Turkish Journal of Electrical Engineering & Computer Sciences

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

Research Article

Turk J Elec Eng & Comp Sci
(2017) 25: 3443 – 3454
© TÜBİTAK
doi:10.3906/elk-1609-163

E-MFDBSCAN: an evolutionary clustering algorithm for gene expression time series

Atakan ERDEM[*], **Taflan İmre GÜNDEM**
Department of Computer Engineering, Boğaziçi University, İstanbul, Turkey

Received: 19.09.2016 • **Accepted/Published Online:** 05.02.2017 • **Final Version:** 30.07.2017

Abstract: DNA microarray experiments are frequently used because they have various advantages. However, gene expression data from DNA microarray experiments are noisy, and, consequently, the computations that are based on such noisy data may lack accuracy. In this paper, an evolutionary uncertain data-clustering algorithm, *E-MFDBSCAN*, and a prediction model using *E-MFDBSCAN* for uncertain data are proposed. The proposed methodology may be successfully applied to noisy gene expression data. In this methodology, global patterns of time series data can be extracted using our evolutionary clustering approach. These patterns are used to infer future projections. In the proposed methodology, an autoregressive time series function (using these patterns) used to predict the similarities among sets of gene expression clusters is constructed. The algorithms are tested with two different gene expression time series datasets.

Key words: Microarray experiment, gene expression, evolutionary clustering, prediction, uncertain data, time series

1. Introduction

Due to the biological variations and the nature of probe-level measurements, missing and noisy data are major problems in microarray experiments. Several studies have focused on this problem. One of the relevant studies [1] proposes a technique applied to microarray experiment datasets in which the fuzzy set logic is used for gene selection. In the work by Troyanskaya et al. [2], three missing-data estimation methods, K-nearest neighbors, singular value decomposition, and row average, were tested and evaluated using three different gene expression datasets. To predict the missing gene expression time series data, linear interpolation was proposed by Aach and Church [3]. Similarly, D'haeseleer [4] proposed a spline interpolation method to estimate the missing time points. Bar-Joseph et al. [5] modeled the gene expression profiles as a cubic spline. Using spline curves, the missing data are estimated. However, because the measured data are fuzzy, even if the missing data are estimated somehow, they are still uncertain. Therefore, Sivriver et al. [6] proposed a dynamic modeling approach, DynaMiteC, to overcome both noisy and missing data problems. The proposed approach enables time course gene expression profiles to be modeled and clustered using biologically meaningful parameters (e.g., point of induction). With the same motivation, another modeling approach, an extension of the Gaussian mixture model named PUMA-CLUST, was proposed by Liu [7]. Because of the use of manual methods in some parts of gene expression experiments, the reliability of the data is low. If these data are directly used as inputs to a data-mining algorithm or a model to evaluate gene expression data, then the adverse effects on the desired results will be inevitable. To eliminate the aforementioned adverse effects and reduce the fuzziness, the data are represented using sample datasets generated using uncertain data-management techniques.

[*]Correspondence: atakan.erdem1971@gmail.com

To vigorously understand and analyze the global patterns of gene activities time series-based dynamic data processing should be used to extract relevant information. Thus, recent studies have focused on time series-based dynamic modeling and clustering. Holter and Maritan [8] proposed a method that uses a time translational matrix. Using a time translational matrix, temporal relationships can be modeled. However, because this method focuses on the time points that are sampled at the lowest common frequencies, the available expression data are only partially used. Zhao et al. [9] proposed statistical models to extract the genes that are regulated by the cell cycle. For each periodically analyzed specific dataset, a custom model is generated, which is the distinct weakness of their proposed method. Aach and Church [3] focused on aligning gene expression time series. These authors propose time-warping methods instead of clustering. Global pattern extraction and prediction are not possible with these algorithms. Therefore, evolutionary clustering and unsupervised learning approaches are more suitable.

In this paper, a prediction model for gene expression time series is proposed. We developed an evolutionary uncertain data clustering algorithm, evolutionary M-FDBSCAN (E-MFDBSCAN), and used it in a model to predict the similarity value of a gene expression cluster for the next time point. In addition, because E-MFDBSCAN is a density-based clustering algorithm, it returns more accurate results for noisy datasets [10] compared to the other algorithms that are based on different approaches. The other strength of the proposed algorithm is the global clusters that are generated according to the time-based evolutionary information. The proposed prediction model allows for the prediction of the next time points by modeling the similarity information of the set of global clusters for n time points.

The paper is organized as follows. In Section 2, the related works are presented. In Section 3, elaborated explanations of the proposed evolutionary clustering algorithm and the prediction model are given. In Section 4, the results of tests using two different gene expression time series datasets are discussed. In Section 5, the conclusion of the study is given.

