Taylor & Francis
Taylor & Francis Group

Automatika

Journal for Control, Measurement, Electronics, Computing and Communications

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/taut20

Spatial clustering based gene selection for gene expression analysis in microarray data classification

P. Edwin Dhas, Lalitha S, Annalakshmi Govindaraj & B. Jyoshna

To cite this article: P. Edwin Dhas, Lalitha S, Annalakshmi Govindaraj & B. Jyoshna (2024) Spatial clustering based gene selection for gene expression analysis in microarray data classification, Automatika, 65:1, 152-158, DOI: 10.1080/00051144.2023.2284027

To link to this article: https://doi.org/10.1080/00051144.2023.2284027

Published online: 10 Dec 2023.

AUTOMATIKA
2024, VOL. 65, NO. 1, 152–158
https://doi.org/10.1080/00051144.2023.2284027

Spatial clustering based gene selection for gene expression analysis in microarray data classification

P. Edwin Dhas[a], Lalitha S[b], Annalakshmi Govindaraj[c] and B. Jyoshna[d]

[a]Department of Computer Science and Engineering, Jayaraj Annapackiam CSI College of Engineering, Nazareth, India; [b]Dept. of Electronics and Communication Engineering, B.M.S. College of Engineering, Bengaluru, India; [c]Department of CSE, Koneru Lakshmaiah Education Foundation, Hyderabad, India; [d]Department of CSE, Keshav Memorial Institute of Technology, Hyderabad, India

ABSTRACT

A typical application of categorization in data mining is to uncover interesting distributions and significant patterns in the information that underlies it using density-based spatial clustering for workloads with noise. In these conditions, it is anticipated that the classification of the microarray gene expression database will have the necessary clustering property that may be utilized to emphasize the effects of the alterations. The proposed method typically guarantees that the subsequent identification of gene clusters' best global arrangement of genes. It provides an iterative method for figuring out the precise number of clusters needed for each data collection. The technique is based on practices frequently used in statistical tests. The key idea is to coordinate gene redistribution optimization across clusters with the search for the optimal number of groups. An experiment that finds the most effective number of genes over time was used to evaluate the effectiveness of the suggested strategy. It used this stringent statistical test to show that our technique accurately clusters more than 95% of the genes. Finally, since the basic principles of gene development and gene cluster assignment have been well characterized by earlier studies and the technique was verified using real gene expression information.

ARTICLE HISTORY
Received 28 August 2023
Accepted 6 November 2023

KEYWORDS
Gene selection; feature selection; microarray gene expression; spatial cluster optimization algorithm

1. Introduction

Rapid advances in transgenic technology in recent years have enabled the simultaneous monitoring of thousands of gene presentation characteristics. What experimental conditions should letters be addressed in? The evolution of computer methods for comprehending such data and efficiently structuring it in system-level conceptual structures has been fuelled by the availability of increasingly precise and freely accessible expression data. The study of genome-wide expression information has now used computationally clustering on the expression of genes profiles.

These techniques are founded on the fundamentals biological premise that genes with similar pattern of expression are co-regulated and may serve a similar function or pathway. Even though this assumption may be overly simplistic and not necessarily true, it has been demonstrated that the study of clusters is effective for analyzing gene expression data. Numerous clustering techniques are now available, but many significant questions remain unresolved. Due to problems with robustness, distinguishing features, and optimal results of nonlinear ordering in the horizontal cluster algorithms, for example, it is difficult to interpret the generated structures. However, algorithms that focus on maximizing a particular cost function are unable to guarantee that the final outcome fulfils the global optimum as opposed to a localized one. The ideal number of clusters is a problem for each of these prominent clustering algorithms. It is up to the observer to assess topographies of trees and identify branch points that divide physiologically meaningful groupings in order to resolve this issue. The total number of clusters has been added as a constant, external parameter of the procedure in optimization-based techniques.

The contribution of the work is

- This strategy often ensures that the best globally organization of genes in gene clusters is identified later.
- The basic concept is to align gene redistribute optimization across clusters alongside the search for the optimal number of groups.
- Furthermore, because previous studies have adequately described the basic principles of gene growth and gene cluster task, the technique was validated using real gene expression data.

