Pak. J. Biotechnol. Vol. 16 (2) 101-104 (2019) pISSN: 1812-1837 eISSN 2312-7791 www.pjbt.org
http://doi.org/10.34016/pjbt.2019.16.1.15

ANALYSIS OF GENE EXPRESSION VALUE USING BAT ALGORITHM WITH MULTIFACTOR NON-NEGATIVE MATRIX FACTORIZATION

J. Sumitha

Department of Computer Technology, Dr. SNS Rajalakshmi College of Arts & Science
Saravanampatti, Coimbatore, India. E. mail: sumivenkat2006@gmail.com

Article received 6.5.2019, Revised 10.6.2019, Accepted 17.6.2019

ABSTRACT

The main aim is to identify the disease-affected gene from microarray data and predict the results from the gene expression value. Many computer-assisted algorithms are proposed to investigate a gene is done using machine learning and other bio-inspired algorithms. In this paper, hybridization of Bat algorithm with Multifactor Non-negative Matrix Tri-factorization method are proposed and compared with Multifactor Non-negative Matrix Tri-factorization method to estimate the efficiency of hybridization of bio-inspired algorithm to prove its optimization. The proposed Multifactor Non-negative Matrix Tri-factorization method with Bat algorithm is used to optimize these predictive results and proven its effectiveness and efficiency in detecting the disease-causing gene than ever before.

INTRODUCTION

Gene expression in microarray data has been used for detecting genes affected diseases to be found in the DNA of human body. Inherited diseases have been affecting human beings to a greater extent and many databases related to this exist on the web for research (Sumitha, 2017). The dataset used for this research is Wisconsin dataset and the software which is used for this research is Matlab. This dataset consists of the information which is related to the patients who are affected by breast cancer with gene ID (Sumitha, 2017) and the results are predicted based on this gene value. There are two types of breast cancer datasets: one is diagnostic cancer and another is prognostic (Sumitha, 2016). The breast cancer dataset used for this research is a diagnostic dataset and the performance metrics can be calculated on the basis of confusion matrix method in terms of accuracy, precision, recall and F- measure (Guha, 2016). The main objective is to detect the disease-affected gene using an optimized algorithm. In this paper, hybridization of Multifactor Non-negative Matrix Tri-factorization method with Bat algorithm is developed to optimize the performance of the other existing algorithm prevailed in this field.

The aim of this paper is to detect the disease-affected gene on top of the gene expression value and predict the performance of the optimized Bat algorithm with Multifactor Non-negative Matrix Tri-Factorization over the data. This paper is organized as follows: Section 2 presents proposed methods to solve the task of identifying the cancer-causing gene. Section 3 contains results obtained and discussions. Finally, section 4 contains conclusions.

METHODOLOGY

Dataset Description: Wisconsin breast cancer dataset taken from the URI repository had 567 data with gene ID or gene expression value i.e., Input Data. Among this, 378 data are taken as training data and 189 data are taken as testing data. But this performance metrics for the classifiers in disease identification (Del, 2015a) is on the basis of confusion matrix to scrutinize the performance of the proposed algorithm have been computed from this dataset. The number of attributes taken for predicting the efficiency from this dataset is thirty-four and Matlab is the software that has been used for implementing this work. The UCI Machine Learning repository link of this prognostic type of breast cancer
dataset: http://archive.ics.uci.edu/ml/datasets/breast cancer+wisconsin+%28.

Multifactor Non-negative Matrix Factorization (MNMF) without Bat algorithm: The proposed MNMF with Bat is used to detect the cancer gene in the human body when non matrix factorization is tried with two or more number of factors, it is termed as Multifactor Non negative matrix factorization (Del, 2015b). The goal of non-negative matrix factorization (NMF) is to approximate a non negative matrix V with the product of two non negative matrices, as V $\approx$ W1W2.

This work addressed that the multi-factor NMF (MNMF) problem, where a nonnegative matrix V is approximated with the product of K $\geq$ 2, non-negative matrices, V$\approx \prod_{k=1}^{K} W_k$. It has been argued that using more factors in NMF can imp-rove the algorithm's stability, especially in the case of

ill-conditioned and badly scaled data (Jiho, 2010). The pseudo code for Multifactor Non-negative matrix factorization is depicted in figure-1. MNMF seeks optimal factors that minimize the total difference between V and $\prod_{k=1}^{K} W_k$.

```
[A,Y] = nmf(traindata,k);
% Then clustering.
cls = kmeansclust(traindata,k,Y);
disp('NMF result')
disp(cls)
P = traindata';
T = cls'; S1=10;
% Create Network
net = newelm(P,T,S1);
% Train Network
Pseq = con2seq (P);Tseq = con2seq(T);
%  net.trainParam.epochs = 50;
net = train(net,Pseq,Tseq);
disp(' training structure')
disp(net)
Pseq = con2seq (testdata');
y = sim(net,Pseq);
y = cell2mat(Y); y = round (double(y));
y (y == 0) = 1;
```

