Pak. J. Biotechnol. Vol. 16 (2) 115-118 (2019) pISSN: 1812-1837 eISSN 2312-7791 www.pjbt.org
http://doi.org/10.34016/pjbt.2019.16.1.17

ANALYSIS OF GENE EXPRESSION VALUE USING BIO-INSPIRED ALGORITHMS

J. Sumitha

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

Article received 3.5.2019, Revised 11.6.2016, Accepted 18.2019

ABSTRACT:

Breast cancer is one of the most common and rigorous cancers among women and continues to be a health problem all over the world. Due to genetic mutations and changes in the life styles, food habits, radiations, BRCA gene become oncogene. In this paper, Orthogonal Non-Negative Tri-factorization method and Bat algorithm was proposed for finding the disease – causing BRCA gene. Mostly Bat algorithm is used for optimizing the efficiency of the other bio-inspired algorithms. These algorithms are compared with each other to determine the efficiency in detecting the diseases from gene expression value. The results proved that the Bat algorithm performs better than Orthogonal Non-Negative Tri-factorization and other Bio-inspired algorithms than ever before.

KEYWORDS: Bat, Breast Cancer, Confusion matrix, Identification of BRCA gene, MNTF, ONMTF.

INTRODUCTION

Cancer is the second leading cause of death with more than 1,500 people per day being affected by this disease (Carl and John 2013). An average of 1,479,350 new cancer cases is being diagnosed all over the world (Erin 2014). Breast cancer is a leading type of cancer in the world (Sumitha and Devi 2016). Approximately 182,000 new cases of breast cancer are being diagnosed and 46,000 women are estimated to die due to breast cancer each year (Carl and John 2013). Nearly 192,370 new cases of invasive breast cancer have been diagnosed among women and thus, the incidence and mortality of breast cancer are very high (Erin 2014, Laura 2012). Early detection is the first significant step towards treating breast cancer and the most common screening methods such as mammography and sonography that doctors use to detect, diagnose and evaluate breast cancer (Guha et al., 2016). This technique has been in use for about forty years and has some disadvantages in detecting breast cancer (Sumitha and Devi 2016). Such as gives false positive results i.e., diagnosing a benign tumor as cancer, it is not accurate in detecting breast cancer. Hence, it is essential to develop software which could give reliable diagnostic results in prevent these drawbacks.

The objective of this paper is to detect BRCA gene with the help of gene expression value using machine learning algorithms. The algorithms which used in this research are the ONMTF and Bat and the prediction is done on the basis of confusion matrix method. This paper is organized as follows: Section 2 presents proposed methods to solve the task of detecting breast cancer. Section 3 contains results obtained and discussions and section 4 for conclusions.

METHODS

The methods used for this research is categorized as the ONMTF and Bat algorithm.

Data set Description: The URI repository is the dataset used in this research in which the diagnostic breast cancer dataset is taken for predicting the stage of the gene. The dataset used to test the algorithm is taken from the website of University of Wisconsin dataset and created by general surgery department and computer science. The experiment is conducted on barcode generated gene expression value which is then given for classification. The breast cancer dataset used in this research is taken from the URI repository with 567 data with a gene expression value shown in fig 1. From this data, 380 data are taken as training data while the remaining is taken as testing data. 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+winconsin+%28.

Accuracy, precision, recall and f-measure are the parameters used for predicting efficiency. But the performance measure for these classifiers in identifying disease is calculated on the basis of confusion matrix method, which has been computed from this breast cancer dataset.

