Greedy-Based Triclustering of Genetic Algorithm Using Computational Intelligence Techniques (CIT)

Dr. N. Narmadha,
Assistant Professor,
PG & Research Department of Computer Science,
Sri Sarada College for Women (Autonomous), Salem-16.
mahanarmadha@gmail.com

Dr. R. Rathipriya,
Assistant Professor,
Department of Computer Science,
Periyar University,
Salem – 11.
rathipriyar@gmail.com

Abstract- **Triclustering is a popular data mining technique for three dimensional data (3D) analysis. It can reveal hidden and unknown interesting patterns in a 3D data. The main source of 3D data is microarray data which has wide scope in the bioinformatics field especially in drug analysis. Genetic algorithms is the traditional and robust bioinspired optimization technique for data mining tasks. In this work, GA is used to extract tricluster from 3D microarray data efficiently by exploring a large solution space and find highly coherent triclusters. The proposed new strategy for triclustering namely Correlation-Based Triclustering Using Genetic Algorithm (CorrTriGA) that combines the power of GA with a greedy strategy to overcome the shortcomings of poor convergence of GA approaches. A greedy-based technique is specifically incorporated into the GA framework to improve the search process. Fuzzy logic, neural networks, and swarm intelligence are types of computational intelligence (CIDs) that improve tricluster identification process by making the evolutionary process more dynamic. The experimental results carried out on the CDC15 Database to assess the performance of the proposed work. It is shown that CorrTriGA has ability to extract the higher volume coherent tricluster.**

Keywords - Computational Intelligence, Triclustering, Genetic Algorithm, Gene Expression Data.

1. INTRODUCTION

Triclustering is a valuable data mining technique used to discover hidden patterns in three-dimensional datasets. Unlike traditional data mining methods, triclustering simultaneously clusters genes, samples, and conditions, providing insights often missed by other algorithms. The main scope for triclustering techniques are bioinformatics, image analysis and social network mining to understand complex phenomena in these domains. Genetic algorithms (GAs) for triclustering tasks find optimal tricluster solutions across large 3D solution spaces. However, the main proven limitation of standard GA-based methods is slow converge rate for large datasets. To address this issue, Correlation-Based Triclustering Using Genetic Algorithm (CorrTriGA) is proposed that combines computational intelligence techniques (CIT) with genetic algorithms (GAs) for faster converges of the optimal solution and avoid suboptimal solutions. This makes triclustering processes for 3D microarray data better in terms of tricluster quality and convergence speed.

The remaining topics of the papers are as follows: Section 2 illuminates the literature study needed for the study. Section 3 discusses the methodologies and materials. Section 4 discusses the results and discussion. The conclusion of this work is exemplified in section 5.

2. LITERATURE REVIEW

This section deals with the various literature based on the genetic algorithm. To optimise the performance of the specialist-generalist utilising a classification system, minor adjustments based on clustering processes should be made (Gnatyshak, 2014). According to Yangyang Li (2014), the binary Particle Swarm Optimisation (BPSO) algorithm enhances search efficiency by integrating a pattern-driven local search operator. Mainly employed for gene selection is the BPSO encoding sensitivity of genes to classes (GCS). Highly accurate predictions of microarray data are made using nearest neighbour (KNN) and support vector machine (SVM) classifiers, which promote efficient gene selection. In FeiHan (2015), data on gene expression is subjected to bicluster removal using multi-objective particle swarm optimization.

2.1 Issues in the existing Algorithms:

- GA based clustering is very slow. It can find the best solution, but it can't find the exact solution.
- Most of these algorithms used MSR, and MSR_{3D} measures to evaluate the bicluster, but the measures fail to extract the correlated pattern from the data matrix.

3. METHODS AND MATERIALS

3.1 Description of Greedy Algorithm

An algorithm that uses heuristics to solve problems by looking for the local optimal solution at each step in order to discover the global optimum is called a greedy algorithm. This chapter defines triclustering as an optimisation issue that uses a greedy strategy to find the largest volume tricluster with a high Tri_{MCV}.

This is accomplished by creating the initial seed or initial population using two-way K-Means clustering methods (Diyar Qader Zeebaree, 2017). Initially, this population was fed into the Greedy Triclustering (GTri) algorithm in order to determine the ideal tricluster. A separate list is kept for genes, samples, and time points that are not part of the tricluster in the suggested greedy triclustering algorithm. Narmadha N. (2019) states that additional genes, samples, and time points are added to each tricluster independently.

