IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

nCORETech 2017

LBS College of Engineering, Kasaragod

Vol. 6, Special Issue 3, March 2017

Efficient Feature Selection and Classification for Gene Expression Data

S. Venkatesh

Assistant Professor, Department of Computer Science, Sri Kaliswari College, Sivakasi, India

Abstract: Microarray data has been widely applied to cancer classification, where the purpose is to classify and predict the category of a sample by its gene expression profile. The cancer classification is required to identify the significant genes that are a good subset of original features. For evaluating the goodness of a subset of features, the feature selection methods fall into two broad categories: the filter approach and the wrapper approach. Since wrapper methods are computationally very intensive, filter approach is chosen for selecting the most informative genes. DNA microarray is a gene chip which consists of expression levels for a huge number of genes a relatively small number of samples. However, only a small number of genes contribute in accurate classification of cancer. Therefore, the challenging task is to identify a small subset of informative genes which has maximum amount of information about the class. The feature selection method is used to find the informative genes which helps to minimize the classification errors. The hybrid correlation methods are used to find out the correlated and negative correlated features. The classifier Support Vector Machine along with Decision Tree Algorithm is proposed to classify the features. The result is compared with the performance of neural network classifier which gives better accuracy in positive correlated features than hybrid correlated and negative correlated features.

Keywords: DNA Microarray, Classification, Correlation, Neural Network, Backpropagation Algorithm.

I. INTRODUCTION

1.1 Data Mining

Data mining refers extracting or "mining" knowledge from large amounts of data. The data mining should have been more appropriately named "Knowledge mining from data, which is shortly termed as "Knowledge mining". The popularly used term is Knowledge Discovery in Database, or KDD. There are many other terms carrying a similar or slightly different meaning to data mining, such as knowledge mining from databases, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. The interesting patterns are presented to the user, and may be stored as new knowledge in the knowledge base.

1.2 Gene Expression Data

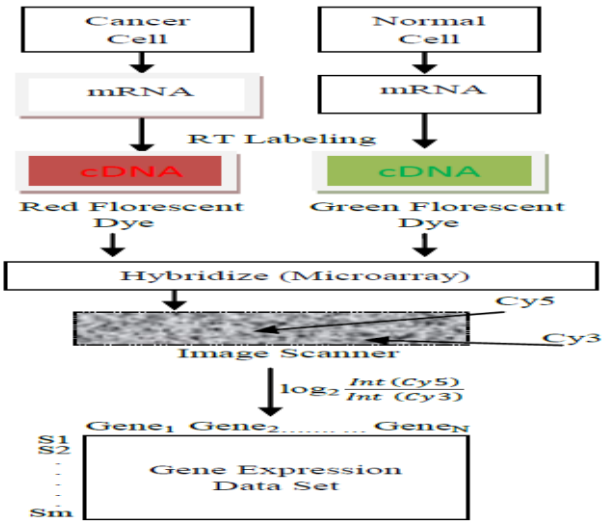

Fig. 1 Process of forming DNA Microarray and acquiring gene expression data

Gene expression is the process by which inheritable information from a gene, such as the DNA sequence, is made into a functional gene product, such as protein or RNA. DNA microarrays are created by robotic machines that arrange

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
nCORETech 2017
LBS College of Engineering, Kasaragod
Vol. 6, Special Issue 3, March 2017

thousands of gene sequences on a single microscope slide. Actually in our body the entire cell contains identical genetic material, but the same genes are not active in every cell. To determine which genes are turned on and which are turned off in a given cell, it is required to collect the mRNA molecules present in that cell, and then label each mRNA molecule by using a reverse transcriptase enzyme (RT) which generates a complementary cDNA to the mRNA. During that process fluorescent nucleotides are attached to the cDNA. A red spot indicates that the specific gene is more expressed in tumor, a green spot indicates that the specific gene is more expressed in the normal tissue and a yellow spot means that the specific gene is equally expressed in normal and tumor.

