International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-11, September 2019

Gene Based Disease Prediction using Pattern Similarity Based Classification

P.Venkateswari, P.Umamaheswari, K.Rajesh, J.Glory Thephoral

Abstract— In today's biology research in a single experiment scientist can simultaneously measure the expression of levels of thousands of genes. The molecular level of the cell is represented in gene expression profile. And it helps for medical diagnosis tools. For addressing the fundamental harms which helps to diagnosis and discovery gene expression data along with diseases classification is included. Monitoring of large number of gene expressions is possible because of this DNAmicroarray technique. Using this large quantity of gene data, experts are trying to find the probability of disease classification using gene expression dataset. Number of technique has been formed with comfortable results over these years. Still there are issues which need to be address and understood. To overcome this disease classification difficulty, it is required to review at the problem, the related issues and proposed solutions together. This paper presents a comprehensive clustering method and classification method such as Partial Swarm Optimization algorithm, K-NN classification algorithm and estimate them based on theirclassification accuracy,evaluation time and to reveal biological information about genes. Based on our multiclass classification method to diagnosis the diseases and also find severity levels of diseases. Our experimental results show that proposed semi supervised classifier performance improved in accuracy percentage.

Keywords— Microarray, Disease Classification, PSO algorithm

I. INTRODUCTION

Medical era has developed with a terrible level of achievements in disease pattern prediction, prevention and cure with the advancements of data mining techniques. Among various mining techniques feature selection occupies an indispensable role for the sake of improving accuracy in any kind of forecasting or cure of diseases. Thus it is treated as an essential earlier work of any kind of mining techniques. The familiar mining techniques namely, Support vector machine, Decision Tree, Random Forest, Navie Baysian, K-NN or Adaboost showed an improved level of accuracy with an earlier work called feature selection[1]. Succeeded by mining concepts machine learning shifted medical stage to the next level to develop diagnostic tools for early warning of disease possibility. Especially in the area of chronic disease possibility identification helps human life to take precaution action towards disease occurrence. Thus it serves as an active alert mode for human lives. This machine learning tools show increased performance efficiency by using dimensionality reduction where as feature selection should be carried out for accuracy improvement. Moreover, classification supports in establishing robust, well-informed and computerized diagnostic tool in identification of possibility of chronic type of diseases [2]. Even though classification award promising results, its computational efficiency can further be expanded by parallel classification systems. Later advancements root by keeping up patience's electronic medical data in the form of e-records. Clustering administers good level of feedback in disease pattern prediction and applied on non-linear health records [3]. It equally treats the history of health record as well as hierarchy of disease nature. The disease forecasting next initiated travel with polygene diseases like obesity and lung and prostate cancer and to discover blending several similarities of genes and diseases. Thus disease gene may be extracted from molecular level by test gene set exist in DisGeNET database. A score calculation is accomplished for genes not only depend on the topological similarity but also on functional and it is derived as a gene gravity algorithm [4]. With the text of documents of patience as source, association of gene and disease may be derived by adopting cosine similarity in mathematical vector [5]. Thus gene based prevision of disease likeness or possibility boost up the various stages of medical fields. In our proposed work, Microarray technology is highlighted as a success factor for genes based prevision of disease. Microarray technology is a hybridization-based method that can measure the expression levels thousands of genes in a single experiment.Clustering is used toanalysis tool in this context for identifying co-expressed gene networks [6]. Fuzzy c-means algorithm is also studied in addition to soft computing-based approaches. Support Vector machine (SVM) is used in supervised learning which improves the clustering technique. A Fragment of data points picked from distinct clusters based on their proximity to the corresponding centers, used for training SVM. Two Standard microarray data sets have been used in proposed approaches for effectiveness.

Manuscript published on 30 September 2019.
***Correspondence Author(s)
P.Venkateswari, Asst. Professor, Department of CSE, SRC SASTRA Deemed to be University, Kumbaonam, Tamil Nadu, India.
(E-mail: venkat_eswar2000@src.sastra.edu)
P.Umamaheswari, Asst. Professor, Department of CSE, SRC SASTRA Deemed to be University, Kumbaonam, Tamil Nadu, India
(E-mail: pum@it.sastra.edu)
K.Rajesh, Asst. Professor, Department of CSE, SRC SASTRA Deemed to be University, Kumbaonam, Tamil Nadu, India.
(E-mail: rajesh@src.sastra.edu)
J.Glory Thephoral, Asst. Professor, Department of CSE, SRC SASTRA Deemed to be University, Kumbaonam, Tamil Nadu, India.
(E-mail: gloryjs81@src.sastra.edu)

