IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-8, Aug 2019 E-ISSN: 2347-2693

Pattern Similarity Based Classification Using K-Nearest Neighbor and PSO Model for Cancer Prediction with Genetic Data

T. Sneka[1*], K. Palanivel[2]

[1,2]Department of Computer Science, A.V.C. College (Autonomous) Mayiladuthurai, Tamil Nadu, India

[*]*Corresponding Author: snekathogai1995@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v7i8.2731 | Available online at: www.ijcseonline.org

Accepted: 13/Aug/2019, Published: 31/Aug/2019

***Abstract*—** Data mining techniques can be used by Health organizations to predict different types of Cancer disease using individual Gene expression data. By using DNA (Deoxyribo Nucleic Acid) Microarray technology, thousands of genes can be articulated simultaneously. The objective of this research is to look closer on the classification issues in handling microarray data by introducing Semi-Supervised KNN (K-Nearest Neighbor) algorithm and Particle Swarm Optimization (PSO) as feature selection to cluster large amount of genetic microarray data. Also, using the predicted type of cancer, the severity level of cancer is diagnosed. Classifier performance is evaluated and it is shown in pie-chart and graph with improved accuracy. The proposed Semi-supervised learning method provides 10% improved accuracy in predicting cancer than the existing Supervised and unsupervised learning methods.

***Keywords*—** Medical Data Mining, Cancer Prediction, Gene sequence, Clustering, Classification.

I. INTRODUCTION

Microarray innovation has turned out to be one of the key device that numerous scholars use to screen genome wide articulation levels of qualities in a given creature especially individual person genetic data. A microarray is the ordered arrangement of DNA molecules in a glass like slide where each location of DNA called as spots or features. Data mining techniques is used in the genome research by biologists to predict the diseases based on given microarray data. The primary goal of this paper is to experiment the data that is collected in the form of DNA microarray and to predict the class on which the uploaded data status lies. The important contributions of this paper are:

- To extract helpful classified accuracy for predicting the type of cancer and its severity.
- Comparison of variety of existing results in the supervised and unsupervised Classification with the proposed semi-supervised KNN Classification result.
- Identify the best overall performance algorithm for prediction of cancer diseases.

The figure 1 shows the gene symbol structure.

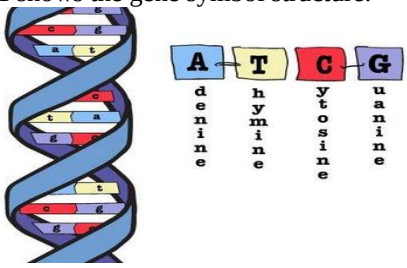

Figure 1. Gene symbol

In this paper, microarray dataset were utilized for clustering and classification technique. The steps include collection of dataset essential for proposed work such as accuracy calculation, classification and the comparison of results. The dataset has been used to classify the diseases based on the ACTG DNA combination i.e., adenine-cytosine-thymine-guanine pairs. Various ACTG combinations are calculated based on their protein values compared to genetic code table .The proposed semi-supervised KNN classification carried out based on these ACTG combinations of genes. The proposed work follows as section II discuss the related work with respect to disease prediction, section III state the some fundamental concepts of clustering and classification algorithm, section IV describes the experimental results of the classification algorithm for DNA microarray pattern. Finally, section V the conclusion of this research work.

II. RELATED WORK

Booma P. M. and Prabhakaran S., identified normal or abnormal genes for identifying various types of disease using data mining techniques, such as Heuristic Approach for analyzing biological changes on genes based on physical and logical pattern which is called Biological process of Physiological data (BPPD).Also uses heuristic search algorithm to identify genes which undergoes the biological process. This work uses Divisive algorithm, Hierarchical Clustering algorithm for identifying similar genes. Cancer datasets were used for measuring classification accuracy. Bi-Clustered Anti Optimized Feature Relational Sequencing Method (BAOFRS) were used and genes pattern was