1.3 Selecction Methods

Selection is an important topic in data mining, especially for high dimensional datasets. Feature selection (also known as subset selection) is a process commonly used in machine learning, wherein subsets of the features available from the data are selected for application of learning algorithms are used as three methods as follows.

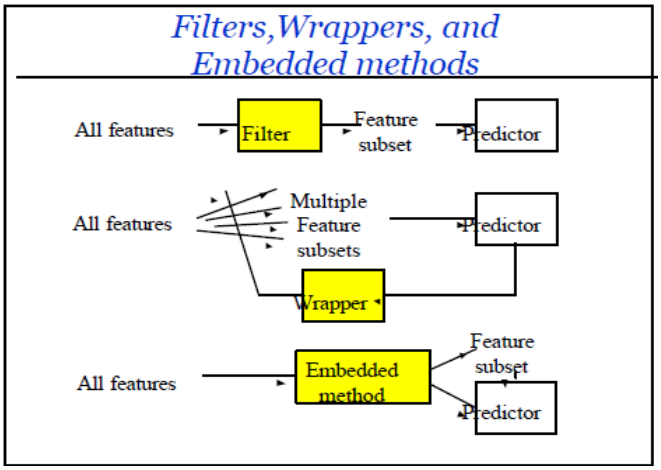

Fig. 2 Filters, Wrappers and Embedded Methods.

II. METHODOLOGY

2.1 Gene Expression Dataset Preprocessing

Gene Expression Dataset preprocessing of microarray gene expression data. Before preprocessing, do the Pre-Analyses module. It automatically selects or unselects different options and gives useful information to the user. After checked pre analyses performed by the Pre-Analysis module, the real data processing starts. The list of available preprocessing functions includes.

2.1.1 File Format

It checks that all the patterns have the same number of conditions. It adds extra missing values or removes extra ones if needed. It also checks that all the patterns have a valid identifier. It adds a unique one when necessary. It also keeps a list of symbols interpreted as missing values and displays this information.

2.1.2 Dimension of the Dataset

The server expects a dataset with more rows (genes) than columns (Experimental conditions). It displays a warning message if it founds more conditions than patterns, as could happen if the data matrix is transposed.

2.1.3 Scale of Expression Patterns

The server plots the histogram of values found in the data set and looks for negative values. It assumes the user did not log-transform the dataset if none is found. It displays a warning message, log-transforms the data internally to continue with the analysis, auto-selects the option for log-transforming the dataset and plots the new histogram of log-transformed values.

2.1.4 Replicated Genes

The server looks for replicated genes and merges them into their average pattern. The list of replicated genes as well as the number of them is displayed. The server also plots an additional histogram of distances to median of replicates to guide user in selecting a good threshold for removing inconsistent replicates.

2.1.5 Missing Values

The server obtains the number of missing values for this dataset and suggests the best imputation method among the

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
nCORETech 2017
LBS College of Engineering, Kasaragod
Vol. 6, Special Issue 3, March 2017

available ones depending on the amount of missing values, the amount of complete patterns and the number of conditions. It also plots a histogram of the number of missing values by pattern.

2.2 Applying for Pre-Filter Methods

Filter approaches are easily scalable to very high-dimensional datasets and accurate, they are computationally simple, and classifier independent .It is divided in 4 sections, start with pre-filtering, then define ideal feature vectors for selecting correlated features, then find the hybrid negative correlated features and lastly check the recognition accuracy of features on a classifier.

