International Journal of Research Studies in Biosciences (IJRSB)
Volume 3, Issue 1, January 2015, PP 201-204
ISSN 2349-0357 (Print) & ISSN 2349-0365 (Online)
www.arcjournals.org

A BBO Based Feature Selection Method for DNA Microarray

Ammu P K

SCT College of Engineering
Kerala, India

Siva Kumar K C, Sathish Mundayoor

Rajiv Gandhi Centre for Biotechnology
Kerala, India

Abstract: *Feature selection is often employed prior to classification using machine learning techniques to alleviate the dimensionality problem of data sets. The dimensionality problem can lead to classification error. DNA Microarray data sets represents the differentially expressed genes that play a major role in various biological processes. Many of the microarray data sets possess the dimensionality problem, i.e. the number of genes is much larger compared to the number of samples. In this paper we have proposed a hybrid feature selection approach for DNA microarray data sets.*

Keywords: *BBO, feature selection, k-nearest neighbor*

1. INTRODUCTION

Feature selection from microarray data sets is usually performed prior to classification. Feature selection techniques can be generally classified as filter, wrapper and embedded methods. Filter methods filter features based on some univariate or multivariate measure and select the best features [1][2]. Wrapper methods usually employ evolutionary algorithms to select the best features [3][4][5] whereas embedded methods embeds the search procedure within a classifier[6]. The advantage of filter method is its simplicity but it is not efficient in identifying the relevance of genes in combination with other genes. Wrapper and embedded methods overcome this disadvantage whereas they are computationally intensive. In this paper we propose a hybrid approach for feature selection which includes two phases. In the first phase a filter approach is employed and in the second phase a wrapper approach is employed.

2. METHODS

The proposed method calculates the information gain values of individual genes in the first phase and selects the genes with non zero information gain values. In the second phase biogeography-based optimization algorithm with immigration refusal [7] is applied on selected genes from the former phase and the best subset of genes is selected. The algorithm is described in detail in the following sections.

2.1. Information Gain

Information gain of a feature indicates the expected reduction in entropy associated with a feature. The formula for calculating information gain is given in (1).

$$Gain(S, A) = Entropy(S) - Sum(|Sv| \ / \ |S|) * Entropy(Sv) \quad (1)$$

where S is a sample of training examples, Gain(S, A) is the expected reduction in entropy due to sorting S on attribute A, S_v is the set of training instances remaining from S after restricting to those for which attribute A has value v. Weka [8] is employed in this study for calculating the information gain. Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand.

2.2. Biogeography-Based Optimization with Immigration Refusal

Biogeography-based optimization (BBO) algorithm was proposed by Dan Simon in 2008 [9]. The algorithm simulates the migration of species among different islands based on their fitness, for survival. Every island is represented as a collection of certain suitability index variables (SIV). SIV represents the suitability of the habitat for the residence of species. The fitness of each island is represented by the term habitat suitability index (HSI). Mutation can also performed in the algorithm depending on the choice of the implementer. Dawei Du made some changes to BBO and introduced biogeography-based optimization with immigration refusal (BBO/RE). The basic procedure of the algorithm is as follows

1. Initialize the population of solutions.
2. Evaluate the fitness of population members.
3. Compute immigration rate (λ), emigration rate (μ).
4. Perform migration based on the following migration criteria.
   - The fitness of immigrating island is lower than that of the emigrating island.
   - The fitness of emigrating island is greater than some threshold.
5. If the fitness criterion is not satisfied, go to step 3; otherwise terminate.

Immigration rate and emigration rate for each island i is calculated as in (2) and (3).

$$\lambda_i = I * Rank(i) / n \tag{2}$$

$$\mu_i = E * (1 - Rank(i) / n) \tag{3}$$

Where I is the maximum immigration rate and E is the maximum emigration rate. Migration procedure can be detailed as follows

---

Select habitat H_i with probability $\alpha \lambda_i$

If H_i is selected

For j=1 to n

Probabilistically select a habitat H_j based on μ_j such that $\mu_i < \mu_j$ and $\mu_j >$ threshold

If H_j is selected

Randomly select an SIV s_i from H_j

Replace a random SIV in H_i with s_i

End

In our experiments every island was represented as a collection of genes. The fitness function was calculated through two fold cross-validation using K-nearest neighbor algorithm (k-NN). In two fold cross-validation, the sample data is partitioned into two complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set). K-NN is a classification algorithm that classifies features based on the closest training samples in feature space [10]. Using k-NN the classification error rate for each island is calculated. The objective function in BBO is to minimize the classification error rate.

