Reducing the Variation of Gene Expression Patterns: A Grey Model Approach Applied to Microarray Data Classification

Tsun-Chen Lin[1], Ru-Sheng Liu[2], Chen-Chung Liu[3], Shu-Yuan Chen[4]
Department of Computer Science and Engineering, Yuan Ze University
[1]lintsunc@ms01.dahan.edu.tw, [2] csrobinl@saturn.yzu.edu.tw
[3]christia@saturn.yzu.edu.tw, [4]cschen@saturn.yzu.edu.tw

Chieh-Yu Chen
Graduate School of Biotechnology and Bioinformatics, Yuan Ze University
cychen@mars.csie.ntu.edu.tw

Abstract

The gene expression levels measured from microarray spots vary between patients with the same type of tumor. This reduces the performance of microarray classification, especially when the microarray dataset has few samples. Here, we introduce the grey model GM(1,1) of the grey system theory for modeling gene expression patterns of small samples to eliminate variations. To evaluate the application of GM(1,1), we have combined GM(1,1) with GA/MLHD approach to solve the problem of multi-class classification. The GM(1,1)-GA/MLHD model was tested on two published microarray datasets: (1) NCI60 cancer cell lines and (2) the GCM dataset. The experimental results show that the GM(1,1) gave gene expression patterns with less variations and helped the MLHD classifier to improve classification accuracy over the method of GA/MLHD, but they also outperformed many class prediction approaches.

Keywords: GM(1,1), microarray, genetic algorithm, classification, tumor class.

1. Introduction

The development of microarray technologies has provided a powerful tool by which the expression profiles of thousands of genes can be monitored simultaneously. One of the most promising applications of this technology is to provide a useful tool for tumor classification. Several previous works such as Alizadeh et al. [1], Ben-Dor et al. [2], and Golub et al. [14] have been proposed and have given promising results for most binary class data. However, if we consider the classification of gene expression data into more than two classes, the performance of most binary or 3-class methods will decrease significantly. One of the reasons is that a dataset may contain many classes and the sample size is small. Recently, Romualdi et al. [5] tried a simulation approach to control the huge source of variation among and between patients and to evaluate a series of supervised statistical techniques. The simulation results show that all the methodologies have comparable performances when the number of patient samples per tumor is greater than 50, the number of tumors is lower than 4 and the number of discriminating gene is larger than 40. As there might be over a hundred types of cancer and potentially even more subtypes Hanahan. and Weinberg. [6] the microarray experiments are still too costly and time consuming so limit the number of samples, and therefore make variations within a class become more accentuated relatively. This fact suggests that the methodologies used to reduce the variability of gene expression patterns should be encouraged for practical applications.

The grey model GM(1,1) as introduced by Deng. [8-9] has been successfully used in many research areas to filter out the random variations in control systems. This leads to our proposal for using GM(1,1) to reduce the variability and identify regularity of gene expressions. The key issues in using GM(1,1) are based on the accumulated generating operation (AGO), a group of differential equations, and the inverse accumulated generating operation (IAGO) to transform gene expression profiles. In this work, we have made many experiments using the GM(1,1)-GA/MLHD (genetic algorithm/maximum likelihood) approach of Ooi et al. [3] under different parameter settings of GA in order to demonstrate that GM(1,1) has generality to take the advantage not only of transforming a gene expression pattern into a less-noise pattern by using a few samples (at least 4 samples /class), but also of assessing the regularity for patterns to associate their phenotypes unique to a class. Finally, our method provides an average 2.5% improvement in classification accuracy for NCI60 data, and 2% improvement for the GCM dataset. In

addition, the importance of gene selection on microarray data analysis was also emphasized.

