Greedy Two Way K-Means Clustering For Optimal Coherent Triclsuter

N. Narmadha, R. Rathipriya

Abstract: Generally, a grouping of the data can be classified as three ways i) Grouping of data in one dimension is called as clustering ii) Grouping of data in two-dimension is called as biclustering iii) Grouping of data in three-dimensional is called triclustering. Now- a -days, triclustering is the frequently used data mining technique for analysis of 3D gene expression data. A tricluster of a gene expression dataset is a subset of a gene which exhibits similar expression patterns with a subset of condition along with the time point. In this paper Greedy two way K- Means clustering algorithm for optimal coherent tricluster is performed over every time point. This algorithm is taken as seed to generate the tricluster to identify a coherent pattern based tricluster with high MCV and larger volume. The performance study is carried out to test the proposed algorithm. The results show that proposed algorithm identifies larger volume tricluster with high correlation among genes of 3D dataset.

Index Terms: Triclustering, Greedy Approach, Yeast Cell Cycle data, Gene expression data, Optimal Tricluster, 3D data,Correaltion

1 INTRODUCTION

THE microarray technology is used for the measurement of mRNA levels and to measure the rapid growth of thousands of genes. For the past years, monitoring the gene expression data is very difficult. Now-a-days, the gene expression data comprises of thousands of genes with experimental conditions over time point (e.g., various patients, with their tissue types, and their growth environments), is studied for a single experiment. Microarray gene expression data constructs a data matrix in which genes are represented as rows, conditions are represented as columns and the time series is represented as a time point. To represent each entry in the data matrix shows the expression level of (gi, ci, ti) that is the specific gene represents (gi) under particular condition represents the (ci) along the time point represents the (ti). The analyses of microarray gene expression data the genes (gi) are identified that similar behaviour among a subset of condition (ci) over the time point (ti) is said to be tricluster. In clustering, for analyses of gene expression data doesn't provide the entire relevant gene for the particular conditions. But it extracts the few genes that are relevant to the subset of the condition. In biclustering, both gene and conditions are clustered even it is more difficult than clustering and it is represented in two dimensions. It is also a failure to extract the relevant gene from the subset of conditions. Triclustering is also a data mining technique to extract the relevant genes under a subset of conditions along with the time point. A new Greedy based triclustering approach is devised in this paper using correlation measure to extract highly correlated triclusters. Contributions of this paper are as follows:

- A new Mean Correlation Value (MCV) measure is proposed to find scaling and shifting pattern tricluster from 3D dataset.
- To develop a new Greedy based triclustering approach over 3D data.

- *N. Narmadha , Ph. D Research Scholar, Department of Computer Science, Periyar Univarsity, Salem – 636011. Tamil Nadu, India E-mail: mahanarmadha@gmail.com*
- *Dr. R. Rathipriya, Assistant Professor, Department of Computer Science, Periyar Univarsity, Salem – 636011. Tamil Nadu, India E-mail: rathipriyara@gmail.com*

The rest of paper is organized as follows section 2 describes the literature review for this research work. Methods and materials needed for this research work is provided in section 3. Section 4 elaborates the proposed work for the research work section 5 concludes the proposed work.

2 Literature Review

This section provides an overview of related works in the field of 3D microarray gene expression data analysis, in particular, the work related to the Greedy based clustering and biclustering. To propose various potential modifications to the OAC-triclustering algorithms based on the prime operators (Arnold, 2016). To perform slight modifications based on clustering procedures to optimize the performance of the specialist-generalist using classification system (Gnatyshak, 2014).A pattern-driven local search operator is inbuilt with the binary Particle Swarm Optimization (PSO) algorithm is used to improve the search efficiency (Yangyang Li, 2014).BPSO encoding gene-to-class sensitivity (GCS) mainly used to perform gene selection. GCS is used to extract the samples with the help of extreme learning machine (ELM).ELM, K. nearest neighbour (KNN) and support vector machine (SVM) classifiers are used for prediction with high accuracy for microarray data, it gives the efficiency and effectiveness for gene selection method (FeiHan, 2015). The Multi-Objective Particle Swarm Optimization is used for gene expression data to extract the bicluster. The main purpose of this technique is to cover all elements of the gene expression matrix amongst the overlapping bicluster (Mohsen lashkargir, 2009). Biclustering algorithm is used to identify the coherent bicluster with minimum MSR (Mean Square Residue) and with maximum row variance for gene expression data.

