Research Journal of Applied Sciences, Engineering and Technology 11(10): 1071-1074, 2015
DOI: 10.19026/rjaset.11.2121
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: June 12, 2015 Accepted: August 5, 2015 Published: December 05, 2015

Research Article

Feature Selection of Microarray Data using Bacterial Foraging Optimization

Sunita Beniwal and Dharminder Kumar
Department of Computer Science and Engineering, Guru Jambheshwar University of Science and Technology, Hisar-125001, Haryana, India

Abstract: This study aims at finding the genes responsible for lung cancer using bacterial foraging optimization. Microarray datasets can be used to predict the presence of cancer, the type of cancer, its stage etc. Microarray datasets available have large number of features and few samples. Bacterial foraging optimization algorithm has been used in our study for feature selection on lung cancer dataset. BFO algorithm selects few genes from the available set. The reduced dataset is then used for designing a classifier using support vector machines which gives an accuracy of about 99% classifying only one sample inaccurately.

Keywords: Classification, microarray, preprocessing, SVM

INTRODUCTION

Data mining is a term used for extraction of useful information from large amounts of data. Data mining cannot be straight away applied to most of the datasets. Some steps needs to be carried out before mining patterns from data. Those steps applied before data mining are termed collectively as preprocessing. Steps like data integration, data cleaning, data selection, data transformation constitute data preprocessing (Kumar and Beniwal, 2013). Data of patients suffering from one or other disease is widely available. Data mining can be applied to the data for finding out the cause and effects of the disease. A classifier can be used for classification of data. It can be used for detecting presence of disease, its type, its stage etc.

DNA microarray offers the ability to measure levels of expressions of thousands of genes simultaneously. The use of microarrays to discover genes, which are differentially expressed between two or more groups of patients has many applications. These include the identification of disease biomarkers that may be important in the diagnoses of the different types and subtypes of diseases (Margalit *et al*., 2005).

To extract knowledge and useful information from microarray gene expression datasets an accurate method is required for using the data in diagnosis (Aguilar-Ruiz, 2005). Microarray gene expression datasets usually consists of large number of genes and few samples which makes it very difficult to extract information and patterns present in the data. So gene selection is carried out to select the genes with maximum variations that is the most informative and relevant genes. The genes whose value don't vary much in the samples and remains approximately same are removed and are not used for building the classifier. Recently research has focused on application of data mining for classification of microarray data and to identify genes that are differentially regulated during different disease (Hewett and Kijsanayothin, 2008).

In the present paper Bacterial Foraging Optimization (BFO) technique has been used for gene selection. BFO algorithm proposed by Passino (2010) is an optimization technique based on the foraging behavior of Bacteria *E. coli*. Given suitable conditions and sufficient food, *E. coli* bacteria grows at a very fast rate. The bacteria moves very fast into nutrient areas and tries to go away from noxious substances. The motions of bacteria are known as taxes. The bacterial foraging process consists mainly of four mechanisms namely chemotaxis, swarming, reproduction and elimination-dispersal. Bacteria have a tendency to gather to the nutrient-rich areas by activity called Chemotaxis. An *E. coli* bacterium can move in two different ways: it can run or tumble and alternates between these movements throughout its travel in search of food. In BFO, a unit walk with random direction represents a tumble and a unit walk with the same direction in the last step indicates run. A bacterium at times of stress release attractants to signal the bacteria to swarm together. Each bacterium also releases repellant to signal the others to be at a minimum distance from it. Thus all of them will have a cell to cell attraction via attractant and cell to cell

Corresponding Author: Sunita Beniwal, Department of Computer Science and Engineering, Guru Jambheshwar University of Science and Technology, Hisar-125001, Haryana, India

repulsion via repellant. After the completion of chemotactic steps, reproduction takes place. Fitness value of the bacteria is stored in ascending order and half of them with the worst value are eliminated. The remaining half is duplicated so as to maintain a fixed swarm size. The bacteria are diversified either gradually or suddenly to eliminate the probability of bacteria being stuck around the initial or local optima positions or global optima is obtained.