2. Methods

2.1. Data Transformation of GM(1,1)

GM(1,1) is one of the GM(m,n) models of grey system theory. The m=1 and n=1 inside the parentheses indicate the 1st-order AGO and the number of variables of the differential equation, respectively. Our works, following the GM(1,1) model, would initially treat a given gene expression pattern of uncertain (high or low) expression values as a serial output of a variable/gene from different patients with the same type of tumor. When constructing a grey model, the GM(1,1) first applies a 1st-order AGO to this pattern to provide the middle message to weaken the variation tendency. Next, a group of grey differential equations were used to give function to this AGO-generated gene expression pattern. Finally, the GM(1,1) requested a 1st-order IAGO (Inverse-AGO) to predict the outputs of gene expressions from GM(1,1)-treated sequence. Therefore, the proposed method is composed of the operations: AGO, GM(1,1), and IAGO for the purpose of transferring raw gene expressions to a new gene expression pattern by steps described in detail as follows.

Step 1. Let $y_i^{(0)}$ be the gene expression levels for a given gene *i* to a specific tumor type

$$y_i^{(0)} = (y_i^{(0)}(1), y_i^{(0)}(2), ..., y_i^{(0)}(k)), k \geq 4$$

where *k* denotes the number of samples.

Step 2. To reduce the variations, we define the 1st-order AGO on $y_i^{(0)}$ by following operation.

$$y_i^{(1)}(k) = AGO \bullet y_i^{(0)} \equiv \sum_{j=1}^{k} y_i^{(0)}(j), k = 1,2,....n$$

The number "1" in the parentheses on the superscript denotes 1st-order AGO.

Step 3. To derive the exponential 1st-order grey function of $y_i^{(1)}$, the GM(1,1) defines the grey differential equation as

$$y_i^{(0)}(k) + az_i^{(1)}(k) = u \qquad (1)$$

where *a* is the development coefficient and *u* is the grey control variable of GM(1,1). We also define

$$z_i^{(1)}(k) = \alpha(y_i^{(1)}(k) + y_i^{(1)}(k-1)), k = 2,3,....n$$

where the parameter α=0.5 means a MEAN operation on $y_i^{(1)}$. Next, we can build up the whitening equation corresponding to Equation (1) as

$$\frac{dy_i^{(1)}(k)}{dx} + ay_i^{(1)}(k) = u \qquad (2)$$

For an approximate solution to determine the *a* and *u* of Equation (1), the least squares estimation method by pseudo-inverse matrix $B\hat{\theta} = Y$ is applied and yields

$$\hat{\theta} = \begin{bmatrix} a \\ u \end{bmatrix} = (B^T B)^{-1} BY$$

where $B = \begin{bmatrix} -z_i^{(1)}(2) & 1 \\ -z_i^{(1)}(3) & 1 \\ . & \vdots \\ . & \\ -z_i^{(1)}(n) & 1 \end{bmatrix}$ $Y = \begin{bmatrix} y_i^{(0)}(2) \\ y_i^{(0)}(3) \\ \vdots \\ \\ y_i^{(0)}(n) \end{bmatrix}$

Substituting *a* and *u* into Equation (2), the $\hat{y}_i^{(1)}(k)$ can be obtained and further be expressed as

$$\hat{y}_i^{(1)}(k) = \left(y_i^{(0)}(1) - \frac{u}{a} \right) \cdot e^{-a(k-1)} + \frac{u}{a} \qquad (3)$$

where the symbol "^" means the prediction value.

Step 4. To predict sequence $\hat{y}_i^{(0)}$ from $\hat{y}_i^{(1)}$, the corresponding IAGO is defined as

$$\hat{y}_i^{(0)}(k) = \hat{y}_i^{(1)}(k) - \hat{y}_i^{(1)}(k-1) \qquad (4)$$

Equation (5) is calculated by substituting Equation (3) into Equation (4) so that the $\hat{y}_i^{(0)}$ with respect to the data sequence $y_i^{(0)}$ is obtained by

$$\hat{y}_i^{(0)}(k) = \left(y_i^{(0)}(1) - \frac{u}{a} \right) \cdot (1 - e^a) \cdot e^{-a(k-1)} \qquad (5)$$

Step 5. Following the method mentioned above, a dataset then can be preprocessed by following procedures to filter out the noises and will be used by the classifier.