Soft computing is a union of methodologies that work concurrently and provides, in one form or another, and ability for information processing and for handling real life ambivalent situations. Evolutionary algorithms, fuzzy set comprising simulated annealing genetic algorithms, differential evolution, etc., and artificialneural networks, are the soft computing components. Major approaches in molecular biology are combined with approaches in genomic technologies in past years, which led to rise in biological information given by the scientific community. Computational methods have been used by Bioinformatics and computational biology to handle biological data. Development of new algorithms and modules is one of the important fields for new useful information from biological data. Structurally andfunctionally is performed for analysis of the macromolecules.DNA microarrays technique has been developed for gathering substantial amount of gene expression in recent years.The performance of three soft computing techniques compares the genetic algorithm simulated annealing and differentialevolution for fuzzy clustering of micro array gene expression data. Differential evolution based fuzzy clusteringconstantly performs better thansimulatedannealing based fuzzy technique and genetic algorithm is resulted on the different gene expression datasets. For performance improvement in clusteringtechnique further, RBF kernel is used along with SVM classifier.

The feature selection is a mechanism in data mining, also known as variable subset selection, variable selection, attribute selection, is the method of selecting a subset of relevant features (predictors, variables) for use in construction of model [7]. In the current emerging world, the amount of data needed to analyze a problem was grown in bioinformatics. The FSS algorithm plays a vital role in the area of bioinformatics by performing with great accuracy thereby helping to solve the problem taken. And also it favors the process of presenting the applications of the technique used to explain the feature in bioinformatics and coming up with the appropriate method to make use of the applications. This can be achieved by first selecting the feature of the problem given and then filtering the unwanted things, after that process it by giving small sample sizes as the input and note down how it works on it. If it works well on the small size samples then we have to precede the task further with large size sample on the same pattern. After being successful in both of them, we have to compare the result obtained from the current feature selection mechanism and existing univariate filtering mechanism widely used in bioinformatics and finalize the result with proper experimental result. And also provides the currently emerging areas in bioinformatics where feature selection can be applied and is able to achieve better performance. Moreover, 'Feature Subset Selection (FSS)' which is one of the major problems in classifier algorithms popularly used in data mining, which helps to provide much understandable result without changing the performance of it. Usually standard sophisticated algorithms are used to handle this but they are very complex so it is better replace to replace it with the FSS by applying it on a high dimensional datasets i.e. a large number of attributes are taken under testing it. To achieve this stochastic algorithm based on GRASP meta-heuristic [8] is introduced, which might be very helpful to minimize the number of wrapper evaluations there by speeding up the FSS mechanism. In general GRASP is a method consisting of 2 stages in total: 1, coming up with a result to the given problem and 2, improving the found result by iteratively processing it. Here main goal is to make the FSS mechanism work faster even on high dimensional datasets, so limited set of attributes are used as input which isselected in a pseudo-random way at each iteration. For instance, X tends to the attributes used and Y tends to the correct output for the problem, we are going to identify a classifier C, where C: X1*X2*.....*Xn→C. Here C tells about how the FFS mechanism will perform with high accuracy when the training datasets are met with an unseen but possible worst situation in a reasonable way. In the 2nd stage, while improving the result obtained in the 1st stage of the GRASP method, focus is not only on how to improvise the result obtained previously but also need to look over the perfectness of the previously obtained non-dominating result by applying FSS on it. Finally the performance measure of the FSS mechanism is calculated using the GRASP method with the deterministic algorithms performance and proved that FSS mechanism also a kind which works well.

In Classification, feature selection mechanism is important. The twofold is for classifier's generalization ability and other is to reduce complexity. The reason is visible, high- dimensional feature vector have high data acquisition and also high cost. Low-dimensional feature vector have huge risk overfitting. Feature Selection determines the subset of features and represents the dimensionality reduction problem and a good model is built for classification. A novel wrapper a technique [9] for feature selection is introduced using support vector machines and kernel functions. Support Vector Machine (SPM) is a best approach for classification that overcomes the absence of local minima, generalization and representation that required few parameters. The binary classification of SPM is used in this paper. Sequential backward elimination is used here. First filtering and wrapping is done to know the validation error measures have an ability to bring good dataset and the over.On each iteration avoiding overfitting is doing a random split of data sets. But the split may also remove important features and this affects the negatively classifier algorithm. To avoid this run the algorithm 3 to 4 times. Remove the features that a have been removed more than one run. This method can be used with any kernel functions, and it is generalized with various SVM. This method relays on backward feature eliminations which is treatable but it requires long input and high cost. To improve the performance filter method for feature selection should be done first the wrapping should be performed next. Micro array technologies have a high impact for finding tumor cells. There are many mechanisms are there to find the gene expression data. Gene expression data will help us to understand the genes at molecular level.