To solve this kind of problem various optimization approaches are used namely 1) Nelder Mead with Levy Flight and 2) Tabu search with Nelder Mead are proposed and compared. NM with Levy Flight shows better performance and it gives a better global optimal solution when compared to Tabu search with Nelder Mead (Kavitha M, 2016). Biclustering algorithm is used to cluster the gene expression data, to improve the residue function for this algorithm. This improved function is more appropriate for the stochastic heuristic algorithm. The parallel genetic algorithm (GA) is

used for biclustering optimization algorithm; it can prevent the local convergence in the optimal algorithm and make the probability for global convergence bigger (Wei Shen, 2012). EDA- GA hybrid is to analyze the gene expression data it not only gives converge quickly but it provides the global solution (Feng Liu, 2006). PSO with GA is used to solve the biclustering problem and it provides high accuracy (Baiyi Xie, 2007). Binary Particle Swarm Optimization (BPSO) is used to retrieve the global optimal bicluster from the web usage data. It provides the relationship between the web users and webpage (R.Rathipriya, 2011). Firstly small disjoint tightly correlated submatrices are generated using the K-Means clustering algorithm. Secondly, the greedy search algorithm mainly used to enlarge the seeds. The output of the greedy search algorithm is used as an initial population of binary PSO, these steps are used to identify bicluster (Shyama Das, 2010). To solve the classification of gene expression data to implement the improved binary particle swarm optimization (IBPSO) for feature selection and K-nearest neighbour (K-NN) as an evaluator of IBPSO. These methods are helpful to reduce the total number of features as required (Li-Yeh Chuang, 2007). The gene expression data clustering K-means, FCM and hierarchical techniques are used for clustering microarray data. But PSO based K-means gives better performance for clustering microarray data (Lopamudra Dey, 2014). Biclustering algorithm used shifting and scaling pattern on the merit function it is mainly used to grow the bicluster. But this measure has its own demerits for identifying scaling pattern and coherent evolution to grow bicluster (K.Thangavel, 2011). Particle Swarm Optimization (PSO) is used for the best subset generation and for evaluating the subset to uses k-means as wrapper algorithm. The algorithm gives the good quality of cluster with an accuracy of 70-80 % (Deepthi P S, 2015). These Greedy techniques are implemented only implemented with the clustering and biclustering techniques. Pros and cons of using greedy approach for clustering and biclustering are stated clearly. Greedy technique is the basic technique to find out feasible solution based on the given criteria. Hence, therefore, Greedy techniques with other method are broadly used for gene expression data analysis and also it is used in the web usage data to extract the quality bicluster. This paper introduces the greedy approach with triclustering technique with coherent pattern is used to find the quality of tricluster.

3 Methods And Materials

This Section expounds the basic concepts for triclustering approach.

3.1 Triclustering Definition

Three dimensional (3D) Microarray dataset is a dataset contains 3 types of variables (gene, sample, and time point). In general, each cell mijk in a 3D dataset represents the value of ith row under jth column at kth time space. It can also be viewed as a two-dimensional matrix, such that each cell mi,j contains the time series with respect to ith row under jth column.

