A Comprehensive Study on Gene Selection and Tissue Samples Techniques

D. Ramyachitra
Assistant Professor
Department of Computer Science
Bharathiar University
Coimbatore

M.Sofia
M.Phil Research Scholar
Department of Computer Science
Bharathiar University
Coimbatore

ABSTRACT

The fundamental power of microarrays lies within the ability to conduct parallel surveys of gene expression using microarray data. The classification of tissue samples based on gene expression data is an important problem in medical diagnosis of diseases such as cancer. In gene expression data, the number of genes is usually very high compared to the number of data samples. The difficulty lies in the data are of high dimensionality and the sample size is small.The combination of filter,wrapper and embedded methods in generalis to improve the accuracy performance of gene expression data classification.The genes identified are subsequently used to classify independent test set samples. The different feature selection methods are investigated and most frequent features are selected among all methods.This paper provides a comparative study of gene selection strategies for multi-class classification that can be used to reach high prediction accuracies with a tiny low number of selected genes.

Keyword

Microarray, Gene selection, Tissue samples,Classification, Random forest

I INTRODUCTION

The classification of gene expression data is challenging due to the enormous number of genes to the number of samples. It is common that a large number of genes are either irrelevant or redundant [1]. For that, it is significant to reduce the number of genes in order to get a good accuracy for the classification. Filter approach, wrapper approach and embedded approach are widely used for feature selection of genes [2]. Filter methods arethe ones that select features as a pre-processing step. That is, they select features without considering theclassification accuracy. Filter type are essentially data pre-processing or data filtering methods. Features are selected based on the intrinsic characteristics, which determine their relevance or discriminate powers with regard to the targeted classes [2]. Based on mutual information the simple method can give effective statistical tests [5][6]. They also have the virtue of being easily and very efficiently computed.

In filters, the characteristics in the feature selection are uncorrelated to that of the learning methods, therefore they have better generalization property[1].The filters, wrapper and embedded are then analyzed to identify the most frequently appearinggenes which would correspond to the most predictive genes[2]. TheGA combined with a SVM classifier is used for selecting predictive genesand for final gene selection and classification. The analysis of gene expression data isto identify the sets of genes as classification or diagnosis platforms.Machine learning techniques, such as artificial neural networks (ANNs), present a more flexible`model-free' approach for classification and frequently yield good results[6].The advantage of selecting a combination of genes with small redundancy,favors the selection of mutually uncorrelated genes. The selected set of paired genes was used as a new feature set for the classification.

In wrapper type methods, feature selection is "wrapped" around a learning method a feature is directly judged by the estimated accuracy of the learning method [11]. One can often obtain a set with a very small number of non-redundant features, which gives high accuracy, because the characteristics of the features match well with the characteristics of the learning method[11]. The wrapper methods use the predictive accuracy of an algorithm to evaluate the possible subset of features and select the subset of features that provides the highest accuracy.

Embedded methods differs from other feature selection methods in the way feature selection and learning interact[39].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[22].

