A Hybrid Gene Selection Method for Multi-category Tumor Classification using Microarray Data

Xiaobo Li[1,*], Huijuan Lu[2], Mingjun Wang[1]

[1]*Department of Computer Science*
College of Engineering
Lishui University
Lishui 323000, China
E-mails: oboaixil@126.com, mingjun_w@163.com

[2]*Department of Computer Science*
College of Information Engineering
China Jiliang University
Hangzhou 310018, China
E-mail: hjlu@cjlu.edu.cn

[*]*Corresponding author*

Received: September 08, 2013 **Accepted: December 06, 2013**

Published: December 20, 2013

Abstract: *Microarray technology allows molecular classification of tumors and identification of tumor markers, and it has been used widely in the field of cancer research. Although the problem of binary tumor classification has been addressed extensively, it lacks in-depth research on multi-category tumor classification. In this paper, informative gene selection method, which is a critical step of multi-category tumor classification, was studied. We present a hybrid gene selection strategy aiming to take advantage from the combination of different gene selection algorithms. Top ranked genes of Chi-squared and SVM-RFE algorithms are fused to generate a gene pool, and a genetic algorithm further explores the search space for reduced gene subsets. We tested the proposed model on the multi-category lung cancer microarray gene expression data set. Compared with each individual gene selection algorithm, our hybrid model was able to obtain highest classification performance with much smaller sized subsets of informative genes.*

Keywords: *Multi-category tumor classification, Gene selection, Hybrid model, Genetic algorithm, Microarray data.*

Introduction

Over the last few years, microarray technology has been used extensively in the field of cancer research, and it allows molecular classification of tumors and identification of tumor markers, according to their gene expression profiles [4]. Microarrays can simultaneously measure the expression level of thousands of genes, providing a high-throughput and systematic research platform in cancer research. However, microarrays contain a larger number of genes (generally greater than 10 000), and a smaller sample size (generally less than 100), therefore, it is a critical issue to select a small number of informative genes from thousands of genes for accurate classification [12]. The objective of gene selection is to eliminate noisy and redundant genes, reduce the calculation burden in subsequent classification task and improve the prediction performance of learning model. In addition, an optimal smaller subset of genes may contain biomarkers, which could be more convenient for verification in the subsequent molecular biological experiments, and thus allow for a better understanding of the molecular mechanisms of tumor development [5].

The problem of binary tumor classification (the class number of tumors is 2, for example, classification between tumor and normal tissue samples) has been studied extensively, and achieved satisfactory results [18]. However, for multi-category tumor classification (the class number of tumors is more than 2), it is lack of in-depth research, the classification accuracy is low in the overall published literature, and the classification accuracy decreases greatly as the tumor category number increases [13, 18].

A few new approaches have been proposed for multi-category tumor classification. Liu et al. [14] proposed to combine genetic algorithm (GA) and all paired support vector machine (SVM) methods for multiclass cancer classification. Zhou et al. [20] proposed the MSVM-RFE algorithms, which are four expansions of the well-known SVM method based on recursive feature elimination (SVM-RFE) algorithm. Wang et al. [18] reported a Chi-square-statistic-based Top Scoring Genes (Chi-TSG) classifier for informative gene selection and multi-class cancer classification. However, it is possible to obtain higher classification accuracy when choosing fewer genes by using more powerful dada mining algorithms.

In this paper, a hybrid gene selection method is presented for multi-category tumor classification on microarray data. Section 2 introduces the proposed hybrid model. The experimental analysis is presented in Section 3. Conclusions are drawn in Section 4.

The hybrid model

There are three common methods to gene selection [16]: filter, wrapper and embedded methods. Filter method is independent of the classifiers, despite its computationally simple and fast, it also has several shortcomings: firstly, it ignores the interaction with the classifiers. Secondly, many filter algorithms are univariate, and ignore the dependency between genes. Wrapper method has advantages over filter method which are the interaction between genes and classifiers, and the ability to take into account gene dependencies. However, its computational cost is relatively high, and the selected gene subset has a higher risk of over-fitting [10]. Embedded method takes into account the internal characteristics of the classifier (such as support vectors in support vector machine classifier). It is able to obtain a high accuracy by coupling with the classifier, but its performance depends greatly on the classifier, and the adaptability of the learning model need to be validated on the other classifiers. The SVM-RFE algorithm is one example of embedded method [6].

