Optimal Genes Selection with a New Multi-objective Evolutional Algorithm Hybriding NSGA-II with EDA

Luo Fei
School of Computer
Wuhan University
Wuhan, China
luofei_whu@hotmail.com

Liu Juan
School of Computer
Wuhan University
Wuhan, China
liujuanjp@163.com

Abstract

Recent studies on molecular level classification of tissues with DNA microarray technology have produced remarkable results. It is believed that the subtypes of cancer can be distinguished by a set of discriminative genes. To achieve this goal, it not only requires high enough classification accuracy, but also a minimal number of genes as much as possible to lower cost. Meanwhile, the number of samples from different tissues may differ greatly. Therefore, it should also avoid classification bias due to unbalance sample number in different classes. In this paper, we propose a new multi-objective evolutional algorithm (MOEA) framework to select optimal genes, which has both advantages of the non-dominated sorting genetic algorithm II (NSGA-II) and the estimation of distribution algorithm (EDA). Finally, experiment on the data is done, The result shows that our method has good performance.

1. Introduction

The process of transcribing a gene's sequence into RNA that serves as the template for protein production is know as gene expression. A DNA microarray (also commonly known as gene or genome chip, DNA chip, or gene array) is a collection of microscopic DNA spots attached to a solid surface forming an array for the purpose of monitoring expression levels for thousands of genes simultaneously, or for comparative genomic hybridization. With this promising technology, much biological knowledge can be obtained from the gene expression profiling, such as understanding cellular processes, discovering gene functions, inferring gene regulatory network, and searching drug targets. In addition, DNA microarray can also be used as cancer diagnosis and classification platform at molecular level [1] [2]. Traditional ways for cancer diagnosis have been primarily based on morphological appearance of the tumor, but this has serious limitations. While the histopathology of two cells may appear very similar, their clinical behavior, such as their response to drugs can be drastically different. Since transcriptional changes accurately reflect the status of disease including cancers, gene expression profiling can be used to classify subtypes of cancers. Although DNA microarray technology provides a new approach for cancer diagnosis and classification, there are still some challenges in practice. One of them is the curse of dimensions that genes are much more than samples. Many of them are irrelevant to the distinction of different tissues types and thus add noise in the classification process which draws out the contributions of relevant genes. Another challenge is that DNA microarray data contain technical and biological noises. The technical noise can be produced at some stages of the experiment, such as production of the DNA array, preparation of the samples, hybridization between cDNA and array, and signal analysis and extraction of the hybridization results. The biological noise can come from non-uniform genetic background of the samples being compared, or from the impurity or misclassification of tissue samples. Therefore, optimal genes selection is crucial for the cancer diagnosis and classification.

The main purpose of optimal genes selection is to find out the subset of genes most suitable for differentiating classes. To achieve the best possible performance with a particular classifier on a particular domain, gene subset selection methods should consider how the classifier and the training data interact. Optimality of a gene subset should only be defined with respect to a particular classifier and only be guaranteed by exhaustive search of all possible subsets. However, this way is infeasible for the large gene sets due to the combinational explosion of the number of subsets. Simple gene-ranking technique may substitute to perform gene subset selection [3] [4]. A fixed number of top ranked genes may be selected for further analysis or to design a classifier. Alternatively, a threshold can be set on the ranking criterion. Only the genes whose criterion exceeds the threshold are retained. However,

978-0-7695-3118-2/08 $25.00 © 2008 IEEE
DOI 10.1109/BMEI.2008.313

these methods based on the index-ranking do not find out compact gene subsets because the genes with high rank usually may be redundant and complementary genes that individually do not separate well the data are missed without taking the interrelationships between genes into account. In another point of view, based on whether or not the selection is done independently of the classification procedure, selection algorithms can be categorized into two types, filter approaches and wrapper approaches. The former is usually used as a preprocessing step independent of the classifier to filter out some irrelevant genes based on some criterions, while the later searches a good subset using classification algorithm itself as part of evaluation function.

