ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

CHALLENGES IN CANCER CLASSIFICATION

Mr.S.Senthil Kumar*, Mr.M.Suresh Kumar, Mrs.S.Kalai Selvi

* mca.,m.phil assistant proessor, department of commerce with computer applications dr.sns rajalakshmi college of arts and science (autonomous),Coimbatore, tamil nadu

m.com (ca)., m.phil., mba., m.a. assistant proessor, department of commerce with computer applications dr.sns rajalakshmi college of arts and science (autonomous), coimbatore, tamil nadu

m.sc.,m.phil.,b.ed assistant proessor, department of commerce with computer applications dr.sns rajalakshmi college of arts and science (autonomous), coimbatore, tamil nadu

ABSTRACT

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 expression. Studies have shown that gene expression changes are related with different types of cancers.

Most proposed cancer classification methods are from the statistical and machine learning area, ranging from the old nearest neighbor analysis, to the new support vector machines. There is no single classifier that is superior over the rest. Some of the methods only works well on binary-class problems and not extensible to multi-class problems, while others are more general and flexible. One thing to note for most of those proposed algorithms on gene classification is that the authors are only concerned with the accuracy of the classification and did not pay much attention to the running time(in fact, most gene classifier proposed are quite computationally expensive).

Recently, the neural network has become a popular tool in the classification of Cancer Dataset. This is particularly due to its ability to represent the behavior of linear or nonlinear functions multidimensional and complex.

KEYWORDS:

INTRODUCTION

Cancer classification using gene expression data stands out from the other previous classification data due to its unique nature and application domain. Through this survey, we hope to gain some insight into the problem of cancer classification in aid of further developing more effective and efficient classification algorithms.

With the advancement of microarray technology, gene expression profiling has shown great potential in outcome prediction for different types of cancers. Microarray cancer data, organized as samples versus genes fashion, are being exploited for the classification of tissue samples into benign and malignant or their subtypes. They are also useful for identifying potential gene markers for each cancer subtype, which helps in successful diagnosis of particular cancer type.The classification of cancer is a major research area in the medical field. Such classification is an important step in determining treatment and prognosis. Therefore, classification

using neural networks that is classifies objects rather simply they take data as input, derive rules based on those data, and make decisions.

CLASSIFICATION

Classification is a task of assigning an item to a certain category, called a class, based on the characteristic

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

features of that item. This task in any classification system is performed by a classifier that takes a feature vector as an input and responds with a category to which the object belongs. A feature vector is a set of features extracted from the input data. In our study the feature vector represents the twelve features extracted for each nucleus as illustrated above in the feature extraction phase. Here we make use of neural network classifiers that are a collection of neurons (systems with many inputs and one output that are trained to fire, or not, for particular input patterns) that are connected one to another. Each connection is assigned an initial weight during the training process which is then adjusted to give a proper answer.

The final decision is made based on the interaction of weights and the feature vector. The classification step was realized using four well known supervised classification algorithms: support vector machine, learning vector quantization, probabilistic neural networks and multilayer perceptron using back-propagation algorithm.

CHALLENGES IN CLASSIFICATION

There have been extensive studies done in the past on the classification problem by the statistical machine learning and database research community. But gene classification as a new area of research poses new challenges due to its unique problem nature. Here we elaborate on some of these challenges.

First challenge comes from the unique nature of the available gene expression data set. Though the successful application of cDNA microarrays and the high-density oligonucleotides have made fast simultaneous monitoring of thousands of gene expressions possible and inexpensive, the publicly available gene expression data set size still remains small. Most of these data, such as the Colon tissue samples, the Leukemia data set, etc., has sample size below 100, On the contrary, the attribute space, or the number of genes, of the data is enormous: there are usually thousands to hundred thousands of genes present in each tuple. If the samples are mapped to points in the attribute space, then the samples can be viewed as very sparse points in a very high dimensional space. Most existing classification algorithms were not designed with this kind of data characteristics in mind. Such a situation of sparseness and high dimensionality is a big challenge for most classification algorithms. Overfitting is a major problem due to the high dimension, while the small data size makes it worse. Also, with so many genes in the tuple, it will be a big challenge on the computation time. Therefore, developing an effective and efficient classification algorithm for cancer classification is not an easy task.

Second challenge comes from the presence of noise inherent in the data set. These noises can be categorized into biological noise and technical noise. Biological noise refers to the noise introduced by genes that are not relevant for determination of the cancer classes. In fact, most of the genes are not related to the cancer classes. Technical noise refers to the noises that are introduced at the various stages of data preparation whereas biological noises are associated with the non-uniform genetic backgrounds of the samples or the misclassification of the samples. Coupled with small sample size, the presence of noise makes accurate classification of data difficult.