These limitations advise us to propose a hybrid method [15, 19] that aims at taking advantage from the combination of different types of algorithms. The proposed hybrid model can be illustrated as follows: In the first step, m ($m \geq 2$) gene selection criteria are applied on initial microarray dataset to generate m different ranked gene lists; In the second step, for each ranked gene set, n top ranked genes from m different ranked gene list are input into the gene pool, each gene pool is further evaluated by a genetic algorithm that could search an optimal subset with smaller size and better classification performance.

Most of genes detected in microarray are irrelevant to classification. In the first step, a particular ranking criterion is used to evaluate individual genes, and each gene is assigned a score according to its relevance to the target class. An ordered gene list is generated, and the genes are listed in descending order of relevance. A gene with a high ranking score indicates that the gene contains information potentially useful for classification.

In this study, two different types of gene selection algorithms (i.e. Chi-squared, SVM-RFE) are used. In particular, Chi-squared algorithm is a type of filter method, whereas SVM-RFE is an embedded method. Their definitions are briefly stated below.

Chi-squared method

The Chi-squared method evaluates ranking scores of each individual gene by measuring the Chi-squared statistics (χ^2) with respect to the classes [8]. The χ^2 value of each gene is calculated as

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}\frac{(A_{ij}-E_{ij})^2}{E_{ij}}, \tag{1}$$

where A_{ij} is the total number of patterns in the *i*-th interval, *j*-th class, E_{ij} is the expected frequency of A_{ij} , k is the number of intervals, and *n* is the number of classes.

SVM-RFE method

SVM-RFE method removes recursively the genes that are of least significance to the classifier from the gene set [6]. The significance of each gene to the classifier is defined by the sum square of the weight vector *w*, which is calculated as

$$w = \sum_{i=1}^{n}\alpha_i y_i x_i \ , \tag{2}$$

where α_i is estimated from the training set, $y_i \in [-1, +1]$ is the class label of sample *i*, and x_i is the gene expression vector of a sample *i* in the training set. The training vectors with non-zero α_i are support vectors [12].

In the second step, *n* top ranked genes are selected from each ranked gene lists, and it is important to set an appropriate threshold to retain only the informative genes. Since the choice of *n* is somewhat arbitrary for the filter method, several parameters *n* for choosing top *n* ranked genes are thus used. Genes selected from multiple gene selection algorithms are then input into the gene pool.

Filter algorithms often do not take into account the correlation between genes, and the gene pool may contain redundant genes. A genetic algorithm is applied to find a gene subset including sufficient classification information as possible, but involving fewer genes. GA is a heuristic search method which mimics the process of natural selection in computer [3, 17]. A genetic algorithm simulates the ways in which organisms adapt to natural environment, and the search space is mapped to the genetic space. A possible solution (called individual) to each problem is encoded as a binary string, and a set of individuals is called population.

A GA starts from a population of randomly generated individuals, evaluates the fitness of every individual in the population. The more fitted individuals are retained in the current population, and each individual is modified to form a new generation. A GA is an iterative process which evolves toward better solutions until it meets certain predetermined optimization targets.

The basic components of our GA are described as follows:

- **Representation of individual**
  Each individual is encoded by *n*-bit binary strings, where the bit "1" represents the corresponding gene in the subset being selected, while the bit "0" means the opposite.
- **Fitness function**
  Our genetic algorithm is designed to minimize classification error rate of the chosen classifier. Each individual in a population is evaluated by the classification error rate of a SVM classifier, i.e., SMO (sequential minimal optimization algorithm) classifier in WEKA [7].
- **Genetic operators**
  Roulette wheel selection is used as selection operator, single-point crossover is used as crossover operator, and bit flip mutation is used as mutation operator.