Determination of the number of genes included in the discrimination subset seems dilemmatic. Including too few genes will not discriminate in a detailed enough manner to classify test data correctly. Having too many genes is not optimal either, as many of the genes are largely irrelevant to the diagnosis and mostly have the effect of adding noise, decreasing the 'information criterion'. Meanwhile, in the implementation of multi-classes discrimination, for datasets with many samples for some classes and few for others, relatively high accuracy may be achieved simply by labeling unknown samples according to the largest class，which may lead to high test error for the classes with less samples. Therefore, it should also avoid classification bias due to unbalance sample number in different classes.

Based on the analysis above, the conflictions among high classification accuracy, the small number of genes included in the discriminative set, and balance accuracy among different classes determine that optimal genes selection is a suitable candidate for the multi-objective modeling. Meanwhile, exhaustive search of all possible gene subsets is impossible. Evolutionary Algorithm (EA) maintains a population of solutions is less likely to be restricted by interdependencies among genes and may speedily perform efficient searches in high dimensional spaces. Thus EA offers a particularly attractive approach to the gene selection problem. Therefore, in this paper, we propose a new Multi-objective evolutional algorithm (MOEA) framework to select optimal genes, which has both advantages of the non-dominated sorting genetic algorithm II (NSGA-II) and the estimation of distribution algorithm (EDA). In order to achieve both the computation efficiency and the final optimal gene subset, we first use filter strategy to decrease the amount of initial gene dataset and then use wrapper strategy to get the final gene subset. The rest paper is organized as following. Part II is the introduction of multi-objective evolutional algorithm. Part III is our proposed MOEA hybriding NSGA-II with EDA. An experiment based on SRBCT is done to evaluate our method in part IV. Finally the conclusion is draw in part V.

2. Multi-objective evolutional algorithm

In the sense of gene subset size minimization and performance maximization, gene selection can be viewed as a multi-objective optimization problem. Formally, each gene subset s_i is associated with a vector evaluation function

$$F(s_i) = (F_1(s_i), ..., F_m(s_i)) \quad (1)$$

Where m is the number of objectives, F_k is the kth objective. The task of the multi-objective optimization involves simultaneous optimization of all of the F_k objectives. In this work, we use 3 heuristic fitness criteria: the misclassification rate of the classifier, the difference in error rate among classes, and the size of the subset, which are denoted as F_1, F_2, and F_3 separately. Assume there are c classes in the considered domain, n samples are used to evaluate the objectives, and e_i ($i \in 1, \ldots, c$) is the error rate on the ith class of the classifier, we define

$$F_1 = \frac{n_{err}}{n}, F_2 = \frac{2\sqrt{\sum_{i=1}^{c}\sum_{j=i+1}^{c}(e_i - e_j)^2}}{c(c-1)}, F_3 = \frac{Gene_{sub}}{Gene_{total}} \quad (2)$$

The meanings of F_1 and F_3 are obvious; they are two main objectives of the gene selection. F_2 is used to avoid classification bias due to unbalanced test samples in different classes. For example, if a dataset consists of 98 samples from class A and only 2 from class B, the F_2 prevents EA from training classifier to always predict class A and thus achieve 2% apparent error rate.

Unlike the single optimization problem, the objectives in the multi-objective optimization are usually conflicted. How to evaluate solutions is every important. There are three kinds of fitness assignment strategies in MOEA: aggregation methods, population-based non-Pareto methods and Pareto-based methods. Aggregation methods combine the objectives into a higher scalar function that is used for fitness calculation. In fact, it converts the multi-objective optimization to single objective one. In population based non-Pareto methods, the search is guided in several directions at the same time by changing the selection criterion during the reproduction phase. The main idea of the Pareto-based MOEAs is that non-dominated solutions in a population have the advantage of being survivors. Pareto-based approaches explicitly use the Pareto dominance in order to determine the reproduction probability of each individual. For usual problems, the Pareto-based

MOEAs are considered to be superior to non-Pareto-based MOEAs at least to the extent of searching Pareto optima.

2.1 NSGA-II

Over the past decade, a number of multi-objective evolutionary algorithms (MOEA) have been suggested such as MOGA [5], NPGA [6] and NSGA-II [7].These algorithms demonstrated the necessary additional operators for converting a simple EA to a MOEA. Two common features on all these operators were the following: i) assigning fitness to population members based on non-dominated sorting and ii) preserving diversity among solutions of the same non-dominated front. Among the existing MOEAs, NSGA-II is well study and widely used, which is characterized by the use of the three characteristics while generating the optimal solution.