2.2.1 Normalization and Pre-Filtering

Normalization is a data analysis technique to design a database system. It allows the database designer to understand the current data structures within an organization. Furthermore, it aids any future changes and enhancements to the system. The end result of normalization is a set of entities, which removes unnecessary redundancy (ie duplication of data) and avoids the anomalies discussed earlier. Normalization follows a staged process that obeys a set of rules.
For this unit you only need to be aware of the above three forms. However, the process to transform them is not assessed and, therefore, not described. Finally, from the normalized tables, a data model is produced.
Earlier is an example of a data model containing several entities. The normalization process helps us determine the entities required in the system being modeled. These entities can then be represented in a data model. Later, following the normalization process to determine the entities, the data model itself is constructed.
First, the process of normalization (to 3NF) is demonstrated by use of an example. Normalization is a bottom-up technique for database design, normally based on an existing system (which may be paper-based). By analyzing the documentation, eg reports, screen layouts, etc from that system, extract gene expression data set and normalize it in the range of [0, 1] using min-max normalization.
ei= (ei-min(ei)) /Max(ei)-Min(ei) where i=1 to N

2.2.2 Ideal Feature Vectors

In M x N training microarray dataset, where M is the number of samples and N is the number of genes, any gene th vector can be expressed in column matrix as
gi=(e1,i e2,i e3,i . . . eJ ,i . . . . eM,i) ,where i=1 to N
Ideal feature vector are depend upon the class of the sample. If in M samples J samples belongs to class “1” and others on class “o” then ideal feature vector can be defined as.

IFVc1 =1 1 1 1 1 1 1 1 1... 1jth 0 0 0 0 0 0 0...M
IFVc2=0 0 0 0 0 0 0 0 0.. 0jth 1 1 1 1 1 1 1...M

Where IFVc1 and IFVc2 are ideal feature vectors for class c1 and c2.

2.2.3 Hybird Negative Correlated Feature Selection

For finding hybrid negative correlated features, all the features (genes) which are high correlated with IFVc1 from three feature selection techniques then same process is repeated for IFVc2.After finding the correlation arrange all the informative features in decreasing order with respect to their correlated values. For PC and CC arrange all features from high values to low values and for ED low values to high values. By combining all high ranked features from all independent feature selection techniques with respect to both ideal feature vectors then the obtained result is negative correlated informative feature (NC) sets, many work has already being done on this negative correlated features.
By taking 25 high ranked features for each set (IF1 to IF6) after arranging the features in decreasing order of their correlated values. For finding hybrid negative informative features set (HNIF1), combine the Pearson coefficient and Euclidean distance based features by uniting subsets IF1 and IF3 such that, first take common features then equally remaining high ranked features from both. Same process is repeated with IF2 and IF4 and combined with IF1 and IF3 outcome. The maximum gap of 10 features for selecting common informative genes is considered.

2.3 Classification of Genes – SVM & Decision Tree

Decision tree support vector machine, which combines SVM and decision tree using the concept of dichotomy, is proposed. DNA microarray experiments generating thousands of gene expression measurements are being used to gather information from tissue and cell samples regarding gene expression differences that will be useful in diagnosing disease. A new method to analyze this kind of data using support vector machines (SVMs) is developed. This analysis consists of both classification of the tissue samples, and an exploration of the data for mislabeled or questionable tissue results. As a result of computational analysis, a tissue sample is discovered and confirmed to be wrongly labeled. Upon correction of this mistake and the removal of an outlier, perfect classification of tissues is achieved, but not with high confidence. Then identify and analyze a subset of genes from the lymphoma dataset whose expression is highly differentiated between the types of tissues. To show robustness of the SVM method, two previously published datasets

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
nCORETech 2017
LBS College of Engineering, Kasaragod
Vol. 6, Special Issue 3, March 2017

from other types of tissues or cells are analyzed. The results are comparable to those previously obtained. Other machine learning methods also perform comparably to the SVM on many of those datasets.

2.4 Prediction of Cancer Genes

The standard strategy is to identify a molecular signature (i.e., the subset of genes most differentially expressed in patients with different outcomes) in a training set of patients and to estimate the proportion of misclassifications with this signature on an independent validation set of patients.
By expanding this strategy (based on unique training and validation sets) by using multiple random sets, the stability of the molecular signature and the proportion of misclassifications is studied.

