A SURVEY ON GENE EXPRESSION CLASSIFICATION SYSTEMS

[1]Praveen Tumuluru M.Tech (Ph.D), [2]Dr. Bhramaramba Ravi, BE, MS, Ph.D
[1]Research Scholar, [2]Associate Professor,
[1]Department of Computer Science and Engineering, [2]Department of Information Technology,
[1,2]GITAM, Visakhapatnam, India.
praveenluru@gmail.com
bhramarambaravi@gmail.com

Abstract — Identification of cancer and classification for diagnostic and prognostic determinations is generally based on pathological analysis of tissue sections, resulting in particular interpretations of data. For the determination of accurate identification of cancer subtypes, several studies have been directed recently in order to identify genes which force creates cancer. However, gene classification is a new research area poses new challenges owing to its distinctiveness of problem. Different gene selection and gene classification methods are existing. Gene selection consists of exploring for gene subsets which are able to differentiate tumor tissue from normal tissue.

Conventional classification techniques are greatly dependent on the morphological appearance of tumors, parameters that are attained from clinical observations, and other biochemical techniques. Their applications are controlled by the uncertainties present and their prediction accuracy requires additional improvement. Deoxyribonucleic or DNA microarray technologies yield cancer researchers a new technique to examine the pathologies of cancer from a molecular angle right under a systematic framework, and furthermore, to extract more accurate prediction in prognosis and treatment. Some major issues related to cancer classification using microarray data are: robustness of gene selection and gene ranking, understanding of issues related to feature selection and performance evaluation of the selected genes.

Keywords: **Cancer classification, gene classification, gene selection, DNA, Microarray data.**

I. INTRODUCTION

The survey of literature is based on several techniques for the classification of microarray data into normal or abnormal. In order to ease radiologist's assessment and the classification of cancer, various techniques have been proposed by researchers for the past two decades. Early detection of cancer from microarray data depends on the feature extraction and expert analysis of data. Micro array data analysis has been efficiently employed in a number of investigations over a wide group of biological disciplines, which consists of cancer classification by class identification and prediction, detection of the unknown effects of a particular therapy, recognition of genes that are suitable for a particular diagnosis or therapy, and cancer identification [10].

Many algorithms have been developed for the recovery of data, as it is not economic and time consuming for the purpose of repeating the experiment. Cancer classification is important for the consequent diagnosis and treatment. In the absence of the correct detection of cancer types, there is little possibility to offer helpful therapies and accomplish anticipated results. Several techniques are used for the classification of cancer. Here, a brief introduction is given for various approaches used for gene selection and classification. Also the motivation, objectives, and contribution of the present research work are discussed.

II. REVIEW OF GENE CLASSIFICATION SYSTEMS

An accuracy based effective ensemble approach for gene classification [1] has been proposed. In this, feature selection models are used to mitigate the effect of over fitting. From, the biological point of view, the choice of feature selection methods that preserve the semantics of the features 'genes' and further select more informative & relevant features. The method enhances better accuracy and applies the technique to more cancer types. It has been testing the performance of the total ensemble rate, base classifiers error, Area under curve (AUC) and Bayesian credible interval (BCI) are calculated and are performed against three benchmark cancer datasets; namely leukaemia, colon and breast cancer datasets. Moreover, this can be applied to multiclass datasets.

An approach to analyse microarray datasets and classify cancer diseases [2] has been proposed, for feature selection that employs Information Gain and uses genetic Algorithm for feature reduction and finally utilizes Genetic Programming for cancer disease classification. The method improves classification accuracy of cancer classification by reducing the number of features and preventing the Genetic Algorithm (GA) from being trapped in a local optimum. It has been verified by considering seven cancer Gene Expression (GE) datasets and for each test two important measures are used for observational assessment of the performance evaluation as a number of selected genes and predictive accuracy on selected gene.

A novel hybrid evolutionary algorithm in- telligent dynamic genetic algorithm (IDGA) [3] is proposed. It mainly consists of two steps, a score-based method is used to reduce the dimensionality and provide statistical significant genes to the next. In the second, scoring methods are used, the Fisher score and the Laplacian score. Next an integer-coded genetic algorithm with variable-length genotype, adaptive parameters and modified genetic operators is proposed. The method was performed independently on each dataset using either of the filter methods separately. Three widely used classifiers, namely, Support Vector Machines (SVM), Naïve Bayes (NBY), and K-Nearest Neighbor (KNN) are used to measure the performance of IDGA. The IDGA are performed independently seven times for each dataset and the results after each run are evaluated using leave-one-out cross-validation (LOOCV) for assessing the performance.

A weighted graph based-GEG classifier for classifying microarray datasets [4] has been proposed, the main contributions of the method is ability to classify samples to the corresponding class and detect out-of-class samples, and it is effective in clinical diagnostic applications because it reduces the rate of detecting false-positives.

The Robust principal component analysis (RPCA) [5] is applied to extract a subset of genes associated with a special biological process and SVM is used to classify the tumour samples based on the extracted features. A modified method is also proposed to enhance the classification performance based on RPCA and linear discriminate analysis. Three kinds of measures are used to evaluate the performance of the method, the first measure is the leave-one-out cross-validation performance is estimated and the second one is the accuracy which measures the classification performance by using the percentage of correctly classified samples. The third one is the Area under Curve of Receiver Operating Characteristic (ROC) which is suitable for evaluating the performance of binary classification.

A novel hybrid feature selection strategy [6] which combines the Mutual information maximization (MIM) and the Adaptive genetic algorithm (AGA) and name it as MIMAGA-Selection algorithm, to eliminate the redundant samples and reduce the dimension of the gene expression data for classification and the reduced gene expression data set provided highest accuracy compared to traditional algorithms. The classification accuracy rates comparison with other existing feature selection algorithms shows the

effectiveness of the MIMAGA-Selection algorithm. The time complexity of the MIMAGA-Selection algorithm can be largely improved on cloud platforms.

