STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LIX**, Number 1, 2014

A STUDY ON DYNAMIC CLUSTERING OF GENE EXPRESSION DATA

ADELA-MARIA SÎRBU

Abstract. Microarray and next-generation sequencing technologies allow measuring the levels of expressions of thousands of genes simultaneously. One of the most popular procedures used to analyze gene expression data is clustering. To study biological processes which evolve over time, researchers can either perform re-clustering from scratch every time new gene expression levels are available, which would be very time consuming, or adapt the previously obtained partitions using a dynamic clustering algorithm. This paper aims to investigate a couple of heuristics for centroids identification within a dynamic k-means based clustering algorithm that was previously introduced for clustering of gene expression data. Computational experiments on a real-life gene expression data set are provided, as well as an analysis of the obtained results.

1. Introduction

The emergence of microarray and next-generation sequencing technologies that allow measuring the levels of expressions of thousands of genes has lead to an exponential increase of the amount of gene expression data. In order to extract useful biological information from this data, exploratory analyses are performed. A first step in these analyses is clustering.

Clustering refers to creating a set of groups (clusters) and assigning each instance of a data set to one of these groups, according to a certain similarity measure. From a biological perspective, clustering represents an important step in determining gene functions, assuming that genes having similar expression levels under the same conditions may also have similar functions.

Received by the editors: December 8, 2013.

2010 *Mathematics Subject Classification.* 68P15, 68T05.

1998 *CR Categories and Descriptors.* I.2.6 [**Computing Methodologies**]: Artificial Intelligence – *Learning*; I.2.8 [**Computing Methodologies**]: Problem Solving, Control Methods, and Search – *Heuristic methods*.

Key words and phrases. Bioinformatics, Dynamic clustering.

We have previously introduced in [1] a novel approach for solving the dynamic problem of clustering gene expression data, when new features (expression levels for new points in time) are added to the genes within a data set. In this paper we aim to study two heuristics for centroids identification within the Core Based Dynamic Clustering of Gene Expression (CBDCGE) algorithm [1] and to analyze the influence on the initial centroids on the obtained results. The evaluations are performed on a data set that was used in [4].

The rest of the paper is organized as follows. Section 2 introduces the problem of dynamic gene expression clustering and presents an overview of the CBDCGE algorithm, together with two heuristics for centroids identification. A comparative study of the results, including experimental evaluations and analysis, is presented in Section 3. Section 4 outlines our conclusions and further work.

2. Background

2.1. **Dynamic Gene Clustering.** Gene expression data analysis is important within biology and medicine as the degree in which genes are expressed in different types of cell dictates cellular morphology and function. One of the most widely used data mining techniques used for this analysis is clustering.

Biological processes are mostly dynamic and in order to study and model them, scientists usually need information about gene expression at different moments in time, as the processes evolve. The resulting data sets are called time series data sets and they consist of gene expression data characterizing samples of cells or tissues which are extracted from the same individual at different moments in time, during the progression of the biological process. Thus, each gene is measured at several distinct time points and its expression levels are recorded. In the end, the time series data set consists of thousands of targeted genes (instances), each one being identified by a set of attributes (features): the values of its expression (quantified as real numbers) at all the considered time points.

The dynamism of gene expression data can be regarded from different perspectives: when new genes (instances) are added into the data set and when new gene expression levels (features), for the existing genes, are added into the data set. While for the first perspective there are several approaches in literature like k-means algorithms [9], artificial neural networks [10], particle swarm optimisation [11], for the second one, to our best knowledge, there are only two models that were previously introduced in [1] and [3].

Some biological processes only last for a short time, but there are other processes that may take months, even years (e.g. diseases). For the latter

ones, waiting until the process is finished to aquire all the necessary data is not feasible. An option would be to collect the data as the process evolves and apply the clustering algorithm each time new information is added. However, this technique could often be slow and inefficient, especially as the data sets contain thousands of instances and this increases the running time of the algorithm.

