A Genetic Programming Ensemble Approach to Cancer Microarray Data Classification

Supoj Hengpraprohm [a] and Prabhas Chongstitvatana [b]
[a] *Faculty of Science and Technology, Nakhon Pathom Rajabhat University, Nakhon Pathom, Thailand.*
[b] *Department of Computer Engineering, Chulalongkorn University, Bangkok, Thailand.*
supojn@yahoo.com, prabhas@chula.ac.th

Abstract

This paper presents a method for building an ensemble of classifiers for cancer microarray data. The proposed method exploits the advantage of a clustering technique, namely K-means clustering, combined with a feature selection technique, namely SNR feature selection. An evolutionary algorithm, namely Genetic Programming, is used to construct a number of classifiers which are assembled into an ensemble. The performance of the proposed method was tested on six cancer microarray data sets. The experimental results indicate that the proposed method yields a good prediction accuracy with a small standard deviation.

1. Introduction

Cancer, a generic term for a group of more than 100 diseases, is a leading cause of death worldwide. In 2005, cancer accounts for 13% of all deaths in the world [1]. Although there are a number of researchers that have been studying how to specify and identify these tumors for many years, there is no any efficient method for diagnosis or treatment of malignancy. The emergence of the microarray technology allows us to investigate an organism in the molecular level for many thousands of genes simultaneously and understand the mechanisms of life in details [2].

The microarray data consist of a small and high dimensional data. It is very complex and difficult to analyze. There are many methods to analyze such data, for examples clustering, classification and feature reduction [3].

Classification is a major challenging problem for cancer microarray data analysis that aims to identify the presence of cancer or to distinguish among specific cancer. The objective of this task is to maximize the classification accuracy. The accuracy is very important issue for diagnosis and treatment of patient with tumor. There are many ways to improve the accuracy of prediction such as feature selection and ensemble approaches [4-8].

Recently, the ensemble methods have been proposed as a mean to improve the classification accuracy. A number of classifiers which are different are used together and their outputs are combined to give the result. The performance of the ensemble approaches is significantly better than using any single classifier [5-7].

Genetic Programming [9] is an algorithm in machine learning that is widely used in various problem domains including cancer microarray data classification. A number of researches reported that classification by means of genetic programming provide a good result of prediction accuracy, see [8] for example.

In this paper, we propose an ensemble approach to build a genetic programming classifier for cancer microarray data. To build an ensemble, we employ data clustering combined with feature selection and promote diversity among many classifiers. The data clustering is a crucial step in our proposal. It determines that different classifiers will use different features. Therefore, there will be many classifiers that have "different point of views". The method was tested on 6 cancer microarray data sets: Lymphoma, Ovarian, Colon, Prostate, Leukemia and Lung data sets.

The rest of paper is organized as follows. Section 2 presents the classification task with Genetic Programming. Section 3 describes the method implemented for the ensemble in this research. Section 4 shows the results of the experiment, and the discussion is presented in Section 5.

2. Classification by means of genetic programming

Genetic Programming [9] is a search method that imitates natural evolution. It is developed from Genetic Algorithms [10] and is differed by the way the solution is represented in a tree structure instead of a fixed length binary string. The solution comprises of nodes from a function set and a terminal set. A function set is a set of operators designed for the problem such as arithmetic operators, logical operators, and functions. A terminal set is a set of operands of functions such as constants and variables. The algorithm of Genetic Programming is shown in Figure 1.

1. *Generate an initial population of solutions.*
2. *Evaluate each solution by a fitness function.*
3. *If the termination criterion is not met*
   3.1 Create a new population by genetic operators
   - *Reproduction*
   - *Crossover*
   - *Mutation*

   3.2 Go to 2.
4. *Return the solution with the best fitness value.*

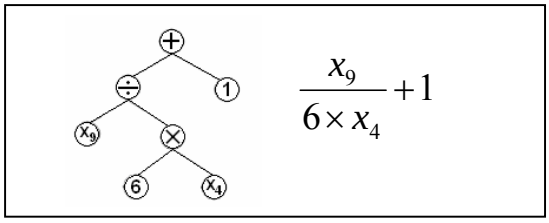
Figure 1. The algorithm of Genetic Programming

In a classification task, a solution of Genetic Programming is represented by a classification tree. The tree represented an arithmetic expression or logical expression (in this research we used the arithmetic expression as shown in Figure 2). The tree consists of symbols from the function set F and the terminal set T. In our experiment, the function set F comprises of arithmetic operators and the terminal set T comprises of 10 constants and a number of variables defined as follows: $F = \{+, -, \times, \div\}$ and $T = \{ 0..9, x_1.. x_n \}$. The variables represent the value of the selected attributes.

