Microarray Data Classification using Artificial Neural Network

K. Kalyani

Abstract*: The accurate cancer classification is very important task for cancer treatment. Recently the informative genes are identified from the thousands of genes for correct cancer classification. The collection of microscopic Deoxyribo Nucleic Acid (DNA) microarray is attached in the solid surface. In this study, DNA microarray data is used for cancer classification. The system uses Artificial Neural Network (ANN) for DNA Microarray Data Classification (MDC). Initially, the preprocessing step is made by using log transformation method to remove the raw data and feature selection. These selected features are classified by using ANN. REctified Linear Unit (RELU) activation function is used as the activation function in each ANN layer. Softmax is used for classification. The performance of the system is made by using leukemia dataset. MDC system produces the classification accuracy of 91.65% by using ANN

Keywords: Microarray Data Classification, Flat Pattern Filtering, Feature Selection, Artificial Neural Network

I. INTRODUCTION

Dual Tree M-Band Wavelet Features (DTMBWF) based MDC is described in [1]. Initially the input microarray data is given to DTMBWF. The K-Nearest Neighbor (KNN) classifier is used for classification. Gene selection using MDC using single and multiple filters is discussed in [2]. At first, the microarray data is given to single filter single wrapper and multiple filter multiple wrapper to improve robustness. The classification is made by KNN, Support Vector Machine (SVM) and weighted voting classifier.

MDC based on supervised attribute clustering is discussed in [3]. The input microarray data is preprocessed by gene clustering method. Then the classification is made by naïve bayes, KNN and SVM. MDC using data dependent kernel machines is described in [4]. The input microarray data is given to bootstrapping based resampling. The MDC is made by using SVM, KNN and uncorrelated linear discriminant classifiers.

MDC for cancer classification based dimension reduction is discussed in [5]. The input microarray data is given to singular value decomposition. The features are selected by recursive feature elimination. SVM classifier is used for classification. MDC using genetic algorithm is described in [6]. The input microarray data is given to Pearson, spearman, cosine coefficients, mutual information, signal to noise ratio and information gain to select the efficient features. The genetic algorithm based classifier is used for classification.

SVM mapreduce based MDC is described in [7]. Initially the gene filtering is performed to remove gene ranking, noisy data and filtering informative genes. The SVM mapreduce is used for classification of the genes. Initially the features are selected by using feature subset selection method. Classification is made by SVM. Gene expression data using associative classification is discussed in [8]. At first, the gene expression data is given to gene filtering. Then discritization is done for data preprocessing. Associative classification is used for classification.

MDC for high dimensional data based feature selection is discussed in [9]. Initially the gene filtering is used to select and reduce the features. The C4.5 classifier is used for classification. DNA based MDC for cancer classification is discussed in [10]. The input microarray data is given to novel strategy for extraction. Then gene rank selection is performed. The classification is made by simple rule based ensemble classifiers.

MDC for gene selection algorithm and combination of ranking and clustering is discussed in [11]. The genes are selected by clustering and non-clustering gene selection. MDC is made by using SVM classifier. MDC using fuzzy inference system is described in [12]. The features are selected by t-statistics method. The selected features are classified by using fuzzy inference system.

An efficient approach for MDC system using ANN is discussed in this study. The organization of paper is as follows: Section 2 describes the methods and materials used for MDC system. Results and discussion of MDC system is explained in section 3. The last section concludes the MDC system.

II. METHODS AND MATERIALS

The ANN based MDC system is shown in figure 1. At first the input microarray data is given to preprocessing step using log transformation to remove raw data to get clear data and also it selects the efficient features. Then ANN is used for the classification of MDC system.

A. Gene data preprocessing

The preprocessing of microarray data is an essential stage to remove raw data in the dataset. In this study, the log transformation technique is used for preprocessing. It is a popular transformation technique to transform the data into normal data. The log-normal distribution is followed by original data. The log transformation technique is used in this study to remove raw data and select the efficient features for classification. The highly skewed distributions are reduced by log transformation. It can be used in the data to help the assumptions of inferential statistics. The log transformation is given by,

$$R = Trans(q) \quad (1)$$

where q is the pixels of the input image and R is the pixels of the output image.

Revised Manuscript Received on December 12, 2019.

***** Correspondence Author

K Kalyani*, Assistant Professor, Department of Computer Science, Marudupandiyar College of Arts and Science (Affiliated to Bharathidasan University), Thanjavur.

The transformation function is used to map the each value of q in the R. In this study, log transformation is used for preprocessing of input microarray data.

