International Journal on Emerging Technologies **10**(3): 64-68(2019)

ISSN No. (Print) : 0975-8364
ISSN No. (Online) : 2249-3255

Genes Expression Classification using Improved Deep Learning Method

Rajit Nair[1] and Amit Bhagat[2]
[1]Department of Computer Application, Maulana Azad National Institute of Technology, Bhopal, India
School of Engineering and Technology, Jagran Lakecity University, Bhopal, India.
[2]Department of Computer Application, Maulana Azad National Institute of Technology, Bhopal, India.

(Corresponding author: Rajit Nair)
(Received 08 June 2019, Revised 25 August 2019 Accepted 05 September 2019)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: In the last few years, there is tremendous research in the area of genes expression. So many classification algorithms had been implemented for classifying the thousands of expression of different types of genes. In this work we will present how different classification algorithm works on the genes expression and we will also introduce the concept of deep learning in the form of Sigmoidal Neural Network. Neural network has been applied because many times it has been observed that classification accuracy has not been improved by traditional machine algorithms like Naïve Bayes, Support Vector Machine, Logistic Regression, Random forest etc. However, the large number of genes and the complexity of biological networks greatly increase the challenges of comprehending and interpreting the resulting mass of data, which often consists of millions of measurements. In this approach we will show how deep learning concept will improve the classification accuracy by deep analysis. The main reason behind performing this classification is to predict the disease by genes expression analysis.

Keywords: Genes, Classification, Artificial Neural Network, Sigmoid, Neurons, Precision, Recall, Accuracy

I. INTRODUCTION

Nowadays technology is changing day by day and we can say upcoming technologies are far better than the previous one mainly in terms of performance. Same concept occur in the area of biological field in which most of the people are using machine learning classification algorithm [1] but these algorithm are not performing very effectively in some cases. So to improve the performance we are using the concept of Artificial Neural Network [2]. The biological area on which our analysis has been done is gene expression [3]. Thousands of genes generate the different expression levels so we have to analyze this expression first for further research. Comparing our approach with traditional approach in the area of genomic research, most of the work is done by examining the data of single genes but this time we are focusing on multiple genes. The computational methods that predict genes expression from various genes are highly desirable for understanding their combinatorial effects in gene regulation. The knowledge helps us to develop 'epigenetic drugs' like diabetes, cancer. This work allows automatic extraction of complex interactions among important features. We propose an approach for classification using neural network during genes expression analysis.

By gene regulation the gene expression can be controlled to become high or low. Cells normally regulate genes by using wide range of mechanism and it increases or decrease the gene products through translation such as proteins. Many factors are there which regulate genes at the DNA level [4] and these can be range from mutation in DNA sequences to various proteins binding to them. One of the major issues during gene expression classification is feature selection [5]. We have already seen in the past years that Naïve Bayes [6], Support Vector Machine [7], Logistic regression [8] and K-Nearest Neighbour (KNN) [9] are used to develop classifier model for gene expression data. We cannot compare all subsets of genes, it is an unfeasible approach. We cannot examine all the combinations directly, so we need an efficient method to sample from less combination to find the optimal solution. The major issue in gene expression classification is feature selection [10], and we have observed that genes expressions are high dimensional data and it is difficult to process these type of data, so we must reduce the number of features before performing classification. Huerta *et al.* [11] has proposed a method based on fuzzy logic for finding informative gene combinations and to find relevant subtype. Genes expression data are not only high dimensional but at the same time they contain redundant information and noises also.

Some of the issues like identifying the informative genes are also involved in genes expression analysis which actually gives the disparity between the number of genes measured and number of individuals sampled [12]. Actually those genes are known as informative genes whose expressions are strongly correlated with the class label distinction. In most of the approaches we always try to find the informative genes after finding this rest of the genes are considered as noise in the dataset. The feature selection in gene expression reduces the training time and it is very much needed process due to its high dimensionality characteristic [13]. Feature selection in genes expression improves the classification accuracy also because it removes the irrelevant genes from the

dataset.
Deep learning models are widely used in the area of bioinformatics which includes biomedical signal processing, biomedical imaging and omics [14]. However, there is very less research done in the area of genes expression using deep learning methods. That's why in this work we focused on the deep learning model to classify genes expression data.