2.5 Performance of Evolution

Constantly improving gene expression data's are expected to provide understanding and insight into cancer-related cellular processes. Classify cancer class of patients by small subset of informative genes which kept the maximum amount of information about class and minimize the classification errors, illustrate a classification framework that combines a pair of hybrid negative correlated features from the combination of three feature selection methods. Finally get the result, the obtained informative gene subsets have good classification accuracy and also take less computational time as concern of classifier. But one of the limitations of this approach is, it does not take into account the correlation between genes, which reduces the usefulness of the selected genes for cancer classification. Then identify and analyze a subset of genes from the lymphoma dataset whose expression is highly differentiated between the types of tissues.

III.EXPLANATION

The filter method is used for feature relation. In gene expression data preprocessing the dataset is uploaded with large dimension and with required format. A warning message is displayed if there are too many of them, as could happen if the file uploaded is in a wrong format.

In normalization, the uploaded dataset may have some null values. These null values are replaced by zero. In pre filtering, the minimum value and the maximum value is taken from each sample and then it is calculated. The values in the pre filtering will be only in 0's and 1's.The steps of normalization are:

- Select the data source and convert into an un normalized table (UNF)
- Transform the un normalized data into first normal form (1NF)
- Transform data in first normal form (1NF) into second normal form (2NF)
- Transform data in second normal form (2NF) into third normal form (3NF)

Occasionally, the data may still be subject to anomalies in third normal form. In this case, perform further transformations.

- Transform third normal form to Boyce-Coded normal form (BCNF)
- Transform Boyce-coded normal form to fourth normal form (4NF)
- Transform fourth normal form to fifth normal form (5NF)

Then apply variance based filter technique to filter out the features which have nearly same values for all samples. In ideal feature vectors, the profited data is taken. The dataset is divided into two classes. The ideal feature vector class1, is arranged in ascending order and then the ideal feature vector class2, is arranged in descending order. In negative correlated feature, these are three feature selection methods.

- Pearson co-efficient (PC)
- Cosine co-efficient (CC)
- Euclidean distance (ED)

The PC, ED and CC are calculated with the ideal feature vector classes. Each technique has two classes like PC class1, PC class2, ED class1, ED class2, CC class1 and CC class2. In hybrid negative correlated feature, the combination of feature selection techniques takes place. The PC, ED and PC, CC are combined to get HNIF1, HNIF2.
After selecting the required features, the classification is performed. The classifier used is Support Vector Machine (SVM) classifier with decision tree algorithm. The classification performance of Decision tree SVM highly depends on its structure, to cluster the multi-classes with maximum distance between the clustering centers of the two sub-classes, genetic algorithm is introduced into the formation of decision tree, so that the most separable classes would be separated at each node of decisions tree. In SVM classifier, the classification is done to select the required dataset. The Euclidean distance, Distance metric, Entropy metric and the Entropy is calculated.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

nCORETech 2017

LBS College of Engineering, Kasaragod

Vol. 6, Special Issue 3, March 2017

In Check Accuracy, the test dataset is checked whether it has cancer or not and the cancer genes are shown. The list of genes identified as predictors of prognosis was highly unstable; molecular signatures strongly depended on the selection of patients in the training sets. Gene expression data is also expected to significantly aid in the development of efficient cancer diagnosis and classification platforms. In Performance evolution, the positive correlated feature, negative correlated feature and hybrid negative correlated feature are taken to represent the bar chart. The maximum accuracy is obtained in positive correlated feature.

IV. RESULTS & DISCUSSION

In Check Accuracy, the test dataset is checked whether it has cancer or not and the cancer genes are shown. The list of genes identified as predictors of prognosis was highly unstable; molecular signatures strongly depended on the selection of patients in the training sets. Gene expression data is also expected to significantly aid in the development of efficient cancer diagnosis and classification platforms. In Performance evolution, the positive correlated feature, negative correlated feature and hybrid negative correlated feature are taken to represent the bar chart. The maximum accuracy is obtained in positive correlated feature.

