Sādhanā (2020)45:160
https://doi.org/10.1007/s12046-020-01399-x

An improved ant-based algorithm based on heaps merging and fuzzy c-means for clustering cancer gene expression data

HASAN BULUT[1,*], AYTUĞ ONAN[2] and SERDAR KORUKOĞLU[1]

[1]Department of Computer Engineering, Ege University, Izmir, Turkey
[2]Department of Computer Engineering, Izmir Katip Celebi University, Izmir, Turkey
e-mail: hasan.bulut@ege.edu.tr; aytug.onan@ikcu.edu.tr; serdar.korukoglu@ege.edu.tr

MS received 12 May 2016; revised 5 April 2020; accepted 20 May 2020

Abstract. The microarray technology enables the analysis of the gene expression data and the understanding of the important biological processes in an efficient way. We have developed an efficient clustering scheme for microarray gene expression data based on correlation-based feature selection, ant-based clustering, fuzzy c-means algorithm and a novel heaps merging heuristic. The algorithm utilizes the feature selection algorithm to overcome the high-dimensionality problem encountered in bioinformatics domain. Based on extensive empirical analysis on microarray data, clustering quality of the ant-based clustering algorithm is enhanced with the use of fuzzy c-means algorithm and heaps merging heuristic. The performance of the proposed clustering scheme is compared with k-means, PAM algorithm, CLARA, self-organizing map, hierarchical clustering, divisive analysis clustering, self-organizing tree algorithm, hybrid hierarchical clustering, consensus clustering, AntClass algorithm and fuzzy c-means clustering algorithms. The experimental results indicate that the proposed clustering scheme yields better performance in clustering cancer gene expression data.

Keywords. Gene expression data; gene selection; clustering; ant-based clustering; correlation-based feature selection; hybrid algorithms.

1. Introduction

Microarray gene expression data is an indispensable technique of research in biomedical informatics. Microarray technologies make large amount of gene expression data available and enable to monitor these genes in parallel. The prediction of an event's outcome based on the associated genes and the identification of sub-classes of a particular disease are two typical research directions in microarray data analysis [1]. Clustering is an essential task of data mining and knowledge discovery. It aims to assign a set of objects into the groups based on proximity or similarity measures such that the objects within the same group are more similar to each other, whilst the objects from different groups (clusters) are different as much as possible. Clustering techniques are viable tools for microarray gene expression. Clustering can be helpful in understanding gene functions, identification of gene regulation, cellular processes and cell/disease sub-types [2]. The conventional clustering algorithms have been successfully applied on microarray gene expression data [3]. The conventional algorithms may suffer from a number of shortcomings, such as being very sensitive to the initial state and converging to the local optima [4]. Clustering algorithms can be broadly divided into five categories as hierarchical, partitional, density-based, model-based and grid-based methods proposed by Boulesteix and Strimmer. Hierarchical clustering methods include agglomerative and divisive hierarchical clustering methods. Partitional clustering methods include error minimization based approaches (such as K-means and partitioning around medoids). Density-based clustering algorithms include the expectation maximization and DBSCAN algorithms. Model-based approaches to clustering include neural networks and decision trees. In addition, soft computing based methods (such as fuzzy clustering and swarm intelligence based approaches) can be employed for cluster analysis. Ant-based clustering and fuzzy c-means are several representatives of soft computing based approaches.

There are many studies on the application of clustering algorithms on gene expression data. The related work is briefly introduced here. Alon *et al* [5] utilized a two-way clustering algorithm to analyze the gene expression patterns of tumor and normal colon tissues. Golub *et al* [6] applied self-organizing map algorithm to cluster tumors by gene expression. Alizadeh *et al* [7] utilized a hierarchical clustering method to assign group tumor and cell samples into clusters based on the similarities in gene expressions. Dudoit and Fridlyand [8] presented a resampling based method to determine the number of clusters in a dataset and

*For correspondence
Published online: 23 June 2020

the performance of the proposed approach has been evaluated on gene expression data obtained from cancer microarray experiments. Datta and Datta [9] empirically evaluated several clustering algorithms, such as hierarchical clustering with correlation, k-means, DIANA, FANNY, model-based clustering and hierarchical clustering with partial least squares on microarray gene expression data. Costa *et al* [10] examined the efficiency of five clustering algorithms (agglomerative hierarchical clustering, CLICK, dynamical clustering, k-means and self-organizing maps) on gene expression time series. Iam-on and Boongoen [11] presented a locally weighted measure for k-means clustering and the approach have been evaluated on microarray cancer data. Castellanos-Garzon and Diaz [12] presented a hybrid clustering algorithm for gene expression data which combines a hierarchical clustering method with genetic algorithms. Binu [13] modelled the clustering algorithm as an optimization problem and developed three different objective functions based on cumulative summation of fuzzy membership and kernel space. Ant-based and swarm-based algorithms have been successfully utilized in clustering to enhance the performance and quality of clustering while overcoming the aforementioned problems of conventional algorithms. Liu and Pham [14] surveyed the application of fuzzy clustering for microarray data analysis. Bhattacharya *et al* [15] examined the effectiveness of several clustering and bi-clustering algorithms on gene expression data. Datta and Mukhopadhyay [16] presented a fuzzy c-means clustering based scheme for identification of in silico identification of human promoters.

In another study, Bhattacharya and De [17] presented a clustering algorithm based on correlation concept for analysis of gene-expression data. In this scheme, correlation matrix is utilized to assign genes such that all genes in a particular cluster have highest average correlation with the genes of the same cluster. In another study, Bhattacharya and De [18] presented a bi-correlation clustering algorithm which obtains a set of bi-clusters of co-regulated genes. Similarly, Bhattacharya and De [19] presented a clustering algorithm based on correlation clustering concept. In this scheme, genes with common transcription factors and similar pattern of variation in their expression values are assigned into the same clusters.

In addition the clustering algorithms discussed in advance, bi-clustering algorithms have been extensively employed in the microarray data analysis to identify groups of genes with high correlated expression patterns. For instance, Turner *et al* [20] presented a biclustering algorithm based on the least squares and the binary constraints. In another study, Santamaria *et al* [21] examined several internal and external cluster validity indices and employed these indices to biclustering microarray data. Similarly, Filippone *et al* [22] employed stability indices to examine the clustering quality of fuzzy biclustering algorithms. In another study, Ayadi *et al* [23] presented a heuristic approach to bicluster microarray data. In the presented scheme, average correspondence similarity index was utilized to examine the coherence of a particular bicluster. More recently, Saber and Elloumi [24] presented a biclustering algorithm for binary microarray data based on iterative row and column clustering combination. A comprehensive survey on biclustering microarray data examines the conventional methods in the field [25].

Ant-based clustering is a swarm intelligence approach which is inspired by the ant behavior in ant colonies. In the basic model, ants are modeled as agents and they move randomly and pick up or drop off an item with regard to density within the environment. There are several extensions of the basic ant-based clustering [26]. AntClass algorithm [26, 27] is a hybrid clustering scheme which integrates ant-based clustering with k-means algorithm. The algorithm starts with ant-based clustering. This is followed by k-means algorithm. Afterwards, ant-based clustering is applied once again on the same partition. Finally, the final cluster assignment is determined at the fourth stage by means of k-means algorithm. The main motivation of this study is to develop an efficient clustering scheme for cancer gene expression data. The paper presents an efficient clustering scheme which utilizes correlation-based feature selection, ant-based clustering, fuzzy c-means algorithm and heaps merging heuristic. The feature selection method, ant-based clustering and fuzzy c-means algorithms are well-established techniques of machine learning research. Based on extensive empirical analysis on sixteen cancer gene expression data, the best (the higher) clustering quality in terms of adjusted rand index is obtained with the utilization of correlation-based feature selection in conjunction to ant-based clustering refined by fuzzy c-means. Hence, this scheme is utilized as the base clustering configuration. In addition, the experimental results indicated that the proposed clustering scheme suffers from over cluster generation problem. In order to overcome this problem, a heaps merging heuristic approach is presented. The contributions of the paper can be summarized as follows:

- Microarray gene expression data is a high-dimensional data and some of the features may degrade the performance of clustering/classification algorithm [28]. In order to obtain an efficient clustering or classification scheme, the identification of appropriate subset of features is a crucial task. The paper presents an extensive empirical analysis on six feature selection algorithms (empirical Bayes moderated t-test, partial least squares based feature selection, random forest based feature selection, significance analysis of microarrays, correlation-based feature selection and ensemble feature selection) on cancer gene expression data clustering.
- The development of efficient/robust ensemble clustering/classification systems is an important task in machine learning research. Besides, the identification

of efficient feature selection methods and the determination of components to be included in the clustering or classification scheme are promising research directions. In this regard, the paper presents a six-staged clustering scheme based on extensive empirical analysis on 18 microarray gene expression data. The experimental results indicate the superiority of the proposed clustering scheme.