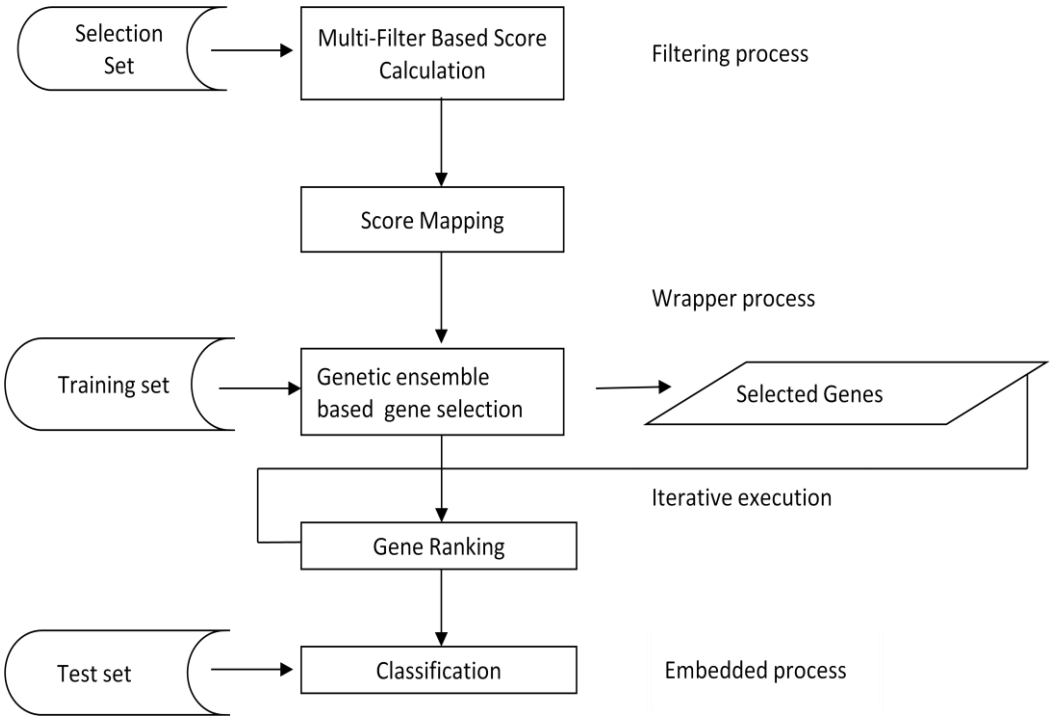

Figure 1.marginal filters, wrappers and embedded methods.

1.1 THE UNIVARIATE FILTER

The univariate filter paradigms are simple and efficientbecause of the high dimensionality of most microarray analyses,fast and efficient techniques such as univariate filter methods have attracted most attention [22]. The prevalence of these univariatetechniques has dominated the field, and now comparativeevaluations of different classification techniques over gene and tissuemicroarray datasets only focused on the univariate case [12] . This domination of the univariate approach can be explainedby a number of reasons:

P. Duda, The output provided by univariate feature rankings is intuitiveand easy to understand [22].

D. Stork The gene ranking output could fulfill the objectives andexpectations that bio-domain experts have when wanting tosubsequently validate the result by laboratory techniques or inorder to explore literature searches [29].

P. Hart, the possible unawareness of subgroups of gene expressiondomain experts about the existence of data analysis techniquesto select genes in a multivariate way [22].

1.2 The Multivariate Paradigm

The application of multivariate filter methods ranges from simplebivariate interactions [10] towards more advanced solutionsexploring higher order interactions, such as correlation based feature selection (CFS) [12, 13] and several variants of the Markovblanket filter method . The Minimum Redundancy- Maximum Relevance (MRMR) [26] and Uncorrelated ShrunkenCentroid (USC) algorithms are two other solid multivariatefilter procedures, highlighting the advantage of using multivariatemethods over univariate procedures in the gene expression domain.

M. Daly, J. Rioux, feature selection using wrapper or embedded methods offers analternative way to perform a multivariate gene subset selection,incorporating the classifier's bias into the search and thus offering anopportunity to construct more accurate classifiers[21].

1.3 Mass Spectra Analysis

Mass spectrometry technology (MS) is emerging as a new andattractive framework for disease diagnosis and protein-basedbiomarker profiling [22]. A mass spectrum sample is characterizedby thousands of different mass/charge ratios on the x-axis,each with their corresponding signal intensity value on the y-axis. For data mining and bioinformatics purposes, it can initially beassumed that each $\frac{m}{z}$ ratio represents a distinct variable whosevalue is the intensity [12].

E. Petricoin and L. Liotta.a feature extraction step is thus advisable to set the computational costs of many feature selection techniques to a feasible size in these MS scenarios.

As Somorjai et al, the dataanalysis step is severely constrained by both high dimensional inputspaces and their inherent sparseness, just as it is the case with geneexpression datasets.