2. Related work

In the case of gene expression clustering, several techniques are widely used. Hierarchical clustering is the most commonly used technique. The main drawbacks of this approach are a lack of robustness, nonuniqueness, complicated hierarchy interpretations due to inversion problems, and local decision-based grouping without clustering reevaluation capability [11]. The nonuniqueness problem of hierarchical clustering approach is studied in [12]. For this purpose, a leaf ordering algorithm is proposed for preserving the clustering result based on the dynamic programming concept. Another interesting study [13] proposed a noise-aware method (a derivation of nonnegative matrix factorization (NMF) called PNMF) for clustering and classifying microarray data, but the solution is not a time-aware solution. The other widely used technique is density-based clustering. In density-based clustering algorithms, the essential task is to discover high-density regions that are separated by low-density regions in a data space. This approach is more robust against noisy and diffused data. Thus, the unambiguous clustering performance is better than that of the other approaches. There are several density-based clustering algorithms, including K-means [14], self-organizing maps (SOMs) [15], and density-based spatial clustering of applications with noise (DBSCAN) [16]. One of the most popular density-based gene expression clustering algorithms is SOMs. To observe the pattern interpretation performance of SOMs, Tamayo et al. [11] developed the computer package GENECLUSTER. According to their execution results, SOMs are well suited for exploratory data analysis. Similarly, Fang et al. [17], developed a computer package, supraHex, to train, analyze, and visualize omics data, an implementation of a derivation of the SOMs algorithm. Despite its popularity, the SOMs approach has two major drawbacks. One drawback is the number of incorrect clusters in the case of noisy and missing data and the other is the high computational cost.

In the present study, the evolutionary clustering algorithm E-MFDBSCAN, which is a density-based clustering algorithm, was developed. E-MFDBSCAN is devised for uncertain data clustering by modeling the uncertainty using a Gaussian probability density function. Thus, the cluster accuracy performance is better than that of SOMs. The other advantage of E-MFDBSCAN is the enhanced computational time performance. Because E-MFDBSCAN is a derivation of M-FDBSCAN [18], which is devised for multicore systems, clustering issues can be resolved in each subdataset concurrently by splitting a gene expression dataset into subdatasets.

Gene expressions change, especially during transcription and translation. If clustering is performed for only static time points, the generated clusters help to deduce only instantaneous and local information about the overall cellular events. Therefore, to see the whole picture, time-aware global clustering approaches are required.

Subhani et al. [19] proposed an evolutionary clustering algorithm that combines the approaches of expectation maximization and the multiple alignments of gene expression profiles to cluster microarray time series data. The algorithm uses the k-means clustering algorithm, which is a centroid-based clustering algorithm. However, the main drawback of the centroid-based representation is that the number of clusters must be specified in advance. Because of the nature of the data domain, determining the number of clusters before the clustering process is difficult.

E-MFDBSCAN generates time-aware global clusters. Thus, it enables future prediction using the extracted time-based global patterns. In this algorithm, uncertainty is modeled by generating several sample points for each gene expression data point using a Gaussian probability distribution function. According to the number of sample points, the original dataset enlarges. However, the sample data approach not only reduces the ratio of fuzziness but also increases the output generation time due to an increase in the amount of processed data, which is directly proportional to the cardinality of the sample dataset. Because E-MFDBSCAN supports parallel computation, this increase in data size can be managed effectively.

3. Materials and methods

In this section, the proposed evolutionary clustering algorithm E-MFDBSCAN and the proposed prediction model are presented.

3.1. E-MFDBSCAN

E-MFDBSCAN is an evolutionary uncertain-data–clustering algorithm based on the DBSCAN clustering approach. The uncertain-data–clustering method proposed in [20] and the parallel computation approach proposed in [18] are used in E-MFDBSCAN. According to the density-based clustering approach, there are two main constraints: i) each cluster has at least a number of μ members, and ii) the distance between any two members in a cluster is not larger than ε. In the context of gene expression, the cluster membership and the distance between two data objects are fuzzy. In [20], to address the uncertainty, a data-sampling technique is proposed. According to their proposal, for each uncertain data object x, a sequence of s sample points, $< x_1, \ldots, x_s >$ is derived using a Gaussian probability density function. The proposed clustering method is based on the sample matrix concept. A sample matrix is used to determine the neighbors of the fuzzy data objects. If a fuzzy data object has no neighbor, then it is flagged as an outlier fuzzy data object.

Each row o_i in the sample matrix represents the derived sample data points of o, where $\{o_i| <$

$o_1, \ldots, o_i, \ldots, o_s >\}$. Each column D_j represents the jth database instance, where $\{x_j| < x_1, \ldots, x_j, \ldots, x_s > \epsilon D \wedge x_j \neq o_j\} \cup o_i$. Finally, each matrix element is $m_{i,j} = |N_{\varepsilon,Dj}(o_i)|$, where $N_{\varepsilon,Dj}(o_i)$ denotes the set $\{x_j|d(o_i, x_j \leq \varepsilon \wedge x_j \epsilon D_j\}$, and $d(.,.)$ denotes the distance function.

In Figure 1, a sample 4 × 4 sample matrix is shown. For example, according to the figure, $m_{11}= 4$, indicating the existence of 4 fuzzy data objects in D, of which the first sample points are in the ε-neighborhood of the first sample point o_1 of o.

$$\mathbf{SM(o)}$$

$$\begin{array}{c} \\ \mathbf{o1} \\ \mathbf{o2} \\ \mathbf{o3} \\ \mathbf{o4} \end{array} \begin{array}{cccc} \mathbf{D1} & \mathbf{D2} & \mathbf{D3} & \mathbf{D4} \\ \end{array} \left[\begin{array}{cccc} 4 & 7 & 5 & 9 \\ 6 & 11 & 8 & 5 \\ 7 & 3 & 9 & 7 \\ 2 & 0 & 5 & 6 \end{array}\right]_{s=4}$$

Figure 1. A 4 × 4 sample matrix.

