ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

Review on Feature Selection of Gene Expression Data for Autism Classification

Pushpa.M [1], Swarnamageshwari.M [2]

Assistant Professor, Department of MCA & SS, VLB Janakiammal College of arts and science, Coimbatore, Tamilnadu, India[1]

Associate Professor, PG and Research Department of Computer Science, Government Arts College, Coimbatore, Tamilnadu, India[2]

ABSTRACT: The objective of microarray experiments is to identify the genes whether differentially transcribed with respect to different biological conditions of cell cultures or tissue samples. With the development of the microarray technology, the necessary processing and analysis methods grow increasingly critical. The microarray analysis has modernized the approach of biology research in such a way that scientists can now measure the expression levels of thousands of genes simultaneously in a single experiment. Gene expression profiles, which represent the state of a cell at a molecular level, have great potential as a medical diagnosis tool. But compared to the number of genes involved, available training data sets generally have a fairly small sample size for classification. These training data limitations constitute a challenge to certain classification methodologies. Feature selection techniques can be used to extract the marker genes which influence the classification accuracy effectively by eliminating the unwanted noisy and redundant genes This paper presents a review of feature selection techniques that have been employed in microarray data based Autism classification.

KEYWORDS: Microarray, Gene Expression, Feature selection, Autism classification, training data.

I.INTRODUCTION

DNA microarray is a prominent high throughput technology that allows the expression levels of thousands of genes has been monitored simultaneously. Today, the analysis of gene expression data is one of the major topics in health informatics [1]. For instance, the classification of DNA micro array data allows the discovery of hidden patterns in expression profiles and opened possibility for accurate autism classification. The main challenge in classifying gene expression data is the curse of dimensionality problem. There is large number of genes (features) compared to small sample sizes [2, 3]. To overcome this, feature selection is used to identify differentially expressed genes and to remove irrelevant genes. Gene selection remains as a critical task to improve the accuracy and speed of Classification systems [4].

In general, feature selection can be organized into three categories: Filter, Wrapper and Embedded methods. They are categorized based on how a feature selection technique combines with the construction of a classification model. A considerable amount of literature has been published on gene selection methods for building effective classification model. In this paper we present a review of feature selection techniques for autism classification

DNA MICROARRAY

Microarrays offer an efficient method of gathering data that can be used to determine the expression pattern of thousands of genes simultaneously. The mRNA expression pattern from different tissues in normal and diseases states could reveal which genes and environmental conditions can lead to disease. The experimental steps of typical microarray began with extraction of mRNA from a tissues sample or probe. The mRNA is then labeled with fluorescent nucleotides, eventually yielding fluorescent (typically red) cDNA. The sample later is incubated with similarly

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

processed cDNA reference (typically green). The labeled probe and reference are then mixed and applied to the surface of DNA microarrays, allowing fluorescent sequences in the probe-reference mix to attach to the cDNA adherent to the glass slide.

The attraction of labeled cDNA from the probe and reference for a particular spot on microarray depends on the extent to which the sequences in the mix (probe -reference) complement the DNA affixed to the slide. A perfect compliment, in which a nucleotide sequence on a strand of cDNA exactly matches a DNA sequence affixed to the slide, is known as hybridization. Hybridization is the key element in microarray technology.

The populated microarray is then excited by a laser and the consequential fluorescent at each spot in the microarray is measured. If neither the probe nor the reference samples hybridize with the gene spotted on the slide, the spot will appear in the black color. However, if hybridization is predominantly with the probe, the spot will be in red (Cy5). Conversely, if hybridization is primarily between the reference and DNA affixed to the slide, the spot will fluoresce green (Cy3). The spot can also incandescent yellow, when cDNA from probe and reference samples hybridize equally at a given spot, indicating that they share the same number of complementary nucleotides in particular spot. Using image processing software, the red-to-green fluorescence will be digitized and providing the ratio values output indicating the expression of genes. The process of microarray experiment is llustrated in Figure1.

