A Gene Expression Data Biclustering Algorithm Using Large Average Submatrix Based Fcm Classification System

M.Ramkumar, G.Nanthakumar

Abstract:In this paper, the biclusters on data from gene expression tend to group or cluster identical data under multiple conditions on gene expression. Therefore, the biclustering method is very necessary if the matrix lines and columns are clustered instantaneously. At first, the set of sub-matrices are identified using Large Average Submatrix. This is based on a simple significance score which transcends the size and average value of a matrix. Large Average Submatrix is used in an iterative way, where a link between the maximum value and the minimum description length is established. With the total number of data from gene expression growing, the matrix will increase and the clustering problem will be deficient. In this stage, the use of the biclustering algorithm leads to severe problems as data is increased. We are therefore using Large Average Submatrix to improve the biclustering performance. This compresses or removes irrelevant or less correlated ones for improved clustering performance. We also use FCM to verify that for further calculation the number of rows and columns in the submatrix can be added. The method is calculated with regard to consistency of elements and submatrices capacity.

Keywords: Biclustering Algorithm, Gene Expression Data, FCM, Large Average Submatrix

1. INTRODUCTION

Microarray technology has recently played an important role in biomedical and biological research [1]. Due to the development of microarray technology, the research in various genetic studies under various conditions has been improved and this allows a vast amount of data to be analyzed.

The main approach for the investigation of the resulting information is the clustering process, which are among the main techniques considered. The transcriptional reaction of genes was certainly conducted under a few experimental cases. For their performance, genes are detected in some contexts by several clustering techniques. In any case, the subset of genes connected to under few subset conditions is difficult to discover these strategies. In addition, no more clusters of genes are assigned to [2]. Furthermore, in certain environments, some of the subsets of genes offer comparative behaviors offering independent conduct under various other conditions [3].

Revised Manuscript Received on June 10, 2019.

M.Ramkumar, Research Scholar, Sri SatyaSai University of Technology & Medical Sciences, Madhya Pradesh, India (ramacumenmec@gmail.com)

Dr.G.Nanthakumar, Associate Professor, AnjalaiAmmalMahalingam Engineering College, Tamilnadu, India (gan_nand@yahoo.com)

The researchers introduced the process of biclustering [4] and it has been initially used on gene - expression data [3] to reduce the drawbacks associated with the process of gene expression data clustering. The biclustering process involves the identification of the group or clusters of genes with a similar behaviour. This is therefore regarded as an NP – Hard [6]. Several techniques can be applied to this issue and the search space can be explored using heuristic approaches [5].

In this paper, we use Large Average Submatrix with FCM based clustering to find similar genes under particular conditions and the redundancies in large gene data elements is thus eliminated [6].

2. BICLUSTERING

The samples can be defined in various ways and take a variety of forms. The simplest way to identify gene expression data associations is through a multivariate, clustering procedure, independently, of the data matrix lines and columns [6]. The result is a partitioning of the data matrix into non-overlapping rectangular cells when rows and columns of a data matrix are rearranged in order to form each cluster into a contiguous group. Sample variable associations are then searched for cells which, on average, positive or negative [7]. The results can, in some cases, be improved by at the same time clustering samples and variables, a procedure known as co-clustering, in the independent column - row clusters [8].

Independent row - column clustering is now a standard tool for viewing and exploring data from microarrays, but indirectly addressing the sample-variable links finding problem. In contrast, the biclustering methods look straight for or, more precisely, for the U sub-matrices of the X data matrix whose entries satisfy a predetermined criterion. The biclusters are referred to as sub-matrices which fulfill the criterion. It should be noted that a bicluster should not be contiguous by its rows and columns. In the literature on the gene expression dataset, a series of criteria have been studied for the definition of bicluster.

2.1. Large Average Submatrix

This paper provides an important approach for biclustering dataset and evaluates it. We assign a significant value score to each sub-matrix U in the data material using a simple Gaussian null model

for the observed data with the p value corrected by Bonferroni based on the sizes and average U inputs. The Bonferroni correction provides multiple comparisons, which occur when searching for a large average value among many sub-matrices. Furthermore, the correction is a penalty that monitors the size of the sub-matrices found. The motivation for the LAS algorithm [11] is an additive model sub - matrix, in the form of a summary K constant and sub-matrices overlapping with noise.