1.4 Tissue Samples

The expression levels for genes in tissue or cell samples consists of a relatively small number of tissue in the analysis of similar datasets[1].The number of features is large compared to the number of samples. Tissues encoded in the expression levels of few genes, are able to understand the biological significance of these genes[3]. Moreover, a major goal for tissue research is to develop

diagnostic procedures based on inexpensive microarrays that have enough probes to detect the tissues. Thus, it is crucial to recognize whether a small number of genes are good for classification. The problem of feature selection received a thorough treatment in pattern recognition and machinelearning [4]. The gene expression data sets are problematic in that they contain a large number of genes(features) and thus methods that search over subsets of features can be prohibitively expensive. Moreover,these data sets contain only a small number of samples, so the detection of irrelevant genes can sufferfrom statistical instabilities.
Golub *et al.* (1999) also describe gene selection methods for improving classification accuracy [5].

P. Hart an interval-valued analysis method based on a rough set technique to selectdiscriminative genes and to use these genes to classify tissue samples of microarray data [7].

The remaining sections of this paper are organized as follows section 2 describes the gene selection techniques, section 3 describes the performance metrics, section 4 describes the databases and datasets of gene selection techniques. Finally section 5 gives the conclusion.

2.GENE SELECTION TECHNIQUES

2.1. Using clustering for classification

When applied to expression patterns, clustering techniques attempt to partition the set of elements into clusters of patterns, so that all the patterns within a cluster are similar to each other and different from patterns in other clusters.

Alon suggests that if the labeling of patterns is correlated with the patterns then unsupervised clustering of the data would cluster patterns with the same label together and separate patterns with different labels[3].

Ben-dor et al involves gene expression patterns from colon samples that include both genes and normal tissues. Applying a hierarchical clustering procedure to the data.

Alon *et al.* observe that the topmostdivision in the selection divides samples into two groups, one predominantly tissue, and the other predominantly normal.

2.1.1The clustering algorithm

en-Dor CAST algorithm, implemented in the ioClust analysis software package takes as input a threshold parameter, which controls the granularity of the resulting cluster structure, and a similarity measure between the tissues [5].

BWratio [9] clustering techniques are computationally efficient, but they fail to deal with redundancy between the selected genes.

Hastie *et al.* generate eight genes having clusters, withsizes chosen by the gap-statistic. Since the genes havingclusters are (almost) orthogonal, simple LDA classifiers is used in conjunction with supervised genes[12].

Li and Hong, cluster genes and use cluster centroids with ‘soft-max’ sample classification there is no selection of centroids, all of them are used for classification [4].

2.1.2 Clustering based classification

As Alon and Eisen, the threshold parameter determines the cohesiveness of the resulting clusters as well their number [6]. A similar situation occurs in other clustering algorithms. In hierarchical clustering algorithms, the cutoff “level” of the tree controls the number of clusters.

Alon *et al.*, In any clustering algorithm, it is clear that attempting to partition the data into exactly two clusters will not be the optimal choice for predicting label[12]s.

Eisen *et al.*For the purpose of determining the right parameter to be used in clustering data that contains some labeled samples a measure of cluster structure *compatibility* with a given label assignment[6].

Everittalthough the clustering algorithm is unsupervised, in the sense that it does not take into account the labels [13].

2.1.3Large-margin classifiers

The cluster-based approach attempts to inherent structure inthe data and uses this structure for prediction.In the clustering method a decision surfacethat separates the positively labeled samples from the negatively labeled samples.The literature of supervised learning discusses a large number of methods that learn decision surfaces.

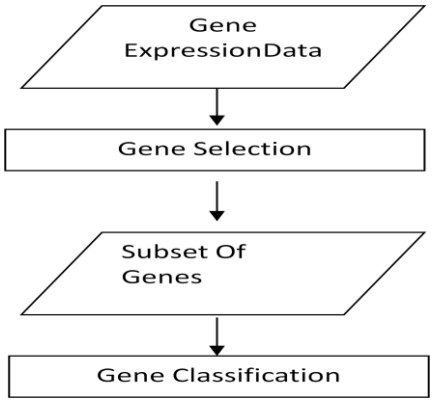

Figure 2.A model for gene selection and classification

2.2 The 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.

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].

Ma and Huang [13] the metaheuristic methods like local search, genetic and memetic algorithms (MAs) can be designed to deal with gene selection and classification of microarray data.