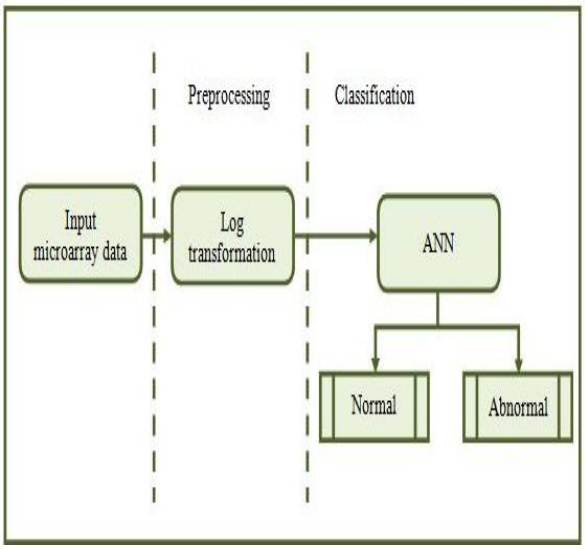

Fig. 1.MDC system using ANN

B. MDC using ANN

ANN model is based on structure functions. ANN structure may change due to the input and output. ANN has input, output and hidden layer. It is also used in hand moment classification [13], heartbeat classification [14] and electrocardiogram signal classification [15]. The input and output patterns with complex relationship are found in the nonlinear statistical data modeling tools. It has a group of nodes which is interconnected by the neurons in the brain. The node in one layer is connected with the other layer. Figure 2 shows the ANN architecture.

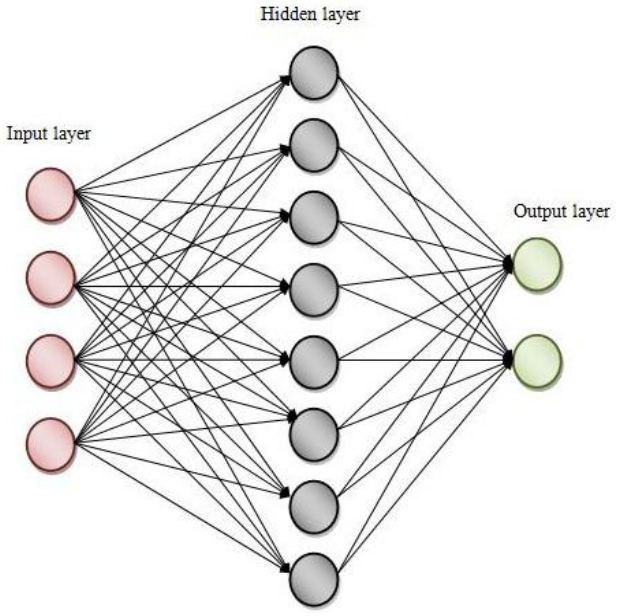

Fig. 2.ANN architecture

In figure 2, the artificial neurons are represented as circular nodes and the connection of artificial neurons are made by using arrows from input to output layer. The first layer neuron is connected to the next layer. The neuron in each layer is connected to the next layer.

ReLU activation function is used in each layer of ANN for the activation. The ReLU activation function is defined by the following equation.

$$k = \max(0, l) \tag{1}$$

ReLU activation function is most commonly used activation function in neural networks. The softmax layer is used for classification. The n-dimensional vector is taken from the real values with the range of 0 and 1. It is used for binary classification with two or more classes. The softmax layer is defined in equation 2.

$$Soft(v_i) = \frac{g^{v_i}}{\sum_j g^{v_j}} \tag{2}$$

In this study, the ANN is used for the classification of microarray data in MDC system. The ReLU activation is used in each layer of ANN for the classification. Finally the softmax layer in the fully connected layer is used for the classification of MDC system.

III. RESULTS AND DISCUSSION

Performance of MDC system is analyzed by using gene expression dataset using ANN classifier and publically available cancer microarray dataset. It consists of leukemia dataset [16]. The table 1 shows the microarray dataset description.

Table- I: Leukemia gene dataset-Description

Dataset	No of cancer cases	No of normal cases	Total no of cases	No of Attributes
Leukemia	47	25	72	7129

The input microarray is given to log transformation for preprocessing to remove raw data and also it selects the efficient features. Then the selected features are used for classification using ANN. Table 2 shows the classification accuracy of MDC system using ANN for ten attempts.

Table- II: Classification accuracies obtained at ten attempts using ANN for MDC system

No. of attempts	Classification accuracy (%)	No. of attempts	Classification accuracy (%)
1	93.50	6	96.00
2	93.00	7	91.50
3	91.50	8	94.50
4	89.00	9	92.00
5	88.50	10	90.00
		Average accuracy (%)	**91.65**

From the above table, it is observed that the overall classification accuracy of 91.65% obtained by using ANN for MDC system. The higher classification accuracy of 96% and the lowest classification accuracy is 88.5% obtained by using ANN for MDC system. Figure 3 shows the performance of ten attempts for MDC system using ANN.

