ISSN: 1832-5505

Applied GIS

Vol-10 Issue-02 April 2022

GENE BASED DISEASE PREDICTION USING PATTERN SIMILARITY BASED CLASSIFICATION

Mrs.M.MANGALAM.T.ANITHA
Assistant Professor[1,2]
MRK INSTITUTEOF TECHNOLOGY

Abstract— Due to the development of DNA microarray technology, biology study has become more modern. Now, scientists can test thousands of genes at the same time using this technology. Gene expression maps can show what's going on with a cell at the molecular level and could be very useful for diagnosing health problems. It is known that using gene expression data to classify diseases is the key to fixing the main problems that come up with detection and finding. With the new development of the DNA microarray method, it is now possible to watch a lot of gene activity at the same time. With this huge amount of gene expression data, scientists have begun to look into how gene expression data could be used to classify diseases. In the past few years, a lot of different ways have been planned with the hope of getting good results. But there are still some problems that need to be fixed and understood. To understand the problem of disease classification, we need to take a closer look at the problem, the suggested answers, and the problems that go along with them. In this project, we show a complete grouping and sorting method, including Particle Swarm Optimization (PSO) and the K-NN classification algorithm. We rate them based on how long they take to evaluate, how well they sort, and how well they can reveal gene information that is medically important. Based on our multiclass classification method, we can diagnose illnesses and figure out how bad they are. The results of our tests show that the predictor works better when we use plots to show its success.

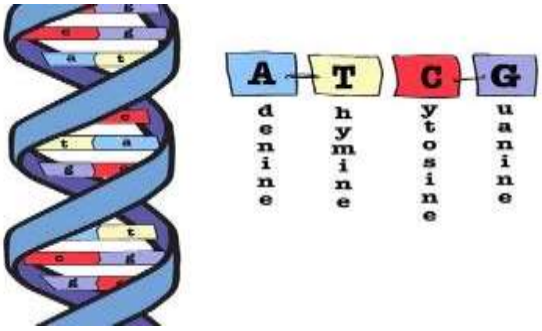
Index Terms— Biomedical study, Gene sequence, DNA microarray, Clustering, and Classification

Introduction

Microarray technology is now one of the most important tools scientists use to keep an eye on how genes are expressed across an organism's whole genome. A microarray is usually a glass slide that has DNA molecules stuck to it in a neat pattern at certain places, known as spots (or features). There could be thousands of spots on a microarray, and each spot could have a few million copies of the same DNA molecules that each relate to a different gene. In a certain area, the DNA could be chromosomal DNA or a short stretch of oligo-nucleotide strands that are linked to a gene. A robot prints the spots on the glass slide, or they are made through a process called photolithography. In many ways, microarrays can be used to measure gene expression. One of the most common uses is to compare the expression of a group of genes from a cell that is kept in a certain condition (condition A) to the expression of the same group of genes from a reference cell that is kept in normal conditions (condition B). Techniques for grouping things together have helped us learn about gene function, gene control, biological processes, and different types of cells. Genes that have similar expression patterns (called co-expressed genes) can be grouped with genes that do similar things in cells.

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] found that figuring out which genes are normal or not is important for clinical research and evaluation. A new system for studying gene data was created and built in this work. To do this, first bio-information from gene expression data was looked at by setting up a heuristic search [BPPD] to look at biological processes. The BPPD method found the biological process on physiological data by using a heuristic search technique in rough set theory to look at gene-expression data. This method took gene expression data and pulled out the cellular process. The suggested method found the cellular process using a predictive search program and was split into two stages. Starting up was the first step, and making changes over and over again was the second. In these two steps, physiological data on gene expression datasets are used to figure out the biological process of each gene and the choice of genes for a dataset. We do tests to see how well heuristic search-based analysis of physiological data works with standard benchmark gene expression data sets from research sources like the Broad Institute in terms of the size of the gene expression datasets. Finally, the suggested Bi-clustered Ant Optimized Feature Relational Sequencing (BAOFRS) method was used to solve problems that came up when trying to extract bi-cluster-based gene expression information. The similarities between the sequences were found by the traits that were used to find the related sequences. When the BAOFRS method used the K-mers related

understanding process to find the traits of the relationships. The Jaccard similarity measure was used to find the value of similarity on relationship traits.

