International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August 2019

Biclustering Gene Expression Data Using Genetic Simulated Annealing Algorithm

M. Ramkumar, G. Nanthakumar

ABSTRACT--- DNA microarray technology produces gene expression matrix that consists of an inexorably missing entries due to poor experimental procedures. The missing values are predicted in the matrix for gene expression data are considered to be essential, since most algorithms analyse the gene expression that usually needs a matrix without missing values. In order to address this issue, the present study biclustering Genetic based Simulated Annealing (Genetic SA) algorithm to predict the items that are missing in the gene expression data. The present study uses biclustering method that is considered to be essential for clustering the gene expression data. The performance evaluation shows that the proposed Genetic SA for gene data expression predicts the missing items in an accurate manner than the existing methods.

Keywords: Gene Expression Data, Biclustering Algorithm.

I. INTRODUCTION

In biological and biomedical research, microarray technology has become a key role [1]. The development in microarray technology has led to a revolution in gene testing and it allows large data volumes to be analyzed under various conditions.

The clustering is one of the main approaches to analyzing the gene data expression. Indeed, it was considered as a first method in several experimental conditions to react transcriptionally to genes. Methods of clustering were recognized in several contexts for their success. These techniques are not, however, targeted at finding a subset of genes correlated in a sub-set of conditions. Moreover, more than one cluster cannot be assigned to a given gene [2] [16]. In addition, a subset of genes can only conduct in certain circumstances when independent behavior is displayed in other conditions [3] [19].

Hartigan introduced biclustering in the seventies in order to avoid some of the cluster inconveniences [4]. This technique was first applied to data from Cheng and Church on gene expression [5]. The biclustering involves discovering groups of genes that have similar behavior under certain conditions.

Revised Manuscript Received on August 30, 2019.

* Correspondence Author

M.Ramkumar*, Research Scholar, Sri Satya Sai University of Technology & Medical Sciences, Madhya Pradesh, India (Email: ramacumenmec@gmail.com)

Dr.G.Nanthakumar, Associate Professor, Anjalai Ammal Mahalingam Engineering College, Tamilnadu, India (Email: gan_nand@yahoo.com)

The NP-Hard [6] is this problem. The majority of algorithms are therefore based on heuristics for the partial exploration of the combinatorial search area [7], [8], [9], [10], [17], [18].

For the analysis of gene expression data, biological and social networks, collaborative filters, phenotype of growth data, structural genomic variations, texts data, chemical data, among others applications, biclustering was applied. Notwithstanding the importance of biomedical and social tasks, the fundamental reality on how the statistical importance of biclustering solutions can be guaranteed is still not accepted. This is because most of the existing approaches are guided by merit functions to ensure homogeneity of biclusters, but usually do not require sound statistical assessment. The optimization of the homogeneity level is clearly not enough because in the sample data (generally observed for small biclusters) good levels of homogeneity can happen [10].

DNA microarray technology produces gene expression matrix that consists of an inexorably missing entries due to poor experimental procedures. The missing values are predicted in the matrix for gene expression data are considered to be essential, since most algorithms analyse the gene expression that usually needs a matrix without missing values. In order to address this issue, the present study biclustering Genetic based Simulated Annealing (Genetic SA) algorithm to predict the items that are missing in the gene expression data. The present study uses biclustering method that is considered to be essential for clustering the gene expression data.

II. RELATED WORKS

Zhang, Y, et al. [11] presented here two unique features: (i) an average of 82% enhanced efficiency by refactoring, optimizing and transferring the C source code of the QBIC and (ii) extensive functions, including qualitative (discretizing) expression data, query-based bicluster expansion, bicluster expansion, bicluster, and other.

The new biclustering algorithm, known as the evolutional biclustering algorithm (BP-EBA), was proposed by Huang, Q et al. [12]. The first phase concerns the development of columns and rows, and the second phase is the development of biclusters. Both phase interaction ensures reliable search guidance and speeds up the convergence towards good solutions. In addition, a conventional hierarchical clustering (HC) strategy to discover Bicluster seeds is used to initialize the population. The authors have developed an evolutionary search for biclusters in greater depth in parallel, seed-based implementation.

Compared with several popular biclustering algorithms using synthetic datasets and real micro array datasets the performance of the proposed algorithm. Zhang et al. [13] has introduced a new approach called Binary Artificial Fish Swarm algorithm in order to combine traditional artificial fish swarm with binary forms. The study used a fitness function based on linear correlation to find genes with shifting and scaling patterns. The BAFSA-based biclustering algorithm was applied to Mice Protein Expression and biclusters showing excellent performance are found.