Because the proposed evolutionary clustering model is developed for gene expression time series data, the sample matrix concept is extended by adding the time dimension. The new matrix, called the t-sample matrix, is a three-dimensional $m \times s \times s$ matrix, where m denotes the number of time points in the time dimension, and s denotes the number of derived sample points for the fuzzy data objects. In E-MFDBSCAN, for each fuzzy data object x, a sequence of $m \times s$ samples, $< x_{11}, \ldots, x_{m1}, \ldots, x_{ms} >$ is derived using a Gaussian probability density function. Similar to the sample matrix, in the t-sample matrix,o_{ki} represents the ith derived sample point for the kth time point, such that $\{o_{ki}| < o_{11}, \ldots, o_{1s}, \ldots, o_{ki}, \ldots, o_{ms} > \epsilon o\}$, and D_{kj} represents the jth database instance for the kth time point, where $\{x_{kj}| < x_{11}, \ldots, x_{1s}, \ldots, x_{kj}, \ldots, x_{ms} > \epsilon D \wedge x_{kj} \neq o_{kj}\} \cup o_{ki}$. Finally, each matrix element is $m_{k,i,j} = |N_{\varepsilon,Dkj}(o_{ki})|$, where $N_{\varepsilon,Dkj}(o_{ki})$ denotes the set $\{x_{kj}|d(o_{ki}, x_{kj}) \leq \varepsilon \wedge x_{kj} \in D_{kj}\}$, and $d(.,.)$ denotes the distance function. In Figure 2, to give an idea, a 5 × 3 × 3 t-sample matrix is illustrated.

In the first step of the algorithm, the t-sample matrices are constructed for all of the fuzzy data objects. A t-sample matrix T-SM(o), where $o$is any fuzzy data object, is a data structure that is used to compute the reachability probabilities between $o$and the other fuzzy data objects in the database. In short, the reachability probability with respect to $o$and $x$denotes the probability of $o$and $x$being in the same cluster. Assigning an

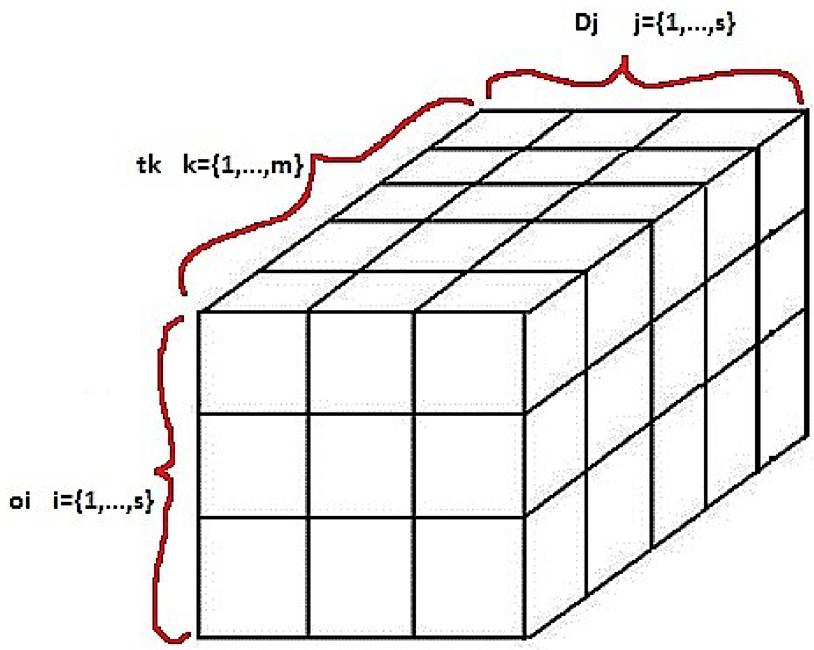

Figure 2. A t-sample matrix with $m = 5$ time points and $s = 3$ sample data points.

object to a cluster or flagging it as an outlier is performed according to the value of the reachability probability.

$$P^{reach}(x,o) = P^{core}(o) \times P_d(x,o)(\varepsilon) \tag{1}$$

In Eq. (1), $P^{core}(o)$ represents the core object probability of o, which denotes the probability of $o$having at least $\mu{-}1$ neighbors in ε-neighborhood, except for x. This probability value is derived using the t-sample matrix. The derivation procedure is customized as follows:

```
Core object probability
Begin
  1. For each time point $t_k$, where $\{t_k| < t_1, \ldots, t_{m>}\}$ :
     1. Decrease $m_{k,i,j}$ by 1 for which $d(o_{ki}, x_{kj}) \leq \varepsilon$ holds.
     2. Count the number of elements in the t-sample matrix T-SM($o$) that contain values greater
        than or equal to $\mu-$ 1.
     3. Normalize the result by $s^2$.
     4. Assign the final normalized result to $P^{core}(o)_k$
  2. Count the core probabilities that are greater than or equal to 0.5
  3. Normalize the result by $m$.
  4. Assign the final normalized result to $P^{core}(o)_{final}$.
End
```