3.2 Types of Tricluster

Triclusters have different patterns. They are:

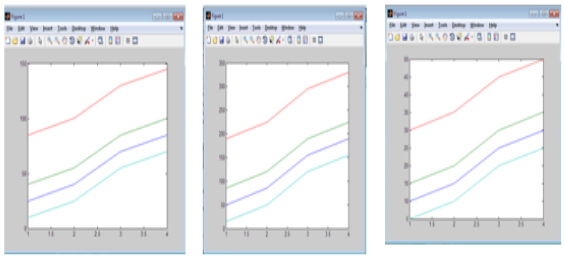
- Tricluster with Additive pattern
- Tricluster with Multiplicative Pattern
- Tricluster with Coherent Pattern
- Tricluster with Coherent Evolution Pattern

Additive Tricluster

g/s	s1	s2	s3	s4
g1	6	7	10	5
g2	7	8	11	6
g3	9	10	13	8
g4	10	11	14	9

g/s	s1	s2	s3	s4
g1	61	62	65	60
g2	62	63	66	61
g3	64	65	68	63
g4	65	66	69	64

g/s	s1	s2	s3	s4
g1	112	112	115	110
g2	112	113	116	111
g3	114	115	118	113
g4	115	116	119	114



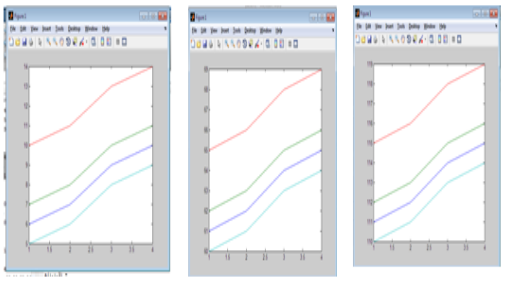
t1 t2 t3

Multiplicative Tricluster

g/s	s1	s2	s3	s4
g1	10	15	30	5
g2	15	20	35	10
g3	25	30	45	20
g4	30	35	50	25

g/s	s1	s2	s3	s4
g1	25	40	85	10
g2	40	55	100	25
g3	70	85	130	55
g4	85	100	145	70

g/s	s1	s2	s3	s4
g1	50	85	190	15
g2	85	120	225	50
g3	155	190	295	120
g4	190	225	330	155



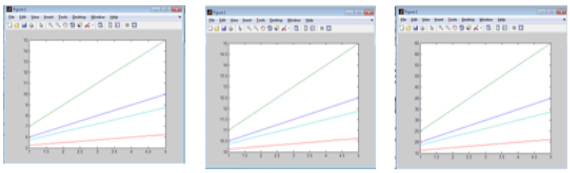
t1 t2 t3

Coherent Tricluster

g/s	s1	s2	s3	s4
g1	6	7	5	6
g2	7	9	6	7
g3	8	11	6	7
g4	9	13	6	8

g/s	s1	s2	s3	s4
g1	11	11	10	10
g2	11	12	10	11
g3	12	13	10	11
g4	12	14	11	12

g/s	s1	s2	s3	s4
g1	20	25	16	19
g2	25	35	18	23
g3	30	45	19	26
g4	35	55	20	30



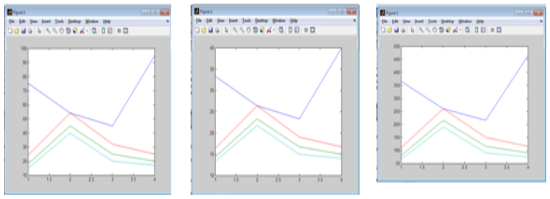
t_1 t_2 t_3

Coherent Evolution Tricluster

g/s	s1	s2	s3	s4
g1	365	80	110	65
g2	260	215	260	190
g3	215	115	150	90
g4	465	90	115	75

g/s	s1	s2	s3	s4
g1	33	14	16	13
g2	26	23	26	22
g3	23	17	19	15
g4	40	15	17	14

g/s	s1	s2	s3	s4
g1	75	18	24	15
g2	54	45	54	40
g3	45	25	32	20
g4	95	20	25	17



t_1 t_2 t_3

There are various types of triclusters are available, but here, greedy based triclustering approach concentrates only on tricluster with the coherent pattern.

