BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis
Volume 9(16)

Feature Selection for high Dimensional DNA Microarray data using hybrid approaches

Ammu Prasanna Kumar* & Preeja Valsala

Sree Chitra Thirunal College of Engineering, Pappanamcode, Trivandrum, Kerala; Ammu Prasanna Kumar - Email: ammupk99@gmail.com; *Corresponding author

Received August 31, 2013; Accepted September 03, 2013; Published September 23, 2013

Abstract
Feature selection from DNA microarray data is a major challenge due to high dimensionality in expression data. The number of samples in the microarray data set is much smaller compared to the number of genes. Hence the data is improper to be used as the training set of a classifier. Therefore it is important to select features prior to training the classifier. It should be noted that only a small subset of genes from the data set exhibits a strong correlation with the class. This is because finding the relevant genes from the data set is often non-trivial. Thus there is a need to develop robust yet reliable methods for gene finding in expression data. We describe the use of several hybrid feature selection approaches for gene finding in expression data. These approaches include filtering (filter out the best genes from the data set) and wrapper (best subset of genes from the data set) phases. The methods use information gain (IG) and Pearson Product Moment Correlation (PPMC) as the filtering parameters and biogeography based optimization (BBO) as the wrapper approach. K nearest neighbour algorithm (KNN) and back propagation neural network are used for evaluating the fitness of gene subsets during feature selection. Our analysis shows that an impressive performance is provided by the IG-BBO-KNN combination in different data sets with high accuracy (>90%) and low error rate.

Background:
Microarray experiments help to identify the significant genes that play a major role in determining various types of cancers. The experiment makes use of a microarray chip which is a collection of known DNA spots. For identifying the relevant genes, samples are collected from normal and cancerous tissues of patients. The samples are coloured differently by labelling them with fluorescent nucleotides and are mixed together. The mixture is then applied to the DNA microarray chip. Based on the amount by which the samples hybridize with the DNA spots in the array, various colours will appear on the chip. By measuring the intensity of colours, the microarray expression data set is prepared. The major problem of microarray data set is that the number of samples in the data set is much smaller compared to the number of genes, so while this data is employed with a classifier, the classifier may overfit. To alleviate this problem feature selection is employed prior to classification. During feature selection every gene in the data set is considered as a feature or attribute and the feature selection procedure aims at reducing the number of features.

Recently many gene selection and classification techniques are proposed. Huang *et a*l. **[1]** proposed an improved decision forest for the classification of genes that incorporates a built-in feature selection mechanism for fine-tuning. Li *et al.* **[2]** proposed an algorithm for mapping the microarray data to a low dimensional space. Ding *et al.* **[3]** proposed a minimum redundancy maximum relevance method for feature selection. Feature selection techniques can be generally classified as filter, wrapper and embedded methods. Filter methods ranks genes based on some univariate measure, thus features that accurately present the whole data set can be found out. On the other hand, filter methods doesn't consider the relevance of genes in combination with other genes. Filter methods include correlation-based feature selection (CFS), t-test, information gain **[4]**, mutual information **[5]**, entropy-based methods **[6]**, Euclidian distance, signal to noise ratio, correlation coefficient and significant analysis of microarray. Wrapper methods try to find out the best combination of genes that may provide high classification accuracy. Wrapper methods include hybrid genetic algorithm **[7]**, particle swarm optimization **[8, 9]**, ant colony optimization and tabu search **[10, 11].** In the case of embedded methods, the feature selection procedure is inbuilt to a classifier. Classification trees like ID3, random forest etc are examples of embedded methods. The hybrid methods proposed in this paper are combinations of filter and wrapper methods. These methods employ Pearson Product Moment Correlation (PPMC) and information gain (IG) as filters and Biogeography based optimization (BBO) as the wrapper approach. K nearest neighbour algorithm (KNN) and artificial neural network (ANN) are employed in this study for the fitness evaluation and classification. The hybrid methods are as follows: (1) Hybrid approach employing PPMC, BBO and ANN (PPMC-BBO-ANN); (2) Hybrid approach employing PPMC, BBO and KNN (PPMC-BBO-KNN); (3) Hybrid approach employing IG, BBO and ANN (IG-BBO-ANN); (4) Hybrid approach employing IG, BBO and KNN (IG-BBO-KNN).