CONTACT Edwin Dhas edwindhas.au@gmail.com Department of Computer Science and Engineering, Jayaraj Annapackiam CSI College of Engineering, Nazareth, Tamilnadu 628617, India, Dept. of Electronics and Communication Engineering, B.M.S. College of Engineering, Bengaluru 560019, India

2. Related work

Density-based spatial grouping for heterogeneous gene expression information has been the subject of numerous investigations. (2020) Vijayalakshmi and Co. They used particle swarm optimization to optimize non-dominated organization, and metaclassifier techniques such the k-nearest neighbour algorithm, rapid decision tree, and assessment of kernel density. To get the best degree of prediction accuracy for breast cancer, the Bayes theorem was applied to the results. With the help of the suggested particle swarm optimization, the most crucial variables affecting a malignant breast patient's prognostic were identified; they were not sorted using a classification approach model. The problem model's objective is influenced by the attributes picked. Considerations for responsiveness, exactness, precision, and temporal sophistication were made [1].

In order to reduce labelling costs while producing a more accurate classification, Nawel Zimal et al. (2020) provided a set of structures for merging the active learning (AL) and particle swarm optimization (PSO) techniques. Using sixteen benchmark datasets, the proposed solution was compared to three well-known classifiers: assistance vector machine (SVM), the extreme machine learning (ELM) with supervised instruction, and the transudative support vector machine (TSVM) with partially supervised learning. The Margi and Nave Base categories are utilized by each of these active learning algorithms. Studies demonstrated that their proposed method might reduce the time and labour professionals expended annotating clinical data in order to produce a viable classification [2,3].

Chun Guan et al. (2019) developed A dataset clustering method for finding groupings of any shape is spatial clustering based on density for application with noise (DBSCAN). To address the drawbacks of DBSCAN, the researchers developed the novel particle swarm optimized density-based clustering and classification (PODCC) method. Particle swarming optimization (PSO), an acknowledged evolving and swarm algorithm (ESA), is used to tackle various optimization issues, including data analytics. The recommended fitness function can be used as an argument to PODCC to determine the cluster count. The suggested technique was evaluated using ten artificially generated datasets and ten benchmark datasets from various open-access organizations [4].

Abdolreza Hatamlou and Asgarali Bouyer (2018): The K-means algorithm separates commodities into more discrete groups that share the most similarities with different goods in a particular category and the greatest variances from those products in other categories. Partition data clustering refers to this. K-harmonic means (KHM) developed as a groundbreaking categorization technique when combined with improved cuckoo search (ICS) and particle swarm optimization (PSO). By dynamically and cleverly adjusting the radius, ICS was able to use the Levy fly method to identify the best global solution. The suggested method was quicker than the conventional cuckoo chase. The ICS was prevented from reaching neighbourhood optima by PSO. Over time, the suggested ICMPKHM algorithm could address the KHM optimal localization problem more successfully [5].

Shubhra Biswal and Santosh Kumar Majhi (2018): An amalgamation clustering approach contingent on the K-means approach and ant-lion optimization has been demonstrated for efficient cluster analysis. The ant-lion optimizing (ALO) model is a stochastic global optimizing model. The effectiveness of the suggested method has been compared to that of the mathematical clustering approaches K-means, K-means-PSO, K-means-FA, DBSCAN, and enhanced DBSCAN using a number of performance indicators. Eight datasets were used in the experiment, and statistical evaluation was used to determine the outcomes. The total number of intra clustered interconnections and F-measure were higher for the K-means & ant-lion optimization algorithms than for the other two approaches [6].

According to Chun Guan (2018), a form of clustered approach that can find clusters of any design is density-based segmentation. A widely recognized density-based clustering method is the spatial clustering based on for applications with noise (DBSCAN) approach. Genetic algorithmic algorithms (GAs), particle swarm optimization (PSO), differentiated evaluations (DE), and artificial colonies of bees are a few examples of ESAs (ABC). In order to get over DBSCAN's limitations, the ESA-DCC architecture was combined in order to provide the appropriate settings for aggregated depending on density and categorization. The effectiveness of the K-means algorithm and the DBSCAN techniques were assessed against the outcomes of the ESA-DCC technique [7].

The cuckoo search approach with an adaptive density's development geographical clustering application with disturbances parameter was suggested by Limin Wang et al. (2018) as a quick solution to the overall optimization problem. The most effective global variable can be located using the cuckoo search technique. The improved approach can enable mechanization of the method of clustering and do away with the need for people to engage in it. The approach is capable of choosing a suitable Eps parameter value and providing incredibly precise clustering results, according to what they have learned from simulations [8–10].

