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

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

A Comparative Study on Data Mining Algorithm for Gene Cancer Analysis

M. Sofia[1], S. Diviya[2]

Assistant Professor, Dept. Computer Science, Pioneer College of Arts & Science, Coimbatore, Tamilnadu, India [1,2]

Abstract: DNA microarray data now permit scientists to screen thousand of genes simultaneously and determine whether those genes are active or silent in normal and cancerous tissues. With the advancement of microarray technology, new analytical methods must be developed to find out whether microarray data have discriminative signatures of gene expression over normal or cancerous tissues. Fuzzy C-Means is a method of clustering which allows one piece of data to belong to two or many clusters. This method is frequently used in pattern recognition. It is based on minimizing functions. Fuzzy Partitioning is carried out through an interactive optimization of the objective function, with the update of membership the cluster centers. Fuzzy c-means is one of them and it is used widely in such applications as a clustering algorithm. In this study, we applied a different clustering algorithm, an artificial immune system (AIS), for data reduction process. We realized the performance evaluation experiments on standard Chain link and Iris datasets, while the main application was conducted by using Wisconsin Breast Cancer dataset and Pima Indians dataset which were taken from the UCI Machine learning repository.

Keywords: K-mean, Fuzzy C-Means, Microarray, Gene selection, Classification

I. INTRODUCTION

Microarray technology permits coincident activity of the expression levels of thousands of genes inside a biological tissue sample. Gene expression is to classify samples according to their gene expression profiles. Gene selection ways are classified into three types: Filter technique, Wrapper technique and Embedded ways. Filter technique valuate a set of genes by viewing the intrinsic characteristics of knowledge. Wrapper technique valuate the goodness of a sequence set by the accuracy of its learning or classification. Gene choice is embedded within the construction of the classifier. Microarray expression experiments permits the recording of expression levels of thousands of sequence at the same time. These experiments primarily consists of either observing every sequence multiple times below several conditions or alternately evaluating every sequence in an exceedingly single atmosphere however in numerous genes attributable to common expression patterns whereas the later experiments have shown promise in classifying tissues sorts and within the identification of genes whose expression are good diagnostic indicators. Clustering analysis groups genes that have interconnected patterns. It provides gene to gene interactions and gene function. The k-nearest neighbors and genetic technique is employed for choosing a set of predictive genes from a large data. Different theoretical measures like t-test, entropy and mutual information's are wide used.

A typical microarray dataset is extremely sparse in the sense that the dataset usually comes with only several dozens of tissue samples but with thousands or even tens of thousands of genes. This extreme sparseness and small sample size remain a bottle- neck in obtaining robust and accurate classifiers. As a result, the ability to extract gene markers while removing irrelevant or redundant genes is crucial for cancer classification. This is also helpful for biologists to find cancer related genes, and hence to develop better diagnostic methods or find better therapeutic treatments. Feature selection and cancer classification are two closely related problems. Most existing approaches handle them separately by selecting genes prior to classification.

Feature selection is an important pattern recognition problem. Successful feature selection has several advantages for microarray data. First, dimension reduction to reduce the computational cost. Second, reduction of noises to improve the classification accuracy. Finally, more interpretable features or characteristics that can be helpful to identify and monitor the target diseases. Biologically, only a few genetic alterations correspond to the malignant transformation of a cell. With today's improving technology are very high level, Data recording opportunities are also expanding and providing lots of ways for information flow. For large datasets, data mining techniques are affected in three ways: computing time, predictive or descriptive accuracy and representation of the data mining model. Thus some preliminary data pre-processing steps should be conducted before mining in data. Several approaches can be taken into consideration for data reduction for example random sampling of current dataset. Clustering is another alternative to reduce the number of samples by taking only cluster representative sample for all samples in a cluster.

II. GENE SELECTION METHODS

Many methods are used for gene selection and tissue sample classification using microarray.

