Journal of Engineering Design and Analysis
Volume 3, Issue 2 - 2020, Pg. No. 85-87
Peer Reviewed Journal

Article

A General Study on Feature Selection for Cancer Classification

Santosini Bhutia[1], Biswajit Tripathy[2]

[1]Indira Gandhi Institute of Technology, Sarang, Dhenkanal, Odisha, India.
[2]Gandhi Institute for Technological Advancement, Bhubaneswar, Odisha, India.

I N F O

Corresponding Author:
Santosini Bhutia, Gandhi Institute for Technological Advancement, Bhubaneswar, Odisha, India.
E-mail Id:
biswajit_cse@gita.edu.in
Orcid Id:
https://orcid.org/0000-0001-6069-5658
How to cite this article:
Bhutia S, Tripathy B. A General Study on Feature Selection for Cancer Classification. *J Engr Desg Anal* 2020; 3(2): 85-87.

Date of Submission: 2020-09-10
Date of Acceptance: 2020-12-10

A B S T R A C T

In the context of cancer research microarray experiments are the most powerful mechanism for the diagnosis of disease. It has the ability to identify the characteristics of gene expression pattern. But DNA microarray experiment produces a huge number of features or genes which is usually more than thousands for a few number of samples or subjects which is less than hundreds.[1] To date this problem there are various efficient classification and good feature selection methods are implemented to reduce the complexity and advance the cost. In this paper we on the methodologies for feature selection to identify important genes that improve the accuracy of classification.

Keywords: Microarray Data, Feature Selection, Classification Method

In the last two decades a newfangled area of research has been enlarging in Biology, Bioinformatics and Machine Learning. These fields of interest are analyzed by microarray gene expression data. It is a big challenge for the researchers to investigate from a large number of genes using traditional methods. DNA microarray is one of the popular technologies for the solution. It enables the researchers to examine and find out the condition of the growth and development of life and also analyze the genetic cause of abnormality arising in the functioning of the individual. In recent studies, one of the application DNA microarray technologies is to learn about various diseases like heart disease, mental illness, infectious disease and the study of cancer. Formerly the types of cancer are classified by the organs in which the tumors have been developed. But now it is possible in microarray technology to classify the types of cancer by observing the pattern of gene activity present in the tumor cells. Here we principally focus on tumor classification using gene expression data. In modern days it is a contemporary research area for the researchers to explore in the field of biological and medical science.

DNA microarray cancer gene expression normally brings about enormous number of features which varies from 2000 to 60000, but the datasets assign to minute no of samples that is varies from 20 to 80.[2] For the sake of enormous number of gene expression and minute number of samples, the classification problem is a challenging work for the researchers. Hence to reduce the ramification and to enhance the efficiency, a robust model is required for feature selection and classification. The classification issue is associates with two distinct type of activity: binary classification and multiclass classification. In binary classification it analyzes the given sample is cancerous or not, and in multiclass classification it analyzes different varieties of tumors.

In the recent years many researchers focus on obtaining the relevant features to improve the classifier accuracy

and present significant report to the medical science. The feature selection method selects the optimal features from a huge number of features that are available in real life application. The two basic feature selection methods are filter method and wrapper method.

There are recent studies that focus on the cancer reading through the microarray gene appearance to help the doctors in their diagnosis process. In the last decades, Support Vector Machine (SVM) attracted more attention for classification of binary microarray data from the researchers, and for multiclass microarray data there have been various proposed methods these are DAGSVM, Evolutionary SVM (ESVM), Genetic Algorithm based SVM (GASVM), and Fuzzy SVM (FSVM).

Feature Selection

Feature selection is a crucial matter in classification, which is the process of reducing the dimensionality of huge data to pick up the best path for concession solution. It can also reduce the complexity of the data and relevant for the microarray data. It helps in understanding the higher classification accuracy which is a major research in this area. It involves in anticipating the class membership of the data, generating the correct label of the training data and also predicting the labels of the unknown data.