BFO has been used with PSO and GA for tuning controllers (Anguluri *et al*., 2011; Kim *et al*., 2007) and has also been used to train neural networks (Zhang *et al*., 2010) and gave good results. However BFO has not been reported till date for gene mining. Keeping this in view the present work reports application of BFO for feature selection in microarray data.

The present work focuses on analysis of microarray data for identifying the presence or absence of cancer. In the present work, lung cancer dataset was used. Before using the data, dataset was preprocessed and then used for training a classifier. Preprocessing of the dataset is required to eliminate the irrelevant genes and to reduce the dimensionality of the dataset. Relevant genes were selected using Bacterial Foraging Optimization (BFO) algorithm and then support vector machine was used to train the classifier on reduced dataset.

METHODOLOGY

Dataset used in the present work is a microarray dataset consisting of 96 samples, out of which 86 samples are of lung cancer patients and 10 samples are of normal healthy persons (Beer *et al*., 2002). Each sample consists of 7129 genes. Out of the 96 samples, 48 were used for training and remaining samples were used for testing the accuracy of classifier. The whole work has been carried out using Matlab. The dataset used in the analysis has large number of dimensions and very few samples. So to avoid over fitting by building a classifier using so many features, feature selection was done. Feature selection is selection of only relevant genes i.e., only those genes are selected which plays a role in cancer and are differently expressed in cancer patients.

Feature selection: For selection of relevant features we have used BFO algorithm. It gives us a small subset of total genes which can then be used for building a classifier. Fitness function used by BFO depends on mean of all features of both classes and total number of samples in each class. The values of attractant and repellant are taken so as to signal other bacteria. Fitness of each gene is calculated using mean value and variance in both classes.

The values of attractant and repellant also effects the chances of each gene. After calculating fitness of each gene, a unit vector is added to the gene's value and fitness is recalculated. If the fitness improves the gene is further processed in the same way i.e., the bacteria swim else the gene's fitness is no more calculated i.e., bacterium tumbles. The bacterium swim length is initialized beforehand. The whole process is done with each gene and number of chemotactic steps tells the number of times the process is carried out. For reproduction the genes are arranged in the descending order. Best fit (top 50%) genes reproduce by making their identical copies and least do not reproduce. After completely executing the BFO algorithm, a threshold value is calculated for retaining the genes. All genes whose fitness value is less than the threshold value are removed from the list and only the genes with best fitness are retained.

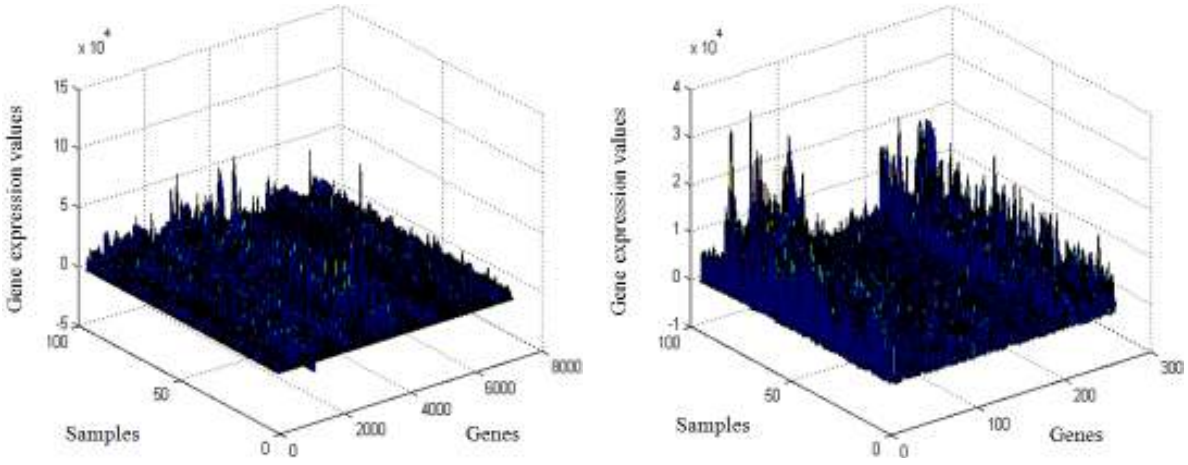
Classification: For checking whether relevant genes have been selected or not by BFO, the genes need to be used to classify the dataset. Support Vector Machines (SVM) are used for classification in the present research. The dataset is divided into two sets : training and test set for designing a classifier. The method used for dividing the datset into two is crossvalidation with holdout option which divides the dataset in such a way that the training and test dataset contain equal number of samples. In our study the training and test set contains 48 samples each.

