Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2019) 27: 4328 – 4339

doi:10.3906/elk-1809-86

Rough fuzzy cuckoo search for triclustering microarray gene expression data

Swathypriyadharsini PALANISWAMY[1*], **Premalatha KANDHASAMY**[2]
[1]Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Erode, India
[2]Department of Information Science and Engineering, Bannari Amman Institute of Technology, Erode, India

Received: 13.09.2018 • **Accepted/Published Online:** 13.07.2019 • **Final Version:** 26.11.2019

Abstract: Analyzing time series microarray gene expression data is a computational challenge due to its three-dimensional characteristics. Triclustering techniques are applied to three-dimensional data for mining similarly expressed genes under a subset of conditions and time points. In this work, a novel rough fuzzy cuckoo search algorithm is proposed for triclustering genes across samples and time points simultaneously. By applying the upper and lower approximation of rough set theory and the objective function of fuzzy k-means, rough fuzzy k-means was incorporated into a cuckoo search to handle the uncertainty of the data. The proposed method was applied to three real-life time series gene expression datasets. This work was evaluated using four validation indices and correlation analysis was performed to indicate the cluster quality. The proposed work was also compared with the existing triclustering algorithms and it outperformed the other methods.

Key words: Tricluster, Rough fuzzy cuckoo search, microarray gene expression data, rough fuzzy k-means, gene ontology, time series data analysis

1. Introduction

Due to the rapid growth of biomedical data, it is hard for medical workers to find important information. The clustering technique is used for identifying the useful knowledge behind large-scale biomedical data and thus it plays a major role in organizing biomedical data. Clustering analysis works in full-dimensional space by grouping the genes that are similar under all conditions or samples [1]. Biclustering groups subsets of genes which are similarly exhibited in only a subset of samples [2]. In addition, it suggests the possibility of mining gene expression patterns across time. Hence, triclustering finds subsets of genes that are coexpressed across a subset of conditions or samples over a subset of time points.

Cuckoo Search is a metaheuristic optimization algorithm that is used for triclustering gene expression data. A novel rough fuzzy cuckoo search algorithm is proposed with which the uncertainty and vagueness in time series microarray data are handled. In this approach, both rough and fuzzy concepts are incorporated with the cuckoo search to improve the performance of the clustering.

2. Materials and methods

2.1. Rough sets

Rough set theory controls the uncertainty and vagueness of imperfect knowledge with the concept of an approximation space [3]. If sets cannot be correctly classified using the available set of attributes, rough

*Correspondence: swa.pspd@gmail.com

sets can be used with fuzzy boundaries. Consider that there is a set of objects U called the universe; let $R \subseteq U \times U$ and X be a subset of U and R the equivalence relation. The most basic concept in rough set theory is the indiscernibility relation, generated by information about objects of interest. The indiscernibility relation is intended to express that due to lack of knowledge, we are unable to discern some objects employing the available information. It means that it is unable to deal with single objects, a fundamental concept of rough set theory [4].

The lower approximation of a set X with respect to R is the set of all objects that can be classified as X with certainty (certainly X with respect to R). It can be defined as follows in Equation (1):

$$R_*(x) = \bigcup_{x \in U} \{R(x) : R(x) \subseteq X\}. \tag{1}$$

The upper approximation of a set X with respect to R is the set of all objects that can be possibly classified as X (possibly X with respect to R). It can be defined as follows in Equation (2):

$$R^*(x) = \bigcup_{x \in U} \{R(x) : R(x) \cap X \neq \Phi\}. \tag{2}$$

The boundary region of a set X with respect to R is the set of all objects that can be classified neither as X nor as not X. If the boundary region is empty, then X is a crisp set. If the boundary region is nonempty, then X is a rough set.

2.2. Cuckoo search

Cuckoo search is a nature-inspired metaheuristic optimization algorithm which imitates the breeding parasitism of cuckoo species. In general, the cuckoo birds lay their eggs in the nests of other host birds, obligating its breeding parasitism [5]. If a host bird finds that the eggs are not their own, it will evict these foreign eggs or simply abandon its nest and build a new nest elsewhere [6]. The algorithm initiates with a number of nests in which each egg in the nest represents a solution. The Lévy flight method is applied for generating a new solution [7]. It replaces the worst solutions in the nest by newly produced better solutions. The best nests with good quality eggs will be carried on to the next generation. The exploration of the search space is carried out by Lévy flights for creating a new solution $x_i(t+1)$ using the solution $x_i(t)$ as shown in Equation (3):

$$x_i(t+1) = x_i(t) + \alpha \oplus Levy(\lambda). \tag{3}$$

The symbol $\oplus$ indicates an entry-wise multiplication. Lévy flights provide a random walk while their random steps are drawn from a Levy distribution for large steps given in Equation (4):

$$Levy \sim u = t^{\lambda}, \tag{4}$$

which has an infinite variance with an infinite mean. Here the consecutive jumps of a cuckoo search structures a random walk process which obeys a power-law step-length distribution with a heavy tail.

