www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 2 Issue 6 June, 2013 Page No. 1845-1849

'CLASSIFICATION OF EXACT IDENTIFICATION OF CANCER USING EXPRESIONS OF SUPPORT VECTOR MACHINES WITH FUZZY C-MEANS CLUSTERING

A. Nirmala, G. SudhaAnanthi,

Assistant Professor, Dr. N.G.P Arts & Science College Coimbatore, Tamilnadu, India
Research Scholar, Dr. N.G.P Arts & Science College, Coimbatore, Tamilnadu, India

Abstract:

Microarray technique is used in gene for classification and identification of cancer parts. As many machine learning and data mining techniques are implemented for finding the cancer using the expression of gene but those technique were not sufficient. Microarray data have a high degree of noise. The disadvantage of existing technique is that it work out with the drawbacks such as noise. Ranking of gene method overcome the problems in proposed technique. Commonly developed Gene ranking techniques would wrongly predict the rank when large database used. Inorder to overcome the drawbacks in the exisiting technique the paper proposes a technique called Score for ranking the gene .The classifier used in the proposed technique is Support Vector Machine (SVM) with Fuzzy C-Means Algorithm. The experiment is performed on data set lymphoma and the result shows the accuracy of classification is best when compared to the older method.

Keywords: Support vector machines, Fuzzy c-means clustering, machine learning, Lymphoma data set.

1. INTRODUCTION

In the field of medicine cancer is the important research area. To predict the accurate cancer of various types by giving the treatment better and minimizing the toxicity to the patients. In Olden days the classification of cancer is based only by the method of clinical and morphological. In previous cancer classification method there are so many restrictions in diagnosing the cancer. The therapies are used according to the tumor types and can be distinguished by patterns so that it can maximize efficiency of patients. The exisiting technique is found to be heterogeneous and it follows different clinical methods.

Cancer is one of the diseases found in the human being and it is a challenge to make a research in this area. Cancer (Alter *et al*., 2003) is fundamentally described by an abnormal, uncontrolled growth that may demolish and attack other healthy body tissues. There are billions of

cells in the human body and most of the cells have an inadequate life-span. Each cell is capable of replicating themselves. Millions of cell divisions and duplication occur daily in the body. To gain a better remedy into the problem of cancer classification, systematic approaches that is been based on gene expression analysis. The expression level of gene contains the key to address fundamental problems relating to the prevention and cure of diseases. The advanced microarray technology has allowed the sequence monitoring of thousands of genes, which paved a way in the development of cancer classification using gene expression data. Classification problem has been extensively studied by researchers in the area of statistics, machine learning and databases. Many classification algorithms have been proposed in the past, such as the decision tree methods, the linear discrimination analysis, the Bayesian network, etc. For the last few years, researchers have started paying attention to the cancer classification using gene expressions. Studies have shown that gene expression changes are related with different types of cancers.

The gene expression data is very different from any of the data these methods had previously dealt with. First, it has high dimensionality, usually contains thousands to tens of thousands of genes. Second, data size is very small. Third, most genes are irrelevant to cancer distinction. The existing classification methods were not designed to handle this kind of data efficiently and effectively. Cancer classification using gene expression data enhance the performances of the classifiers in classifying gene expression data. Through this paper, hope to gain some insight into the problem of cancer classification in aid of further developing more effective and efficient classification algorithms.

2. RELATED WORK

Using exisiting training examples from cancer and normal patients, the approach build a classifier suitable for diagnosis, as well as drug discovery. Previous attempt address this type of problems. The new method of gene selection utilizing Support Vector Machine methods based on Lymphoma data.It is experimentally demonstrated that the genes selected by our techniques yield better classification performance and are biologically relevant to cancer. Hernandez *et al*. (2007) presents a Genetic Embedded Approach for Gene Selection and Classification of microarray data requires the selection of subsets of relevant genes to achieve good classification performance. that performs the selection task for a SVM classifier.

