Supplementary Information
Classification of Two and Multi-Class Cancer Data Reliably Using Multi-Objective Evolutionary Algorithms

Kalyanmoy Deb and A. Raji Reddy

Kanpur Genetic Algorithms Laboratory (KanGAL)
Indian Institute of Technology Kanpur
Kanpur, PIN 208 016, India
{deb,arreddy}@iitk.ac.in
http://www.iitk.ac.in/kangal

1 CLASS PREDICTION PROCEDURE

1.1 Preprocessing the Gene Expression Values

Most oftenly the DNA microarray data is arranged in a matrix $X_{m \times p}$ form, with rows for the m genes and columns for the p biological samples in which the gene expression has been monitored. Here, we represent x_i^j as the raw expression value of a gene i in a sample j. A row x_g is the *gene expression vector* of a gene g over all the samples. A column x^i is the *sample expression vector* of the expression values of all genes in a sample i. In most of the supervised learning algorithms, the available samples p in a data set is divided into two groups - one is training set $T_{m \times n}$ ($n \in p$) used for generating a classifier and the remaining samples are assigned to the test set. It is observed that, in most of the data sets these expression values are available in raw-form, hence these are need to be processed before starting the identification task. Let us briefly consider these steps here.

The top box in Figure 1 shows a sketch of gene expression values corresponding to a gene g for all training samples. These expression values are need to be modified according to a threshold parameter θ and a ceiling parameter ϕ, meaning that if any value in data set is less than θ or greater than ϕ is replaced with θ or ϕ, as the case may be. This process is expected to get rid of the problem of negative expression values in the data set. Thenceforth, we eliminate the genes with not wide variation in their expression values based on some thresholds on β and Ω ($\beta = max(x_g) - min(x_g)$ and $\Omega = \frac{max(x_g)}{min(x_g)}$ where x_g is raw expression value of gene g). For this purpose, we calculate the actual values for β and Ω for each gene and those genes violating the above conditions are excluded from the data set. Thereafter, the logarithm of the expression values are calculated and then normalized such that the mean of the expression values in row becomes 0 and the standard deviation becomes 1. If there are any missing values in a data set, they are imputated using simple k-nearest neighbor approach.

1.2 Two-Class Classification

In this work, we consider the weighted voting (WV) approach for predicting the class of any sample in a data set. In this approach, the class identification of any sample in a data set is completely depends on the *weighted votes* of a set of informative genes G and which in turn depends on the set of samples