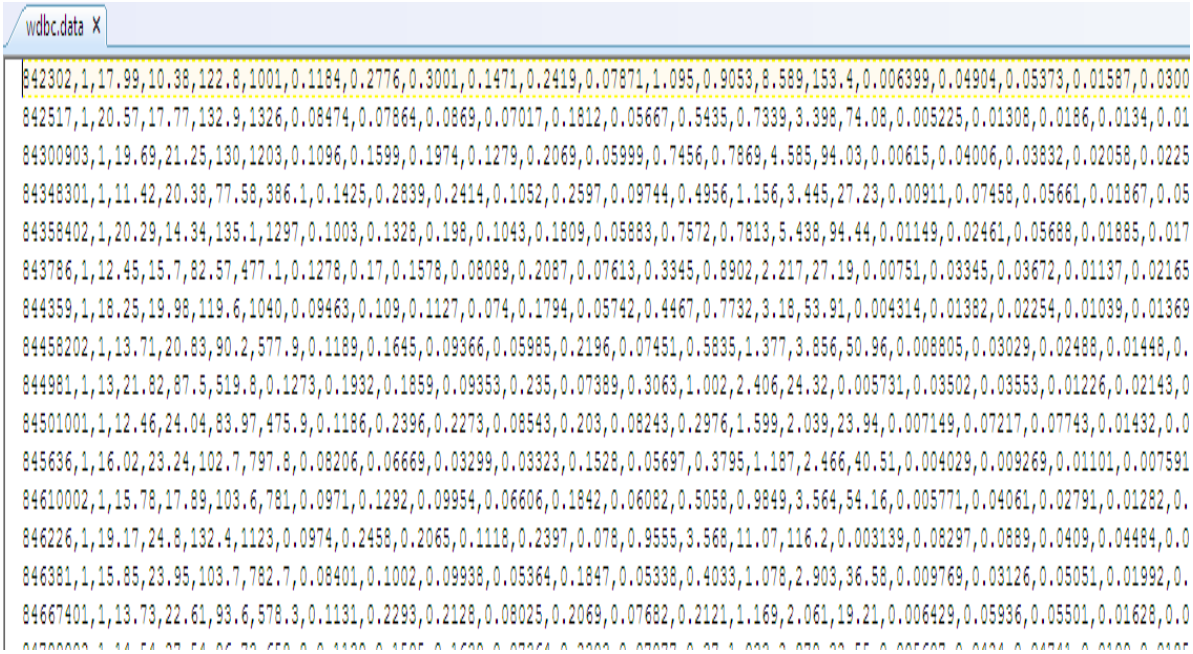

Figure-1: Breast Cancer Dataset

ONMTF: The rapid growth of the Internet and recent technologies have brought data connecting objects of various types that are associated to each other, known as Multi-Type Relational data (Jiho and Seungjin 2010). Traditional clustering methods for single-type data rarely work well on them, thus calling for new clustering techniques, known as high-order co-clustering (HOCC) to simultaneously deal with the multiple types of data. A major challenge in developing HOCC methods including both inter-type and intra-type relationships are effectively using the available information contained in a multi-type relational data set. Meanwhile, because many real world data sets are often large, clustering methods with computationally-efficient solution algorithms are of great practical interest. The Orthogonal Nonnegative Matrix Tri-factorization (O-NMTF) for simultaneous clustering of multi-type relational data employs nonnegative matrix tri-factorization (NMTF) to simultaneously cluster different types of data using the inter-type relationships and incorporates intra-type information through manifold regularization, where, the importance of the orthogonal bind of the factor matrices of NMTF is emphasized (Guha et al., 2016, Jiho and Seungjin 2010). The example for matrix factorization-based methods is that to carry out document clustering and co-clustering words, the documents provide a term-document matrix.

The model of this orthogonal NMTF (ONMTF) algorithm is $X \approx USV^T$ and the update rule of the algorithm is $U \leftarrow U * XVS^T / USV^TX^TU$, $V \leftarrow V*X^TUS/VS^TU^TXV$, $S \leftarrow S*U^TXV/U^TUSV^TV$, where U is the row-coefficient matrix, V is the column-coefficient matrix and S is the block value matrix (Jiho and Seungjin 2010). Given X, a term-document matrix, factor matrices are determined by decomposition. Combining the first two factor matrices or the last two factor matrices relates the 3-factor decomposition to the 2-factor decomposition (Ingrid et al., 2002). While the column vectors in US are associated with basis vectors for the column space of X, the row vectors in SV^T correspond to the basis vectors for the row space of X. The factor matrices U and V denote the degrees of the rows and columns of X are associated with their clusters (Jiho and Seungjin 2010). While U_{il} corresponds to the degree to which term t_i belongs to term-cluster b_l, V_{jk} corresponds to the degree document d_j is associated with document-cluster c_k.

