MICROARRAY GENE BASED DISEASE PREDICTION USING PATTERN SIMILARITY BASED SVM CLASSIFICATION

[1]D.Santhakumar, [2]Dr.S.Logeswari

[1]Assistant Professor, Department of CSE, CK college of Engineering and Technology
[2]Professor and Head, Department of CSE, Bannari Amman Institute of Technology

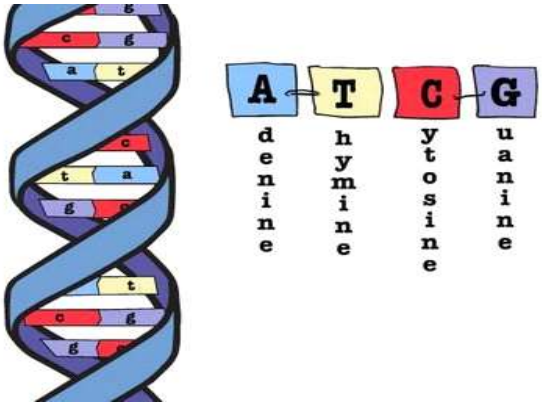
***Abstract*: The DNA microarray technology has modernized the approach of biology research in such a way that scientists can now measure the expression levels of thousands of genes simultaneously in a single experiment. Gene expression profiles, which represent the state of a cell at a molecular level, have great potential as a medical diagnosis tool. Diseases classification with gene expression data is known to include the keys for addressing the fundamental harms relating to diagnosis and discovery. The recent introduction of DNA microarray technique has complete simultaneous monitoring large number of gene expressions possible. With this large quantity of gene expression data, experts have started to discover the possibilities of disease classification using gene expression data. Quite a large number of methods have been planned in recent years with hopeful results. But there are still a set of issues which need to be address and understood. In order to gain insight into the disease classification difficulty, it is necessary to get a closer look at the problem, the proposed solutions and the associated issues all together. In this project, we present a comprehensive clustering method and classification method such as Particle Swarm Optimization (PSO), K-NN classification algorithm and estimate them based on their evaluation time, classification accuracy and ability to reveal biologically meaningful gene information. Based on our multiclass classification method to diagnosis the diseases and also find severity levels of diseases. Our experimental results show that classifier performance through graphs with improved accuracy.**

***Index Terms*: Bio-medical research, DNA microarray, Gene sequence, Clustering, Classification**

• Introduction

Microarray technology has become one of the indispensable tools that many biologists use to monitor genome wide expression levels of genes in a given organism. A microarray is typically a glass slide on to which DNA molecules are fixed in an orderly manner at specific locations called spots (or features). A microarray may contain thousands of spots and each spot may contain a few million copies of identical DNA molecules that uniquely correspond to a gene. The DNA in a spot may either be genomic DNA or short stretch of oligo-nucleotide strands that correspond to a gene. The spots are printed on to the glass slide by a robot or are synthesized by the process of photolithography. Microarrays may be used to measure gene expression in many ways, but one of the most popular applications is to compare expression of a set of genes from a cell maintained in a particular condition (condition A) to the same set of genes from a reference cell maintained under normal conditions (condition B). Clustering techniques have proven to be helpful to understand gene function, gene regulation, cellular processes, and subtypes of cells. Genes with similar expression patterns (co-expressed genes) can be clustered together with similar cellular functions. This approach may further understanding of the functions of many genes for which information has not been previously available. Furthermore, co-expressed genes in the same cluster are likely to be involved in the same cellular processes, and a strong correlation of expression patterns between those genes indicates co-regulation. Searching for common DNA sequences at the promoter regions of genes within the same cluster allows regulatory motifs specific to each gene cluster to be identified and cis-regulatory elements to be proposed. The inference of regulation through the clustering of gene expression data also gives rise to hypotheses regarding the mechanism of the transcriptional regulatory network. Finally, clustering different samples on the basis of corresponding expression profiles may reveal sub-cell types which are hard to identify by traditional morphology-based approaches.