Duval and Hao to the first kind of embedded methods since the exploration of gene subsets is guided by information provided by the classifier [5].

Trent Higgs et al., present a feature based re sampling genetic algorithm to refine structures that are outputted [6].

A. Tantara et al., proposeda parallel hybrid genetic algorithm (GA) which is used to efficiently solve the problem. It is used by defining not only the gene selection but also the [12]
accuracy.

2.2.1 Crossover

Crossover is simply a matter of replacing some of the genes in one parent by the corresponding genes of the other. One-point crossover has been described as two-pointcrossover[27]. Two cross points are chosen at random from the numbers and a new solution produced by combining the pieces of the original 'parents'.

Eshelman et al an early and thorough investigation of multi-point crossovers [23] the biasing effect of traditional one-point crossover, and considered a range of alternatives.

DeJong and Spears [23] produced a theoretical analysis that was able to characterizethe amount of disruption introduced by a given crossover operator exactly.

Booker [23] reported significant gains from using an adaptive crossover rate therate was varied according to a characteristic called percent involvement.

2.2.2 Mutation

The concept of mutation is even simpler than crossover, and again, this can easily be represented as a bit-string. There are different ways of implementing this simple idea that can make a substantial difference to the performance of a GA [29]. The naive idea would be to draw a random number for everygene in the string and compare it to but this is potentially expensive in terms of computation if the strings are long and the population is large. An efficient alternative is to draw a random variate from a Poisson distributionwith parameter where is the average number of mutations per chromosome [12].

Perhaps it isbest to say that the balance between crossover and mutation is often a problem-specificone, and definite guidelines are hard to give.

Fogarty [27] experimented with different mutation rates at different loci.

Reeves [29] varied the mutation probability according to the diversity in the population sophisticated proceduresare possible and anecdotal evidence ofdiversity maintenance policy.

2.3 The Support Vector Machine Classifier

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 hyperplane between the classes. When the classes are linearly separable, the hyperplane 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 hyperplane; in this case it tries to maximize the positive genes but allow some classification errors to the constraint that the total error is less than a negative gene. There are several possible approaches;

In this support vector machine method "one against- one" approach, as implemented in "libsvgm" [12] Chan CC 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 informativepatterns[4].

Sujun Hua et al[23] ., represented a new approach to supervised pattern classification applied to a pattern recognition problems, including object recognition, speaker identification, gene function prediction with microarray expression profile, etc.

Minh N. Nguyen et al[14]., investigates the multi-class SVM methods involved to resolve a much larger optimization problem and are applicable to small datasets. The multi-class SVM methods are more suitable for prediction[25] than the other methods. Duval extend the prediction accuracy by adding a second-stage multi-class SVM to capture the information among the genes [12].

2.3.1Cross-validation:

Cross-Validation (CV) is very helpful in evaluating and comparing learning algorithms. It is a statistical technique used during the training process of the classifier where its task is to divide the train dataset into two segments; one is used for training and the other is used for validation [12].

2.3.2SVM-classifier:

Support vector machine uses the SVMstructureto classify the test data into the predefined classes.As the cross validation and SVM parameters are accurately chosen, asthe classification accuracy of the test samples increases [18].The cross-validation coupled with the SVM-trainer runsseveral times in a continuous loop until reaching maximumtrain classification accuracy [22].