A hybrid gene selection approach, namely Genetic Bee Colony algorithm [7] combines the use of a Genetic algorithm along with Artificial Bee Colony algorithm. It can be used to solve classification problems that deal with high dimensional datasets. The algorithm, adopts four modifications of the original ABC algorithm in order to combine the exploration and exploitation capabilities of the GA and the ABC algorithm. First, it pre-processed the microarray dataset using an mRMR filter method to reduce the computational time and cost of the classic algorithm. Algorithm is improved by adding a uniform crossover operation and subsequently increased the number of scout bees to improve the movement speed. The GBC considered as a general framework that can be used to solve various optimization problems.

Techniques for Mining Gene Expression Data Focusing Cancer Therapeutics: A Digest [8] reviewed conventional and advanced techniques in gene expression data analysis from cancer perspective. Previously, the complicated genetic disease is diagnosed by non-molecular characteristics like kind of tumor tissue, clinical phase and pathological characteristics. Currently, such diseases are diagnosed through microarray data expressions. Various algorithms such as hierarchical clustering, k-means clustering, principal components analysis, and discriminate analysis and classification tree based gene expression data analysis was employed for effective disease diagnosis.

A hybrid technique namely, designing a fuzzy expert system using microarray data classifier [9]. To find the membership function, an algorithm, Ant Bee Algorithm (ABA), was presented, combining the benefits of two optimization algorithms, Ant Colony Optimization (ACO) and Ant Bee Colony (ABC). For the detection of informative genes, Mutual Information is utilized. The accuracy of the classification was improved, as the classifier had minimum false positive rate and large discrimination power. Although the accuracy is improved, a large number of genes can increase the error rate.

III. CONCLUSION

A summary on cancer classification system using microarray data carried out by earlier researchers is presented. It is evident from the literature survey that the classification of microarray data was mainly based on statistical based features. Several methods are used for the classification of cancer. In this work, cancer classification technique uses the machine learning approaches like Genetic and Evolutionary Algorithms such as Ensemble approach, hybrid evolutionary approach, MIMAGA approach and Genetic Bee Colony Algorithm and so on are discussed. Further, advanced algorithms to be commence in order to obtain better gene selection for cancer classification and the performance of the classification is measured with the metrics such as detection rate, false alarm rate (FAR) and accuracy.

REFERENCES

[1] Sara Tarek, Reda Abd Elwahab and Mahmoud Shoman, "Gene expression based cancer classification," Egyptian Informatics Journal, December 2016.

[2] Hanaa Salem, Gamal Attiya and Nawal El-Fishawy, "Classification of Human Cancer Diseases by Gene Expression Profiles," Applied Soft Computing, Vol. 50, pp.124-134, January 2017.

[3] M. Dashtban, Mohammadali Balafar, "Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts," Genomics, Vol.109, pp.91–107, 2017.

[4] Sun-Yuan Hsieh and Yu-Chun Chou, "A Faster cDNA Microarray Gene Expression Data Classifier for Diagnosing Diseases," IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol.13, No.1, pp.43 – 54, 2016.

[5] Jin-Xing Liu, Yong Xu, Chun-Hou Zheng, Heng Kong and Zhi-Hui Lai,"RPCA-based Tumor Classification Using Gene Expression Data,"IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol.12, No.4, pp. 964-970, 2015.

[6] Huijuan Lua, Junying Chena, Ke Yana, Qun Jina, Yu Xuec, Zhigang Gao, "A hybrid feature selection algorithm for gene expression data classification," Neurocomputing, pp.1–7, 2017

[7] H. M. Alshamlan, G. H. Badr and Y. A. Alohali, "Genetic Bee Colony (GBC) Algorithm: A New Gene Selection Method for Microarray Cancer Classification", Computational Biology and Chemistry, Vol. 56, pp.49–60, 2015.

[8] Shaurya Jauhari and S.A.M. Rizvi "Mining Gene Expression Data Focusing Cancer Therapeutics: A Digest", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 11, No. 3, 2014.

[9] P. G. Kumar, T. A. A. Victoire, C. Renukadevi and D. Devaraj, "Design of fuzzy expert system for microarray data classification using a novel Genetic Swarm Algorithm," Expert Systems with Applications, vol. 39, pp.1811–1821, 2012.

[10] P Tumuluru, "Dijkstra's based Identification of Lung Cancer Related Genes using PPI Networks", International Journal of Computer Applications (0975 – 8887), Vol. 163, No. 10, pp. 1-10, April 2017.

AUTHORS PROFILE

1. Praveen Tumuluru, received the M.Tech degree in Computer Science and Engineering from Koneru Lakshmaiah College of Engineering, Acharya Nagarjuna Univerisity, in 2008. He is a research Scholar in GITAM, Visakhapatnam, Andhra Pradesh, India, working in Data Mining Techniques for Bioinformatics, Cancer classification, Protein-Protein Interaction. He has been working as Assistant Professor, in the Department of Electronics & Computer Engineering, Prasad V. Potluri Siddhartha Institute of Technology, Vijayawada since 2008.

2. Dr. Bhramaramba Ravi, presently working as Associate Professor in GITAM, Visakhapatnam, Andhra Pradesh, India. She has a total of 12 years of research experience and 17 years of teaching. She received her Ph.D from Jawaharlal Nehru Technological University, in 2011 and MS degree in Software Systems from BITS, in 1999. She has published 22 papers in various National and International Journals/ Confer current research interests are in the areas of Data Mining Techniques for Bioinformatics.