Conclusions Classification Methods

Linear Support Vector Machine

Support Vector Machine (SVM) is a powerful data mining technique which is developed by Vapnik.[3] It has been broadly used in the ongoing research field of computational biology. Depending on statistical learning theory, SVM results high classification accuracy and resilience in modeling distinct source of data. It is mostly convenient for analyzing microarray gene expression data and also used in binary classification and regression.[4] Let us assume we have to develop a classifier whose function is to generate a hyper plane. This hyper plane is used to distinguish the positive and negative samples. But generally in real time problem it is not an ideal solution to distinguish the negative samples from the positive samples. To solve this problem, SVM is provided with a set of training samples which is used to map into the possible high dimensional feature space that generate a hyper plane which separates the samples. To generate a separating hyper plane the kernel function.[5] is:

$$x_i \in R^d, \{ x_i, y_i \}, i = 1,.....N, \{y_i / y_i \in \{-1,1\}\}$$

By using the kernel function SVM maps the data to a high dimensional space.

Sequential Minimal Optimization (SMO)

In 1998 Sequential Minimal Optimization (SMO)[6] algorithm is invented by John Platt. This algorithm is generally used for training the SVM. The quadratic problem which is generated through the training of SVM can also be solved by SMO algorithm. In this paper Sequential minimal optimization problem is used to train the SVM and to construct the maximum margin hyper plane.

Multilayer Perceptron (MLP)

Multilayer perceptron is a class of feed forward artificial neural network. It consists of three layers i.e. input layer, hidden layer and output layer. Here each node is a neuron which uses a nonlinear activation function except the input node. For training MLP uses a supervised learning technology called back propagation. MLP can separate the data which are not linearly separable. But when it is having a single hidden layer then it is same as ordinary neural network.

Related Work

There are several filter methods recycled with many Searching algorithms to evaluate the ability of each feature and calculate the important features from the input microarray dataset provided. There after different neural network classifiers have been applied to evaluate the prediction of accuracy.

CFSES optimization Feature Selection with neural network classification for microarray data analysis is proposed by B Patra and S S Bisoyi.[7] In this study, Correlation Based Feature Selection (CFS) with Elephant Search (ES) Algorithm is used for selecting significant genes and then Linear SVM, SMO-SVM and MLP neural network classifiers are applied to the reduced datasets to evaluate the accuracy for cancer classification.

Elephant Search with Deep Learning for Microarray Data Analysis[8] is proposed by M Panda. In this study, the effectiveness of microarray gene expression profiling is determined with FFS and ES based deep learning respectively. Deep learning works well for almost in all datasets besides a complicated model is chosen to learn from an easy problem.

Conclusion

In this paper, we discuss on feature selection methods to eliminate the insignificant and redundant genes. There after the neural network classifiers are applied to the reduced dataset. Hence it reduces the complexity of the problem and then come to a good accurate classification decision based on the experiment.

References

1. Lakhani, Sunil R, Ashworth A. Microarray and histopathological analysis of tumours: the future and the past?." Nature Reviews Cancer 1.2 (2001): 151-157.
2. Harrington, Christina A, Rosenow C et al. Monitoring

gene expression using DNA microarrays. *Current opinion in Microbiology* 2000; 3(3): 285-291.
3. Vapnik V. Statistical learning theory., in: 1998
4. Dash S, Patra BN. A Hybrid Data Mining Technique for Improving the Classification Accuracy of Microarray Data Set. I.J. *Information Engineering and Electronic Business* 2012; 4(2): 43-50.
5. Chow TWS, Cho SY. Neural networks and computing : learning algorithms and applications, Imperial College Press, Distributed by World Scientific, *London*, Singapore, Hackensack, NJ 2007.
6. Platt J. Fast training of support vector machines using sequential minimal optimization. Advances in Kernel Methods Support Vector Learning. 1998, MIT Press.
7. Patra BN, Bisoyi SS. CFSES optimization feature selection with neural network classification for microarray data analysis. International Conference on Data Science and Business Analytics (ICDSBA) 2018.

ISSN: 2582-5607