A time series gene expression dataset has G number of genes, S number of samples, and T number of time points. Each egg in a nest is represented by a binary string with three parts. An egg encodes a possible tricluster. Therefore, a nest has m number of bits corresponding to the genes, n number of bits corresponding to the samples, and k number of bits corresponding to the time points. Each string is represented by $m+n+k$

bits that have a value either 1 or 0. If the value is 1 then the corresponding gene or sample or time point is present in the tricluster. Figure 1 shows an encoding representation of a tricluster where genes are marked as rows, samples as columns, and time periods as depth. In a row, 1 represents the presence of gene/sample at a time point t for framing a tricluster and 0 represents the absence of gene/sample at a time point t when structuring a tricluster. For example, $\{G1, G3, \ldots, S1, \ldots Sn\}$ forms tricluster at time point $T1$.

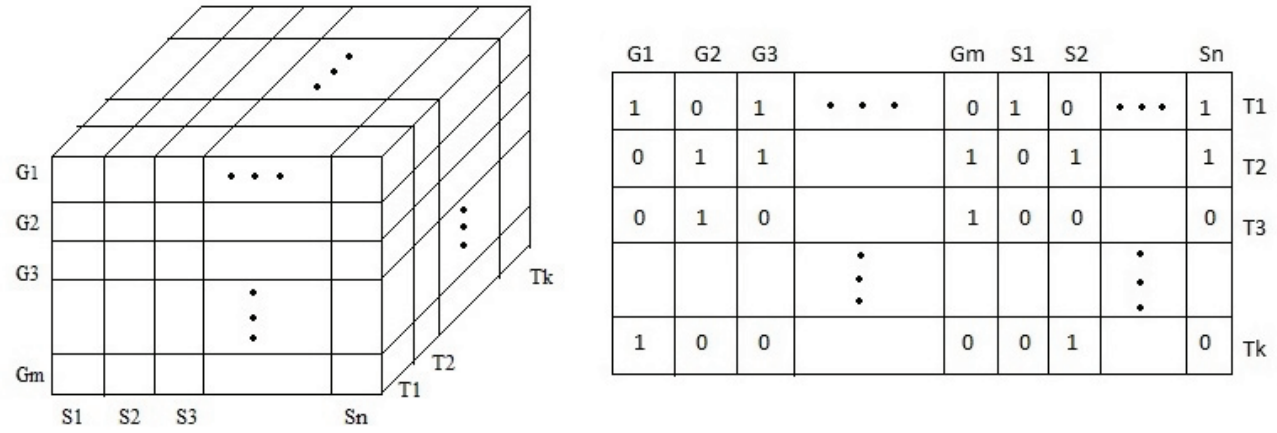

Figure 1. The solution representation of triclustering.

2.3. Rough fuzzy cuckoo search (RFCS) algorithm

The rough fuzzy cuckoo search algorithm initiates with n number of host nests in three dimensions such as genes, samples, and time points. Each cuckoo subset of triclusters is randomly chosen. The fuzzy concept is incorporated in the cuckoo search by calculating a membership matrix for each nest and the fitness for each nest is calculated by using the objective function of the rough fuzzy k-means. Let $X = \{x_1, x_2, \ldots, x_n\}$ be a set of objects to be clustered, the i^{th} class be denoted by w_i, its centroid v_i, the number of the cluster k, m a parameter, and A_i the upper approximate limit characterizing the border of all objects possibly belonging to the i^{th} class. If any object is greater than the upper approximate limit, then the object does not belong to the tricluster. The objective function of the RFCS algorithm is calculated using Equation (5) [8]:

$$F = \sum_{j=1}^{n} \sum_{i=1, x_j \in \overline{R} w_i}^{k} u_{ij}^m d_{ij}^2, \tag{5}$$

where F is the sum of squared error for the set of fuzzy clusters represented by the membership matrix U, and the associated set of cluster centers V. $||.||$ is some inner product induced norm. In the formula, $||x_k – v_i||^2$ represents the distance between the data x_k and the cluster center v_i. The squared error is used as a performance index that measures the weighted sum of distances between cluster centers and elements in the corresponding fuzzy clusters. The number m governs the influence of membership grades in the performance index, where m is any real number greater than 1. u_{ij} is degree of membership function with constraint conditions such as $u_{ij} \in [0,1], 0 \leq \sum_{j=1}^{n} u_{ij} \leq N, \sum_{i=1, x_j \in \overline{R} w_i}^{k} u_{ij} = 1, d_{ij} = \|x_j - v_i\|$. The centroids and membership matrix for each

nest can be evaluated using the Equations (6) and (7):

$$V_i = \sum_{j=1}^{n} u_{ij}^m x_j \Big/ \sum_{j=1}^{n} u_{ij}^m, \tag{6}$$

$$u_{ij} = \frac{1}{\sum\limits_{l=1, x_j \in \overline{R} w_i}^{k} \left(\frac{d_{ij}^2}{d_{lj}^2} \right)^{\frac{1}{m-1}}}. \tag{7}$$

If the membership of the object is greater than the upper approximation A_i, then the object is excluded from the tricluster. Otherwise, it is retained in the tricluster. Then a fraction of nests is aborted, keeping only the best solutions. Then, the current best tricluster solution is passed on to the next generation and the process is continued until it reaches a maximum number of generations or a stopping criterion.