- As mentioned in advance, the number of clusters obtained at the end of AntClass algorithm may exceed the actual number of clusters in datasets [29]. Therefore, further improvement on the performance of the proposed clustering scheme is achieved with the utilization of presented heaps merging heuristic.

2. Materials and methods

This section briefly explains the feature selection and clustering methods utilized in the proposed clustering scheme.

2.1 *Feature selection methods*

Microarray gene expression data is a high-dimensional data. Feature selection is the process of identifying an appropriate feature subset so that non-informative features can be eliminated, whilst improving the understanding of the data and reducing the computational time [30]. Empirical Bayes moderated t-test feature selection (eBayes) is a univariate filter-based method which utilizes empirical Bayes method and t-test to determine appropriate feature subsets [31]. In this method, an empirical Bayes method is utilized to diminish gene-wise sample variances towards a common value and to augment the degrees of freedom for the individual variances. Partial least squares based feature selection (PLSS) is a univariate filter-based method which ranks the features in terms of their importance based on the magnitudes of the weight vector defining the first latent component which defines the first latent component in a partial least square classifier [32]. Random forest based feature selection (RF) is a feature ranking method which utilizes random forest classifier to determine the relevance of the features [33]. In the method, the mean decrease in accuracy or the mean decrease in the Gini index node impurity measure can be used to evaluate the merit of a feature [29]. For this approach, the evaluation measure is computed from parent nodes to their child nodes. This computation is repeated for all nodes in the decision tree. Based on the value for this measure, a ranking list for the features is obtained. Significance analysis of microarrays (SAM) is a feature selection method which assigns a value to each gene based on the changes in gene expression [34]. In the method, the standard deviation among different measurements is computed for each gene. Based on this measure, the genes above a user-defined threshold value are filtered. Then, false discovery rate is utilized to identify proportion of genes selected by chance. Correlation-based feature selection (CFS) is a filter-based feature selection method which examines feature subsets via a correlation-based heuristic evaluation function [35]. In correlation-based feature selection, both the merit of individual features for predicting the class label and inter-correlation levels of features are taken into account. The heuristic evaluation function is calculated as given by Eq. (1) [35]:

$$M_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{1}$$

where MS is the heuristic merit of a feature subset S containing k features, $\overline{r_{cf}}$ is the mean feature-class correlation ($f \in S$) and $\overline{r_{ff}}$ is the mean feature-feature inter-correlation. A ranking of the feature subsets in the search space of all possible enumerations is obtained with the use of Equation (1) [35]. The method is generally utilized in conjunction with metaheuristic search algorithms, such as greedy best first search. Ensemble feature selection (ENS) aims to obtain a robust feature selection scheme by integrating multiple feature selection methods. In this ensemble feature selection method, empirical Bayes moderated t-test feature selection, partial least squares based feature selection, random forest based feature selection and significance analysis of microarrays methods are combined and the final ranking of the features is obtained by summing the ranks obtained by individual feature selection methods [29].

2.2 *Clustering algorithms*

Cluster analysis is the process of assigning data objects into clusters based on similarities or distances. Gene expression data clustering can be broadly divided into three groups as gene-based, sample-based and subspace clustering [36]. In gene-based clustering, genes are regarded as data objects and samples are regarded as features. In contrast, samples are regarded as data objects, while genes are treated as features in sample-based clustering. Subspace clustering aims to identify all clusters in all subspaces. In subspace clustering, genes or samples can be considered as objects or features symmetrically [36]. In the experimental analysis, the performance of conventional gene-based clustering algorithms is evaluated. Conventional clustering algorithms can be broadly divided into two groups as partition-based and hierarchical clustering [37]. Partition-based clustering algorithms divide data objects into clusters so that each cluster must contain at least one data object and an object cannot be assigned to several groups [38]. K-means algorithm (KM) is a partition-based clustering algorithm. It is frequently used in cluster analysis due to its easy implementation, simplicity and efficiency [39]. The algorithm

takes the number of clusters as input parameter. The algorithm initiates with the random selection of k objects as cluster centers. The remaining objects are assigned to the clusters with closest centers. The algorithm continues to compute new mean of clusters till stopping criterion. The algorithm gives good results especially for clusters with well aligned and compact shapes [38]. It is an efficient and scalable algorithm. There are two main drawbacks of the utilization of k-means algorithm as a gene-based clustering algorithm [36]. First, the identification of optimal number of clusters involves fine-tuning which may be costly for large gene expression data. In addition, gene expression data contains generally huge amount of noise. However, k-means algorithm is sensitive to the noise [36]. Fuzzy c-means clustering algorithm (FCM) is a popular fuzzy clustering algorithm. Fuzzy clustering is mainly based on fuzzy set theory and fuzzy logic proposed by Zadeh in 1965. Fuzzy clustering becomes very useful especially when objects in datasets cannot be partitioned into well-separated clusters [40]. K-medoids based algorithms pick an actual object to represent the clusters, using one representative object per cluster and assign each of the remaining object to the most similar representative objects. Since clusters are represented via representatives instead of cluster centers, K-medoids based algorithms are characterized by being less sensitive to outliers and extreme values. PAM (Partitioning around medoids) algorithm takes the number of clusters in the data set and a data set with n objects [40]. It starts with arbitrarily choosing k objects in data set as the initial representative objects or seeds. Each remaining object is assigned to the cluster with the nearest representative object. Then, a non-representative object is randomly selected and total cost of swapping representative object with randomly selected non-representative object is computed. Based on computed total cost, representative object is whether swapped with randomly selected non-representative object or not to form new set of representatives. The process continues until no change [39]. The algorithm works efficiently on only small data sets, but it cannot scale well for large data sets [41]. In order to overcome this inefficiency, another K-medoids based algorithm, named CLARA (Clustering large applications) was introduced [42]. CLARA algorithm works on multiple samples of the data set and applies PAM to the sample [42]. Compared to PAM algorithm, CLARA can handle with larger data sets properly. However, the performance of CLARA is badly affected while increasing the number of clusters [43]. Self-organizing map (SOM) is a partition-based clustering algorithm which obtains a two-dimensional grid with nodes that function as cluster centers in the high dimensional space. Inner nodes of the grid are connected by an edge. For the data vectors, the nearest neighbors are determined by the learning algorithm [44]. Hierarchical clustering provides a natural representation scheme for gene expression data [36]. Hierarchical clustering algorithms can be either agglomerative or divisive. In the agglomerative clustering algorithms, the individual objects are merged into clusters on the basis of similarities, whereas the objects within single group are divided into subgroups in divisive clustering algorithms [45]. Divisive analysis clustering (DIANA) is a divisive hierarchical clustering algorithm [38]. The algorithm starts with an initial cluster containing all of the objects. Then, a cluster with the maximum diameter is selected and the splinter group is initialized. For each data objects, an average distance to the all other objects is computed. Based on the computed value, data objects are assigned into the splinter groups with the maximum difference. Self-organizing tree algorithm (SOTA) is a hybrid clustering algorithm which combines hierarchical clustering and self-organizing maps [44]. In the method, nodes are mapped into a binary tree based topology instead of a two-dimensional space. In SOTA, the number of nodes is not kept fixed at the start. Instead, the tree structure expands during the clustering process. Hybrid hierarchical clustering (HYBRID) is a clustering algorithm which combines the divisive hierarchical clustering with the agglomerative hierarchical clustering [45]. The algorithm starts with a bottom-up hierarchical clustering step which is followed by a top-down clustering. Consensus clustering (CONS) algorithm [29] aims to enhance clustering quality by combining the information obtained from several individual clustering algorithms. In the method, simulated annealing algorithm is utilized for combining the results of individual algorithms.

2.3 *AntClass algorithm*

Monmarché *et al* [26] proposed a hybrid method, called AntClass algorithm, to combine the stochastic and exploratory features of ant colony algorithm with the deterministic and heuristic features of K-means algorithm to improve convergence. The algorithm mainly consists of four steps. Initially, the algorithm starts with ant-based algorithm for clustering objects. This step is followed by K-means algorithm. Then, ant-based clustering algorithm is applied once again and the algorithm terminates with applying K-means algorithm on objects once more [27]. Each data contains a vector of n real values. The distance between two data objects is measured by Euclidean distance. Data objects are scattered randomly on a grid similar to other ant clustering algorithms. One basic difference of the algorithm is that ants can create, build or destroy heaps, containing two or more objects. A heap can be located on a single grid cell. Given a particular heap H consisting of $n_H$objects, maximum distance between two data objects $D_{\max}(H)$ is determined as follows [17]:

$$D_{\max}(H) = \max_{\gamma_i,\gamma_j \in E} D(\gamma_i, \gamma_j) \tag{2}$$

where γ_i and γ_j denote data object i and j, respectively.

