International Journal for Modern Trends in Science and Technology
Volume: 03, Issue No: 08, August 2017
ISSN: 2455-3778
http://www.ijmtst.com

Genetic Algorithm Based Model for Early Detection of Cancer

R.Gayathri

M.Phil Scholar in Department of Computer Applications, Alagappa University, Karaikudi, Tamilnadu, India.

To Cite this Article
R.Gayathri , "Genetic Algorithm Based Model for Early Detection of Cancer", *International Journal for Modern Trends in Science and Technology*, Vol. 03, Issue 08, August 2017, pp.-28-31.

ABSTRACT

Cancers are a large everyday of diseases that involve uncharacteristic cell growth with the potential to spread to other parts of the body. A cancer syndrome in any of its forms represents a major cause of death universal. In cancer diagnosis, classification of different cancer types is of the greatest significance. Accuracy for prediction of various cancer types gives better handling and minimization of toxicity on patients. Consequently, creating methodologies that can effectively differentiate between cancer subtypes is essential. There are various conventional methods to predict early cancer still needsimprovements to solve the issues and challenges like accuracy, sparsity and Scalability. Hence, this paper presents a new methodology based on genetic algorithm to classify and predict Human cancer diseases tested using real-world cancer datasets. This methodology combines both genetic algorithm and neural network to classify and predict cancer earlier. The proposed system is evaluated by classifying and prediction cancer diseases in various cancer datasets and evaluation measures. The results are compared with latest methods for performance benchmark.

Keywords: *Cancer, Classification, Data mining, Prediction, Genetic Algorithm*

I. INTRODUCTION

Data mining is the process in which valued information is extracted from the large dataset. It has reached the high growth over past few years. Due to the usefulness of data mining approaches in health world, it has become the good technology in healthcare domain.

Cancer is a hypothetically final disease caused mainly by conservational issues that mutate genes encoding critical cell regulatory proteins. Resultant Many features of the modern Western diet (high fat, low fiber content) will increase cancer frequency. (S. Josh et al.,) Identifying cancer is still uncharacteristic cell behavior indications to extensive commonalities of abnormal cells that destroy neighboring ordinary tissue and can spread to vital organs resulting in disseminated disease, commonly a suggestion of pending patient death. More sensitively, globalization of unhealthy lifestyles, particularly cigarette smoking and the implementation stimulating for the doctors in the field of medicine. Even now the actual reason and complete cure of canceris not invented.

II. RELATED WORKS

Li, Eldon Y. [1]. Medical data mining can exploit the hidden patterns present in voluminous medical data which otherwise is left undiscovered. Data mining techniques which are applied to medical data include association rule mining for finding frequent patterns, prediction, classification and clustering.

The research work done in data mining medical fields given as: **Evans et. al [2]** proposed a system based on data mining techniques to detect the hereditary syndromes.

Pradhan and Prabhakaran [3] proposed an approach through association rule mining to mine high-dimensional, time series medical data for discovering high confidence patterns.

DoronShalvi and Nicholas Declares, [4] discussed medical data mining through unsupervised neural networks besides a method for data visualization. They also emphasized the need for preprocessing prior to medical data mining.

Krzysztof J. Cior [5], bioengineering professor, identified the need for data mining methods to mine medical multimedia content.

Tsumoto [6] identified problems in medical data mining. The problems include missing values, data storage with respect to temporal data and multi-valued data, different medical coding systems being used in Hospital Information Systems (HIS).

Barmier and Banshee [7] explored and analyzed two programming models such as neural networks, and linier genetic programming for medical data mining. **Abide and Hoe [8]** proposed and implemented a symbolic rule extraction workbench for generating emerging rule-sets.

Abide et al. [9] explored the usage of rule-sets as results of data mining for building rule-based expert systems.

Olukunle and Ehikioya [10] proposed an algorithm for extracting association rules from medical image data. The association rule mining discovers frequently occurring items in the given dataset.

Traditionally data mining techniques were used in various domains. However, it is introduced relatively late into the Healthcare domain.

Raja, Chital [11] et al. normally the necessary part of any human body is blood since it keeps one alive. It executes many vital functions such as to transfer oxygen, carbon dioxide, mineral and etc. to the complete body in order to keep metabolism. Blood consists of three main components which RBC, WBC and Platelets. Insufficient amount of the blood could affect the metabolism critically which could be very hazardous if early treatment is not taken. One of the normal blood disorders is Leukemia. Leukemia is the common type of cancer in children. All cancers start in body cells, and leukemia is a type of cancer that starts in blood cells.