Experimental results

Data set

The lung cancer data set [1] was used to validate the proposed model, the data set was downloaded from the website: http://www.pnas.org/content/98/24/13790/suppl/DC1. This data set is a multi-class tumor gene expression data set, which contains a total of 203 samples and six categories (i.e., 4 subtypes of lung cancer, 1 subtype of extrapulmonary metastasis and 1 subtype of normal tissue). The 203 samples include histologically defined lung adenocarcinomas (*n* = 127), squamous cell lung carcinomas (*n* = 21), pulmonary carcinoids (*n* = 20), SCLC (*n* = 6) cases, and normal lung (*n* = 17) tissues. The other 12 specimen suspected to be extrapulmonary metastases were not included in this experiment. Each sample contains 12 600 gene expression values.

Experimental platform

The experimental data preprocessing was divided into two steps: removal of housekeeping genes and normalization. After removal of the housekeeping gene, 12,533 gene expression values of each sample remained. The gene expression values are normalized so that the gene expression value of each sample has a mean of 0 and a standard deviation of 1.

The experiments were carried out by using the WEKA [7] (http://www.cs.waikato.ac.nz/ml/weka/) platform, which provides a variety of gene selection algorithms as well as the genetic search and the classification model. The SMO classifier was used to perform the classification task. The SMO classifier contains 4 kinds of kernel functions, including NormalizedPolyKernel, PolyKernel, RBFKernel and StringKernel, and the polynomial kernel function (PolyKernel) was selected. The optimal adjustment of parameters in training SVM classifier is very time-consuming, and the parameters were specified in a fixed manner hereby. Since the data was already normalized, the "FilterType" parameter was set to "standardize training data" option. The penalty parameter *C* was set to 100. The 10-fold cross-validation was used to evaluate the performance of a SMO classifier. In the 10-fold cross-validation, the data is randomly partitioned into 10 subsets of (approximately) equal size. The classifier is trained 10 times, each time 9 subsets are used as training data, and a single subset is retained as the validation data for computing the classification accuracy. The 10 results from the folds are then averaged.

The GA parameters used in our experiment were set as follows:

- probability of crossover = 1,
- probability of mutation = 0.02,
- number of generations = 50,
- population size = 30.

In fitness evaluation, classification error rate was performed by a 10-fold cross-validation for SMO classifiers.

Experimental results

Both the Chi-squared and SVM-RFE algorithms generate a gene list where the gene scores rank from high to low, and subsets of top *n* genes were used to assess the performance of the classifiers. As shown in Table 1, in general, SVM-RFE gives better performance, and it achieves 100% accuracy with 30 genes. It is because SVM-RFE is an embedded method, and it has a good coupling with SVM classifier.

Table 1. 10-fold accuracy of Chi-squared and SVM-RFE algorithms on lung cancer data set

Top *n* genes	**Chi-squared (%)**	**SVM-RFE (%)**
10	83.77	95.81
20	82.72	97.91
30	90.58	100.00
50	92.15	99.48
100	94.76	100.00
200	95.29	99.48
500	96.34	99.48
1000	96.86	98.95
2000	97.38	98.43
5000	96.86	97.91
12533	94.24	94.24

To validate our proposed hybrid model, we tested several gene pools, each of which contains a different number of top *n* genes chosen from the Chi-squared and SVM-RFE algorithms, where *n* is set to 10, 20, 30, 50 and 100. As the genetic algorithm is a stochastic search method, 10 trials were performed on each gene pool, and the results were averaged. The top-20 genes given by the two algorithms are shown in Table 2, and the only overlapped gene is VAMP2.

As shown in Table 3, when top-20 genes were searched, the genetic algorithm was capable of finding smallest size of subset and achieves 100% classification accuracy. The average subset size of 14.6 genes is much less than SVM-RFE method while it needs 30 genes to obtain the same accuracy.

All of the 10 subsets selected from the gene pool of top-20 genes achieve 100% accuracy, and 3 subsets out of 10 contain minimum number of genes (n = 13). The selected genes of the minimum subsets are shown in Table 4.