The center of mass of all objects in H, $\gamma_{center}(H)$ and the mean distance between all objects in H, $D_{mean}(H)$ are determined by the following equations [26]:

$$\gamma_{center}(H) = \frac{1}{n_H} \sum_{\gamma_i \in H} \gamma_i \tag{3}$$

$$D_{mean}(H) = \frac{1}{n_H} \sum_{\gamma_i \in H} D(\gamma_i, \gamma_{center}(H)) \tag{4}$$

where γ_i refers to a particular data object i. The most dissimilar object in a heap, $\gamma_{dissim}(H)$ is an object which maximizes the distance to center of mass of all objects in the heap.

Ant-based clustering stage of AntClass algorithm starts with random initialization of ants' positions on the grid. Ants can perform several actions depending on the state. Each ant moves at each iteration. If an ant does not carry any object, then eight cells in the ant's neighborhood are examined and a single object from the neighborhood is picked up according to the pick-up probability. Otherwise, if an ant carries a data object γ, then eight cells in the ant's neighborhood are examined and the object is dropped according to the dropping off probability [26, 27]. The two main actions performed by the ants are picking up and dropping off objects. Since ants are able to build or destroy heaps, there are several conditions to be considered in picking up and dropping off objects. An unladen ant looks for a possible object by examining the eight cells in its neighborhood and if one object or heap is found, then there are three cases to consider. The first case is that the ant may find one object alone. In this case, data object is picked up according to a fixed probability. The second case is that the ant finds a heap with two objects. In this case, one of the objects from the heap is removed with a probability$P_{destroy}$.. The other case is that the grid cell contains a heap H with more than two objects. In this case, the ant picks up the most dissimilar object in the heap if the condition given by Eq. (5) is satisfied [26, 27]:

$$\frac{D(\gamma_{dissim}(H), \gamma_{center}(H))}{D_{mean}(H)} > T_{remove} \tag{5}$$

where T_{remove} represents a threshold parameter for removing an object.

An ant carrying an object also considers eight cells in its neighborhood and there are three conditions to examine; the cell may not contain any object, may contain only one object or may contain a heap. If the cell is empty, carried object will be dropped off with a probabilityP_{drop}. If the cell contains only one object, carried object will be dropped off and a heap with two objects will be generated if the condition given by Eq. (6) is satisfied [26, 27]:

$$\frac{D(\gamma, \lambda)}{D_{\max}} < T_{create} \tag{6}$$

where λ is the object in the cell, γ is the object carried by the ant and T_{create} is a threshold parameter. If the cell contains a heap H, then carried object is dropped off if the condition given by Eq. (7) is satisfied [26]:

$$D(\gamma, \gamma_{center}(H)) < D(\gamma_{dissim}(H), \gamma_{center}(H)) \tag{7}$$

At the end of the ant-based clustering stage of AntClass algorithm, there are some objects which are not assigned to any heap at all or wrongly assigned to any heap. In order to overcome this problem, K-means algorithm is applied to the partition obtained by ant-based clustering stage. The algorithm operates on grid positions. The first two steps of AntClass algorithm end with applying ant-based clustering and K-means algorithms. After that, ant-based clustering algorithm is applied once more but instead of single objects, the algorithm runs on heaps. In this step, ants are able to pick up or drop off previously created heaps [27]. At the end of this step, K-means algorithm is applied once again and the algorithm terminates.

3. Proposed clustering scheme

In this paper, a hybrid clustering algorithm is presented for clustering microarray gene expression data. Hybrid clustering algorithms may enhance the clustering quality for gene expression data. Microarray gene expression data is a high dimensional data with irrelevant/redundant features. The number of clusters obtained at the end of AntClass algorithm can exceed the actual number of clusters in the original dataset [46]. Fuzzy clustering algorithms can perform well on microarray gene expression data. Taking all these issues into account, the presented hybrid clustering algorithm contains six stages which combine feature selection, ant-based clustering, fuzzy clustering and heaps merging. The stages of the proposed clustering scheme are feature selection, ant-based clustering, fuzzy c-means clustering, ant-based clustering, fuzzy c-means clustering and heaps merging, respectively. As it will be presented in section 4, correlation-based feature selection yields better results for microarray gene expression data among other feature selection methods, such as empirical Bayes moderated t-test feature selection, partial least squares based feature selection, random forest based feature selection, significance analysis of microarrays and ensemble feature selection. Hence, clustering scheme starts with correlation-based feature selection to obtain an appropriate feature subset. AntClass algorithm is a four-staged scheme: ant-based algorithm for clustering objects, k-means algorithm based clustering on the initial partition of the ants, ant-based clustering on the formed heaps and final refinement of clustering by k-means algorithm. In this paper, ant-based

algorithm for clustering objects and the heap structure of AntClass algorithm is preserved. Based on empirical analysis on the cancer gene expression data with conventional clustering algorithms, the higher clustering quality is obtained with the utilization of fuzzy c-means clustering algorithm in conjunction with the ant-based clustering algorithm. Hence, we have utilized the fuzzy c-means algorithm instead of k-means algorithm.

In figures 1 and 2, illustrations of clustering on Golub-1999-v1 dataset for k-means and the proposed scheme are presented, where ellipses corresponds to heaps. The data is a 3D data and partitioned into nonoverlapping clusters. When showing the data in a 2D graph, some elipses might appear as coinciding. In this example, the main purpose is to demonstrate how heaps are changed and merged over iterations with the clustering approach. The clustering quality of conventional clustering algorithms (such as k-means) and swarm-based clustering algorithms (such as ant-based clustering algorithm) may be degraded with the overmuch clusters at the end of the final clustering. The number of clusters (classes) for this dataset is two. As it can be observed from figure 1, k-means partitions data instances into four clusters. The same pattern is valid for the other conventional clustering schemes utilized in the empirical