Methodology:

Feature selection using correlation coefficient

Correlation measures the relationship between variables. The most common measure of correlation in statistics is the Pearson Product Moment Correlation (PPMC), which shows the linear relationship between two variables. Formula for calculating Pearson correlation between features xi and yi is given in equation 1.

Correlation = (Σ (xi – mean (xi) * (yi-mean(yi))/(n * Stantard deviation(xi) * Stantard deviation(yi)) → (1)

The two stages of the hybrid approach based on correlation coefficient are as follows:

Stage 1: Pearson correlation coefficient between attributes is found out. Attributes having low inter-correlation are selected. The idea here is to reduce redundancy among features by selecting uncorrelated features; Stage 2: On the filtered attributes BBO is applied to find out the best set of attributes.

Feature selection using information gain

Information gain (IG) of a feature indicates how much informative the feature is for classification. IG is calculated using equation 2:

Gain(S, A) =Entropy(S) – (sum (|Sv|/|S|)*Entropy (Sv)) → (2)

Where S is a sample of training examples, Gain(S, A) is the expected reduction in entropy due to sorting S on attribute A, Sv is the set of training instances remaining from S after restricting to those for which attribute A has value v. The two stages of the hybrid approach based on information gain are as follows: Stage 1: Information gain values of individual attributes are found out using weka (a machine learning software in java). Attributes having non-zero information gain are selected; Stage 2: BBO is applied on the filtered attributes to find out the best set of attributes.

Biogeography based optimization

Biogeography based optimization (BBO) is a population based optimization algorithm introduced by Dan Simon in 2008 **[12].** BBO got its inspiration from the theory of island biogeography. The algorithm uses species migration between islands and the process of mutation to reach the global minimum. Every solution in the solution space is considered as an island. An island is any habitat that is geographically isolated from other habitats. Every island is a collection of certain suitability index variables (SIV's). In other words SIV's are the decision variables in the solution. Every solution has its associated immigration and emigration rate. Immigration and emigration rates of solutions indicate their willingness to accept features from other solutions. Immigration rate and emigration rate for solution 's' are calculated as follows:

λ_s= I* Rank (s)/n → (3)
μ_s = E* (1-Rank (s)/n) → (4)

Ranks are assigned to the solutions based on their fitness. The fitness of each island represents the habitat suitability index (HSI) of the island. HSI indicates the suitability the island for species residence. Islands with high HSI have high emigration rate and low immigration rate. Islands with low fitness will be having low emigration rate and high immigration rate. The basic procedure of BBO algorithm is shown in **Figure 1.** Migration and mutation procedures are shown in **Figure 2 & Figure 3**.

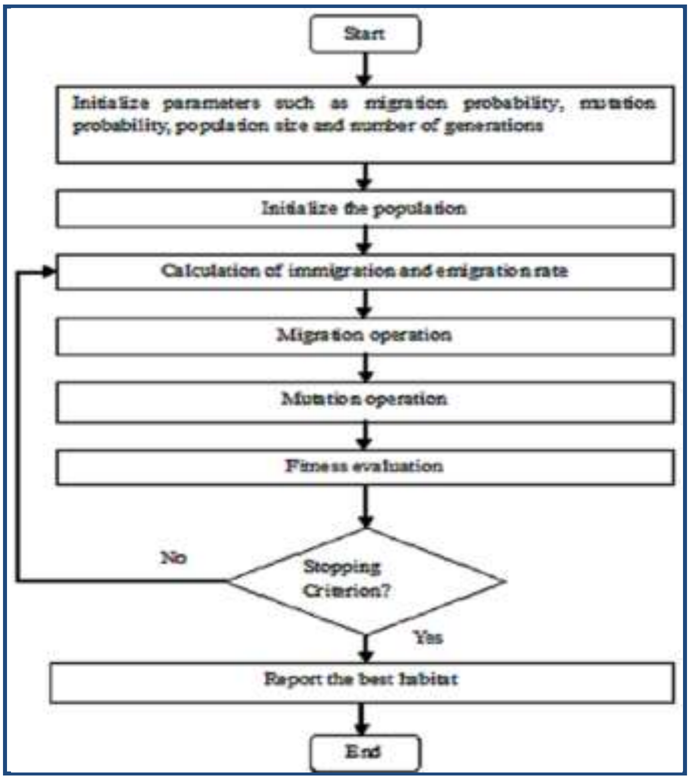
Start
Initialize parameters such as migration probability, mutation probability, population size and number of generations
Initialize the population
Calculation of immigration and emigration rate
Migration operation
Mutation operation
Fitness evaluation
No
Stopping Criterion?
Yes
Report the best habitat
End