To evaluate the fitness of a candidate, its expression is evaluated. The variables ($x_1.. x_n$) are data from the microarray data. If the result of evaluating an expression is more than 0, it is classified as Class 1. Otherwise it is classified as Class 2. An expression is evaluated with data from the training set. The total number of the correct classification, C, is counted as the fitness value of the expression. The term 1/size is included as a penalty for the solution that has a large expression and to encourage a compact solution (Eq. 1). The higher fitness value indicates the better solution. The fitness function is defined as follows:

$$fitness = C + \frac{1}{Size} \qquad (1)$$

The population of solutions is evolved by means of genetic operations. Three genetic operators: reproduction, crossover and mutation are used. They sample the current population and generate offspring which become the next generation. A description of the working details of these operators can be found in [8].



$$\frac{x_9}{6 \times x_4} + 1$$

Figure 2. (left) A tree represented an arithmetic equation. (right) The expression derived from the tree

Step 2 and 3 in Figure 1 are repeated until the termination criteria are met. A run is limited to the maximum number of generations. Throughout generations, the quality of solutions is improved. The result from each run is different as the search for a solution is probabilistic and the solution for this problem is not unique. We denote classification by means of Genetic Programming "GPC" (Genetic Programming Classifier). The parameters used in the experiment are shown in Table 1.

Table 1. The parameters used in Genetic Programming Classifier

Population Size	1,000
Maximum Size of Tree	500
Maximum number of Generation	500
Reproduction Rate	10%
Crossover Rate	80%
Mutation Rate	10%
Termination Criteria: Correctly classify the training data 100% or exceed the maximum number of generations	

3. Method Implemented for The Proposed Ensemble Approach

In the ensemble approach, the key of the success is the variation in the member of ensemble. Each member must provide a good performance. To create the diversity among classifiers, the proposed method selects and distributes different sets of features to different classifiers. Firstly, the features are clustered by K-means clustering method (see [11] for the details of the algorithm). Then, the features are selected by applying SNR ranking. The algorithmic description of our method is as follows:

1. Cluster gene with K-means clustering. 30 clusters are used. The maximum number of iteration is 10.

2. Apply SNR ranking (Eq.2) to select features in each cluster.

$$F = \left| \frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2} \right| \qquad (2)$$

where μ_1 and μ_2 denote the mean expression level for the samples in Class 1 and Class 2 respectively. σ_1 and σ_2 denote the standard deviation for the samples in each class.

3. For each cluster in the emsemble i, select the i^{th} best feature with SNR score of each cluster to form a set of features, S_i, and train the i^{th} genetic programming classifier GPC_i by S_i.
4. Combine the outputs of all classifiers using the weighted voting approach where each GPC_i votes for its output with w_i.

$$w_i = \frac{T_i \times E_i}{\sum_i (T_i \times E_i)} \qquad (3)$$

where E_i be the training error of GPC_i and T_i be the sum of SNR score of S_i. The class with the highest vote is the output of the ensemble.

Using this method, each set of features S_i comprises of informative genes but it differs from the other sets. The learning algorithm like GPC will be trained with difference useful knowledge. As a result, there is diversity among the member of the ensemble of GPC because of the different set of features.

4. Experimental Results

Six data sets of cancer microarray data from Bio-medical Data Analysis web site [12] are used to test the proposed method. The details of each data set are shown in Table 2.

To evaluate the performance of an ensemble of classifiers, a 10-Fold cross validation method is used. There are N records of data. The records are divided into 10 subgroups with randomly chosen members (without replacement). Nine subgroups are used as training set and the rest subgroup is used as a test set. We exchange a test set of data through all subgroups and evaluate an expression in terms of its accuracy defined as follows:

$$Accuracy = \frac{TP + TN}{N} \qquad (4)$$

where N is the total number of test cases, TP is a total number of affected subjects correctly classified, TN is a total number of normal subjects correctly classified, and $TP+TN$ is the total number of subjects correctly classified.

Table 2. The details of data set used in this work

Data Set	No. of Gene	No. of Instance (Class)
Leukemia	7,129	38 (27 ALLs, |11 AMLs)
Colon Cancer	2,000	62 (40 cancers, 22 normals)
Ovarian Cancer	15,154	253 (162 cancers, 91 normals)
Prostate Cancer	12,600	102 (52 cancers, 50 normal)
Lung Cancer	12,533	32 (16 MPMs, 16 ADCAs)
Lymphoma	4,026	47 (24 GCBs, 23 ACBs)

The proposed method (denoted ClusSNR) are compared with two other ensembles obtained from conventional methods. The first method uses all genes to build the ensemble (denoted Reg). The second method uses SNR feature selection alone with the best 30 features by SNR score (denoted SNR). We also compare our method with two well-known ensemble methods such as Bagging [13] and AdaBoost [14]. The number of ensemble used in this work is 9. The results are reported from the average of 10 runs. Using 10-fold cross validation, the total number of experiment in each data set is 100. The classification accuracy and its standard deviation are shown in Table 3.

The results show that the proposed method can provide the best prediction accuracy in all data sets. They also indicate that the standard deviation of the proposed method is small. In some data set such as Colon, Prostate and Leukemia data set, it provides the smallest value of SD.

