Comput Syst Sci & Eng (2020) 1: 39–50

International Journal of
Computer Systems
Science & Engineering

Impact of Fuzzy Normalization on Clustering Microarray Temporal Datasets Using Cuckoo Search

Swathypriyadharsini P[1]*, K.Premalatha[2†]

[1] *Research Scholar, Anna University, Chennai, India*
[2] *Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Erode, India*

Microarrays have reformed biotechnological research in the past decade. Deciphering the hidden patterns in gene expression data proffers a prodigious preference to strengthen the understanding of functional genomics. The complexity of biological networks with larger volume of genes also increases the challenges of comprehending and interpretation of the resulting mass of data. Clustering addresses these challenges, which is essential in the data mining process to reveal natural structures and identify interesting patterns in the underlying data. The clustering of gene expression data has been proven to be useful in making known the natural structure inherent in gene expression data, understanding gene functions, cellular processes, and molecular functions. Clustering techniques are used to examine gene expression data to extract groups of genes from the tested samples based on a similarity criterion. Subspace clustering broadens the traditional clustering by extracting the groups of genes that are highly correlated in different subspace within the dataset. Mining the temporal patterns in high dimensional data is done with computational effort and thus normalization is needed. In this work, normalization using fuzzy logic is applied to the data before clustering. The multi-objective cuckoo search optimization is implemented to extract co-expressed genes over different subspaces. The proposed methods are applied to the real life temporal gene expression datasets in which it extracts the genes that are responsible for the disease grouped in a same cluster. The experiment results prove that the impact of fuzzy normalization on the dataset improves the clustering.

Keywords: Fuzzy normalization; Cuckoo search; Multi-objective optimization; Gene ontology; Temporal gene expression data.

1. INTRODUCTION

Microarray technology has been very effective in the examination of the expression of thousands of genes at a time and it has revolutionized the study of gene expression data. The activity of all genes measured for a number of biological conditions at each time point is referred to as three-dimensional datasets. The temporal datasets in microarray technology has been used to measure the expression values of thousands of genes under a huge variety of experimental conditions across different time points in a single experiment. As the volume of data is huge, several computational methods are needed to analyze such datasets. Therefore, Normalization is essential as a preprocessing technique before analyzing the datasets. The results of analysis will vary depending upon the normalization method and analysis method used for the same dataset. In this work, normalization using fuzzy logic is applied to the dataset. The fuzzification concept is applied in some gene expression profile analysis methods (Lim and Wong, 2014) (Geistlinger, 2011) and also in proteomic profile analysis methods (Goh et al., 2015) (Goh et al., 2016). However, the idea of fuzzification is used as a component of those methods but its role and effectiveness is utilized very less as a normalization procedure (Abha and Limsoon, 2016) (Kim et al., 2006).

Clustering is one of the unsupervised approaches to identify the coexpressed genes. Clustering algorithms aim to maximize similarity within the clusters as well as to minimize similarity between the clusters, based on a distance measure. The traditional clustering algorithms fail to find the group of genes that are similarly expressed over the subset of experimental

*E-mail: swa.pspd@gmail.com
†E-mail: kpl_barath@yahoo.co.in

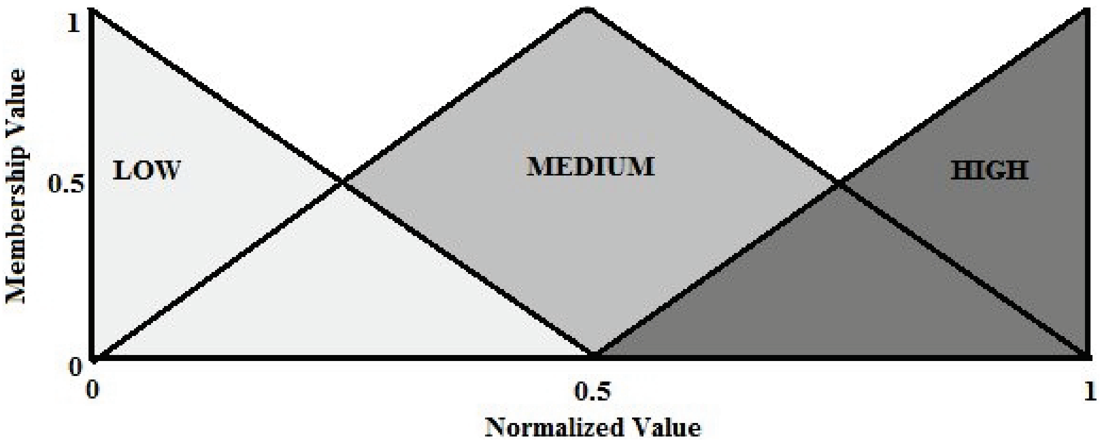

Figure 1 Fuzzy membership function.

conditions. Subspace clustering solves the problem by finding clusters with different subspace within a dataset. Subspace clustering algorithms like PROCLUS (PROjected CLUStering algorithm)(Aggarwal et al., 1999) and ORCLUS(arbitrarily ORiented projected CLUSter generation) (Aggarwal and Yu, 2000) deal with high dimensional datasets by finding clusters in different possible subset of dimensions using projective clustering. PROCLUSfinds axis-aligned subspaces by selecting k-mediods and then iteratively improves clustering by applying hill-climbing techniques. ORCLUS extends PROCLUS by considering non-axis parallel subspaces. This algorithm finds arbitrarily oriented clusters using singular value decomposition method and then determines the subspaces and finally merges it. Other subspace clustering algorithms like CLIQUE(Clustering In QUEst) (Agrawal et al., 1998) and ENCLUS(ENtropy based CLUStering) (Cheng et al., 1999) find the clusters within subspace but do not scale well with high dimensions and have low coverage. DOC (Document Clustering) (Procopiuc et al., 2002) fails to mine coherent patterns from microarray datasets. CLIFF (Clustering using Iterative Feature Filtering) (Xing and Karp, 2001) iterates between gene filtering and sample partitioning. Initially it finds k best genes according to the intrinsic discriminability. Then it partitions the samples with these features holding minimum normalized weights and iterates the entire process until convergence. COSA(Clustering on Subsets of Attributes)is an iterative algorithm that assigns weights to each instance using K nearest neighbour method. Each cluster may exist in different subspaces of different sizes but in similar dimension.

