DNA Gene Expression Classification with Ensemble Classifiers Optimized by Speciated Genetic Algorithm

Kyung-Joong Kim and Sung-Bae Cho

Department of Computer Science, Yonsei University,
134 Shinchon-dong, Sudaemoon-ku, Seoul 120-749, South Korea
{kjkim, sbcho}@cs.yonsei.ac.kr

Abstract. Accurate cancer classification is very important to cancer diagnosis and treatment. As molecular information is increasing for the cancer classification, a lot of techniques have been proposed and utilized to classify and predict the cancers from gene expression profiles. In this paper, we propose a method based on speciated evolution for the cancer classification. The optimal combination among several feature-classifier pairs from the various features and classifiers is evolutionarily searched using the deterministic crowding genetic algorithm. Experimental results demonstrate that the proposed method is more effective than the standard genetic algorithm and the fitness sharing genetic algorithm as well as the best single classifier to search the optimal ensembles for the cancer classification.

1 Introduction

The ensemble classifier, a combination of several feature-classifier pairs, has been regarded as promising due to the incompleteness of classification algorithms, the defects of data, and the difficulty of setting parameters. With the ensemble classifier, we can obtain more reliable solutions than with a single feature-classifier alone. However, because not all ensembles yield good classification performance, it is necessary to find the optimal\ ensembles in order to classify the samples accurately. In the neural network domain, it is well known that many ensembles instead of all neural networks were better. Therefore, forming an ensemble of all the feature-classifier pairs is not a good heuristic. A straightforward way of finding the optimal ensemble is to compare all the ensembles and simply select the best one. However, the possible number is too huge. In this paper, we used 42 feature-classifier pairs, indicating 2^{42} possible ensembles. It would be almost impossible to enumerate all the ensembles with even the most powerful computer.

Kuncheva et al. used the GA approach to design the classifier fusion system which was tested on a non-biological benchmark dataset [5]. In this paper, we propose the deterministic crowding genetic algorithm (DCGA) to search the optimal ensemble classifier. The reasons why we use DCGA rather than standard genetic algorithm (SGA) and fitness sharing genetic algorithm (FSGA) are summarized as follows:

(1) Geometry of the ensemble space is not known,
(2) Feature (gene) space is very huge,

S.K. Pal et al. (Eds.): PReMI 2005, LNCS 3776, pp. 649 – 653, 2005.

(3) SGA tends to converge to only one local optimum of the function, and
(4) DCGA does not require prior knowledge.

The proposed method is unique to search the optimal ensemble evolutionarily from huge number of ensembles, whereas most of the other ensemble methods combine small number of classifiers according to specific rule.

2 Backgrounds

Studies for cancer classification based on gene expression data using ensemble approaches are summarized in Table 1. Most of the previous studies generate the base classifiers using different subsets of features to obtain diverse classifiers. Most of these studies have explored a small part of the ensemble space. However, the objective of this paper is to make a huge ensemble space and search optimal ensembles evolutionarily, which is the main contribution of this paper.

Table 1. Studies using ensemble approach for gene expression data classification

Researcher	Feature selection	Classifier	Ensemble method	Remark
Cho *et al.* [1]	Several methods	MLP KNN SVM SASOM	Majority voting Weighter voting Bayesian combination	Systematical compariso n of features, classifiers and ensemble methods
Tan *et al.* [2]	Fayyad and I rani's discreti zation	C4.5	Bagging Boosting	Resampling
Tsymbal *et al.* [3]	N/A	Simple Bayesian classifier	Cross-validation majority Weighter voting Dynamic selection Etc.	Various ensemble method
Cho *et al.* [4]	Correlation Analysis	MLP	Majority voting	Ensemble classifier trai ned in mutually exclusi ve feature spaces

3 Proposed Method

GA (Genetic Algorithm) is stochastic search method that has been successfully applied in many search, optimization, and machine learning problems. However, standard GA has a defect which tends to converge to local minimum. DeJong's crowding is one of the niching methods which have been developed to reduce the effect of genetic drift resulting from the selection operator in the standard GA.