We also compare our results with the results reported in [5-7] (only the best of the average value of each reported figure in the paper) on five data sets such as Lymphoma, Ovarian, Colon, Leukemia and Lung data set. The result shows in Table 4. Although we can not compare the results directly because there are differences in the experiment setting, e.g. the testing method used (leave-one-out, 5-fold or 10-fold) or the training set and the test set generation, the results suggest that our approach can provide a good result for the ensemble of GPC. Especially in the Ovarian and Leukemia data set, the proposed method provides the best result. The results also show that the SD. of the results in our method is smaller than the values reported in [6] (only one of the all three works that reported the SD. value).

Table 3. Classification accuracy and standard deviation of each ensemble method for GPC

	Reg	SNR	Bag	AdaBoost	ClusSNR
Lymphoma	86.38±4.28	91.70±***3.54***	82.55±4.58	88.50±6.03	***92.12***±3.62
Ovarian	96.99±0.86	98.57±***0.46***	96.79±0.57	97.98±0.63	***99.21***±0.53
Colon	86.12±2.54	81.93±5.31	80.16±2.85	79.19±3.84	***87.09***±***1.70***
Prostate	78.72±3.14	79.39±2.63	64.60±3.35	68.72±3.53	***87.15***±***1.08***
Leukemia	93.33±2.05	79.99±3.02	88.47±3.71	89.02±3.84	***96.95***±***1.58***
Lung	98.67±***0.70***	96.40±0.91	97.79±1.04	97.62±1.04	***99.22***±0.83

Table 4. Comparison of the accuracy of the proposed method with other methods

Authors.	**Lymphoma**	**Ovarian**	**Colon**	**Leukemia**	**Lung**
Jin-Hyuk Hong (2006) [5]	***97.6***	98.0	-	-	***99.4***
Sung-Bae Cho (2007) [6]	93.0±10.9	-	***87.9***±17.0	95.9±6.4	-
Kyung-Joong Kim (2006) [7]	85.2	-	74.8	92.8	-
Our Approach	92.1±3.6	***99.2***±0.5	87.0±1.7	***96.9***±1.5	99.2±0.8

5. Discussion

In our previous work [4], K-means clustering and SNR ranking are used in the classification task with very good results in terms of the classification accuracy. We continue to use these techniques in conjunction with an ensemble approaches reported here. Due to clustering technique, features which are similarly expressed will be grouped into the same cluster. Therefore, after applying SNR and selected features with the best score of SNR it is assured that a set of features selected has no redundant features. So, each feature can provide useful information to learning algorithms.

In forming an ensemble, each classifier has received the information from different set of features. The i^{th} classifier received the i^{th} best feature from each cluster. In this way, the diversity of the classifiers is promoted. As a result, we can obtain a good classifier but very different from the others to form the ensemble.

6. References

[1] The World Health Organization (WHO) web size: http://www.who.int/mediacentre/factsheets/fs297/en/, Retrieved on 13 Dec. 2007.

[2] Radha Shyamsundar, et al., "A DNA microarray survey of gene expression in normal human tissues", *Genome Biology*, vol. 6 (3), 2005.

[3] Soumya Raychaudhuri, et al., "Basic microarray analysis: grouping and feature reduction", *TRENDS in Biotechnology*, vol. 19 (3), pp. 189-193, May 2001.

[4] Supoj Hengpraprohm and Prabhas Chongstitvatana, "Selecting informative genes from microarray data for cancer classification with genetic programming using K-Means clustering and SNR ranking", *proceeding of Frontiers in the Convergence of Bioscience and Information Technology (FBIT)*, Jeju Island, Korea, October 11th – 13th, 2007.

[5] Jin-Hyuk Hong and Sung-Bae Cho, "The classification of cancer based on DNA microarray data that uses diverse ensemble genetic programming", *Artificial intelligence in Medicine*, 2006

[6] Sung-Bae Cho and Hong-Hee Won, "Cancer classification using ensemble of neural networks with multiple significant gene subsets", *Applied Intelligence*, vol. 26 (3), pp. 243-250, June 2007.

[7] Kyung-Joong Kim and Sung-Bae Cho, "Ensemble classifiers based on correlation anlysis for DNA microarray classification", *Neurocomputing* (70), pp. 187-199, 2006.

[8] Supoj Hengpraprohm and Prabhas Chongstitvatana, "Diffuse large B-cell lymphoma classification using genetic programming classifier", *proceeding of IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, La Jolla, CA, USA, November 14th – 15th, 2005.

[9] Koza, J. , *Genetic Programming*, MIT Press, 1992.

[10] Holland, J., *Adaptation in Natural and Artificial System*, Ann Arbor, Michigan : University of Michigan Press, 1975.

[11] J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations", *proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 1:281-297, 1967.

[12] http://sdmc.lit.org.sg/GEDatasets/

[13] Leo Breiman, "Bagging Predictors", *Machine Learning*, 24:123-140, 1996.

[14] Yoav Freund and Robert E. Schapire, "Experiments with a New Boosting Algorithm", *Machine Learning: Proceedings of the Thirteenth International Conference*, pp. 148-156, 1996.