1. For each class C_j in gene expression profiles
2. For each gene *i* in gene expression profiles
3. For each sample's order *k*>1 in C_j
4. Calculate Equation (5)

2.2. Reviews of GA/MLHD

Examinations to the GM(1,1) model, applied to microarray classification, were performed by using an available implementation of the GA toolbox for gene selection and the MLHD classifier for discrimination analysis from Ooi et al. [3]. In order to work with an ensemble of different gene subspaces (sets of predictor genes), this GA toolbox provides two selection methods: (1) stochastic universal sampling (SUS) and (2) roulette wheel selection (RWS). It also provides two tuning parameters, *Pc*: crossover rate and *Pm*: mutation rate, used to tune one-point and uniform crossover operations to evolve the population in the mating pool for choosing the optimal genes, consisting of chromosomes, to work with the MLHD classifier.

For the discrimination analysis to a chromosome, the chromosome is designed by a string S_i, S_i = [R g_1, g_2 ... $g_{i=Rmax}$], where R value denotes the number of genes, and g_1, g_2...$g_{i=Rmax}$ denotes the indices of

predictor genes. In the process of pattern classification, the first R genes out of $g_{1,}\, g_{2\ \ldots}\, g_{i=Rmax}$ are then used to form dataset of sample patterns for all samples and will be fed into the MLHD classifier. The essence of the MLHD classifier is based on a discriminant function to estimate the discrimination score of genes in classifying tumor samples, and it is given by

$$f_q(\vec{e}) = \vec{\mu}_q^T \sum^{-1} \vec{e} - \frac{1}{2} \vec{\mu}_q^T \sum^{-1} \vec{\mu}_q$$

where $\vec{\mu}_q = (\mu_{q,1}, \mu_{q,2}, ...\mu_{q,R})^T$ is the class mean vector, $\mu_{q,i}$ is the average expression level of gene i for all samples belonging to class q, $\sum$ means the common covariance matrix [10] between classes and is defined as

$$\Sigma = \frac{1}{M_t - Q} \sum_{q=1}^{Q} \Sigma_q$$

where $\sum_q$, $q \in \{1,2,\ldots\ldots,Q\}$ is the class covariance matrix of the selected R genes for all training samples belonging to class q, and M_t is the number of all training samples.

To predict the class of an query sample pattern $\vec{e} = (e_1, e_2, ..., e_R)^T \in$ class q, where the element e_i is the expression level of gene i, the classification rule of the MLHD classifier is defined as

$$\mathrm{C}(\vec{e}) = q, \text{ where } f_q(\vec{e}) > f_r(\vec{e}) \quad (6)$$

for $q \neq r$, $r \in \{1,2 \ldots \mathrm{Q}\}$, and Q is the possible classes of the experiment dataset. The MLHD classifier also defines a fitness function as $f(S_i) = 200 - (E_C + E_I)$, where E_C is the error rate of the Leave-One-Out Cross-Validation (LOOCV) test on the training data , and E_I is the error rate of independent test on the test data. By calculating *Ec* and *Ei* under the classification rule as Equation (6), the returning fitness value, which is in turn, will be used by GA to evolve better gene subsets.

2.3. Prediction Errors

In a domain of Q classes, the success rate estimations through GM(1,1)-GA/MLHD method begin with setting 100 runs by following the program, with each run beginning with a different initial gene pool in order to have an unbiased estimate of classifier performance. The maximum generations for each run are set to 100, which each generation produces 100 and 30 chromosomes with size of genes ranging from R_{min}=11 to R_{max}=15 and from R_{min}=5 to R_{max}=50 in a chromosome corresponding to the NCI60 and the GCM dataset respectively.

According to the gene indices in each chromosome, only the first R numbers of genes out of $g_{1,}\, g_{2,\ldots}\, g_{nmax}$ are picked to form sample patterns for all samples and hereby assume that we are given a dataset of reduced dimensions and to be evaluated by MLHD classifier.