Table 2. Top-20 genes selected from the Chi-squared and SVM-RFE algorithms (the word in parentheses is probe ID)

Gene selection algorithms	Top-20 genes
Chi-squared	VAMP2(32254_at), SFN (33322_i_at), CHGB (33426_at), SCG5 (34265_at), PAR-SN /// SNORD107 /// SNRPN /// SNURF (34842_at), NOS1AP (35531_at), APLP1 (36148_at), PTPRN2 (36160_s_at), SCG2 (36924_r_at), SYP (37182_at), INA (37210_at), MAPRE2 (37406_at), SCAMP5 (37545_at), SV2A (38032_at), S100A11 (38138_at), PSD (38174_at), SNAP25 (38484_at), TERF2IP (38982_at), CHGA (40808_at), KCNK3 (41325_at)
SVM-RFE	SMAD6 (1955_s_at), ERBB3 (2089_s_at), NDUFS7 (31638_at), FXYD1 (32109_at), VAMP2 (32254_at), ADH7 (33529_at), SH3BP1 (34046_at), PNPLA6 (34874_at), PCYT1B (35552_at), DSP (36133_at), BRD2 (36209_at), CBX7 (36894_at), TUBA3C /// TUBA3D (38350_f_at), EMP3 (39182_at), KCND3 (39266_at), ISL1 (39990_at), UBE2S (40619_at), MAPRE3 (40825_at), HMGN2 (41231_f_at), FKBP1A (880_at)

Table 3. 10-fold accuracy of the proposed model on lung cancer data set

Top *n* genes	Average accuracy (%)	Average subset size
10	98.00	9.60
20	100.00	14.60
30	100.00	17.30
50	100.00	27.00
100	100.00	57.30

The genes with star (*) are selected from the SVM-RFE method, and the other genes are chosen from the Chi-squared method. The minimum informative gene subsets are combined by the genes from multiple outcomes of gene ranking algorithms. It is observed that 6 genes of SMAD6, FXYD1, ADH7, PCYT1B, HMGN2, and FKBP1A are picked in all of the three trials. Out of the 6 genes, SMAD family member 6 (SMAD6) [9], and high mobility group nucleosomal binding domain 2 (HMGN2) [2] were reported to be associated with lung cancer. SMAD 6 is a member of SMAD family of proteins, and it plays a role in BMP and TGF-beta signaling pathway. High expression of SMAD6 was reported to be associated with a reduced survival in lung cancer patients [9]. HMGN2 is probably involved in control of chromatin structure and transcription. Recently, microarray analyses suggested that HMGN2 acts as a positive modulator of nuclear factor κB signaling pathway in lung cancer cells [2].

Table 4. The selected genes of the minimum subset on lung cancer data set (the word in parentheses is probe ID)

Trials	Selected genes
1	SMAD6 (1955_s_at)*, ERBB3 (2089_s_at)*, FXYD1 (32109_at)*, CHGB (33426_at), ADH7 (33529_at)*, SH3BP1 (34046_at)*, PCYT1B (35552_at)*, DSP (36133_at)*, MAPRE2 (37406_at), KCND3 (39266_at)*, MAPRE3(40825_at)*, HMGN2(41231_f_at)*, FKBP1A(880_at)*
2	SMAD6 (1955_s_at)*, NDUFS7 (31638_at)*, FXYD1 (32109_at)*, ADH7 (33529_at)*, PNPLA6 (34874_at)*, PCYT1B (35552_at)*, BRD2 (36209_at)*, INA (37210_at), SCAMP5 (37545_at), S100A11 (38138_at), KCND3 (39266_at)*, HMGN2 (41231_f_at)*, FKBP1A (880_at)*
3	SMAD6 (1955_s_at)*, ERBB3 (2089_s_at)*, FXYD1 (32109_at)*, ADH7 (33529_at)*, PNPLA6 (34874_at)*, PCYT1B (35552_at)*, APLP1 (36148_at), BRD2 (36209_at)*, MAPRE2 (37406_at), PSD (38174_at), TUBA3C /// TUBA3D (38350_f_at)*, HMGN2 (41231_f_at)*, FKBP1A (880_at)*

Conclusions

In recent years a remarkable progress has been seen in the use of high-throughput techniques such as microarrays for molecular classification of tumors. The development, invasion and metastasis of tumor is a multi-stage, multi-gene regulated, multi-pathway process [11] which results in tumor heterogeneity and multi-category subtypes of tumors. The problem of multi-category tumor classification remains a challenge in the field of machine learning.