2.1.1 Non-domination One solution s_i is said to dominate another solution s_j iff:

$$\forall k \in \{1,2,...,m\}, F_k(s_i) \succeq F_k(s_j) \text{ and } \exists k', F_{k'}(s_i) \succ F_{k'}(s_j)$$

$\succeq$ Means better or equal, $\succ$ means better. Otherwise the two solutions are non-dominated to each other.

2.1.2 Distance diversity estimation Because there may be several solutions in the same non-dominated front, in order to preserver their diversity in the population, a measure called crowding distance is used in NSGA-II. The crowding distance computation requires sorting the population according to each objective function value in ascending order of magnitude. Then this assigns the highest value to the boundary solutions and the average distance of two solutions [(i + 1)th and (i − 1)th] on either side of solution i along each of the objectives.

2.1.3 Crowded tournament selection Crowded tournament selection operator is defined as follows. A solution s_i wins tournament with another solution s_j if any one of the following is true:
a) Solution s_i has better non-dominated rank. b) Both the solutions are in the same non-dominated rank, but solution s_i is less densely located in the search space.

The detailed procedures of finding non-dominated set, distance diversity estimation and crowded tournament selection can be found in [8].

3. Hybridizing NSGA-II with EDA to select genes

Estimation of distribution algorithms (EDAs) are a class of novel stochastic optimization algorithms, which have recently become a hot topic in field of evolutionary computation. EDAs acquire solutions by statistically learning and sampling the probability distribution of the best individuals of the population at each iteration of the algorithm. EDAs have introduced a new paradigm of evolutionary computation without using conventional operators such as crossover and mutation. According to the complexity of probability models for learning the interdependencies between the variables from the selected individuals, EDAs can be categorized to dependency-free, bivariate dependencies and multivariate dependencies.

3.1 Comparison of NSGA-II and EDA

The same with other general EAs, NSGA-II uses traditional variation operators to generate the offspring population. Assignment of parameters, like replacement rate, crossover rate, and mutation rate, determines the final performance of the algorithm. Setting proper parameters is not easy for inexperienced users. On the other hand, neglect of the relations between variables in an individual may lead to blindness of searching for optimal individuals. These two main shortcomings of NSGA-II motivate us to combine NSGA-II and EDA, whose advantages just make up shortcomings of NSGA-II. Compared to the rest of EAs including NSGA-II, the characteristic of EDAs is that EDAs replace the application of variation operators in order to generate the next population from the current one at each iteration by learning and subsequent simulation of a joint probability distribution for those individuals selected from the current population by means of the selection method. It results in two important advantages of EDAs over NSGA-II: The sometimes necessary design of variation operators tailored to the particular optimization problem at hand is avoided, and the number of parameters to be assessed by the user is reduced. A further advantage of EDAs is that the relationships between the random variables can be explicitly expressed through the joint probability distribution learnt from them, instead of being implicitly kept by the individuals of successive populations as building blocks.

3.2 Hybriding NSGA-II with EDA to Selection Genes

3.2.1 Initial gene pool. Because the number of genes in the dataset is always huge and some of them are not useful for classification, it is necessary to narrow down genes from many thousand down to the order of 10^2 to speed up searching with MOEAs which may cost a lot of running time. There are many filter approaches can do this. Instead of using principle component analysis

(PCA), which is most common way, we use the ratio of between-groups to within-groups sum of squares (BSS/WSS) given by Dudoit[9].

$$Rank(x) = \frac{BSS(x)}{WSS(x)} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{c} I(y_i = j)(\bar{x}_{\cdot j} - \bar{x})^2}{\sum_{i=1}^{n}\sum_{j=1}^{c} I(y_i = j)(x_i - \bar{x}_{\cdot j})^2} \quad (3)$$

Here, for gene x, x_i is the expression value on the ith sample, n is the number of samples and c is the number of sample types, $\bar{x}$ is the average expression value in all n samples, $\bar{x}_{\cdot j}$ is the average expression value in samples belonging to class j. $I(y_i = j)$ is a judging function, whose value equals one when the class label of sample is j, otherwise it returns zero. Rank(x) calculates the ratio of between-groups to within-groups sum of squares and its value will be big if gene x can perfectly distinguish the c types on all n samples.