Cheng and Church proposed the first biclustering algorithm that was used to analyse gene expression datasets and it used a greedy search heuristic approach to retrieve largest possible bicluster having Mean Squared Residue (MSR) under a predefined threshold value δ (δ-bicluster) (Cheng and Church, 2000). Feng et al. (2004) proposed a time-frequency based full-space algorithm using a measure of functional correlation set between time course vectors of different genes. Jiang et al. (2006) proposed an algorithm to mine biologically meaningful coherent gene clusters using Spearman rank correlation similarity measurement and extended the clique search technique for the third dimension (Jiang et al., 2006). Yin et al. (2007) has given a new definition of coherent cluster for time series gene expression data called ts-cluster. The ts-cluster algorithm is able to detect a significant amount of clusters of biological significance. Aviles et al. (2014) also proposed TriGen algorithm which implements genetic algorithm for mining triclusters in temporal gene expression data. This algorithm implements the genetic algorithm, an optimization technique in order to retrieve the triclusters. Bhar et al. (2015) proposed EMOA-δ-TRIMAX, multi-objective optimization algorithm by implementing genetic algorithm. Liu et al. (2015) proposed fuzzy triclustering algorithm to mine triclusters based on the membership function for each dimension but it has computational efforts. Guigoures et al. (2016) also applied triclustering approach to track patterns in time-varying graphs. Anidha and Premalatha (2017) proposed a feature selection method for identifying the biomarker genes involving in causing cancer from microarray data. The data is normalized using fuzzy Gaussian membership function before classification which yielded higher accuracy. Prema and Premalatha (2018) applied fuzzy normalization to the microarray data before applying data mining techniques. Clustering using cuckoo search algorithm for three dimensional microarray data is performed for different encoding representations. Karmakar et al. (2019) proposed tight clustering for larger microarray gene expression data by applying k-means algorithm on several sub sampling of genes.

2. FUZZY NORMALIZATION

Zadeh (1974) proposed fuzzy set theory which is based on the intuitive reasoning by considering human subjectivity and imprecision. The idea of fuzzy logic is to hold the vagueness of human thinking and expressing it with mathematical tools. The classical set is a point-to –point control where as fuzzy set is a range-to-range control. In gene expression data analysis, the crisp values of the dataset are transformed into fuzzy values and the process is called fuzzification. A fuzzy set is constructed by dividing each gene of the dataset into three intervals namely low, medium and high. Figure 1 shows the common membership function of a fuzzy set.

The fuzzy membership function is used to represent the vague linguistic terms. The membership function value of each gene is calculated using Gaussian membership function which is given in the equation (1).

g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	s_1	s_2	s_3	s_4	s_5	s_6	t_1	t_2	t_3	t_4
1	0	1	0	1	1	1	0	0	1	0	1	0	1	0	1	1	0	1	1

Figure 2 Encoding representation of an egg.

Algorithm 1 Pseudo code for Cuckoo Search

Generate an initial population of n host nests representing genes, samples and time points;
while t< (MaxGeneration) or (stop criterion)
Get a cuckoo subset matrix randomly (i) and replace its solution by applying Levy flights;
Evaluate its fitness F_i
Choose a nest among n (j) randomly;
if ($F_i < F_j$)
Replace j by the new solution
end if
A fraction (p_a) of the worse nests is abandoned and new ones are built;
Keep the best solutions/clusters;
Rank the solutions and find the current best cluster;
Pass the current best cluster to the next generation;
end while

$$\mu_{A^i} = exp^{\left(-\frac{(x-c_i)^2}{2\sigma_i^2}\right)} \tag{1}$$

Where C_i is the centre and σ_i is the width of the i^{th} fuzzy set A^i. After applying the Gaussian membership function, it is represented with linguistic variables namely low, medium and high. The membership function value of each gene will range from 0 to 1.

3. CUCKOO SEARCH

Cuckoo Search is an optimization algorithm which is inspired from the breeding parasitism of cuckoo species. Some cuckoo species lay their eggs in the nest of other host birds in obligating its breeding parasitism (Fister et al., 2014). If a host bird discovers the eggs which are not their own, it will either throw these foreign eggs away or simply abandon its nest and build a new nest elsewhere (Payne et al., 2005). It initiates with number of nests in which each egg in the nest represents a solution. A new solution is generated by Levy flight (Yang and Deb, 2009). It aims to produce better solutions by replacing the worst solutions in the nest. Each cuckoo lays one egg at a time and throws down its egg in a randomly chosen nest. The best nests with good quality of eggs will be carried on to the next generation. The number of host nests is fixed, and a host can discover a foreign egg with a probability $p_a \in[0, 1]$. The host bird then throws away the egg or abandons the nest and also it finds the worst nest which is to be replaced.

When generating new solutions $x_i(t+1)$ for a cuckoo subset matrix i, a Levy flight is performed using the following equation (2).

$$x_i(t+1) = x_i(t) + \propto \bigoplus Levy(\lambda) \tag{2}$$

The symbol $\oplus$ is an entry-wise multiplication. Basically, Levy flights provide a random walk while their random steps are drawn from a Levy distribution for large steps given in equation (3).

$$\text{Levy} \sim u = t^{-\lambda} \tag{3}$$

which has an infinite variance with an infinite mean. Here the consecutive jumps of a cuckoo essentially form a random walk process which obeys a power-law step- length distribution with a heavy tail. The step size $\alpha > 0$ is related to the scale of the problem of interest but in most cases $\alpha = 1$ is maintained. $x_i(t+1)$ is the next location which depends on $x_i(t)$ is the current location and the second term in the equation is the transition propability. The Lévy flight based random walk is more efficient to explore the search space through longer step length. Here the consecutive jumps of a cuckoo essentially form a random walk process which obeys a power-law step- length distribution with a heavy tail.

CS is one of the best optimization algorithms that use elitism method with passing the best solutions to the next generation. The randomization through Lévy flight gives CS a random walk that is characterized by a probability density function. In case of Particle Swarm Optimization algorithm that depends on the inertia weight which needs development to incorporate the elitist concept of CS. Convergence to optimal solution is insensitive to the algorithm dependent parameters. CS has advantage over other algorithms by having only one parameter p_a needs to be adjusted.

3.1 Encoding

Each egg in a nest is represented by a binary string with three parts. An egg encodes a possible cluster. A time series gene expression dataset has G number of genes, S number of samples and T number of time points. Therefore, a nest has the first m bits corresponding to the genes, the next $n$bits corresponding to the samples and the last k 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 cluster. For example, a gene expression dataset having 10 genes, 6 samples and 4 time points, a string {10101110010101011011} represents that genes $\{g_1, g_3, g_5, g_6, g_7, g_{10}\}$, samples $\{s_2, s_4, s_6\}$ and time points $\{t_1, t_3, t_4\}$ are the members of the cluster as shown in Figure 2.

3.2 Fitness Function

Cluster is given as $C = I, J, L = cc_{ijl}$ where $i \in I,\ j \in J$, and $l \in L$. The cuboid C represents subset of genes which have similar expression values over a subset of samples during the subset of time points.