In this paper, we present a hybrid feature selection model for gene expression-based multi-category tumor classification. Genes from multiple outcomes of gene ranking algorithms are combined, and a genetic algorithm is applied to search an optimal subset. Compared with each individual ranking gene selection algorithm, our hybrid model is capable of finding much smaller sized subsets of informative genes and obtaining highest classification performance. We leave for future work the investigation of combining additional gene ranking algorithms, as well as the other state of the art classifiers.

Acknowledgments

This study was supported by grants from the National Natural Science Foundation of China (Grant No. 61373057 and 61272315), and a grant from the Zhejiang Provincial Natural Science Foundation of China (Grant No. Y1110763).

References

1. Bhattacharjee A., W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E. J. Mark, E. S. Lander, W. Wong, B. E. Johnson, T. R. Golub, D. J. Sugarbaker, M. Meyerson (2001). Classification of Human Lung Carcinomas by mRNA Expression Profiling Reveals Distinct Adenocarcinoma Subclasses, Proc Natl Acad Sci USA, 98(24), 13790-13795.
2. Deng L. X., G. X. Wu, Y. Cao, B. Fan, X. Gao, L. Luo, N. Huang (2011). The Chromosomal Protein HMGN2 Mediates Lipopolysaccharide-induced Expression of Beta-defensins in A549 Cells, The FEBS Journal, 278(12), 2152-2166.
3. Dessi N., B. Pes (2009). An Evolutionary Method for Combining Different Feature Selection Criteria in Microarray Data Classification, Journal of Artificial Evolution and Applications – Special Issue on Artificial Evolution Methods in the Biological and Biomedical Sciences, 2009, Article ID 803973, http://dx.doi.org/10.1155/ 2009/803973.
4. Golub T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, E. S. Lander (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, Science, 286(5439), 531-537.
5. Guyon I., A. Elisseeff (2003). An Introduction to Variable and Feature Selection, Journal of Machine Learning Research, 3, 1157-1182.
6. Guyon I., J. Weston, S. Barnhill, V. Vapnik (2002). Gene Selection for Cancer Classification using Support Vector Machines, Machine Learning, 46(1-3), 389-422.
7. Hall M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten (2009). The WEKA Data Mining Software: An Update, SIGKDD Explorations Newsletter, 11(1), 10-18.
8. Huan L., S. Rudy (1995). Chi2: Feature Selection and Discretization of Numeric Attributes, Proceedings of the 7th International Conference on Tools with Artificial Intelligence, IEEE Computer Society, 388-391.
9. Jeon H. S., T. Dracheva, S. H. Yang, D. Meerzaman, J. Fukuoka, A. Shakoori, K. Shilo, W. D. Travis, J. Jen (2008). SMAD6 Contributes to Patient Survival in Non-small Cell Lung Cancer and its Knockdown Reestablishes TGF-beta Homeostasis in Lung Cancer Cells, Cancer Research, 68(23), 9686-9692.
10. Leung Y., Y. Hung (2010). A Multiple-filter-multiple-wrapper Approach to Gene Selection and Microarray Data Classification, IEEE – ACM Transactions on Computational Biology and Bioinformatics, 7(1), 108-117.
11. Li X. B., J. Chen, B. J. Lu, S. H. Peng, R. Desper, M. D. Lai (2011). -8p12-23 and +20q Are Predictors of Subtypes and Metastatic Pathways in Colorectal Cancer: Construction of Tree Models Using Comparative Genomic Hybridization Data, OMICS A Journal of Integrative Biology, 15(1-2), 37-47.
12. Li X., S. Peng, J. Chen, B. Lu, H. Zhang, M. Lai (2012). SVM-T-RFE: A Novel Gene Selection Algorithm for Identifying Metastasis-related Genes in Colorectal Cancer using Gene Expression Profiles, Biochemical and Biophysical Research Communications, 419(2), 148-153.
13. Li X., S. Peng, X. Zhan, J. Zhang, Y. Xu (2011). Comparison of Feature Selection Methods for Multiclass Cancer Classification based on Microarray Data, Proceedings of the 4th International Conference on Biomedical Engineering and Informatics (BMEI), 3, 1692-1696.
14. Liu J. J., G. Cutler, W. Li, Z. Pan, S. Peng, T. Hoey, L. Chen, X. B. Ling (2005). Multiclass Cancer Classification and Biomarker Discovery using GA-based Algorithms, Bioinformatics, 21(11), 2691-1697.

