IEIT Journal of Adaptive & Dynamic Computing

IEIT Journal of Adaptive & Dynamic Computing, 2011(4), 34-39, 2011

doi: 10.5813/www.ieit-web.org/IJADC/2011.4.5

The Classification of Microarray Data Using Evolutionary Classifier Ensemble System

Kun-Hong Liu*

Software School of Xiamen University,Xiamen 361005,Fujian Province,China

Keywords: Classification; Genetic Algorithm; DNA

Abstract: Microarray data prediction is a hard task due to the small sample and high dimension property. This paper proposes a classifier fusion approach to solve this problem based on genetic algorithm (GA). In this fusion strategy, the GA is applied to select proper feature subsets and weight value for the fusion of classifiers. The experimental results show that this scheme can improve the prediction accuracy.

1. Introduction

With the development of microarray technology, it is possible to diagnose and classify some particular cancers directly based on these DNA microarray data nowadays. There have been some successful examples: molecular classification of accurate leukemia [4]; clustering analysis of cancer and normal colon tissues [1]; classification and analysis of breast cancers [11]. Although it is important to design an efficient and accurate method for classifying the samples through analyzing the microarray data, it is still difficult to build a new efficient prediction system because microarray gene expression data usually contains a large set of gene features including many irrelevant ones with a small number of samples.

Feature selection is an important aspect of classification problems. With the aid of feature selection techniques, the irrelevant features can be removed and important features can be identified by applying certain selection criteria. There are two kinds of feature selection methodologies: filter approach and wrapper approach. The filter approach has been proposed to rank the features using statistical or information-theoretic functions, such as T-test, signal-to-noise statistics, U-statistics, entropy-based measures. Usually by selecting the top-ranked features, an accurate classifier system can be built. However, as all the filter methods are motivated by various hypotheses which would not be true in many cases, different method would pick up quite different feature subsets even for a same feature set. At the same time, one method performing well for a data set would be bad for other data sets. What's more, as the filter approach is independent to the prediction rule, it is impossible to decide which filter method could be applied to improve the performance of a classifier system immediately. So it is not easy for us to select a proper filter method. On the contrary, the wrapper approach would be a better choice for selecting the features as the wrapper methods, such as sequential floating forward selection and Genetic Algorithm (GA), take the performance of the classifier into account. So usually the wrapper approach can lead to better prediction results. While a main drawback of the wrapper approach lies in that it requires much longer time to evaluate

*Corresponding author: Kun-Hong Liu
E-mail Address: author@affiliation.org

features compared with the filter approach, especially when dealing with the microarray data problem. Based on this observation, we try to combine the two approaches together. That is, a feature pool is set up by applying some filter methods to selecting a part of feature subsets firstly, and then a wrapper method is applied for further filtering the features in the feature pool.

At the same time, it should be also noted that one single classifier can not always lead to good prediction performance. Instead, a multiple classifier system would be a better choice, and it is proved that the classifier ensembles can be more accurate than an excellent single classifier in many fields [6]. Multiple classifier system is currently a hot research area. Dietterich has pointed out that the integration of multiple classifiers is one of the four most important directions in machine learning research [2]. However, the investigation for microarray data prediction based on multiple classifiers is still just at the beginning.

After eliminating the irrelevant features, the generalization of multiple classifier system will be further improved, which is named as ensemble feature selection [9]. And it is obvious that the exploration of the ensemble feature selection problem for microarray data is necessary and urgent. The evolutionary approaches are proved to be efficient in the search of high quality results, and there are some classifier fusion approaches proposed based on them [7, 9]. But to the best of our knowledge, there is still no paper discussing the ensemble feature selection problem for microarray data problem based on evolutionary approaches. And in this paper, a GA based classifier fusion approach is proposed for the prediction of microarray data. This GA is designed to implement the feature subset selection by combining valuable outcomes from multiple filter feature selection methods, and discover the proper feature subset ensemble method for the fusion of multiple classifiers. Then, the algorithm would assign proper weights to different classifiers and produce the results by majority vote method. The simulation results show that our approach is efficient and applicable to microarray data sets.