1. FOR each generation *G*=1 to *G*=100
2. FOR each chromosome *C*=1 to C=100
3. FOR each training sample $\vec{e} \in class\ q$
4. Build up discriminant model with remaining training samples for LOOCV tests
5. IF ($f_q(\vec{e}) \leq f_r(\vec{e})$)
6. *Xc*Error = *Xc*Error + 1 // sample misclassified
7. END FOR
8. FOR each unknown sample $\vec{e} \in class\ q$
9. Build up discriminant model with all training samples for independent tests
10. IF ($f_q(\vec{e}) \leq f_r(\vec{e})$)
11. *Xi*Error = *Xi*Error + 1 // sample misclassified
12. END FOR
13. *Ec*ErrorRate = *Xc*Error / Total training samples
14. *Ei*ErrorRate = *Xi*Error / Total test samples
15. Fitness[G][C]=200–(*Ec*ErrorRate + *Ei*ErrorRate)
16. END FOR
17. END FOR
18. Findmax (Fitness) // best chromosome

In the running of above program through 100 generations and 100 individual runs, the chromosome with the best fitness, chosen from the simulation to arrive at the optimal operation will be based on the idea that a classifier need not only work well on the training samples, but also work equally well on previously unseen samples. Therefore, the optimal individuals of each generation were sorted in ascending order by the sum of the error number on both tests. The smallest number then determines the chromosome that contains discriminatory genes and the number of genes needed in the classification as well as gives the classification accuracy obtained by our methods.

3. Datasets

There are two published microarray datasets from human cancer cell lines will be used in this paper. Before the datasets were used in our experiments, the data was preprocessed by following steps.

1. The spots with missing data, control, and empty spots were excluded.
2. For each sample array in both datasets, the gene expression intensity of every spot was normalized by subtracting the mean expression intensity of control spots and dividing the result by the standard deviation of control spots.

3. A preliminary selection of 1000 genes with the highest ratios of their between-groups to within–groups sum of squares (BSS/WSS) was performed. In our case, the BSS/WSS ratios for NCI60 data are ranging from 0.4 to 2.613 and from 0.977 to 3.809

for the GCM. For gene *i*, x_{ij} denotes the expression level from patient *j*, and the ratio is define as

$$\frac{BSS(i)}{WSS(i)} = \frac{\sum_{j=1}^{M_t}\sum_{q=1}^{Q} I(c_j = q)(\mu_{qi} - \mu_{\bullet i})^2}{\sum_{j=1}^{M_t}\sum_{q=1}^{Q} I(c_j = q)(x_{ij} - \mu_{qi})^2}$$

where M_t is the training sample size, Q is the number of classes and $I(\bullet)$ is the indicator function which equal 1 if the argument inside the parentheses is true, and 0 otherwise. $\mu_{\bullet i}$ denotes the average expression level of gene *i* across all samples, μ_{qi} denotes the average expression level of gene *i* across all samples belonging to class *q*. This is the same gene preselection method as the paper of Dudoit et al [12].

3.1. The NCI60 Dataset

The NCI60 dataset Ross et al. [7] were measured with 9,703 spotted cDNA sequences among the 64 cell lines from tumors with 9 different sites of origin from the National Cancer Institute's anti-cancer drug screen and can be downloaded from http://genome-www.stanford.edu/sutech/download/nci60/dross_array_nci60.tgz. During the data preprocessing, the single unknown cell line and two prostate cell lines were excluded due to their small sample size, leaving a matrix of 1000 genes × 61 samples. These genes are henceafter referred to by their index numbers (1 to 1000) in our experiments. To build the classifier and run GM(1,1) in a small size of training samples, this dataset was divided into a learning and test set (2:1 scheme, 41 samples for training and 20 for testing). The 41 patient samples are gene expression levels composed of 5 breast, 4 central nervous system (CNS), 5 colon, 4 leukemia, 5 melanoma, 6 non-small-cell-lung-carcinoma (NSCLC), 4 ovarian, 5 renal, and 3 reproductive.