identified. Jaccard similarity coefficient was applied to identify the similarity value on relational features [1]. Natarajan A. and Balasubramanian R., proposed Fuzzy Parallel Island Model Multi-Objective Genetic Algorithm (FPIMMOGA) for identifying the best features in genetic data. SVM (Support Vector Machine)classifier is used to evaluate the result. Island model is used for generating the best population. Multiple Islands were executed in parallel to reduce the execution time. In this work standard microarray breast cancer data sets are taken from Kent Ridge Biomedical Data Set Repository [2]. Bennet, et al.,proposed an ensemble feature selection technique which is a combination of Recursive Feature Elimination (RFE) and Based Bayes error Filter (BBF) for gene selection and Support Vector Machine (SVM) algorithm for classification. Leukemia dataset were used for SVM classification. This approach can play a vital role in accurate cancer classification thus, eliminating the morphological and clinical means of diagnosis. Time complexity is the limitation to be considered in this work [3].Nagpal Rashmi and Shrivas Rashmi address the problem in selecting genes with correlation techniques. A new method of gene selection utilizing Elitism Particle Swarm Optimization (EPSO) based on Recursive Feature Reduction (RFR).Cancer Classification is the main goal of this work. EPSO is used for selection of small informative genes from huge data sets. This work uses breast cancer datasets for both feature selection and classification accuracy. 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. This work is helpful in mining High-dimensional data [4].Thangaraju P. and Mehala R. Provided a study of various technical and review papers on lung, liver and Breast cancer data sets and explores that data mining techniques offer great promise to uncover patterns hidden in the data that can help the clinicians in decision making. This work also addresses the challenges in predicting the existence of liver cancer at the early stage. Also seven classification algorithms with respect to time and accuracy are experimented with the help of dataset. From the above study it is observed that the accuracy for the diagnosis analysis of various applied Data mining Classification techniques. Implementation of the techniques is highly acceptable and can help the medical professionals in decision making for early diagnosis and to avoid biopsy. In the case of the data sets used it is observed that the accuracy reached up to 100% when there is more number of attributes and the accuracy is decreasing as the number of attributes decreased[5].

III. METHODOLOGY

This paper has analyzed three types of classification approaches namely supervised, unsupervised and semi-supervised KNN Classification to predict which classification is more accurate in predicting the cancer diseases based on DNA microarray pattern. The dataset has to be uploaded in the testing phase to extract the features. Best features are referenced with the Normal DNA (Protein) values. The steps in the proposed work always compared the normal persons data and the uploaded data i.e., individual who is free from cancer is compared with affected DNA pattern. The goal of this proposed work will predict cancer using semi-supervised learning methodology and to classify the cancer severity level based on the DNA microarray pattern dataset. The outcome of this study will provide information regarding the efficiency of the machine learning techniques, in particular a KNN (K Nearest Neighbor) 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 cancer with severity levels and predict various types of cancer. Figure 2 shows proposed framework.

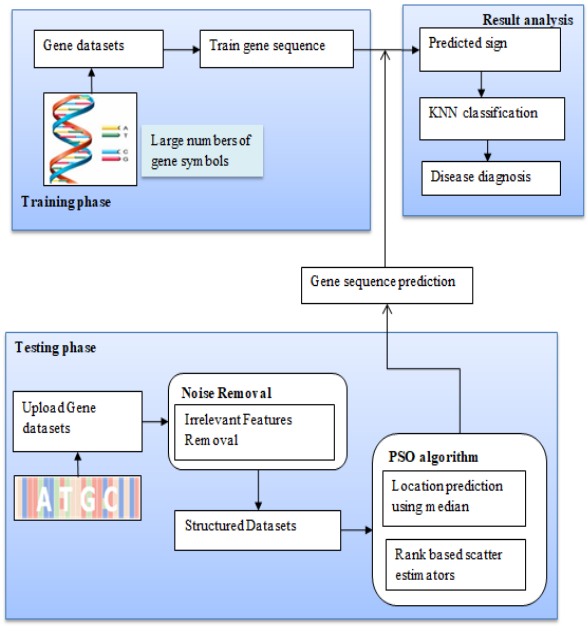

Figure 2. Proposed Framework

PSO (Particle Swarm Optimization)

Particle Swarm Optimization 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 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.

Cancer Prediction

Factual approval is important before models can be utilized, particularly in clinical settings. K nearest neighbor calculation is utilized to order the different kinds of illnesses from quality articulation of DNA data. Order is finished with the assistance 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. When the classes are non-separable, the optimal hyper plane is the one that minimizes the probability of classification error. 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. The KNN demonstrate various attractive features such as good generalization ability compared to other classifiers. The algorithm steps that generally compare the Normal and uploaded data are given as follows:

for all the unknown samples UnSample(i)
for all the known samples Sample(j)
compute the distance between
Unsample(i) and Sample(j)
end for
find the k smallest distances
locate the corresponding samples
Sample (j1),….,Sample(jK)
Assign UnSample(i) to the class which appears more frequently
end for

Also the proposed algorithm steps are:

Input: *Gene dataset*
Output: *Classification of data set based on type of cancer disease predicted otherwise as normal also Severity of the cancer*
Step 1: Input the data set
Step 2: Apply pre-processing techniques
Step 3: Cluster the features based on values obtained after applying PSO as feature selection
Step 4: Discard redundant features
Step 5: Apply semi-supervised KNN on predominant features
Step 6: Measure the performance of the KNN+PSO model

KNN performance based on semi-supervised learning is evaluated based on the results taken from the related work done on cancer prediction.