After the dataset is divided into two sets, the SVM classifier is designed on the training dataset. The accuracy of the classifier designed is then tested on the test dataset.

RESULTS AND DISCUSSION

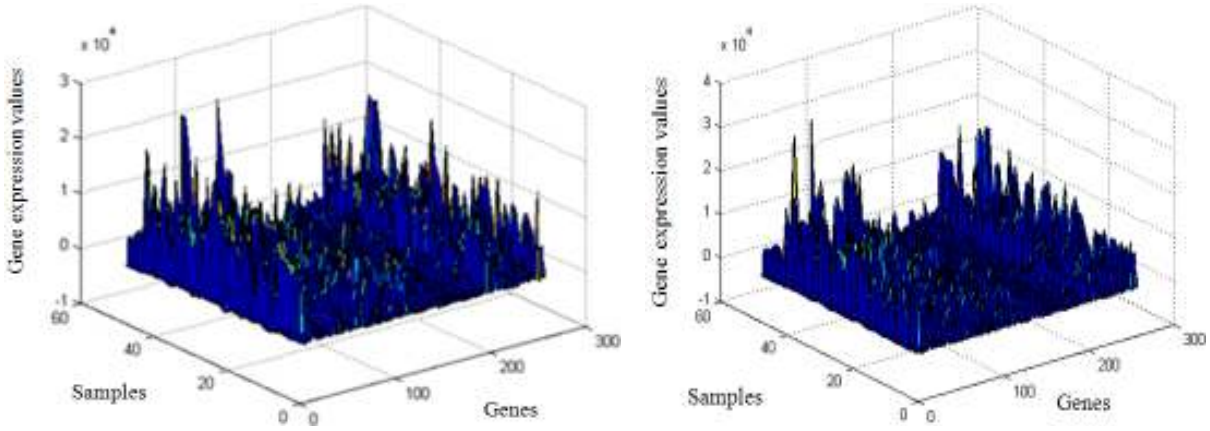
BFO algorithm when used for gene selection selected 260 genes out of 7129 genes. Figure 1 shows the bar graphs of original features and reduced features of the dataset. The x-axis shows the number of samples, the y-axis is expression value of genes in various samples with the z-axis showing the gene number. As can be seen from Fig. 1a the values of genes in the original dataset ranges from negative values to values in lakhs. Figure 1b shows the reduced dataset i.e., after only relevant genes are left. After reduction the number of features is less. So it can be seen that BFO selected lesser number of genes as compared to original dataset. The genes have come down from 7129 to only 260.