In order to surpass this drawback, a dynamic approach of clustering gene expression data has been introduced in [1], which is capable of adapting the previously obtained partition instead of re-clustering from scratch when new expression levels are added into the data set.

2.2. **Core Based Dynamic Clustering of Gene Expression.** In this section we present an overview of the CBDCGE model that was previously introduced in [1].

Let us denote by $\mathcal{G} = \{G_1, G_2, \ldots, G_n\}$ the set of genes to be classified. Each gene is measured at m moments in time and is therefore described by an m-dimensional vector $G_i = (G_{i1}, G_{i2}, \ldots, G_{im}), G_{ik} \in \Re, 1 \leq i \leq n, 1 \leq k \leq m$. An element G_{ik} from the vector characterizing the gene G_i represents the expression level of gene G_i at time point k.

Let $\{K_1, K_2, \ldots, K_p\}$ be the set of clusters discovered in data by applying the *k-means* algorithm. Each cluster is a set of genes, $K_j = \{G_{i_1}, G_{i_2}, \ldots, G_{i_{l_j}}\}$, $1 \leq l_j \leq n$, $1 \leq i_k \leq n \; \forall 1 \leq k \leq l_j$, where l_j is the number of genes from cluster j. The centroid (cluster mean) of the cluster K_j is denoted by f_j, where $f_j = \left(\frac{\sum_{k=1}^{l_j} G_{i_k 1}}{l_j}, \ldots, \frac{\sum_{k=1}^{l_j} G_{i_k m}}{l_j} \right)$.

In order to measure the distance between genes, we have chosen the *Euclidian distance*, because it takes into account the magnitude of the changes in gene expression, therefore preserving more data [5].

The measured set of attributes consisting of m gene expression levels (coming from m consequent measurements) is afterwards extended with s ($s \geq 1$) new attributes, numbered as $(m+1), (m+2), \ldots, (m+s)$. After extension, the genes' feature vectors become $G'_i = (G_{i1}, \ldots, G_{im}, G_{i,m+1}, ...G_{i,m+s})$, $1 \leq i \leq n$.

We analyzed in [1] the problem of recalculating the genes' grouping into clusters, after gene extension and starting from the current partitioning. Our goal was to obtain a better performance with respect to the partitioning from scratch process.

We denoted by $K'_j, 1 \leq j \leq p$, the set containing the same genes as K_j, after the attribute set extension. By $f'_j, 1 \leq j \leq p$, we denote the mean (center) of the set K'_j.

The sets $K'_j, 1 \leq j \leq p$, will not necessarily represent clusters after the attribute set extension, as the newly arrived attributes can change the genes' arrangement into clusters. But there is a considerable chance, when adding one or few attributes to genes and when the attributes have equal weights and normal data distribution, that the old arrangement into clusters is close to the new actual one.

The actual clusters could be obtained by applying the *k-means* clustering algorithm on the set of extended genes, but this process is computationally expensive. That is why we tried to avoid this process and replace it with one less expensive but not less accurate. CBDCGE algorithm [1] starts from the partitioning obtained before the attribute set extension and adapts it considering the newly added gene expression levels. This way, the clustering of genes at intermediate time points during the experiment can be more efficiently exploited and the final result could be achieved in smaller amounts of time. More details about the CBDCGE algorithm and its characteristics may be found in [1].

For identifying the most appropriate number p of clusters in the gene expression data set, the following heuristic is used. We have determined p representative genes, i.e., a representative gene for each cluster. First, the initial number p of clusters is set to 0. Then the first representative gene is chosen as being the most "distant" gene from the set of all genes (the gene that maximizes the average distance from all other genes). The number p of chosen representatives becomes now 1. In order to choose the next representative gene we reason as follows: for each remaining gene (that was not already chosen), we compute the average distance ($davg$) from the gene and the already chosen representative genes. The next representative gene is chosen as the gene g that maximizes $davg$ and this distance is greater than a positive given threshold ($distMin$), p is increased, and another representative gene is chosen again (the iterative process is performed again). If such a gene does not exist, it means that g is very close to all the already chosen representatives and should not be chosen as a new representative. In this case, the iterative process of selecting the initial centroids stops.