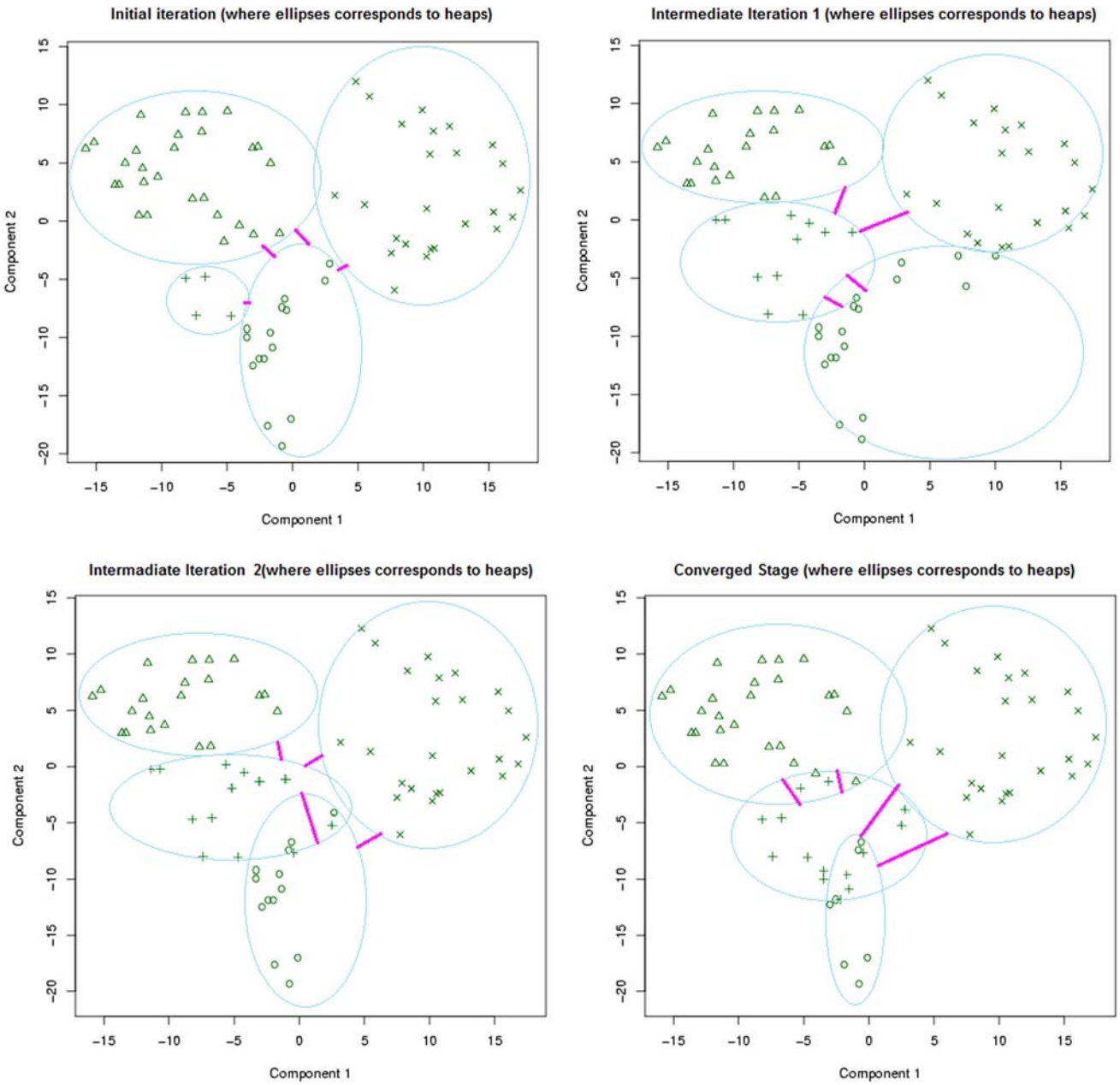

Figure 1. PCA clustering plots for K-means on Golub-1999-v1 dataset.

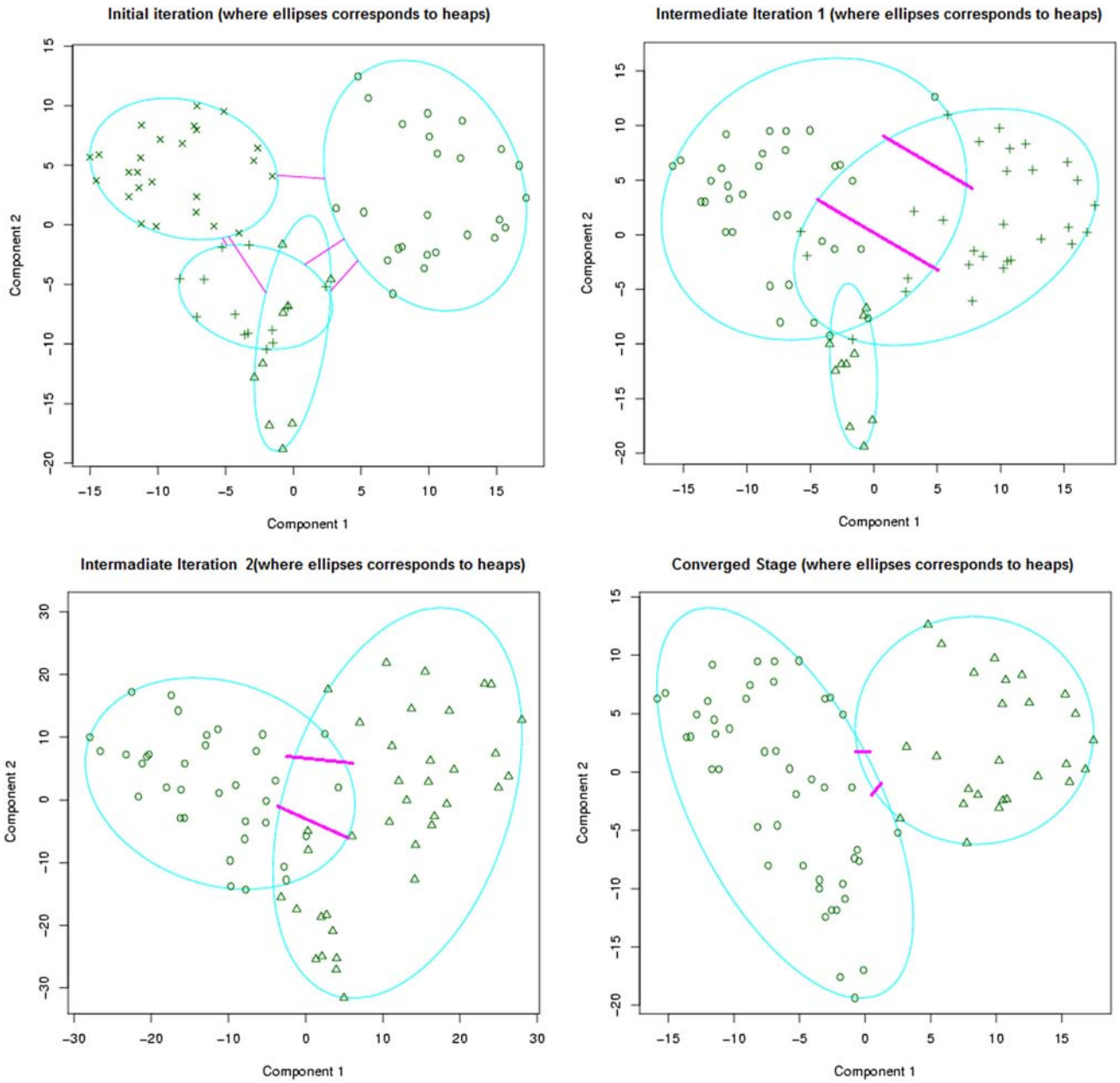

Figure 2. PCA clustering plots for the proposed scheme on Golub-1999-v1 dataset.

analysis. In contrast, the proposed clustering scheme based on heaps merging obtains a two-clustered partition for the same dataset. In this way, the proposed scheme aims to enhance the clustering quality on gene expression data.

As indicated above, our approach attempts to reduce the number of heaps (clusters) at the end of the fifth stage of the clustering scheme by means of a heaps merging based method [46, 47].

In heaps merging, distance between the objects of heaps and center of clusters is examined. If there are a number of objects in any heaps whose distances to their heap's center are larger than their distances to other heap's center, these two heaps can be merged into a single cluster. Heaps merging stage starts with determining the size of the neighborhood based on the number of clusters obtained at the end of the fifth stage of clustering scheme. For cases of eight heaps or less, all the heaps are taken as the neighbors to be examined. For cases of 8–16 heaps, the neighborhood size is taken as 8, whereas for other cases the size is taken as 16. Then, the maximum number and the average number of objects in the heaps are identified. The threshold value for merging criterion is computed based on the two numbers by subtracting the maximum number from average number and dividing this value to five. Then, the distances

of each object to its cluster center and other heaps' centers are computed. Then, the heaps are merged based on the merging criterion. The process is continued until there is no heaps satisfying the required conditions for merging [46].

In heaps merging, we need to determine the appropriate number of objects satisfying the merging criterion and the appropriate number of heaps in the neighborhood to be examined for merging. Since the characteristics, the attributes and the number of clusters differ in various datasets, the number of heaps generated and the number of objects in each heap may vary accordingly.

Table 1 depicts the number of objects, the number of heaps, the average heap sizes and the largest heap sizes after the first two steps of the algorithm in for each dataset when heaps merging approach is used. In order to obtain a better approximation to the number of objects for merging criterion, a number of different scenarios are conducted for various number of heaps (i.e., 2, 3, 4, 5, 6, 7, and 8) and various number of neighborhood (i.e., 2, 4, 8, and 16). The last two columns of table 1 indicate the number of heaps to merge and the size of neighborhood when the best results among all the different combinations are obtained from the given data sets. Based on these values, a relationship between the number of objects for merging criterion, the number of objects in maximum heap and the average number of objects in heaps is derived as given by Eq. (8):

$$n_{appr} = \frac{n_{H_{\max}} - n_{average}}{5} \tag{8}$$

where n_{appr} is number of objects for merging criterion, $n_{H\max}$ is number of objects in heap with maximum objects and $n_{average}$ is average number of objects in heaps.

In figure 3, the general structure for the proposed clustering scheme is outlined. In addition, figure 4 summarizes the block diagram for the proposed clustering scheme.

Figure 5 summarizes the phases of proposed clustering scheme on a toy dataset. For the dataset utilized in this illustration, there are three clusters, denoted by purple, blue and red dots (instances). Initially, ant-based clustering (Phase 1) has been employed on the instances of dataset. In this stage, the positions of ants have been randomly initialized, as denoted by black dots in figure 5. Then, ant agents move. By picking up an object or dropping an object as set by the rules indicated in figure 3, ants obtain an intermediate clustering of instances. In Phase 2, cluster centers obtained in phase 1 have been utilized to initialize cluster centers for fuzzy c-means. Then, a fuzzy c-means based partitioning of instances has been obtained. In Phase 3, ant-based clustering has been employed once again on the instances obtained as a result of fuzzy c-means clustering algorithm. Here, cluster centers of fuzzy c-means based partitioning have been utilized to initialize the positions of ants. This stage has been followed by fuzzy c-means clustering. Then, the partitions obtained in the earlier stage have been merged based on heaps merging criterion in Phase 5.

