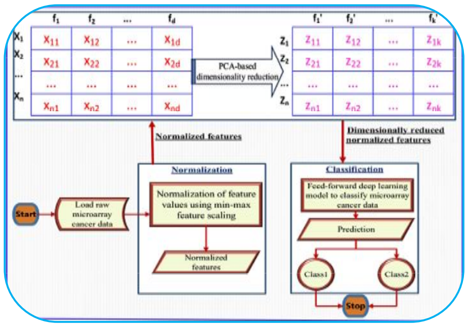
REVIEW OF RESEARCH
ISSN: 2249-894X
IMPACT FACTOR : 5.7631(UIF)
VOLUME - 11 | ISSUE - 4 | JANUARY - 2022

CLASSIFICATION OF MICROARRAY CANCER DATA USING DEEP LEARNING

Pallavi Madhukarrao Tekade, Dr. Vinod M. Vaze and Dr. Ram B. Joshi

ABSTRACT:

Cancer is a group of diseases characterized by abnormal growth of cells. In a healthy body, the growth of cells is controlled in such a way that they grow and die systemically. Internal and environmental factors can impair the genetic make-up of cells leading to the continuous growth of cells to form tumors. Improper cell division and loss of deoxyribonucleic acid are major internal factors, while exposure to substances such as tobacco smoke, radiation and the sun's ultraviolet rays are important environmental factors that cause cancer.



KEY WORDS: *diseases characterized , continuous growth, major internal factors.*

INTRODUCTION:

The diagnosis of cancer and gene expression is differentiated using profiles. Microarray gene expression data analysis is one of the most challenging areas of research in the fields of computational biology, genomics, statistics and pattern classification. The main challenge of microarray cancer analysis is related to the high curse and small sample size of the existing due to irrelevant and meaningless genes. In addition, medical datasets are generally noisy, with variations in feature values and an unbalanced number of classes leading to over-fitting and low classification accuracy. The need for research in microarray data analysis, in particular the classification of cancer, helps to identify and understand the features that contribute to the development of cancer. An important role played by the microarray data classification method is the identification of genes that contribute to specific biological effects and the use of such genes to predict new observations. It helps in early detection of cancer so that domain experts can create treatment plans to increase the survival rate of cancer patients. Therefore, the problem requires careful construction of a model that takes the input pattern that represents the objects and estimates the range of the object under consideration and, therefore, develops an accurate estimation model based on the given test data.

Microarray cancer data classification which has major functions such as data collection from its source, pre-processing, feature selection, classification and post-classification analysis. Feature selection is the process of selecting important genes from thousands of highly correlated and informative genes and providing these filtered data component classifiers to achieve better classification accuracy. The selection of features plays an important role in the classification of cancer data to identify optimal and relevant subsets of features, thereby increasing classification accuracy and computational stability. Analysis of microarray cancer data analysis plays an important role in obtaining good information about the disease which ultimately helps in planning conclusive measures and improving the cancer diagnosis process. In our work, we propose the use of in-depth learning-based classifiers to classify microarray cancer data. Intensive learning takes a large amount of data to learn the behavior of features during training and estimates the class of unseen data. To validate the proposed method, we have considered eight standard microarray cancer datasets: prostate, colon, central nervous system (CNS), ovarian, leukemia and lung cancer dataset. Feature values are calculated using a minimum-maximum approach to overcome the decision bias in favor of high value features.

PROPOSED METHOD FOR CLASSIFYING BINARY CLASS DATASETS:

The proposed work involves various phases, including feature scaling, amplitude reduction, and deep feed forward neural network-based classification methods, including parameter settings. The proposed approach framework includes key tasks such as loading RAW microarray cancer data, then generalizing using the Min-Max method, reducing the dimensions as presented in Figure 3.1, and in-depth learning-based classification.

In the field of pattern recognition and machine learning, it is a known fact that feature scaling techniques are explored to normalize data. This process brings all data components to the same scale to avoid outliers and thus increases the quality of the estimate. Due to the high variability in features in microarray cancer datasets, we propose exploring feature scaling as one of the pre-processing techniques for data normalization. Feature scaling is done using the Min-Max method to fit into the sigmoid activation function as it considers values between 0 and 1 with a threshold value of 0.5 for binary classification during model training. In our work, the Min-Max feature scaling technique, which measures data in the range of [0, 1], has been considered to make it suitable to fit into the sigmoid function during the training of the proposed model.