3.2. The GCM Dataset

The GCM dataset Ramaswamy et al. [14] were measured by Affymetrix Genechips containing 16063 genes among 198 samples with 14 different classes of tumor, and can be obtained from http://www-genome.wi.mit.edu/mpr/publications/projects/Global_Cancer_Map/. In our data preprocessing, the dataset left a matrix of 1000 genes × 198 samples. These genes are referred to by their index numbers (1 to 1000) in our experiments. This dataset originally contains 144 samples for training, and 54 for testing. The 144 patient samples are gene expression levels composed of 8 breast, 8 prostate, 8 lung, 8 colorectal, 16 lymphoma, 8 bladder, 8 melanoma, 8 uterine, 24 leukemia, 8 renal, 8 pancreatic, 8 ovarian, 8 mesothelioma, and 8 brain.

4. Results and Discussions

4.1. Classification Accuracy

In this section, the classification accuracy of MLHD classifier, using GM(1,1)-treated NCI60 dataset, will be compared to its performance without using GM(1,1)-treated data. From Table 1, the best predictive accuracies are achieved using the Uniform crossover and SUS selection strategy with GAs. The best predictor set obtained from GM(1,1)-GA/MLHD method exhibits a cross-validation success rate of 87.8% while the success rate of GA/MLHD is 83%. Even in diagnosing blind test samples our method needs only 11 predictive genes to produce a success rate of 95% (overall success rate = 91.4%), whereas GA/MLHD needs 14 predictive genes to produce *Ei* = 95% (overall success rate = 89%). If we take only the average performance over different parameter settings on independent test data, the mean test error rate in the model of GM(1,1)-GA/MLHD is 7.5%, and in the model of GA/MLHD it is 10%. This indicates that our method is better than GA/MLHD and has improved 2.5% better classification accuracy.

Table 1. Recognition error rate (%) and the parameters used in the NCI60 dataset.

NCI60 dataset (1000 genes)				GM(1,1) - GA/MLHD				GA/MLHD			
Pc	Pm	Crossover	Selection	*Ec*	*Ei*	*Ea*	*R*	*Ec*	*Ei*	*Ea*	*R*
1	0.002	Uniform	SUS	12.19	5	8.59	11	17.07	5	11	14
0.7	0.005	One-point	SUS	9.75	10	9.87	11	17.07	10	13.53	13
0.7	0.001	Uniform	RWS	19.5	5	12.25	11	26.82	10	18.41	14
0.8	0.02	One-point	RWS	19.5	10	14.75	11	21.95	15	18.47	13

Ea: overall error rate ((*Ec* + *Ei*) / 2); *R*: optimal number of predictive genes.

Having obtained good performance on the NCI60 dataset with 9 classes, we next tested the proposed method on a more complicated dataset consisting of 14 classes with each class containing more samples to examine the generality of our method. By using the experience with the NCI60 dataset, we also employed the Uniform and SUS strategies and set the *Pc* = 1, 0.8, 0.8, and *Pm* = 0.002, 0.02, 0.001 respectively. From Table 2, the best outcomes that selected an optimal gene set of 17 elements producing *Ec* = 5.5% and *Ei* = 18.5% (over all rate = 12%) of our method still outperformed

GA/MLHD method, which selected an optimal geneset of 35 elements, producing *Ec* = 19.4% and *Ei* = 18.5% (over all rate = 18.95%). When the NCI60 and the GCM datasets have become two popular benchmark data used by many classification algorithms, we list the performance differences for the NCI60 and the GCM datasets among various methods. More detailed discussions to these methods can be found in the papers of Ooi et al. [3], Yeang et al. [4], and Peng et al. [13]. In our comparisons with these methods, we found the best model of our methods yielded clear improvements compared to other approaches listed in the Table 3 with respect to the dataset they used.

Table 2. Recognition error rate (%) and the parameters used in the GCM dataset.