Figure-1: Pseudocode for Multifactor Non-negative Matrix Factorization

MNMF with BAT: This algorithm determines multi-variant approach over the data and predicts the value of the gene (Del, 2015a). The multi-variant may be more than one factor taken as a predictive element for processing the results (Mishra, 2012). It is evaluated with multiple factors that correlate the efficiency and takes more than one factor to predict results in the dataset other than a single factor in the dataset.

Algorithm: Multi-label Nonnegative Matrix Factorization (MNMF)

1: Input: Nonnegative matrix X and binary label matrix Y; Weighting parameter for label correlation Ω; the number of bases J;

2: Output: Non –negative matrices U and S minimizing ||X-USY||2 F + Ωtr (SLST);

3: Initialize U and S by random positive values;

4: $U = U * \frac{XYTST}{USYYTST}$

5: $S = S * \frac{UTXYT + \Omega S}{UTUSYYT + \Omega SD}$

6: repeat 4 and 5.

7: until convergence criterion met results are evalua ted with hybridization of BAT (Yubao, 016) with MNMF algorithm. The accuracy is estimated by comparing the results of MNMF and MNMF with BAT algorithm.

RESULTS AND DISCUSSION

Figure-2 depicts the results of the MNMF with BAT algorithm which are compared with MNMF without BAT algorithm. It can be observed that the MNMF with BAT algorithm has 95% of accuracy, 0.83 as precision value, 1.00 as recall value and 0.91 as f-measure value showing the greatest efficiency than MNMF without BAT algorithm with 93% of accuracy, 0.81 as precision value, 0.91 as recall value and 0.86 as f-measure value. The results depicted in the Table-1 shows that the bio-inspired bat algorithm is a good optimizer and also it is hybridized with other algorithms, the predicted results are gets optimized. Hence, Bat is considered to be as a good optimizing technique among bio-inspired algorithms.

Table-1: Results of MNMF without Bat algorithm and MNMF with Bat algorithm

Algorithms / Parameters	MNMF without Bat (%)	MNMF with Bat (%)
Accuracy	93	95
Precision	0.81	0.83
Recall	0.91	1.00
F-measure	0.86	0.91

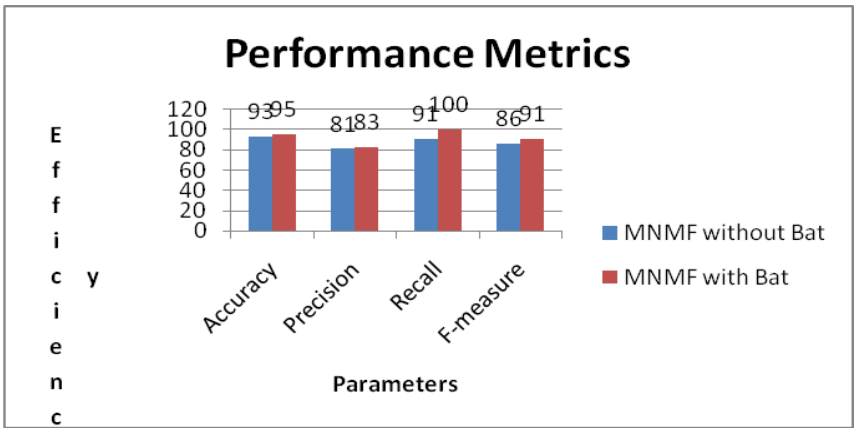

Figure-2: Performance Graph of MNMF without Bat and MNMF with Bat

Multifactor NMF applied to predict breast cancer gene based on Confusion Matrix Method: Multifactor Non-negative Matrix Factorization applied into breast cancer dataset taken from UCI repository contains 567 data. When MNTF is applied, based on the confusion matrix, benign and malignant stages are predicted. The prediction results of Multifactor Non-negative Tri-matrix Factorization algorithm are represented in the table.2.