3. Comparative study

In this section we aim at providing an analysis of CBDCGE algorithm developed for dynamic clustering of gene expression data. The case study used

in our experiment, the evaluation measures, as well as the obtained results are presented and analysed in the following.

3.1. **Comparison criteria.** It is well known that a problem of the k-means based clustering algorithms is that they are sensitive to the selection of the initial centroids and may converge to a local minimum of the squared error value if the initial centroids are not properly chosen [12]. Consequently, it is very likely that the initial centroids may have an impact on the accuracy of the obtained results.

Thus, our comparative analysis is oriented to different methods for centroids' identification within the CBDCGE algorithm, as follows:

Heuristic 1. Te first heuristic method for selecting centroids is the one used in [1].

Heuristic 2. The second heuristic method for selecting the appropriate number p of clusters is based on selecting p representatives genes, as follows.

(i) The initial number p of clusters is set to 0.
(ii) The first representative gene chosen is the most "distant" gene from the set of all genes (the gene that maximizes the average distance from all other genes). The number p of chosen representatives becomes 1.
(iii) In order to choose the next representative gene we reason as follows. For each remaining gene (that was not already chosen), we compute the minimum distance ($dmin$) from the gene and the already chosen representative genes. The next representative gene is chosen as the gene g that maximizes $dmin$ and this distance is greater than a positive given threshold ($distMin$), p is increased, and step (iii) is performed again. If such a gene does not exist, it means that g is very close to all the already chosen representatives and should not be chosen as a new representative. In this case, the iterative process of selecting the initial centroids stops.

Random. The third way of choosing centroids is a random selection of p centroids, p being the number of clusters heuristically identified as above.

3.2. **Experiments.** In order to test the performance of the CBDCGE algorithm, we used a real-life data set, taken from [4] which contains the levels of expression of 6400 genes belonging to organism *Saccharomyces cerevisiae* during its metabolic shift from fermentation to respiration.

We have chosen this dataset from the following reasons: it is time series gene expression dataset, it is publicly available, it was used in several approaches from the literature giving us the posibility to compare our results

with the existing ones, and allows us to perform an evaluation from a biological perspective.

Gene expression levels were measured at seven time points during the diauxic shift. First, a pre-processing step was applied, in which the genes having small variance over time or very low absolute expression values, as well as genes whose profiles have low entropy are removed .

Considering an initial number of features (denoted by m) characterizing the genes from the considered data set, the experiments are conducted as follows:

(1) The number of clusters nc and the initial centroids are identified in the data set using different selection criteria (Subsection 3.1). The *k-means* clustering algorithm is applied on the data set consisting of m-dimensional genes, starting from the identified centroids and a partition $\mathcal{K}$ is provided.
(2) The set of features is now extended with s ($s \geq 1$) new attributes, numbered as $(m+1), (m+2), \ldots, (m+s)$. The $CBDCGE$ adaptive algorithm is now applied, by adapting the partition $\mathcal{K}$ and considering the instances extended with the newly added s features.
(3) The partition into clusters provided by $CBDCGE$ algorithm (denoted by $\mathcal{K}_{CBDCGE}$) is compared with the one provided by the $k-means$ algorithm applied from scratch on the $m+s$-dimensional instances (denoted by $\mathcal{K}'$). We mention that the initial centroids considered in the partitional clustering process are the centroids identified at step 1. The comparison of the obtained partitions is made considering the evaluation measures presented in Subsection 3.2.1 (both from the clustering and biological point of view), as well as the number of iterations performed by the clustering algorithms.