3. Proposed methodology

Spatial clustering methods in genomics have proven to be highly effective in identifying patterns and relationships in gene expression data. By considering the spatial organization of genes within cells or tissues, these

methods can reveal important insights into the biological processes and functions of genes. Spatial clustering methods are particularly useful when studying single-cell RNA sequencing data, which provides high-resolution information about gene expression within individual cells.

It will go by means of all of the previously described clustering and clustering approaches, as well as the methods of inquiry used in this study, in this part. In an unstructured learning process called clustering, data items or trends are grouped based on similarity metrics. Objects exist as data points in Rd space. The features of entities inside a cluster are more comparable than those of entities within other clusters. In certain cases, such as collecting relevant articles for investigation, identifying protein and gene structures with similar activities, or compressing data, a technique called as clustering analysis is used to compress or better understand data in Figure 1 Proposed spatial cluster gene method.

The method for clustering is a widely used method for data analysis. Data retrieval, image processing, and pattern identification in machine learning are some uses for algorithms for clustering. It can also be used for data analysis. Each clustering algorithm has its own set of advantages and disadvantages. Like earlier clustering techniques, the density-based approach is simple and effective. Using density-based clustering techniques created to find clustering of any shape in noisy datasets, a cluster can be characterized as an extremely densely populated sector divided by low-density portions in data space.

3.1. Gene preprocessing

Preprocessing gene is an important step in studying microarray gene expression datasets [11–18]. Missing values in raw data from microarray studies are common, and can arise for a variety of reasons, including experimental error or technical limits. This can have an impact on the accuracy and dependability of gene expression analyzes. Failure to correctly preprocess the data can result in incorrect analysis and misleading outcomes. As a result, sufficient data preparation is required to provide a relevant study of gene expression patterns. The data pretreatment procedures employed are noted in the following lines. Dealing with a noisy dataset requires a combination of domain knowledge, data preprocessing techniques, and appropriate modelling approaches to obtain meaningful and reliable insights from the data.

3.2. Train-test split

This is a foundational ML technique for assessing the effectiveness of a model on unknown data. We can train the model on one part of the data and then evaluate how it performs on another fraction that the model has never seen before by splitting the data. The purpose is to determine how effectively the model will generalize to new, previously unseen data. Train-test splitting is employed in this work. This technique separates the data being gathered into two sections: one to be evaluated and the other for training. It assures that each set contains an equal number of examples from each type. The data has been separated in an 80–20 ratio, with 80% set aside for training and 20% set aside for testing.

3.3. Imputation of missing genes

Missing values can be found in some datasets. The KNN technique is used in this research to impute missing values by matching them with the mean figures of their closest neighbours in the training set. Two occurrences are considered similar in this example if their existing gene values are similar. If an instance lacks a class label, it is often optimized rather than imputed.

3.4. Data normalization

The purpose of data normalization is to convert a dataset into a standard format, which is often used to compare variables with varied units, sizes, or distributions. The precision and effectiveness of machine learning models can be enhanced by changing the values of features in a dataset to produce the same scale without misrepresenting fluctuations in value categories or losing information.

3.5. Clustering by simulated annealing

The aggregate amount of M time points per temporal-course gene manifestation profile is denoted by the letter N. Since we are primarily interested in the contours that depict the patterns of expression than exact quantities of production, each profile is normalized such that the total amount expressed ranges from 0 to 1. An M-dimensional vector, ei1, ei2, … , eiM, is used to represent each ith profile, with component eim corresponding to the normalized communication level of gene i at time point m ($0 \leq e^{i}{}_{m} \leq 1$). The similarity metric we use is the Euclidean distance, di j, between vectors i and j:

$$d_{ij} = \left[\sum_{m=1}^{M} e_m^i - e_m^j \right]^{1/2} \tag{1}$$

Using Equation (2), we optimize the distribution of profiles over a given number of clusters, K, by minimizing the sum of distances di, j inside clusters.

$$E(K) = \frac{1}{k} \sum_{k=1}^{k} \left[\sum_{j \in Ck} \sum_{j \in Ck} d_{i,j} \right] \tag{2}$$

where i ε Ck stands for vector i that belongs to the cluster number k. In Equation (2), the E-value must