Dataset	Description	Type	Size
Breast cancer dataset GSE7561	It aims at finding IGF-I stimulated MCF-7 cells which is breast cancer cell line that is highly responsible to IGFs. Cells were stimulated in triplicate with or without IGF for 3 or 24 hrs.	Homo Sapiens	22277 genes
Ovarian cancer dataset GSE6653	It identifies the genes that show the changes after insertion of TGFb1 in IOSE which is derived from normal ovarian cells.	Homo Sapiens	54675 genes
Thyroid hormone dataset GSE24793	It has the target genes of thyroid hormone in cerebellum neurons of new born wild type mouse (mus musculus). To include maximum number of target genes, several cultures were treated or left untreated as controls for 4 time points such as 6, 16,24 and 48 hrs and the results are compared pairwise for each time point.	Mus musculus	45101 genes

Mean Square Residue (MSR) of the cluster can be modelled as

$$MSR = \frac{\sum_{g\in G, s\in S, t\in T} \quad r_{gst}^2}{|G| \times |S| \times |T|} \tag{4}$$

$$\begin{aligned} r_{gst} &= TS_v(t,g,s) + M_{GS}(t) + M_{GT}(s) + M_{ST}(g) \\ &\quad + M_G(s,t) - M_S(g,t) - M_T(g,s) - M_{GST} \end{aligned} \tag{5}$$

Where $M_{GS}(t)$ is the mean of genes under samples at a time point, $M_{MT}(S)$ is the mean of the genes over time under $M_{ST}(g)$ a sample, 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 cluster.

The first objective is to calculate the MSR value of the cluster which is given in equation (4) and (5).

$$f_1 = MSR \tag{6}$$

The low MSR value denotes there is strong coherence in the cluster. This includes only the trivial cluster when there is no fluctuation. The row variance is calculated in order to include the non trivial cluster. The second objective function is to calculate the row variance which is given in equation (7) and (8) in which a_{ij} is the value of a gene in the cluster, a_{ij} is the mean of ith row in cluster for all j conditions and a_{ij} is the mean of ith row for all k time points.

$$f_2 = \frac{1}{|I|}\sum_{i\in I} var_i \tag{7}$$

$$var_i = \frac{1}{|J|}\sum_{j\in J}\left(a_{ij} - a_{iJ}\right)^2 + \frac{1}{|K|}\sum_{k\in K}\left(a_{ik} - a_{iK}\right)^2 \tag{8}$$

The third objective function is the volume of the cluster which is calculated using the following equation (9).

$$f_3 = \frac{|I| \times |J| \times |K|}{|G| \times |S| \times |T|} \tag{9}$$

Where $(|I| \times |J| \times |K|)$ is the volume of the cluster and $(|G| \times |S| \times |T|)$ is the volume of the dataset. And this objective function is to be maximized in order to have increased size of the cluster.

The aim of this work is to find the clusters which should have a lower MSR score and a higher variance and a larger volume of the cluster. Thus, the first objective function is to be minimized and the second and third objective function is to be maximized in order to accomplish the goals. Therefore the optimal solution of the objective function is different from each other. Pareto optimal solutions solve this problem by considering set of constraints to get the optimal solution (Yang and Deb, 2001). It is based on the dominance criteria where a solution $x^{(1)}$ is said to dominate other solution $x^{(2)}$ if it holds the conditions such as the solution $x^{(1)}$ is no worse than $x^{(2)}$ in terms of all the objectives and the solution $x^{(1)}$ is strictly better than $x^{(2)}$ in at least one objective. The set of solutions which are not dominated by any others are called Pareto optimal front. Thus, the solutions are selected based on the pareto optimal front. Figure 3 shows the flowchart for the proposed work.

4. RESULTS

4.1 Description of Datasets

The proposed method is implemented on three different real life datasets. All the datasets are obtained from Gene Expression Omnibus (GEO). Two datasets are experiments for humans (Homo Sapiens) for different diseases and the third one is an experiment for human mouse (Mus musculus). All the experiments are time course experiments that have the behaviour of genes under samples for different time points.

4.2 Results and Discussion

Table 1 shows the parameters and values for the proposed work and it is constantly maintained for all the three datasets. The traditional CS algorithm uses fixed value for Lévy distribution coefficientλ, probability of discovery rate of the eggsp_a and the step size α and the same values are assigned here for performing

Table 1 Parameters and values considered for cuckoo search.

Parameter	Value
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	20

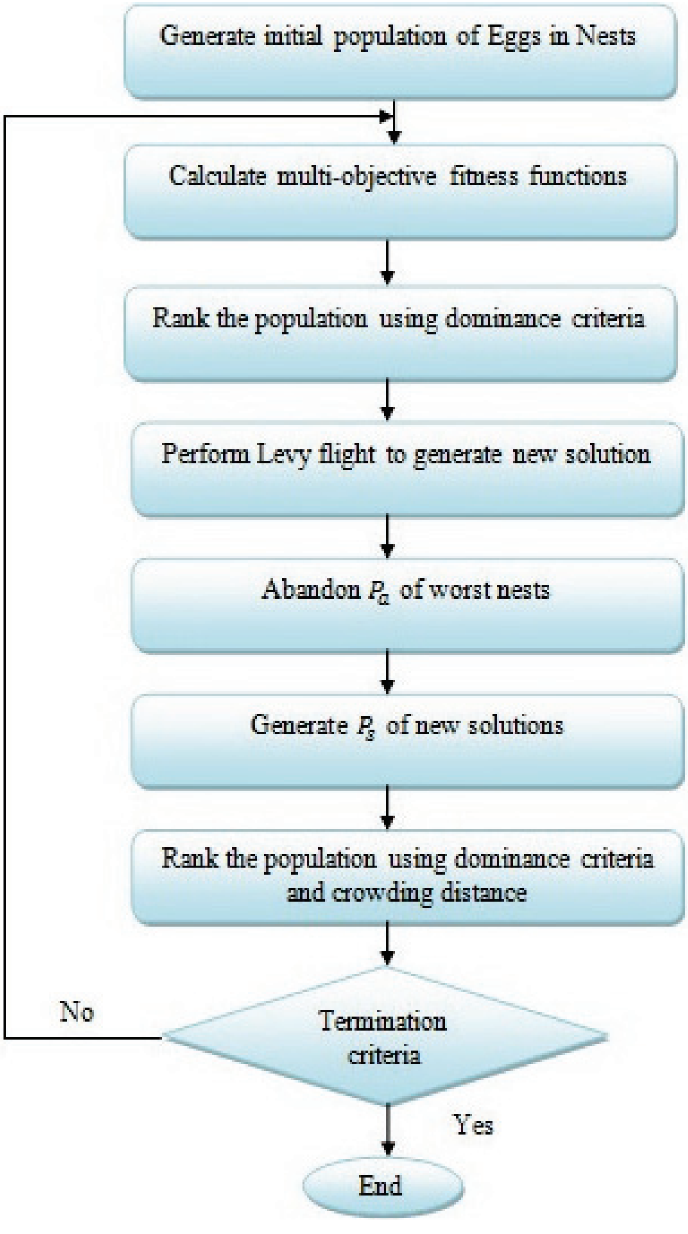

Figure 3 Cuckoo search with multi objective function.

the clustering Initially the entire dataset is normalized using fuzzy logic. Then the cuckoo search algorithm is applied for clustering the temporal datasets. To compare its performance, the other normalization methods such as quantile and z-score normalization methods are also applied to the entire dataset before clustering. During fuzzy normalization method, each gene of the dataset is considered as a fuzzy set and the fuzzification process is applied to all the genes. In the fuzzification process, the Gaussian membership function which is given in equation (1) is applied to all the genes where each gene is divided into three intervals namely low, medium and high. Figure 4 shows the Gaussian membership function of four random genes. The raw data has gene expression value ranges from 0 to 1600 which is hard to compute further. After fuzzy normalization, each gene value ranges between 0 and 1.