2. The design of genetic algorithm

GA is inspired by mechanisms of evolutions in nature, and has been proved to be successful at tackling the optimization or feature selection problem. In this paper, GA is proposed to construct an efficient classifier fusion system. Our GA is based on the GEATbx toolbox [10].The details of the design of GA are as follows.

2.1 Ensemble feature selection

In our experiment, four filter methods are applied to select the gene features first. These methods are based on T-test, entropy, Chernoff bound and U-statistic respectively, and are provided in the bioinformatics toolbox in Matlab 2006. These methods are feature independent. To simplify the feature selection problem, for each microarray data set, only the 20 top gene features are selected for each method, and the selected 80 gene features are kept to construct a feature pool. Then the features in the pool are further selected by the GA. In this way, the search space is relatively small and then we can obtain satisfying results by the GA within only a few generations. And it is obvious that if there are some more candidates in the gene pool, it is of greater probability to discover better gene feature sets with higher prediction ability.

In our experiments, all the filter methods would share some features with others, so some features will appear in the pool more than once. For example, in our experiments, we find that for the colon data set [1], there are only 49 unique features in the total 80 features. Instead of simply ignoring the reduplicative ones, we argue that these features may be of much greater significance compared with those selected only once. So we hope that the information of significance can be applied to guide the feature selection process.

Based on this consideration, we developed the coding method proposed in [7]. The

method is proposed for classifier fusion with feature selection based on GA, which is named as selection of disjoint feature selection [7]. With this method, assume that there are L classifiers in the fusion system, and the chromosome consists of n positions so as to represent n features. Each position can take value [0, L], and the corresponding feature is ignored when scored as 0, is used by kth classifier when it is valued as k ($1 \le k \le L$). It is an efficient ensemble method, but since this method requires that a feature can only be applied to a classifier, it is too restricted for many real world applications. While it is intuitive to find that this drawback can be compensated by allowing all the features in the pool to get an equal selection probability. In this way, the possibility for each feature to be assigned to a classifier is equal to the frequency of the feature appearing in the pool. As a result, the features shared by more filter methods are given more chances to be picked up by different classifiers, and then the feature selection process can be under the guidance of the feature significance. If a same feature is assigned to a same classifier more than one time, the redundant one(s) would be ignored. Figure 1 gives an example of a possible coding scheme. Feature F_i is shared by three filter methods so there are three copies of F_i in the feature pool. But as the values of the corresponding positions are as 5, 0, 5 respectively, F_i is finally only picked up by 5^{th} classifier once according to the decoding rule described above. While feature F_j is assigned to 4^{th} and 5^{th} classifier. With this design, the ensemble feature selection can be more efficient.

The different feature subset can already ensure the diversity among the classifiers, so in [7], only the accuracy of the individual classifier is adapted as the fitness function.

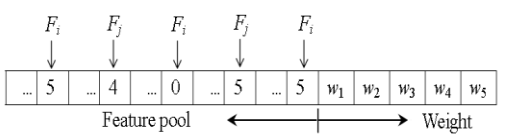

Figure 1 The structure of chromosome.

However, we find that in our experiments, this fitness function can not work well due to the small samples size. And it is necessary to consider the diversity in the fitness function to avoid overfitting. And the fitness function is adjusted by taking the diversity into consideration, and for i-th chromosome, it can be expressed as:

$$\text{Fitness}_i=\text{accuracy}_i+\text{diversity}_i \quad (1)$$

Where accuracy_i is the prediction accuracy of the classifier fusion system, and diversity_i is the diversity of the ensemble classifier system which is quantified by the plain disagreement measure in our paper. In detail, for classifier n and m, the plain disagreement measure is:

$$\text{diversity}_i=\sum_{n=1}^{L}\sum_{m=n+1}^{L}\sum_{k=1}^{N} Dif(C_{ni},C_{mi}) /((L-1)\times L\times N) \quad (2)$$

where $\text{Dif}(C_{ni},C_{mi})=0$ if the outputs of classifier n and m are the same based on the feature subset decided by chromosome i, 1 otherwise; N is the total number of samples. In this way, the plain disagreement varies from 0 to 1