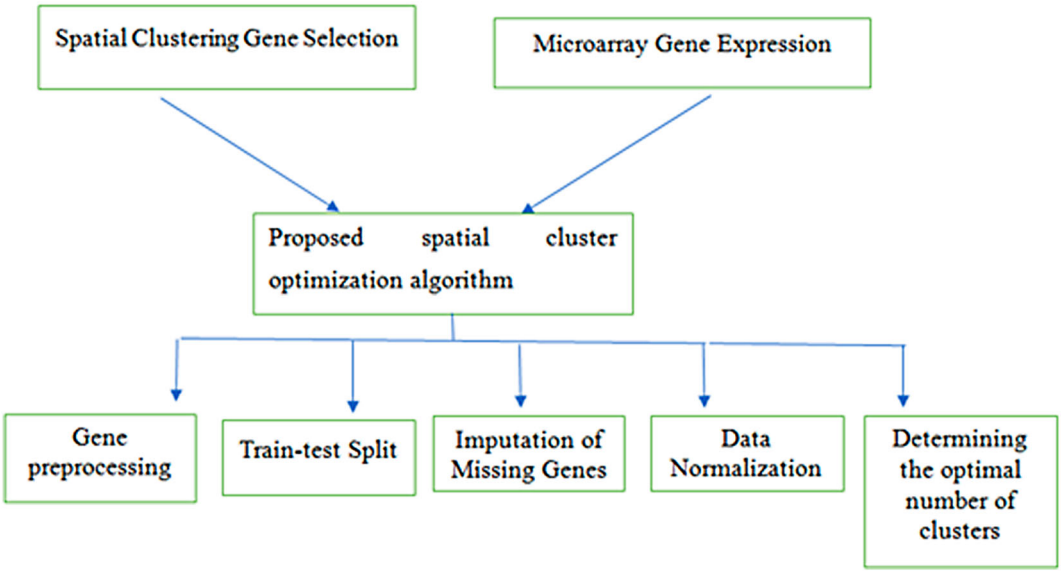

Figure 1. Proposed spatial cluster gene method.

be minimized. Vectors of information are initially distributed among clusters in an undetermined order. During each iterative phase, a randomly selected vector is removed from its neighbourhood and assigned to a different cluster. Eold, the previous value, is computed, and Enew is compared. The new vector allocation is used as the starting point for the subsequent repeat if Eold is bigger than Enew. Alternatively, if the E-value is used as the "energy," the newly assigned value is accepted with probability exp [(Enew Eold)/T], where T can be thought of as the "temperature" of the system in question. This method guarantees that every component of the system complies with the Boltzmann distribution at a particular temperature after initialization. Therefore, if T decreased gradually enough, the system would reach the global minimum of the E function despite avoiding local minima. We frequently employ an increasing cooling approach. $T_{n+1} = cT_n$, where n is the step number and the value $1 - c$ is positive and close to zero. It proved that, if the simulated annealing process were used to minimize the E function Equation (2), the E-value and the associated Agreements optimal distribution for genes over groupings would not depend on the selection of the random seed $1-c \leq 10^{-6}$.

3.6. Conceptual framework for determining the optimal number of clusters

The variance in attributes across a given data set clearly determines the appropriate number of clusters. One method to measure this variation is to use the function of distribution p(d) of the Euclidean distance between the vectors that represent the qualities in the understanding set. The normalization of the function p(d) results in a summing over all d of one. The greater the number of clusters necessary to generate tight clusters with distinct expression patterns, the larger the coefficient p(d). The standard deviation inside clusters may eventually grow to be less than the measurement error because K may someday grow to be so huge and closely grouped. However, increasing K beyond a certain amount of clusters K is often meaningless.

It specifies a cutoff difference D and assume that the assumption that each of the variables i and j are associated with the exact same cluster is false in order to address the difficulty of determining the appropriate number of subgroups in a quantitative manner if $d^i_j \geq D$. The significance of D might be better understood by taking an additional look at the connection connecting the variation in the value of the parameter p(d) and the measurement distance D. Assume that each gene has a single cluster. The integral of D determines the percentage of incorrect numerical pairs f (D, K = 1) for an instance of D.

$$f(D, K = 1) = \int_D^{\infty} p(x)dx \tag{3}$$

Specifically, the likelihood of locating two vectors that are spaced apart by an amount of time D or larger. Integral (3) offers the upper bound (in terms of the number among clusters K) for the percentage of improper vector pairings f (D, K). The probability of identifying an incorrect vector pair may often be estimated using the average, or weighted, fraction of faulty vector pairings for a given number of clusters K and an ideal vectors-to-cluster assignment:

$$f(D, k) = \frac{1}{K}\sum_{K=1}^{K} \frac{number\ of\ incorrect\ vector\ pairs\ in\ cluster\ \#\ k}{total\ number\ of\ vector\ pairsincluster\ \#\ k} \tag{4}$$

As the total amount of clusters K rises, this likelihood declines monotonously.