In order not to miss discriminative genes as much as possible, we consider to calculate genes' BSS/WSS rank in the conditions of all distinct combinations of c sample types. For each combination, best S genes will be reserved in the gene pool. For example, supposing the total number of sample classes is 3 and donated as 1, 2 and 3, all distinct combinations are {1,2},{2,3},{1,3}and {1,2,3}, thus genes' BSS/WSS rank will be calculated in these four conditions. In that way, those genes which can separate only some sample types are able to be kept. After all, this is a preprocessing step and an optimal gene subset may be composed of complementary genes that individually do not separate well the data.

3.2.2 Hybriding NSGA-II with EDA. Given a gene pool ζ ,each possible optimal gene subset is represented by an individual with fixed sized binary string whose length equals the number of genes included in ζ . If the value of the ith bit in the individual is one, it means that the corresponding gene i in ζ will be included in the optimal gene subset. The fitness functions used in our method is given by the formula (2).The new hybrid MOEA for optimal gene subset selection is shown in table I.

TABLE I. HYBRID MOEA FOR OPTIMAL GENE SUBSET SELECTION

1.Generate a population PO_u composed of Q uniformly generated individuals. u=1 .
2. **while** the stopping condition is not met **do**
3. Calculate the multi-objective fitness functions .
4. Rank the population using dominance criteria .
5. Calculate crowding distance.
6. Perform crowding tournament selection to generate learning dataset D_u , | D_u |=M<Q.
7. Perform EDA to generate offspring population O_u by learning D_u . | O_u |=Q.
8. Combine PO_u and offspring O_u
9. Rank the mixed population .
10. Calculate the crowding distance using dominance criteria.
11. Generate PO_{u+1} with the best Q members of the combined population PO_u and offspring O_u by crowding tournament selection .
12. u++ .
13.**Return** the best individuals found so far .

Steps From 1 to 5 are standard operators in the NSGA-II. After the step 5, a set of best individuals D_u with high fitness value and good diversity are prepared for learning dataset used in EDA. EDA constructs probabilistic model and a joint probability distribution $p_u(X)$ for $X = (x_1, \ldots, x_{|\zeta|})$ is induced. Here we use the algorithm UMDA (univariate marginal distribution algorithm) to construct probabilistic model. The joint probability distribution $p_u(X)$ is induced by

$$p_u(X) = p_u(X \mid D_u) = \prod_{i=1}^{|\zeta|} p_u(x_i) = \prod_{i=1}^{|\zeta|} \frac{\sum_{j=1}^{M} \delta_j(x_i = a_i \mid D_u)}{M} \quad (4)$$

$$\delta_j(x_i = a_i \mid D_u) = \begin{cases} 1, if\ x_i = a_i \\ 0, otherwise \end{cases}$$

Here, a_i is the value of the ith bit, which equals one or zero. M is the number of individuals included in the D_u. After the joint probability distribution $p_u(X)$ is available, offspring O_u with Q individuals is generated by randomly sampling with the joint probability $p_u(X)$. Finally according to crowded tournament selection, select best Q individuals as next population PO_{u+1} from the combination of PO_u and O_u. The stop condition can be specified by users freely. One common way is to set the iteration times.

4. Experiment

We code our method with VC++ 6.0 and Matlab and evaluate it on the dataset of SRBCTs (the small, round blue-cell tumors) published by Khan in [10]. The SRBCTs of childhood, which includes four sample types of neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL) and the Ewing family of tumors (EWS), are so named because of their similar appearance on routine histology. The dataset is a collection of 2308 genes whose red intensity more than