Retrieval Number: K25240981119/19©BEIESP
DOI: 10.35940/ijitee.K2524.0981119
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

For analyzing the gene data set many mathematical tools are required. And these tools such have to handle huge amount of data and the complexity should be reduced.The clustering method is one of the effective methods to findto identify the genes with similar pattern. Many mechanisms are there to cluster the gene expression data. Some of them are Hierarchical
clustering (HC), affinity propagation (AP), self-organizingmaps (SOM), and non-negative matrix factorization (NMF). Each of these techniques has their own pros and cons. In HC the temporal expression patterns are analyzed and lymphoma is predicted and breast cancer is portrayed in molecular form. But the problem in this method is tree structure data is imposed and it is highly sensitive to use. Similarly SOM helps us to identify the sub types of leukemia, but this method is infirm. All these methods have a major disadvantage, i.e. the sample can be grouped under one class. They may not be identical to the case some times. To solve that group of original sample is formed as a "metasample" using Penalized Matrix Decomposition (PMD) method [10]. Here small no. of metasample is extracted. The Extracted metasample can capture the structure of sample which belong to the same class. The samples are mapped automatically to the metasample that are extracted. The PMD uses molecular discovery pattern and it detects compound tumor cells which the conventional method cannot find. The PMD can find samples with complex classes. The experimental result shows that PMD is better than those conventional methods.

II. OPTIMIZATION AND CLASSIFICATION ALGORITHM RESULTS

Now days, Cancer prediction takes the vital role for the development in the Medical Field. Tumor types were prediction accurately as much as possible thereby has huge value in identifying efficient minimization and treatment of toxicity for the patients. Statistical and machine learning area's popular and unique classification methods have been put on to cancer classification, but there is a chance of risk because of few issues that may cause a nontrivial task. The gene expression data have unique nature from the data, used by the existing methods previously. The unique nature include very high dimensionality but generally possess thousands to tens of thousands of genes, Data size is limited when it was available in public, all below 100 and most of the genes are not similar to cancer distinction. These factors ensure that the predefined classification techniques were not designed to process this variety of data efficiently and productively. Doing Gene Selection Prior, to classify Cancer was the idea followed by some researchers, because it assists to minimize data size thereby reducing the running time of the process. Existing system follows a method where classification accuracy, computation time and ability of various cancer classification methods are given as an overview and evaluate them to acknowledge biologically meaningful gene information. Various gene selection methods which provide an essential preprocessing step for cancer classification are evaluated and applied on the research. In order to acquire a complete content of cancer classification, we need to concentrate on various issues related to cancer classification process, which includesthe gene contamination problem, the biological significance vs. statistical significance of a cancer classifierand the asymmetrical classification errors for cancer classifiers.

The simultaneous study of genes that comprises a large part of the genome was approved by Microarray technology as a unique and modern way of biological research. Classification methods and gene selection techniques are being evaluated for better utility of classification algorithm in the case of microarray gene expression data for rapid development of DNA Micro array technology. Microarrays are able to obtain the expression levels of thousands of genes at the same time. Samples are classified into categories such that it is considered as one of the major applications of gene expression data. Individual patient's clinical management decision e.g. in oncology is simple when classification methods combine with this. When the number of variables p (genes) exceeds the number of samples n, then the Standard statistic methodology in prediction or classification may fail to work especially in the case of gene microarray expression data. The main goal of this proposed system was to use supervised learning to predict and classify diseases, using the gene expressions obtained from microarrays. According to the gene patterns, known set of data was selected and utilized in order to train the machine learning protocols which is used to categorize diseases. The result of this study states the information regarding the efficiency of KNN method which is one of the machine learning techniques. The type of kernel function that is used in the process, decides the efficiency of classification. Various kernel functions performance was analyzed here. Finally prediction of various diseases with their severity levels was done.

A. ROBUST MODEL-BASED LEARNING (RMBL) VIA SPATIAL-PSO ALGORITHM

A microarray database is a repository containing microarray gene expression data.The data sets will be unstructured. Then implement preprocessing steps to eliminate the irrelevant symbols and it is converted to structured datasets. The filtered dataset with amino acid of a normal dataset is specified in figure 1.In PSO algorithm, can analyze coverage of the data before clustering begins. And propose an algorithm [Fig.2], which modifies the nearest centroid sorting and the transfer algorithm, of the spatial medians clustering. Matched with gene code, gene patterns are derived.Reduce dimensionality to cluster the genes with labels.