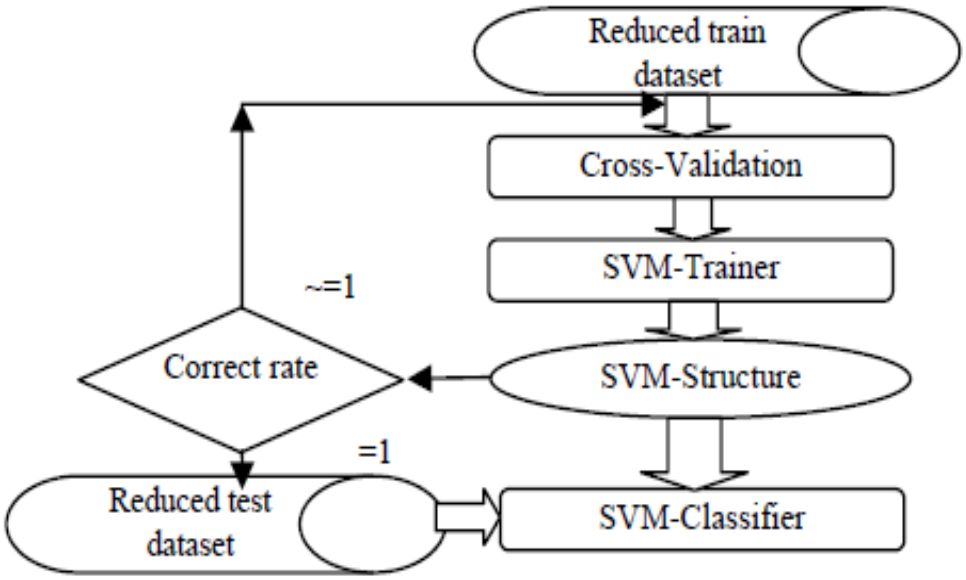

Figure. 3. SVM general scheme.

2.4 Random Forest

Random forest returns several measures of variableimportance. The most reliable measure is based on thedecrease of classification accuracy when values of a variablein a node of a tree are permuted randomly [13],and this is the measure of variable importance .The measure of variableimportance is not the same as a non-parametric statistic. Other measures ofvariable importance are available, for the performance of differentmeasures of importance [26].

Lee et al generates a gene expression classification profiles which can discriminate between different known cell types or conditions [27].

Leo Breiman [13] that uses sample genes forclassification [14, 16] in a random subset of the variables.

Adele Cutler each experiment is repeated for the average accuracy, sensitivity, and specificity of the trails [28].

P. Hart classifier model is constructing multiple decision trees, each of which uses a subset of genes randomly selected from the whole original set of gene [21].

2.4.1Standard Random Forest Gene Selection Method

In the standard random forest gene selection, the selection of the genes is done by using both the backward gene elimination and the selection based on the accuracy [31]. The backward elimination is done for the selection of small sets of non redundant variables, and the accuracy for the selection of large, potentially highly correlated variables [22].

Leo Breimanusing the default parameters [3], all forests that result from iterative elimination based on fraction dropped value, a fraction of the least importance. The default fraction.dropped value which allows for relatively fast operation is consistent with the idea of an aggressive gene selection approach, and increases the resolution as the number of variables considered becomes smaller.

2.4.2 Improvement Made to the Random Forest Gene Selection Method

The random forest gene selection, which includes automated dataset input that simplifies the task of loading and processing of the dataset to an appropriate format so that it can be used for accuracy [27]. Integration of the different approaches into a single function with parameters as an option allows greater usability while maintaining thecomputation time required [29].

Random forest can be used for problems arising from more than two classes (multi class) as stated by Díaz-Uriarte R & Alvarez de Andrés[18].

2.5 Sam Algorithm

SAM identifies genes with statistically significant changes in expression by assimilating a set of gene-specific *t* tests [30]. Each gene is assigned a score on the basis of its change in gene expression relative to the standard deviation of repeated measurements for that gene. Genes with scores greater than a threshold are deemed potentially significant [31].

Tusher VG, Tibshirani R says that SAM is arguably the most widely utilized method in the microarray analysis field [30].

Harris MA, Clark J suggested that one standard method for carrying out such analysis is theuse of gene-grouping, such as Gene ontology classification [9].

As Irizarry RA when comparing two of the time points during the induction of sense were unable to observe any effect of selection within SAM on the yield of significant genes [31].

2.6 Rough Set Theory

Gene selection algorithm based on roughest theory for gene expression data is composed of t-test and feature selection based on the rough set theory [12].T-test is helpful for reducing dimensionality. The algorithm without the t-test preprocessing will get worseperformance. After feature ranking, top ranked n genes are selected to form the feature set. The values of all continuous features are discredited. Roughest theory-based feature selection method starts with the full set and consecutively deletes one feature at a time until a reduction[14].

L.Shenproposed the rough sets theory and its applications are often computationally efficient techniques for addressing problems such as hidden data discovery, data reduction, data significance evaluation and decision rule generation[12].