A. Balasubramanian et al.,[2] suggested a fuzzy logic-based preparation method to get rid of unnecessary data and group genes that are similar from a lot of microarray data. The suggested Parallel Island Model GA is put into action for the process of choosing gene features. The multiobjective genetic algorithm is used to execute our proposed feature selection method. A different

operator called the multi objective operator is used for this. To find the Pareto best methods for sorting, the multi-objective aspect is used. The island model was suggested because the search space is big and needs a lot of different things. Finally, the Fuzzy Based Parallel Island Model GA has been put into action with the Open MP parallelization tool.The simultaneous form of the SVM classifier is used to figure out how fit each gene group is, and this FPIMMOGA is used to move them forward. We used the fuzzy preparation method to cut down on the amount of data we had to work with and put our FPIMMOGA into action using Open MP parallel programming.In a short time, the best features are chosen. A duplicate form of the SVM Classifier is used to rate the best gene groups that have been found. This method is better at classifying things than other ways. The island model is used in this method to make the best people. Because the various islands are built at the same time, the action time for choosing the best feature has been cut down by a large amount. Kent Ridge Biomedical Data Set Repository is where the normal microarray breast cancer data sets used in this work come from.

Bennet et al. [3] came up with this method. It involves searching through the space of genes and judging the quality of each gene group by estimating the accuracy percentage of the classifier that will be used. The classifier is then trained only with the genes that were found. According to the claim, this method gets more accurate predictions than the previous method. One problem with this method is that it is more likely to overfit than filter techniques and takes a lot of computing power. Instead, it uses the relationship between gene selection and classification models, which makes it different from other models that are already out there. The use of gene expression data to classify cancer is an interesting area of study in the field of data mining. In this study, a mixed gene selection method that blends SVM-RFE and BBF is suggested as a way to choose genes. Using the leukemia dataset as an example, it was found that SVM-RFE and BBF paired with SVM worked better for classification than other similar works in terms of gene selection and classification. SVM-RFE sorts the genes into groups, and BBF gets rid of duplicates in the top genes. Also, a number of gene selection methods were tested against a range of algorithms. This method can be very helpful for correctly classifying cancer, which would get rid of the need for physical and clinical methods of detection.

Nagpal et al. [4] suggested a way to group genes into groups, which is a very important and difficult task. A lot of experts have already worked in this area, so they planned a lot of algorithms for data mining, such as the genetic algorithm, decision tree methods, and linear discrimination analysis. Most of the suggested ways to classify cancer are in the area of data mining or soft computing. Such as fuzzy logic analysis, nearest neighbor analysis, and back propagation network analysis. Most of the time, methods work well for binary-class problems but not so well for multi-class problems. Most experts were only interested in how well the sorting worked. Another problem is that the suggested gene detectors are very expensive to run, so not everyone can afford them. Correctly putting cancers into groups based on microarray gene patterns is very important for

to help the doctor choose the right medicine. Using DNA microarrays to get gene expression data has helped scientists learn more about tumors and connect expression patterns with clinical outcomes for patients with different stages and types of diseases. One of the ways for choosing features or attributes is EPSO, which stands for "Elitism Particle Swarm Optimization." Feature selection is a key method for finding genes that are useful in microarray datasets.existing methodologiesCancer 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.