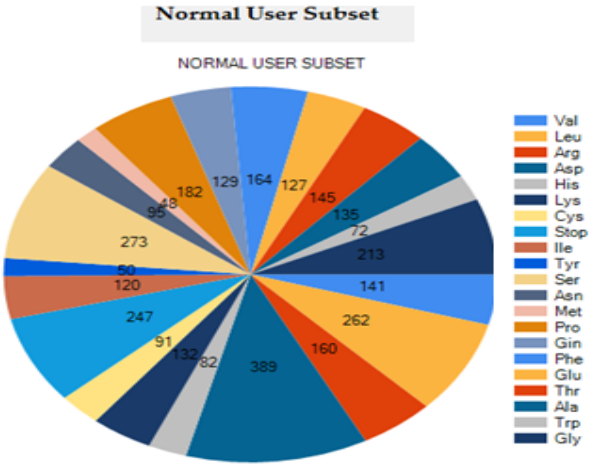

Figure 1 Normal User Subset [Training Phase DataSet]

B. SEVERITY ANALYSIS

During the classification of disease, the top five amino acid counts are extracted separately.Based on that Amino acid count, the severity range[Tab. 1] of the disease is identified. The difference in amino acid range of normal data set and tested data is represented in the form of graph in figure3. For every disease, there exists a default threshold value which plays a role in the process of severity analysis. If the amino acid

Disease name	Severity level		
	Low	Medium	High
Lymphoma cancer	38	50	>50
Blood cancer	40	68	>68
Lung cancer	28	78	>78
Eye disease	18	35	>75
Liver disease	30	55	>55

Table 1: Diseases Severity Level

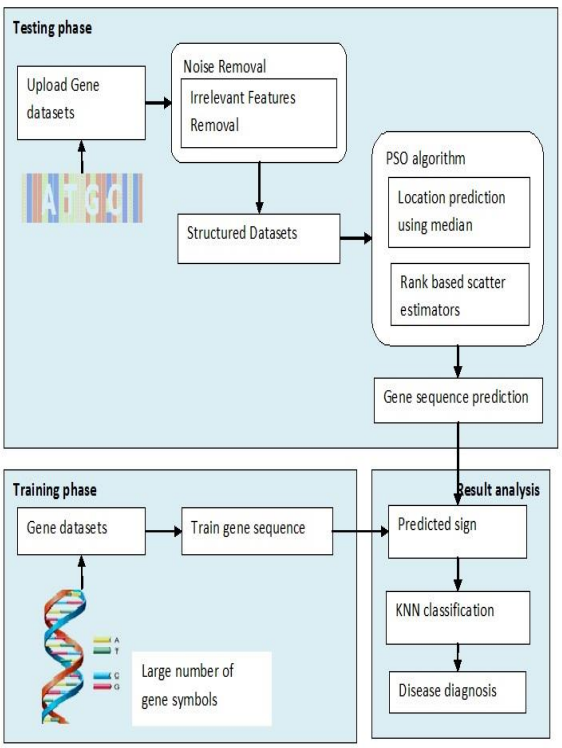

Figure 2 Architecture of RMBL via Spatial-PSO Algorithm

count range crosses the threshold value then the severity was considered as normal. The severity was provided by means of percentage and for every disease the system provides the prescription to the patients.

C. EVALUATION CRITERIA

Evaluation criteria can provide the performance measure of the system, based on the accuracy of each module. It calculates the time taken by each module in separate, based on the time required by the mechanism / algorithm involved in the module. The calculated time taken by each module of this system was compared with the time complexity of the existing disease prediction system and provides the result if it is improved here or not. Based on that, the accuracy of using this algorithm for disease prediction was detailed [Figure 4] with the percentage on the improvement in accuracy.