The rough fuzzy cuckoo search algorithm steps are as follows:

Initial population of n host nests representing genes, samples, and time points

While $t <$ (*MaxGeneration*) *or* (*StopCriterion*)

Get a cuckoo subset matrix of Triclusters randomly (i)
and by applying Lévy flights using Equation (3).
The fitness of each nest F_i is evaluated using Equation (5) and the centroids are determined using Equation (6).

Evaluate the membership matrix for each nest using Equation (7).
Choose a nest among $n(j)$ randomly.

if $(F_i < F_j)$
Replace j by new solution by implementing Lévy flight and update the centroids
end if

if $u_{ij} > A_i$ (*UpperApproximate*)
Remove j from the tricluster.
end if

Fraction (p_a) of the worse nests is abandoned and new ones are built.
Retain the best triclusters.
Rank the solutions and choose the current best tricluster.
Pass the current best tricluster to the next generation.

end while

The fuzzy concept is incorporated by calculating the fuzzy membership u_{ij} for all nests and the fitness of each nest is evaluated using the objective function of the fuzzy k-means method. Then, the rough set concept is applied in the cuckoo search by evaluating the upper approximation A_i and if any object membership value is higher than the upper approximation, then the object is excluded from the tricluster. The selection of upper approximation is critical and a great deal of experiments is required to attain the optimal value for A_i. The value of the upper approximation limit A_i is a user-defined value based on the empirical analysis. The value

of the upper and lower approximation ranges between 0 and 1. The upper approximation may range from 0.7 to 0.9 and lower approximation ranges between 0.1 and 0.3. Thus, the upper approximation limit may be set between 0.7 and 0.9 and it depends on the data clusters and its residue value. The mean square residue value is calculated for each of the resultant clusters derived for different Ai and the best value with lowest MSR is chosen as the upper approximate limit. The value of A_i should not be too high or too low. If A_i is too large, then the rough set theory becomes meaningless by including all the objects. If A_i is too small, then the clustering error rate may be high.

3. Results and discussion

The proposed method was implemented on three different real-life time series datasets. All the datasets were obtained from the Gene Expression Omnibus (GEO). The dataset GDS5192 is a *Homo sapiens* experiment and holds 22158 Affymetrix human genome U133A 2.0 probe IDs. Cultures of primary fibroblasts from neonatal skin treated by Egr1 and Tgfb1 were measured at 24 and 48 h with two biological replicates per time point. The second dataset GDS1489 is an experiment on *Mus Musculus* growth hormone treatment and it holds 12,488 genes in an Affymetrix Murine Genome U74A Version 2 Array. It includes expression profiling of adipocytes treated with growth hormone and control for 30 min, 4 h, and 48 h in three independent experiments. The third dataset GDS2335 is an experiment on *Mus Musculus* in which exercise training-induced expression changes in the cardiac muscle of mice were studied. Muscle samples from training and diabetic training were collected for 5 weeks. The dataset holds 12,488 genes of Affymetrix Murine Genome U74A Version 2 Array.

The rough fuzzy cuckoo search algorithm was implemented in MATLAB version 7.14 (R2012a). The traditional cuckoo search algorithm uses fixed values for Lévy distribution coefficient λ, probability of discovery rate of the eggs p_a and step size α and the same values are assigned for optimization in the rough fuzzy cuckoo search and for clustering, the fuzzifier and weighting factor are fixed throughout. Table 1 shows the parameters and values considered for the proposed work, which were constantly maintained for all the datasets.

Table 1. Parameters and values of rough fuzzy cuckoo search.

Parameter	Value
Fuzzifier (m)	2
Weighting factor (ω)	0.95
Number of nest (n)	50
Discovery rate of alien eggs (pa)	0.25
Step size (α)	1
Levy distribution coefficient (λ)	1.5
Number of iterations	50

3.1. Performance comparison

The performance of the proposed work was compared based on four validation indices such as mean square residue [9], coverage of the triclusters [10], average row variance [11], and average correlation value [12]. These validation measures are used to evaluate biclustering, so they were remodeled for triclustering by extension with the time point. they were remodelled for triclustering by extension with the time point.

The mean square residue (MSR) of the tricluster is given in Equations (8) and (9) below. The low MSR value denotes that there is strong coherence in the tricluster [13].

$$MSR = \frac{\sum_{g\in G, s\in S, t\in T} r_{gst}^2}{|G| \times |S| \times |T|}, \tag{8}$$

$$r_{gst} = TS_v\,(t,g,s) + M_{GS}\,(t) + M_{GT}\,(s) + M_{ST}\,(g) - M_G\,(s,t) - M_S\,(g,t) - M_T\,(g,s) - M_{GST}, \tag{9}$$

where $M_{GS}\,(t)$ is the mean of genes under samples at a time point, $M_{GT}(s)$ is the mean of the genes over time under a sample, $M_{ST}(g)$ is the mean of a gene in time under the samples, $M_G(s,t)$ is the mean of the genes under a sample and a time point, $M_S(g,t)$ is the mean of the values of a gene at a time point under samples, $M_T(g,s)$ is the mean of a gene under a sample at all time points and M_{GST} is the mean value of all values in the tricluster.