It is based on normal CDF and sensitive to variations in the empirical distribution of expressive values from normality arising from heavy tails of the LAS score. Outliers can produce sub-matrices with very few samples or variables, although they are highly significant. As a first step in the algorithm, we regard the standard plot for the empirically distributed data matrix entries against the normal standard CDF.

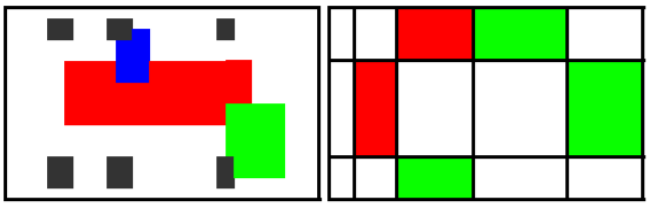


Fig. 1. Left: bicluster overlap and Right: row–column clustering

2.2. FCM Algorithm

In the proposed system, FCM is used to verify for calculation of the number of rows and columns in the submatrix. The conventional cluster algorithms tend to attribute data to a cluster without taking into account the extent to which data is part of a cluster. On the other hand, the fuzzy clustering has established a membership grade which enables each data point to belong to several clusters with different membership degree.

Every cluster is represented by the parameter vector that oscillates in FCM algorithm$\theta_j$where j = 1,2,…,c and c is the total number of clusters. In FCM, the assumption is that a data point from the X dataset does not belong exclusively to a group, but can be part of more than one cluster at a certain degree simultaneously. The u_{ij} variable represents a x_i membership level in the C_j cluster. The data point is more susceptible to the cluster with a higher membership value. In all clusters of a given data point, the total membership value is considered as 1. An additional parameter called fuzzifier$q$ (≥1) (fuzzifier) is used for the algorithm. The value preferable of the fuzzifier unit is considered as 2. However, then the study observe the difference with various other values. The higher the q value is, the lesser is the generalization of the FCM algorithm.

FCM algorithm is formed from the cost minimization function, which is expressed as follows [9] [10:

$$J(\theta,U)=\sum_{i=1}^{c}\sum_{j=1}^{c}u_{ij}^{q}\left\|x_i-\theta_j\right\|^2 \tag{1}$$

FCM algorithm is considered as the most popular algorithm, which is considered as an iterative process and it has got some initial estimates. The following are the important steps used in the iteration process:

The membership degree, u_{ij} of an image x_j is represented in terms of a cluster C_j, i=1,2,…,N, and j=1,2,…,c, which is computed using Euclidean distance of x_i over all θ_j'.

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{d\left(\vec{x}_i,\vec{\theta}_j\right)}{d\left(\vec{x}_i,\vec{\theta}_k\right)}\right)^{\left(\frac{1}{q-1}\right)}} \tag{2}$$

Then θ_j is considered as a representative, which is updated at regular instance based on the weighted means of the image vectors.

$$\theta_j=\frac{\sum_{i=1}^{n}\left(u_{ij}\right)^q\vec{x}_i}{\sum_{i=1}^{n}\left(u_{ij}\right)^q} \tag{3}$$

A number of methods can be used to terminate the FCM algorithm. The algorithm can be terminated if there is a small difference in values between θ_j or membership grade between two successive iterations. However, there are predetermined numbers of iterations.