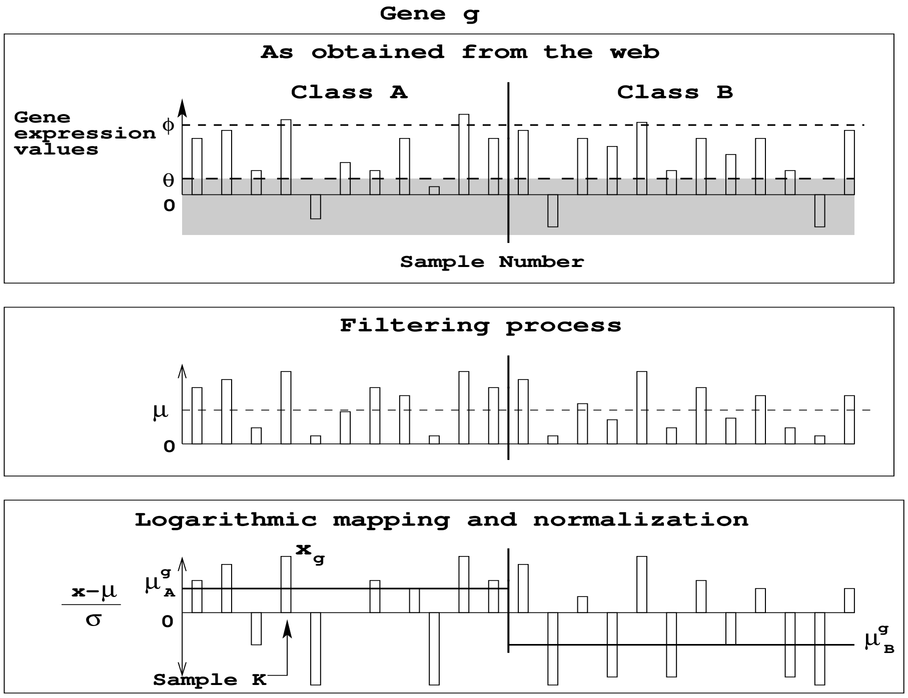

Fig. 1. Preprocessing of gene expression values for cancer class prediction.

S in a data set with known class labels. However, the identification of a set of informative genes is one of the crucial task in understanding the metabolic pathways in biological point-of-view and there exist not many studies for this purposes. Here, we have been observed that EAs in conjunction with WV approach are ideal candidates in identifying the most discriminative genes for reliable classification of cancer disease data.

In the WV approach, each such informative gene $g \in G$ casts a vote for one of the classes, the magnitude of each vote is dependent on the gene expression values in that sample ($\bar{x}_g^i$) and the degree of correlation with the class distinction among the available training samples in a data set. So, the magnitude of the vote for a sample i of gene g is defined as the product of the weighting factor and the weighted vote (mathematically, $S(i,g) = w_g^i \times v_g^i$), where w_g^i is a correlation coefficient or weighting factor that reflects how well the gene is correlated with the class distinction and v_g^i is a weighted vote which reflects the deviation of the expression level $\bar{x}_g^i$ from the average of μ_A^g and μ_B^g. This is illustrated in Figure 2. Thereafter, this process is repeated for all the genes ($g \in G$) and the votes are added to determine the winning class. The sample i is assigned to the class with the highest vote total. This concept is further elaborated in the next paragraph.

For a given gene subset G, the class of a sample i can be predicted with respect to a known set of S samples in the following manner. Let us say that S samples are composed of two subsets S_A and S_B, belong to class A and Class B respectively. For each gene $g \in G$, we first calculate the mean μ_A^g and the standard deviation σ_A^g of the normalized gene expression values of all S_A samples. The same is repeated for the class B samples and μ_B^g and σ_B^g are found. Thereafter, the class of i can be predicted by using