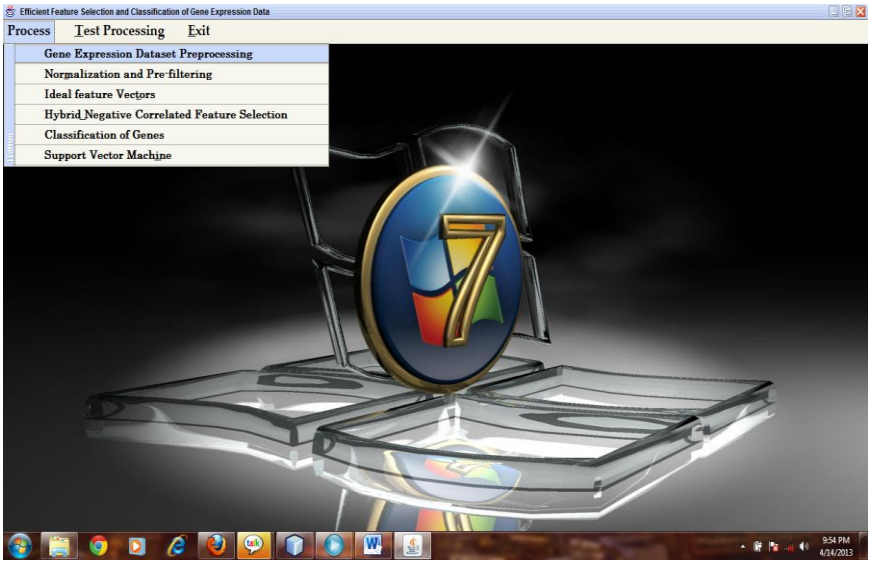

Fig 3. Main Screen for Gene Expression Data

In this figure is designed using net beans tool. Here using these tools to provide the GUI for our processing and placed on the buttons, labels, panel, frame, text area etc. Use this GUI to load the gene expression dataset into our process. This significance of this work is to classify and predict the category of the simple by its gene expression profile. There are some techniques used to predict the category. It shows the screen shot for main screen for gene expression data.

Fig. 4 Feature selection technique cc calculation

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

nCORETech 2017

LBS College of Engineering, Kasaragod

Vol. 6, Special Issue 3, March 2017

This module consists of whole techniques which are to be calculated one by one. The selected option redirects to the corresponding calculation or module. Here under the process menu there are five methods are available. After analyzing the results of each gene and show the information gene accession number, affecting gene in filter method and wrapper method and embedded method. To show robustness of the SVM method, two previously published datasets from other types of tissues or cells are analysts. The log ratio between the two intensities of each dye is used as the gene expression data.

The feature selection techniques is calculated there are three types of feature selection techniques are available. It is Euclidean distance measurement is displayed. Euclidean examined the root of square difference between co-ordinates of a pair of objects. It also referred as normal distance between two variables. This methods PC and ED, CC and PC are combined to give the hybrid negative correlated features. HNIF1 combines ED and PC calculation which arrange all features from low values to high values. HNIF2 combines CC and PC calculation which arrange all features from high values to low values. The SVM classifier is used to reduce the highly dimensional subset into small ranged subsets with the required conditions.