3.3 MSR3D Vs. MCV

In this study, the two evaluative measures will be taken for evaluation of tricluster. They are :

i) Three dimensions MSR (MSR3D)

Three dimensions of Mean Square Residue (MSR_{3D}) that measures the homogeneity of triclusters which contain subgroups of genes, conditions, and time points (1,2014). This measure is said to be MSR3D the formulae is given in equaiton 1.

ii) Mean correlation value (MCV)

The range of MCV is from 0 to 1. Value close to '1' signifies high correlated tricluster otherwise low or null correlated tricluster the formula for MCV is shown in equation 4. Therefore, MCV based fitness is designed to extract the optimal tricluster from 3D dataset (Narmadha.N, 2018)

TABLE 1: *MSR_{3D} Vs MCV*

Measures	Descriptions and its Formula
MSR_{3D}	$MSR_{3D}(TC) = \frac{\sum_{g \in G, c \in C, t \in T} r_{gct}^2}{\quad}$ (1) Where r_{gct} can be defined as $r_{gct} = TC_V(g,c,t) + M_{GT}(g) + M_{GC}(T) - M_G(c,t) - M_C(g,t) - M_T(g,c) - M_{GCT}$ (2)
MCV	$\sum_m \sum_n (A_{mn} - \bar{A})(B_{mn} - \bar{B}) \;/\; \left[\left(\sum_m \sum_n (A_{mn} - \bar{A})^2\right)\left(\sum_m \sum_n (B_{mn} - \bar{B})^2\right)\right]$ (4) Where $\bar{A} = \frac{\sum_m \sum_n (A_{mn})}{\quad}$, $\bar{B} = \frac{\sum_m \sum_n (B_{mn})}{\quad}$ The range of MCV is [0 ,1]

The table 1 shows the MSR3D Vs. MCV measure with the description. Most of the literatures are used the MSR3D as a homogeneity measure to evaluate the quality of tricluster.When compared to MSR3D, MCV shows a similar pattern. But in MSR3D has high magnitude with high value than MCV. If MCV value is close to '1' signifies high correlated tricluster otherwise if it is '0' signifies low or null correlated tricluster is shown in table 2. From the study clearly shows the MCV as best evaluation measure so this research work carries the MCV as a correlation measure.

Table 2: *Types of Tricluster with the values of MSR_{3D} Vs MCV*

Tricluster	Additive	Multiplicative	Coherent	Coherent Evolution
MSR_{3D}	$5.0759\ e^{+04}$	$6.1162e^{+07}$	$9.3726\ e^{+03}$	$9.8709\ e^{+07}$
MCV	1	1	1	1

3.4 Seed Generation

During the seed generation step, two way K-means clustering algorithm is applied along the two dimensions of A(G, S, T) to generate k_g and k_s cluster and combined these clusters to get $k_g * k_s$ initial bicluster for every time point 't' in T. These biclusters are encoded as binary string of size $n_b*(nG+nS)$. Figure 1 represents the encoded bicluster and algorithm 2 describes the seed formation using Greedy Approach.

g_1	g_2	.....	g_{nG-1}	g_{nG}	s_1	s_2	.....	s_{nS}

Fig 1: Encoded Bicluster of length nG+nS

3.5 Fitness Function

The main objective of this work is to discover larger volume triclusters with high MCV. The following fitness function F (I, J, K) is used to extract optimal tricluster is shown in equation 3.

$$F(I,J,K) = \begin{cases} |I| * |J| * |K|, & \text{if MCV(Tricluster)} \geq \delta \\ 0, & \text{Otherwise} \end{cases} \quad (3)$$

where |I|, |J|, |K| are number of rows, columns and the time points of Tricluster and MCV threshold 'δ' which range from 0 to 1. In this study, δ is set as 0.91 to 0.95.