In order to accurately evaluate CBDCGE algorithm, we considered the same initial centroids when running k-means for the initial and feature-extended gene set (m and m + s number of features).

3.2.1. *Evaluation measures.* In order to measure the quality of the obtained partitions we use four *evaluation measures*. The first three measures (*IntraD*, *Dunn* and *Dist*) evaluate a partition from the clustering point of view, while the last one (*Z-score*) evaluates a partition from a biological point of view.

In the following, let us consider a partition $K = \{K_1, \ldots, K_p\}$, where each cluster consists of a set of genes. In the following $d(G_i, G_j)$ denotes the euclidean distance between G_i and G_j.

A. Intra-cluster distance of a partition - IntraD. The *intra-cluster distance* of a partition K, denoted by *IntraD(K)*, is defined as:

$$IntraD(K) = \sum_{j=1}^{p} \sum_{k=1}^{l_j} d(G_{i_k}, f_j)$$

where the cluster K_j is a set of genes $\{G_{i_1}, G_{i_2}, \ldots, G_{i_{l_j}}\}$ and f_j is the centroid (mean) of K_j.

From the point of view of a clustering technique, smaller values for $IntraD$ indicate better partitions, meaning that $IntraD$ has to be minimized.

B. Dunn Index - Dunn. The *Dunn index* [6] of a partition K is defined as:

$$Dunn(K) = \frac{d_{min}}{d_{max}}$$

where d_{min} represents the smallest distance between two genes from different clusters and d_{max} is the largest distance among two genes from the same cluster. The *Dunn index* takes values from the interval [0,∞]. The greater the value of this index, the better a partition is, therefore the *Dunn index* should be maximized.

C. Overall distance of a partition - Dist [1]. The *overall distance* of a partition K, denoted by *Dist(K)*, is defined as:

$$Dist(K) = \sum_{j=1}^{p} d_j$$

where d_j is defined as the sum of distances between all pair of genes from the cluster K_j, i.e

$$d_j = \sum_{G_1, G_2 \in K_j} d(G_1, G_2)$$

From the point of view of a clustering technique, smaller values for $Dist$ indicate better partitions, meaning that $Dist$ has to be minimized.

D. Z-score. Z-score [7] is a figure of merit, indicating the relationship between a clustering result and the functional annotation of the used genes, within the gene ontology developed by the Gene Ontology Consortium [8]. A higher value of the z-score indicates that the obtained clusters are more biologically relevant and therefore a more accurate clustering. To compute the z-score for a partition we used the ClusterJudge software, which implements the algorithm described in [7].

No.	**No. of clusters**	**No.of iterations**	**IntraD**	**Dunn**	**Dist**	**Z-score**
1	Heuristc 1 $distMin = 3.47$ $nc = 44$					
$\mathcal{K}'$	44	21	448.1514	0.1586	392.0747	7.7170
$\mathcal{K}_{CBDCGE}$	40	14	445.0609	0.1894	412.4337	9.9510
2	Heuristc 2 $distAvg = 1.13$ $nc = 44$					
$\mathcal{K}'$	44	11	440.1101	0.2686	20685.6718	6.0420
$\mathcal{K}_{CBDCGE}$	41	12	448.1529	0.2102	17133.5209	7.6540
3	Random $nc = 44$					
$\mathcal{K}'$	44	15	424.8114	0.1363	10912.5659	7.8844
$\mathcal{K}_{CBDCGE}$	43	14	427.7379	0.1535	11593.0304	9.2200

TABLE 1. Results for the first experiment.

3.3. **Results and Discussion.** In order to decide the most appropriate heuristic for selecting the initial centroids in the adaptive clustering process, we conducted two experiments. In each one we started from a different number of initial features and then we added the rest of the attributes (up to seven, which is the total number of attributes).

In both experiments the centroids were identified in three ways: using *Heuristic 1*, *Heuristic 2* and randomly (Section 3.1). For the randomly chosen centroids, an average obtained by five consequent runs was provided.