GCM dataset (1000 genes)				GM(1,1) GA/MLHD				GA/MLHD			
Pc	Pm	Crossover	Selection	*Ec*	*Ei*	*Ea*	*R*	*Ec*	*Ei*	*Ea*	*R*
1	0.002	Uniform	SUS	5.5	18.5	12	17	19.4	22.2	20.8	27
0.8	0.02	Uniform	SUS	5.5	20.4	12.9	15	26.4	22.2	24.3	29
0.8	0.001	Uniform	SUS	6.3	18.5	12.4	20	19.4	18.5	18.95	35

Ea: overall error rate ((*Ec* + *Ei*) / 2); *R*: optimal number of predictive genes.

Table 3. Classification accuracy comparisons among different approaches.

NCI 60 dataset	LOOCV (%)	Independent test (%)	Overall (%)	Genes needed	Reference
GM(1,1)-GA/MLHD	88	95	92	11	[This paper]
GA/MLHD	83	95	89	14	[3]
GA/SVM/RFE	88	—	—	27	[13]
GCM dataset					
GM(1,1)-GA/MLHD	94	82	88	17	[This paper]
GA/SVM/RFE	85	—	—	26	[13]
GA/MLHD	79	82	82	32	[3]
OVA/SVM	81	78	80	16063	[4]
OVA/KNN	73	54	63	100	[4]

4.2. Perturbed Versions of Learning Data

In this section, we examine a possible effect that may influence the tumor classification using the GM(1,1) model. As we have mentioned, for a given gene, GM(1,1) generated new pattern, which was transformed from the original gene expression levels across a set of micrarrays through Equation (5) depending on the development coefficient, the grey control variable and the first value of the original sequence. We questioned whether if we reorder the order of input sequence, GM(1,1) may be capable of selecting better patterns for discrimination analyses. Therefore, we took the NCI60 dataset for example and tried to randomly reorder the learning dataset into 30 different perturbed versions of the same size as the original learning set without changing the test data in order to examine the effect of sequence order on the results of classification. Strikingly, with the Uniform and SUS strategy of GA, we found that the best version of training dataset produced a cross-validation error rate equal to 2.4% and a test error rate equal to 5%. In Table 5, we list the results performed by the best and worst versions of the datasets, as well as the average performances over 30 learning sets. Although this procedure is computationally more expensive, it is valuable for the pattern recognition to select discriminatory genes and thus improve the accuracy in the classification.

Table 5. Classification results of 30 different training sets for the NCI60 data.

NCI60 dataset (1000 genes)				Best				Worst				Average			
Pc	Pm	Crossover	Selection	*Ec*	*Ei*	*Ea*	*R*	*Ec*	*Ei*	*Ea*	*R*	*Ec*	*Ei*	*Ea*	*R*
1	0.002	Uniform	SUS	2.4	5	3.7	11	12.2	5	8.6	13	2.6	5.7	4.15	11

Ea: overall error rate ((*Ec* + *Ei*) / 2); *R*: optimal number of predictive genes.

4.3. Gene Expression Patterns

According to the assumption of Yeang et al. [4], individual genes for the same type of tumor may share some expression profile patterns unique to their class. The usage of GM(1,1) to provide gene expression profiles with less variations within a class but more discriminatory information among classes then will help gene expression patterns catch internal regularity and become more tightly associated to class phenotype for samples in the same tumor class. Figure 1 illustrates 11 discriminatory gene patterns indexed by 539, 493, 11, 470, 460, 975, 326, 341, 112, 626, 996 corresponding to the NCI60 training dataset, and compared to original gene expression patterns. Clearly, the GM(1,1) method inflates the variance of observations and hence assess the smaller variability within a class. This is the reason why GM(1,1) is useful to form gene patterns as good candidates for molecular fingerprints in tumor classification. And we also believe that classification methods, by carefully choosing predictor genes used by a classifier from better quality gene expression patterns, will thus help classification analysis.