II. BACKGROUND

Artificial Neural Networks (ANN) is inspired by the computational paradigm of the biological neural networks which is basically connection of interconnected nodes. A gene is an instance of DNA that contains all the necessary information that is required to create protein in our body. Cells generate different types of expression that depends on the type of cell. Microarray expression generates the thousands of gene expression. Mainly these experiments consist of expression of each gene under different conditions or it can be computing each gene in the same condition under different tissues, sometime these tissues are cancerous tissue. Most of the gene classification problem we have set of genes with their corresponding class label. During gene expression classification, we try to find the relation between the genes who belongs to same class. The objective of classifier is to maps the object to its class label and this can be done by analyzing the samples generated by the set of features which are actually called as input features. It is also each sample is already mentioned with their class label. Our work is also based on binary classification and it is used to develop a classifier using training samples from benign and malign tumor samples. Many research papers have already discussed about the classification of genes expression on the basis of machine learning algorithms but there are very papers based on deep learning for genes expression classification. This work will present how deep learning has improvised the classification of genes expression [15]. The classification is done on the basis of malignant and benign tumors cells in our body or we can say that it is mainly applied for cancer and non cancer classification. We have implemented the deep learning [16] concept with the help of log function also called the sigmoid function [16] was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment.
Although in this work we will also present how other machine learning algorithms work on the same dataset with their accuracy score, confusion matrix including the classification report. Firstly we will split the training dataset and test dataset separately for separate training and test accuracy. Here we have used sigmoid function. Basically the sigmoid function is an activation or transfer function and that is added with the output end of any neural network. It is used to determine the output of the network which can be 0 to 1 or can be -1 to 1 and many more. The activation function can be linear and non linear but in our work we have taken sigmoid function that is having values 0 to 1 and it is mostly used for prediction purpose like yes or no. If we are using linear activation function in that case we can stack as many hidden layers as we need.

III. PROPOSED WORK

In the given below figure we had shown our approach for genes classification using sigmoid neural network. In this approach we have taken two dataset i.e. training and testing dataset, we have used sigmoid function which is actually the inverse of log function. In this work sigmoid function is used because it predicts the value between 0 to 1. Output is generated in the form probability value. If predicted output is needed in the form of value between 0 to 1, sigmoid is the preferred choice. In this function is differentiable which indicated that slope of the sigmoid curve can be found using two points. The function is monotonic but function's derivative is not. The logistic sigmoid function can cause a neural network to get stuck at the training time.

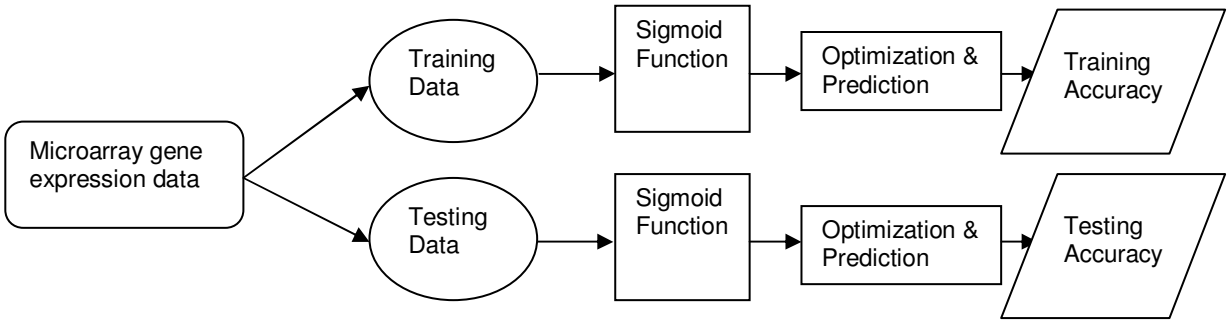

Fig. 1. Flow chart of proposed work.

A. Sigmoid Neural Network (SNN)

The proposed approach takes micro array genes expression for classification using deep learning methods, which has shown the improved accuracy than other existing algorithm, here we have used sigmoid function as activation function. The gene expression data set are mainly high dimensional dataset so most of the time for improved classification we need feature selection but in this work we do not need any feature selection method separately. It is already applied with the help of proposed algorithm. In this approach first we have split the training and testing dataset into train and test set, the training dataset is cancer_data.csv and testing dataset is test_cancer_data.csv. The dataset consist of 31 features including their class label. In neural network we will take input nodes equal to number of features i.e. 30, one is used as output node, each node has assigned some weight, the values of this weight changes till the predicted accuracy has achieved. After that we will use sigmoid function as activation

function then at last we get output from output layer. The predicted output will be compared with test value and then the accuracy will be achieved. The loss is computed after each iteration and we have taken 190500 numbers of iterations.