3.3.1. *Experiment 1.* In this experiment, the initial data set contains the first five features ($m = 5$) features and the remaining two features ($s = 2$) are added subsequently. The obtained results are presented in Table 1 and Figures 1a-3a.

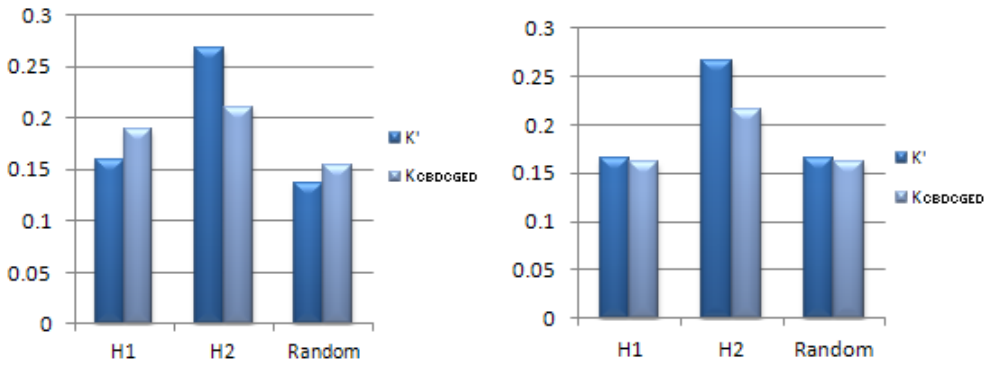
From these results we can conclude the following:

- The minimum number of iterations and the smallest *Dist* value are acheived by using Heuristic 1, both in K' and K_{CBDCGE}.
- The smallest $IntraD$ value is achieved by randomly choosing centroids, both in K' and K_{CBDCGE}.
- The highest *Dunn* value is achieved by using Heuristic 2, both in K' and K_{CBDCGE}.
- The highest $Z-score$ value is achieved by randomly choosing centroids in K' and using Heuristic 1 in K_{CBDCGE}.
- From a biological point of view (considering the $Z - score$ evaluation measure), in all three cases the adaptive clustering outperforms the re-clustering from scratch process.

No.	**No. of clusters**	**No.of iterations**	**IntraD**	**Dunn**	**Dist**	**Z-score**
1	Heuristc 1 $distMin = 4.66$ $nc = 42$					
$\mathcal{K}'$	42	23	451.5669	0.1655	446.6040	8.2300
$\mathcal{K}_{CBDCGE}$	40	23	438.2164	0.1621	351.8567	8.5044
2	Heuristc 2 $distAvg = 1.51$ $nc = 42$					
$\mathcal{K}'$	42	18	445.1501	0.2664	21649.7054	7.288
$\mathcal{K}_{CBDCGE}$	40	18	436.7869	0.2163	16133.5829	9.244
3	Random $nc = 42$					
$\mathcal{K}'$	42	15	430.6806	0.1434	11368.6995	9.0144
$\mathcal{K}_{CBDCGE}$	41	14	428.3622	0.1949	11848.0839	9.0853

TABLE 2. Results for the second experiment.

3.3.2. *Experiment 2.* In this experiment, the initial data set contains the first six features ($m = 6$) features and the remaining feature ($s = 1$) is added afterwards. The obtained results are presented in Table 2 and Figures 1b-3b.