Third challenge involves dealing with a huge number of irrelevant attributes (genes). Though irrelevant attributes are present in almost every kind of data sets researchers have dealt with previously, but the ratio of irrelevant attributes to the relevant attributes is not as huge as that in the gene expression data. In most gene expression data set, the number of relevant genes only occupies a small portion of the total number of genes. Most genes are not cancer related. The presence of these irrelevant genes interferes with the discrimination power of those relevant attributes. This not only incurs extra computation time in both the training and testing phase of the classifier, but also increases the classification difficult. One way to handle this is to incorporate a gene selection mechanism to select a group of relevant genes. Then cancer classifier can be built on top of these selected genes. Another way is to incorporate the selection of relevant genes inside the training phase of the classifier. Performing cancer classification efficient and effective using either way is a nontrivial process, thus requiring further exploration.

Fourth challenge arises from the application domain of cancer classification. Accuracy is important in cancer classification, but it is not the only goal we want to achieve. Biological relevancy is another important criterion, since any biological information revealed during the process can help in further gene function discovery and other biological studies. Some useful information can be gained from the classification process is the determination of the genes that work as a group in determining the cancerous tissues or cells or the genes that are under-expressed or over-expressed in certain tissues or cells. All these would help biologists in gaining more understanding about

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

the genes and how they work together and interact with each other. Therefore biologists are more interested in classifier that not only produce high classification accuracy but also reveal important biological information.

When a classification problem is defined by features, the number of features can be quite large, many of which can be irrelevant. A relevant feature can increase the performance of a classifier while an irrelevant feature can deteriorate it. Therefore, in order to select the relevant features, it is necessary to measure the goodness of selected features using a feature selection criterion. The class separability is often used as one of the basic selection criteria.

PATTERN CLASSIFICATION

Pattern classification is an important field of study in prospect of identifying problem domain in machine learning and has a wide range of application in engineering ,sciences ,medical and other analyzing fields whether a set of attributes aggregately decide the nature and class of the sample being analyzed [5]. Various intelligent techniques are used for classification purpose are decision tree, Neural Network, Genetic Algorithm.

With the development of science and technology, Neural Network is one of the techniques to solve pattern classification problem with reliability and intelligent manner with reduced fault as deciding parameters are spread throughout the network, collapse of one of the node have negligible impact on overall network. In this paper a Neural Network is used to classify Benign (non-cancerous) and Malignant (cancerous) cells.

BREAST CANCER CLASSIFICATION PROBLEM

Classification is a task that is often encountered in everyday life. A classification process involves assigning objects into predefined groups or classes based on a number of observed attributes related to those objects. The use of data mining approaches and learning machine has revolutionized the process of classification breast cancer data set getting from Wisconsin university.

The classification dataset was created based on the "Breast Cancer Wisconsin" problem dataset from UCI repository of machine learning databases from Dr. William H. Wolberg [1]. This problem tries to diagnosis of breast cancer by trying to classify a tumor as either benign or malignant based on cell descriptions gathered by microscopic examination.

IRIS CLASSIFICATION PROBLEM

This is a classical classification dataset made famous by Fisher [2], who used it to illustrate principles of discriminant analysis. This is perhaps the best-known database to be found in the pattern recognition literature.

THYROID CLASSIFICATION PROBLEM

This dataset was created based on the "Thyroid Disease" problem dataset [3] from the UCI repository of machine learning database. The dataset was obtained from the Garvan Institute. This dataset deals with diagnosing a patient thyroid function. Each pattern has 21 attributes and can be assigned to any of three classes which were hyper-, hypo- and normal function of thyroid gland.

DIABETES CLASSIFICATION PROBLEM

This dataset was created based on the _Pima Indians Diabetes' problem dataset [4] from the UCI repository of machine learning database. From the dataset doctors try to diagnose diabetes of Pima Indians based on personal data (age, number of times pregnant) and the results of medical examinations (e.g. blood pressure, body mass index, result of glucose tolerance test, etc.) before decide whether a Pima Indian individual is diabetes positive or not.

CONCLUSION

In this paper, I discussed about the pattern classification, iris classification problem, thyroid classification problem, diabetes classification problem, breast cancer classification based on neural network technique.

The objective of this study is to create an effective tool for building neural models to help us making a proper classification of various classes of cancer.

REFERENCES

[1] Mangasarian O. L., Wolberg W. H.: Cancer Diagnosis via Linear Programming. SIAM News, vol. 23(5), pp.1-18 (1990)
[2] Fisher R.A.: The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, vol. 7: pp.179□188 (1936).
[3] Coomans D, Broeckaert I, Jonckheer M, Massart DL.: Comparison of Multivariate DiscriminationTechniques for Clinical Data—Application to The Thyroid

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

Functional State. Methods of Information Medicine, vol. 22, pp. 93–101. (1983)

[4] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes: Using the ADAP LearningAlgorithm to Forecast the Onset of Diabetes Mellitus. Proceedings of the Symposium on Computer Applications and Medical Care. IEEE Computer Society Press. pp. 261–265 (1988).

[5] Schuermann, Juergen (1996). Pattern Classification: A Unified View of Statistical and Neural Approaches. New York: Wiley. ISBN 0-471-13534-8.