According to Ron Kohavi's research, the best subset in feature selection may not be a reduct and the reduct may lead to the unfavorable performance when being used to train classifiers [12].

Shannon [16], a useful mechanism for characterizing the information content in various modes and applications in many diverse fields.

Pawlak pointed out that one of the most important and a fundamental role of the rough sets philosophy is the need to discover redundancy and dependencies between features [19].

3. PERFORMANCE METRICS

3.1 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:

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

Whereμi and σi are the mean and standard deviation of the gene expression values of gene i for all the patients of class (+) or class (−), $i = 1, \ldots n$.

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

3.2 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 discriminant classifier:

$$D(\mathbf{x})=(\mathbf{x}-\mu)(3) \quad (3)$$

where **w** is defined in

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

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

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

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 discriminant 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.

3.3 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:

$$DJ (i)= (1/2) \partial 2 J/\partial w2i (Dwi)2 \quad (6)$$

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 pruningcriterion. For linear discriminant 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

$$J =(1/2)||w||v \quad (7)$$

3.4 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 $F1 \subset F2 \subset \cdots F$.

3.5 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)/(σg1+σg2) (9)

where μg1, σg1 and μg 2, σg2 are the mean and standard deviation for values of gene g among trainingsamples of class 1 and 2, respectively. Genes with the most positive and most negative G-S correlationvalues are selected in parallel and grouped together in equal number in the final classifier [4]. Thismethod tends to not select genes for which class values have large standard deviations with respectto the training data, though some of those are most relevant and biologically informative.

4.DATABASES AND DATASETS

4.1 Database

4.1.1Blast

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.

4.1.2 Rat Genome Browser

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

4.1.3Gene Annotator

The Gene Annotator takes a list of gene symbols,RGD IDs ,GeneBank accession numbers, Ensembi 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.

4.1.4Genome Viewer

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

4.2 Datasets

4.2.1Leukemia (LEU)

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

4.2.2 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 chroniclymphocytic leukemia (B-CLL), follicular lymphoma (FL) and large B-celllymphoma.

4.2.3NCI 60 (NCI60)

The cell lines were derived from various tumor tissues: breast, central nervoussystem (CNS), colon, leukemia, and melanoma, no small cell lung carcinoma(NSCLC), ovarian, prostate, renaland unknown. The full dataset composedof samplesand genes. Because the size of some classes was too small toperform discriminant analysis, used a subset with genes and six classes whichwas 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.

4.2.4 Colon cancer (COLON)

A gene expression study of tumorand normal colontissue samples which were analyzed with an Asymetrix oligonucleotide array complementaryto more than humangenes. A selection of genes with highest minimal intensity across the samples has been made and this gene expression data collected with size of samplesand genes.

4.2.5Small round blue cell tumor (SRBCT)

The data, consisting of expression measurements on genes, were obtained from glass-slidecDNA microarrays, which were prepared according to the standard of National HumanGenome 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 setwith size of samplesand genes.

4.2.6 Yeast

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

5. 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 datafor 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] M. Sharma, G. Singh, R. S. Virk and G. Singh, "Design and comparative analysis of DSS queries in distributed environment," *Computer Science and Engineering Conference (ICSEC), 2013 International*, Nakorn Pathom, 2013, pp. 73-78. doi: 10.1109/ICSEC.2013.6694756
[30]http://www.bioinformaticsweb.net/datalink.html
[31]http://www.science.co.il/Biomedical/Structure-Databases.asp
[32]http://scop.mrc-lmb.cam.ac.uk.
[33]http://www.bioinformaticsweb.net/data.html
[34]file:///F:/Untitled%20Document.htm
[35]file:///F:/allover/algorithms/extra/PDBsum%20entry%20%201g8p.htm
[36]http://www.bioinformaticsweb.net/toollink.html
[37]http://www.bioinformaticsweb.net/tools.html
[38]G.B. Fogel and D.W.Corne, "Evolutionary Computation in Bioinformatics", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, Vol. 36, No. 5, September 2006
[39] www.ncbi.nlm.nih.gov