The neural network mainly consists of three layers which are input layer, hidden layer and output layer. There can be multiple hidden layers and we have observed that more hidden layers can provide good accuracy but it totally depends on the dataset and the number of features involved in the dataset and the size of the dataset. If we are working with binary classification there will be one output neuron and if we are working on multiclass classification then there will be multiple output neurons.

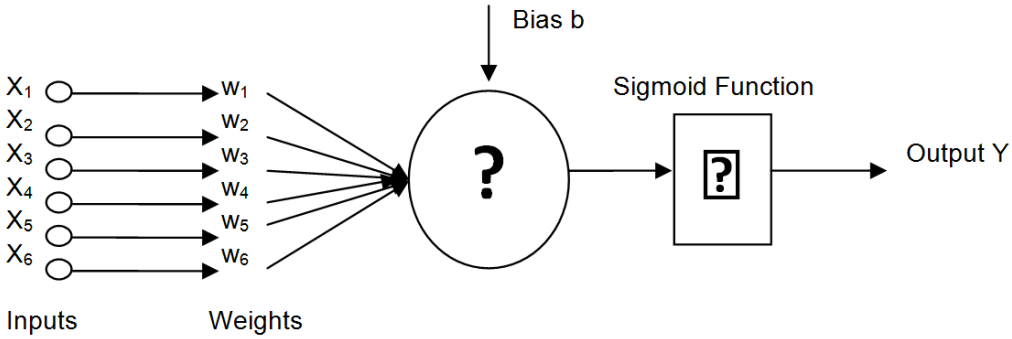

Fig. 2. Working of Sigmoid function on input features.

An artificial neuron calculates a "weighted sum" of its input, adds a bias and then activation function decides whether it should be yes or no. Let us take an equation

$$Y=\sum (\text{weight} * \text{input}) + \text{bias} \quad (1)$$

Consider the above equation in which Y can be anything ranging from –infinity to +infinity. The neuron does not know the exact value bounds. So how can we decide whether neuron should predict or not. That's why in neural network we have added "activation functions" for this purpose. To check whether the Y value produced by a neuron and take decisions that whether outside connections should consider this neuron as active or not.

B. Sigmoid Function

It looks step like function but actually it is non linear in nature even its combination are also non linear but it has smooth gradient. Next advantage of sigmoid function is its output range in between 0 to 1 unlike linear function which is -∞ to +∞. This means that we have activation which is bounded in a range. This is why it is the most widely used activation function today and the corresponding equation is given below. Now we will see mathematically, a Sigmoidal function is a map $\boldsymbol{\sigma} : R \rightarrow R$ where R is the set of all real numbers, having the following limits for the argument to the function (x) tending to $\pm\infty$.

$$\boldsymbol{\sigma}(x) = \frac{1}{1+e^{-x}} \quad (2)$$

In the above formula e is the Euler's number Let us discuss some of the points related to sigmoid function:

- It produces output which is between 0 and 1. If the sigmoid function computes value greater than or equal to 0.5 then it gives 1 as output.
- It does not form a jerk on its curve, means it is smooth.

C. Methods

Most of the deep learning algorithms work on images and in this area had shown significant improvement, but in this paper we deal with gene expression data that is in text form and how they are classified, on the basis of classification we will compute the train accuracy as well as test accuracy also. We have separate dataset for training as well as for testing and it is named as Wisconsin dataset. We have taken the dataset from UCI machine learning repository, in which we have taken the dataset that is based on genes expression namely they are cancer_data.csv and test_cancer_data.csv. Out of these features diagnosis is the target label which is having value 0 or 1. The 0 is for cancerous and 1 is for non cancerous. Our work contains two separate dataset in which one is training dataset (cancer_data.csv) and the other one is testing dataset (test_cancer_data.csv). So we will calculate the training accuracy as well as the testing accuracy. In this process we will split both the training and testing dataset into train data and test data. Most of the cases train data contains 70% of the total dataset and test data contains 30% but there is no such rule for that and it is totally depending on us. Our training dataset contains 419 instances and test dataset contains 150 instances.