Figure 1: Various steps in the BBO algorithm.

Data description

Two different data sets are employed in this study. The data sets are colon tumor and prostate tumor data sets. Colon tumor data set is obtained from Kent Ridge Biomedical Data Set Repository. The data set include samples of 62 patients and 2000 genes. Among the 62 samples, 40 tumor biopsies are from tumors and 22 normal biopsies are from healthy parts of the colons of the same patients. Prostate tumor data set was downloaded from http://www.gems-system.org. This data set contains 102 samples and 10,509 genes. Among them 52 are tumor samples and 50 are non tumor samples.

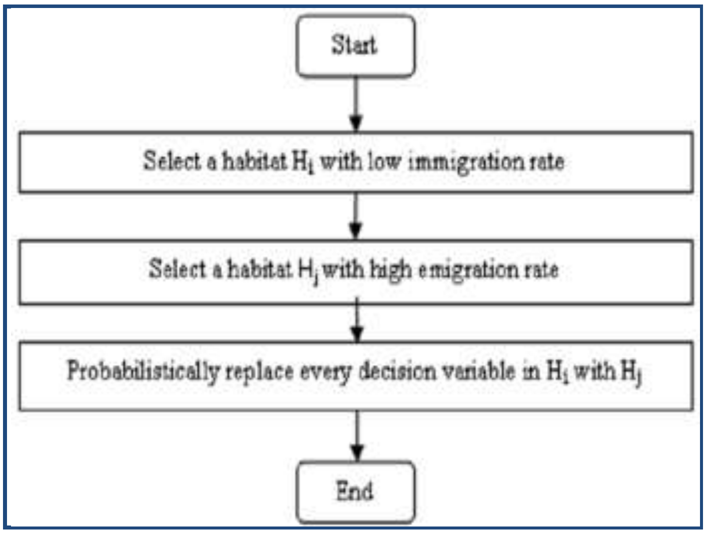
Start
Select a habitat H_i with low immigration rate
Select a habitat H_j with high emigration rate
Probabilistically replace every decision variable in H_i with H_j
End

Figure 2: Various steps in the migration procedure.

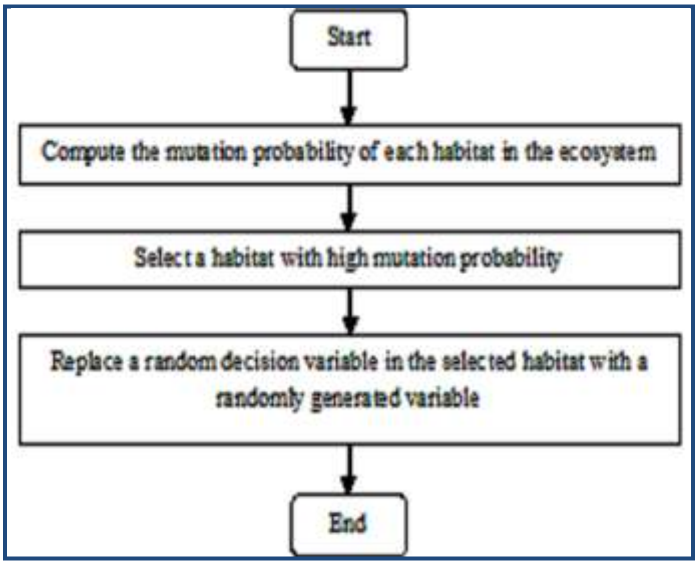
Start
Compute the mutation probability of each habitat in the ecosystem
Select a habitat with high mutation probability
Replace a random decision variable in the selected habitat with a randomly generated variable
End

Figure 3: Various steps in the mutation procedure.