A. K-NEAREST NEIGHBORS

K- nearest neighbor is a non parametric classification method ,that predicts the sample of a test case[7].To apply K- nearest neighbor each sample was represented by a pattern of expression that consists of D genes. Each sample was then classified according to the class memberships of its k nearest neighbors, as determined by the Euclidean distance in the d-dimensional space. Dudoit S.Fridly says that the number of neighbors used is chosen by cross validation[14].By using the prediction top features are extracted and the method is used to classify unknown samples. When unclassified is accepted as a possible output, one needs to consider the various outcomes in analyzing the value of a classification[8].

B. GENETIC ALGORITHM

A genetic algorithm (GA) is a global optimization procedure that uses the genetic evolution of biological organisms. It generates a new population from the current population using cross over and mutation methods [13]. Genetic algorithm is an intelligent technique used to find a useful subset. Since genetic algorithm has been shown to be effective in searching complex high-dimensional space. As Holland and Goldberg adapted Genetic algorithm as search tool[7].Each 'chromosome' consists of d distinct genes that are initially randomly selected from all genes. A set of chromosomes is constructed to from a 'population' or a 'niche'. The genes to be selected is correspond to the features attributes.[2],[3].

C. SUPPORT VECTOR MACHINES:

The ability of support vector machine is to deal with high dimensional data. The four different kernels are used for testing the genes. SVM try to find an optimal gene separating hyper plane between the classes. When the classes are linearly separable, the hyper plane is located so that it has maximal margin which should lead to better performance on data not yet seen by the SVM. When the data are not separable, there is no separating hyper plane; in this case it tries to maximize the margin but allow some classification errors subject to the constraint that the total error is less than a constant. There are several possible approaches; In this method "one against- one" approach, as implemented in "libsvgm"[12]Chan CC. 200 genes as predictors tended to perform as well as, or better than, smaller numbers. Guyon used the support vector machine as a tool for discovering informative patterns[4].

D. FUZZY C-MEANS ALGORITHM

Fuzzy C-mean algorithm is also called as ISODATA. It was most frequently used in pattern recognition. Fuzzy C-mean is the method using in clustering. It is using one piece of data to belong to two or more clusters. It always based on minimization of objective functions to achieve a good classification. Fuzzy partitioning is carried out through an iterative optimization of the objective function display above, with the updates of membership .

E. K-MEANS ALGORITHM:

K-Means is a well known partitioning algorithm used for grouping. Objects are classified as belongings to one of the k groups, the k chosen a priori.

The most common algorithm uses an iterative technique. Due to its ubiquity it is often called the **k-means algorithm**; it is also referred to as **Lloyd's algorithm**, mainly in the data mining community. These initial set of k means $m_1^{(1)},\ldots,m_k^{(1)}$, the algorithm proceeds by alternating between two steps:

Assignment step: Assign each observation to the cluster whose mean yields the least within-cluster sum of squares (WCSS). Since the sum of squares is the squared Euclidean distance, this is intuitively the "nearest" mean. (Mathematically, this means partitioning the observations according to the Voronoi diagram generated by the means).

Update step: This method calculate the new means to be the centroids of the observations in the new clusters. These was an arithmetic mean is a least-squares estimator, this also minimizes the within-cluster sum of squares (WCSS) objective.

III. PERFORMANCE METRICS

A. Feature ranking with correlation coefficients

For gene selection testing is not possible to achieve an errorless separation with a single gene. These methods include correlation methods and quantitative relation methods [6]. Moreover, complementary genes that severally don't separate well the information are incomprehensible. The coefficient used is defined as:

$$wi=(\mu i(+)-\mu i(-))/(si(+)+si(-))(2) \quad (1)$$

Where μi and si are the mean and standard deviation of the gene expression values of gene i for all the patients of class (+) or class (-), i = 1, . . . n.