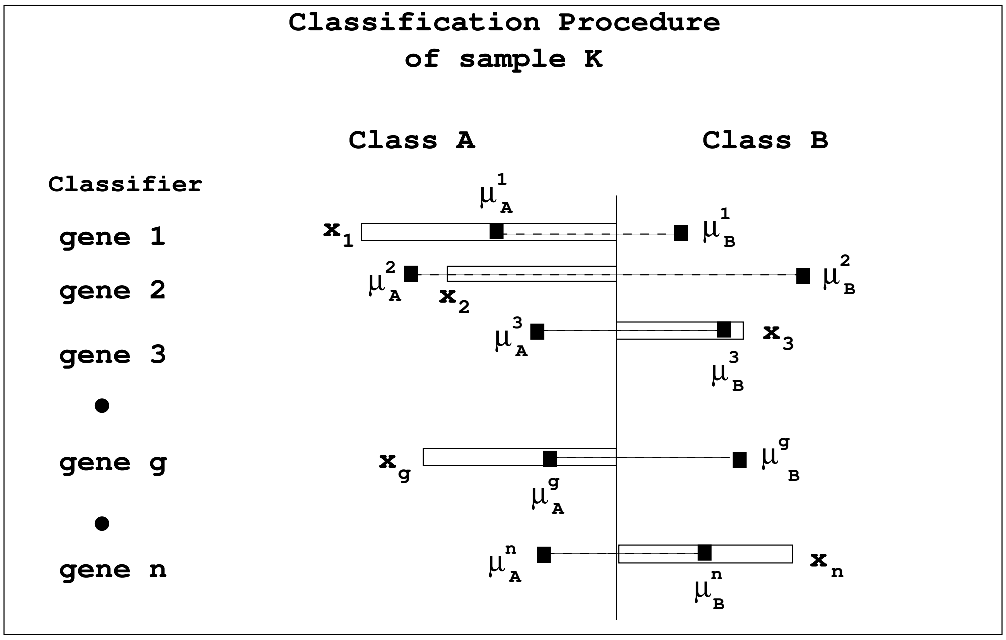

Fig. 2. The variation of the gene expression values from its corresponding means

a simple GS method proposed by Golub et al.[5, 4].

$$\text{class}(i) = \text{sign}\left\{ \sum_{g \in G} \left(\frac{\mu_A^g - \mu_B^g}{\sigma_A^g + \sigma_B^g} \right) \left(\bar{x}_g^i - \frac{\mu_A^g + \mu_B^g}{2} \right) \right\}, \tag{1}$$

where $\bar{x}_g^i$ is the normalized expression value of a gene g in a sample i. If the right term of the above equation is positive, the sample belongs to class A and if the quantity is negative, it belongs to class B. This concept is expected to be more elaborated with an example. For this purpose, we consider a 50-gene Leukemia data set with two different class of samples (ALL and AML). We have used 38 training samples for constructing the classifier. Thereafter, the statistic $S(i, g)$ (the term inside the summation in equation 1) values are calculated for each g of the 50 genes in two different class of the test samples (x and y) and they are plotted in Figure 3 and 4 respectively. For a sample x, the equation 1 finds the right side value is positive, thereby declaring x to be class A (ALL), and similarly for a sample y, the value is negative, thereby declaring y to be class B (AML). These predictions are matches with their actual classes. Hence, the samples x and y are said to be *correctly classified* samples.

1.3 Classification With Confidence

Although the above classification procedure is found to provide an efficient and reliable gene subsets with a reasonable classification accuracy. One of the difficulties with this approach is that the classification process is solely based on the sign of the right term in Equation 1, and is checked to predict the class of a sample (i.e., if the sign is positive the sample belongs to one particular class and if the sign is negative, it belongs to another class). So, the gene subset emerged from this approach may not truly reflect the class distinction among the available samples. This difficulty can be expected to bypass by introducing a

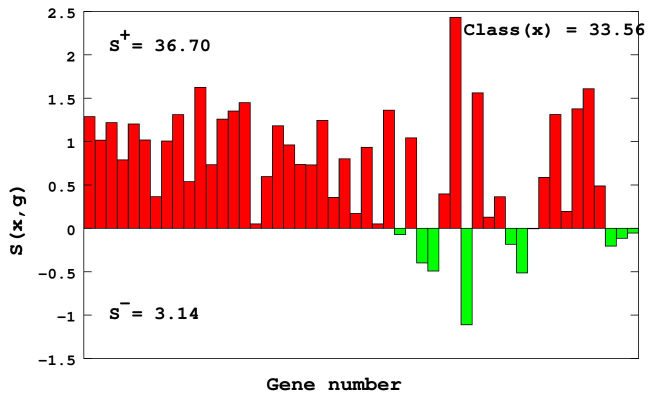

Fig. 3. WV approach: Class A sample.

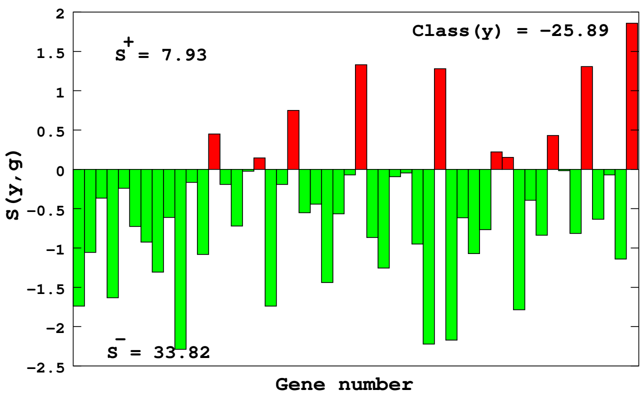

Fig. 4. WV approach: Class B sample.