IV. RESULTS AND DISCUSSION

Our implementation is done in Python 3.6 with 4GB RAM, we have taken Python, because it is one of the best tool used for prediction purpose. Here we have taken the parameters like precision, recall and accuracy to represent the performance of the existing and proposed classifier. The parameters are computed on the training dataset and testing dataset separately. In case of our classifier we have also taken two other parameters known as batch size and epochs. Updation of weight depends upon the number of observation specified by batch size. Epoch is nothing but the total number of iterations. There is no specific rule for choosing the batch size and epoch.

Table 1: Classification report on training dataset.

Classification Algorithm	Accuracy Score (%)	Precision	Recall
Logistic Regression	89.68	0.89	0.84
Naive Bayes	85.71	0.94	0.72
Support Vector Machine	81.71	0.99	0.54
K Nearest Neighbor	87.30	0.85	0.82
Sigmoidal Neural Network	91.6	0.99	0.84

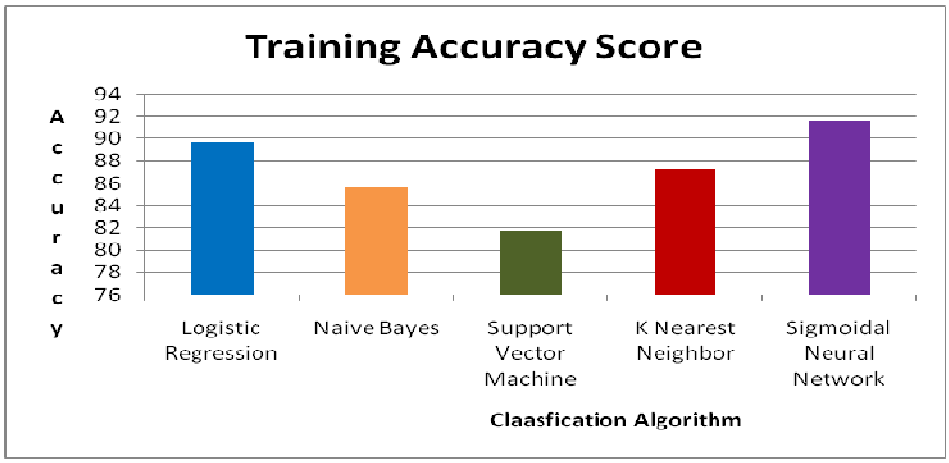

Fig. 3. Comparison graph of the training accuracy of the proposed SNN with other classifiers.

Table 2 Classification report on testing dataset.

Classification Algorithm	Accuracy Score (%)	Precision	Recall
Logistic Regression	93.33	0.99	0.75
Naive Bayes	91.11	0.94	0.66
Support Vector Machine	88.8	0.99	0.54
K Nearest Neighbor	91.11	0.9	0.75
Sigmoidal Neural Network	95.62	0.87	0.87

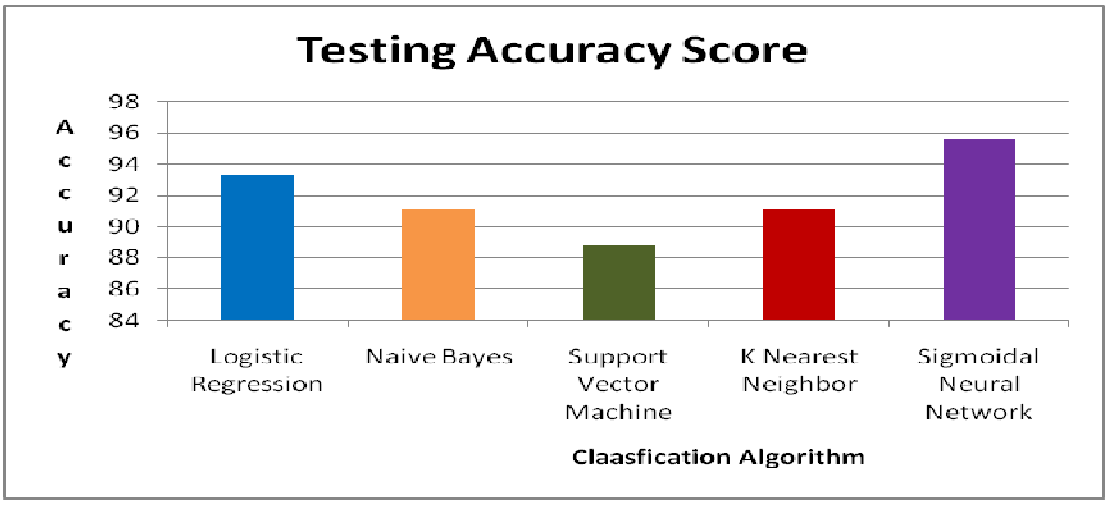

Fig. 4. Comparison graph of the testing accuracy of the proposed SNN with other classifiers.