The structure of a chromosome is very important in the GA. In this paper, a chromosome corresponds to an ensemble. The chromosome is composed of 48 bits string, each of which indicates whether the corresponding FC (Feature-Classifier pair) is joined to the ensemble or not. In this paper, MLP (Multilayer-Perceptron), KNN (K-nearest neighbor), SVM (Support Vector Machine), and SASOM (Structure Adaptive

Self-Organizing Map) [1] are used as classification models. PC (Pearson Correlation), SC (Spearman Correlation Coefficients), ED (Euclidean Distance), CC (Cosine Coefficient), IG (Information Gain), MI (Mutual Information), SN (Signal-to Noise Ratio) and PCA (Principal Component Analysis) [1] are used as feature selection.

$$C_{fitness} = ENS_{accuracy} - k\, N_p$$

where k is a constant, N_p is the number of participant FCs to the ensemble and $ENS_{accuracy}$ is the accuracy of the corresponding ensemble on the validation data set.

$$ENS_{accuracy} = \frac{\#\text{ of exactly classified samples}}{\#\text{ of total validation samples}}$$

The crossover operation changes individual FCs partly between mated chromosomes, and the mutation operation either adds new FC to current chromosome or deletes a FC from it. The similarity between chromosomes is computed using hamming distance of genotypes. The algorithm of the procedure is described in Fig. 1. The majority voting scheme is used for combining the classifiers in an ensemble in this paper. This is a very popular combination scheme because of both its simplicity and its performance on real data.

```
Initialize $P$ individuals; {each individual represents an ensemble}
Insert $P$ individuals into $Q$; {$Q$ holds individuals of the population}
for ($i = 0$ ; $i < MAX\_GEN$; $i$++)
  Shuffle $P$ individuals in $Q$;
  while ($Q$ is not empty) do
    Delete ($p_1$, $p_2$) from $Q$;
    $q_1$, $q_2$=Crossover $p_1$, $p_2$;
    $r_1$, $r_2$=Mutate $q_1$, $q_2$;
    if (distance ($p_1$, $r_1$)+distance ($p_2$, $r_2$) < distance ($p_1$, $r_2$)+distance ($p_2$, $r_1$)) then
      if (fitness ($p_1$) < fitness ($r_1$)) then Insert ($r_1$) to $Q2$; else Insert ($p_1$) to $Q2$; end if
      if (fitness ($p_2$) < fitness ($r_2$)) then Insert ($r_2$) to $Q2$else Insert ($p_2$) to $Q2$; end if
    else
      if (fitness ($p_1$) < fitness ($r_2$)) then Insert ($r_2$) to $Q2$ else Insert ($p_1$) to $Q2$; end if
      if (fitness ($p_2$) < fitness ($r_1$)) then Insert ($r_1$) to $Q2$ else Insert ($p_2$) to $Q2$; end if
    end if
  end while
  $Q$=$Q2$;
end for
```

Fig. 1. A pseudo code for crowding algorithm

If there are m classes and k classifiers, the ensemble result by majority voting is determined as follows:

$$c_{ensemble} = \underset{1 \le i \le m}{\arg\max} \left\{ \sum_{j=1}^{k} s_i(classifier_j) \right\}$$

where c_i is the class i, $i = 1, \ldots, m$, $s_i(classifier_j)$ is 1 if the output of the j-th classifier $classifier_j$ equals to the class i, 0 otherwise.

4 Experimental Results

B cell diffuse large cell lymphoma (B-DLCL) is a heterogeneous group of tumors, based on significant variations in morphology, clinical presentation, and response to treatment. Gene expression profiling has revealed two distinct tumor subtypes of B-DLCL: germinal center B cell-like DLCL and activated B cell-like DLCL. This lymphoma cancer dataset consists of 24 samples of GC B-like and 23 samples of activated B-like (http://genome-www.stanford.edu/lymphoma).

For the feature selection, we have selected top 25 genes or principal components (for PCA) considered as informative since a preliminary study suggested the optimal number of genes as 25~30 [1]. We have calculated PC, SC, ED and CC between each gene vector and 'ideal gene vector' in which the expression level is uniformly high in class 1 and uniformly low in class 2. IG, MI and SN are calculated using the feature values and the class label. Only PCA does not use the label information of samples.