1.1 Challenges in gene clustering:



- Related work

Booma,et.al,…[1] identifyed normal or abnormal genes is important for clinical analysis and diagnosis. In this work, a novel framework for analyzing gene data was designed and developed. For this, initially, Bio-information from gene expression data was evaluated with the establishment of analyzing biological process using heuristic search [BPPD]. BPPD method identified the biological process on physiological data using heuristic search algorithm in rough set theory for gene-expression data analysis. This method extracted the biological process on gene expression data. The proposed method used heuristic search algorithm for identifying the biological process and processed based on two phases. The first phase was initialization phase and another was iterative adjustment phase. With respect to these two phases, the biological process of each gene and gene selection for a dataset is identified in terms of physiological data on gene expression datasets. Experimental evaluations are conducted for heuristic search based analysis of biological process on physiological data with standard benchmark gene expression data sets from research repositories such as broad institute in terms of size of gene expression datasets. Finally, issues in presence of extract the bi-cluster based gene expression information was addressed with proposed Bi-clustered Ant Optimized Feature Relational Sequencing (BAOFRS) method. The features used to identify the relational sequences also computed the similarity value between the sequences BAOFRS method used the K-mers relational knowledge sequence to identify the relational features. Jaccard similarity coefficient was applied to identify the similarity value on relational features.

Balasubramanian,et.al,…[2] proposed fuzzy logic based preprocessing technique to reduce the redundant information and grouping the similar genes from large amount of microarray data. The propose Parallel Island Model GA is implemented for gene feature selection process. Our propose feature selection algorithm is implemented based on multi objective genetic algorithm. This uses a different operator called multi objective operator. Multi objective aspect is defined to find the pareto optimal solutions for ranking. Since the search space is large and requires a good diversity, island model has been proposed. Finally the Fuzzy Based Parallel Island Model GA has been implemented by using parallelization tool Open MP.This FPIMMOGA is used to progress the gene subsets and whose fitness is calculated by parallel version of SVM classifier. We have done Fuzzy preprocessing technique for reducing the input data and implemented our FPIMMOGA using parallel programming (Open MP) .The best features are selected in short time. The best identified gene subsets are evaluated by parallel version of SVM Classifier. The method has given good classification accuracy than other methods. This method uses the island model for generating the best population. The multiple islands are implemented in parallel, which has significantly reduced the execution time in the process of best feature selection. In this work standard microarray breast cancer data sets are taken from Kent Ridge Biomedical Data Set Repository

Bennet,et,al,… [3] implemented this method, a search is conducted in the space of genes, evaluating the goodness of each gene subset by the estimation of the accuracy percentage of the specific classifier to be used, training the classifier only with the found genes. It is claimed that this approach obtains better predictive accuracy estimates than the previous approach. A common drawback in this method is that they have a higher risk of over fitting than filter techniques and are very computationally intensive. In contrast, it incorporate the interaction between genes selection and classification model, which make them unique compared to existing ones. Classification of cancer based on gene expression data is a promising research area in the field of data mining. In this paper, hybrid gene selection technique which combines SVM-RFE and BBF has been proposed for gene selection. Based on the experimental results on leukemia dataset it is found that the performance of SVM-RFE and BBF combined with SVM for classification was superior to the previous related works in terms of gene selection and classification. SVM-RFE ranks the genes and BBF is applied to remove redundancy on top ranked genes. Moreover, several gene selection methods against different classifiers were compared. This approach can play a vital role in accurate cancer classification thus, eliminating the morphological and clinical means of diagnosis.