2.2 The fusion of the outputs

An efficient method for the fusion of label outputs is the key to the high prediction ability of a multiple classifier system. The majority vote is a common used method, while as proved in [6], the weighted majority vote can lead to higher accuracy than simple majority vote. In this paper, both the simple majority vote and the weighted majority vote are applied as the fusion method for the purpose of comparison. And the GA with simple majority vote method is named as GA-1.

Usually the weight assignment is based on the estimate of the performance of the classifiers the training data sets. After the evaluation of the prediction accuracy of the i-th classifier pi, the weight bi for the classifier can be calculated with the following formula [6]:

$$b_i=\frac{\log p_i/(1-p_i)}{\sum_{i=1}^{L}\log p_i/(1-p_i)} \quad (3)$$

However, it is impossible to embed this weight assignment method in the process of

ensemble feature selection as the selection of features is based on the corresponding accuracy of a classifier or a classifier system. There are two methods for the realization of weight majority vote: a). after the process of feature selection, each feature subset would be assigned to a corresponding classifier, then the weight can be calculated and assigned to a classifier by evaluating its accuracy according to (3), with the aid of cross validation or bootstrap technique; b). the weight assignment can also be regarded as a set of solution which can be optimized by GA. The first method is still a standard method to implement the weighted majority vote without the influence on the ensemble feature selection. While the second one is different from the first one in that it combines the weight value search with the feature selection scheme. To realize the second method, assume that w classifiers are adopted to construct a classifier ensemble system, and there should be additional w positions in each chromosome to represent the weight of each classifier, as shown in Figure 1. These positions will take value within the interval [1, 10]. Denote a position indicating the weight for a classifier i as wi, and the corresponding weight ci is calculated by

$$c_i = w_i / \sum_i w_i \qquad (4)$$

In this way, a convex combination can be achieved. So for the second method, the total length of a chromosome is 80+w. We try other two GAs with the two different weighted majority vote methods: GA-2, with the first weighted majority vote scheme; GA-3, with the second weighted majority vote scheme. It should be noted that the length of chromosome for GA-1 and GA-2 is 80, and as we use five classifiers in the experiments, the length of chromosome is 85 for GA-3.

3. Experimental results and discussion

We use three publicly available microarray datasets: colon cancer data [1], hepatocellular carcinoma data [5] and high-grade glioma data [8]. In these three datasets, there are 40, 33, 21 samples for classifier training, and 22, 27, 29 samples for test respectively. Five classifiers are used as base classifiers in this paper: fishier classifier, binary decision tree, nearest mean classifier (nmc), support vector classifier (svc), the nearest neighbor classifier (1-nn). All of them are provided by the PRTools [3], and all the corresponding parameters of these base classifiers are set to the default values. In all the experiments, we applied external 10-fold cross-validation to evaluate the accuracy of each classifier, so that the prediction results are tested on the independent test data sets. The GAs run with 5 independent runs, and only the best results are reported. For all the GAs, the population is divided into three subpopulations with 25, 15, 10 individuals respectively, and the number of generation is set to be 10 since usually good results can be achieved within the 10 generations and it would not benefit with some more generations. The stochastic universal sampling selection, randomly exchange mutation operator, and discrete recombination operator are adapted. The recombination rate and migration rate are both set as 1, and the mutation rate is set as 0.5. The generation gap is 0.9, and the selection pressure is 1.7. These settings can guarantee the diversity.