$$(\mu i(+)-\mu i(-))2/(si(+)2+\mu i(-)2) \quad (2)$$

B. Ranking criterion and classification

One possible use of feature ranking is the design of a class predictor based on a pre-selected subset of features. Each feature that is correlated with the separation of interest is by itself such a class predictor, an imperfect one. This suggests a simple method of classification based on weighted voting: the features vote proportionally to their correlation coefficient, the method being used [12]. The weighted voting scheme yields a particular linear discriminate classifier:

$$D(x)=(x-\mu)(3) \quad (3)$$

where w is defined in

$$\mu = (\mu\ (+) + \mu\ (-))/2. \quad (4)$$

It is interesting to relate this classifier to Fisher's linear discriminate. Such a classifier is also of the form of Eq. (3), with

$$w = S\text{-}1\ (\mu\ (+) - \mu\ (-)) \quad (5)$$

and where µ is the mean vector over all training patterns.Coefficients are denoted by X (+) and X (-) the training sets of class (+) and (-). This particular form of Fisher‗s linear discriminate implies that S is invertible. It retains some validity if the features are uncorrelated, that is if the expected value of the product of two different features is zero, after removing the class mean. Approximating S by its diagonal elements is one way of regularizing it.

C. Feature ranking by sensitivity analysis

For classification problems, the ideal objective function is the expected value of the error. The OBD algorithm approximates DJ (i)by expanding J in Taylor series to second order [7]. At the optimum of J, the first order term can be neglected, yielding:

DJ (i)= (1/2) .2 J/.w2i (Dwi)2 (6)

The change in weight Dwi =wi corresponds to removing feature i. The authors of the OBD algorithm advocate using DJ(i) instead of the magnitude of the weights as a weight pruning criterion. For linear discriminate functions whose cost function J is a quadratic function of wi these two criteria are equivalent. This is the case for example of the mean-squared-error classifier (Duda, 1973) with cost function

J =(1/2)||w||v (7)

D. Recursive Feature Elimination

A good feature ranking criterion is not a good feature subset ranking criterion. The criteria DJ(i) or (wi)(wi) estimate the effect of removing one feature at a time on the objective function. It will become very sub-optimal when it comes to removing several features at a time, which is necessary to obtain a small feature subset. This problem can be overcome by using the following iterative procedure that as Recursive Feature Elimination [12] . Optimize the weights wi with respect to J.

(DJ(i) or (wi)(wi). (8)

This iterative procedure is an instance of backward feature elimination. In such a case, the method produces a feature subset ranking, as opposed to a feature ranking. Feature subsets are nested.

E. Ranking with correlation coefficients

The classification of genes with the best separation between means for the two classes was by .G-S correlation. metric are chosen: GS-correlation

(g)=(µg1-µg2)/(sg1+sg2) (9)

where µg1, sg1 and µg 2, sg2 are the mean and standard deviation for values of gene g among training samples of class 1 and 2, respectively. Genes with the most positive and most negative G-S correlation values are selected in parallel and grouped together in equal number in the final classifier [4]. This method tends to not select genes for which class values have large standard deviations with respect to the training data, though some of those are most relevant and biologically informative.

IV DATABASES AND DATASETS

Blat

Blast uses a heuristic algorithm to detect relationships among sequences which share regions of similarity. The National Center for Biotechnology Information (NCBI) at the National Institutes of Health was created in 1988 to develop information systems for many resources that can be accessed through the NCBI home page at www.ncbi.nlm.nih.gov. This complex database receives data from three sources: direct submissions from external investigators, internal collecting efforts and collaborations or agreements, with data providers and research consortia (both national and international). Within NCBI operates the Online Mendelian Inheritance in Man (OMIM) database and , a catalog of human genes and genetics disorders; it contains information about linkage data, phenotypes and references on all inherited or heritable human known disorders. The OMIM comprises about diseases, genes with an associated phenotype and genes (including microRNAs) with known sequence. The information provided covers bibliography, structure, and function, association with disease and animal models.

Rat Genome Browser

Using a sequence name, gene name, locus, oligoncleotide or other landmark can search for its location on the rat genome. Links between the rat, human and mouse Genome Browse facilitate cross species comparisons.

Gene Annotator

The Gene Annotator takes a list of gene symbols, RGD IDs ,Gene Bank accession numbers, Endemic identifiers, or a chromosomal region, and retrieves annotation data from RGD. The tool will retrieve annotations from any or all ontologies usetrieve annotations from any or all ontologies used at RGD for genes and their orthologs, as well as links to additional information at other databases.