Nagpal, et.al,… [4] proposed the system to classify the genes which is the very critical and challenging job. Many researchers has already practice in this field so they planed many algorithms related to data mining such that decision tree methods, the linear discrimination analysis, the RBF network, Genetic algorithm etc. Most proposed cancer classification methods are related to the data mining or soft computing area. Like nearest neighbor analysis, Back propagation network analysis, Fuzzy logic analysis. Mostly of the methods are work fine on binary-class problems and not provide well result in multi-class problems. Most researchers only concerned with the accuracy of the classification. One another problem is gene classifiers proposed are quite computationally expensive they cannot afford to the all people. Exact classification of cancers based on microarray gene expressions is very crucial for doctor to select a proper treatment. Gene expression data, obtained by DNA micro arrays, has been used to investigate the biological terms of tumors and to concatenate expression patterns with clinical results for patients in various stages and different types of diseases. EPSO (Elitism Particle Swarm Optimization) is one of the Features selection or Attribute selection methods. Feature selection is important techniques for identifying informative genes in microarray datasets.

- existing methodologies

Cancer research is one of the major research areas in the medical field. Accurate prediction of different tumor types has great value in providing better treatment and toxicity minimization on the patients. Different classification methods from statistical and machine learning area have been applied to cancer classification, but there are some issues that make it a nontrivial task. The gene expression data is very different from any of the data these methods had previously dealt with. First, it has very high dimensionality, usually contains thousands to tens of thousands of genes. Second, publicly available data size is very small, all below 100. Third, most genes are irrelevant to cancer distinction. It is obvious that those existing classification methods were not designed to handle this kind of data efficiently and effectively. Some researchers proposed to do gene selection prior to cancer classification. Performing gene selection helps to reduce data size thus improving the running time. In this existing system, we present a comprehensive

overview of various cancer classification methods and evaluate them based on their computation time, classification accuracy and ability to reveal biologically meaningful gene information. We also introduce and evaluate various gene selection methods which we believe should be an integral preprocessing step for cancer classification. In order to obtain a full picture of cancer classification, we also discuss several issues related to cancer classification, including the biological significance vs. statistical significance of a cancer classifier, the asymmetrical classification errors for cancer classifiers, and the gene contamination problem.

- GENE BASED DISEASE PREDICTION

Microarray technology has made the modern biological research by permitting the simultaneous study of genes comprising a large part of the genome. In response to the rapid development of DNA Micro array technology, classification methods and gene selection techniques are being computed for better use of classification algorithm in microarray gene expression data. Microarrays are capable of determining the expression levels of thousands of genes simultaneously. One important application of gene expression data is classification of samples into categories. In combination with classification methods, this tool can be useful to support clinical management decisions for individual patients, e.g. in oncology. Standard statistic methodologies in classification or prediction do not work well when the number of variables p (genes) far too exceeds the number of samples n which is the case in gene microarray expression data. The goal of our proposed project will be to use supervised learning to classify and predict diseases, based on the gene expressions collected from microarrays. Known sets of data will be used to train the machine learning protocols to categorize diseases according to their gene patterns. The outcome of this study will provide information regarding the efficiency of the machine learning techniques, in particular a KNN method. The efficiency of classification depends on the type of kernel function that is used. So here we will analyze the performance of various kernel functions used for classification purpose. Finally predict the diseases with severity levels and predict various types of diseases. Fig 2 shows proposed framework.

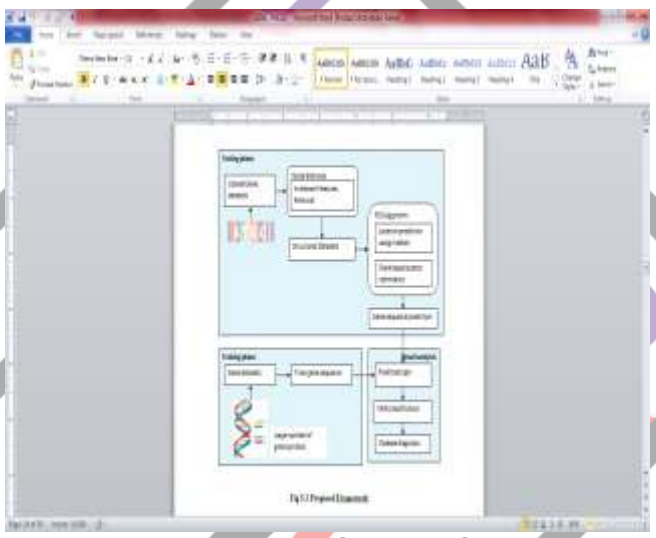


Fig 2: Proposed Framework

DATASETS ACQUISITION

In this module, upload the datasets. The dataset may be microarray dataset. A microarray database is a repository containing microarray gene expression data. Then implement preprocessing steps to eliminate the irrelevant symbols.