Coverage for the triclustering algorithm can be given as in Equation (10):

$$Coverage = \left(\frac{g_{a\,lg} \times c_{a\,lg} \times t_{a\,lg}}{G \times C \times T}\right) \times 100, \tag{10}$$

where $g_{a\,lg}$,$c_{a\,lg}$, and $t_{a\,lg}$ denotes the total number of genes, samples and time points retrieved by the triclustering algorithm. G,C and T represents the total number of genes, samples and time points in the original datasets.

Average row variance (AVR) is defined as in Equation (11) below:

$$AVR_i = \frac{1}{|J|}\sum_{j\in J}(a_{ij} - a_{iJ})^2 + \frac{1}{|K|}\sum_{j\in J}(a_{ik} - a_{iK})^2, \tag{11}$$

where a_{ij} is the value of a gene in the tricluster, a_{iJ} is the mean of ith row in tricluster for all j conditions and a_{Ij} is the mean of ith row for all k time points.

Average Correlation value is given in Equation (12):

$$ACV = \max\left\{\frac{\sum_{i_1=1}^{|I|}\sum_{i_2=1}^{|I|}|r\_row_{i_1i_2}| - |I|}{|I|^2 - |I|}, \frac{\sum_{j_1=1}^{|J|}\sum_{j_2=1}^{|J|}|r\_col_{j_1j_2}| - |J|}{|J|^2 - |J|}, \frac{\sum_{k_1=1}^{|K|}\sum_{k_2=1}^{|K|}|r\_tim_{k_1k_2}| - |K|}{|K|^2 - |K|}\right\}, \tag{12}$$

where $r\_row_{i_1i_2}$,$r\_col_{j_1j_2}$, and $r\_tim_{k_1k_2}$ refers to the correlation between any pair of rows i1,i2 or columns j1,j2 or time points k1,k2 , according to the Pearson coefficient.

Table 2 shows the performance comparison of the various sample triclusters for all the datasets. The MSR value is low, which indicates that there exists good coherence among the triclusters. For GDS5192, the MSR value is very low, implying that the best triclusters were extracted. The average row variance is minimized, which indicates that a cluster is trivial. The ACV value lies between the range of 0 to 1. The higher ACV indicates that the genes clustered in the triclusters are highly correlated. For all the datasets, ACV is close to 1, indicating that the genes are highly correlated within a cluster.

Scatterplots are used for determining the relationship between two quantitative variables measured for the same individuals. All scatterplots determine two properties, such as strength and direction. The strength of a correlation is determined by its numerical value and the direction indicates whether the variables have positive

Table 2. Performance comparison.

Dataset	Clusters	Mean square residue	Coverage	Average row variance	Average correlation value
GDS5192	Tricluster 1	3.5927e-04	69.5109	0.0011	0.9999
	Tricluster 2	3.5395e-04	69.6191	0.0011	0.9999
	Tricluster 3	3.6737e-04	70.2321	0.0011	0.9999
	Tricluster 4	1.3355e-04	69.7543	4.0068e-04	0.9999
	Tricluster 5	2.8182e-04	70.2186	8.4550e-04	0.9999
GDS1489	Tricluster 1	0.2499	70.9216	0.7761	1.0000
	Tricluster 2	0.1417	70.2354	0.4333	1.0000
	Tricluster 3	0.1541	69.9712	0.4779	1.0000
	Tricluster 4	0.1469	70.5237	0.4525	1.0000
	Tricluster 5	0.1468	69.6749	0.4496	0.9999
GDS2335	Tricluster 1	0.1478	69.7550	0.4660	0.9999
	Tricluster 2	0.1567	70.1794	0.5018	0.9999
	Tricluster 3	0.1483	69.0423	0.4683	0.9999
	Tricluster 4	0.1527	70.2995	0.4850	0.9999
	Tricluster 5	0.1473	70.2594	0.4623	0.9999

correlation or negative correlation. If both the variables move in the same direction, then they are positively correlated, otherwise they are negatively correlated. In Figures 2–4, it is clearly seen that the variables are positively correlated. Only a very few outlier points are seen in all the figures. Then, the slope of the scatterplot provides information on the strength of the relationship between two variables and it can clearly be seen from Figures 2 and 4 that the variables have a strong linear relationship.

3.2. Comparison with other triclustering algorithms

The performance of the proposed rough fuzzy cuckoo search algorithm was compared with other triclustering algorithms based on two measures, the Triclustering Quality Index (TQI) and Statistical Difference from Background (SDB) [14]. TQI for the extracted triclusters is given in Equation (13):

$$TQI = \frac{MSR_i}{volume_i}, \tag{13}$$

where MSR_i is the mean squared residue of the i^{th} tricluster and $volume_i$ is the volume of the i^{th} tricluster representing the number of genes, samples, and time points. The value of TQI should be low for the quality of the triclusters to be high.