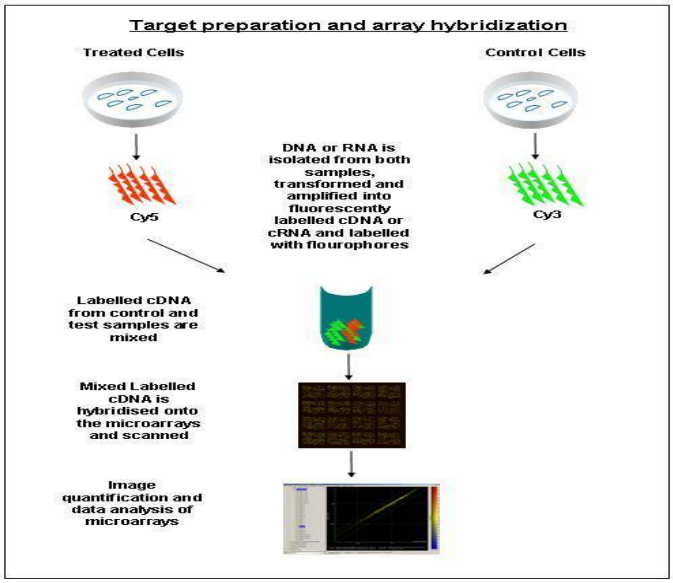

Figure 1: Microarray Experiment

Finally, the gene expression data set can be noted by the following matrix M _{ wij |1 __ i __n,1 __j __ m}, where the rows (G { g1 ,…,gn }) from the expression patterns of genes, the columns (S { s1 ,…,sm }) from the expression profiles of samples, and wij is the measured expression level of gene i in sample j . Thus, M is defined as:

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

$$M = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1m} \\ w_{21} & w_{22} & \cdots & w_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nm} \end{bmatrix} \leftarrow g_i, i = 1, \cdots, n$$

$$\uparrow$$

$$s_j, j = 1, \cdots, m.$$

Due to its high throughput nature, microarray data poses new challenges for data analysis. Although the type of analysis depends on the research questions posed, typical steps in the analysis of microarray data are: i) pre-processing and normalization, ii) detection of genes with significant fold changes, iii) classification and clustering of expression profiles.

II. SOME CHALLENGES FACED IN MICROARRAY DATA ANALYSIS

Many challenges in microarray need to be addressed before new knowledge about gene expression can be revealed. Some of the problems are:

1. Microarray data is high dimensional data characterized by thousand of genes in few sample sizes, which cause significant problems such as irrelevant and noise genes, complexity in constructing classifiers, and multiple missing gene expression values due to improper scanning. Moreover, most of studies that applied microarray data are suffered from data over fitting, which requires additional validation.
2. Mislabeled data or questioned tissues result by experts also another types of drawback that could decrease the accuracy of experimental results and led to imprecise conclusion about gene expression patterns.
3. Biological relevancy result is another integral criterion that should be taken into an account in analyzing microarray data rather than only focusing on accuracy of autismr classification. Although there is no doubt gaining high accuracy classification results are important in microarray data analysis, but revealing the biological information during the process of autism classification is also essential. For instance determination of genes that are under expressed or over expressed in autism cells could assists domain experts in designing and planning more appropriate treatments for autism patients. Therefore, most of domain experts are interested in classifiers that not only produce high classification accuracy but also reveal important biological information.

III. RELATED WORK

Autism is a neurodevelopment disorder characterized by impaired social interaction, verbal and non-verbal communication and restricted and repetitive behavior. Parents usually notice signs in the first two years of their child's life [5].These signs often develop gradually, though some children with autism reach their developmental milestones at a normal pace and then regress [6]. The diagnostic criteria require that symptoms become apparent in early childhood, typically before age three [7]. Gene expression is nothing but level of production of protein molecules defined by a gene. Monitoring of gene expression is one of most fundamental approach in genetics. The technique for measuring gene expression is to measure the mRNA instead of protein, because mRNA sequences hybridize with their complementary RNA or DNA sequence while this property lacks in protein. The DNA arrays are novel technologies that are designed to measure gene expression of tens of thousands of genes in a single experiment.
Gene expression data usually contain a large number of genes in thousands and a small number of experiments (in dozens). In machine learning terminology, these data sets are usually of very high dimensions with undersized samples. The purpose of Gene selection is to find a set of genes that best discriminate biological samples of different types. The

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

selected genes are “biomarkers,” and they form a “marker panel” for analysis. For analyzing the marker panel rank based scheme such information gain was used. It was observed that the information gain with large group was not accurate, therefore in paper [8] they proposed model-based approach to estimate the entropy on the model, instead of on the data themselves. Here, they used multivariate Gaussian generative models, which model the data with multivariate normal distributions.

Feature selection method

There are three types of feature selection methods have been studied: Filter Methods [9], Wrapper Methods [10] and Embedded Methods [11].

1. Filter methods are essentially data preprocessing or data filtering methods. Features are selected based on the intrinsic characteristics that determine their relevance or discriminative powers with regard to the target classes.
2. In wrapper methods, feature selection is “wrapped” around a learning method: the usefulness of a feature is directly judged by the estimated accuracy of the learning method. Wrapper methods typically require extensive computation to search for the best features.
3. Embedded methods differ from other feature selection methods in the way feature selection and learning interact. Filter methods do not incorporate learning. Wrapper methods use a learning machine to measure the quality of subsets of features without incorporating knowledge about the specific structure of the classification or regression function, and can therefore be combined with any learning machine. In contrast to filter and wrapper approaches, in embedded methods the learning part and the feature selection part cannot be separated - the structure of the class of functions under consideration plays a crucial role.