Finally, the next probability that must be computed is the distance distribution probability $P_d(x,o)(\varepsilon)$ with respect to x and o, which denotes the probability of x being in the ε-neighborhood of o. As in the core object probability computation, the computation of $P_d(x,o)(\varepsilon)$ is customized for *E-MFDBSCAN* as follows:

```
Distance distribution probability
Begin
  1. For each time point $t_k$ of the fuzzy data object $x$, where $\{t_k| < t_1, \ldots, t_{m>}\}$ :
      1. Count the number of events $d(o_{ki}, x_{kj}) \leq \varepsilon$.
      2. Normalize the result by $s^2$.
      3. Assign the final normalized result to $P_d(x, o)(\varepsilon)_k$
  2. Count the distance distribution probabilities that are greater than or equal to 0.5
  3. Normalize the result by $m$.
  4. Assign the final normalized result to $P_d(x, o)_{final}$.
End
```

Thus, the final reachability probability is

$P^{reach}(x, o)_{final} = P^{core}(o)_{final} \times P_d(x, o)_{final}$.

If $P^{reach}(x, o)_{final} \geq 0.5$, then it is accepted that $x$is reachable from o, also indicating that $x$and $o$are in the same cluster. If for all x values $P^{reach}(x, o)_{final} < 0.5$, then $o$is flagged as an outlier. With this new perspective proposed by E-MFDBSCAN, time-based global clusters and outliers can be generated.

Because standard distance functions, such as Euclidian distance functions, are mostly used for objects in n-dimensional spaces ($n \geq 2$), they are not adequate for defining the correlations among the gene expression data, which is in 1D space. Therefore, the Pearson correlation function has been used in many gene expression studies [6,21,22]. Thus, the Pearson correlation was used to measure distance.

As seen in Figure 3, for each time point, the gene expression data lie on a vertical line, and each fuzzy data object is represented with a sequence of derived sample points.

If E-MFDBSCAN is assumed as a function, then it returns a set of sets of clusters SC such that $SC = \{SC_1, \ldots, SC_i, \ldots, SC_m\}$, where m denotes the number of time points, and SC_i denotes the set of clusters that are generated at the time point t_i using the database $D_i \subseteq D$, where D_i denotes the database in which the data for the time points $< t_1, \ldots, t_i >$ are stored.

3.2. The prediction model

The stable pattern of gene expression time series data is an advantage with respect to near-future time prediction. For most of the other data domains, as mentioned in [23], the risk of erroneous prediction is high due to the instability of the trend of the time series. Recently, several studies have focused on gene expression profile prediction. The dynamic model proposed by Holter and Maritan [8] is capable of class prediction. The class prediction model proposed by Sorlie et al. [24] is based on prediction analysis of microarrays, which is a variant of the nearest-centroid classification approach. Another classification and prediction solution based on the nearest-centroid approach was proposed by Tibshirani [25]. The common restriction of these models is that they are applicable only when the classes are known and defined. Unfortunately, in most cases, the classification of gene expression profiles is challenging. As mentioned in [25], because the number of genes to be classified and predicted is much greater than the number of samples that are analyzed by microarray experiments, selecting the best-fit classes for each gene is nearly impossible. In addition, the significant identification of the genes that

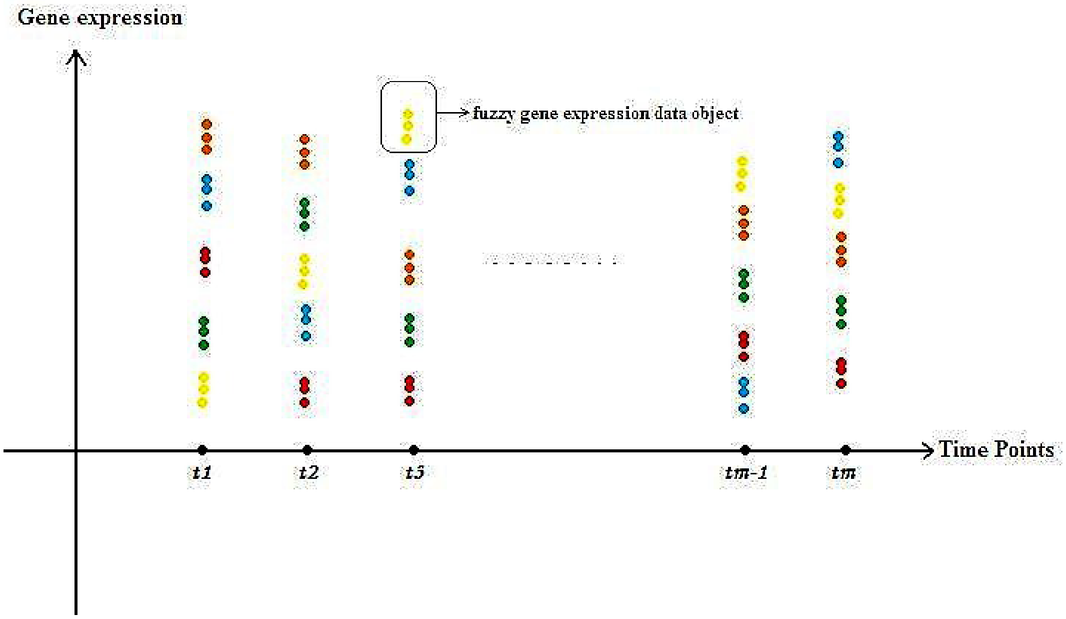

Figure 3. A gene expression time series dataset.

contribute to this classification is difficult. For these reasons, clustering-like unsupervised learning methods are widely used to group gene expression profiles.