The normalized dataset which has I genes, J samples and K time points is given as input and that considers all

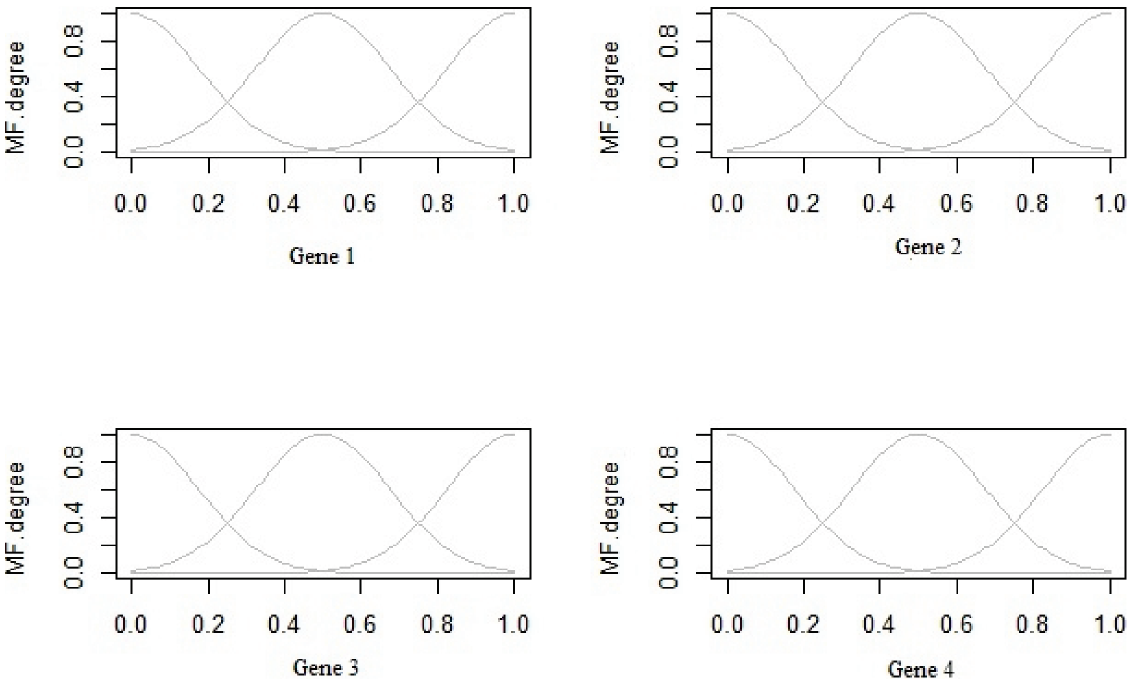

Figure 4 Gaussian membership function.

Table 2 Performance comparison of fuzzy normalization impact on clustering.

Data		Volume of the cluster	Best MSR value of the cluster
Breast cancer	Raw	124490	0.0320
	Quantile	126610	0.0239
	Z-Score	126470	0.0204
	Fuzzy	141640	0.0087
Ovarian cancer	Raw	190760	0.0112
	Quantile	191760	0.0096
	Z-Score	207880	0.0081
	Fuzzy	229970	0.0043
Thyroid hormone	Raw	189530	0.1077
	Quantile	199280	0.0849
	Z-Score	199320	0.0723
	Fuzzy	228450	0.0037

genes, samples and time points. It produces number of clusters which has i genes, j samples and k time points for which $i \in I$, $j \in J$ and $k \in K$. Table 2 shows the performance comparison of Fuzzy normalization with other existing normalization methods in terms of the volume of the cluster and best MSR value obtained from number of iterations. The aim is to get lower MSR value even in the cluster with larger volume. In all three datasets, quantile and z-score normalization has little improvement from the raw dataset. But Fuzzy normalization gives the lowest MSR value from a large volume of the cluster which proves that it outperforms the other methods.

Box plot is the effective way to assess the distribution of the data graphically. It splits the dataset into quartiles. The inter quartile which is the box in the plot represents the range of the data. The median of the data is marked in the box which divides the box in to two halves. The whiskers and the outliers can also be easily interpreted. In all the three datasets, skewness is witnessed in the raw dataset before normalization which shows the asymmetric distribution of data. The degree of dispersion and outliers are high in the raw dataset. Therefore, the dataset is to be normalized.

In quantile normalization, the skewness and outliers are slightly controlled. Z-Score performs better than quantile normalization, but there are outliers. But after Fuzzy normalization, all the data falls within the range, skewness is controlled and there are no outliers found. In Figure 5 breast cancer dataset, the median of the data varies slightly even in fuzzy normalization, but in Figures 6 and 7, the median of the data almost falls in the same range and also the degree of dispersion is very low. In addition, the fuzzy normalization scales the expression value which lies within 1 for all the genes.

Figures 8, 9 and 10 shows the sample five clusters obtained for breast cancer, ovarian and thyroid dataset respectively. Hundred genes are taken as sample for viewing the gene expression profile of the clusters. It shows that all the genes are correctly grouped in the clusters which have similar expression values. Only a very few fluctuations are seen in the clusters. This may be due to the missing values that are replaced by the random number but most of them remain in the cluster without violating.

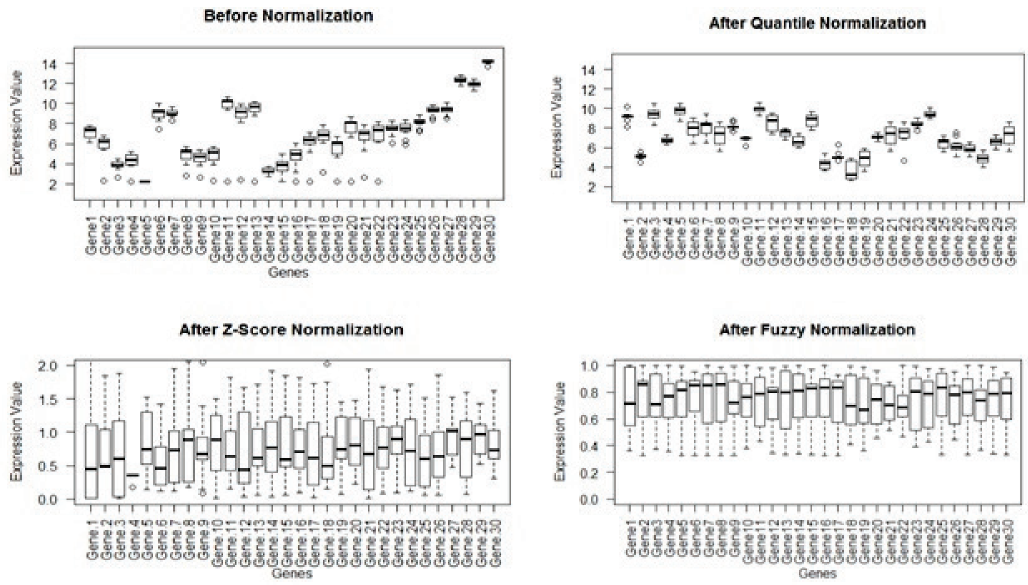

Figure 5 Box plots of breast cancer dataset.