threshold in prediction strength and also expected to predict the class of a sample with more confident. Although the term *classification with confidence* is not new, perhaps it has been used more frequently in many of the statistical algorithms [6]. In the current study, we investigate the concept of *classification with confidence* for the gene subset identification task and is described in lapidary manner with an example. The statistic $S(i, g)$ values are calculated for each g of the 50 genes in a particular Leukemia sample and they are plotted in Figure 5. For this sample, a correct prediction has been made.

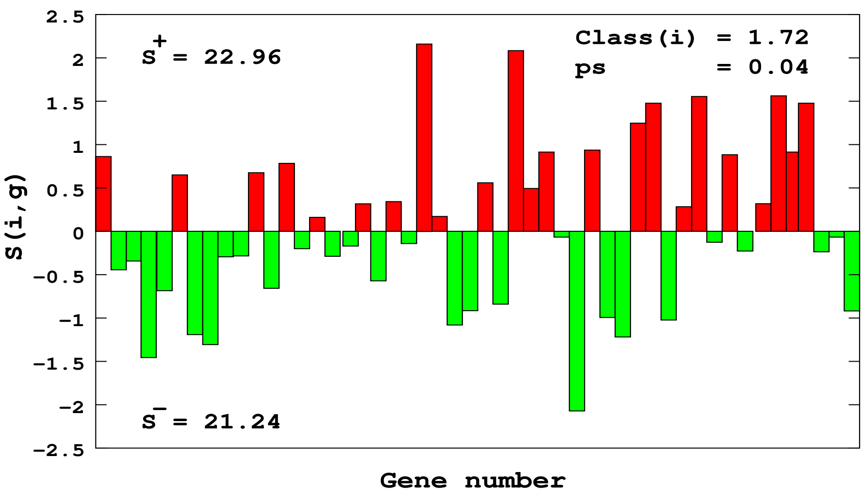

Fig. 5. The statistic $S(i, g)$ of a sample for 50-gene Leukemia data.

Although these 50 genes in a subset can predict the class of a sample to its belonging class. It can be seen from the Figure 5 that for 27 genes, negative values of $S(i, g)$ emerged, thereby classifying individually that the sample belongs to AML (class B), whereas only 23 genes detects the sample to be ALL (class A). Equation 1 finds the right side value to be 1.72, thereby declaring the sample to be an ALL sample and in fact which is correct. But it has been intuitively argued elsewhere [5] that a correct prediction with such a small strength does not allow to make the classification with a reasonable confidence.

For a more confident classification, one can fix a prediction strength threshold θ and modify the classification procedure accordingly. Let us say that the sum of the positive $S(i,g)$ values is S^A and the sum of the negative $S(i,g)$ values is S^B (in absolute). Then, the prediction strength (ps) is defined as follows:

$$ps = \left| \frac{S^A - S^B}{S^A + S^B} \right| \tag{2}$$

Thereafter, the prediction strength ps is compared with θ. If ps is more than θ, the classification is accepted, else the sample is considered to be undetermined for class A or B. For simplicity, in this study we consider these undetermined samples to be identical to the mismatched samples and include this sample to increment τ_{train} or τ_{test}, as the case may be. This way, a 100% correctly classified gene subset will ensure that the prediction strength is greater than θ for all the samples. Thus, the classification procedure with multiple training sets along with inclusion of a threshold on prediction strength is expected to be significantly aid in generating a generic classifier for reliable classification of cancer data sets.