Machine Learning is considered as a subfield of Artificial Intelligence and it is concerned with the development of techniques and methods which enable the computer to learn. In simple terms development of algorithms which enable the machine to learn and perform tasks and activities. Machine learning overlaps with statistics in many ways. Over the period of time many techniques and methodologies were developed for machine learning tasks [1]Support Vector Machine (SVM) was first heard in 1992, introduced by Boser, Guyon, and Vapnik in COLT-92. Support vector machines (SVMs) are a set of related supervised learning methods used

for classification and regression [1]. They belong to a family of generalized linear classifiers. In another terms, Support Vector Machine (SVM) is a classification and regression prediction tool that uses machine learning theory to maximize predictive accuracy while automatically avoiding over-fit to the data. Support Vector machines can be defined as systems which use hypothesis space of a linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory. Support vector machine was initially popular and now is an active part of the machine learning research around the world. SVM becomes famous when, using pixel maps as input; it gives accuracy comparable to sophisticated neural networks with elaborated features in a handwriting recognition task [2]. It is also being used for many applications, such as hand writing analysis, face analysis and so forth, especially for pattern classification and regression based applications. The foundations of Support Vector Machines (SVM) have been developed by Vapnik [3] and gained popularity due to many promising features such as better empirical performance. The formulation uses the Structural Risk Minimization (SRM) principle, which has been shown to be superior,to traditional Empirical Risk Minimization (ERM) principle, used by conventional neural networks[4]. SRM minimizes an upper bound on the expected risk, where as ERM minimizes the error on the training data. It is this difference which equips SVM with a greater ability to generalize, which is the goal in statistical learning.[6] SVMs were developed to solve the classification problem, but recently they have been extended to solve regression problems.

The very active cancer research over the last decades has brought a better understanding of this complex and dynamic disease. It has been shown that cancer arises from a trans-formation of a single cell that acquires abnormal proliferation and invasion abilities. The probability of such a malignant change in the genome may be increased by inherited factors as well as environmental agents (exposure to chemicals such as tobacco, certain infections such as HPV or Hepatitis B and C).The transformation of a healthy cell into a malignant tumor cell is typically a slow and multi-step process. A cell has to undergo several steps of mutations in order to overcome the body defence mechanisms. Thus Hanahan and Weinberg [2000] think that the genome of cancer patients must have an “increased mutability in order for the process of tumour progression to reach completion in several decades”. Parametric models have nearly entirely been replaced by a semi parametric model. It is named after the British statistician Cox who first introduced this less parametric approach to proportional hazard . [5,9]Several machine learning Techniques have been developed for the examination of Microarray data.[7]The grouping of gene microarray Method and machine learning technique assures new Approaches into mechanisms of living schemes. An Application field where these methods are likely to create key Contributions is the identification of cancers depends on Clinical phase and biological activities. Such

classifications have a huge contribution on diagnosis and treatment. This technique enhances the performances of the classifiers in classifying gene expression data [8]. Xiyi et al., [10] given a cancer classification technique by sparse representation using microarray gene expression data.

3. STAGES IN PROPOSED TECHNIQUE

There are two stages included in the proposed technique. In the first stage, each gene in the training data are ranked with a scoring technique .In the second stage, the classification of combination of the selected genes is tested with the help of a classifier called Support Vector Machine with fuzzy c means clustering.This stage defines the importance of ranking top hundred genes with scoring technique.

3.1 Identifying the gene subset : This stage classifies the given data set using single gene after choosing several top ranked genes in the ranking list. Each gene that is selected is given as an input to the classifier. When good accuracy is not found, it is needed to categorize the data set with combination of two genes within the selected genes. Even if the good accuracy is not found, this procedure is repeated with all of the three gene combinations and so on until the better accuracy is obtained.

3.2 Ranking the gene with the help of score

This stage determines the necessity of ranking the gene with the help of scoring phenomenon.

As support vector machines are linear classifier that has the capability of finding the optimal hyper plane that increases the separation among patterns, this characteristic creates support vector machines as a potential means for gene expression data examination purposes. The fold cross validation (CV) is performed for support vector machine in the training data set. First split the entire dataset randomly as training (t1) and testing (t2) data . The genes are ranked with the help of samples of T1. The combination (TC1) is produced with the help of 2 genes from 100. Then TC1 is split into 5 folds (Tc1, Tc2, Tc3, Tc4 and Tc5). Among these folds one fold is chosen for testing. The other 4 folds are used as a classifier for SVM with fuzzy c-means. This combination is occurs sequentially and stops only when the better accuracy is achieved. At last with the fitted SVM with fuzzy c-means, the prediction can be carried out.