III. System Architecture

The Figure 1 explains the architecture of the proposed Genetic Algorithm Based Model for Early Detection of Cancer system

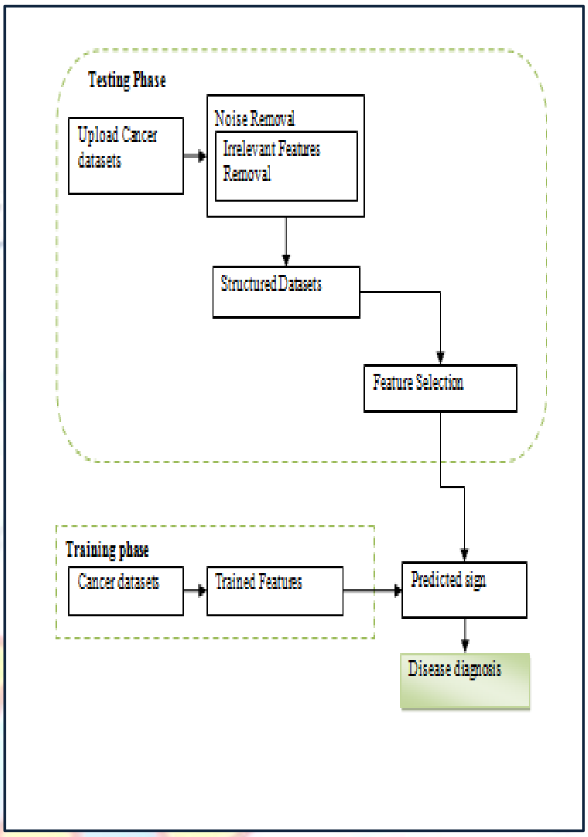

Fig: 3.1 System Architecture

Next step is classification of condensed set using neural network. .Based on the values acquired from training phase, the performance of the NN network is analyzed Genetic algorithm calculates the best fitness function. Neural network classified the matched and unmatched values. At last accuracy limitations are used for accuracy. FRR and FAR is evaluated.

The GA technique performs the dimensionality decrease process for obtaining the dataset with small size. Initially, the dimensionality reduction process is carried out on the UCI Repository cancer gene dataset for moving back the complexity in the gene classification. This procedure is performed because the dataset size is lofty dimensional, which increases the processing time and does not produce accurate result for the arrangement process. The fitness function is carried out to choose the best chromosomes among the generated chromosomes.

To obtain appropriate values for testing phase. In order to find the optimum structure, the NN network performance has been analyzed for the optimum number of hidden nodes and epochs. For this situation, the epochs will be set to a firm preset value. Then, the NN network was trained at the appropriate range of hidden nodes. The number of hidden nodes that have given the best performance is then selected as the optimum hidden nodes. After that, by fixing the best number of hidden nodes, the epochs will be evaluated in a similar way to achieve the best number of epochs that can give the highest or best accuracy.

IV. System Implementation

The following actions are carried out in the proposed system. They are;

- Dataset Acquisition
- Preprocessing
- Feature Selection
- Disease Diagnosis
- Evaluation Criteria

4.1 Dataset Acquisition

In this module, upload the datasets. The dataset may be microarray dataset. Gather the data from hospitals, data centers and cancer research centers. The collected data is pre-processed and stored in the knowledge base to build the model.

4.2 Preprocessing

Data pre-processing is an important step in the data mining process. The phrase "compost in, garbage out" is mainly applicable to data mining and machine projects. Data-gathering methods are often insecurely controlled, resulting in out-of-range values, impossible data combination, missing values, etc. Analyzing data that has not been carefully screened for such problems can produce ambiguous results.

4.3 Feature Selection

In this module is used to select the features of the given dataset. Attribute selection was performed to determine the subset of features that were highly correlated with the class while having low intercorrelation.

4.4 Disease Diagnosis

Based on the values acquired from training phase, the performance of the NN network is analyzed to obtain appropriate values for testing phase. In order to find the optimum structure, the NN network performance has been analyzed for the optimum number of hidden nodes and epochs. For this situation, the epochs will be set to a definite preset value. Then, the NN network was trained at the appropriate range of hidden nodes. The number of hidden nodes that have given the best performance is then selected as the optimum hidden nodes. After that, by fixing the optimum number of hidden nodes, the epochs will be analyzed in a similar way to obtain the optimal number of periods that container give the highest or best accuracy.

4.5 Evaluation Criteria

In this module, the performance of the proposed Genetic algorithm is extensively analyzed with that of some existing supervised and unsupervised gene clustering and gene selection algorithms using various statistical measures.