The Statistical Difference from Background (SDB) score is the second measure which signifies how many of the triclusters are statistically different from the background data matrix. The SDB score is given in Equation (14):

$$SDB = \frac{1}{n}\sum_{i=1}^{n}\frac{MSR_i}{\frac{1}{n}\sum_{j=1}^{r} RMSR_j - MSR_i}, \tag{14}$$

where n is the total number of triclusters extracted, MSR_i represents the mean squared residue of the i^{th} tricluster, and $RMSR_i$ is the mean square residue of the j^{th} random tricluster having the same number of

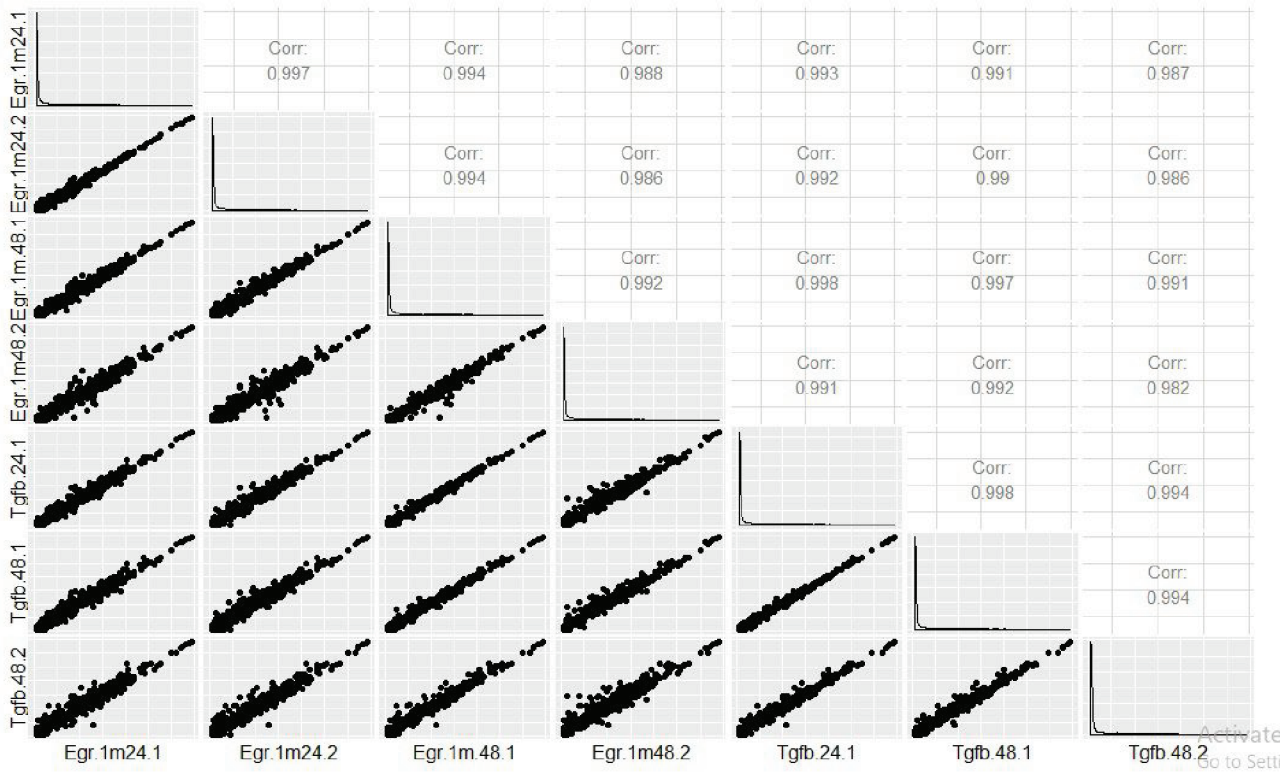

Figure 2. Scatterplot of a sample tricluster from GDS5192.

genes, samples, and time points as the i^{th} resultant tricluster. A lower SDB score implies better performance for the algorithm. Table 3 shows a comparison of the performance of various algorithms in terms of SDB and TQI indices. It is clearly shown that the proposed method outperforms the other algorithms, yielding the lowest values for both SDB and TQI.

Table 3. Performance comparison with other triclustering algorithms.

Algorithm	SDB	Average TQI
Tricluster using rough fuzzy cuckoo search algorithm	0.14703	1.48e-10
Tricluster using traditional cuckoo search algorithm	0.20945	2.01e-09
δ-TRIMAX	0.46709	3.08e-05
TRICLUSTER	0.47753	3.35e-05

Figure 5 shows the heatmaps of the sample triclusters extracted from GDS5192. The rows are the genes and the columns are samples from different time points. C1 to C3 belong to the first time point, C4 to C6 belong to the second time point, C7 to C9 belong to the third time point, and C10 to C12 belong to the fourth time point. The heatmap shows that the resultant triclusters are highly correlated.