4. Results and discussion

The performance of the proposed hybrid clustering scheme is assessed on eighteen microarray gene expression datasets summarized in table 2. In addition, the performance of the clustering algorithms are evaluated on three yeast datasets [17]. The empirical analysis regarding the feature selection methods and clustering algorithms have been conducted on ArrayMining software [29], whereas the proposed clustering scheme is implemented by Java programming language [47]. In the experimental analysis, the parameter values of ant-based clustering are assigned based on the reference papers [26, 27]. We conduct a set of experiments on an Intel Core i7 CPU 3.40 GHz with 8.0 GB RAM. In the experimental analysis, the performance of several feature selection methods, namely correlation-based feature selection, empirical Bayes moderated t-test feature selection, partial least squares based feature selection, random forest based feature selection, significance analysis of microarrays and ensemble feature selection is examined. In table 3, the number of features for the preprocessed microarray gene expression datasets with and without feature selection methods have been summarized. For feature selection methods, maximum feature subset size parameter has been set to 100.

In order to evaluate the clustering quality, thirteen different clustering algorithms are utilized. These algorithms are K-means algorithm (KM), PAM (Partitioning around medoids) algorithm, CLARA (Clustering large applications), self-organizing map (SOM), hierarchical clustering (HIER), divisive analysis clustering (DIANA), self-organizing tree algorithm (SOTA), hybrid hierarchical clustering (HYBRID), consensus clustering (CONS), AntClass

Table 1. The descriptive information regarding the datasets.

Dataset	Number of heaps	Average heap size	Largest heap size	Number of heaps merging	Neighborhood size
Alizadeh-2000-v1	5	25	42	3	8
Bittner-2000	32	21	60	8	16
Dyrskjot-2003	24	8	16	2	16
Golub-1999-v1	7	25	63	8	4

Input: A set of n genes $X=\{x_1, x_2, \ldots, x_n\}$, for each of which m expression values are given.

Output: K disjoint clusters $C_1, C_2, \ldots, C_k$.

Stage-1: Ant-based clustering stage

1. Initialize randomly the positions of ants
2. Repeat
3. For each ant ant_i Do
   a. Move ant_i,
   b. **(Picking up an object)** If ant_i does not carry an object Then look at the 8 cells in the neighborhood of ant_i location and possibly pick up an object:
      i. Label the 8 cells around ant_i as "unexplored"
      ii. Repeat
         1. Consider the next unexplored cell c around ant_i with the following order: cell1=in front of ant_i, cell 2=to the left, cell 3=to the right, etc.
         2. If c is not empty, THEN do one of the following action only:
            a. If c contains a single object O, THEN load O with probability P_{load}, Else
            b. If c contains a heap of two objects, THEN remove one of the two objects only with a probability $P_{destroy}$, Else
            c. If c contains a heap H of more than 2 objects, THEN remove the most dissimilar object $O_{dissim(H)}$ from H provided that: $\frac{D(O_{dissim}(H),O_{center}(H))}{D_{mean}(H)} > T_{remove}$

               where the most dissimilar object $O_{dissim(H)}$ is the object of heap which maximizes the distance to center ($O_{center(H)}$), $D_{mean(H)}$ is the mean distance between the objects of H and the center of mass and T_{remove} is the threshold value for removing an object.
         3. Label c as explored.
      iii. Until all the 8 cells have been explored or one object has been loaded.
   c. **(Dropping an object)** Else look at the 8 cells in the neighborhood of ant_i and possibly drop an object O:
      i. Label the 8 cells around ant_i as "unexplored"
      ii. Repeat
         1. Consider the next unexplored cell c around ant_i with the following order: cell1=in front of ant_i, cell 2=to the left, cell 3=to the right, etc. and possibly perform one of the following:
            a. If c is empty THEN drop O on this cell with a probability P_{drop}, Else
            b. If c contains a single object O' THEN drop O and O' to create a heap but provided that division of the distance between O and O' to maximum distance value is smaller than a creation threshold value.
            c. If c contains a heap H, THEN drop O on H but provided that:
               1. $D(O_{dissim}(H), O_{center}(H)) > D(O, O_{center}(H))$
         2. Label c as explored.
      iii. Until all the 8 cells have been explored or one object has been dropped.
4. Until stopping criterion

Stage-2: Fuzzy c-means clustering

1. Use cluster centers obtained in Stage-1 to initialize cluster centers by Fuzzy C-means algorithm.
2. Cluster the data using Fuzzy C-means algorithm.

Stage-3: Repeat Stage-1 (Ant-based clustering stage)

Stage-4: Repeat Stage-2 (Fuzzy c-means clustering)

Stage-5: Heaps merging

1. Take the partition of data set generated by fuzzy c-means algorithm and k heaps.
2. Determine the number of heaps in the neighborhood as follows: If $k < 8$, add all heaps to neighborhood. If $8 < k < 16$, neighborhood size is equal to eight and if $k > 16$, neighborhood size is equal to sixteen.
3. Compute optimal number of objects (n_{opt}) for merging criterion by subtracting maximum number of objects in heaps from average number of objects in heaps and dividing this to five.
4.
5. Repeat

   (a) Calculate the distance of each object in a heap to its center and center of other heap,

   (b) Compare calculated distances in (a).

   (c) If there are at least n_{opt} objects whose distances to their heap's center are larger than their distances to other heap's center, then merge two heaps. Otherwise, proceed with next heaps.

   Until stopping criterion

Figure 3. The general structure of the proposed clustering scheme.

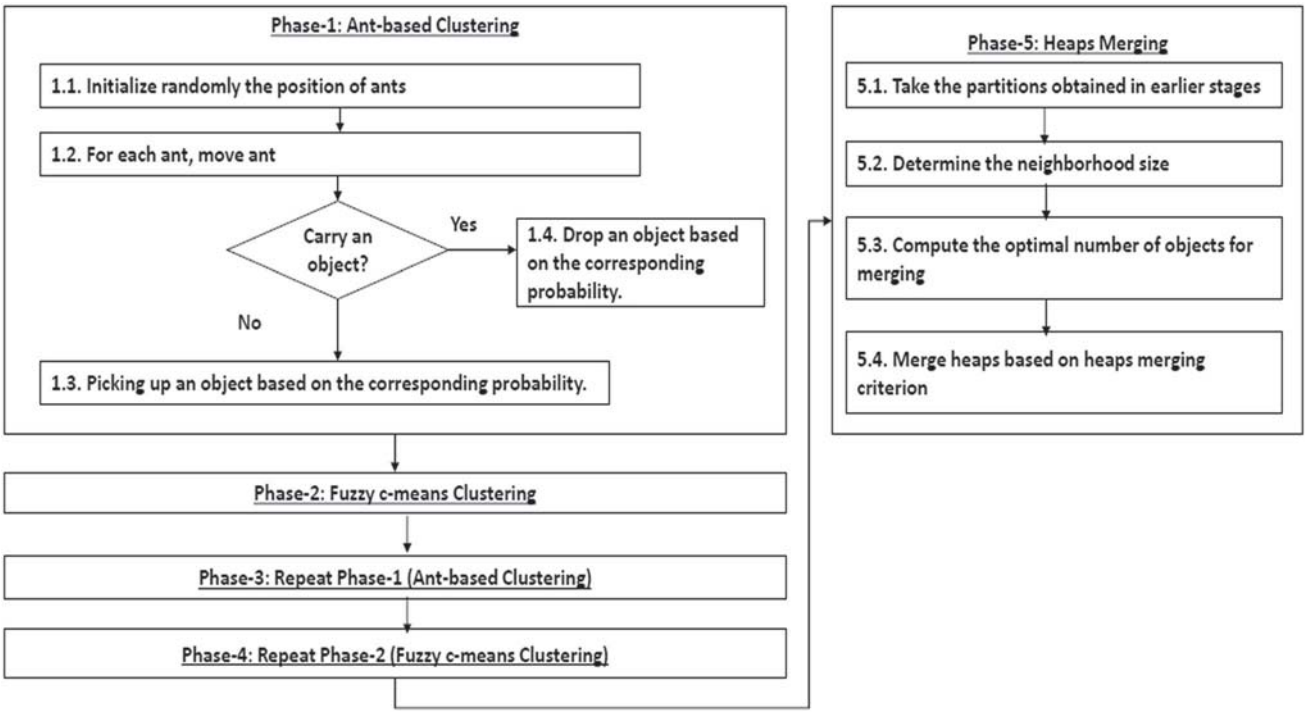

Figure 4. The block diagram of the proposed clustering scheme.