In several studies, evolutionary clustering approach-based prediction models have been proposed. The evolutionary clustering approaches in which the time parameter is not considered, such as in EvoCluster, proposed by Ma et al. [26], are not suitable for future time prediction models.

In the present study, a prediction model for gene expression time series is proposed. As illustrated in Figure 4, one of the crucial benefits of *E-MFDBSCAN* is the demonstration of the temporal evolution of the gene expression clusters.

To use the proposed prediction model, a similarity measure between two timely adjacent sets of clusters is defined. The obtained similarity values determine the level of affinity of the two sets of clusters. The proposed prediction model is represented by the time-series function, which is constructed according to the autoregressive model concept and the achieved time-based similarity values.

$$\begin{aligned} &Sim(E-MFDBSCAN_{i+1}, E-MFDBSCAN_i) = A + B \times Sim(E-MFDBSCAN_i,\\ &E-MFDBSCAN_{i-1}) + E_{i+1}(i = 1, \dots, m) \end{aligned} \tag{2}$$

In Eq. (2), *Sim* represents the BestMatch similarity function as defined by Goldberg [27], which is used to compute the similarity values between two timely adjacent sets of clusters. A and B are the coefficients of the time series function, and E is the noise function. The index i represents the ith time point. To observe the prediction performance of the model, the similarity value for the next future time period $[t_{m-}t_{m+1}]$ is obtained from the time series function. Then the obtained similarity value is compared to the similarity value by executing E-MFDBSCAN for the time points t_m and t_{m+1}. Notice that, although t_{m+1} represents the next future time point, the gene expression data at this time point are known and reserved in the database for testing issues. The difference between these two similarity values determines the performance of the prediction model.

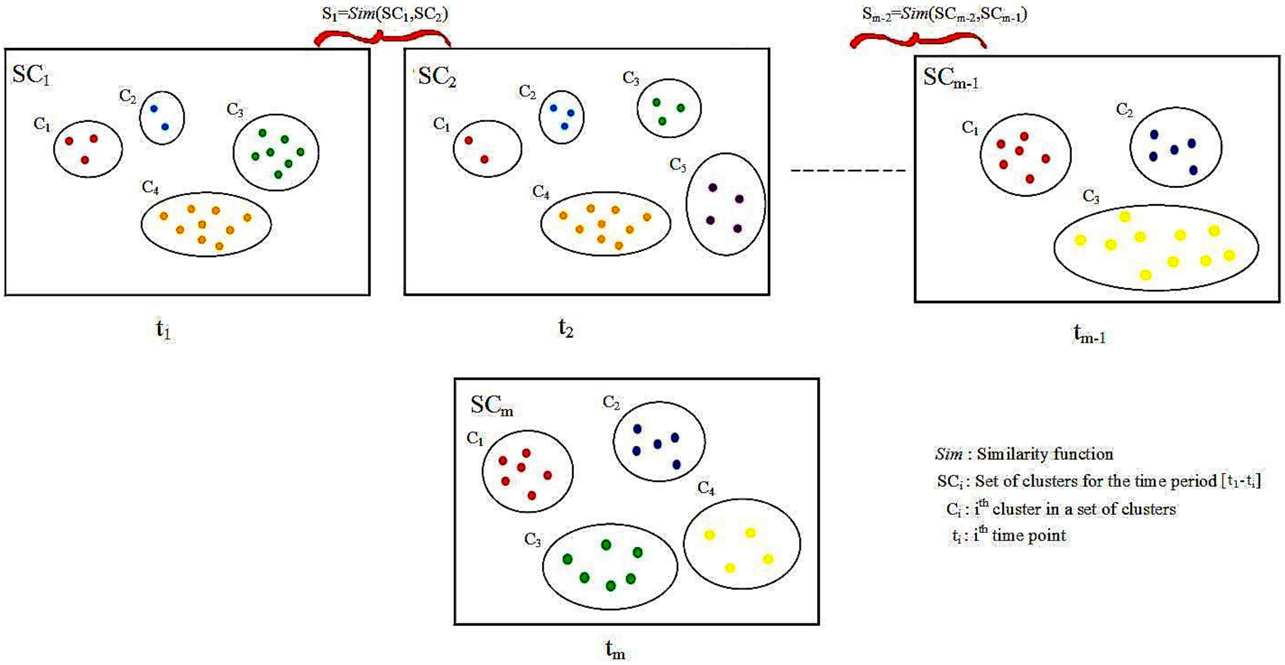
S1=Sim(SC1,SC2)
Sm-2=Sim(SCm-2,SCm-1)
SC1
C1
C2
C3
C4
t1
SC2
C1
C2
C3
C4
C5
t2
SCm-1
C1
C2
C3
tm-1
SCm
C1
C2
C3
C4
tm
Sim : Similarity function
SCi : Set of clusters for the time period [t1-ti]
Ci : ith cluster in a set of clusters
ti : ith time point

Figure 4. Time-based evolutionary changes of sets of clusters.