15. Rothlauf F., J. Branke, S. Cagnoni, E. Costa, C. Cotta, R. Drechsler, E. Lutton, P. Machado, J. Moore, J. Romero, G. Smith, G. Squillero, H. Takagi, E. Huerta, B. Duval, J.-K. Hao (2006). A Hybrid GA/SVM Approach for Gene Selection and Classification of Microarray Data, Applications of Evolutionary Computing, Springer Berlin Heidelberg, 34-44.
16. Saeys Y., I. Inza, P. Larranaga (2007). A Review of Feature Selection Techniques in Bioinformatics, Bioinformatics, 23(19), 2507-2517.
17. Tan F., X. Fu, Y. Zhang, A. A. Bourgeois (2008). Genetic Algorithm-based Method for Feature Subset Selection, Soft Computing, 12(2), 111-120.
18. Wang H., H. Zhang, Z. Dai, M. S. Chen, Z. Yuan (2013). TSG: A New Algorithm for Binary and Multi-class Cancer Classification and Informative Genes Selection, BMC Medical Genomics, 6(1), 1:S3.
19. Yin H., W. Wang, V. Rayward-Smith, L. Cannas, N. Dessi, B. Pes (2011). A Hybrid Model to Favor the Selection of High Quality Features in High Dimensional Domains. Intelligent Data Engineering and Automated Learning – IDEAL 2011, Springer Berlin Heidelberg, 228-235.
20. Zhou X., D. P. Tuck (2007). MSVM-RFE: Extensions of SVM-RFE for Multiclass Gene Selection on DNA Microarray Data, Bioinformatics, 23(9), 1106-1114.

Assoc. Prof. Xiaobo Li, Ph.D.
E-mail: oboaixil@126.com

Dr. Li received his B.Sc. in Microelectronics (1990) from Nankai University (China), Master of Engineering (Research) (2004) from The University of Sydney (Australia) and Ph.D. in Pathology and Pathophysiology (2012) from Zhejiang University (China). Now he is a full-time Associate Professor at Department of Computer Science, College of Engineering, Lishui University, China. His current research interests include different aspects of bioinformatics, machine learning and data mining.

Prof. Huijuan Lu, Ph.D.
E-mail: hjlu@cjlu.edu.cn

Prof. Lu received her B.Sc. in Automation (1986) from China University of Mining and Technology, M.Sc. in Electrical Engineering (1995) from Zhejiang University and Ph.D. in Information and Electrical Engineering (2012) from China University of Mining and Technology. Now she is a professor and director of Computer Software Research Institute of China Jiliang University. She became a senior visiting scholar in Temple University in the United States in 2012. Since 2012 she is an executive director of the Computer Society of China. Since 2011 she is the teaching master of Zhejiang Province. Her current research interests include pattern recognition, bioinformatics, data mining, things of Internet, cloud computing and logistics engineering.

Dr. Mingjun Wang, Ph.D.
E-mail: mingjun_w@163.com

Dr. Wang received his B.Sc. in Computer Science and Technology from Beijing Jiaotong University (2004), M.Sc. in Computer Application (2006) from China University of Geosciences (Beijing) and Ph.D. in Pattern Recognition and Intelligent System (2010) from Donghua University. Now he is a full-time lecturer of informatics at Engineering College, Lishui University. His current research interests include different aspects of bioinformatics and data mining.