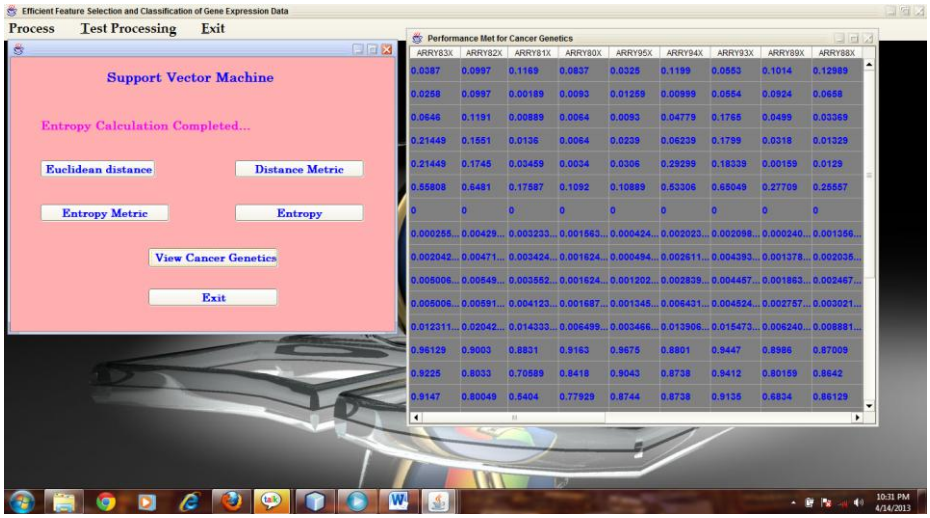

Fig 5. Performance met for cancer genetics

In this figure is designed using net beans tool. Using these tools to provide the GUI for our processing and placed the buttons, labels, panel, frame, text area etc. It is one of the feature selection techniques which represent the relationship between two variables that are measured on the same interval or ration scale. It ranges from -1 to +1. A value of +1 is the result of perfective positive relationship. When all points of a scattered plot fall directly on a line with an upward inline then it means perfect positive correlation.

Table 1. Analysis with Positive Correlated, Negative Correlated (NC) and Hybrid Negative Correlated Features

Feature Sets	Number of Test Samples	Matched Samples	Recognition Accuracy
IF1	31	21	67.8%
IF2	31	24	77.4%
IF3	31	23	74.2%
IF4	31	24	77.4%
IF5	31	26	83.9%
IF6	31	24	77.4%
NC1	31	26	83.9%
NC2	31	27	87.1%
NC3	31	28	90.3%
HNIF1	31	28	**90.3%**
HNIF2	31	29	**93.5%**

Table 2. Analysis of Leukemia Dataset with Positive Correlated, Negative Correlated (NC) and Hybrid Negative Correlated Features

Feature Sets	Number of Test Samples	Matched Samples	Recognition Accuracy

IF1	34	27	79.4%
IF2	34	32	94.1%
IF3	34	28	85.5%
IF4	34	33	97.1%
IF5	34	27	79.4%
IF6	34	32	94.1%
NC1	34	32	94.1%
NC2	34	33	97.1%
NC3	34	32	94.1%
HNIF1	34	33	**97.1%**
HNIF2	34	33	**97.1%**

V. CONCLUSION

In this work to classify the gene small subset of informative genes, which having maximum amount of information are find out which leads to minimize the classification errors. The most predominantly used feature selection technique, filter method is used to preprocess the data which normalize and transform the data into required format. The correlated features are determined by using three techniques such as ED, PC and CC method. As a result the Positive correlated, Negative correlated and Hybrid negative correlated features are found out. For classifying the result the classifier support vector machine (SVM) classifier with decision tree algorithm is used. The SVM can easily deal with large number of features and also with large number of features. It is efficient to analyze broad pattern of gene expression from microarray data. The performance evaluation of this classifier is compared with the Feed Forward Neural network classifier. While comparing, the result shows that it eliminates the over fitting problem and also it takes into account about the correlation between genes which increases the usefulness of the selected genes for cancer classification.