4. Results and discussion

The tests were done on a server with an Intel Xeon X5650 2.67 GHz 24 Core CPU and 72 GB RAM. The operating system was 64-bit Linux Centos 5.11. The algorithm and all other coding issues were implemented in ANSI C programming language.

Because the time series of gene expression is generally very short (i.e. 4 to 20 samples) and unevenly sampled, extracting the time-based evolutionary patterns is difficult. In this study, two different time series gene expression datasets are used to test the issues surrounding evolutionary clustering and prediction. The evolutionary clustering tests are used to observe the patterns of the generated global clusters. For each time point i, the set of global clusters SC_i is established, including the time period $[t_1 - t_i]$. Thus, for $m$time points, m sets of global clusters are established such that $\{SC_i|\ SC_1, \ldots, SC_i, \ldots, SC_m\}$. For predicting and obtaining the similarity values of each of the two timely adjacent sets of global clusters, the coefficients and the noise of the time series function as given in Eq. (2) are defined. The performance tests of the prediction capability of the model are performed due to the similarity values that are obtained from the time series function. Then these values are compared to the similarity values using E-MFDBSCAN for two timely adjacent time points. The comparison results for the used datasets are given in Figures 5 and 6.

The first dataset is a budding yeast dataset, for which the gene expression time series data were obtained during the yeast cell cycle. The dataset also includes the class information of the achieved data. The expression levels were observed for 6220 yeast genes. At 10 min intervals, from time points 0 to 160 min, a total of 17 distinct temporal data were obtained for each gene. The used clustering algorithm is described in [28]. This dataset was extended by deriving 8 sample points for each gene expression data point. The input parameters ε (maximum distance between two members in a cluster) and μ (minimum number of members in a cluster) are derived from the given class information. The values of the parameters that were derived from the class information are $\varepsilon = 0.005$ and $\mu = 5$. Starting from the 0th time point, for each time point of 17 time points,

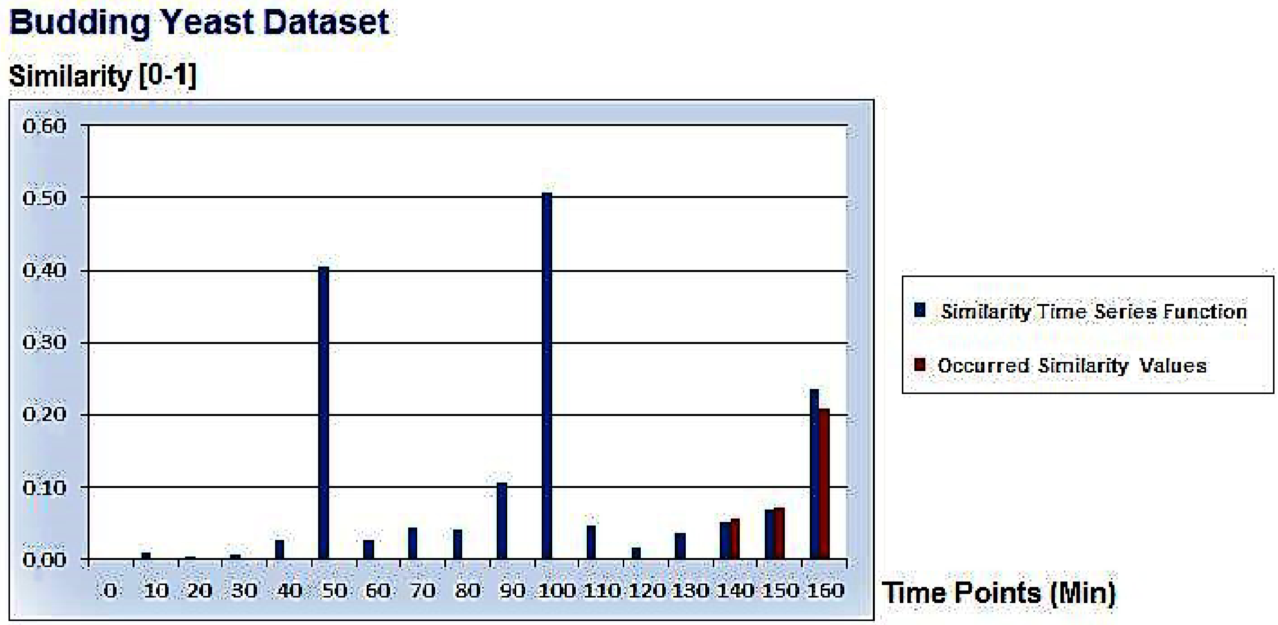

Figure 5. Similarities versus time graph for a budding yeast dataset.