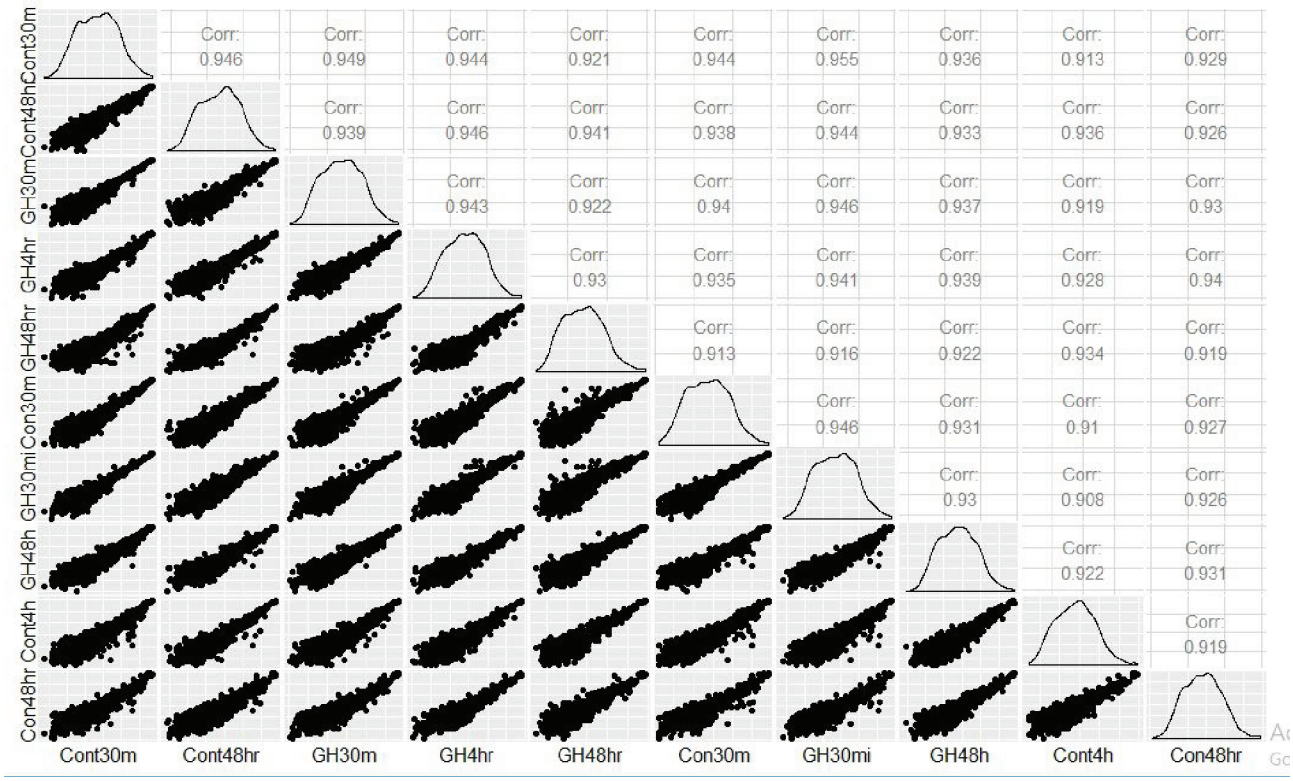

Figure 3. Scatterplot of a sample tricluster from GDS1489.

3.3. Biological significance

The biological significance of the genes belonging to the extracted triclusters from all the datasets was identified by performing Gene ontology and KEGG pathway enrichment analysis. The David ontology tool which is freely available on the internet was used for the gene enrichment analysis. Table 4 shows the biological process, cellular component, and molecular function for the best tricluster in all the three datasets. The Benjamini Hochberg method [15] was used for adjusting the P-values. The significant genes that have a P-value below a threshold of 0.05 were selected. Hence, the statistically enriched genes were retrieved for each tricluster.

For the gene enrichment analysis, the limma package [16] in R was used. In this package, the FDR-BH corrected P-value cut-off 0.5 was used. The gene ontology biological process and KEGG pathway enrichment analysis was performed. Table 5 shows the GOBP and KEGG pathway terms with the corrected P-values for all the datasets.

Figure 6 shows a comparison of the proposed work with the existing algorithms regarding biological significance. The biologically significant genes were compared in terms of the hit ratio, which is the number of genes in the triclusters that hit the enrichment analysis with the lowest P-value. The rough fuzzy cuckoo search gave a higher percentage of genes in the tricluster with lower P-value than the other algorithms. The inverse trend of genes hitting the analysis is noted with the rough fuzzy cuckoo search, which has the largest population at the lowest P-values, whereas the other algorithms have increasing populations with increasing P-values. Therefore, the triclusters extracted by the hybrid cuckoo search with clonal selection algorithm outperformed all other methods.