3.6 Tricluster Formation

The biclusters generated from the seed generation phase are used to create an initial tricluster population (npop) for Greedy approach. First, to select the top 'npop' bicluster based on their Average Correlation Value from the entire set of biclusters. Secondly, to create initial tricluster population, generate random time point of size n_b*nT and add these points with the 'n_b' biclusters. Figure 2 shows single binary-encoded tricluster A(nG+nS+nT) as binary string of length nG+nS+nT. Algorithm 1 describes the initialization of random tricluster with the size of 50 genes, 16 samples and 6 time points.

g_1	g_2	…..	g_{nG-1}	g_n	s_1	s_2	…..	s_{nS}	t_1	t_2	……	t_{nT}
50					16				6			

Fig 2: Encoded tricluster of length nG+nS+nT

4. Proposed Work

This section provides the proposed work, which discusses the application of Greedy approach algorithm for triclustering of gene expression data.

Algorithm 1: Random Initialization of Tricluster

```
// Random Initialization of Tricluster
data= randi ([0 1], ng,ns,nt)
// r represents the no. of genes
// c represents the no. of samples
// t represents the no. of time points
[r,c,d]=size(data)
size(data) = [ng,ns,nt]
```

Algorithm 2: Seed Formation using Greedy Approach

```
Input      : Initialization of (nG+nS)
Output: Gene Enlargement and Refined tricluster
Step 1: Generate random population using the algorithm 2
Step 2: For each gene
                i) Call gene Enlargement (gene (G', S', T'))
                ii) Call gene Refinement (gene (G', S', T'))
Step 3: Return the Gene Enlargement and Refined tricluster
Step 4:
 // Sub functions of Gene Enlargement and Refined tricluster
Call gene Enlargement (gene (G', S', T'))
Step 1: Set of genes 'g' not in G'
Step 2: Set of samples 's' not in S'
Step 3: Set of time point 't' not in T'
Step 4: For each node g/s/t
          If MCV (union (gene, (g/s/t)))> MCV (gene (G,S,T)) then
              1. Add g/s/t to gene (G,S,T)
              2. End (if)
        End (for)
Step 5: Return Enlarged gene set
Call gene Refinement (gene (G',S', T'))
Step 1: For each node g/s/t in Enlarged gene
          Remove node g/s/t in Enlarged gene
          G''/S''/T' be set of rows /columns/timepoint in G'/S'/T' but not
contained g/s/t
          If MCV  Enlarged  gene (G'',S'', T'')> MCV (Enlarged gene
(G'/S'/T'))
          Update G'/S'/T'
          End (if)
        End (for)
Step 2: Return refined gene set G'',  and A(G'',S',T') as refined
tricluster.
```

Enlargement and Refinement of Tricluster Using Greedy Approach

Algorithm 2 describes the enlargement and refinement of tricluster using greedy approach. The main objective of this algorithm is clarified in step by step (R.Rathipriya, 2011).

- In this step, seeds are enlarged and refined by adding /removing the rows and columns to enlarge their volume and to improve their quality of tricluster.
- The main objective of the greedy search procedure is to maximize the volume of the tricluster without degrading the quality measure.
- Here, MCV is used as excellence function to grow the seeds.
- Insert/Remove the genes from the tricluster if it increases MCV of the tricluster.

TABLE 3: *Characteristics of Tricluster for random population*

Tricluster ID	Threshold value	No. of Genes	No. of Sample	No. of Time point	Volume	MCV
Tricluster 1	0.91	30	11	6	648	0.9122
Tricluster 2	0.92	24	9	5	648	0.9323
Tricluster 3	0.93	32	8	3	648	0.9235
Tricluster 4	0.94	23	7	5	648	0.9423
Tricluster 5	0.95	30	10	4	648	0.9515

The characteristics of the optimal tricluster for each threshold are given in table 3. The volume of the tricluster, number of genes, number of samples, number of time points and their MCV of the optimal tricluster are shown in this table clearly. From the study to analysis the correlation values for each tricluster is various based on the threshold value for the random population. It is clearly proved that when threshold value increases the MCV value are also be increased, at the same time there is no changes in the volume. The correlation coefficients are very high, in most cases the values are close to one. This indicates almost perfect homogeneity between the genes, samples and times points of the tricluster.