The best element is chosen and added to the tricluster using this algorithm from the gene list, sample list, or time point list. Included in the element are Tri_{Vol} and Tri_{MCV} values that indicate the quality of the tricluster. It is believed that the optimal element is one that increases the Tri_{MCV} added to the tricluster. Prior to a decrease in the Tri_{MCV} value of that tricluster, the original tricluster grows from the gene list, sample list, and time point list. In order to create an ideal tricluster with a higher Tri_{MCV} value, this is referred to as the greedy strategy, which chooses the subsequent gene, sample, or time point. Algorithm 1 describes avaricious triclustering.

Algorithm 1: G_{Tri} (Greedy Triclustering)

Input: Initial gene nG and sample sets nS
Output: Enlarged set of genes and refined tricluster

Step 1: To generate a new random population by means of Algorithm 4.1.
Step 2: For each gene:
  i) To enlarge the set of genes.
  ii) To refine the set of genes.
Step 3: To return the Enlarged set of genes and Refined tricluster.

//Subfunction: To enlarge the set of gene (GeneEnlargement)
Step 1: To identify a set of genes that are not present in the current set (G').
Step 2: To identify a set of samples that are not present in the current set (S').
Step 3: To identify a set of time points that are not present in the current set (T').
Step 4: For each (G/S/T) gene/sample/time point that is not in the current set:
  If possible, adding it advances the Tri_{MCV} (Mean Correlation value of the tricluster):
    Increase it to the current set.
  End if
End for
Step 5: The Enlarged gene set is to returned.

// Subfunction: To refine a set of gene sets (GeneRefinement)
Step 1: For each node in the Enlarged set of genes:
  To remove the node from the Enlarged set of genes.
  To define G"/S"/T" as the each sets of rows/columns/time points in G'/S'/T' without the removed node.
  If refining the set then progresses the Tri_{MCV} of the Enlarged set of gene:
    Update G'/S'/T'.
  End if
End for
Step 2: To return the refined set of genes as (G") and also return the refined tricluster.

3.2 Correlation-Based Triclustering Using Genetic Algorithm ($CorrTri_{GA}$)

$CorrTri_{GA}$ is used to determine the optimal global tricluster using inputs from a database to predict the behaviour of genes concerning the subset of samples and time points. The proposed work is divided into two stages, namely the seed generation phase and Tricluster phase optimisation.

Two-$way_{K\text{-}Means}$ clustering and Greedy Triclustering methods are the two methods used to produce seeds from the 3D gene expression data in the seed generation phase. The following subsections define and explain the approaches. In the second step, GA is used as a global optimizer to identify the global optimal tricluster as defined in algorithm 2.

This work is presented to find maximum volume tricluster with a high degree of correlation among the genes. As mentioned in section 2, the fitness method for the existing GA based triclustering methods is typically built with the use of the MSR_{3D} score as the main component to measure the coherence of the tricluster. Although the MSR_{3D} is widely used as a quality metric, with such shifting and scaling, interesting patterns cannot be identified. Despite representing quality patterns, the MSR_{3D} is active in identifying triclusters with shifting patterns but not approximately patterns with scaling trends. (Shreya Mishra, 2018). It is evidenced that the MSR_{3D} is not a suitable measure for discovering patterns in 3D data when the variance of gene values is very high, that is, when the users have existing scaling patterns.

Algorithm 2: Correlation Based Triclustering Using Genetic Algorithm ($CorrTri_{GA}$)

Input: Initialise the population, max iteration
Output: Optimal Tricluster (Tri_{Opt})
Step 1: Set t=0, max iteration
Step 2: Initialise the population 'pop' using seeds obtained from the seed generation phase
Step 3: while (t≤ max iteration) do
// Select individual for reproduction
Selection (Tri_{pop})
// Recombine individuals (crossover)
Crossover (Tri_{pop}, Tri_{cp})
// Apply Mutation
Mutate (Tri_{pop}, Tri_{mp})
// Calculate the fitness of offspring
Evaluate fitness of Tri_{pop}
Reinsert offspring into population
t=t+1
end (while)
Step 4: Return Tri_{Opt} as the optimized tricluster.