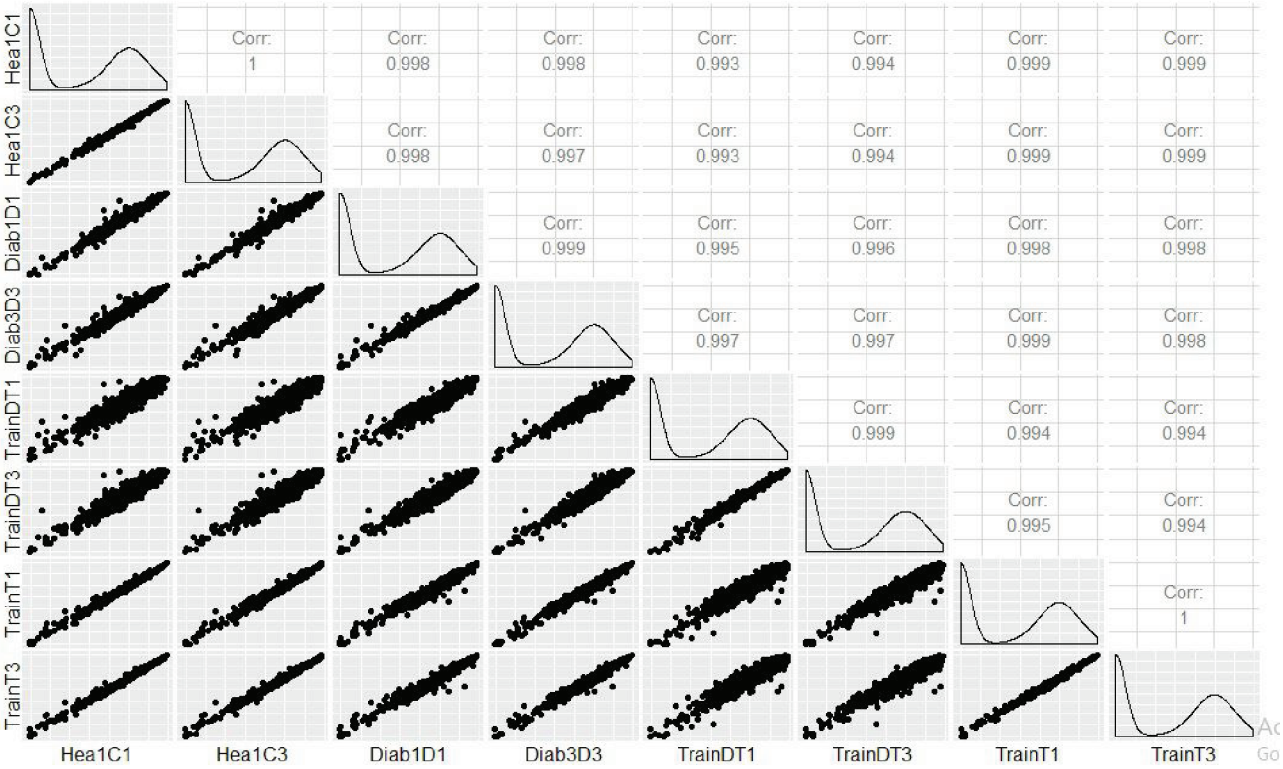

Figure 4. Scatterplot of a sample tricluster from GDS1489.

Table 4. Gene ontology for best tricluster in all three datasets.

Dataset	Biological process	Cellular component	Molecular function
GDS5192	Ras protein signal transduction, positive regulation of GTPase activity, transforming growth factor beta receptor signalling pathway	Axon Basolateral plasma membrane, sarcolemma	Ras quanyl-nucleotide exchange factor activity, signal transducer activity, androgen receptor binding
GDS1489	Positive regulation of transcription from RNA polymerase II promoter, negative regulation of cell proliferation, in utero embryonic development, apoptotic process	Cytoplasm, extracellular exosome cytosol	Protein binding, poly (A) RNA binding, protein homodimerization
GDS2335	Response to hypoxia, multicellular organism development, osteoblast differentiation,	Mitochondrion, protein complex, cell-cell adherens junction	Nucleotide binding, chromatin binding, cadherin binding involved in cell-cell adhesion

Table 5. GOBP and KEGG pathway terms.

Dataset	GOBP term	KEGG pathway terms
GDS5192	GO:0070060 actin filament nucleation (1.35 e-08)	path:hsa 00472 D-Arginine and D-ornithine metabolism (0.021)
GDS1489	GO:0046361 oxobutyrate metabolic process (2.37 e-06)	path:hsa00785 liopic acid metabolism (0.05)
GDS2335	GO:1902559 adenylyl sulphate transmembrane transport (1.18 e-02)	path:hsa00290 valine, leucine and isoleucine biosynthesis (0.625)

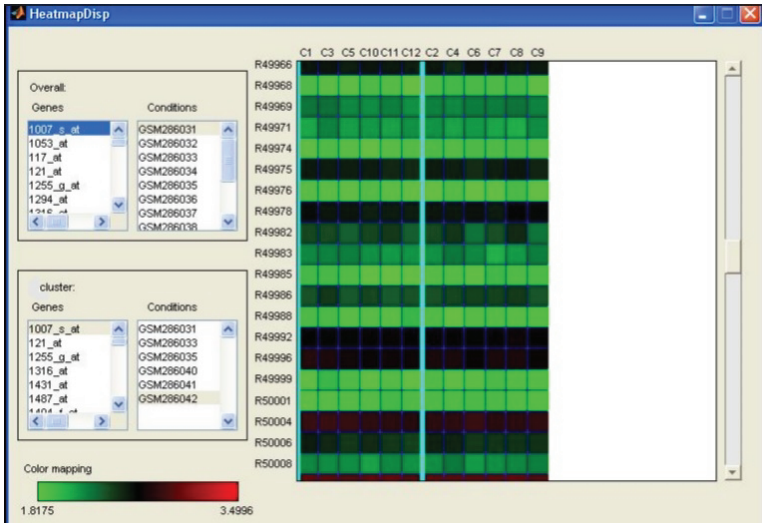

Figure 5. Sample heatmaps of the triclusters from GDS5192.

