Advances in Environmental Biology, 11(5) May 2017, Pages: 13-18

AENSI Journals

Advances in Environmental Biology

ISSN-1995-0756 EISSN-1998-1066

Journal home page: http://www.aensiweb.com/AEB/

A Survey on different classification methods for microarray data analysis

[1]Ms. S. Jayanthi and [2]Dr. C. R. Rene Robin

[1]Department of Computer Science/Agni College of Technology, Chennai.
[2]Department of Computer Science/Jerusalem College of Engineering, Chennai.

Address For Correspondence:
Ms.S.Jayanthi, Agni College of Technology, Chennai.
E-mail: krishnanradha_1976@yahoo.com

Received 8 January 2017; Accepted 28 April 2017; Available online 24 May 2017

ABSTRACT
Accurate classification of diseases is an essential task for treatment. The microarray data allows the researcher to study the genome-wide patterns of gene expression in any given cell type, at any given time and under any given set of conditions. It also provides more valuable information about diseases pathology, progression, and resistance to treatment and response to cellular micro environments. This paper provides brief survey on various microarray data classification methods for various (cancer) Data sets. This paper provides comparative analysis of various feature selection method and classifier.

KEYWORDS: Gene Expression, Microarray data

INTRODUCTION

Microarray data is a gene expression provides unrivalled opportunity to study about specific gene involved in complex biological functions, such as development, signal transduction and diseases. Gene selection plays vital role in identification and publication of diseases. The main objective of this paper to study about 1) Difference between biological and clinical categories, 2) various classification techniques for cancer identification, 3) Analyzing their accuracy. Clinical dataset are set of information recorded by any organization or system for specific area of case, disease or services are at patient level.

Microarray data analysis involves two steps: 1.Feature Selection and 2. Classification.

Microarray data gathered by fabrication, hybridization and image processing. This process adds the uncertainties in the form of noise. The Microarray dataset contains both relevant and irrelevant information which increases the high dimensionality. This degrades the performance of classification. In addition to that, the dataset contains only few trained data (e.g. Breast Cancer Contains 24481 genes and only 97 samples), which leads to critical problem in dimensionality. The Feature Selection is used to avoid dimensionality curse.

The Feature Selection is process of selecting a relevant attributes or predictor for model construction. Feature selection involves various machine learning algorithm. The feature selection method classified into three categories: 1) Filter Method 2) Wrapper Method and 3) Embedded Method.

In Filtering Method, the attribute selection is only depend on the general characteristics of trained data (i.e. distance between classes or statistical dependencies). This method does not depend on any induction algorithm. The best feature selection is analyzed by ANOVA, Chi Square, Pearson's Correlation and LDA.

In Wrapper method, the selection process is considered as search problem, where different combinations of subset are prepared, evaluated and compare with other combination of subset. The predictive model is used to

Advances in Environmental Biology, 11(5) May 2017, Pages: 13-18

evaluate the combination and assign the score based on model accuracy. The famous wrapper methods are forward feature selection, backward feature elimination and recursive feature elimination.

In embedded method, the qualities of filter and wrapper methods are implemented for feature selection.

After feature selection, which removes non-feature term, results in cleansed training datasets can be used for effective classification.

Classification of microarray data into normal and abnormal is done by designing hybrid classifier based on neural networks, Bayesian and support vector machine. In this the extracted feature is compared with references obtained during training stage.

Classification Techniques:

The Traditional classification methods are statistical approaches were inflexible classification systems that are unable to classify a sample, if the expressions of genes are slightly different from the predefined profile. Table 1 shows summarization of classification methods developed for processing microarray data in past years.

Table 1: Summarization of Classification methods for microarray data.

Source	Tools
Friedman et al. (2000)	Bayesian Network
Li.et al (2001)	Genetic algorithm and k-nearest neighbors
Khan et al. (2001)	Artifical neural network
Zhang et al. (2001)	Binary Decision Tree
Nguyen and Rocke (2002)	t statistics, partial least squares and logistic discrimination analysis
Albrecht et al. (2003)	Perceptron and stimulation annealing
Antoniadis et al (2003)	Logisitic discrimination analysis
Jorsten and Yu (2003)	Linear discrimination analysis
Lee et al (2003)	Bayesian Model
Desper et al (2004)	Phylogenetic Model
Simek et al (2004)	Singular value decomposition and SVM
Asyali and Alci (2005)	Fuzzy c-means and normal mixture model
Georgii et al (2005)	Quantitative association rule
Qiu et al. (2005)	Esemble dependence Model
Martella (2006)	Factor mixture models
Wu (2006)	Penalized linear regression model.

These techniques are well known and assemble them together to deal with microarray data. The latter approach can construct a method that is easier to learn and apply for analyzing microarray data.

At [3] Gene expression classification has taken new era for two-stage classification methods. The first stage is selects the pre-specified number of genes. The selected number of genes is less than the instance of original genes. The selected genes are passed to the second stage for classification. The gene selection mechanism involves Individual gene ranking and gene subset ranking. The Classification tools proposed for second stage was K-Nearest Neighbors and Partial Least squares for dimensionality reduction. The accuracy is measured by performing t test and it shows the selection works well for selected gene expression, not for the entire gene. This method proposed for the selected genes at the first stage rather than the original gene expression.