REFERENCES

1. HuilinXiong, Ya Zhang and Xue-Wen Chen, "Data- Dependent Kernel Machines for Microarray Data Classification", IEEE/ACM Transactions on Computational Biology and Bio-Informatics.
2. Jiexun Li, Hua Su, Hsinchun Chen and Bernad W. Futscher, "Optimal Search-Based Gene Subset Selection for Gene Array Cancer Classification", IEEE Transactions on Information Technology in Bio-Medicine,Vol.11, No. 4, July 2007.
3. Jin-Mao Wei, Shu-Qin Wang and Xiao-Jie Yuan, "Ensemble Rough Hypercuboid Approach for Classifying Cancers", IEEE Transactions on Knowledge and Data Engineering, Vol.22, No.3
4. Kai-Bo Duan, Jagath C. Rajapakse, Haiying Wang, Francisco Azuaje, "Multiple SVM-RFE for Gene Selection in Cancer Classification with Expression Data", IEEE Transactions on NanoBioscience, Vol.4, No.3, September 2005.
5. R. Kohavi and G. John, "Wrappers for Feature Subset Selection", Artif. Intell., 1-2(1997)
6. Lipo Wang, Feng Chu and Wei Xie, " Accurate Cancer Classification Using Expressions of Very Few Genes", IEEE/ACM Transactions on Computational Biology and Bio-Informatics, Vol.4
7. Li Shen and Eng Chong Tan, "Dimension Reduction-Based Penalized Logistic Regression for Cancer Classification using Microarray Data", IEEE/ACM Transactions on Computational Biology and Bio-Informatics, Vol.2, No.2, June 05
8. LubomirHadjiiski, BerkmanSahiner, Heang-Ping Chan, Nicholas Petrick and Mark Helvie, "Classification of Malignant and Benign Masses Based on Hybrid ART2LDA Approach", IEEE Transactions on Medical Imagine, Vol.18, No.12,
9. PradiptaMaji and Chandra Das, "Relevant and Significant Supervised Gene Clusters for Microarray Cancer Classification", IEEE Transactions on Nano-Bio-Science, Vol.11, No.2,
10. Zexuan Zhu, Yew-Soon Ong and Jacek M. Zurada, "Identification of Full and Partial Class Relevant Genes", IEEE/ACM Transaction on Computational Biology and Bio-Informatics, Vol.7, No.2, June 2010,s
11. RuiXu, Georgios C. Anagnostopoulos and Donald C.Wunsch II, "Multiclass Cancer Classification using Semisupervised Ellipsoid ARTMAP and Particle Swarm Optimization with Gene Expression Data", IEEE/ACM Transactions on Computational Biology and Bio-Informatics, Vol.4, No.1, March 2007.
12. Runxuan Zhang, Guang-Bin Huang, NarasimhanSundararajan and P.Saratchandran, "Multicategory Classification using an Extreme Learning Machine for Microarray Gene Expression Cancer Diagnosis", IEEE/ACM Transactions on Computational Biology and Bio-Informatics, Vol.4, No.3, September 2007.
13. Sujata Dash, BichitranandaPatra, B.K. Tripathy (2012), "A Hybrid Data Mining Technique for Improving the Classification Accuracy of Microarray Data Set", I.J. Information Engineering and Electronic Business, 2012, 2, 43-50.
14. SantanuGhorai, Anirban Mukherjee, SanghamitraSengupta and Pranab K. Dutta, "Cancer Classification from Gene Expression Data from NPPC Ensemble", IEEE/ACM Transactions on Computational Biology and Bio-Informatics, Vol.8, No.3, June 2011.
15. SarasSaraswathi, Suresh Sundaram, NarasimhanSundararajan, Michael Zimmermann and MaritNilsen-Hamilton, "ICGA-PSO-ELM Approach for Accurate Multiclass Cancer Classification", IEEE/ACM Transactions on Computational Biology and Bio-Informatics, Vol.8, No.2, April 2011.
16. Yuchun Tang, Yan-Qing Zhang, Zhen Huang, Xiaohua Hu and Yichuan Zhao, "Recursive Fuzzy Granulation for Gene Subsets Extraction and Cancer Classification", IEEE Transactions on Information Technology in BioMedicine, Vol.12, No.6, November 2008.
17. M.Akila, S.SenthamaraiKannan, "Hybrid Local Feature Selection in DNA Analysis Based Cancer Classification", IJCSE, ISSN: 0976-5166, Vol.3, No.3,July2012.