4. EXPERIMENTAL RESULTS

The experimentation on the proposed method is carried on lymphoma data set . In the lymphoma data set, there are 42 samples derived from Lymphoma, nine samples from Lymphoma and 11 samples from Lymphocytic. The expression data of 4026 genes are included in the entire data set. Very few parts of data are missing in this data set. For filling those missing values k-nearest neighbor algorithm was applied.

5. CONCLUSION AND FUTURE WORK

Cancer is one of the important aspect in the field of bio-medicine. Exact prediction of several tumor kinds has greater value in offering treatment and toxicity on the patients. In the olden days, cancer categorization is generally depends

on morphological and clinical analysis. These methods used before for cancer classification techniques stated to have many disadvantages in their diagnosis.

The gene ranking technique is used to support the paper. This paper uses scoring phenomenon for ranking the gene. Then the classifier is trained with that dataset. The classification of gene for identifying the cancer is been obtained. The classifier used in this paper is support vector machine with fuzzy c-means clustering. The experiment is performed with the help of lymphoma data set. The experimental result shows that the proposed technique results in better accuracy and consumes less time for classification when compared to the previous techniques. The performance of the proposed approaches is evaluated based on measures such as accuracy. The experiments are performed in data sets namely lymphoma cancer data set. The experimental results show that the proposed fuzzy c- means clustering approach shows significant performance in terms of classification accuracy. This is due to the salient features of the proposed fuzzy c - means clustering approach which provides better performance because of the advantages of SVM.

The datasets are used Lymphoma in the experimental result shows that the proposed method performs the cancer classification with exact accuracy. In order to improve the efficiency, this research requires some future enhancement. In future, better neural network techniques can be incorporated with the present research work for less complexity and better learning capacity. Moreover, better Neuro fuzzy techniques could also be used to improve the classification rate and accuracy. Better machine learning techniques can also be implemented for better learning capability and speed.

6. BIBLIOGRAPHY

1. Boser, B., Guyon, I., & Vapnik, V. (1992). An training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (pp. 144–152). Pittsburgh: ACM.

2.Spellman,P.T,Sherlock,G.,Zhang,M.Q,IyerV.R, Anders,K.,Eisen,M.B,Brown,P.O,Botstein,D.&Fu tcher,B.(1998) Mol.Biol.Cell 9, 3273–3297.

3. Vapnik, V. N. (1998). *Statistical learning theory*. Wiley Interscience. Walsh, J. H. (1999). Epidemiologic evidence underscores role for folate as foiler of colon cancer. *Gastroenterology,116*, 3–4.

4.Bishop,C.(1995)NeuralNetworks forPatternRecognition (OxfordUniv.Press, NewYork).

5. Quinlan,J.(1997)in Programs for MachineLearning,SeriesinMachineLearning (MorganKaufmann,SanFrancisco)

6. Weston, J., Muckerjee, S., Chapelle, O., Pontil, M., Poggio, T., & Vapnik, V. (2000). Feature selection for SVMs. In *Proceedings of NIPS 2000*

7.Pavlidis, P., Weston, J., Cai, J., & Grundy, W. N. (2000). Gene functional analysis from heterogeneous data.

8 Ramaswamy, S., Tamayo, P. and Rifkin, R., "Multiclass Cancer Diagnosis using Tumor Gene Expression Signatures", PNAS, Pp. 15149–15154, 2001

9.Kohavi, R. & John, G. (1997). Wrappers for feature subset selection. *Artificial Intelligence, 97:12*, 273–324.

10. Xiyi Hang, "Cancer Classification by Sparse Representation using Microarray Gene Expression Data", IEEE International Conference on Bioinformatics and Biomeidcine Workshops, Pp. 174-177, 2008.