20 in 88 samples. For the classifier module, we use KNN, which has such advantages that easy realization, high running speed and being suitable for datasets with multi-modal classes. The same with khan's work, use 63 samples as the training set, and 25 samples as the test set. 2308 genes in the training data are first preprocessed by BSS/WSS. For some top genes may repeatedly appear in some combination situations, the number of the final gene pool is usually less than $S* 2^{(k-1)}$. Here we donate the top genes number S as 30 and Q as 50. Thereafter, as for our MOEA-Classifier component, the training data is randomly shuffled and generate two subsets, one subset containing 70% of the samples for constructing the classifier and the other 30% of the samples serving as tuning set to evaluate the classifier and then calculate the fitness values of individuals. During generating offspring at each iteration, best genes M=10 selected by crowding tournament selection to establish learning dataset D_u. After 10 runs, we use leave one out cross validation (LOOCV) to estimated the predication accuracy of the classifier built by our selected optimal gene subset, in which one sample in the training set is withheld, the remaining samples of the training set are used to build a classifier to predicate the class of the withheld sample, and the cumulative error rate is calculated.

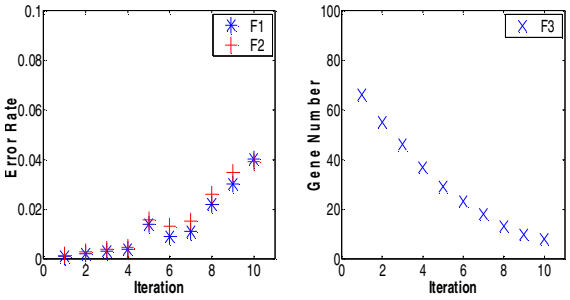

Figure 1. The fitness average values of population in the 10 iteration.

Figure.1 is the three fitness average values of populations in the 10 iterations. We can see the method decreases the number of genes under the acceptable accuracy. Finally, we find the best soluteion whose objectives F_1 is 0, F_2 is 0, F_3 is 7 and LOOCV accuracy is 0.048. With these seven genes to predict the 25 samples in the test dataset, 100% accuracy is acquired except for the normal samples.

5. Conclusion

In this paper, we proposed a new MOEA framework hybridizing NSGA-II with EDA to select optimal genes. In the preprocessing step we narrow down genes' number in a filter way, and sequentially use our new MOEA method to find optimal gene subsets in a wrapper way. The experiment proves its good performance. In future, we will discuss the combination of NSGA-II with EDA in bivariate dependencies, and multivariate dependencies situations and do experiments on more datasets.

6. Acknowledgement

This work is supported by National Nature Science Foundation of China grant by NO. 60773010.

7. References

[1] Golub T.R., Slonim D.K., et al. Molecular Classification of Cancer:Class Discovery and Class Prediction by Gene Expression Monitoring, Science, 286(15):531-537,1999.

[2] Tibshiranit, R., Hastie, T., Narasimhan, B., Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Nat. Acad, 6567–6572 . Sci. 99. 2002

[3] Ben-Dor A,, Bruhn L. et al., Tissue Classification with Gene Expression Profiles, Journal of Computational Biology, 7(3/4):559- 583, 2000.

[4] Park P.J et al. , A Nonparametric Scoring Algorithm for Identifying Informative Genes from Microarray Data. PSB2001,6:52-63,2001

[5] C. M. Fonseca and P. J. Fleming, Genetic algorithms for multiobjective optimization: Formulation, discussion and generalization,in Proceedings of the Fifth International Conference on Genetic Algorithms, pp. 416–423. 1993

[6] J. Horn, N. Nafploitis, and D. E. Goldberg, A niched Pareto genetic algorithm for multiobjective optimization, in Proceedings of the First IEEE Conference on Evolutionary Computation, Z. Michalewicz,Ed. Piscataway, NJ: IEEE Press, , pp. 82–87. 1994.

[7] K Deb, A Pratap, S Agarwal, T Meyarivan,A fast and elitist multiobjective genetic algorithm: NSGA-II,IEEE Transactions On Evolutionary Computation, Vol. 6, No. 2, 2002

[8] S Mitra, H Banka, Multi-objective evolutionary biclustering of gene expression data ,Pattern Recognition, 2006

[9] Dudoit, S., Fridlyand, J., Speed, P.,. Comparison of discrimination methods for classification of tumors using gene expression data. J. Amer. Statist. Assoc. 97, 77–87. 2002

[10] Khan, J., Wei, J. S., Ringner, M., et al..Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med. 7, 673-679.2001