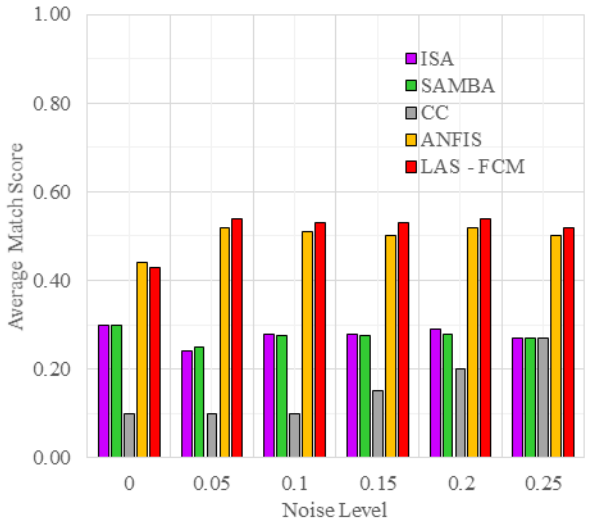
The FCM algorithm is considered to be sensitive during the outlier presence, since its requirement is based on following expression:

$$\sum_{j=1}^{m}u_{ij}=1 \tag{4}$$

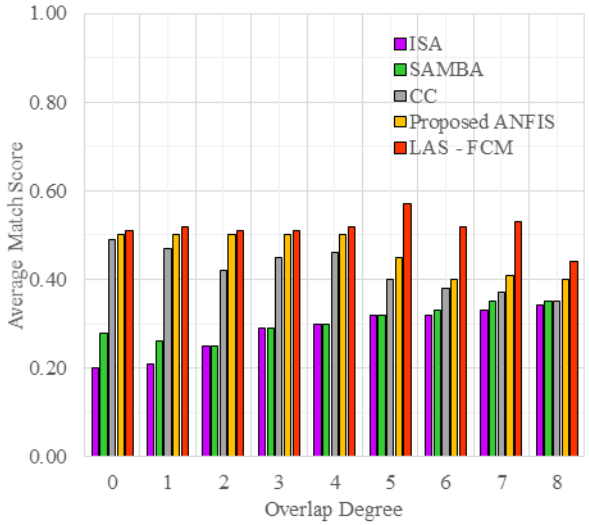
The Eq. (4) is used to represent the noise point, which is to be accounted in order to acquire higher membership grade in a cluster.

3. PERFORMANCE EVALUATION& RESULTS

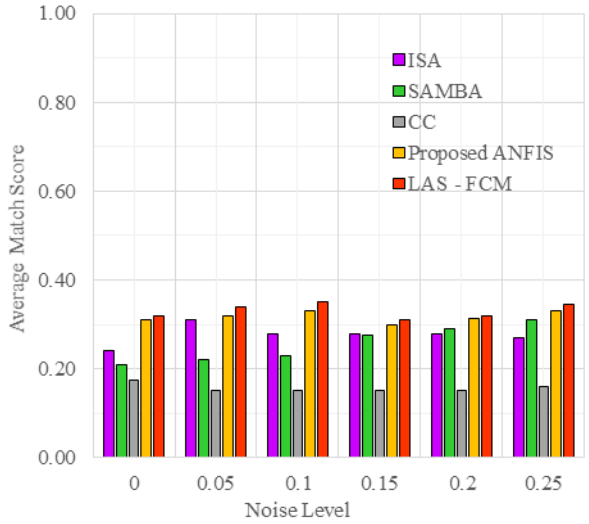
To evaluate the quality of the extracted bicluster, the performance of the proposed Biclustering Algorithm is evaluated and tested. Synthetic datasets are used to test the methods suggested. The synthetic data matrix is used to investigate the capacity recovery of biclusters and to compare it against other biclusters: ISA, SAMBA, and CC. For reference, the Biclustering Analysis Toolbox (BicaT) is a software platform for data analysis through clustering and integrating all biclustering algorithms. Figures 2 and 3 illustrate the performance of different biclustering algorithms w.r.t noise for continuous and additive biclusters. Different biclustering algorithms are presented in figures 4 and 5, without noises, for constant and additive biclusters. The results indicate that, for modules not covered by sound in the absence of noise, the ISA, SAMBA and proposed approach identify more than 85% biclusters than the CC method. The method proposed exceeds and retains a higher percentage than other biclustering methods in the case of noise.