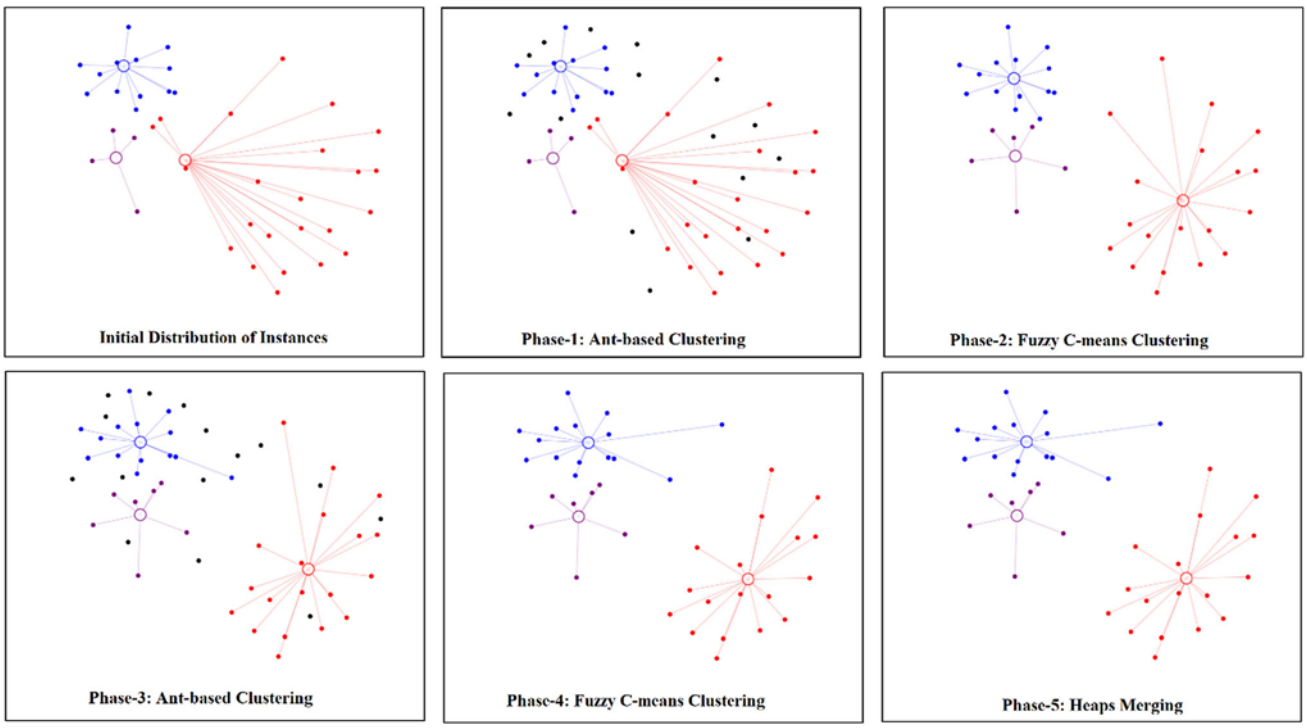

Figure 5. The illustration of the stages of the proposed clustering scheme.

algorithm (ANT), fuzzy c-means (FCM), hybrid clustering (ANT+FCM) and the proposed clustering which combines ant-based clustering with FCM scheme (ANT+FCM+HM).

Table 2. The descriptive information regarding the datasets.

Dataset name	Array type	Tissue	Total samples	Number of classes
Alizadeh-2000-v1	Double channel	Blood	42	2
Alizadeh-2000-v2	Double channel	Blood	62	3
Armstrong-2002-v1	Affymetrix	Blood	72	2
Armstrong-2002-v2	Affymetrix	Blood	72	3
Bittner-2000	Double channel	Skin	38	2
Bredel-2005	Double channel	Brain	50	3
Chen-2002	Double channel	Liver	179	2
Chowdary-2006	Affymetrix	Breast, colon	104	2
Dyrskjot-2003	Affymetrix	Bladder	40	3
Garber-2001	Double Channel	Lung	66	4
Golub-1999-v1	Affymetrix	Bone Marrow	72	2
Golub-1999-v2	Affymetrix	Bone Marrow	72	3
Gordon-2002	Affymetrix	Lung	181	2
Khan-2001	Double channel	Multi-tissue	83	4
Laiho-2007	Affymetrix	Colon	37	2
Liang-2005	Double channel	Brain	37	3
Nutt-2003-v1	Affymetrix	Brain	50	4
Nutt-2003-v2	Affymetrix	Brain	28	2

Table 3. Number of features for microarray gene expression datasets.

	Without FS (pre-processed)	CFS	eBayes	PLSS	RF	SAM	ENS
Alizadeh-2000-v1	1095	45	100	100	100	100	100
Alizadeh-2000-v2	2093	57	100	100	100	100	100
Armstrong-2002-v1	1081	60	100	100	100	100	100
Armstrong-2002-v2	2194	53	100	100	100	100	100
Bittner-2000	2201	46	100	100	100	100	100
Bredel-2005	1739	41	100	100	100	100	100
Chen-2002	85	19	85	85	85	85	85
Chowdary-2006	182	18	100	100	100	100	100
Dyrskjot-2003	1203	43	100	100	100	100	100
Garber-2001	4553	86	100	100	100	100	100
Golub-1999-v1	1877	58	100	100	100	100	100
Golub-1999-v2	1877	64	100	100	100	100	100
Gordon-2002	1626	84	100	100	100	100	100
Khan-2001	1069	62	100	100	100	100	100
Laiho-2007	2202	72	100	100	100	100	100
Liang-2005	1411	28	100	100	100	100	100
Nutt-2003-v1	1377	72	100	100	100	100	100
Nutt-2003-v2	1070	40	100	100	100	100	100

To evaluate the performance of proposed clustering algorithm and feature selection methods, adjusted rand index and jaccard coefficient (J) are utilized as the evaluation function. Adjusted rand index (ARI) and jaccard coefficient (J) are calculated as given by Equations (9) and (10), respectively [48]:

$$RI_{adj} = \frac{2(ad - bc)}{(a+b)(b+d) + (a+c)(c+d)} \tag{9}$$

$$J = \frac{a}{a+b+c} \tag{10}$$

where a denotes the number of objects that are assigned to the same cluster in V and U, b denotes the number of pairs in the same cluster in U, but not in V, c denotes the number of pairs in the same clusters in V, but not in U and d denotes the number of pairs at the different clusters in U and V. For adjusted rand index, higher value of it indicates better quality of clustering. Jaccard coefficient takes values in the range of [0–1] and higher value of it indicates better quality of clustering. The first concern of the experimental analysis is to identify which feature selection method obtains the optimal feature subset for microarray gene expression data. In order to answer this query, we have analyzed the

performance of datasets without feature selection and with several feature selection methods on clustering algorithms. The average adjusted rand index values and jaccard coefficient values obtained by feature selection methods are summarized in tables 4 and 5, respectively. As it can be seen from the adjusted rand index values presented in table 4, the lowest average adjusted rand index values for microarray gene expression data are obtained when feature selection is not applied. Among the configurations with feature selection, the highest performance has been generally achieved by correlation-based feature selection (CFS). For consensus clustering, hierarchical clustering and hybrid clustering, random forest based feature selection achieves the highest performance. For AntClass algorithm, DIANA algorithm and fuzzy c-means algorithm, the best performance has been achieved by ensemble feature selection method. Regarding the jaccard coefficient results presented in table 5, the highest performance has been generally achieved by correlation-based feature selection (CFS).

The second concern of the study is to identify which clustering algorithms yield better results on microarray gene expression data and whether ant-based clustering algorithm or the proposed clustering scheme achieves comparable results to the conventional clustering algorithms. The results presented in table 4 indicate that the highest performance in terms of adjusted rand index has been achieved by the proposed clustering scheme (ANT+FCM+HM). This is followed by ANT+FCM and ANT configurations. The performance of clustering algorithms varies based on the applied feature selection method.

To further evaluate the performance of clustering algorithms on the datasets, we have presented in tables 6 and table 7, the adjusted rand index values and jaccard coeffient values obtained on the individual gene expression datasets when correlation-based feature selection is applied, respectively. These results also indicate that the proposed clustering scheme can effectively cluster microarray gene

Table 4. Average ARI results for feature selection methods.