The lower threshold of the function f (D, K) can be calculated using an easy comparison, the pre-assignment of the *p*-value in typical statistical tests. The most vulnerable barrier is the maximum permissible possibility for the cluster to identify a pointless scalar pair. As a result, the conceptual framework that we have created in accordance with Equation (5) defines the ideal number of clusters as follows:

$$F(D, K) = P \tag{5}$$

The aforementioned equation can be solved given the parameters D and P by gradually increasing the number of subgroups K and using to reduce the amount technique of functional analysis (2) for each value of K to increase the percentage of mistaken a vector pairings f (D, K) until the *P*-value is reached. It is crucial to remember that the smaller the smallest allowed length D, the more clusters are needed to achieve the same fraction of incorrect vector pairings. The overall amount of clusters K that are produced by this procedure will, of course, depend on the values we choose for the variables D and P. As a result, choosing the right assortment of clusters now just requires balancing those two factors.

4. Results and discussion

Several charts presenting the outputs show the commencement results. The effectiveness of the two suggested technologies is compared in Table 1. It supported the clustering strategy presented here in two instances. The algorithm's effectiveness was thoroughly statistically verified through an indirect constructed experimental in which the right response was known in advance. The algorithm's clustering was shown to be physiologically significant when it was used to a gene expression study set (Cho et al., 1998) where appropriate grouping had previously been acknowledged using a variety of various ways, including examination by sight. Figure 2 shows a reverse engineering exercise. According to the article's remark, (A) the curving curve depicts the normalized distribution based on the separations between the 2000 profiles that were made to construct the 24 clusters. The left breadth and right peak of this function are the intersection of two overlapping curves that represent the total length between profiles of various clusters and the distances between features within regions, respectively. The dotted curve represents the shape of the distribution of the function for the same dataset with shuffled time points, which eliminates similarities between profiles within a cluster. The left maximum vanishes. (B) For the best characteristic distributions over clusters, Equation 4's proportion of incorrect pairs is used as a cluster number indication. Three distinct cutoff distance D values are shown by the narrow lines. The narrow straight line represents a *P*-value of 0.055.

Table 1. The contrast of the expected and calculated distributions of profiles over clusters.

	Expected		Calculated	
Cluster number	Number of profiles	Number of profiles	Missed	Added
1	10	16	2	8
2	20	19	0	4
3	30	20	1	5
4	40	45	0	6
5	50	56	0	4
6	60	58	0	7
7	70	62	1	2
8	80	69	0	5
9	90	75	2	7
10	100	80	2	4
11	110	85	2	5
12	120	89	1	9
sum	780	674	11	66

Experiment with reversal on the relationship bet ween characteristics D and P.

P is the percentage of permitted positives that are false that we randomly distribute within a permissible range. The value that is most frequently used is $P = 0.055$. The ideal value of the threshold range D is as follows after determining parameter P. Let's assume that already know Kopt, the number of clusters should be used for a given data collection. The equation f (D, Kopt) = P can be solved to determine the value of D. We employed the below-described inverse engineering strategy to accomplish this. The first step is the random generation of 24 expression seed structures with each interval being 10 times. A total of 2000 characteristics, with 10–200 characteristics for every arrangement, were created by dividing each pattern into individual accounts prior to clustering them; see Table 1 for the number pf profiles in each cluster. Although the weighted average variance from the beginning of the sequences inside groups is random at this stage, it is controlled to ensure that it stays under predetermined value SD. To approach the average variability mentioned in previous analysis of expression research, we used SD = 0.15 in our study.

It is obvious that the quantity of time points affects characteristic D. We now wonder how to calculate D for more pieces of information with various quantities of information points after computing D using our backwards data from engineering set. To do this, we employ the freely randomized data points and the standardized dispersion function of interactions among profiles. Equation (6) determines the likelihood Q that two separate profiles chosen at chance will be located within or equal to D of one another.

$$Q(D) = \int_0^D g(x)dx \tag{6}$$

if g(x) is the recognized distributions for describing the separations among profiles at two different times. The chance of discovering two randomly clustered qualities in a known set of information is expressed as Q(D).