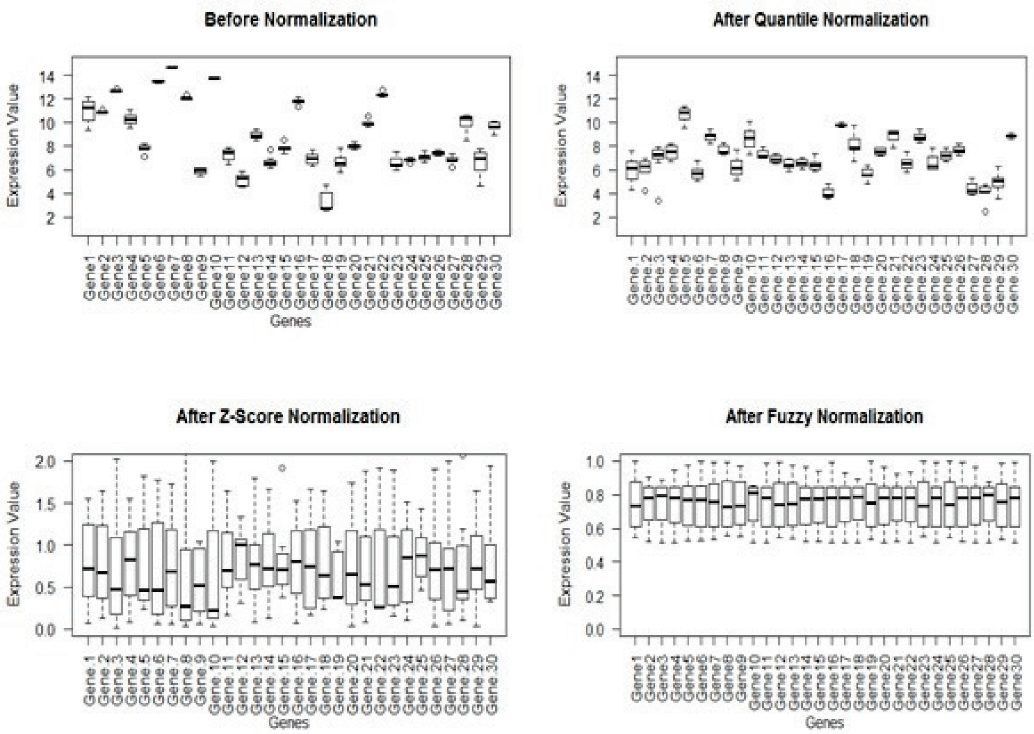

Figure 6 Box plots of ovarian cancer dataset.

Figure 11 shows the correlation analysis of the sample clusters extracted from the proposed work. The Pearson correlation coefficient method is used for performing correlation analysis of all genes in the clusters. Figure 11 A) shows the correlation of genes in breast cancer dataset between all the samples across all the time points that are grouped together in the cluster. It clearly shows that all the genes in the cluster are highly correlated. In Figure 11 B) ovarian cancer dataset, most of the values are 1 which proves that all the genes in the cluster are strongly correlated. In a sample cluster of thyroid dataset, 5 conditions are grouped together in the cluster and their correlation among all the genes is shown in Figure 11C.

4.3 Comparison With Other Clustering Algorithms

The performance of the proposed work is compared with other clustering algorithms based on two validation indexes. The first measure is the Triclustering Quality Index (TQI) which is given in Eq. (10).

$$TQI = \frac{MSR_i}{volume_i} \tag{10}$$

Where MSR_i is the mean squared residue of the cluster i and volume$_i$ is the volume of the cluster i. The volume of the ith

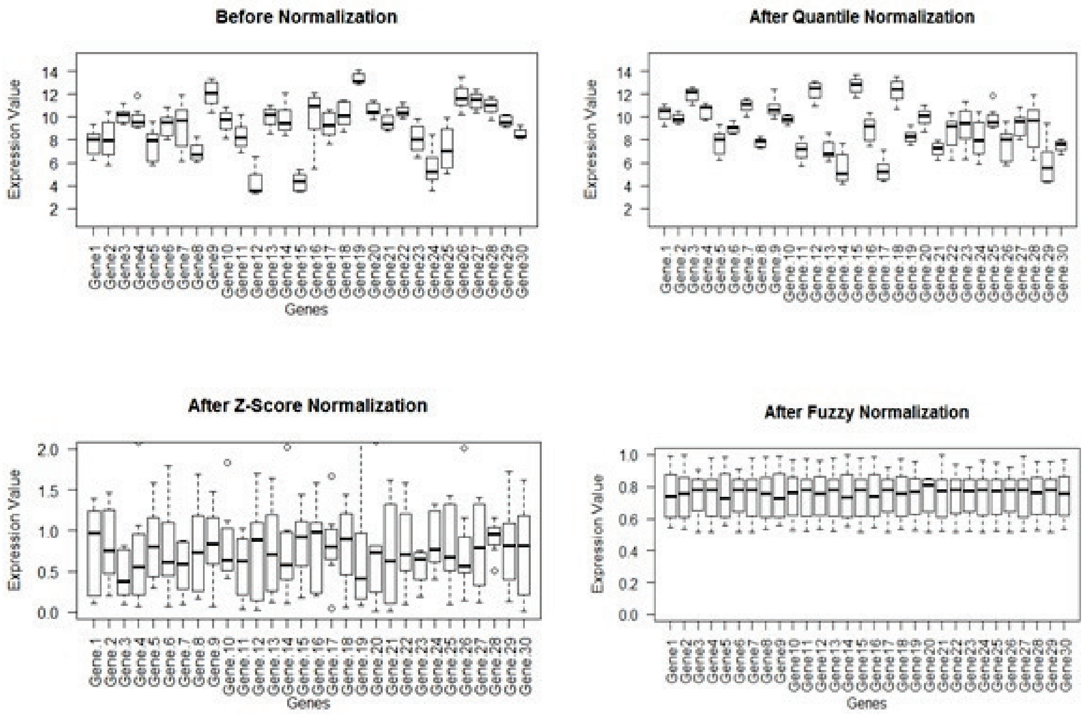

Figure 7 Box plots of thyroid hormone dataset.