4. RESULTS AND DISCUSSION

Table 1: Features of the optimal tricluster using $GTri_{GA}$ for the CDC15 Database

Tricluster ID	Tri_G	Tri_S	Tri_T	Tri_{Vol}	Optimal Tri_{MCV}
1	427	7	16	47824	0.971
2	421	7	16	47152	0.981
3	422	7	16	47264	0.990
4	423	7	16	47376	0.972
5	426	7	16	47712	0.981
6	418	7	16	46816	0.991
7	427	7	16	47824	0.987
8	421	7	16	47152	0.990
S9	428	7	16	47936	0.989
10	422	7	16	47264	0.990
11	421	7	16	47152	0.991
12	425	7	16	47600	0.990
13	422	7	16	47264	0.987
14	422	7	16	47264	0.989
15	425	7	16	47600	0.990
16	418	7	16	46816	0.991
17	421	7	16	47152	0.988
18	426	7	16	47712	0.989
19	418	7	16	46816	0.997

20	424	7	16	47488	0.989
21	427	7	16	47824	0.987
22	421	7	16	47152	0.988
23	421	7	16	47152	0.981
24	427	7	16	47824	0.988

Table 1 shows the features of optimal tricluster using $GTri_{GA}$ for the CDC15 database. It contains tricluster ID, Tri_G, Tri_S, Tri_T, Tri_{Vol}, and Tri_{MCV}. It was observed that the best tricluster ranged from 0.971 to 0.997. For different triclusters, there was a tremendous increase in their Tri_{MCV} value for the different tricluster ID. So, it had a highly correlated Tri_{MCV} value. Although there was no difference in the sample set and time point set, there was a notable difference in the gene set. At the same time Tri_{Vol} of the best triclusters is also increased in their volumes, so they have a large Tri_{Vol}. Therefore, $GTri_{GA}$ has the ability to extract the optimal tricluster with different gene sets for different Tri_{MCV}. Figure 1 shows the representations of Tri_G for $GTri_{GA}$ in the CDC15 database. The volume of the best tricluster (Tri_{Vol}) for $GTri_{GA}$ in the CDC15 database. The tricluster ID = 9 has a high Tri_{Vol} value of 47936. At the same time, $GTri_{GA}$ has a high optimal Tri_{MCV} for all the triclusters when compared to $CorrTri_{GA}$, as shown in Figure 2. Figure 3 shows the MCV of the best tricluster (TriMCV) for the CDC15 database using $GTri_{GA}$