CLASSIFICATION BASED ON DEEP LEARNING:

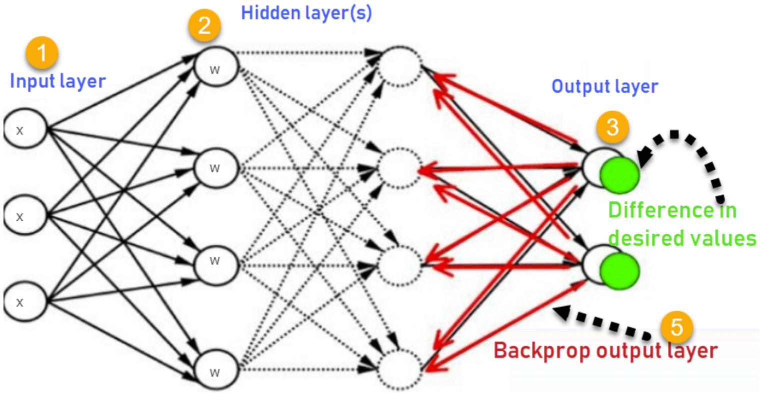
The proposed work involves various phases, including feature scaling, amplitude reduction, and deep feed forward neural network-based classification methods, including parameter settings. The proposed approach framework includes key tasks such as loading raw microarray cancer data, then generalizing using the Min-Max method, reducing the dimensions as presented in Figure 3.1, and in-depth learning-based classification. To explore deep feed-forward neural network-based models for classifying microarray data. An in-depth feedforward network is an important model of in-depth learning. The model is called feedforward neural network because the information travels from x to the evaluated function, through the intermediate computation used to define f, and finally to output y, where x is the input characteristic vector and yi is the approximate class label. There are no feedback connections in which the output of the model is returned to itself. If the feed forward neural network is extended to include feedback connections from the same node, it is not a feed forward model, but a recurring neural network model. The feed forward neural network is represented by various directed functions. A model is a guided graph that describes how functions are connected together.

For example, let us consider the three functions $f_1, f_2\ and\ f_3$ connected in a chain to form a network and thus define $f(x) = \left(f_1, f_2, f_3, (x)\right)$. These chain structures are the most commonly used structures in neural networks. In this case, f_1 is the first level of the neural network, f_2 is the 2nd level and f_3 is the 3rd level and so on.

A fully connected neural network approach is defined to have input levels that start with input features. Furthermore, hidden layers are defined with parameters such as activation and log-loss function. The output of each hidden layer is converted to the next hidden layer by firing with the output of the sigmoid function. Finally, the model converts to a single output layer for the classifier to estimate the class label for each sample. The output layer contains a single neuron that returns as a class one or class two class label.

The proposed method from input characteristics and weights, activation function, output, error calculation, and error backpropagation. To give a single output each input feature is multiplied by its weight which will also be an input for the next level and will give an estimate at the end. The actual class is evaluated by subtracting the label from the estimated class and the difference is re-triggered as backpropagation error. The back progression of the errors is to update the weight and ultimately to give the maximum possible estimate, so the minimum error is registered.

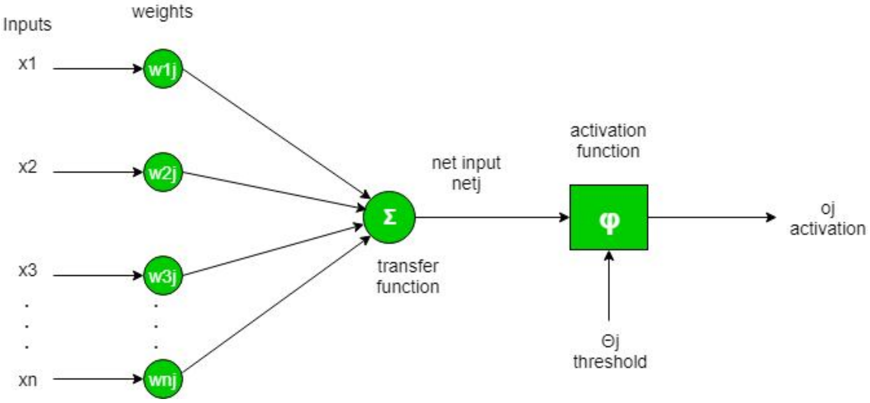
Figure 3.2 Back Propagating Deep Learning Model



As shown in Figure 3.2, the activation function a input data defines the output of the dot product of X_1, X_2,….,X_n and relative weight W_1,W_2,W_3….W_n in the first hidden layer h1. The output of the activation goes to the node next to the hidden layer and finally, the output is triggered to the node of the output layer. Weights are updated by transmitting weights to the previous nodes until an optimal estimate is reached

The proposed in-depth learning approach takes input data in the form X = X_1, X_2,….,X_n where $X \varepsilon \mathbb{R}^{n-k}$ and each input vector is multiplied by its corresponding weight. Therefore, the weight vector for input data is W_1,W_2,W_3….W_n; As shown W_n and bias *b* are added to the loaded input vector. At each level of the model, the loaded input vectors are multiplied by the sigmoid activation function to obtain the intermediate potential results in the hidden layers based on following equation.