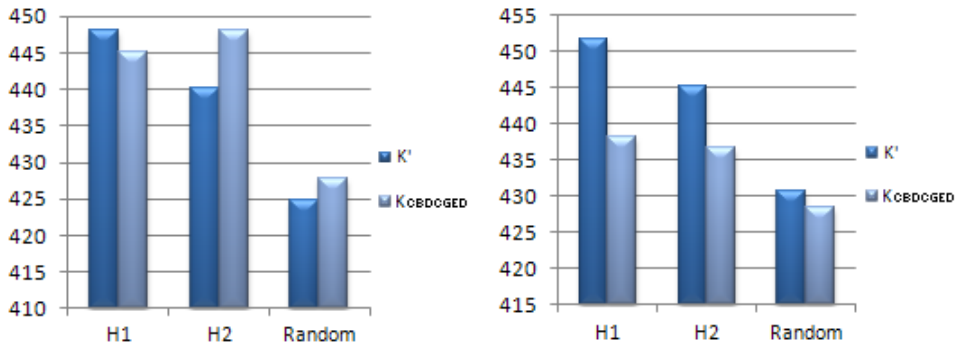


(A) First experiment. (B) Second experiment.

FIGURE 1. Illustration of the values of Dunn index obtained by using Heuristic 1, Heuristic 2 and random centroids, both for K' and K_{CBDCGE}.

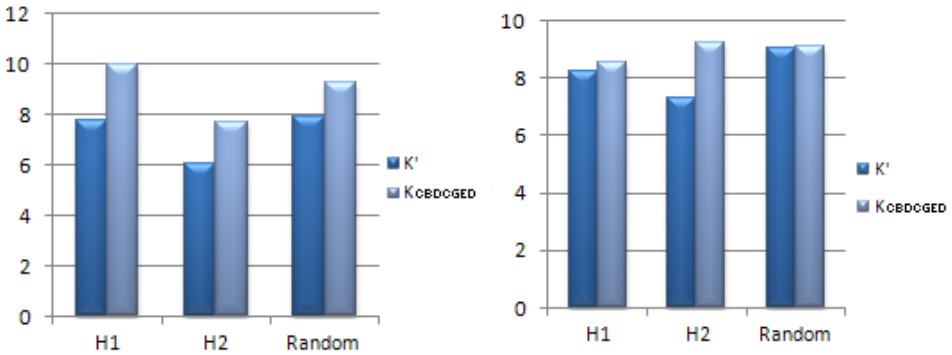
From these results we can conclude the following:

- The minimum number of iterations and the smallest $IntraD$ value are acheived by randomly choosing centroids, both in K' and K_{CBDCGE}.
- The highest $Dunn$ value is achieved by using Heuristic 2, both in K' and K_{CBDCGE}.



(A) First experiment. (B) Second experiment.

FIGURE 2. Illustration of the values of IntraD obtained by using Heuristic 1, Heuristic 2 and random centroids, both for K' and K_{CBDCGE}.



(A) First experiment. (B) Second experiment.

FIGURE 3. Illustration of the values of Z-score obtained by using Heuristic 1, Heuristic 2 and random centroids, both for K' and K_{CBDCGE}.

- The smallest *Dist* value is achieved by using Heuristic 1, both in K' and K_{CBDCGE}.
- The highest $Z-score$ value is achieved by randomly choosing centroids in K' and using Heuristic 2 in K_{CBDCGE}.
- The $Z-score$ evaluation measure, indicates in all three cases that the adaptive clustering outperforms the re-clustering from scratch.

3.3.3. *Statistical analysis.* Since for the problem we approach in this paper, clustering of gene expression data, the most relevant evaluation measure is the biological one, we performed a statistical analysis of Z-score values. We computed 95% Confidence Interval [2] for the average of the differences between the Z-scores obtained using the adaptive and from scratch approaches. All the Z-score values from the two experiments were considered. We obtained the (0.53, 1.95) Confidence Interval for the average. Thus, there is a 95% confidence that the Z-score of the partition obtained adaptively exceeds the Z-score of the partition obtained by applying the k-means from scratch with a value that lies within the specified range.

Due to the variation of the results, we can not conclude which heuristic is the best. It depends on the evaluation measure (e.g. Heuristic 2 is the best from *Dunn* index perspective, but is not the best from $IntraD$ perspective), the type of algorithm (adaptive/from scratch), the number of features added (for the adaptive approach). Even if there are cases in which choosing centroids randomly gives better results than using heuristics, it does not represent a reliable option, as an inappropriate choice could strongly degenerate results.