A novel sequential interpolation-display method for predicting missing values in gene expression data was introduced in Hossain, A., et al. [14]. This method initially created a bicluster for each missing input by selecting a number of correlated genes and samples to that missing position and then apply the approximation technique based on interpolation on this bicluster. This method begins to calculate the minimum number of missing values from the gene and continues to calculate the already imputed values.

III. GENETIC - SA ALGORITHM

In this section, the solution of biclusters in genetic data expressions is proposed to have an improved genetically simulated annealing algorithm. Many meta-heuristic algorithms have been used in the expression of genetic data, such as the genetic algorithm (GA), the

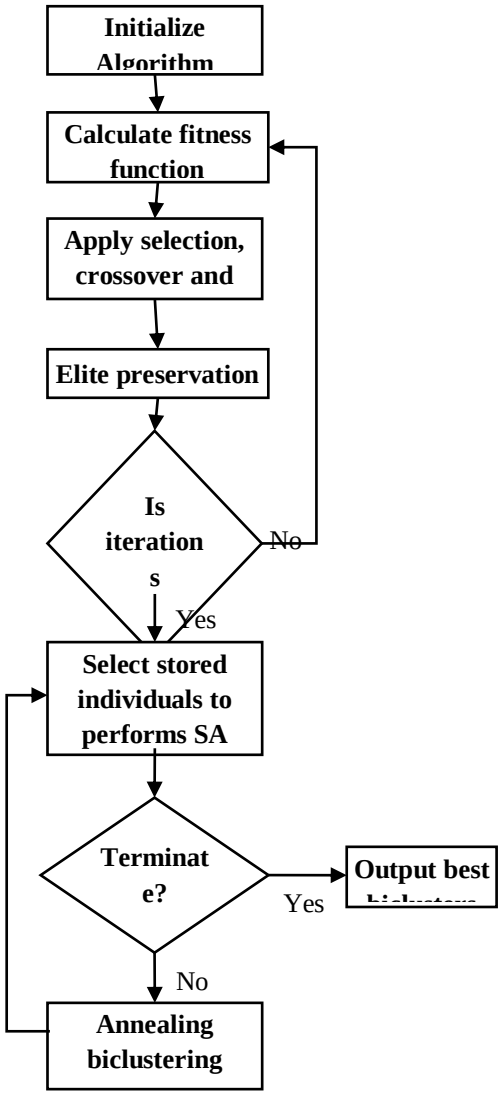

Fig.1. Summary of Genetic-Simulated Annealing Algorithm

simulated-annealing algorithm (SA) etc. The GA solution for optimization can be quickly reached, but it has a fatal deficiency i.e. it is susceptible to a premature convergence in the local context [15].

Fortunately, SA is able to spring from local optimization and seeks the best solution. This paper therefore combines gene algorithm with simulated annealing algorithm. GA is developed to quickly search the solution space for an optimum or near-optimum solution and SA is then used to find the best solution based on the same. In addition, to further increase the efficiency of exploration, an annealing fitness function inspired by the hormone modulation mechanism. In this study, the genetic-simulated annealing algorithm is used for solving the biclustering problem associated with gene data expression, where the architecture of Genetic SA is given in Figure 1.

Genetic Algorithm

The initial population is randomly reported during the GA operation phase. The GA works to produce a new population using basic genetic operations (i.e. selection, crossover and mutation). The following are detailed descriptions of these three operations:

The Selection operation:

The selection operator selects individuals for transmission and mutation based on the individual fitness. It's often not the best fitness value. To determine an appropriate solution space, several selection systems have been developed. A' 2/4 selection' is adopted to preserve the most suitable people at each generation and to preserve the population's diversity. To generate a new population, roulette-wheel selection is used.

The crossover operation:

The combining genes of the selected solution to produce a new solution is legal only if the requirement is satisfied, in accordance with the coding rule. Therefore, a two-point crossover with a crossover probability must be performed randomly in every segment of a chromosome.

The mutation operation:

Given that the crossover process cannot provide new information solutions, a mutation operation with a specified probability is necessary for each segment in order to obtain the solutions with the best fitness. The mutation operation is executed if the specified probability is higher than a random number generated at the 0 to 1 interval using a uniformly distributed rule.

IV. PERFORMANCE EVALUATION & RESULTS

The performance of the proposed Biclustering algorithm is evaluated and tested in order to assess the quality of the extracted bicluster. To test the suggested methods, synthetic data sets are used.

In order to investigate the recovery of biclusters and make the comparison to other biclusters, the synthetic data matrix is used: ISA, SAMBA and CC. The Biclustering Analysis Toolbox (BicaT) is a data analysis software platform that includes all biclustering algorithms. The performance of various biclustering algorithms w.r.t for continuous and additive biclusters is illustrated in Figures 2 and 3. Various algorithms for constant and additive biclusters in Figures 4 and 5 are presented without noise. The results show that the ISA, SAMBA and the approach proposed to identify more than 85% Biclusters in modules not covered by sound in the absence of noise. The proposed method exceeds and maintains a higher percentage in the event of noise than other biclustering methods.