III. CONCLUSION

For cancer classification at the molecular level, Microarray can be used as an important tool. Monitoring of expression levels of large number of genes is done parallelly. 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. Microarray data's important application is to classify biological samples or predict clinical or other outcomes. For multivariate analysis of high throughput as say data such as gene expression data Dimension reduction is a necessary part. Dimension reduction methods are mostly used but their relative performance has not been well studied. Comparing the performance of dimension reduction methods based on results

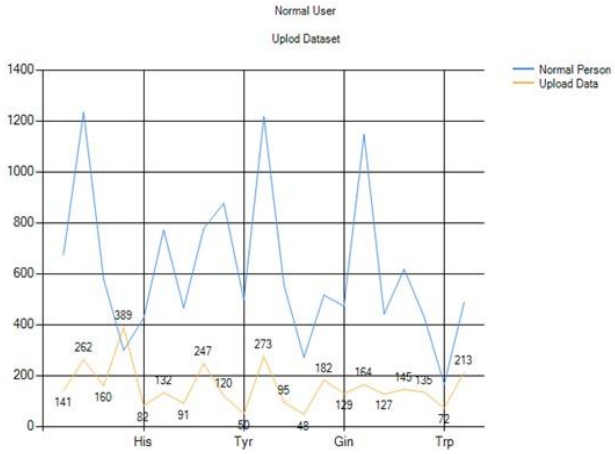

Figure 3 Comparison Graph Tested DataSet vs Normal DataSet

of published studies would be difficult due to differences among the studies in data sets, data preprocessing, and methods of gene selection, model selection and validation. In this project, a hybrid gene selection method has been proposed, which incorporates dimensionality reduction methods to achieve the subsets with irrelevant features.

In this thesis, a hybrid gene selection method has been proposed, which merges a PSO method and KNN classification to achieve high classification performance. The method was designed to address the importance of gene ranking and selection prior to classification, improving the prediction strength of the classifier. The main focus of project is to provide accuracy results with few number of gene subsets, enabling the doctors to predict the type of cancer. Importance of the same classifier used for both the gene selection and classification improves the strength of the model, this can be known from results of various diseases. Then provide severity level for each classified disease.

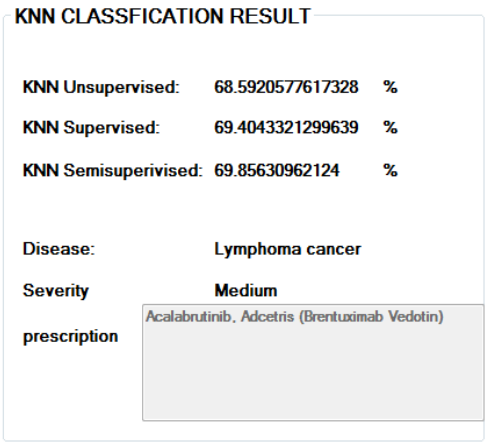

Figure 4 Result Accuracy Level.

REFERENCES

1. Tejeswinee. K, Shomona Gracia Jacob, Athilakshmi. R,"Feature Selection Techniques for Prediction of Neuro-Degenerative Disorders: A Case-Study with Alzheimer's And Parkinson's Disease ", Procedia Computer Science 115 (2017) 188–194
2. Divya Jain , Vijendra Singh," Feature selection and classification systems for chronic disease prediction: A review", Egyptian Informatics Journal 19 (2018) 179–189
3. Mohan Kumar K N, S.Sampath, Mohammed Imran," An Overview on Disease Prediction for Preventive Care of Health Deterioration", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958,Volume-8, Issue-5S, May 2019
4. Limei Lin, Tinghong Yang, Ling Fang, Jian Yang, Fan Yang and Jing Zhao," Gene gravity-like algorithm for disease gene prediction based on phenotype-specific network", Lin et al. BMC Systems Biology (2017) 11:121; DOI 10.1186/s12918-017-0519-9
5. Jie Zhou and Bo-quan Fu," The research on gene-disease association based on text-mining of PubMed", Zhou and Fu BMC Bioinformatics (2018) 19:37; https://doi.org/10.1186/s12859-018-2048-y
6. UjjwalMaulik," Analysis of gene microarray data in a soft computing framework", Applied Soft Computing 11 (2011) 4152–4160
7. Americo Pereira, Jan Otto,"Review of Feature SelectionTechniques in Bioinformatics", Bioinformatics, Volume 23, Issue 19, 1 October 2007
8. Pablo Bermejo, Jose A. Gámez, Jose M. Puerta, "A GRASP algorithm for fast hybrid (filter-wrapper) feature subset selection in high-dimensional datasets", Pattern Recognition Letters 32 (2011) 701–711
9. Sebastián Maldonado, Richard Weber, "A wrapper method for feature selection using Support Vector Machines", Information Sciences 179 (2009) 2208–2217
10. Chun-Hou Zheng, Lei Zhang, Vincent To-Yee Ng, Simon Chi-Keung Shiu, and De-ShuangHuang,"Molecular Pattern Discovery Based on Penalized Matrix Decomposition", IEEE Transactions on Computational Biology and Bioinformatics, Vol.8, No.6, November /December2011