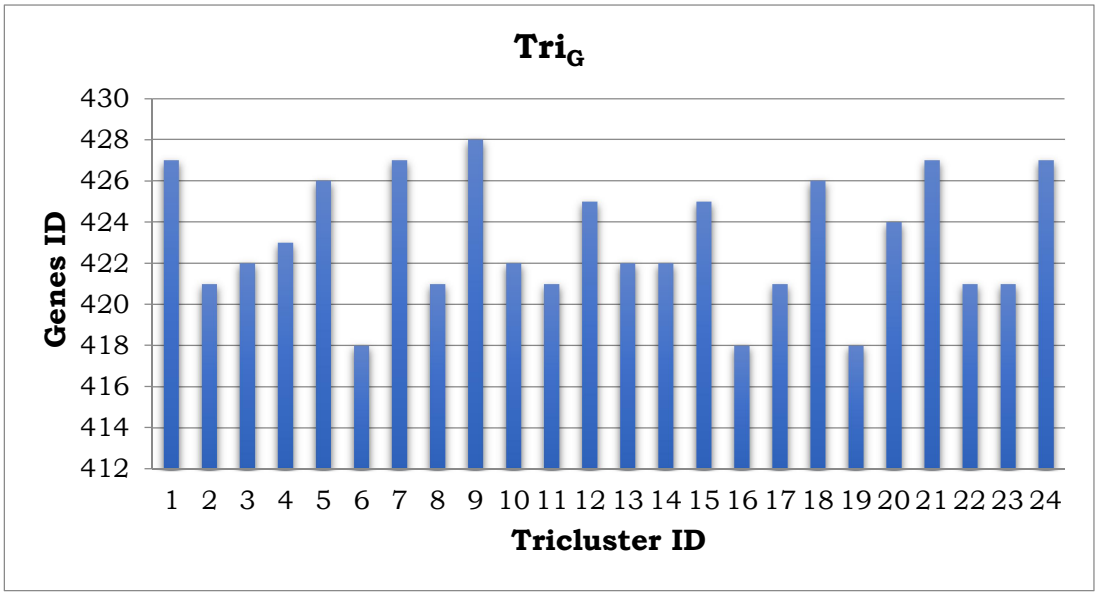

Figure 1: Representations of Tri_G for $GTri_{GA}$ for the CDC15 database

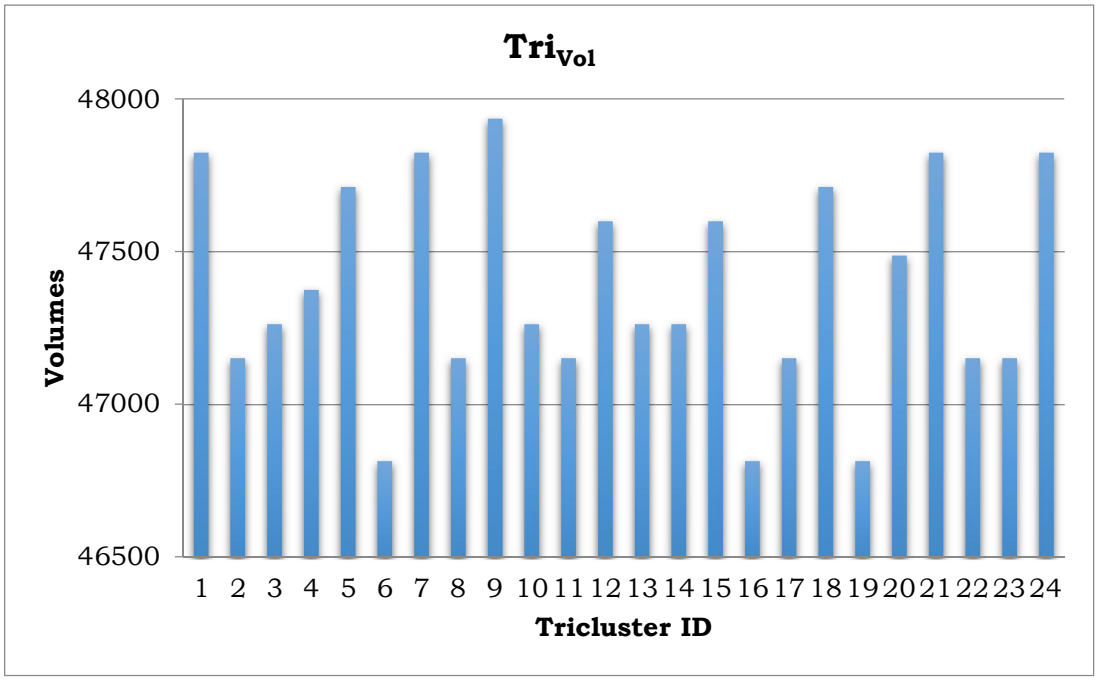

Figure 2: The volume of the best tricluster (Tri_{Vol}) for the CDC15 data using $GTri_{GA}$