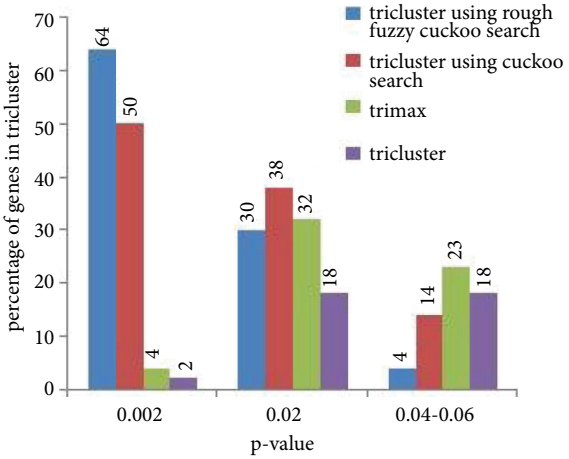

Figure 6. Biological significance comparison analysis.

4. Conclusion

In this work, the rough fuzzy cuckoo search is proposed to cluster three-dimensional datasets. It extracts the triclusters that contain genes that are coexpressed over a subset of experimental conditions across a subset

of time points. The rough fuzzy concept was incorporated into the cuckoo search optimization algorithm for handling uncertainty. The rough fuzzy cuckoo search algorithm was applied to four real-life temporal datasets. The proposed work was evaluated using four validation indices and its performance was also compared with those of other existing triclustering algorithms in terms of SDB and TQI. The experiment results prove that the rough fuzzy cuckoo search outperforms the existing methods. The biological significance of the triclusters that are extracted from this work was analyzed. The gene enrichment analysis was done by identifying the gene ontology and KEGG pathway terms by P-values adjusted using the FDR method.

References

[1] Xu R, Wunsch D. Survey of clustering algorithms. IEEE Transactions on Neural Networks 2005; 16 (3): 645-678. doi: 10.1109/TNN.2005.845141

[2] Cheng Y, Church GM. Biclustering of expression data. In: International Conference on Intelligent Systems for Molecular Biology; San Diego, CA, USA; 2000. pp. 93-103.

[3] Lingras P, West C. Interval set clustering of web users with rough K-Means. Journal of Intelligent Information Systems 2004; 23 (1): 5-16. doi: 10.1023/B:JIIS.0000029668.88665.1a

[4] Rissino S, Torres GL. Rough set theory – fundamental concepts, principals, data extraction, and application. In: Ponce J, Karahoca A (editor). Data Mining and Knowledge Discovery in Real Life Applications. Vienna, Austria: In-Tech. 2009, pp. 35-58. doi: 10.5772/6440

[5] Fister IJr, Yang XS, Fister D, Fister I. Cuckoo search: a brief literature review. Studies in Computational Intelligence 2014; 516(1): 49-62. doi: 10.1007/978-3-319-02141-6_3

[6] Payne RB, Sorenson MD, Klitz K. The Cuckoos. New York, NY, USA: Oxford University Press, 2005.

[7] Yang XS, Deb S. Cuckoo search via Lévy flights. In: 2009 World Congress on Nature & Biologically Inspired Computing; Coimbatore, India; 2009. pp. 210-214.

[8] Zhang JH, Ha MH, Wu J. Implementation of rough fuzzy k-means clustering algorithm in MATLAB. In: IEEE 2010 International Conference on Machine Learning and Cybernetics; Qingdao, China; 2010. pp. 2084-2087.

[9] Pontes B, GirldezJess R, Ruiz SA. Quality measures for gene expression biclusters. Plos one 2015; 10(3): e0115497.

[10] Bhar A, Haubrock M, Mukhopadhyay A, Maulik U, Bandyopadhyay S, Wingender E. Coexpression and coregulation analysis of time-series gene expression data in estrogen-induced breast cancer cell. Algorithm MolBiol 2013; 8 (1): 9. doi: 10.1186/1748-7188-8-9

[11] Haifa BS, Mourad E. DNA microarray data analysis: a new survey on biclustering. International Journal of Computational Biology 2015; 4 (1): 21-37.

[12] Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols 2009; 4 (1): 44-57. doi: 10.1038/nprot.2008.211

[13] Swathypriyadharsini P, Premalatha K. TrioCuckoo: a multi objective cuckoo search algorithm for Triclustering Microarray Gene Expression Data. Journal of Information Science and Engineering 2018; 34 (6): 1617-1631. doi: 10.6688/JISE.201811_34(6).0014

[14] Bhar A, Haubrock M, Mukhopadhyay A, Maulik U, Bandyopadhyay S et al. Multi objective triclustering of time-series transcriptome data reveals key genes of biological processes. BMC Bioinformatics 2015; 16 (1): 200.

[15] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B 1995; 57 (1): 289-300.

[16] Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology 2004; 3 (1): 1544-6115. doi: 10.2202/1544-6115.1027