Clustering algorithm	Without FS	CFS	eBayes	PLSS	RF	SAM	ENS
ANT	0.52	**0.73**	0.70	0.65	0.69	0.69	**0.73**
ANT+FCM	0.57	**0.76**	0.73	0.69	0.73	0.72	0.75
ANT+FCM+HM	0.62	**0.79**	0.76	0.72	0.75	0.74	0.78
CLARA	0.29	**0.55**	0.44	0.41	0.54	0.42	0.40
CONS	0.31	0.58	0.49	0.38	**0.60**	0.50	0.52
DIANA	0.39	0.46	0.45	0.45	0.52	0.47	**0.53**
FCM	0.51	**0.69**	0.65	0.62	0.64	0.66	**0.69**
HIER	0.24	0.50	0.46	0.36	**0.53**	0.51	0.46
HYBRID	0.24	0.50	0.46	0.36	**0.53**	0.51	0.46
KM	0.36	**0.56**	0.45	0.40	0.49	0.44	0.43
PAM	0.28	**0.56**	0.53	0.45	0.51	0.54	0.55
SOM	0.26	**0.57**	0.50	0.39	0.49	0.51	0.54
SOTA	0.26	**0.57**	0.50	0.39	0.49	0.51	0.54

Highest performance obtained by the measure is indicated in bold

Table 5. Jaccard coefficient results for feature selection methods.

Clustering algorithm	Without FS	CFS	eBayes	PLSS	RF	SAM	ENS
ANT	0.66	**0.84**	0.80	0.79	0.79	0.71	0.78
ANT+FCM	0.63	0.73	0.83	0.70	0.72	0.79	**0.86**
ANT+FCM+HM	0.67	**0.90**	0.84	0.83	0.74	0.85	0.77
CLARA	0.42	**0.67**	0.57	0.52	0.67	0.57	0.52
CONS	0.44	**0.65**	0.58	0.52	0.62	0.59	0.58
DIANA	0.49	**0.63**	0.56	0.61	0.61	0.58	0.62
FCM	0.60	**0.79**	0.70	0.76	0.74	0.73	0.79
HIER	0.40	**0.63**	0.59	0.50	0.60	0.60	0.60
HYBRID	0.40	0.57	0.57	0.47	**0.64**	0.61	0.61
KM	0.50	**0.70**	0.52	0.53	0.56	0.58	0.54
PAM	0.40	**0.70**	0.64	0.60	0.56	0.66	0.60
SOM	0.41	**0.68**	0.64	0.54	0.58	0.56	0.62
SOTA	0.42	**0.69**	0.63	0.52	0.63	0.60	0.68

Highest performance obtained by the measure is indicated in bold

Table 6. Average ARI results for clustering methods.

Dataset	PAM	KM	CLARA	SOM	SOTA	HIER	DIANA	HYBRID	CONS	ANT	FCM	ANT+ FCM	ANT+ FCM+ HM
Alizadeh-2000-v1	1.00	0.81	0.65	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Alizadeh-2000-v2	0.95	0.98	0.93	0.95	0.95	0.69	0.31	0.69	0.95	0.81	0.82	0.87	0.98
Armstrong-2002-v1	0.56	0.37	0.37	0.51	0.51	0.33	0.30	0.33	0.37	0.62	0.58	0.61	0.63
Armstrong-2002-v2	0.81	0.51	0.51	0.77	0.77	0.56	0.56	0.56	0.77	0.84	0.59	0.84	0.85
Bittner-2000	0.80	0.80	0.89	0.80	0.80	0.80	0.80	0.80	0.80	0.90	0.88	0.91	0.94
Bredel-2005	0.49	0.50	0.50	0.52	0.52	0.60	0.37	0.60	0.52	0.51	0.52	0.54	0.57
Chen-2002	− 0.01	− 0.01	− 0.01	− 0.01	− 0.01	− 0.01	0.60	− 0.01	− 0.01	0.60	0.61	0.61	0.64
Chowdary-2006	0.07	0.07	0.07	0.11	0.11	0.92	0.92	0.92	0.11	0.95	0.73	0.95	0.97
Dyrskjot-2003	0.68	0.35	0.35	0.61	0.61	0.13	0.11	0.13	0.35	0.71	0.38	0.74	0.76
Garber-2001	0.08	0.27	0.24	0.14	0.14	− 0.01	− 0.01	− 0.01	0.14	0.35	0.42	0.43	0.44
Golub-1999-v1	0.68	0.89	0.89	0.73	0.73	0.48	0.44	0.48	0.89	0.89	0.89	0.92	0.94
Golub-1999-v2	0.82	0.86	0.86	0.95	0.95	0.82	0.93	0.82	0.95	0.89	0.92	0.94	0.97
Gordon-2002	− 0.03	− 0.03	− 0.02	− 0.03	− 0.03	0.84	0.81	0.84	− 0.03	0.83	0.82	0.84	0.87
Khan-2001	1.00	0.83	0.83	0.59	0.59	0.89	0.85	0.89	0.92	0.89	0.84	0.94	1.00
Laiho-2007	0.56	0.34	0.34	0.36	0.36	0.08	− 0.02	0.08	0.50	0.54	0.58	0.60	0.63
Liang-2005	0.57	0.16	0.16	0.61	0.61	0.16	0.16	0.16	0.16	0.54	0.57	0.60	0.62
Nutt-2003-v1	0.16	0.20	0.20	0.28	0.28	− 0.01	0.00	− 0.01	0.22	0.39	0.24	0.39	0.41
Nutt-2003-v2	0.31	0.22	0.22	0.15	0.15	− 0.01	− 0.01	− 0.01	0.15	0.32	0.34	0.36	0.39
Yeast ATP	0.80	0.41	0.41	0.72	0.72	0.15	0.13	0.15	0.41	0.84	0.85	0.87	0.89
Yeast PHO	0.67	0.34	0.34	0.60	0.60	0.13	0.11	0.13	0.34	0.70	0.69	0.70	0.72
Yeast AFR	0.80	0.41	0.41	0.72	0.72	0.15	0.13	0.15	0.41	0.84	0.85	0.87	0.89

Table 7. Average ARI results for clustering methods.

Dataset	PAM	KM	CLARA	SOM	SOTA	HIER	DIANA	HYBRID	CONS	ANT	FCM	ANT+ FCM	ANT+ FCM+ HM
Alizadeh-2000-v1	0.71	0.62	0.52	0.86	0.80	0.83	0.75	0.86	0.72	0.70	0.82	0.89	0.89
Alizadeh-2000-v2	0.86	0.83	0.75	0.78	0.79	0.59	0.24	0.56	0.76	0.63	0.74	0.87	0.88
Armstrong-2002-v1	0.44	0.30	0.27	0.45	0.45	0.24	0.27	0.28	0.27	0.49	0.47	0.51	0.52
Armstrong-2002-v2	0.63	0.43	0.42	0.64	0.55	0.39	0.49	0.39	0.59	0.67	0.43	0.70	0.71
Bittner-2000	0.62	0.72	0.69	0.66	0.59	0.56	0.57	0.70	0.63	0.68	0.64	0.71	0.73
Bredel-2005	0.44	0.38	0.42	0.46	0.36	0.42	0.33	0.43	0.44	0.46	0.38	0.48	0.55
Chen-2002	0.31	0.31	0.33	0.33	0.33	0.33	0.43	0.44	0.45	0.44	0.53	0.49	0.58
Chowdary-2006	0.58	0.59	0.59	0.57	0.59	0.55	0.64	0.67	0.08	0.67	0.59	0.69	0.72
Dyrskjot-2003	0.52	0.31	0.31	0.48	0.44	0.10	0.09	0.09	0.32	0.50	0.31	0.57	0.61
Garber-2001	0.21	0.22	0.23	0.12	0.12	0.32	0.35	0.38	0.39	0.40	0.30	0.37	0.44
Golub-1999-v1	0.61	0.66	0.75	0.62	0.58	0.35	0.34	0.42	0.78	0.77	0.66	0.67	0.79
Golub-1999-v2	0.74	0.76	0.77	0.74	0.68	0.66	0.71	0.65	0.74	0.77	0.73	0.75	0.75
Gordon-2002	0.52	0.55	0.56	0.61	0.61	0.67	0.61	0.61	0.67	0.68	0.58	0.65	0.73
Khan-2001	0.69	0.69	0.71	0.53	0.49	0.73	0.70	0.68	0.62	0.71	0.64	0.72	0.75
Laiho-2007	0.46	0.41	0.45	0.48	0.49	0.47	0.45	0.46	0.39	0.49	0.50	0.52	0.55
Liang-2005	0.51	0.44	0.41	0.51	0.45	0.42	0.44	0.44	0.44	0.51	0.46	0.53	0.56
Nutt-2003-v1	0.13	0.16	0.15	0.22	0.25	0.24	0.25	0.29	0.31	0.30	0.32	0.32	0.36
Nutt-2003-v2	0.27	0.19	0.18	0.11	0.11	0.21	0.24	0.26	0.29	0.33	0.34	0.38	0.42
Yeast ATP	0.68	0.63	0.62	0.57	0.60	0.53	0.57	0.54	0.59	0.65	0.71	0.73	0.76
Yeast PHO	0.52	0.53	0.55	0.53	0.43	0.51	0.53	0.59	0.59	0.61	0.51	0.58	0.64
Yeast AFR	0.68	0.31	0.34	0.54	0.60	0.61	0.49	0.56	0.54	0.64	0.72	0.69	0.73