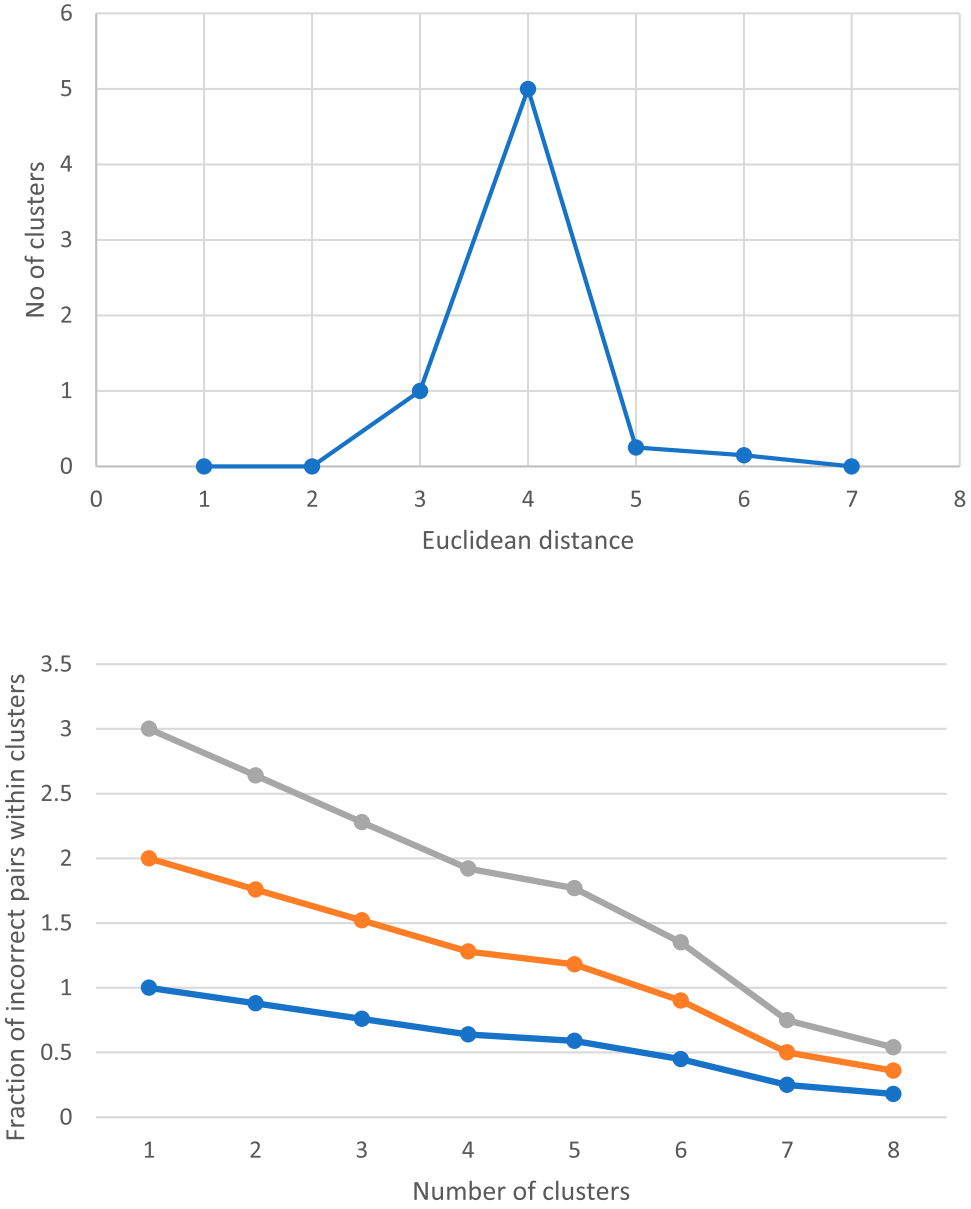

Figure 2. Reverse engineering experiment.