Table-2: Prediction by Multifactor NMF

N= 189 data		Predicted	
		Benign	Malignant
Actual	Benign	39.0	9.0
	Malignant	4.0	137.0

True Positive (TP)= 39.0, False Positive (FP)= 9.0, False Negative (FN)=4.0, True Negative (TN)=137.0

Precision=TP/(TP+FP)=39/(39+9) = 39/48=0.81

Recall=TP/(TP+FN)=39/(39+4)=39/43= 0.906= 0.91

Accuracy= (TP+TN)/(TP+TN+FP+FN)=(39+137) /(189)=179/189=0.931= 0.931 x 100 (when it is converted into percentage)= 93.1% =93%

F-measure=2 x ((Precision * Recall) / (Precision + Recall)) = 2 x ((0.81 * 0.91)/ (0.81+ 0.91)) = 2 x ((0.7371)/ (1.72)) = 2 x 0.4285=0.857 = 0.

The true positive, false positive, false negative, and true negative values are predicted and it is used for calculating the results of the confusion matrix.

Bat with Multifactor NMF applied to predict breast cancer gene based on Confusion Matrix Method: Bat with Multifactor NMF applied into breast cancer dataset taken from UCI repository contains 567 data. When it is applied, based on the confusion matrix, benign and malignant stages are predicted. The prediction results of Bat with Multifactor Non-negative Tri-matrix Factorization algorithm are represented in the table.3.

The results revealed a truth that without bat algorithm, the efficiency becomes less when compared with using along bat algorithm. Hence, Bat algorithm proves that it is a good optimizer when it is used along with bio-inspired algorithms whereas with machine learning algorithms.

Table-3: Prediction by Bat with MNMF

N= 189 data		Predicted	
		Benign	Malignant
Actual	Benign	43.0	9.0
	Malignant	0.0	137.0

True Positive (TP) = 43.0, False Positive (FP) = 9.0, False Negative (FN)=0.0, True Negative (TN) =137.0

Precision=TP/ (TP+FP)=43/(43+9)=43/52=0.826= 0.83

Recall=TP/(TP+FN)= 43/ (43+0) = 43/43 = 1.00

Accuracy= (TP+TN)/ (TP+TN+FP+FN)= (43+ 137) /(189)=0.952 = 0.952 x 100 (when it is converted into percentage) = 95.2% =95%

F-measure = 2 x ((Precision * Recall) / (Precision + Recall)) = 2 x ((0.83)/ (1.83)) = 2 x 0.4536=0.907=0.91

The predicted values of true positive, true negative, false positive and the false negative are used to predict the efficiency of the confusion matrix parameters.

CONCLUSION

After proposing MNMF with bat algorithm over the cancer dataset, the performance of efficiency has been noticed between the MNMF without Bat and MNTF with Bat. The results proved that this hybridized MNMF with Bat algorithm gives better efficiency than the other without using bat on investigating cancer gene. Hence, Bat algorithm proved in itself as the best optimizer on investigating the gene. In the

future, it can enhance with a newly developed algorithm or converge with a new algorithm.

REFERENCES

Del Buono N., Pio G., Non-negative Matrix Tri-Factorization for Co-clustering: An Analysis of the Block Matrix. *Information Sciences* 301(11):13–26 (2015a).

Del N. Buono, Melisew Tefera Belachew, Addolorata Salvatore, NMF-based Algorithms for Data Mining and Analysis: Feature Extraction, Clustering, and Maximum Clique Finding. Wroclaw University of Technology, Poland (2015b).

Guha S., Meyerson A., Mishra N., Motwani R., and O'Callaghan, Clustering Data Streams: Theory and Practice. IEEE T Knowl Data En. 13(4): 88-92 (2016)

Jiho Yoo, Seungjin Choi, Nonnegative Matrix Factorization with Orthogonality Constraints. J. Comp. Sci. and Eng 4: 97-109 (2010)

Mishra S., Shaw K., Mishra D., A New Meta-Heuristic Bat Inspired Classification Approach for Microarray Data, Procedia Technology 4(1): 802–806 (2012).

Sumitha J., Devi T., Breast Cancer Diagonosis in Analysis of BRCA Gene Using Machine Learning Algorithms. Pak. J. Biotechnol. 13(4): 231-235 (2016).

Sumitha J., Devi T., Analysis of Expression Level of Breast Cancer Gene Using Machine Learning Algorithms for Diagnosis of Breast Cancer. Int. J. Pharm. Bio. Sci. 8(1): 79 – 85 (2017).

Sumitha J., Devi T., Ravi D., Comparative Study on Gene Expression for Detecting Diseases Using Optimized Algorithm. Int. J. Hum. Genet. 17(1): 38-42 (2017)

Sumitha J., BRCA gene expression level analysis for identification of breast cancer using computer assisted algorithms. Int. J. Pharm. Bio. Sci. 9(4): (B) 60-64 (2018).

Yubao Liu, Improved Bat algorithm for reliability-redundancy allocation problems. International Journal of Security and Its Applications 10(2): 1-12 (2016).