Genome Viewer

Genome Viewer provides users with complete genome view of gene and QTL annotated to a function, biological process, cellular component, phenotype, disease, or path way. The tool will search for matching terms from the gene Ontology, Mammalian Phenotype Ontology, Disease Ontology or pathway Ontology.

DATASETS

Leukemia (LEU)

Leukemia dataset composed of gene expressions in three classes of leukemia: B -cell, T-cell acute lymphoblastic leukemia and acute myeloid leukemia. The data were obtained after three pre-processing.

Lymphoma (LYM)

In order to examine the extent to which genomic-scale gene expression profiling understanding of B cell malignancies of lymphoma, studied gene expression of three prevalent adult lymphoid malignancies: B-cell chronic lymphocytic leukemia (B-CLL), follicular lymphoma (FL) and large B-cell lymphoma.

NCI 60 (NCI60)

The cell lines were derived from various tumor tissues: breast, central nervous system (CNS), colon, leukemia, and melanoma, no small cell lung carcinoma (NSCLC), ovarian, prostate, renal and unknown. The full dataset composed of samples and genes. Because the size of some classes was too small toper form discriminate analysis, used a subset with genes and six classes which was also used. Based on hierarchical clustering depicted assigned 6 classes and the size of each class respectively. Most of the samples in class are leukemia patients, and CNS is predominant in class.

Colon cancer (COLON)

A gene expression study of tumor and normal colon tissue samples which were analyzed with an Asymetrix oligo nucleotide array complementary to more than human genes. A selection of genes with highest minimal intensity across the samples has been made and this gene expression data collected with size of samples and genes.

Small round blue cell tumor (SRBCT)

The data, consisting of expression measurements on genes, were obtained from glass-slide DNA microarrays, which were prepared according to the standard of National Human Genome Research Institute. The tumors are classified as Burkitt lymphoma, Ewing sarcoma (EWS), neuroblastoma (NB), or rhabdomyosarcoma (RMS). Since this data did not make public, we used training set with size of samples and genes.

Yeast

Gene expression in the budding yeast Saccharomyces cerevisiae was studied during the diauxic shift, the mitotic cell division cycle, sporulation and temperature and reducing shocks. The data matrix consists of genes by slides.

V. CONCLUSION

A study on the method of gene selection and tissue classification based on expression data. The method used to perform a feature selection of genes such as support vector machine, random forest, Sam algorithm and genetic algorithms given. It is informed from the review that the number of gene selection has to be reduced and classification accuracy rate has to be increased. The performance measures such as feature ranking with correlation coefficients, ranking criterion and classification, feature ranking by sensitivity analysis, recursive feature elimination and ranking with correlation coefficients are also studied. And also the gene database tools are listed out in this paper. Based on the database the feature selection of genes is identified easily.

REFERENCES

[1] Roberto Ruiza,, Jose C. Riquelmea, Jesus S. Aguilar-Ruiz, .Incremental wrapper-based gene selection from microarray data for cancer classification., Pattern Recognition 39 (2006) 2383 – 2392.

[2] JinHyukHong,SungBaeCho,.Gene boosting for cancer classification based on gene expression profiles., Pattern Recognition 42 (2009) 1761 – 1767.

[3] Goldberg.D.E,.Genetic Algorithm in search optimization and machine learning., Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA ©1989 ISBN:0201157675

[4] Isabelle Guyon, Jason Weston,Stephen Barnhill, Vladimir Vapnik,. Gene Selection for Cancer Classification using Support Vector Machines., Machine Learning, 46, 389–422, 2002.

[5] Terrence S.furey,Nello cristianini,Nigel Duffy,.Support vector machine classification and validation of cancer tissue samples using microarray expression data., vol.16no.102000.

[6] Leping Li,Clarice R.Weinberg, Thomas A.Darden ,.Gene Slection a study of sensitivity to choice of parameters of the GA/KNN Method,vol.17no.122001.

[7] Fan Liand Yiming Yang,.Gene expressionAnalysis of recursive gene selection approaches from microarray data., Vol. 21 no. 19, 2005.

[8] Xin Zhou and K. Z. Mao1, .Gene expression LS Bound based gene selection for DNA microarray data.,Vol. 21 no. 8 2005.