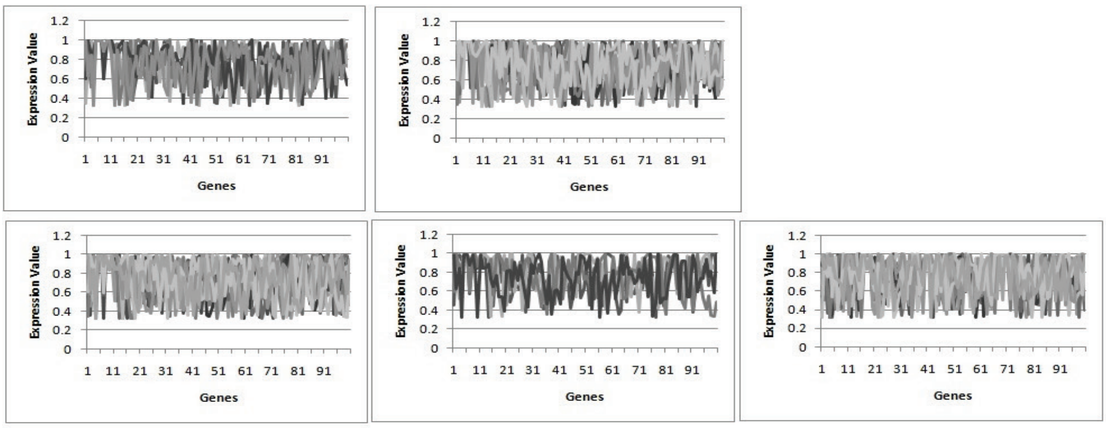

Figure 8 Five clusters obtained for breast cancer dataset.

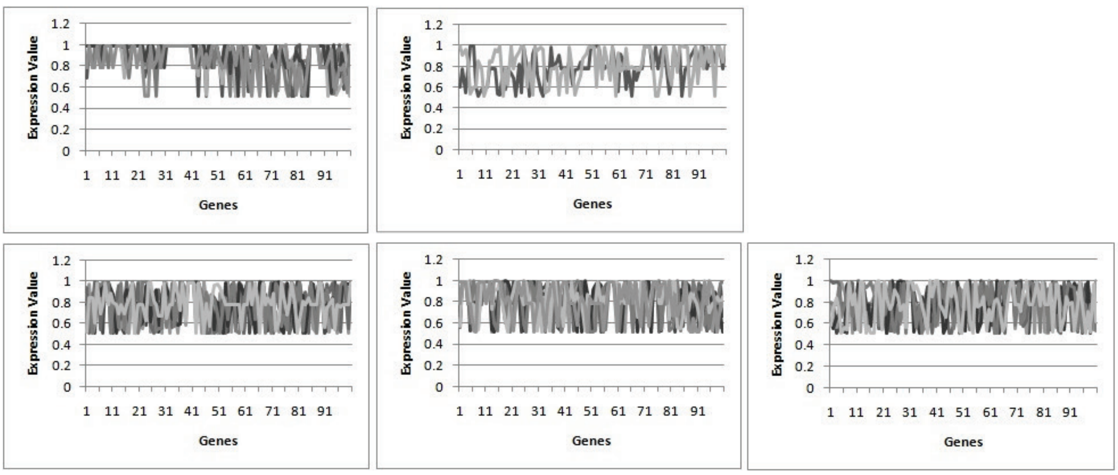

Figure 9 Five clusters obtained for ovarian cancer dataset.

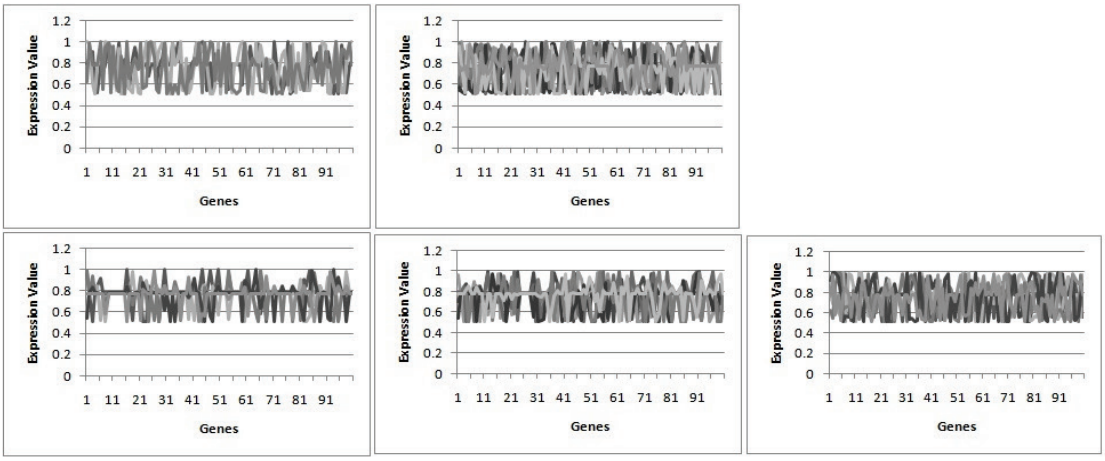

Figure 10 Five clusters obtained for thyroid hormone dataset.

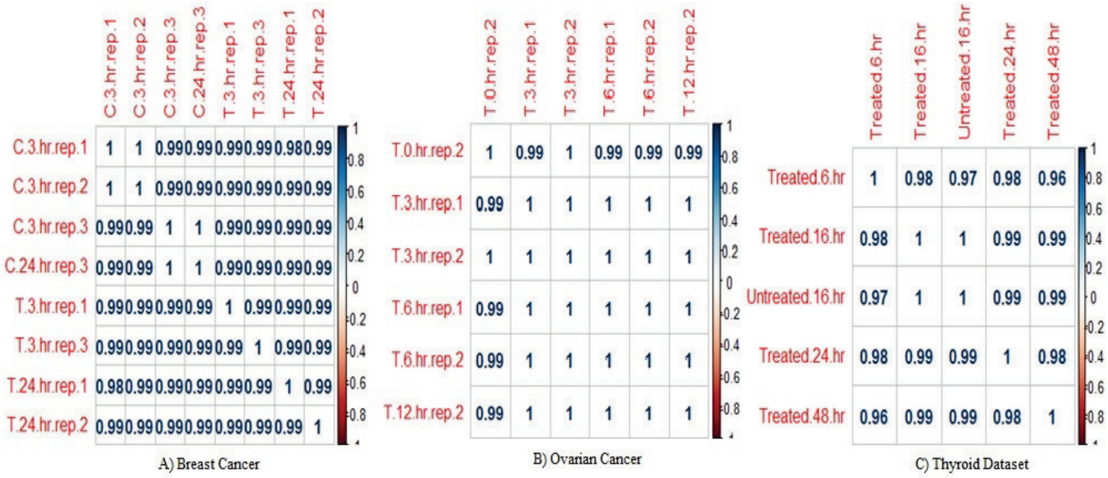

Figure 11 Correlation analysis of sample clusters.

cluster is defined as $(|I_i| \times |J_j| \times |K_k|)$ where $|I_i|$, $|J_j|$ and $|K_k|$ represent the number of genes, samples and time points of the ith cluster. A lower TQI represents the better quality of the clusters (Swathypriyadharsini and Premalatha, 2018).

