12] K. Fukunaga, Introduction to Statistical Pattern Recognition, 2nd ed. New York: Academic, 1991.

[13] I. Daubechies, "Orthonormal bases of compactly supported wavelets," Communications on Pure and Applied Mathematics, vol. 41, 1988, pp. 909-996.

[14] E.L. Kyeong, N. Sha, E.R. Dougherty, M. Vannucci and B.K. Mallick, "Gene selection: a bayesian variable selection approach," Bioinformatics, vol. 19, 2003, pp. 90–97.

[15] S. Bicciato, A. Luchini and C.D. Bello, "PCA disjoint models for multiclass cancer analysis using gene expression data," Bioinformatics, vol. 19, 2003, pp. 571–578.

[16] C.C. Aedin, P. Guy, C.C. Elizabeth, G.C. Thomas and G.H. Desmond, "Between group analysis of microarray data," Bioinformatics, vol. 18, 2002, pp. 1600–1608.

[17] S.A. Armstrong, J.E. Staunton, L.B. Silverman, R. Pieters, M.L. den Boer, M.D. Minden, S.E. Sallan, E.S. Lander, T.R. Golub and S.J. Korsmeyer, "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia," Nat Genet, vol. 30, 2002, pp. 41–47.

[18] D. Singh, P.G. Febbo1, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D'Amico, J.P. Richie, E.S. Lander, M. Loda, P.W. Kantoff, T.R. Golub and W.R. Sellers, "Gene expression correlates of clinical prostate cancer behavior," Cancer Cell, vol. 1, 2002, pp. 203-209.

[19] G.J. Gordon, R.V. Jensen, L.L. Hsiao, S.R. Gullans, J.E. Blumenstock, S. Ramaswamy, W.G. Richards, D.J. Sugarbaker and R. Bueno, "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma," Cancer research, vol. 62, 2002, pp. 4963-4967