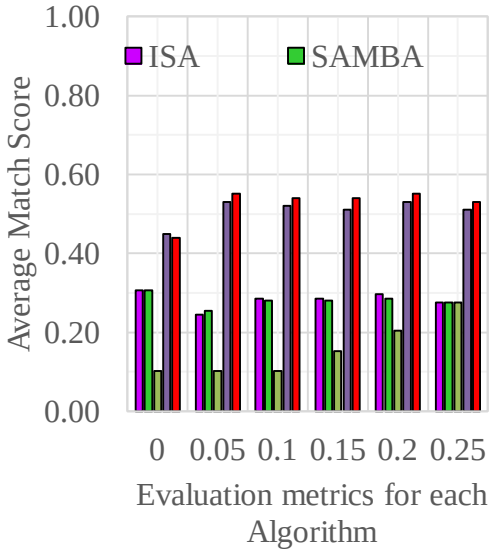

Figure 2: Non-overlapping modules with constant biclusters for increasing noise levels over synthetic datasets

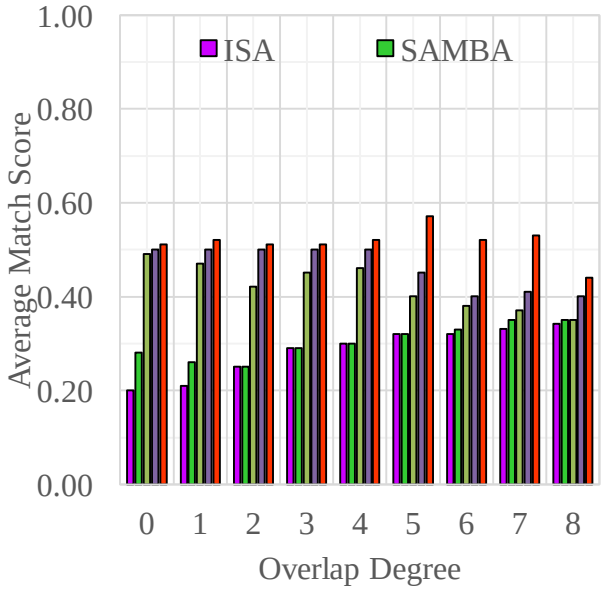

Figure 3: Overlapping modules with constant biclusters for increasing overlap degree over synthetic datasets

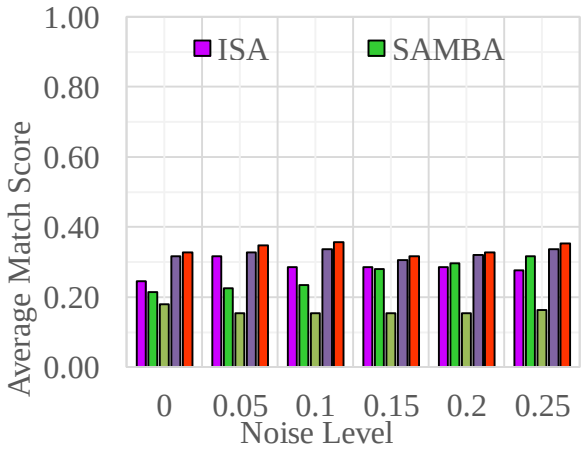

Figure 4: non-overlapping modules with Additive Bicluster for increasing noise levels over synthetic datasets

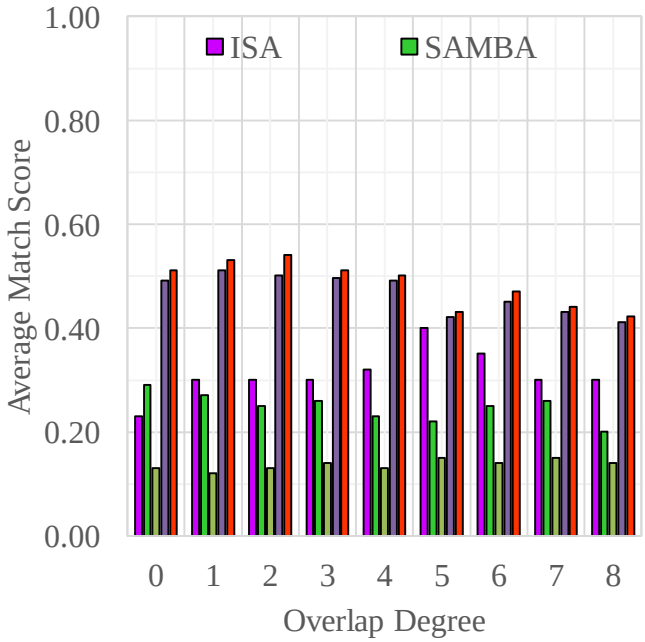

Figure 5: Overlapping modules with Additive Bicluster for increasing overlap degree over synthetic datasets