[9] Christophe Ambroise and Geoffrey J. McLachlan,.Selection bias in gene extraction on the basis of microarray gene-expression data., Proceedings of National Academy of Sciences of United States of America, vol. 99 no. 10, , 6562–6566

[10] Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., and Levine, A.J. 1999,. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotidearrays,. Proc.Nat. Acad. Sci. USA 96, 6745–6750.

[11] Cortes, C. & Vapnik, V. (1995). Support vector networks. Machine Learning, 20:3, 273–297.

[12] 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.

[13] Ying Wai Li , Thomas Wüst, David P. Landau ,. Monte Carlo simulations of the HP model (the .Ising model. of protein folding)., Computer Physics Communications 182 (2011) 1896–1899.

[14] Dina A. Salem, Rania Ahmed A. A. Abul Seoud, and Hesham A. Ali, .A New Gene Selection Technique Based on Hybrid Methods for Cancer Classification Using Microarrays., .International Journal of Bioscience, Biochemistry and Bioinformatics, Vol. 1, No. 4, November 2011.

[15] Kohbalan Moorthy & Mohd Saberi Mohamad,.Random forest for gene selection and microarraydata classification., Bioinformation. 2011; 7(3): 142–146.

[16] Tao Li, ChengliangZhang and Mitsunori Ogihara,.A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression ., Vol. 20 no. 15 2004, pages 2429–2437

[17] Yang Ai-Junand Song Xin-Yuan,.Bayesian variable selection for disease classification using gene expression data., Vol. 26 no. 2, 2010, pages 215–222

[18] Hong Hu, Jiuyong Li, Hua Wang, and Grant Daggard,. Combined Gene Selection Methods for Microarray Data Analysis Knowledge-Based Intelligent Information and Engineering Systems Lecture Notes in Computer Science Volume 4251, 2006, pp 976-983

[19] R.Sivaraj International Journal of Engineering Science and Technology (IJEST),. A Review Of Selection Methods In Genetic Algorithm., Issn : 0975-5462 Vol. 3 No. 5 May 2011 3793.

[20] Mohd Saberi Mohamad · Sigeru Omatu · Safaai Deris Siti Zaiton Mohd Hashim,. A model for gene selection and classifi cation of gene expression data., Artif Life Robotics (2007) 11:219–222

[21] Kohbalan Moorthy and Mohd Saberi Mohamad,. Random Forest for Gene Selection and Microarray Data Classification, Bioinformation. 2011; 7(3): 142–146.

[22] Yu Wanga,, Igor V. Tetkoa, Mark A. Hallb, Eibe Frankb, Axel Faciusa, Klaus F.X. Mayera, Hans W. Mewesa,c,. Gene selection from microarray data for cancer classification—a machine learning approach., Computational Biology and Chemistry 29 (2005) 37–46

[23] Statnikov A, Aliferis C, Tsamardinos I, Hardin D, Levy S (2005),. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis., Bioinformatics 21: 631–643.

[24] Vapnik V (2000) ,.The nature of statistical learning theory., Information Science and Statistics, ISBN 978-1-4757-3264-1

[25] Golub T.R., Slonim D.K. and Tamayo, .Classification of Cancer: Class dis- covery and Class Prediction by Gene Expression Monitoring. Science. 286 (1999) 315-333

[26] Dingfang Li, Wen Zhang ,.Gene Selection Using Rough Set Theory ., Rough Sets and Knowledge Technology Lecture Notes in Computer Science Volume 4062, 2006, pp 778-785

[27] Ben-Dor, A., Bruhm, L. and Friedman,. Tissue Classification with Gene Expression Profiles.Computational Biology., (2000) 559-584.

[28] Jaeger, J., Sengupta, R., Ruzzo,.Improved gene selection for classificationof microarrays. Pacific Symposium on Biocomputing, (2003) 53-64.

[29] http://www.bioinformaticsweb.net/datalink.html

[30] http://www.science.co.il/Biomedical/Structure-Databases.asp

[31] http://scop.mrc-lmb.cam.ac.uk.