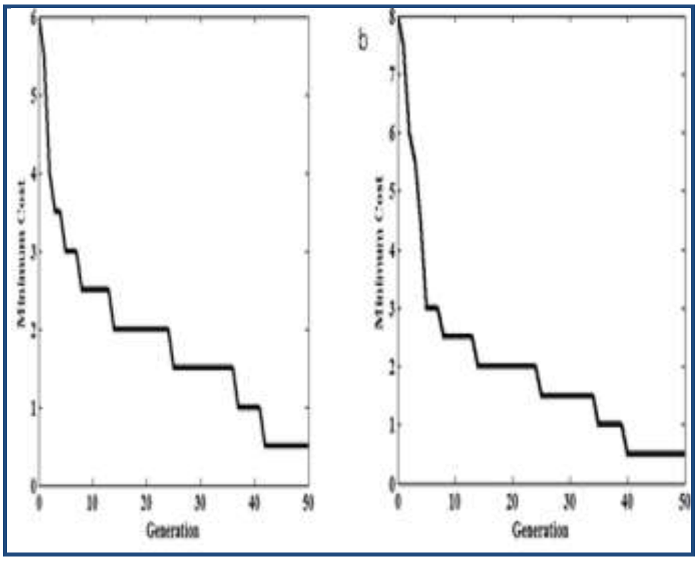

Figure 4: Minimum cost vs. Generation obtained while employing IG-BBO-KNN: **a)** Colon tumor data set; **b)** Prostate tumor data set.

Results & discussion:

Our study started with two main objectives related to feature selection. The first objective was to reduce the computational complexity of BBO (wrapper approach) by using an optimal filter method and the second was to select an optimal classifier that could give good accuracies with the wrapper method. To meet the first objective we have employed a filter method prior to BBO. Filter methods are substantially faster compared to wrapper methods but at the same time doesn't explore the relevance of features in combination with other features but this drawback can be alleviated by employing the wrapper method in the second stage. We have employed Correlation coefficient and information gain as filter approaches and compared their performances when employed with BBO. We have used weka for calculating the information gain values of all the attributes, and have selected the attributes with non zero 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 **[13].** 135 out of 2000 attributes were selected from the colon tumor data set and 2016 out of 10509 attributes were selected from the prostate tumor data set after the filtering phase based on information gain. We have calculated the correlation coefficient matrix between genes in the data sets using equation 1. This matrix represents the normalized measure of strength of linear relationship between genes. Values close to 1 indicate that there is positive linear relationship between the data columns. Values close to -1 indicate that one column of data has negative linear relationship to another column of data (also known as anti-correlation).Values close to 0 indicate that there is no linear relationship between the data columns. We have selected the genes with maximum number of uncorrelations based on the correlation matrix. In the next stage we have used BBO for selecting the best set of genes out of the filtered ones. BBO gave better results with the genes filtered based on information gain.

Recently many research activities have been carried out regarding BBO. BBO has been applied to real-world optimization problems, including sensor selection **[11],** power system optimization, groundwater detection, and satellite image classification **[14].** Many extensions to BBO have also been proposed such as blended BBO **[15],** BBO with immigration refusal and BBO using evolutionary strategies. The various parameters that need to be set for BBO includes probability of migration, probability of mutation, number of generations and population size and we have set the parameters as 1, 0.005, 50 and 50 respectively. Two fold cross-validation is employed in BBO for evaluating the fitness of gene subsets. During two fold cross-validation the entire data set is divided into two folds- test set and training set and the classification over two rounds of evaluation is taken as the fitness measure. We have employed two classifiers such as ANN and KNN for cross validation and compared their performances. Back propagation algorithm is used as the learning method in ANN and Euclidian distance metric is used for finding the nearest neighbours in KNN. The parameter K in KNN was set as 1 since high values of K may make the borders between classes less distinct. For evaluating the performances of the various hybrid methods, we compared the best minima obtained using the various algorithms over 50 generations. IG-BBO-KNN could provide the best minimum with 50 generations compared with the other hybrid methods. The best minimum obtained by IG-BBO-KNN for both the data sets is 0.5. **Figure 4** shows the best minima obtained by IG-BBO-KNN on colon tumor and prostate tumor data sets over various generations. **Table 1** shows the average accuracies obtained with the four hybrid methods over 10 runs. Results from **Table 1 (see supplementary material)** indicate that IG-BBO-KNN gave the best performance on both data sets. The method achieves accuracies of 90% and 96% on colon tumor and prostate tumor data sets respectively. These results indicate that IG-BBO-KNN is the best model among the various methods employed in this study. The various performance parameters of the IG-BBO-KNN model such as sensitivity (true positive rate), specificity (true negative rate), positive predictive value and negative predictive value are shown in **Table 2 (see supplementary material)**. These parameters are calculated as in the following equations:

$$\text{Sensitivity} = TP/(TP + FN) \quad \rightarrow \quad (5)$$

$$\text{Specificity} = TN/(FP + TN) \quad \rightarrow \quad (6)$$

$$\text{Positive predictive value} = TP/(TP + FP) \quad \rightarrow \quad (7)$$

$$\text{Negative predictive value} = TN/(TN + FN) \quad \rightarrow \quad (8)$$

Where TP refers to true positives (people with cancer and tested positive), TN refers to true negatives (people without cancer and tested negative), FN refers to false negatives (people with cancer and tested negative) and FP refers to false positives (people without cancer and tested positive).

Conclusion:

The high dimensionality of the microarray expression data is a concern during gene selection. Therefore, the use of four hybrid feature selection methods (combines filter and wrapper procedures) is discussed. Analysis shows that these hybrid methods effectively simplify feature selection by reducing the number of required features. Genes filtered with information gain proved to be more informative for BBO when compared with Pearson correlation coefficient for use with the classifiers ANN and KNN for cross-validation and classification. The classification error rate obtained by the IG-BBO-KNN combination was the lowest of all the hybrid methods discussed here.

References:

[1] Huang J *et al.* *Comput Biol Med*. 2010 **40**: 698 [PMID: 20591424]

[2] Li B *et al. Comput Biol Med*. 2010 **40**: 802 [PMID: 20864095]

[3] Ding C & Peng H, *J Bioinform Comput Biol*. 2003 **3**: 185 [PMID: 15852500]

[4] Quinlan JR, *Machine Learning*. 1986 **81**: 106

[5] Battiti R, *IEEE Trans Neural Netw*. 1994 **5:** 537 [PMID: 18267827]

[6] Liu X *et al. BMC Bioinformatics*. 2005 **6**: 76 [PMID: 15790388]

[7] Oh IS *et al. IEEE Transn Pattern Anal Mach Intell*. 2004 **26**: 1424 [PMID: 15521491]

[8] Chuang LY *et al. Comput Biol Chem*. 2008 **32**: 29 [PMID: 18023261]

[9] Chuang LY *et al. Methods Inf Med*. 2010 **49**: 254 [PMID: 20135079]

[10] Chuang LY *et al. J Comput Biol*. 2009 **16**: 1689 [PMID: 20047491]

[11] Zhang H & Sun G, *Pattern Recognition*. 2002 **35**: 701

[12] Simon D, *IEEE transactions on evolutionary computation*. 2008 **12**: 702

[13] Frank E *et al. Bioinfomatics*. 2004 **20**: 2479 [PMID: 15073010]

[14] Panchal V *et al. International Journal of Computer Science and Information Security*. 2009 **6**: 269

[15] Ma H & Simon D, *Engineering Applications of Artificial Intelligence*. 2011 **24:** 517

Edited by P Kangueane

Citation: Ammu & Preeja, Bioinformation 9(16): 824-828 (2013)

Supplementary material:

Table 1: Accuracies obtained using various feature selection methods

Data set	PPMC-BBO-KNN	PPMC-BBO-ANN	IG-BBO-KNN	IG-BBO-ANN
Colon tumor	74%	75.16%	90%	82.34%
Prostate tumor	85%	90%	96%	65%

Table 2: Various performance parameters of the IG-BBO-KNN model

Data set	Sensitivity	Specificity	Positive predictive value	Negative predictive value
Colon tumor	88%	96%	97%	81%
Prostate tumor	96%	96%	96%	96%