Severity Analysis

Severity level of diseases is classified using classified data count. A threshold value is set to differentiate the severity level as high, medium and normal. Prescription is given to the patients according to the diseases condition and Severity.

IV. EXPERIMENTAL RESULTS

Implementation of the system is done by uploading the cancer datasets from NCBI Repository from this link https://www.ncbi.nlm.nih.gov/genbank/. Gene clustering and Classification performed 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 given in figure 3.

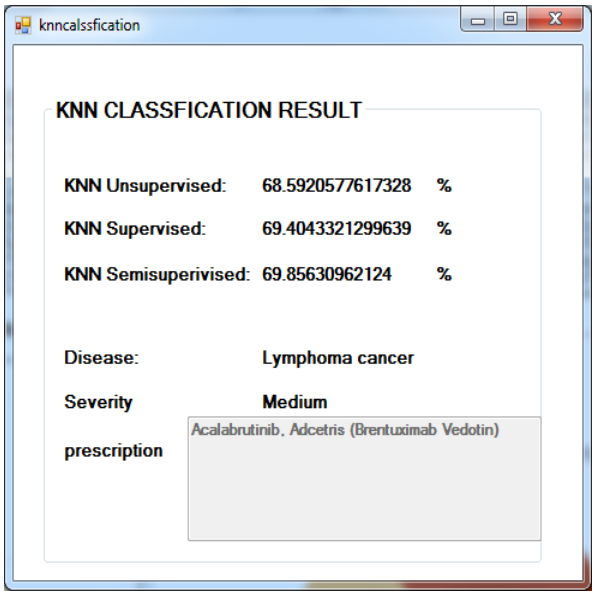

Figure 3. Accuracy rate

Accuracy rate is calculated as

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} *100$$

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

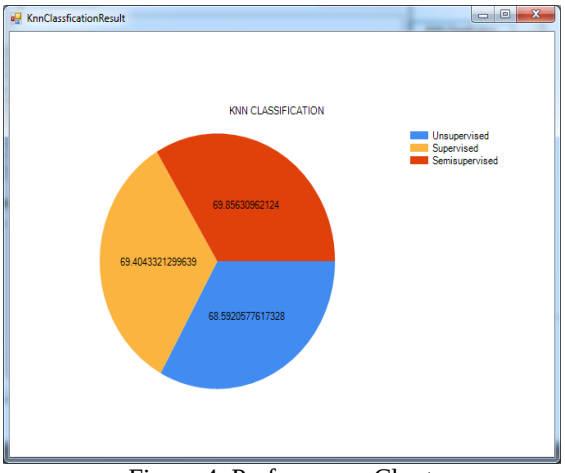

Figure 4. Performance Chart

The performance result is shown as pie-chart in figure 4 demonstrates that the proposed KNN semi-supervised algorithm provides 10% improved accuracy than the existing algorithms.

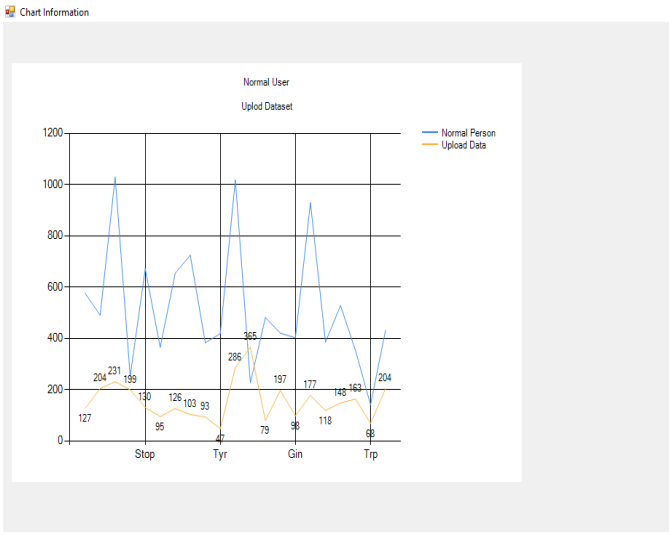

Figure 5. Graph represents variation between Normal person and uploaded Data

The graph shown in figure 5, explains about the training and testing gene datasets. In training stage, train the normal person genes with amino acid labels. In testing stage, upload the cancer datasets and calculate the amino acid labels and plot the graph based on varied patterns.