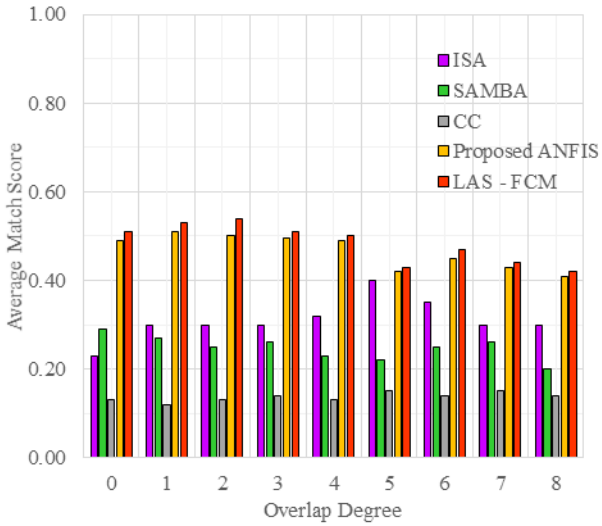
(a) Non-overlapping modules with constant biclusters for increasing noise levels



(b) Overlapping modules with constant biclusters for increasing overlap degree



(c) non-overlapping modules with Additive Bicluster for increasing noise levels



(d) Overlapping modules with Additive Bicluster for increasing overlap degree

Figure 2. Results for the synthetic Datasets

4. CONCLUSIONS

In this paper, we have proposed a Larger Average Submatrix - FCM bi-clustering algorithm to extract microarray biclusters from gene expression datasets. This method is primarily aimed at determining the optimal bicluster of highly correlated genes. Experiments on synthetic data sets are made available to the performance of the proposed method. The experiments demonstrate that the method proposed competes favorably with the given task than any other biclustering algorithm. In future, deep learning methods can improve the quality of the biclustering.

REFERENCES

1. Tanay, A., Sharan, R., & Shamir, R. (2002). Discovering statistically significant biclusters in gene expression data. Bioinformatics, 18(suppl_1), S136-S144.
2. deFrança, F. O., Bezerra, G., & Von Zuben, F. J. (2006, July). New perspectives for the biclustering problem. In Evolutionary Computation, 2006. CEC 2006. IEEE Congress on (pp. 753-760). IEEE.
3. Hartigan, J. A. (1972). Direct clustering of a data matrix. Journal of the american statistical association, 67(337), 123-129.
4. Yip, K. (2003). DB seminar series: Biclustering methods for microarray data analysis, 46–47, 2003.
5. 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.
6. Sancetta, A. (2016). Greedy algorithms for prediction. Bernoulli, 22(2), 1227-1277.
7. Hastie, T., Tibshirani, R., Eisen, M. B., Alizadeh, A., Levy, R., Staudt, L., ...& Brown, P. (2000). 'Gene shaving'as a method for identifying distinct sets of genes with similar expression patterns. *Genome biology*, *1*(2), research0003-1.
8. Weigelt, B., Hu, Z., He, X., Livasy, C., Carey, L. A., Ewend, M. G., ...&van't Veer, L. J. (2005). Molecular portraits and 70-gene prognosis signature are preserved throughout the metastatic process of breast cancer. *Cancer research*, *65*(20), 9155-9158.

9. Kluger, Y., Basri, R., Chang, J. T., & Gerstein, M. (2003). Spectral biclustering of microarray data: coclustering genes and conditions. *Genome research*, *13*(4), 703-716.
10. Yang, M. S., Lin, K. C. R., Liu, H. C., &Lirng, J. F. (2007). Magnetic resonance imaging segmentation techniques using batch-type learning vector quantization algorithms. *Magnetic resonance imaging*, *25*(2), 265-277.
11. Ahsen, Mehmet Eren, Todd P. Boren, Nitin K. Singh, BurookMisganaw, David G. Mutch, Kathleen N. Moore, Floor J. Backes et al. "Sparse feature selection for classification and prediction of metastasis in endometrial cancer." *BMC genomics* 18, no. 3 (2017): 233.
12. Shabalin, A. A., Weigman, V. J., Perou, C. M., & Nobel, A. B. (2009). Finding large average submatrices in high dimensional data. *The Annals of Applied Statistics*, *3*(3), 985-1012.