Figure 2a and b shows the distribution of training samples and test samples respectively. Training and test datset have equal number of samples i.e., 48 samples. As can be seen the distribution of samples is very uniform with both having approximately same distribution. Training dataset is used for building the classifier i.e., SVM in our work. Once the classifier is made, it need to be tested which is done using test dataset. Figure 3 shows the result of SVM being applied to the dataset and its results. It is evident from the figure that only one sample is classified incorrectly. The incorrecly classified sample is of a healthy person and not of a cancer patient. All other samples are



(a) (b)

Fig. 1: Comparison of dataset before and after BFO application; (a): Original dataset; (b): Reduced dataset



(a) (b)

Fig. 2: Distribution of samples; (a): Training dataset; (b): Test dataset

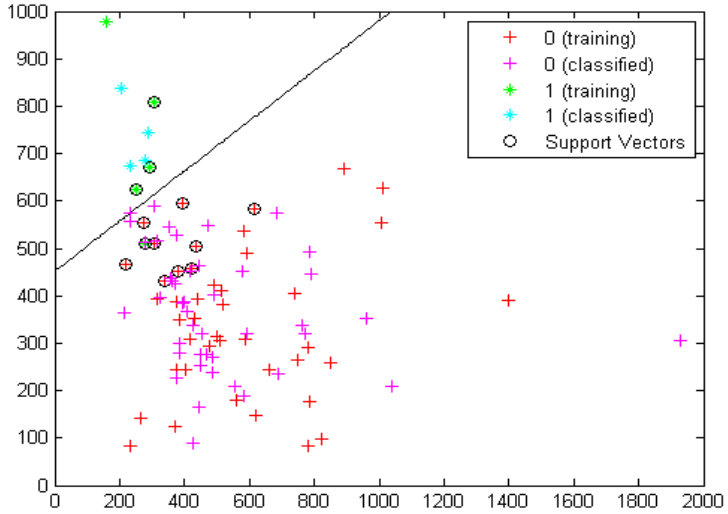


Fig. 3: Output of SVM classifier on the lung cancer dataset

classified correctly by the classifier taking the accuracy to approximately 99%.

CONCLUSION

BFO algorithm was applied to select relevant genes and it selected 260 genes and then those genes were used for designing a classifier with the help of SVM which gave a good accuracy. The number of genes selected by BFO is still very large, but it gives good results as far as accuracy is concerned. BFO algorithm can be combined with some other feature selection techniques to reduce the number of genes selected. Also the efficieny of technique can be tested on other microarray datasets.

REFERENCES

Aguilar-Ruiz, J., 2005. Shifting and scaling patterns from gene expression data. Bioinformatics, 21(20): 3840-3845.

Anguluri, R., A. Ajith and S. Vaclav, 2011. A hybrid bacterial foraging: PSO algorithm based tuning of optimal FOPI speed controller. Acta Montan. Slovaca, 16(1): 55-65.

Beer, D.G., S.L. Kardia, C.C. Huang, T.J. Giordano, A.M. Levin, D.E. Misek, L. Lin, G. Chen, T.G. Gharib, D.G. Thomas, M.L. Lizyness, R. Kuick, S. Hayasaka, J.M. Taylor, M.D. Iannettoni, M.B. Orringer and S. Hanash, 2002. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat. Med., 8(8): 816-824.

Hewett, R. and P. Kijsanayothin, 2008. Tumor classification ranking from microarray data. BMC Genomics, 16(9Suppl. 2): S21.

Kim, D.H., A. Abraham and J.H. Cho, 2007. A hybrid genetic algorithm and bacterial foraging approach for global optimization. Inform. Sciences, 177: 3918-3937.

Kumar, D. and S. Beniwal, 2013. Genetic algorithm and programming based classification: A survey. J. Theor. Appl. Inf. Technol., 54(1): 48-58.

Margalit, O., R. Somech, N. Amariglio and G. Rechavi, 2005. Microarray-based gene expression profiling of hematologic malignancies: Basic concepts and clinical applications. Blood Rev., 19(4): 223-234.

Passino, K.M., 2010. Bacterial foraging optimization. Int. J. Swarm Intell. Res., 1(1): 1-16.

Zhang, Y., W.U. Lenan and S. Wang, 2010. Bacterial foraging optimization based neural network for short-term load forecasting. J. Comput. Inform. Syst., 6(7): 2099-2105.