Due to the small sample size, the task of prediction is quite hard for a single classifier. The prediction results with all 80 features selected by four filter methods for each single classifier are shown in Table 1, and it is obvious that the results are not good enough. For the comparison purpose, we also show the results of the fusion of outputs of the five classifiers with the simple majority vote and weighted majority vote, which are represented as method 6 and 7 respectively. These results are consistent with the observation in [6] that the majority vote can not always do better than a single classifier in the fusion system. From the results, it can be found that without the ensemble feature selection, neither a single classifier nor the direct fusion system could lead to good results. And it should be noted that in our experiments, we find that the results with unique features are very close to those with all features, so we do not list them in Table 1.

Table 1 The Numerical Experiments On The Data Sets

method	Colon data	hepatocellular data
fisher	0.7727	0.5185
Tree	0.6364	0.5926
Nmc	0.7727	0.3704
Svc	0.7727	0.5926
1-nn	0.7273	0.5185
Simple majority vote	0.7727	0.5556
Weighted majority vote	0.7727	0.6667
GA-1	0.7727	0.6667
GA-2	0.8182	0.7407
GA-3	0.8182	0.8148

Compared with them, the GAs can lead to better results by finding proper feature subset for the majority vote. With simply majority vote, the results of GA-1 are not always better than those of method 1-7 except for the glioma data set. GA-2 and GA-3 perform better than GA-1 in all the cases, which show that with weighted majority vote, a gain in accuracy can be obtained. GA-3 leads to the best results, and specially, it achieves surprising good results on the hepatocellular data set, which shows that it can efficiently search the weight assignment schemes.

4. Conclusion

We provide an ensemble feature selection approach based on GA to the construction of classifier fusion system in this paper. This approach is applied to three microarray data sets with three different fusion methods. In our experiments, the GA-3 achieves best results with the aid of weight assignment search. Although the proposed approach is only applied to tackle the prediction problem of two classes in this paper, it can be extended to predict the multi-class microarray data by applying multi-class classification methods, such as error correcting output coding or multi-class classifiers, which is the direction our future search.

References

[1] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," Proc. Natl Acad. Sci. USA, 96, 1999, pp.6745–6750.

[2] T.G. Dietterich, "Machine learning research: four current directions," AI magazine, 18(4), 1997, pp. 97-136.

[3] R.P.W. Duin, P. Juszczak, P. Paclik, E. Pekalska, D. de Ridder, D.M.J. Tax, "PRTools4, A Matlab Toolbox for Pattern Recognition," Delft University of Technology, 2004.

[4] 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.

[5] N. Iizuka, M. Oka, H. Yamada-Okabe, M. Nishida, Y. Maeda, N. Mori, T. Takao, T. Tamesa, A. Tangoku, H. Tabuchi, et al. "Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection," The Lancet, 361, 2003, pp. 923–929.

[6] L.I. Kuncheva, "Combining pattern classifiers: methods and algorithms," Wiley, 2004

[7] L.I. Kuncheva and L.C. Jain, "Designing classifier fusion systems by genetic algorithm," IEEE Transaction on evolutionay computation, vol.4, no. 4, 2000, pp. 327-336

[8] C. L. Nutt, D. R. Mani, R. A. Betensky, P. Tamayo, J. G. Cairncross, C. Ladd, U. Pohl, C. Hartmann, M. E. McLaughlin, et al. "Gene expression-based classification of malignant gliomas correlates better with survival than histological classification," Cancer Res., 63, 2003, pp.1602–1607.

[9] D.W. Opitz, "Feature selection for ensembles," Proc. 16th National Conf. on Artificial Intelligence, 1999, pp. 379-384.

[10] H. Pohlheim, "GEATbx - Genetic and Evolutionary Algorithm Toolbox for use with Matlab," http://www.geatbx.com/, 1994-2006.

[11] van't Veer L.J., Dai H., Van De Vijver M. J., He Y. D., Hart A. A. M., Mao M., Peterse H. L., Van Der Kooy K., Marton M. J., Witteveen A. T., Schreiber G. J., Kerkhoven R. M., Roberts C., Linsley P. S., Bernards R., and Friend S. H., Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer, Nature, 415: 530-536, 2002.