The second measure is the Statistical Difference from Background (SDB) score which signifies whether a set of $n$clusters are statistically different from the background data matrix or not. The SDB score is given in the Eq. (11).

$$SDB = \frac{1}{n}\Sigma_{i=1}^{n}\frac{MSR_i}{\frac{1}{r}\sum_{j=1}^{r} RMSR_j - MSR_i} \tag{11}$$

Where n is the total number of clusters extracted by the algorithm. MSR_i represents the mean squared residue of the ith cluster and $RMSR_j$ is the mean square residue of the jth random cluster having the same number of genes, samples and time points as the ith resultant cluster. The higher the value of the denominator denotes the better the quality of the resultant cluster. Hence, lower SDB score signifies better performance of the algorithm. Table 3 shows the comparison of the performance of various algorithms in terms of SDB and TQI indexes in which clustering using cuckoo search after fuzzy normalization performs better.

4.4 Biological Significance

The biological significance of the genes which belongs to each of the clusters is identified by performing Gene Ontology analysis. David Ontology tool which is freely available in the Internet is used for the analysis (Huang et al., 2009). The p-values are adjusted using Benjamini Hochberg method (Benjamini and Hochberg, 1995). The significant genes that have a p-value below the threshold of 0.05 are selected. The lower p-value represents the higher significance level. Thus, the statistically enriched GO terms belonging to each cluster is extracted. Table 4, 5 and 6 shows the gene ontology of the breast cancer, ovarian and thyroid datasets respectively. The biological process, molecular function and cellular component of the genes in the different clusters are analysed and top three functionalities are given based on the lowest p-values.

5. CONCLUSION

In this work, the impact of the normalization using fuzzy logic is presented to cluster the temporal datasets. These temporal datasets have thousands of genes under several conditions across

Table 3 Performance Comparison.

Algorithm	SDB	Average TQI
Cluster extracted from Breast cancer dataset	0.25772	4.21E-09
Cluster extracted from Ovarian cancer dataset	0.33203	3.27E-09
Cluster extracted from Thyroid hormone dataset	0.20945	2.01E-09
δ-TRIMAX	0.46709	3.08E-05
TRICLUSTER	0.47753	3.35E-05

Table 4 Gene ontology for breast cancer dataset.

Cluster	Biological process	Molecular function	Cellular component	p-value
Cluster 1	Angiogenesis, Viral Process Signal transduction	Protein binding, poly(A) RNA binding, ATP binding	Cytosol, Nucleoplasm, membrane	2.9E-05 3.2E-04 3.0E-04
Cluster 2	Positive regulation of transcription from RNA polymerase II promoter, cell-cell adhesion, heart development	Identical protein binding, enzyme binding, transcription coactivator activity	Golgi apparatus, mitochondrion, cell-cell adherens junction	1.9E-08 1.9E-08 3.3E-07
Cluster 3	Response to drug, negative regulation of apoptotic process, extracellular matrix organization	Protein tyrosine kinase activity, ubiquitin protein liqase binding, receptor binding	Protein complex, transcription factor complex, extracellular space	1.3E-07 2.9E-06 2.3E-06
Cluster 4	Leukocyte migration, heart development, response to hypoxia	Cadherin binding involved in cell-cell adhesion, integrin binding, actin binding	Apical plasma membrane, melanosome, early endosome	2.3E-06 3.1E-05 1.5E-04
Cluster 5	Signal transduction, angiogenesis, cell proliferation	Protein homodimerization activity, protein serine, protease binding	Apical plasma membrane, melanosome, caveola,	3.0E-04 2.8E-04 1.3E-04

Table 5 Gene ontology for ovarian cancer dataset.

Cluster	Biological process	Molecular function	Cellular component	p-value
Cluster 1	Protein phosphorylation, viral process, intracellular protein transport	Transcription coactivator activity, metal ion binding, calmodulin binding	Nucleolus, cell junction, membrane raft	9.9E-08 1.8E-05 5.6E-05
Cluster 2	Cell-cell adhesion, protein poly ubiquitination, response to estradiol	ATP binding, Ras quanyl-nucleotide exchange factor activity, liquase activity	Centrosome, endoplasmic reticulum membrane, chromatin	2.4E-09 8.4E-08 2.4E-08
Cluster 3	DNA repair, cell migration, insulin receptor signalling pathway	Zinc ion binding, PDZ domain binding, cadherin binding involved in cell-cell adhesion	Endoplasmic reticulum membrane, perinuclear region of cytoplasm, lysosomal membrane	3.5E-08 6.4E-05 3.2E-05
Cluster 4	Vesicle-mediated transport, cell division, protein auto phosphorylation	Protein homodimerization, signal transducer activity, protein complex binding	Postsynaptic density, sarcolemma, mitochondrion	2.6E-04 1.2E-04 1.4E-03
Cluster 5	Cell proliferation, regulation of cell cycle, actin filament organization	Microtubule binding, actin filament binding, transcription regulatory region DNA binding	Nuclear speck, actin filament, extrinsic component of membrane	1.8E-05 5.6E-05 1.4E-04

many time points. Therefore, these datasets are to be normalized before clustering. The Gaussian membership function is applied to each gene of the dataset for fuzzification process. Then, the cuckoo search optimization technique is applied to subspace cluster the genes under different samples at many time points. The proposed work is applied to three different real life temporal

Table 6 Gene ontology for thyroid hormone dataset.

Cluster	Biological process	Molecular function	Cellular component	p-value
Cluster 1	Multicellular organism development, cell differentiation, mitotic nuclear division	Hydrolase activity, RNA binding, protein homodimerization	Golgi apparatus, synapse, neuronal cell body	2.3E-05 3.6E-04 1.3E-02
Cluster 2	Cell cycle, protein transport, mRNA processing, RNA splicing	Sequence-specific DNA binding, transferase activity, oxidoreductase activity	Cytoskeleton, perinuclear region of cytoplasm, neuron projection	1.5E-07 1.2E-06 2.5E-06
Cluster 3	Angiogenesis, DNA repair, covalent chromatin modification, mitotic nuclear division	Helicase activity, catalytic activity, ubiquitin protein liqase binding	Endoplasmic reticulum, lamellipodium, cytoplasmic vesicle	1.7E-06 1.4E-05 2.8E-05
Cluster 4	Positive regulation of cell migration, ion transport, axon guidance	Quanyl nucleotide exchange factor binding, magnesium ion binding, GTP binding	Mitochondrion, dendrite, postsynaptic density	2.0E-04 4.1E-03 1.6E-02
Cluster 5	Apoptotic process, cellular response to DNA damage stimulus, brain development	Nucleotide binding, protein kinase binding, transferace activity	Golgi apparatus, myelin shealth, intercalated disc	1.4E-04 1.5E-03 5.8E-03

datasets. The fuzzy normalization is also compared with other existing normalization methods and it outperforms other methods. The biological significance of the clusters that are extracted from the proposed work is analysed. In addition, the correlation analysis is performed to evaluate the results of the clusters which prove that all the genes in the cluster are highly correlated. For instance, the proposed method aims to cluster genes with related function, then the existing functional annotations are used to validate the resultant clusters. Clustering has been consistently applied in the medical sector to identify and analyze several ailments such as cancer, malaria and hormonal problems.