V. CONCLUSION AND FUTURE WORK

In this research, the accuracy of the two existing algorithms namely supervised, unsupervised KNN classification and a proposed algorithm namely semi-supervised KNN Classifier is evaluated using DNA microarray pattern of Normal and uploaded DNA dataset. This research work also focuses on clustering the best feature of DNA pattern by using PSO. Out of the three classification learning algorithms; our proposed semi-supervised KNN classification approach gives 10% improved accuracy in the cancer prediction. The proposed work focus on promising accuracy results with very few number of gene subsets enabling the doctors to predict the type of cancer and the disease severity. In future, weighted classifier is adopted to find out the best classifier for cancer prediction.

REFERENCES

[1] B. Jaison, G. Chilambuchelvan, K. Nirmal, "*A hybrid approach for gene selection and classification using support vector machine",* International Arab Journal Information Technology, Vol.**12**, issue.**6A**, pp.**695-700**, **2015**.

[2] P. M. Booma , S. Prabhakaran, "*Classification of genes for disease identification using data mining techniques"*,Journal of Theoretical and Applied Information Technology, vol.**83**, issue.**3**, pp.**399-414**, 2016**.**

[3] L. George, S. Asha ,W. Haibo , D.F. Michael , R.M. Stephen , N.C.S. Natalie , S. Elaine, R. Timothy, E.T. John, M. Anant, "*Supervised Multi-view Canonical Correlation Analysis (sMVCCA): Integrating Histologic and proteomic features for predicting Recurrent prostate cancer"*, IEEE transactions on Medical Imaging,vol.**34**,issue.**1**, pp.**284-297**,**2015**.

[4] A. Hasseeb, H. Jingyu, X. Yong, A. Russul, *" Lung Cancer Prediction from Microarray data by Gene expression programming"*, IET Systems Biology,vol.**10**,issue.**5**, pp.**168-178**, **2016**.

[5] Liu, Jin-Xing, Yong Xu, Chun-Hou Zheng, Heng Kong, Zhi-Hui Lai , "*RPCA-Based Tumor Classification Using Gene Expression Data"*,IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) ,vol.**12**,issue.**4**, pp.**964-970**, **2015**.

[6] R. Nagpal, R. Shrivas, "*Cancer Classification Using Elitism PSO Based Lezy IBK on Gene Expression Data"*, International Journal of Scientific and Technical Advancements, vol.**1**, issue.**4**, pp.**19-23**, **2015**.

[7] A. Natarajan, R. Bala Subramanian , "*A Fuzzy Parallel Island Model Multi Objective Genetic Algorithm Gene Feature Selection for Microarray Classification"*, International Journal of Applied Engineering Research, Vol.**11**, issue.**4**, pp.**2761-2770**, **2016**.

[8] H. Park, Y. Shiraishi, S. Imoto, S. Miyano, "*A novel Adaptive Penalized Logistic Regression for uncovering biomarker associated with Anti-cancer drug sensitivity"*, IEEE/ACM transactions on Computational Biology and Bioinformatics, vol.**14**, issue.**4,** pp.**771-782, 2017**.

[9] N. Songyot, "*Gene selection using interaction information for Microarray-based Cancer classification",* IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, **2016**.

[10] P. Thangaraju , R. Mehala, "*Novel Classification Based Approaches over Cancer Diseases"*, International Journal of Advanced Research in Computer and Communication Engineering, vol.**4**, issue.3, pp.294-**297**, **2015.**

[11] R. Anupriya, P. Saranya , R. Deepika , "*Mining Health Data in Multimodal Data Series for Disease Prediction*", International Journal of Scientific Research in Computer Science and Engineering ,Vol.**6**, Issue.**2**, pp. **96-99**, **2018**.

[12] Pramod Pardeshi and Ujwala Patil , "*Fuzzy Association Rule Mining- A Survey*",International Journal of Scientific Research in Computer Science and Engineering ,Vol.**5**, Issue.**6**, pp.**13-18**, **2017**.

Authors Profile

T. Sneka has completed her M.Sc. in Computer Science degree at A.V.C. College (Autonomous), Mayiladuthurai. Currently she is doing M.Phil, in Computer Science at A.V.C. College (Autonomous), Mayiladuthurai. She is doing Research in the area of 'Knowledge Discovery in Databases'.

K. Palanivel received his M.Sc.(Computer Science) degree from Bharathidasan University, M.Phil.(Computer Science) degree from Manonmaniam Sundaranar University and Ph.D.degree from Bharathidasan University.He is currently working as Associate Professor in the Department of Computer Science at AVC College (Autonomous), Mayiladuthurai. He has published many research papers in international journals. His research area includes Human Computer Interaction, Machine Learning, a Recommender systems and Data mining.