PSO ALGORITHM

In PSO algorithm, can analyze coverage of the data before clustering begins. And propose an algorithm, which modifies the nearest centroid sorting and the transfer algorithm, of the spatial medians clustering. It has two distinct phases: one of transferring an object from one cluster to another and the other of amalgamating the single member cluster with it's the nearest cluster. Given a starting partition, each possible transfer is tested in turn to see if it would improve the value of clustering criterion. When no further transfers can improve the criterion value, each possible amalgamation of the single member cluster and other clusters is tested.

DISEASE PREDICTION

Classifiers based on gene expression are generally probabilistic, that is they only predict that a certain percentage of the individuals that have a given expression profile will also have the phenotype, or outcome, of interest. Therefore, statistical validation is necessary before models can be employed, especially in clinical settings. In this module implement K nearest neighbor algorithm to classify the various types of diseases from gene expression. Classification is done with the help of KNN classifier. In the recent years, KNN classifiers have established excellent performance in a variety of pattern recognition troubles. The input space is planned into a high dimensional feature space. Then, the hyper plane that exploits the margin of separation between classes is constructed. The points that lie closest to the decision surface are called support vectors directly involves its location. When the classes are non-separable, the optimal hyper plane is the one that minimizes the probability of classification error. Initially input image is formulated in feature vectors. Then these feature vectors mapped with the help of kernel function in the feature space. And finally division is computed in the feature space to separate out the classes for training data. A global hyper plane is required by the KNN in order to divide both the program of examples in training set and avoid over fitting. This phenomenon of KNN is higher in comparison to

other machine learning techniques which are based on artificial intelligence. Here the important feature for the classification is the width of the vessels. With the help of KNN classifier we can easily separate out the vessels into arteries and veins. The KNNs demonstrate various attractive features such as good generalization ability compared to other classifiers. Indeed, there are relatively few free parameters to adjust and it is not required to find the architecture experimentally. The algorithm steps as follows:

for all the unknown samples UnSample(i)

for all the known samples Sample(j)

compute the distance between

Unsamples(i) and Sample(j)

end for

fing the k smallest distances

locate the corresponding samples

Sample(j1),….,Sample(jK)

assignUnSample(i) to the class which appears more frequently

end for

The performance of a KNN classifier is primarily determined by the choice of K as well as the distance metric applied. The estimate is affected by the sensitivity of the selection of the neighborhood size K, because the radius of the local region is determined by the distance of the Kth nearest neighbor to the query and different K yields different conditional class probabilities.

SEVERITY ANALYSIS

Using multi class classification algorithm to classify the severity level of diseases using classified data count. If count is more than threshold means, provide severity as high and count is less than threshold means, consider as normal. Then provide prescription to patients according to the diseases.

- Experimental Results

We can implement this system for uploading the gene datasets from NCBI Repository from this link https://www.ncbi.nlm.nih.gov/genbank/. And we can perform gene clustering and classification using ASP.NET (C#) as Front End and SQL SERVER as Back End for WINDOWS OS with any configuration.
KNN algorithm can be implemented and calculate the performance metrics for accuracy based on True positive rate, False positive rate, True negative rate and False negative rate.