1.4 Classification with Multiple Training Data Sets

In the previous sections and many other studies related to the machine learning, the available data set is partitioned into two groups: training and test sets. In these cases, one particular training set is considered during the learning phase and a classifier is developed. For the Leukemia samples, the training set (38 out of 72 samples) is already suggested and used in the literature, whereas for other two cases (Lymphoma and Colon samples), no particular training set is pre-specified. For these two cases, we had chosen a training set randomly from all available samples and the rest was declared as the test set. Although the developed classifier based on the above classification procedure (with only one training set) is expected to provide some insight into the distinguished genes. It can be observed and suggested in the literature that the developed classifier is completely dependent on the training data. So, to make the classification procedure more generic (perhaps, independent of the training data), the present study introduces the concept classification using multiple training sets (may be of the same size). Moreover, we investigate the importance of using multiple training sets for generating a better classifier.

Let us say there are total H (this study uses $H = 100$ in all the cases) different training sets, instead of one. For each case, these training samples are chosen randomly from all the available samples in a data set and rest of the samples assigned to the test set. Thus, for evaluating a classifier consisting of a few genes in a subset(G), follow the similar classification procedure used in the previous section for all H training sets independently. For each case j, find the number of mismatches (τ^j_{train} and τ^j_{test}) in both training and test samples. Thereafter, we take an average of these mismatches and calculate the two objective values as follows:

$$f_2 = \frac{1}{H}\sum_{j=1}^{H} \tau^j_{train}, \quad f_3 = \frac{1}{H}\sum_{j=1}^{H} \tau^j_{test}. \tag{3}$$

Since a number of training sets are used during the development of the classifier, apparently the obtained classifier is also expected to be more generic than those obtained in the previous sections. However, the development of such a classifier comes at the expense of more computational efforts. Since H different training cases are considered in the evaluation of one classifier, the computational time is also expected to increase by at least H times.

1.5 Multi-Class Classification

In the previous sections, we have examined classification procedure for predicting the class of two-class cancer samples. Although the above classification procedure can be extended to predict the class of multi-class (more than two classes) cancer samples, in this study, we have used one of the variants of above classification, the one-versus-all (OVA) binary pair-wise classification algorithm for predicting the class of the samples in a multi-class data set [2, 3, 6]. The class prediction procedure using OVA binary classification is described here.

For a given gene subset G, the class of a new sample i is predicted in the following manner. Let us say there are total J different tumor types in a data set such that these S samples are composed of J different subsets S_1, S_2, ..., S_j, ..., S_J, belonging to class 1, class 2,..., class j,..., Class J, respectively. The class of i is identified by assuming that the sample i belongs to one particular class j ($j \in J$). For each gene $g \in G$, we first calculate the mean μ_j^g and the standard deviation σ_j^g of the normalized gene expression values of all S_j samples. Similarly, calculate the mean $\mu_{\bar{j}}^g$ and the standard deviation $\sigma_{\bar{j}}^g$ of the remaining $J-1$ class of samples (S_1, S_2, ..., S_{j-1}, S_{j+1}, ...,S_J) together. Thereafter, the class of sample i can be predicted by modifying the above Equation 1 in the following manner.

$$\text{class}(i_j) = \text{sign}\left\{\sum_{g \in G}\left(\frac{\mu_j^g - \mu_{J-1}^g}{\sigma_j^g + \sigma_{J-1}^g}\right)\left(\bar{x}_g^i - \frac{\mu_j^g + \mu_{J-1}^g}{2}\right)\right\}, \tag{4}$$

If the sign of the above equation is positive the sample i belongs to class j, and if the sign is negative the sample i does not belong to that particular class. The above procedure is repeated for all j's ($j =$ 1,2,...,J). Hence, the multi-class classification problem can be thought of as a decomposition of a series of J one-versus-all (OVA) binary classification problems and each sample is tested for its belonging class against J different classes.