Bat: It is a metaheuristic algorithm used for optimization and was motivated by the echolocation performance of microbats, with variable pulse rates of emission and loudness (Komarasamy and Wahi 2012). Bat algorithm (BA) is a genetic algorithm enhanced and also BA has been found out to be better efficient (Minseung and Sung-Hou, 2015). Since this is a revolutionary function, bat algorithm (BA) became dependant on the echolocation popular functions of microbats and Bat set of rules make use of a frequency-tuning approach to enhance the diversity of the treatments (Komarasamy and Wahi 2012, Minseung and Sung-Hou, 2015). The steps for Bat algorithm are depicted in fig 2.

Step 1:Initialization.Set the generation counter t=1 and set pulse rate. The population of n bats are initialized at random and each bat equivalent to a possible solution to the given crisis; identify loudness, pulse frequency, the initial velocities V_i (i= 1, 2, . . ,N).

```
Step 2: Repeat. Create new solutions by regulating frequency, and changing velocities
        and solutions.
          If (rand > pulse rate) then
             Select a solution among the best solutions;
                Produce a confined solution approximately to the preferred the best solution;
          end if
                Generate a new solution by flying randomly;
                If the new solutions are accepted then
                Increase pulse rate and reduce loudness;
          end if
                Position the bats and search the existing best x*;
                t=t+1;
Step 3:  Iterate the loop until the termination criteria is not satisfied.
Step 4: Post-process the results and visualize.
```

Figure 2: Steps for Bat Algorithm

It takes only a few seconds to complete 1000 iterations and generates high performance than other algorithms in terms of accuracy, precision, recall and f-measure.

RESULTS AND DISCUSSION

In this research, the ONMTF and the bio-inspired algorithm bat are applied to analyze the prognostic breast cancer dataset for detecting diseases are shown in the Table-1. Bat algorithm gives better results when compared to the algorithm because of its parameter control, frequency tuning and automatic zooming. It is also applied for Multifactor Non-Negative Matrix Factorization (MNF) method other than ONMTF. The performance can be calculated in terms of accuracy, precision, recall and F-Measure. The criterion for these classifiers in disease detection is based on Confusion matrix. The accuracy percentage of the ONMTF and Bat is 87 and 89 respectively as depicted in Table-1. The results show that the bat algorithm gives percentage 89 percentage of accuracy, 71 percentage of precision and 91 percentage of recall and 80 percentage of F-measure of which is higher than the ONMTF. Hence, Bat is best in predicting the BRCA gene in the dataset. The graph plotted for these results are shown in Figure-3.

The ONMTF and Bat algorithm has been proposed for optimizing performance and the performance parameters of the algorithm are calculated using confusion matrix (Gabere, 2016). Accuracy is the percentage of correct predictions from the dataset and is proportional to the total number of predictions that were correct. It can be Calculated by means of the following equation using the formula given in equation 1,

Accuracy=a1+a4/(a1+a2+a3+a4)...(equ.1)

calculated by means of the following equation using the formula given in equation 1,
Accuracy=a1+a4/(a1+a2+a3+a4)...(equ.1)

The recall is the intention of positive cases that are appropriately identified, as estimated using the formula given in equation 2,

Recall=a3/(a3+a4).....(equ.2)

Precision (P) refers to the proposition of the predicted positive cases that were correct, as evaluvated using the equation given in equation 3,

Precision=a4/(a2+a4).....(equ.3)

The f-measure (F) can be calculated using the formula given in equation -4,

F=2*(Precision*Recall)/(Precision+ Recall)….(equ.4)

Where a1 is the number of correctly predicted genes which are not disease causing gene, a2 is the number of incorrectly predicted which are diseases causing gene, a3 is the number of incorrect prediction of non –diseased gene and a4 is the number of correct predictions of disease causing genes.

In this research, ONMTF is compared with Bat and predicted the efficiency in terms of accuracy, precision, recall and f-measure. The efficiency of ONMTF exposed that the accuracy percentage of 87, precision percentage of 66, recall percentage of 88 and f-measure percentage of 75, whereas the efficiency of Bat exposed that the accuracy percentage of 89, precision percentage of 71, recall percentage of 91 and f-measure percentage of 80. The results predicted that the efficiency metrics measured for Bat algorithm gives higher performance than the ONMTF in predicting the cancer- affected gene.