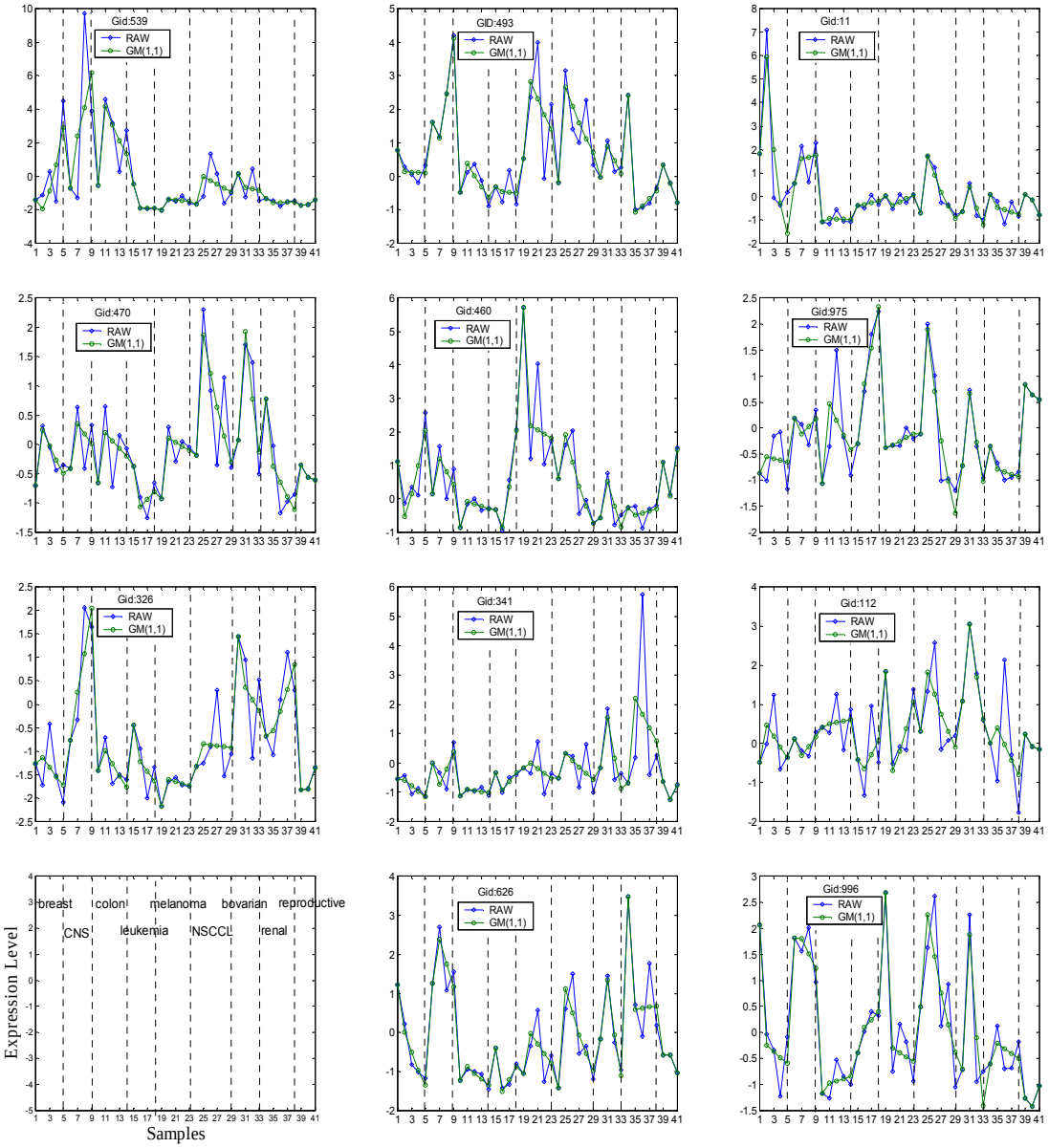

Figure 1. shows 11 gene expression patterns of the best predictor set selected fromt the GM(1,1)-GA/MLHD method and compares to original gene expression patterns. These patterns are gene expression levels composed of 41 samples: 5 breast, 4 central nervous system (CNS), 5 colon, 4 leukemia, 5 melanoma, 6 non-small-cell-lung-carcinoma (NSCLC), 4 ovarian, 5 renal, and 3 reproductive.