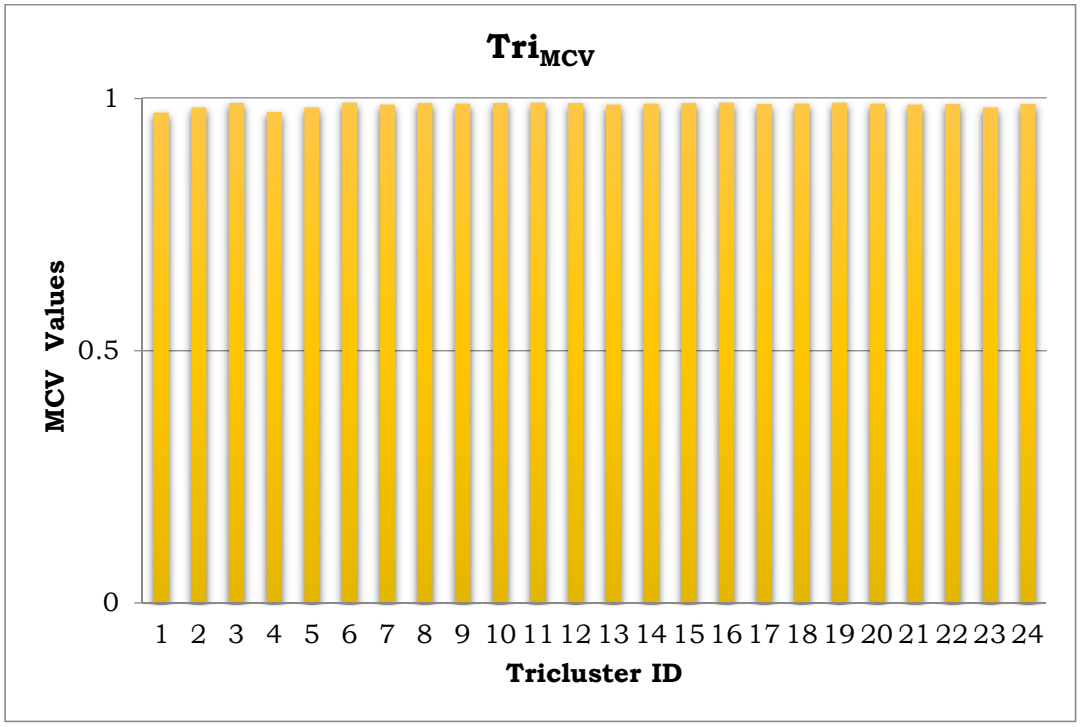

Figure 3: MCV of the best tricluster (Tri_{MCV}) for the CDC15 database using $GTri_{GA}$

5. CONCLUSION

This work proposed the evolutionary triclustering approach, which is used to identify the correlated genes from the gene expression data. In terms of optimal tricluster, this method defines the optimally correlated (i.e. coherent) pattern. A new correlation-based fitness function was the main feature of this work.

Experiments were performed on three different databases of 3Dgene expression data to determine the efficiency of the evolutionary triclustering algorithms. This experimental study showed that the proposed work achieved highly competitive results as compared to other common greedy triclustering algorithms in the literature. It has been observed from this analysis that all the individuals appear to converge to the best solution, which that takes more computational time.

REFERENCES

D.Gutiérrez-Avilés,C.Rubio-Escudero,F.Martínez Álvarez ,J.C.Riquelme (2014).TriGen: A genetic algorithm to mine triclusters in temporal gene expression data", Neurocomputing, Volume 132, 20 May, 42-53.

Duygu Dede and Hasan Oğul. 2014. TriClust: A tool for cross-species analysis of gene regulation. Molecular Informatics 33, 5 (2014), 382–387.

Fei Han (2015). A Gene Selection Method for Microarray Data Based on Binary PSO Encoding Gene-to-class Sensitivity Information.14(1),85-96. IEEE Transactions On Computational Biology And Bioinformatics.

Gnatyshak, D. V. (2014). Greedy Modifications of OAC-triclustering Algorithm. Procedia Computer Science, Elsevier,31, 1116-1123. doi:10.1016/j.procs.2014.05.367

J. Bagyamani, D. K. (2013). Comparsion of Biological Significance of Biclusters of SIMBIC and SIMBIC+ Biclustering Models.3(1),5-8 ACEEE Int.J.on Information Technology.

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

N. Narmadha, R. Rathipriya (2017).Triclustering : An Evolution of Clustering. 2016 Online International Conference on Green Engineering and Technologies , 04 May IEEE

N. Narmadha, R. Rathipriya (2019). Gene Ontology Analysis of 3D Microarray Gene Expression Data using Hybrid PSO Optimization. International Journal ofInnovative Technology and Exploring Engineering (IJITEE) , September 8 (11).

Narmadha.N, Rathipriya.R (2018).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.

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

Shyama Das (2010). Greedy Search-Binary PSO Hybrid for Biclustering GeneExpression Data. International Journal of Computer Applications.2(3),1-5, May.

Yangyang Li., Tian, X., Jiao, L., & Zhang, X. (2014). Biclustering of gene expression data using Particle Swarm Optimization integrated with pattern-driven local search. 2014 IEEE Congress on Evolutionary Computation (CEC). doi:10.1109/cec.2014.6900323