One of the difficulties with the above classification procedure is that the classification procedure is solely based on the sign of the right term of the Equation 4. Thus, it may possible to happen that the sign of the right term of the above equation is positive for more than one classes while predicting the class of a sample i, which ultimately assigns more than one classes for only one sample. One of the easiest way to overcome this difficulty is that the sample i is assigned to the class with the maximum value of $class(i_j)$ (out of all j's). In this study, we have used a more stringent classification procedure for generating a generic classifier for reliable classification of multi-class cancer data. The description of the method is as follows. Let us say, we want to predict the class of a sample i. For this purpose, we calculate the prediction strength (ps) for each j and then find the maximum value (let us say ps_{max}). If ps_{max} is greater than the prediction strength threshold θ, then assign sample i to that particular class, else the sample is added to the mismatch counters τ_{train} or τ_{test} as the case may be. Thereafter, the predicted class of a sample is compared with its actual class. If the predicted class matches with the actual class, then the sample is said to be matched one, else the sample is added to the mismatch counters.

1.6 Cross Validation

Cross validation is a method used for estimating the *generalization error* based on re-sampling [1]. The resulting estimates of the *generalization error* are often used for choosing a best model among various models. In cross validation, the training data set (let say $T_{m \times n}$) is partitioned into k subsets, C_1, C_2, ..., C_k (k is also known as the number of cross validation trails). The number of samples in each subset is kept roughly of the same size. Then a classifier is constructed using $T_i = T - C_i$ samples to predict the class of the samples in C_i. Once the classifier is constructed using T_i samples, each of the sample in C_i is tested for its belonging class A or B using the above classifier. Since these C_i samples are subset of the T training samples, we can compare the predicted class of the C_i samples with the actual class. If there is a mismatch, we increment the training sample mismatch counter τ_{train} by one. This procedure is repeated for all C_i samples in the i^{th} subset. Thereafter, this procedure is repeated for all k subsets and the overall training sample mismatch counter τ_{train} is noted. The advantage of this method is that it matters less how the data gets divided, the classifier is tested on each sample exactly once, the training set for each trail is $(\frac{k-1}{k})$ of the original training set (mathematically, $T_i = \frac{k-1}{k} \times T$). Thus, for large k, we get a relatively unbiased estimate of the classifier for a given training set ($T_{m \times n}$). So, the variance of the resulting estimate is reduced by increasing the value of k. The disadvantage of this method is that the training algorithm has to be rerun from scratch k times, which means it takes k times as much computation to make an evaluation. One of the most commonly used method in cross validation is leave-one-out-cross-validation (LOOCV), in which only one sample in the training set is withheld and the classifier is constructed using the rest of the samples to predict the class of the withhee current study uses the LOOCV to estimate the number of mismatches in the training set. The classifier obtained from all T training samples is used to predict the class of each test sample in the independent test set. The number of mismatches τ_{test} obtained by comparing the predicted class with the actual class of each sample is noted.

References

1. Ben-Dor A., Bruhn A., Friedman N., Nachman I., Schummer M., and Yakhini Z. Tissue classification with gene expression profiles. *Journal of Computational Biology*, 7:559–583, 2000.
2. Allwein. Reducing multiclass to binary: A unifying approach for margin classifiers. In *Proceedings of ICML 2000*, pages 337–348, 2000.
3. Bose R. C. and Ray-Chaudhari D. K. On a class of error correcting binary group codes. *Information Control*, 2:68–79, 1960.
4. Slonim D.K., Tamayo P., Mesirov J.P., Golub T.R., and Lander E.S. Class prediction and discovery using gene expression data. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology (RECOMB)*, pages 263–272, 2000.
5. Golub T. R., Slonim D. K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J. P., Coller H., Loh M. L., Downing J. R., Caligiuri M. A., Bloomfield C. D., and Lander E. S. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

6. Ramaswamy S., Tamayo P., Rifkin R., Mukherjee S., Yeang C. H., Angelo M., Ladd C., Reich M., Latulippe E., Mesirov J. P., Poggio T., Gerald M., Loda M., Lander E. S., and Golub T. R. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Science*, 98(26):15149–15154, 2001.