5. Conclusions

To obtain the ability to eliminate the variations in microarray data and find genetic fingerprints among various tumor classes, we propose using the GM(1,1)-GA/MLHD method for the classification problem of small-sample issues. Based on the success of our method, we conclude the main advantages of GM(1,1). First, GM(1,1) has the ability to smooth data variation by processing discrete numerical data into a pattern with less-noise, while the data in a class are not necessarily distributed normally. Secondly, the GM(1,1), representing a gene expression pattern shared by samples of the same tumor type, only needs a few samples to obtain better gene expression pattern. This is a reversal of traditional data mining techniques, where there are typically more samples than variables.

The work reported here is an expansion of paper Ooi et al. [3]. Our approach combined GM(1,1) and GA/MLHD methods, using the same procedures in classification on the same dataset, and exhibiting a 2.5-4% improvement in accuracy. In the multi-class classification scenario, the currently available datasets containing relatively few samples but a large number of variables make it difficult to demonstrate one method's superiority. While no methods have yet become the standard method to be adopted in this domain, we have shown GM(1,1)-treated gene expression patterns along with classifiers outperform classifiers without GM(1,1)'s. And finally, we anticipate that the use of GM(1,1) would be a helpful tool leading to practical uses of microarray data in cancer diagnosis.

References

[1] A.A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Tran, X. Yu, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, 403, pp.503–511, 2000.

[2] A. Ben-Dor, N. Friedman, and Z. Yakhini, "Scoring genes for relevance," Technical Report AGL-2000-13 Agilent Laboratories, 2000.

[3] C.H. Ooi, and P. Tan, "Genetic algorithms applied to multi-class prediction for the analysis of gene expression data," *Bioinformatics*, 19, pp.37–44, 2003.

[4] C.H. Yeang, S. Ramaswamy, P. Tamayo, S. Mukherjee, R.M. Rifkin, M. Angelo, M. Reich, E.S. Lander, J.P. Mesirov and T.R Golub, "Molecular classification of multiple tumor types," *Bioinformatics*, 17, pp.S316–S322, 2001.

[5] C. Romualdi, S. Campanaro, D. Campagna, B. Celegato, N. Cannata, S. Toppo, G. Valle, and G, Lanfranchi, "Pattern recognition in gene expression profiling using DNA array: a comparative study of different statistical methods applied to cancer classification," *Human Molecular Genetics*, vol. 12, No. 8, pp.823-836, 2003.

[6] D. Hanahan, and R. Weinberg, "The hallmark of cancer," *Cell*, 100, pp.57–71, 2000.

[7] D.T. Ross, U. Scherf, M.B. Eisen,, C.M. Perou, C. Rees, P. Spellman, V. Iyer, S.S. Jeffrey, M.V. de Rijn, M. Waltham, A. Pergamenschikov, J.C. Lee, D. Lashkari, D. Shalon, T.G. Myers, J.N. Weinstein, D. Botstein, and P. O. Brown, "Systematic variation in gene expression patterns in human cancer cell lines," *Nat. Genet*, 24, pp.227–235, 2000.

[8] J.L. Deng, "Control problems of grey system," *System and Control Letters*, 5, pp.288-294, 1982.

[9] J.L. Deng, "Introduction to Grey System Theory," *Journal of Grey System*, vol.1, pp.1-24, 1989.

[10] M. James, "*Classification Algorithms*," Wiley, New York, 1985.

[11] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. Yeang,, M. Angelo, C. Ladd, M. Reich,M, E. Latulippe, J. Mesirov, T. Poggio, et al., "Multiclass cancer diagnosis using tumor gene expression signatures," *Proc. Natl. Acad. Sci USA*, 98, pp.15149–15154, 2001.

[12] S. Dudoit, J. Fridly and T. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *JASA*, Berkeley Stat. Dept. Technical Report #576, 2000.

[13] S. Peng, Q. Xu, X. Ling, X. Peng, W. Du, L. Chen, "Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines," *FEBS Letters*, 555, pp.358-362, 2003.

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