REFERENCES

1. ABHA B and LIMSOON W, *GFS: Fuzzy preprocessing for effective gene expression analysis*, BMC Bioinformatics, **17**(1), pp. 169–184, 2016.
2. AGGARWAL C.C., PROCOPIUC C., WOLF J.L., YU P.S., and PARK J.S., *Fast algorithms for projected clustering,* In ACM SIGMOD Conference, 1999.
3. AGGARWAL C. C. and YU P. S., *Finding generalized projected clusters in high dimensional spaces*, In ACM SIGMOD Conference, 2000.
4. AGRAWAL R., GEHRKE J., GUNOPULOS D. and RAGHAVAN P., *Automatic subspace clustering of high dimensional data for data mining applications*, In Proceedings of the ACM SIGMOD international conference on Management of data, pp. 94–105, 1998.
5. ANIDHA M and PREMALATHA K, *An application of fuzzy normalization in miRNA data for novel feature selection in cancer classification*, Biomedical Research, **28**(9), 2017.
6. AVILES D G, ESCUDERO C R, ALVAREZ F M and RIQUELME J C, *TriGen: A genetic algorithm to mine triclusters in temporal gene expression data*, Neurocomputing, **132**(1), pp. 42–53, 2014.
7. BHAR A, HAUBROCK M, MUKHOPADHYAY A, MAULIK U, BANDYOPADHYAY S and WINGENDER E, *Multiobjectivetri-clustering of time-series transcriptome data reveals key genes of biological processes*, BMC Bioinformatics, **16**(1), pp.200, 2015.
8. BENJAMINI Y. and HOCHBERG Y., *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, Journal of the Royal Statistical Society: Series B., **57**(1), pp. 289–300, 1995.
9. CHENG C.-H., FU A. W. and ZHANG Y., *Entropy-based subspace clustering for mining numerical data*, In Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM Press, pp.84–93, 1999.
10. CHENG Y. and CHURCH G.M., *Biclustering of expression data*, In Proceedings of International Conference on Intelligent Systems Molecular Biology, pp. 93–103, 2000.
11. FENG J., BARBANO P.E. and MISHRA B., *Time-frequency feature detection for timecourse microarray data*, In Proceedings of the ACM Symposium on Applied Computing, pp. 128–132, 2004.
12. FISTER I. Jr., YANG X. S., FISTER D. and FISTER I., *Cuckoo Search: A Brief Literature Review in Cuckoo Search and Firefly Algorithm*, *Studies in Computational* Intelligence, **516**, pp. 49–62, 2014.
13. GEISTLINGER L., CSABA G., KÜFFNER R., MULDER N. and ZIMMER R., *From sets to graphs: towards a realistic enrichment analysis of transcriptomic systems*, Bioinformatics, **27**(13), pp.366–373, 2011.
14. GOH W.W.B., GUO T., AEBERSOLD R. and WONG L., *Quantitative proteomics signature profiling based on network contextualization*, Biology Direct, **10**(1), pp. 71, 2015.
15. GOH W.W.B. and WONG L., *Evaluating feature-selection stability in next-generation proteomics*, Journal of Bioinformatics and Computational Biology, **14**(5), pp. 1650029, 2016.
16. GUIGOURÈS R, BOULLÉ M and ROSSI F, *Discovering patterns in time-varying graphs: a triclustering approach*, Advances in Data Analysis and Classification, Springer Verlag, pp. 1–28, 2016.
17. HUANG D.W., SHERMAN B.T. and LEMPICKI R. A., *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources*, Nature Protocols, **4**(1), pp. 44–57, 2009.
18. JIANG H., ZHOU S., GUAN J. and ZHENG Y., *gTRICLUSTER: A More General and Effective 3D Clustering Algorithm for Gene-Sample-Time Microarray Data*, Lecture notes in computer science, **3916**, pp. 48–59, 2006.
19. KARMAKAR B, DAS S, BHATTACHARYA S, SARKAR R and MUKHOPADHYAY I, *Tight clustering for large datasets with an application to gene expression data*, Scientific Reports, **99**(1), 2019.

20. KIM S Y, LEE J W and BAE J S, *Effect of Data Normalization on Fuzzy Clustering of DNA Microarray Data*, BMC Bioinformatics, **7**(1), 2006.
21. KUO H.C. and TSAI P.C., *Mining Time-delayed Gene Regulation Patterns from Gene Expression Data*, GSTF Journal on Computing (JoC), **2**(1), 2012.
22. LIM K. and WONG L., *Finding consistent disease subnetworks using PFSNet*, Bioinformatics, **30**(2), pp. 189–96, 2014.
23. PAYNE R. B., SORENSON M. D. and KLITZ K., *The Cuckoos*, Oxford University Press, 2005.
24. PREMA R and PREMALATHA K, *Effect of intuitionistic fuzzy normalization in microarray gene selection*, Turkish Journal of Electrical Engineering & Computer Sciences, **26**(3), pp. 1141–1152, 2018.
25. PROCOPIUC C. M., JONES M., AGARWAL P.K. and MURALI T., *A monte carlo algorithm for fast projective clustering*, In ACM SIGMOD Conference, 2002.
26. SWATHYPRIYADHARSINI P and PREMALATHA K, *TrioCuckoo: A Multi Objective Cuckoo Search Algorithm for Triclustering Microarray Gene Expression Data*, Journal of Information Science and Engineering, **34**(6), pp. 1617–1631,2018.
27. XING E. P. and KARP R. M., *CLIFF: clustering high-dim microarray data via iterative feature filtering using normalized cuts*, Bioinformatics, **17**(1), pp. 306–315, 2001.
28. YIN Y., ZHAO Y., ZHANG B. and WANG G., *Mining Time-Shifting Co-regulation Patterns from Gene Expression Data*, Springer-Verlag Berlin Heidelberg, **4505**, pp. 62–73, 2007.
29. YANG X.S. and DEB S., *Cuckoo Search via Levy Flights*, World Congress on Nature and Biologically Inspired Computing, pp. 210–214, 2009.
30. YANG X.S. and DEB S., *Multi objective optimization using evolutionary algorithms*, Published in Wiley, UK, 2001.