V. CONCLUSIONS

In this paper, we proposed a biclustering Genetic based Simulated Annealing (Genetic SA) algorithm to predict the items that are missing in the gene expression data. The present study uses biclustering method that is considered to be essential for clustering the gene expression data. The performance of the proposed method is supplied with experiments on synthetic data sets. The test shows that this approach compared to any other biclustering algorithm, competes favorably with this specific task.

REFERENCES

1. Tanay, A., Sharan, R., & Shamir, R. (2002). Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, *18*(suppl_1), S136-S144.
2. de França, F. O., Bezerra, G., & Von Zuben, F. J. (2006, July). New perspectives for the biclustering problem. In *2006 IEEE International Conference on Evolutionary Computation* (pp. 753-760). IEEE.
3. Yip, K. (2003). DB seminar series: biclustering methods for microarray data analysis.
4. Hartigan, J. A. (1972). Direct clustering of a data matrix. *Journal of the american statistical association*, *67*(337), 123-129.
5. Cheng, Y., & Church, G. M. (2000, August). Biclustering of expression data. In *Ismb* (Vol. 8, No. 2000, pp. 93-103).
6. Ayadi, W., Elloumi, M., & Hao, J. K. (2012). BicFinder: a biclustering algorithm for microarray data analysis. *Knowledge and Information Systems*, *30*(2), 341-358.
7. Ayadi, W., Elloumi, M., & Hao, J. K. (2012, December). Pattern-driven neighborhood search for biclustering of microarray data. In *BMC bioinformatics* (Vol. 13, No. 7, p. S11). BioMed Central.
8. Ayadi, W., Elloumi, M., & Hao, J. K. (2009). A biclustering algorithm based on a bicluster enumeration tree: application to DNA microarray data. *BioData mining*, *2*(1), 9.
9. Ayadi, W., Elloumi, M., & Hao, J. K. (2012, December). Pattern-driven neighborhood search for biclustering of microarray data. In *BMC bioinformatics* (Vol. 13, No. 7, p. S11). BioMed Central.
10. Henriques, R., & Madeira, S. C. (2018). BSig: evaluating the statistical significance of biclustering solutions. *Data Mining and Knowledge Discovery*, *32*(1), 124-161.
11. Zhang, Y., Xie, J., Yang, J., Fennell, A., Zhang, C., & Ma, Q. (2016). QUBIC: a bioconductor package for qualitative biclustering analysis of gene co-expression data. *Bioinformatics*, *33*(3), 450-452.
12. Huang, Q., Huang, X., Kong, Z., Li, X., & Tao, D. (2018). Bi-phase evolutionary searching for biclusters in gene expression data. *IEEE Transactions on Evolutionary Computation*.
13. Zhang, R., Gao, H., Liu, Y., Lu, Y., & Cui, Y. (2018, November). Biclustering of Gene Expression Data Based on Binary Artificial Fish Swarm Algorithm. In *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)* (pp. 247-251). IEEE.
14. Hossain, A., Chattopadhyay, M., Chattopadhyay, S., Bose, S., & Das, C. (2017). A bicluster-based sequential interpolation imputation method for estimation of missing values in microarray gene expression data. *Current Bioinformatics*, *12*(2), 118-130.
15. Dai, M., Tang, D., Giret, A., Salido, M. A., & Li, W. D. (2013). Energy-efficient scheduling for a flexible flow shop using an improved genetic-simulated annealing algorithm. *Robotics and Computer-Integrated Manufacturing*, *29*(5), 418-429.
16. Vivekanandan, P., et al. (2013). An efficient SVM based tumor classification with symmetry non-negative matrix factorization using gene expression data. In 2013 International Conference on Information Communication and Embedded Systems (Icices) (pp. 761-768). IEEE.
17. Sivaram, M., et al. "Advanced Expert System Using Particle Swarm Optimization Based Adaptive Network Based Fuzzy Inference System to Diagnose the Physical Constitution of Human Body." International Conference on Emerging Technologies in Computer Engineering. Springer, Singapore, 2019.
18. Raja R., et al., Analysis on Improving the Response Time with PIDSARSA-RAL in ClowdFlows Mining Platform. EAI Endorsed Trans. Energy Web 5.20 (2018): e2.
19. Dhas, C. S. G., et al., (2018). High-performance link-based cluster ensemble approach for categorical data clustering. The Journal of Supercomputing, 1-24.