For the classifiers, we have set the number of input-hidden-output nodes to 25-8-2. We also set 0.01~0.50 of learning rate, 0.3~0.9 of momentum, and 500 of maximum iterations. We let the back-propagation algorithm stop the training when it reaches to 98% of training accuracy. In the case of *k*NN, we have set the *k* from 3 to 9, and used Pearson correlation coefficients and cosine coefficients for the similarity measures. We have used Joachim's SVM^{light} with linear and radial basis function kernels (http://svmlight.joachims.org/). In SASOM, we have used initial 4×4 map which has rectangular shape. The details of parameters for each classification model can be found in [1].

For the DCGA, we have set the 0.9 of crossover rate, 0.05 of mutation rate, 500 of population size, and 2000 of maximum iterations and employed roulette wheel selection scheme. The value of *k* is set to 0.01. Fitness sharing which is for comparison, we set the sharing radius of 5. Table 2 shows the average accuracies of individual FCs for lymphoma dataset.

Leave-one-out cross-validation (LOOCV) is employed in our experiments both to overcome the number of samples and to just evaluate the proposed method. For LOOCV, dataset is divided into three parts: training samples, validation samples and one test sample. The FCs are trained by training samples. DCGA operates with validation samples to find the optimal ensemble. Finally, the best solution in the last generation is validated by test sample. These are repeated as many times as the number of data.

Table 2. The accuracy of individual FCs for lymphoma dataset

	MLP	SASOM	SVM(L)	SVM(R)	KNN(C)	KNN(P)	AVG
PC	77.6	67.7	66.4	55.6	78.4	78.0	70.6
SC	78.8	67.2	68.0	57.6	78.4	76.8	71.1
ED	75.2	62.8	66.4	64.0	76.0	77.6	70.3
CC	80.0	64.4	72.4	56.4	78.0	78.4	71.6
IG	85.2	75.2	77.6	66.8	81.6	83.2	78.3
MI	80.0	67.6	67.2	58.4	76.4	77.2	71.2
SN	81.2	70.8	68.0	58.4	78.8	79.2	72.7
PCA	87.2	84.0	88.4	58.4	86.0	86.4	81.7
AVG	80.7	70.0	71.8	59.5	79.2	79.7	73.5

Table 3. The comparison of performance (Average of ten runs)

Methods	LOOCV accuracy
Best single feature-classifier pair	95.3±3.05
Ensemble of all classifiers whose accuracy is larger than 80%	94.7±3.16
Simple GA	98.0±2.10
Crowding	98.8±1.93

DCGA found the optimal ensembles which exactly classify every validation sample. To demonstrate the superiority of our method, we have compared it with other methods and the result is in the Table. 3. The accuracy means the rate of exactly classified samples among test samples. Ensemble of good base classifiers (accuracy > 80%) is the combination of good individual FCs for classification. The performances of all GA strategies (SGA, DCGA and FSGA) are the best ensembles among 1 million ones which are generated by their operations. This shows that the combination of similar good ensemble classifiers degrades the performance and the proposed method performs well.

5 Conclusion and Future Work

This paper presents a DCGA-based method of searching the optimal ensemble for cancer classification using DNA microarray data. Though FSGA is also well known for one of good niching methods, we have employed DCGA because FSGA is known that it usually fails on hard problems and it requires prior knowledge for good result. Experiments have supported the use of DCGA for the optimal ensemble to classify cancers. The result of LOOCV also confirms the superiority of the proposed method.

Acknowledgement

This research was supported by Brain Science and Engineering Research Program sponsored by Korean Ministry of Commerce, Industry and Energy.

References

1. Cho, S.–B., and Won, H.–H.: Data mining for gene expression profiles from DNA microarray. Int. Journal of Software Engineering and Knowledge Engineering, 13 (2003) 593-608
2. Tan, A. C. and Gilbert, D.: Ensemble machine learning on gene expression data for cancer classification. Applied Bioinformatics, 2 (2003) s75-s83
3. Tsymbal, A. and Puuronen, S.: Ensemble feature selection with the simple Bayesian classification in medical diagnostics. Proc. of the 15th IEEE Symp. on Computer-Based Medical Systems (2002) 225-230
4. Cho, S.-B., and Ryu, J.-W.: Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features. Proc. of the IEEE, 90 (11) (2002) 1744-1753
5. L. I. Kuncheva, and L. C. Jain, "Designing classifier fusion systems by genetic algorithms," IEEE Transactions on Evolutionary Computation, vol. 4, no. 4, pp. 327-336, 2000.