TABLE 4: *Overall performance of Tricluster for random population*

Threshold value	No. of Genes	Gene Coverage %	No. of Samples	Sample Coverage %	No. of Time Points	Time point Converge %
0.91	30	60 %	11	68.75 %	6	100 %
0.92	24	48 %	9	56.25 %	5	83.33 %
0.93	32	64 %	8	50 %	3	50 %
0.94	23	46 %	7	43.75 %	5	83.33 %
0.95	30	60 %	10	62.5 %	4	66.66 %

Table 4 depicts that overall performance of tricluster for random population. The table shows the different threshold value from 0.91 to 0.95, no. of genes with gene coverage, no. of samples with sample coverage and no. of time points with time point coverage. Here the no. of genes, no. of samples and no. of time points various based on the threshold value. At the same time gene coverage, sample coverage and time point coverage are also be varies based on the no. of genes, no. of sample, and no. of time points . Gene coverage, sample coverage and time point. coverage is calculated using the formulae is given in equation (4, 5, 6)

$$Gene\ coverge = \frac{No.of\ genes\ extracted}{Total\ no.genes} * 100 \quad (4)$$

$$Sample\ coverge = \frac{No.of\ samples\ extracted}{Total\ no.Sample} * 100 \quad (5)$$

$$Time\ point\ coverge = \frac{No.of\ time\ points\ extracted}{Total\ no.time\ points} * 100 \quad (6)$$

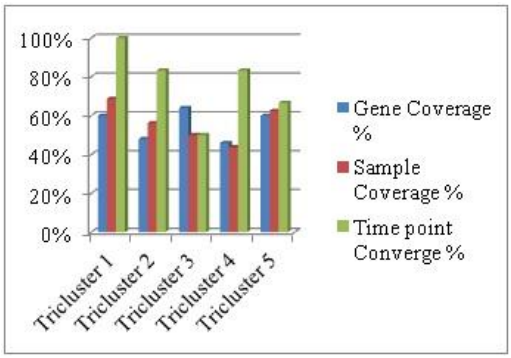


Fig 3: Graphical Representation of Gene coverage, Sample Coverage and Time point coverage

Figure 3 shows the graphical representation of the optimal tricluster for random population with threshold value 0.91 covers the 60 % of genes from gene set G, 68.75% of samples from sample set S and 100% of time points from time set T. The threshold value 0.92 covers 48% of genes from gene set G, 56.25% of samples from sample set S and 83.33% of time points from time set T. The threshold value 0.93 covers 64 % of genes from gene set G, 50 % of samples from sample set S and 50% of time points from time set T. The threshold value 0.94 covers 46% of genes from gene set G, 43.75% of samples from sample set S and 83.33% of time points from time set T. The threshold value 0.95 covers 60% of genes from gene set G, 62.25% of samples from sample set S and 66.66% of time points from time set T respectively.

5 CONCLUSION

In this paper, a new Greedy based triclustering algorithm has proposed to extract the high quality of tricluster with larger volume. The result has shown that the proposed work has performed well to extract the larger volume tricluster with a high coherent pattern from the given 3D dataset.

FUTURE ENHANCEMENT

- It has been observed from the result, the proposed algorithm shows better performance in extracting highly correlated tricluster.
- Thus all the works are empirical study was conducted to test the performance of proposed triclustering algorithm using yeast cell cycle analysis dataset and in bioinformatics
- Bioinformatics is playing an increasingly important role in nearly all aspects of drug discovery, drug assessment and drug development.
- To handles large volumes of data
- Bioinformatics tools to predict, analyze and help interpretation in clinical and preclinical findings.
- Bioinformatics provides a huge support to overcome the cost and time context in various ways.
- CADD - It provides wide range of drug-related databases and software which can be used for various purposes related to drug designing and development process.