Table-1: Results of ONMTF and Bat algorithm

Algorithms / Parameters	ONMTF (%)	Bat (%)
Accuracy	87	89
Precision	66	71
Recall	88	91
F-measure	75	80

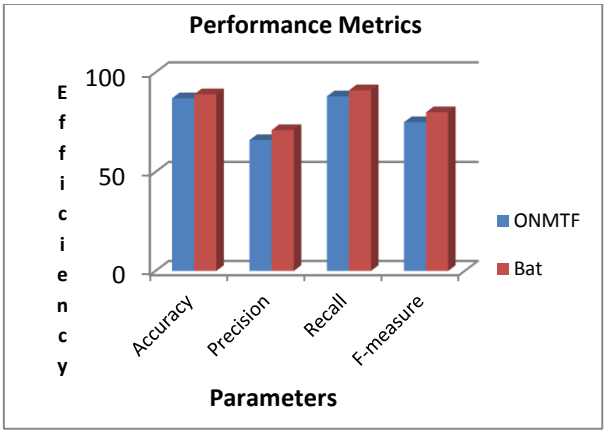

Figure-3: Performance graph of ONMTF and Bat Algorithms

Hence, this algorithm is selected for this research and predicted the accuracy of 89 percentage which is the highest performance than other classifiers in diagnosing of breast cancer. Machine learning algo -rithm is most commonly used in all the research related to this work, but bio-inspired algorithm bat is used specifically in this research to show the performance of bio-inspired algorithm in cancer research. Table-1 concludes that ONMTF predicts high efficiency than Bat. Based on the confusion matrix method, the parameters used for predicting the results are shown in equation 1 to 4.

CONCLUSION

This work is primarily concerned with study and analysis of finding the stage of the breast cancer gene. ONMTF and a bio-inspired algorithm bat have been proposed for predicting its potential and its results have been compared with each other. The efficiency of proposed bat algorithm clearly shows its superiority over the other bio-inspired algorithms.

The algorithms used for predicting the stage of the breast cancer gene were seen to offer better performance as 89% of accuracy, 71% of precision, 91% of recall and 80% of f-measure respectively. The values offered are much higher value when compared with those obtained by other bio-inspired algorithms. Further, this Bat algorithm is a better optimizing algorithm when compared to the other bio-inspired algorithms. In the future, it can either be enhanced with machine learning algo-rithms also to predict the disease causing gene.

REFERENCES

Carl H. Mooney and John F. Roddick, Sequential Pattern Mining-Approaches and Algorithms. *ACM Computing Surveys* (2013)

Erin L. Linnenbringer, Social Constructions, Biological Implications: A Structural Examination of Racial Disparities In Breast Cancer Subtype, University of Michigan (2014)

Gabere, M.N., Algorithm are calculated using confusion matrix (2016)

Guha S., A.Meyerson, N.Mishra, R.Motwani and O'Callaghan, Clustering Data Streams: Theory and Practice. *IEEE Transactions* 15(3): 515-528 (2003) http://dx.doi.org/10.1109/TKDE. 2003.119 8387.

Ingrid A. Hedenfalk, Markus Ringner, Jerey M. Trent and Ake Borg, Gene Expression in Inherited Breast Cancer. *Advances in Cancer Research* 4: 1-34 (2002)

Jiho Yoo and Seungjin Choi, Nonnegative Matrix Factorization with Orthogonality Constraints, *Journal of Computing Science and Engineering* 4: 97-109 (2010)

Komarasamy G. and A. Wahi, An Optimized K-Means Clustering Technique Using Bat Algorithm, *European Journal of Scientific Research* 84(2): 263-273 (2012).

Minseung K. and K. Sung-Hou, Empirical prediction of genomic susceptibilities for multiple cancer classes. *PNAS* 11(5): 1921–1926 (2015)

Sumitha J. and T. Devi, Breast Cancer Diagonosis in Analysis of BRCA Gene Using Machine Learning lgorithms. *Pak. J. Biotechnol. 13(4):* 231-235 (2014)