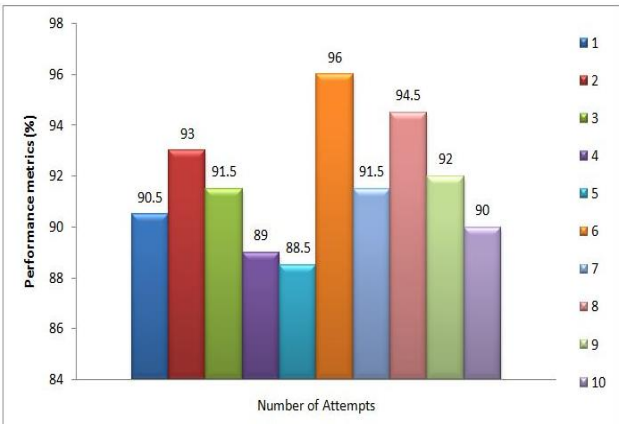

Fig. 3.Classification accuracies of ten attempts using ANN for MDC system

In above figure, it is clearly observed that the higher classification accuracy is 96% and lower classification accuracy is 88.5% using ten attempts for MDC classification system using ANN classifier.

IV. CONCLUSION

An approach for MDC system using log transformation and ANN is described in this study. The performance of the system is made by publicly available gene expression database. The MDC system uses leukemia datasets for performance evaluation. The input leukemia microarray data is given to log transformation to reduce raw data and also to select the efficient features for preprocessing. The preprocessed images are used for further step. ANN classifier is used for the classification of MDC system into normal and abnormal. The average classification accuracy is calculated for ten attempts of classification due to variations in accuracy. The MDC system produces the average classification accuracy of 91.65% obtained by using ANN classifier.

REFERENCES

1. J.M. Sonawane, S.D. Gaikwad, and G. Prakash, "Microarray Data Classification Using Dual Tree M-Band Wavelet Features", *International journal of advances in signal and image sciences*, Vol. 3, No. 1, 2017, pp. 19-24.
2. Y. Leung, and Y. Hung, "A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 7, No. 1, 2010, pp. 108-117.
3. P. Maji, "Mutual information-based supervised attribute clustering for microarray sample classification", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24, No. 1, 2010, pp.127-140.
4. L. Shen, and E.C. Tan, "Dimension reduction-based penalized logistic regression for cancer classification using microarray data", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 2, No. 2, pp. 166-175.
5. X.R. Jenifer, and R. Lawrance, "Classification of microarray data using SVM mapreduce", *IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing*, 2017, pp. 1-6.
6. E. Ahmed, N. El-Gayar, and I.A. El-Azab, "Support Vector Machine ensembles using features distribution among subsets for enhancing microarray data classification", *International Conference on Intelligent Systems Design and Applications*, 2010, pp. 1242-1246.
7. S. Alagukumar, and R. Lawrance, "Classification of microarray gene expression data using associative classification", *International Conference on Computing Technologies and Intelligent Data Engineering*, 2016, pp. 1-8.
8. H. Yu, and S. Xu, "Simple rule-based ensemble classifiers for cancer DNA microarray data classification", International Conference on Computer Science and Service System, 2011, pp. 2555-2558.
9. M.J. Rani, and D. Devaraj, "A Combined Clustering and Ranking Based Gene Selection Algorithm for Microarray Data Classification", *IEEE International Conference on Computational Intelligence and Computing Research*, 2017, pp. 1-5.
10. M. Kumar, and S.K. Rath, "Classification of microarray data using Fuzzy inference system" *International Conference on Recent Trends in Information Technology*, 2014, pp. 1-8.
11. K.P. Shashikala, and K.B. Raja, "Palmprint identification using Log Transformation of Transform Domain Features" *International Conference on Electronics and Communication Systems*, 2014, pp. 1-5.
12. P. Meaney, T. Grzegorczyk, S.I. Jeon, and K. Paulsen, "Log transformation with Gauss-Newton microwave image reconstruction reduces incidence of local minima convergence", *IEEE Antennas and Propagation Society International Symposium*, 2009, pp. 1-4.
13. U. Baspinar, H.S. Varol, and K. Yildiz, "Classification of hand movements by using artificial neural network", *International Symposium on Innovations in Intelligent Systems and Applications*, pp. 1-4.
14. M.H. Song, J. Lee, H.D. Park, and K.J. Lee, "Classification of heartbeats based on linear discriminant analysis and artificial neural network", *IEEE Engineering in Medicine and Biology 27th Annual Conference*, 2005, pp. 1151-1153.
15. D. Gnana Rajesh, "Analysis of MFCC features for EEG signal classification", *International Journal of Advances in Signal and Image Sciences*, Vol. 2, No. 1, 2016, pp. 14-20.
16. T. Furey, N. Cristianini, N. Duffy, D. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data", *Bioinformatics*, Vol. 16, No. 10, 2000, pp. 906–914.

AUTHORS PROFILE

Dr.K.Kalyani, Assistant professor, Department of Computer Science, Marudupandiyar College of Arts and Science, Thanjavur. She has completed her Master of Computer Application at Bharathidasan University , Trichy in the year of 2007 and Master of Philosophy in Computer Science at PRIST University in 2010 . Doctor of Philosophy in Computer Science at A.V.V.M. Sri Pushpam College , Poondi and awarded in the year of 2017. Also She has published more than 15 International journal of Computer Science and presented more than 15 papers in National and International Conferences of various domains