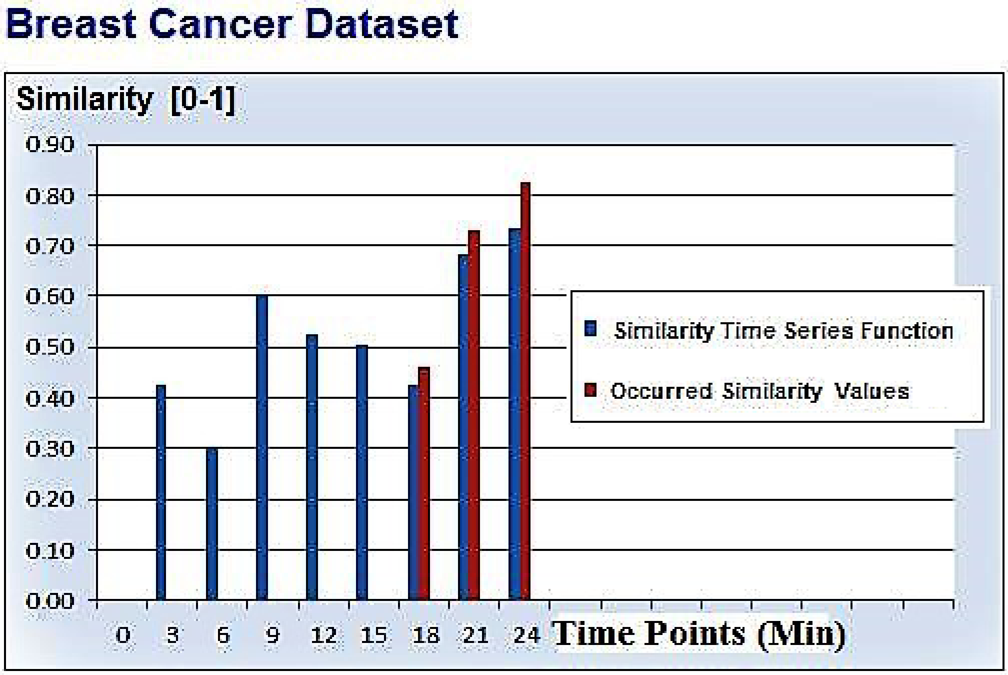

Figure 6. Similarities versus time graph for a breast cancer dataset.

E-MFDBSCAN is used to generate the sets of global clusters. After computing the similarity values for each set of clusters pairs of adjacent time points, we obtain the set of similarity values S. Using S, the unknown coefficients in Eq. (2) are extracted. Finally, a time series function is established for the budding yeast dataset. During the tests, the similarity values for the last three time points are predicted.

As shown in Figure 5, the actual and predicted value bars are approximately the same height. Due to the stable pattern of the gene expression time series data, after the construction of a model with a reasonable number of time points, satisfactory prediction results can be generated.

The next dataset is a breast cancer gene expression time series dataset. The data are the result of an analysis of breast cancer MCF10A-Myc cells at four time points up to 24 h following treatment with dexamethasone to activate the glucocorticoid receptor. This dataset is composed of 22,283 gene expression time series values. For each gene expression data point, again, 8 sample points are derived using a Gaussian probability density function. In contrast to the first dataset, the classes in this dataset are not defined. Thus, the tests are performed using several μ and ε parameter values. The relevant graph is shown in Figure 6. As in the previous experiment, for the last three time points, the actual and predicted values are compared. Because the number of time points is smaller than in the first dataset, the fitting ratio of the time series function is low. Therefore, the differences between the predicted and actual values are slightly greater than in the first dataset.

5. Conclusion

In this study, a methodology to predict the evolutionary patterns of gene expression time series data is proposed. An evolutionary clustering algorithm (E-MFDBSCAN) and a prediction model are developed. The distinct contribution of the methodology is the proposed comprehensive solutions for several problems in gene expression data processing, including uncertain data management, evolutionary pattern extraction, and near-future time prediction. There are few studies that concentrate on a complete solution from this perspective. Most of these studies focus on a specific problem in gene expression data processing. In most of these studies, the prediction is performed by means of class prediction. However, this approach requires the classes to be previously known. Because the samples from microarray experiments are smaller than the genes to be classified in most cases, the provided classes are not sufficient for a comprehensive classification. Therefore, instead of supervised learning approaches, unsupervised learning approaches are preferred in most of the recent studies. In some studies, evolutionary clustering methods have been proposed for gene expression data. However, these methods are incapable of time-based evolutionary pattern extraction.

In the clustering part of the methodology, an evolutionary clustering algorithm is presented. E-MFDBSCAN generates time-based global gene expression clusters. The other features of the algorithm are i) parallel execution capability and ii) uncertain data consideration and management. According to the results of our literature survey, there is no other clustering algorithm that gathers all of the aforementioned features into a single algorithm.

In the prediction part of the methodology, a prediction model is presented. The similarity values between two timely adjacent sets of global clusters, which were generated by E-MFDBSCAN, are used to construct an autoregressive time series function. The model is verified by comparing the predicted and the actual similarity values for the same time period. The predicted similarity values are derived from the function, and the actual similarity values are achieved by executing E-MFDBSCAN for two adjacent time points. The prediction performance of the model depends on the size of the training dataset. Unfortunately, in the context of gene expression data, the time series are generally short, meaning that the training dataset has only a few time points (e.g., 4 to 20). However, despite this essential drawback, due to the stable pattern of gene expression time series data, the prediction results are satisfactory.