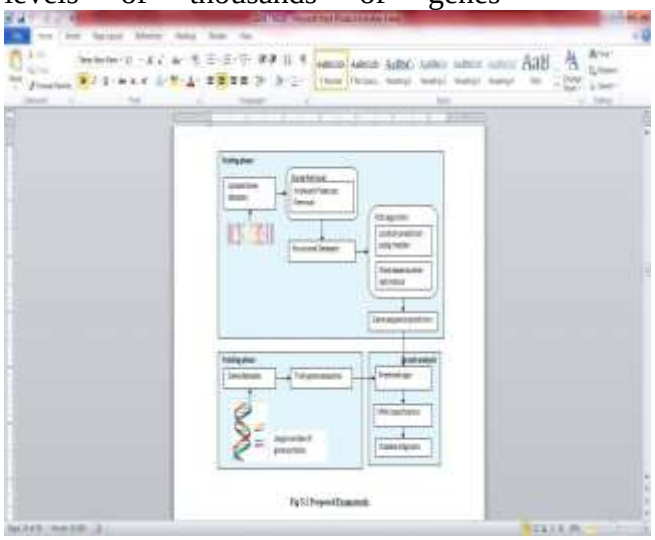


Fig 2: Proposed Framework

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.

ISSN: 1832-5505

Applied GIS

Vol-10 Issue-02 April 2022

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.

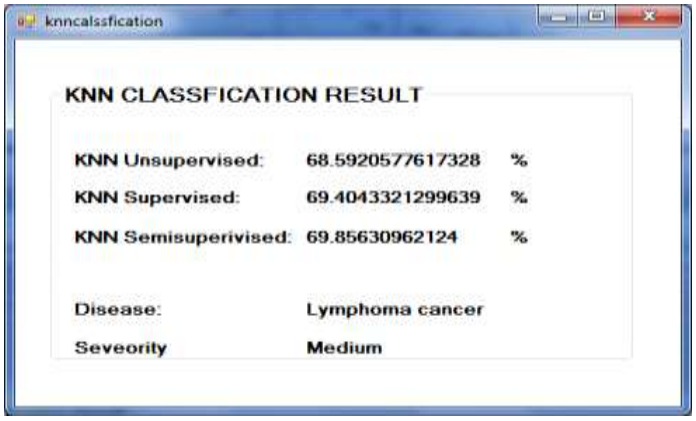


Fig 3 Accuracy rate

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

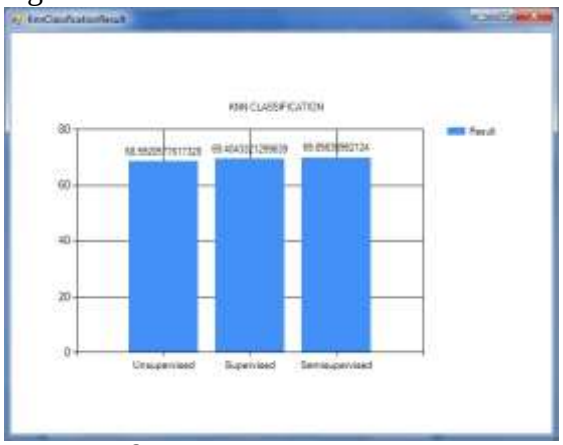


Fig 4 Performance Chart

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

Conclusion

A microarray is a useful tool for sorting different types of cancer at the molecular level. It keeps an eye on the amounts of expression of many genes at the same time. To find genes that can help identify different types of diseased tissues, we need to use the right statistical and machine learning methods on the large amounts of expression data that come from microarray experiments. We have come

up with a new way to choose genes that uses PSO methods and KNN classification to get very good classification results. The method was created to take into account how important it is to rank and choose genes before classifying them, which makes the classifier's predictions more accurate. The project was mostly about getting good results with a small number of gene groups that would help doctors guess what kind of cancer someone has. Because the same predictor is used for both gene selection and classification, the model is stronger. This was shown by results from different disease datasets. Then give an amount of seriousness for each disease that has been identified. As part of future work, the original gene set will be split into separate groups or clusters. This will make sure that the genes within a cluster are strongly linked to the sample categories. In order to make disease forecast more accurate, we can add more work to use different classification methods.

- **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.