V. CONCLUSION AND FUTURE WORK

This work has presented how neural network based classification algorithm works better than other existing machine learning classification algorithms like Naïve Bayes, Support Vector Machine, K-Nearest Neighbor (KNN) and random forest. We have also implemented stochastic based neural network which shows improved accuracy than above said algorithm especially during training dataset classification. We have also shown the parameters like precision and recall for both training dataset and testing dataset. In the future our focus will be on large dataset and try to implement other deep learning algorithms like CNN, RNN and many more. It is already know that dimensionality reduction is not needed in case of deep learning classifier but we will focus on dimensionality reduction method on deep learning classifiers.

ACKNOWLEDGEMENT

The authors are grateful to the authorities of Maulana Azad National Institute of Technology, Bhopal, India.

Conflict of Interest. The authors declare no conflict of interest.

REFERENCES

[1]. Cho, Sung-Bae and Won, Hong-Hee (2003). Machine learning in DNA microarray analysis for cancer classification. *In Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics 2003*-Vol. *19* (pp. 189-198). Australian Computer Society, Inc..

[2]. Jain, A. K. and Mao, J. (1996). Artificial Neural Network: A Tutorial," Communications, 1996.

[3]. B. Lewin, (2004). "Gene Expression," *Gene Expr.,* 2004.

[4]. Karakach, T. K., Flight, R. M., Douglas, S. E., & Wentzell, P. D. (2010). An introduction to DNA microarrays for gene expression analysis. *Chemometrics and Intelligent Laboratory Systems, 104*(1), 28-52.

[5]. Singh, R. K., & Sivabalakrishnan, M. (2015). Feature selection of gene expression data for cancer classification: a review. *Procedia Computer Science*, *50*, 52-57.

[6]. Ahmed, M., Shahjaman, M., Rana, M., Mollah, M., & Haque, N. (2017). Robustification of Naïve bayes classifier and its application for microarray gene expression data analysis. *BioMed Research International,* 2017.

[7]. Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning, 46*(1-3), 389-422.

[8]. Mount, D. W., Putnam, C. W., Centouri, S. M., Manziello, A. M., Pandey, R., Garland, L. L., & Martinez, J. D. (2014). Using logistic regression to improve the prognostic value of microarray gene expression data sets: application to early-stage squamous cell carcinoma of the lung and triple negative breast carcinoma. *BMC medical genomics, 7*(1), 33.

[9]. Parry, R. M., Jones, W., Stokes, T. H., Phan, J. H., Moffitt, R. A., Fang, H., ... & Wang, M. D. (2010). k-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction. *The Pharmacogenomics Journal, 10*(4), 292.

[10]. Ding, C., & Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, *3*(02), 185-205.

[11]. Huerta, E. B., Duval, B., & Hao, J. K. (2008). Fuzzy logic for elimination of redundant information of microarray data. *Genomics, proteomics & bioinformatics, 6*(2), 61-73.

[12]. Narayanan, A., Keedwell, E. C., Gamalielsson, J., & Tatineni, S. (2004). Single-layer artificial neural networks for gene expression analysis. *Neurocomputing, 61*, 217-240.

[13]. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., ... & Bloomfield, C. D. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, *286*(5439), 531-537.

[14]. Lee, B., Baek, J., Park, S., & Yoon, S. (2016). deepTarget: end-to-end learning framework for microRNA target prediction using deep recurrent neural networks. *In Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (pp. 434-442). ACM.

[15]. Krishnan, K. G., Vanathi, P. T., Raj, S. S., Nancy, M. and Parveene, S. S. R. (2019). Image classification using deep learning technique. *International Journal on Emerging Technologies*, Vol. *10*, no. 2, pp. 577–590.

[16]. Chen, Y., Li, Y., Narayan, R., Subramanian, A., & Xie, X. (2016). Gene expression inference with deep learning. *Bioinformatics*, *32*(12), 1832-1839.

How to cite this article: Nair, Rajit and Bhagat, Amit (2019). Genes Expression Classification using Improved Deep Learning Method. *International Journal on Emerging Technologies*, **10**(3): 64-68.