Still, for both experiments we have performed, we can conclude that from a biological point of view (considering the $Z - score$ evaluation measure) a better approach is to use an heuristic for the initial centroids selection, instead of a random choice.

4. Conclusions and further work

In this paper we presented a study on CBDCGE algorithm for dynamic clustering of gene expression data, with focus on the impact of heuristics used for centroids identification on the quality of clustering.

After an analysis of the obtained results, we can conclude, from a biological perspective, that CBDCGE algorithm outperforms the k-means applied from scratch, as it obtained better Z-score values in all the investigated cases.

As further work we plan to extend the CBDCGE method to a fuzzy clustering approach. Moreover, we plan to examine practical applications of the proposed method and to extend the experimental evaluation on other publicly available case studies.

References

[1] Bocicor, Maria Iuliana and Sirbu, Adela and Czibula, Gabriela *Dynamic core based clustering of gene expression data*, International Journal of Innovative Computing, Information and Control, volume 10, no. 3, P.1-13, 2014

[2] Brown, LD, Cat, TT and DasGupta, A *Interval Estimation for a proportion*, Statistical Science, volume 16, P.101-133, 2001

[3] Sirbu, Adela and Bocicor, Maria Iuliana *A Dynamic Approach for Hierarchical Clustering of Gene Expression Data*, Proceedings of 9th International Conference On Intelligent Computer Communication and Processing, P.3-6, 2013

[4] DeRisi, J.L. and Iyer, P.O. and Brown, V.R. *Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale*, International Journal of Innovative Computing, Information and Control, volume 278, no. 5338, P.680-686, 1997

[5] Kim, K. and Zhang, S. and Jiang, K. and Cai, L. and Lee, I.B. and Feldman, L.J. and Huang, H. *Measuring similarities between gene expression profiles through new data transformations*, BMC Bioinformatics, volume 8, no. 29, 2007

[6] M. K. Pakhira and S. Bandyopadhyay and U. Maulik *Validity index for crisp and fuzzy clusters*, Pattern Recognition, volume 37, no. 3, P.478-501, 2004

[7] Gibbons, F.D. and Roth, F.P. *Judging the Quality of Gene Expression-Based Clustering Methods Using Gene Annotation*, Genome Research, volume 12, no. 10, P.1574-1581, 2002

[8] Ashburner, M. and Ball, C.A. and Blake, J.A. and Botstein, D. and Butler, H. and Cherry, J.M. and Davis, A.P. and Dolinski, K. and Dwight, S.S. and Eppig, J.T. and et al. *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium*, Nature Genetics, volume 25, no. 1, P.25-29, 2000

[9] Bagirov, A.M. and Mardaneh, K. *Modified global k-means algorithm for clustering in gene expression data sets*, Proceedings of the 2006 workshop on Intelligent systems for bioinformatics, series WISB '06, P.23-28, 2006

[10] Yuhui, Y. and Lihui, C. and Goh, A. and Wong, A. *Clustering gene data via associative clustering neural network*, Proc. 9th Intl. Conf. on Information Processing, P.2228-2232, 2002

[11] Xiao, X. and Dow, E.R. and Eberhart, R.C. and Miled, Z.B. and Oppelt, R.J. *Gene Clustering Using Self-Organizing Maps and Particle Swarm Optimization*, Proc. 17th Intl. Symposium on Parallel and Distributed Processing, 2003

[12] Mohammed El Agha, Wesam M. Ashour. *Efficient and Fast Initialization Algorithm forK-means Clustering*, MECS Publisher, no. 1, 2012

Babeş-Bolyai University, Department of Computer Science, 1 M. Kogălniceanu Street, 400084 Cluj-Napoca, Romania

E-mail address: adela@cs.ubbcluj.ro