Figure 3.3 Activation Function



$$y_p = f\left(\sum_{i=1}^{n} w_{n_i} . z_{k_i} + b\right)$$

Equation – 3.1

Where,

y_p = Predicted of Dependent Variable

w_{n_i} = Matrices of Weight

z_{k_i} = Vector of Feature

f = Sigmoid Activation Function

We propose to use the sigmoidal activation function during the training of the Deep Feed-Forward Neural Network model to estimate the classification for the new data component. The sigmoid function is used in our model as the activation function which limits the output between 0 and 1 which handles the estimation of the class labels as probability and is given below.

$$f(z_1) = \frac{1}{1+e^{-(z_k.w_k)}}$$

Equation – 3.2

Equation 2.7 handles multi-features on hidden layers by learning the interesting behaviour of data as it moves to the output layer.

$$\begin{cases} 1, \; if\; w_0 + w_1.z_1 + w_2.z_2 +, \ldots + w_n.z_n \geq 0.5 \\ \qquad 0, Otherwise \end{cases}$$

Equation – 3.3

Equation 3.3 is used to find the relationship between a target-dependent variable and one or more independent features. Weight start is done randomly and uniformly, then the weight is updated during the training period. The proposed model compares the probabilities of the actual and approximate class and assigns a class value based on the threshold value of the binary class and assigns a prediction to any class.

In an in-depth learning-based classification, the final estimate error is calculated as the difference between the approximate class label y_p and the given class label y_i using a predetermined objective function called cross-entropy. Errors are re-transmitted across the network to optimize the weight for the minimum error value.

The Deep Feed-Forward approach, which is applicable for classification purposes. These include input characteristics and weight, activation function, output, error calculation, and error return propagation. To give a single output each input feature is multiplied by its weight which will also be an input for the next level and will give an estimate at the end. The actual class is evaluated by subtracting the label from the estimated class and the difference is re-triggered as backpropagation error. The back-propagation of errors is to update the weight and finally give the maximum possible estimate, and therefore the minimum error is reported.

REFERENCES:

1. Bose S, Das C, Banerjee A, Ghosh K, Chattopadhyay M, Chattopadhyay S, Barik A. 2021. An ensemble machine learning model based on multiple filtering and supervised attribute clustering algorithm for classifying cancer samples. Peer Journal Computer Science. 7: e671 DOI 10.7717/peerj-cs.671, DOI 10.7717/peerj-cs.671.
2. Banerjee M, Mitra S, Banka H. Evolutionary rough feature selection in gene expression data. IEEE Trans Syst Man Cybern Part C (Applications and Reviews). 2007;37:622-32.
3. Barnali Sahu, Satchidananda Dehuri and Alok Jagadev (2018), 'A Study on the Relevance of Feature Selection Methods in Microarray Data', The Open Bioinformatics Journal, 11, 117-139.
4. Bhatt, U.; Ravikumar, P.; Moura, J.M.F. Building Human-Machine Trust via Interpretability. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 9919–9920
5. Gulum, M.A.; Trombley, C.M.; Kantardzic, M. A Review of Explainable Deep Learning Cancer Detection Models in Medical Imaging. Appl. Sci. 2021, 11, 4573. https://doi.org/10.3390/app11104573
6. Hanaa Fathi, Hussain AlSalman, Abdu Gumaei, Ibrahim I. M. Manhrawy, Abdelazim G. Hussien, and Passent El-Kafrawy (2021), 'An Efficient Cancer Classification Model Using Microarray and High-Dimensional Data', Computational Intelligence and Neuroscience, pp. 1-14.

7. Karthik S. and M. Sudha (2018), 'A Survey on Machine Learning Approaches in Gene Expression Classification in Modelling Computational Diagnostic System for Complex Diseases', International Journal of Engineering and Advanced Technology (IJEAT), ISSN: 2249 – 8958, Volume-8 Issue-2, pp. 182-191.
8. Karthikeyan B, Sujith Gollamudi, Harsha Vardhan Singamsetty, Pavan Kumar Gade, Sai Yeshwanth Mekala (2020), International Journal of Advanced Trends in Computer Science and Engineering, Vol-9, Issue-2, pp. 981-984