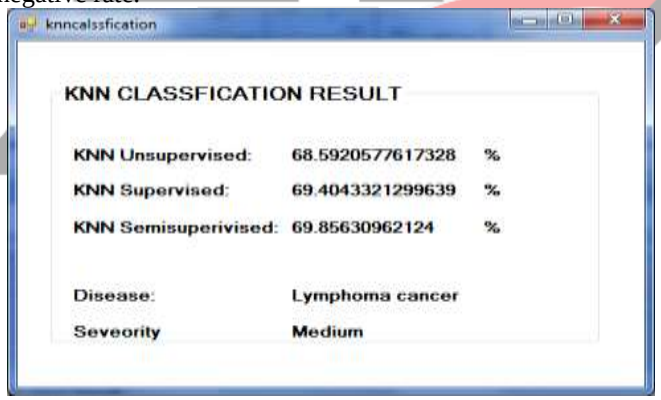


Fig 3 Accuracy rate

Accuracy rate is calculated as

*100

And compare the results with existing unsupervised, supervised algorithms. The proposed semi-supervised algorithm provide improved accuracy rate than the existing algorithms

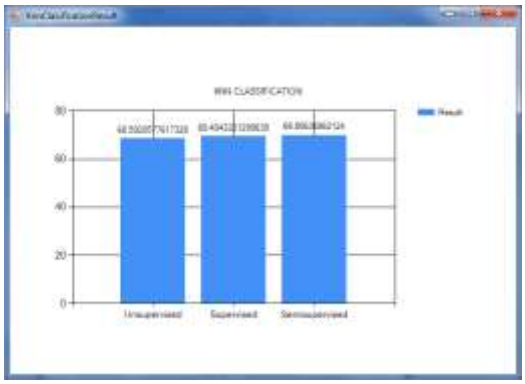


Fig 4 Performance Chart

The performance result is shown in fig 4 and KNN algorithm provides 70% accuracy than the existing algorithms.

Conclusion

Microarray is an important tool for cancer classification at the molecular level. It monitors the expression levels of large number of genes in parallel. With large amount of expression data obtained through microarray experiments, suitable statistical and machine learning methods are needed to search for genes that are relevant to the identification of different types of disease tissues. In this thesis, we have proposed a hybrid gene selection method, which combines a PSO methods and KNN classification to achieve high classification performance. The method was designed to address the importance of gene ranking and selection prior to classification, which improves the prediction strength of the classifier. The project focused on promising accuracy results with very few number of gene subsets enabling the doctors to predict the type of cancer. The results on various disease datasets shows the importance of the same classifier used for both the gene selection and classification can improve the strength of the model. Then provide severity level for each classified diseases. Future work includes partitioning of the original gene set into some distinct subsets or clusters so that the genes within a cluster are tightly coupled with strong association to the sample categories. We can extend the work to implement various classification algorithms to improve the accuracy rate at the time of disease prediction

References

[1] Booma, P. M., and S. Prabhakaran. "Classification of genes for disease identification using data mining techniques." Journal of Theoretical and Applied Information Technology 83.3 (2016): 399.

[2] Natarajan, A., and R. Balasubramanian. "A Fuzzy Parallel Island Model Multi Objective Genetic Algorithm Gene Feature Selection For Microarray Classification." International Journal of Applied Engineering Research 11.4 (2016): 2761-2770.

[3] Bennet, Jaison, ChilambuchelvanGanaprakasam, and Nirmal Kumar. "A hybrid approach for gene selection and classification using support vector machine." Int. Arab J. Inf. Technol. 12.6A (2015): 695-700.

[4] Nagpal, Rashmi, and RashmiShrivas. "Cancer Classification Using Elitism PSO Based Lezy IBK on Gene Expression Data." Journal of Scientific and Technical Advancements 1.4 (2015): 19-23.

[5] Thangaraju, Mr P., and R. Mehala. "Novel Classification based approaches over Cancer Diseases." system 4.3 (2015).

[6] Park, Heewon, et al. "A novel adaptive penalized logistic regression for uncovering biomarker associated with anti-cancer drug sensitivity." IEEE/ACM transactions on computational biology and bioinformatics 14.4 (2017): 771-782.

[7] Nakariyakul, Songyot. "Gene selection using interaction information for microarray-based cancer classification." Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2016 IEEE Conference on. IEEE, 2016.