Although the methodology is applied to the gene expression data domain, it can be used for other data domains, such as social media. In future studies, the model will be tested for social media analysis.

References

[1] Wojcik PI, Ouellet T, Balcerzak M, Dzwinel W. Identification of biomarker genes for resistance to a pathogen by a novel method for meta-analysis of single-channel microarray datasets. J Bioinform Comput Biol 2015; 13: 1-19.

[2] Troyanskaya O, Cantor M. Missing value estimation methods for DNA microarrays. Bioinformatics 2001; 17: 520-525.

[3] Aach J, Church GM. Aligning gene expression time series with time warping algorithms. Bioinformatics 2001; 17: 495-508.

[4] D'haeseleer P, Wen X, Fuhrman S, Somogyi R. Linear modeling of mRNA expression levels during CNS development and injury. In: Proceedings of the 4th Pacific Symposium on Biocomputing; 4–9 January 1999; The Big Island, HI, USA: pp. 41-52.

[5] Bar-Joseph Z, Gerber G, Gifford DK, Jaakkola TS, Simon I. A new approach to analyzing gene expression time series data. In: Proceedings of the Sixth Annual International Conference on Computational Biology; 18–21 April 2002; Washington, DC, USA: ACM. pp. 39-48.

[6] Sivriver J, Habib N, Friedman N. An integrative clustering and modeling algorithm for dynamical gene expression data. Bioinformatics 2011; 27: 1392-1400.

[7] Liu X, Lin KK, Andersen B, Rattery M. Including probe-level uncertainty in model-based gene expression clustering. BMC Bioinformatics 2007; 8: 98-116.

[8] Holter NS, Maritan A. Dynamic modeling of gene expression data. Proc Natl Acad Sci USA 2001; 98: 1693-1698.

[9] Zhao LP, Prentice R, Breeden L. Statistical modeling of large microarray data sets to identify stimulus-response profiles. Proc Natl Acad Sci USA 2001; 98: 5631-5636.

[10] Mumtaz K, Duraiswamy K. An analysis on density based clustering of multi-dimensional spatial data. IJCSE 2010; 1: 8-12.

[11] Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Dmitrovsky E, Lander ES, Golub TR. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA 1999; 96: 2907-2912.

[12] Novoselova N, Wang J, Klawonn F. Optimized leaf ordering with class labels for hierarchical clustering. J Bioinf Comput Biol 2015; 13: 1-19.

[13] Bayar B, Bouaynaya N, Shterenberg R. Probabilistic non-negative matrix factorization: theory and application to microarray data analysis. J Bioinf Comput Biol 2015; 12: 1-25.

[14] Kaufman L, Rousseeuw PJ. Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken, NJ, USA: John Wiley & Sons, 1990.

[15] Kohonen T. Self-Organizing Maps. Berlin, Germany: Springer, 1997.

[16] Ester M, Kriegel H, Sander J, Xiaowei X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining; 2–4 August 1996; Portland, OR, USA: pp. 226-231.

[17] Fang H, Gough J. supraHex: an R/Bioconductor package for tabular omics data analysis using a supra-hexagonal map. Biochem Bioph Res Co 2014; 443: 285-289.

[18] Erdem A, Gündem Tİ. M-FDBSCAN: A multi core density-based uncertain data clustering algorithm. Turk J Elec Eng & Comp Sci 2014; 22: 143-154.

[19] Subhani N, Rueda L, Ngom A, Burden C. Clustering microarray time-series data using expectation maximization and multiple profile alignment. In: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine Workshops; 1–4 November 2009; Washington, DC, USA: IEEE. pp. 2-7.

[20] Kriegel HP, Pfeifle M. Density-based clustering of uncertain data. In: Proceedings of the Eleventh International Conference on Knowledge Discovery and Data Mining; 21–24 August 2005; Chicago, IL, USA: ACM. pp. 672-677.

[21] Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, Boe AF, Boguski MS, Brockway KS, Byrnes EJ et al. Genome-wide atlas of gene expression in the adult mouse brain. Nature 2007; 445: 168-176.

[22] Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 2000; 406: 536-540.

[23] Li G, Cai Z, Kang X, Wu Z, Wang Y. ESPSA: A prediction-based algorithm for streaming time series segmentation. Expert Syst Appl 2014; 41: 6098-6015.

[24] Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci USA 2000; 100: 8418-8423.

[25] Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci USA 2002; 99: 6567-6572.

[26] Ma PCH, Chan KCC, Xin Y, Chiu DKY. An evolutionary clustering algorithm for gene expression microarray data analysis. IEEE T Evolut Comput 2006; 10: 296-314.

[27] Goldberg MK, Hayvanovych M, Ismail MM. Measuring similarity between sets of overlapping clusters. In: Proceedings of the Second International Conference on Social Computing, SocialCom/International Conference on Privacy, Security, Risk and Trust; 20–22 August 2010; Minneapolis, MN, USA: IEEE. pp. 303-308.

[28] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1999; 95: 14863-14868.