Basic feature selection algorithm

Input:

S - Data sample f with features X, |X| = n
J - Evaluation measure to be maximized
GS – successor generation operator
Repeat

L: = Search Strategy (L, GS (J), X);
X’:= {best of L according to J};
If J (X’) =J (Solution) or (J (X’) =J (Solution) and |X’| < |Solution|) then
Solution: =X’;
Until Stop (J, L).
Output:

Solution – (weighted) feature subset
L: = Start Point(X);
Solution: = {best of L according to J };

The discriminating criteria are being used by filter method for feature selection. The correlation coefficient or statistical test like t-test or f-test is used to filter the features in the filter feature selection method.
Many interesting results were obtained by researchers aiming to distinguish between two or more types of cells (e.g., diseased versus normal or cells with different types of disorder), based on gene expression data in the case of DNA microarrays. Since microarray data have large amount of data and attributes, which makes complex for researcher to do analysis. A small subset of genes is easier to analyze as opposed to the set of genes available in DNA microarray chips. Therefore it is important to focus on very few genes to give insight into the class association for a microarray sample. It

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

also makes it relatively easier to deduce biological relationships among them as well as to study their interactions.

In paper [11] they obtained feature selection algorithms for classification with tight realizable guarantees on their generalization error. The proposed approaches are a step toward which are more general learning strategies that combine feature selection with the classification algorithm.

IV. CONCLUSION

In this paper we reviewed about the feature selection techniques that have been employed in autism classification using gene expression profiles. With a high dimensionality input and small sample data size are the main two problems that have been triggers the application of feature selection in microarray data analysis. Number of efforts have been conducted during the past several years in the utilization of feature selection to encounter these problems, which mainly can be grouped into three main approaches; filter, wrapper and embedded approaches .Each algorithm has different behavior which shows that using single algorithm or different dataset is infeasible. The feature selection algorithms are one which decides the accuracy of the classification of different datasets. The feature selection algorithm must select the relevant features and also remove the irrelevant and inconsistent features which cause the degradation of accuracy of the classification algorithms. Feature selection algorithm is playing a major role in accurate classification of large data set like gene expression. Therefore proper Autism classification can be achieved using feature selection algorithms, and on time and accurate treatment may be provided to the patients.

REFERENCES

[1] R.T.Ng and J.Pei, "Introduction to the special issue on data mining for health Informatics", SIGKDD Explore News1,Vol 9 PP1-2 2007
[2] G.Piatetsky-shapiro and P.Tamayo, "Microarray data mining : facing the challenges," SIGKDD Explorations Newsletter Vol 5,PP 1-5 December 2003
[3] M.Rocha, R.Mendas, P.Maria, D.Glez-Pena, and F.Fdez-Reverola, "A Platform for the selection of genes in DNA Microarray data using evolutionary Algorithms," in Proceedings of 8th Annual Conference on Genetic and Evolutionary computation, London, England, 2007, pp 415-423.
[4] C.Shang and Q.shen, "Aiding classification of gene expression data with feature selection: A comparative study", Vol1, 2006, pp 68-76.
[5] Myers SM, Johnson CP (2007). "Management of children with autism spectrum disorders". Pediatrics **120** (5): 1162–82. doi:10.1542/peds.2007-2362.PMID 17967921.
[6] Stefanatos GA (2008). "Regression in autistic spectrum disorders". Neuropsychology Rev**18** (4): 305–19. doi:10.1007/s11065-008-9073-y. PMID 18956241 .
[7] Autism Spectrum Disorder, 299.00 (F84.0). In: American Psychiatric Association.Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition. American Psychiatric Publishing; 2013.
[8] Shenghuo Zhu, Dingding Wang, Kai Yu, Tao Li, and Yihong Gong "Feature Selection for Gene Expression Using Model-Based Entropy" IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2010 , pp 25-36
[9] P. Langley, "Selection of Relevant Features in Machine Learning," Proc. AAAI Fall Symp.Relevance, pp. 140-144, 1994.
[10] R. Kohavi and G.H. John, "Wrappers for Feature Subset Selection," Artificial Intelligence, vol. 97, no. 1-2, pp. 273-324, 1997.
[11]Thomas Navin Lal1 , Olivier Chapelle1 , Jason Weston2 , and Andr´e Elisseeff3 "Embedded Methods" .
[12] Mohak Shah, Member, IEEE, Mario Marchand, and Jacques Corbeil "Feature Selection with Conjunctions of Decision Stumps and Learning from Microarray Data" IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 34, No. 1, January 2012
[13] Mohua Banerjee, Sushmita Mitra "Evolutionary Rough Feature Selection in Gene Expression Data" IEEE Transactions on Systems, Man, And Cybernetics—Part C: Applications And Reviews, Vol. 37, No. 4, July 2007