ACKNOWLEDGEMENT

The first author gratefully acknowledges financial support from UGC- NFSC (Formerly Rajiv Gandhi) Scheme and its award letter- number: F1–17.1/2017–18/RGNF–2017–18–SC–TAM–29799/(SA–III/Website) and dated : 27.07.2017.

REFERENCES

[1] Arnold, S. (2016). A Greedy Algoirhtm to cluster specialists. arxiv.org,cornell University, 13 Sep .

[2] Baiyi Xie, S. C. (2007). Biclustering of Gene Expression Data Using PSO-GA Hybrid. 1st International Conference on Bioinformatics and Biomedical EngineeringIEEE Xplore.

[3] David Gutiérrez-Avilés and Cristina Rubio-Escudero, Mining 3D Patterns from Gene Expression Temporal Data:

A New Tricluster Evaluation Measure, The Scientific World Journal, Volume 2014, Article ID 624371, 16 pages

[4] Deepthi P S, S. M. (2015). PSO Based Feature Selection for Clustering Gene Expression Data. IEEE.

[5] Fei Han, C.-Q. W.-S.-H.-Q.-S. (2015). A Gene Selection Method for Microarray Data Based on Binary PSO Encoding Gene-to-class Sensitivity Information. IEEE Transactions On Computational Biology And Bioinformatics.

[6] Feng Liu, H. Z. (2006). Biclustering of Gene Expression Data Using EDA-GA Hybrid. IEEE Congress on Evolutionary Computation Sheraton Vancouver Wall Centre Hotel

[7] Gnatyshak, D. V. (2014). GreedY Modifications of OAC-triclustering Algoirthm. Procedia Computer Science , 31, 1116-1123.

[8] K.Thangavel, J., (2011). Novel Hybrid PSO-SA Model for Biclustering of Expression Data. International Conference on Communication Technology and System Design by Elsevier.

[9] Kavitha M, D. (2016). A Hybrid Nelder-Mead Method For Biclustering Of Gene Expression Data. International Journal Of Technology Enhancements And Emerging Engineering Research.

[10] Li-Yeh Chuang, H.-W. C.-J.-H. (2007). Improved binary PSO for feature selection using gene expression data. Computational Biology and Chemistry Elsevier Ltd. All rights reserved.

[11] Lopamudra Dey, A. M. (2014). Microarray Gene Expression Data Clustering using PSO based K-means Algorithm. International Journal of Computer Science and its Applications.

[12] Mohsen lashkargir, S. A. (2009). A Hybrid Multi-Objective Particle Swarm OptimizationMethod to Discover Biclusters in Microarray Data. (IJCSIS) International Journal of Computer Science and Information Security.

[13] Narmadha.N, Rathipriya.R, "Triclustering Algorithm for 3D Gene Expression Data using Correlation Measure", International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 5, Issue 2, February 2018

[14] R.Rathipriya, K. J. (2011). Binary Particle Swarm Optimization based Biclustering of Web usage Data. International Journal of Computer Applications.

[15] R.Rathipriya, K. J. (2011).Evolutionary Biclustering of Clickstream Data, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 1, May 2011

[16] Shyama Das, S. M. (2010). Greedy Search-Binary PSO Hybrid for Biclustering GeneExpression Data. International Journal of Computer Applications.

[17] Wei Shen, G. L. (2012). A Novel Biclustering Algorithm and Its Application inGene Expression Profiles. Journal of Information & Computational Science.

[18] Yangyang Li, M. I. (2014). Biclustering of Gene Expression Data Using Particle Swarm Optimization Integrated with Pattern-Driven Local Search. IEEE Congress on Evolutionary Computation(CEC).