5. Conclusion

It presents an easy and reliable clustering technique that seeks to discover not only the best organization to express profiles across clusters, but also the most suitable number of groupings for a given data set. By doing trials backwards and analyzing the fundamental asset real-world experimental data, it assessed the effectiveness of the technique and helped us comprehend the biological implications of the outcomes. These methods use a collection of aggregate transcription pattern data as its input, and the output is a suppressing structure. It is obvious that the degree of complexity or ecological significance of the emerging regulatory network is significantly influenced by both the standard of the clustering procedure and the overall number of selected clusters. It believes that the approach demonstrated here, which addresses both problems, can be a helpful tool for locating and modelling a biologically significant regulatory network. The effectiveness of the proposed technique was evaluated using an experiment that discovers the most effective number of genes over time. This demanding statistical test was utilized to demonstrate that our approach accurately clusters more than 95% of the genes. Finally, because previous research has adequately described the underlying principles of gene growth and gene cluster assignment, the technique was validated using real gene expression data (expression fluctuations during the yeast cell cycle).

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

[1] Bouyer A, Hatamlou A. An efficient hybrid clustering method based on improved cuckoo optimization and modified particle swarm optimization algorithms. Appl Soft Comput. 2018;67:172–182. doi:10.1016/j.asoc.2018.03.011

[2] Guan C, Yuen KKF, Coenen F. Particle swarm Optimized Density-based Clustering and Classification: Supervised and unsupervised learning approaches. Swarm Evol Comput. 2019;44:876–896. doi:10.1016/j.swevo.2018.09.008
[3] Hu X, Liu L, Qiu N, et al. A MapReduce-based improvement algorithm for DBSCAN. J Algorithm Comput Technol. 2018;12:53–61. doi:10.1177/1748301817735665
[4] Majhi SK, Biswal S. Optimal cluster analysis using hybrid K-means and ant lion optimizer. Karbala Intern J Modern Sci. 2018;4:347–360. doi:10.1016/j.kijoms.2018.09.001
[5] Mohan S, Bhattacharya S, Kaluri R, et al. Multi-modal prediction of breast cancer using particle swarm optimization with non-dominating sorting. Intern J Distrib Sensor Netw. 2020;16.
[6] Zemmal N, Azizi N, Sellami M, et al. Particle swarm optimization based swarm intelligence for active learning improvement: application on medical data classification. Cognit Comput. 2020;12:991–1010. doi:10.1007/s12559-020-09739-z
[7] Deepthi P, Thampi SM. PSO based feature selection for clustering gene expression data. 2015 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES). 2015: 1–5.
[8] Mumtaz K, Duraiswamy K. An analysis on density based clustering of multi-dimensional spatial data. Ind J Comput Sci Eng. 2010;1:8–12.
[9] Li X, Zhang P, Zhu G. Dbscan clustering algorithms for Non-uniform density data and Its application in urban rail passenger aggregation distribution. Energies. 2019;12:3722–3722. doi:10.3390/en12193722
[10] Wang D, Tan D, Liu L. Particle swarm optimization algorithm: an overview. Soft Comput. 2018;22:387–408. doi:10.1007/s00500-016-2474-6
[11] Edwin Dhas P, Sankara Gomathi B. BI-clustering algorithm for microarray gene data based on the combination of FCM And Lion Optimization Algorithm. J Electr Eng. 2019;19(1):1–7.
[12] Edwin Dhas P, Sankara Gomathi B. A novel clustering algorithm by clubbing GHFCM and GWO for microarray gene data. J Supercomput. 2019: 1–15. doi:10.1007/s11227-019-02953-z
[13] Zareizadeh Z, Helfroush MS, Rahideh A, et al. A robust gene clustering algorithm based on clonal selection in multiobjective optimization framework. Expert Syst Appl. 2018;113:301–314. doi:10.1016/j.eswa.2018.06.047
[14] Zheng Y, Jeon B, Xu D, et al. Image segmentation by generalized hierarchical fuzzy C-means algorithm. J Intell Fuzzy Syst. 2015;28:961–973. doi:10.3233/IFS-141378
[15] Zhou K, Fu C, Yang S. Fuzziness parameter selection in fuzzy c-means: The perspective of cluster validation. Sci China Inf Sci. 2014;57(11):1–8.
[16] Yu Z, Chen H, You J, et al. Adaptive fuzzy consensus clustering framework for clustering analysis of cancer data. IEEE/ACM Trans Comput Biol Bioinform. 2014;12(4):887–901.
[17] Swathypriyadharsini P, Premalatha K. Triocuckoo: A multi objective cuckoo search algorithm for triclustering microarray gene expression data. J Inform Sci Eng. 2018;34(6):1617–1631.
[18] Suo Y, Liu T, Jia X, et al. Application of clustering analysis in brain gene data based on deep Learning. IEEE Access. 2019;7:2947–2956. doi:10.1109/ACCESS.2018.2886425