The Wavelet method is used [4] for feature extraction and dimensionality reduction of microarray data. In this method detail wavelet coefficients based on wavelet decomposition at different levels is extracted to characterize the localized features of microarray data. The K fold cross validation experiment is conducted to evaluate the performance of this method. It shows 97.73% of accuracy at 2 fold cross-validation and 96.21% 3 fold cross validation. The result suggests that the detail coefficients at the second and third levels are robust to characterize the feature of microarray data. The main drawback of this method was wavelet and feature transformation is knowledge-free, so it is not possible to obtain an "intelligible" result (the selected genes) as using feature selection techniques.

Instead of single wavelet method, several wavelets are used for feature selection [5]. The first step produces two sets of coefficient: approximation coefficients (Scaling coefficients) and detail coefficients (wavelet Cofficients). Then they are split into two parts by using the same algorithm and so on. This process is repeated until the required level is reached. The multi-classifiers are used where each classifier, a SVM, is trained using a different set of detailed coefficients, the classifiers are combined by "sum rule". The goodness of the method is evaluated under the ROC as performance indicator.

The novel approach for feature selection and classification was proposed by combination of discrete wavelet transformation and moving widow technique [6]. In this method, first window size is defined. For each window the wavelet transform is applied and define the level of decomposition. Data are rearranges with threshold to the wavelet coefficient of DWT. The t-test is applied to select the top ranking features. Then hybrid classifiers (KNN, Bayes and SVM) are applied. The Multi-resolution representation of microarray data is achieved by DWT inside a window predefined size as test pattern. The robustness of the system is tested by

three parameters types of wavelet, decomposition level of DWT and window size of MWT. The 100% accuracy of system is depends on the level and window size.

To improve the interpretability of feature selection of gene, binary particle swarm optimization technique proposed [7]. It maps the gene – to – class sensitivity information was encoded. The gene - to – class sensitivity information, extracted from the samples by extreme learning machine, is encoded into the selection process in four aspects: initializing, updating the particles, modifying maximum velocity and adopting mutation operation adaptively. The classifier KNN and SVM is used for classification. This method avoids filtering out the some critical genes; it may increase the computational cost for large initial gene pool.

Datasets:

The cancer gene expression data sets used for the study are described. These datasets are also summarized below.

Colon Datasets: It is derived from colon cancer patient samples. It has 1909 genes of 62 patients among which 40 are colon cancer cases and 22 are normal cases.

CNS Datasets: It contains 60 patient samples out of which 39 are normal cases and 21 are cancer cases.

Leukemai Datasets: It contains the expression level of 7129 genes taken from 72 samples. There are 47 cancer cases and 25 normal cases.

SRBCT Datasets: The dataset consists of four categories of small round blue cell tumors (SRBCT) with 83 samples from 2308 genes.

Conclusion:

This paper specifies various feature selection and classification method that is used to classify the microarray datasets. This paper also shows how hybrid techniques (combining various algorithms) are used to improve the efficiency of classification. This paper expose that the wavelet transforms is an important multi-resolution analysis tool that has applied to various classification systems. In future better wavelet transforms will be used for addressing curse of dimensionality reduction.

REFERENCES

[1] Albrect, A., S.A. Vintwebo and L. Ohno-Machado, 2003. An Epicurean learning approach to gene expression data classification. Artificial Intelligence in Medicine, 28: 75-87.
[2] Antonids, A., S. Lambert-Lacroiix and F. Leblance, 2003. Effective dimension reduction methods for tumor classification using gene expression data, Bioinformatics, 19: 563-570.
[3] Tzu-Tsung wong, Ching-Han Hsu, 2008. Two stage classification methods for microarray data Elsevier Expert systems with applications, 34: 375-383.
[4] Yihui Liu, 2009. Wavelet feature extraction for high – dimensional microarray data Elsevier Neurocomputing, 72: 985-990.
[5] Loris nanni, Alessandra Lumini, 2011. Wavelet selection for disease classification by DNA microarray data. Elsevier Expert systems with applications, 38: 990-995.
[6] Jaison Bennet, Chilambuchelvan Arul Ganaprakasam and Kannan Arputharaj, 2014. A discrete wavelet based feature extraction and hybrid classification technique for microarray data analysis. The Scientific world journal, 9 pages.
[7] Fei Han, Chun yang, Ya-Qi Wu, Jian-Sheng Zhu, Qing-Hua Ling, Yu-Qing Song, De-Shuang, 2015. A gene selection method for microarray data based on binary PSO encoding gene-to-class sensitivity information. IEEE Transaction on computational biology and bioinformatics.

Advances in Environmental Biology, 11(5) May 2017, Pages: 13-18

Advances in Environmental Biology, 11(5) May 2017, Pages: 13-18

Advances in Environmental Biology, 11(5) May 2017, Pages: 13-18