V. Result and Discussion

To analyze the performance of different algorithms, the experimentation is done on leukemia Cancer data sets. The major metrics for evaluating the performance of different algorithms are the class reparability index and classification accuracy of Neural Network rule. The proposed system provide improved accuracy rate in gene classification.

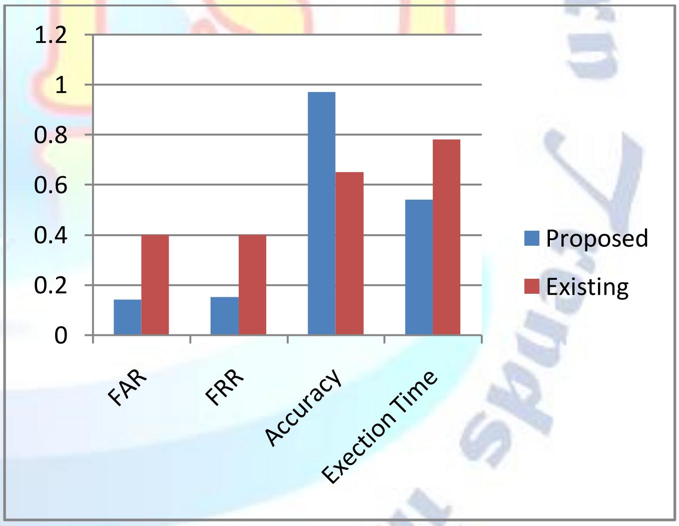

Figure 2. Accuracy of the proposed method

VI. Conclusion

This adversely affects the formation and normal function of blood tissues and cells. NNs are particularly attractive for diagnostic troubles without a linear resolution. Usually physicians analyze clinical and laboratory symptoms of blood cancer qualitatively and finally use bone marrow biopsies as a better formula for assess the nature of disease. Then precise and reliable detection of cancer need more Para clinical tests and costs and take much time. In this consider we apply simple and early clinical and assessment for proper discovery of leukemia. Therefore by using accomplished NN and Genetic Algorithm, we can predict cancer with least medical and laboratory tests and without obligation of much time. Accuracy of the recognition of cancer by the assembled artificial neural network was analyzed by roc and regression analysis. Outputs of trained NN for testing data were used to plot graphs.

The future work of our experimentation will include ever-increasing the number of records of the dataset for training the data in order to get accurate results and the network will be able to learn more professionally with more number of records. In addition to this, larger datasets can be obtained &applied and the approach can be tested to provide higher accuracy results. Different neural networks and other classification techniques can also be tried to obtain better results. Use of support vector machines will be considered in the future work as a classification tool.

VII. Reference

[1] Li, Eldon Y. "Artificial neural networks and their business applications. "Information& Management 27.5 (1994): 303-313.

[2] Xue-wen Chen and Michael McKee," Finding expressed genes using genetic algorithms and support vector machines", Department of Electrical and Computer Engineering, California State University 18111 Nordhoff Street, Northridge, CA 91330, USA,2003.

[3] Marc C. Chamberlain, M.D.," Leukemia and the Nervous System",2003

[4] S. H. Rezatofighi et.al," A New Approach to White Blood Cell Nucleus Segmentation Based on Gram-Schmidt Orthogonalization", International Conference on Digital Image Processing,2005.

[5] Huerta, Edmundo Bonilla, Beatrice Duval, and Jin-Kao Hao. "A hybrid GA/SVM approach for gene selection and classification of microarray data." Applications of Evolutionary Computing. Springer Berlin Heidelberg, 2006. 34-44.

[6] M. Pei," Feature Extraction Using Genetic Algorithms", Case Center for Computer-Aided Engineering and Manufacturing Department of Computer Science Genetic Algorithms Research and Applications Group, 2006.

[7] Yuh-Jye Lee+ and Chia-Huang Chatom," A Data Mining Application to Leukemia Microarray Gene Expression Data Analysis", Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, No. 43, Sec. 4, Keelung Rd., Taipei, 106, Taiwan,2007.

[8] Alba, Enrique, et al. "Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms." Evolutionary Computation, 2007. CEC 2007. IEEE Congress on. IEEE, 2007.

[9] NiponTheera-Umpon," Morphological Granulometric Features of Nucleus in Automatic Bone Marrow White Blood Cell Classification", IEEE transactions on information technology in biomedicine, VOL. 11, NO. 3, MAY 2007.

[10] YvanSaeys and et.al," A review of feature selection techniques in bioinformatics", Vol. 23 no. 19 2007, pages 2507–2517.

[11] Raja, Chital, and JyotiRangole. "Detection of Leukemia in microscopic images using image processing." Communications and Signal Processing (ICCSP), 2014 International Conference on. IEEE, 2014.