3. RESULTS

The performance of the hybrid approach was analyzed on colon tumor and prostate tumor data sets. Colon tumor data set was obtained from Kent Ridge Biomedical data set repository. It contains 62 samples and 2000 genes. Prostate tumor data set was downloaded from http://www.gems- system. org. This data set contains 102 samples and 10,509 genes. The number of generations and the population size for BBO/RE were set as 50. The algorithm was run for 50

generations and the best minimum and average minimum classification error rate were calculated. The results are shown in Table 1. Classification accuracy up to 80% was obtained on colon tumor data set and 85% accuracy was obtained on prostate tumor data set. Fig 1 shows the best minimum obtained on various generations on colon tumor data sets and Fig 2 shows the various results obtained on prostate tumor data sets.

Table1. *Results obtained using the algorithm on various data sets*

	Colon Tumor	Prostate Tumor
Best minimum error rate	2	2.642
Mean minimum error rate	1	2.357
Classification accuracy	80%	85%

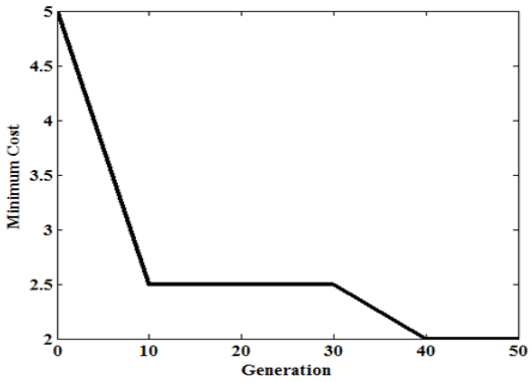

Fig1. *Minimum cost obtained over 50 generations on colon tumor data set*

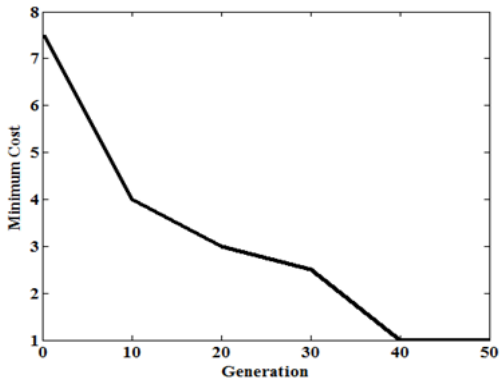

Fig2. *Minimum cost obtained over 50 generations on prostate tumor data set*

4. CONCLUSION

The feature selection approach proposed here combines the advantage of both filter and wrapper approaches. Applying filter approach before the wrapper approach reduces the dimensionality of data set and on the reduced data set, the wrapper approach is applied. This helps in speeding up the wrapper approach. The algorithm proposed here could obtain very good classification accuracies. The feature selection procedure is effectively simplified by our approach.

REFERENCES

[1] Wang Z., Neuro-fuzzy modelling for microarray cancer gene expression data. Conference Proceeding, Int. Symposium on Evolving Fuzzy Systems, Pp 241-246, (2005).

[2] Fung Y M. and Ng VTY., Classification of heterogeneous gene expression data, Pharmacogenomics. 5(2), 69(2003).

[3] Chiang YM., and Lin SY., The application of ant colony optimization for gene selection in microarray-based cancer classification, Conference Proceeding, Int. Conference on Machine Learning and Cybernetics, Pp 12-15, (2008).

[4] Nikumbh S., Ghosh S. and Jayaraman VK., Biogeography-based informative gene selection and cancer classification using SVM and Random Forests, Conference proceeding, IEEE Congress on Evolutionary Computation, Pp 1-6 (2012).

[5] Karzynski M., Mateos A. and Dopazo J., Using a genetic algorithm and a perceptron for feature selection and supervised class learning in DNA microarray data**,** Artificial intelligence review. 20(1), 39(2003).

[6] Yu Y., SVM-RFE algorithm for gene feature selection, Technical report, University of Delaware, (2005).

[7] Dawei Du., Biogeography-based optimization: synergies with evolutionary strategy, immigration refusal, and kalman Filters. Thesis, Cleaveland State University, (2007).

[8] Frank E., Hall M., Trigg L., Holmes G. and Witten IH., Data mining in bioinformatics using weka, Bioinfomatics 20(15), Pp 2479–2481, (2004).

[9] Simon D., Biogeography-based optimization, IEEE Transactions on Evolutionary Computation, 12(6), 702(2008).

[10] Cover T. and Hart P., Nearest neighbour pattern classification, IEEE Transactions on Information Theory, 13(1), 21(1967).