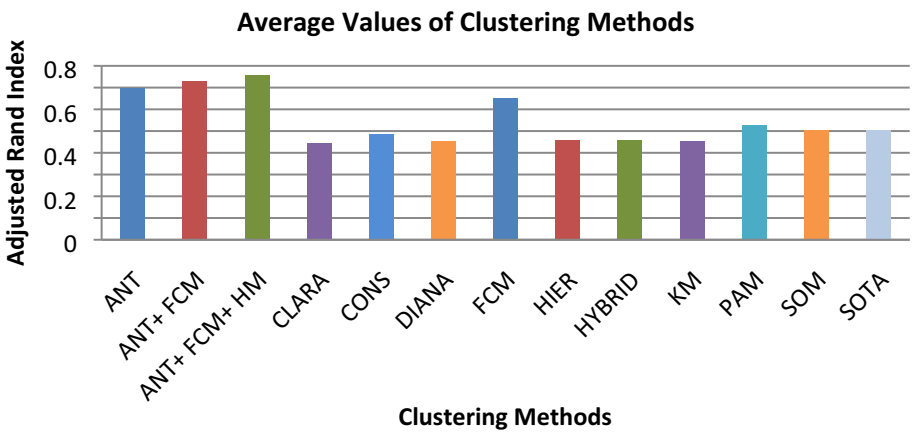

Figure 6. Comparison of average adjusted rand index values for clustering methods.

Table 8. Execution time comparisons for the clustering algorithms.

Clustering algorithm	Execution time (s)
PAM	0.84
KM	0.81
CLA-RA	0.12
SOM	0.90
SOTA	1.65
HIER	0.90
DIANA	0.93
HYBRID	1.59
CONS	1.40
ANT	90.35
FCM	2.21
ANT + FCM	93.46
ANT + FCM+ HM	106.56

expression datasets and yields better adjusted rand index values or jaccard coefficient values from the conventional clustering algorithms. In figure 6, average adjusted rand index values for different clustering algorithms are depicted.

In table 8, the performances of compared clustering algorithms in terms of average running times on gene expression datasets are presented. As it can be seen from the results presented in table 7, conventional clustering algorithms tend to be more effective in terms of running times compared to the metaheuristic clustering algorithms (such as ant-based clustering and the proposed hybrid clustering scheme). However, the clustering quality obtained by the metaheuristic clustering algorithms are more promising compared to the conventional clustering algorithms. Hence, there is a trade-off between execution times and clustering qualities.

For performance comparisons, we have used z-score that is computed on the clusters obtained by the compared clustering algorithms using gene expression datasets. A higher value of z-value indicates that a more biologically relevant and efficient clustering result [18, 19]. In addition, the clustering results of the compared algorithms are analyzed by the functional enrichment analysis [19]. In the functional enrichment analysis, we have used P-values that represent the probability of observing at least a particular number of genes in a cluster are from a specific gene ontology functional category [18, 19]. In Table 9, total number of enriched clusters, total number of enriched attributes and z-scores have been presented. The results reported in table 9 are average results for 21 gene expression datasets. As it can be seen from the results presented in table 9, the proposed clustering scheme (ANT+FCM+HM) achieves higher values for z-score, enriched clusters and enriched attributes. Higher number of functionally enriched attributes and enriched clusters indicate better clustering quality. Hence, the clustering quality obtained by the proposed scheme is more promising compared to the other clustering algorithms.

Microarray gene expression data is characterized by high-dimensional feature space. This characteristic is also valid for other types of expression data (such as RNA and protein sequences). The presented scheme utilizes correlation-based feature selection in conjunction with swarm-based clustering algorithm. As the experimental analysis on gene expression data indicate the clustering quality of conventional clustering algorithms can be improved with the utilization of feature selection. In addition, swarm based approaches to clustering can yield better clustering quality on gene expression data. Conventional clustering algorithms (such as K-means algorithm, hierarchical clustering and model based clustering) have been successfully employed to organize RNA and protein sequences into clusters. Hence, the presented scheme, which combines correlation-based feature selection, ant-based clustering, fuzzy c-means and heaps merging heuristics, can be utilized to cluster RNA and protein sequences.

Table 9. Functional enrichment analysis results for the compared clustering algorithms.

Methods	Total clusters	*Z*-Score	Enriched clusters	Enriched attributes
PAM	7.77	19.46	2.96	18.53
KM	7.12	17.67	3.14	15.83
CLARA	10.85	6.12	1.94	4.46
SOM	48.02	13.15	1.56	8.68
SOTA	42.12	9.08	1.99	20.01
HIER	4.93	7.84	2.13	23.14
DIANA	8.38	6.78	1.10	8.77
HYBRID	5.16	3.82	1.71	25.05
CONS	2.70	2.79	1.88	15.20
ANT	4.10	20.87	3.16	43.97
FCM	5.84	20.28	2.80	22.67
ANT + FCM	6.03	23.06	4.85	68.20
ANT + FCM + HM	5.80	27.06	6.26	76.48

5. Conclusion

The clustering gene expression data is an important research direction in medical informatics. In this study, an improved ant-based clustering algorithm has been presented. The improved ant-based clustering algorithm consists of two consecutive stages of ant-based clustering algorithm followed by fuzzy c-means. In order to solve the problem of generation of over clusters (heaps) at the end of the clustering scheme, a heap merging approach is presented. The proposed clustering scheme is compared to conventional clustering algorithms, such as partitioning around medoids algorithm, self-organizing maps, hierarchical clustering and AntClass algorithm in terms of adjusted rand index on gene expression datasets. The experimental results indicate that ant-based clustering can be utilized as a viable tool in the clustering gene expression data. Based on the experiments with gene expression data, the proposed heaps merging approach improve the clustering quality and outperform conventional clustering algorithms.

References

[1] Dalton L, Ballarin V and Brun M 2009 Clustering algorithms: on learning, validation, performance and applications to genomics. *Current Genomics* 10: 430–445

[2] Daxin J, Tang C and Zhang A 2004 Cluster analysis for gene expression data: a survey. *IEEE Transactions on Knowledge and Data Engineering* 16(11):1370–1386

[3] De Souto M C P, Costa I G, De Araujo D S A, Ludermir T B and Schliep A 2008 Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics* 9: 497

[4] Hasan M J A and Ramakrishnan S 2011 A survey: hybrid evolutionary algorithms for cluster analysis. *Artificial Intelligence Review* 36(3): 179–204

[5] Alon U, Barkai N and Notterman D A 1999 Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America* 96: 6745–6750

[6] Golub T R, Slonim D K and Tamayo P 1999 Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286: 531–537

[7] Alizadeh A A, Eisesn M B and Davis R E 2000 Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403: 503–511

[8] Dudoit S and Fridlyand J 2002 A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology* 3(7):1–21

[9] Datta S and Datta S 2003 Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics* 19(4): 459–466

[10] Costa I G, de Carvalho F A T and de Souto M C P 2004 Comparative analysis of clustering methods for gene expression time course data. *Genetics and Molecular Biology* 27(4): 623–631

[11] Iam-on N and Boongoen T 2012 A new locally weighted k-means for cancer-aided microarray data analysis. *Journal of Medical Systems* 36: 43–49

[12] Castellanos-Garzon J A and Diaz F 2013 An evolutionary computational model applied to cluster analysis of DNA microarray data. *Expert Systems with Applications* 40(7): 2575–2591

[13] Binu D 2015 Cluster analysis using optimization algorithms with newly designed objective functions. *Expert Syst Appl* 42(14): 5848–5859

[14] Liu J and Pham T 2011 Fuzzy clustering for microarray data analysis: a review. *Current Bioinformatics* 6(4): 427–443

[15] Bhattacharya A, Chowdhury N and De R K 2012 Comparative analysis of clustering and biclustering algorithms for grouping of genes: co-function and co-regulation. *Current Bioinformatics* 7: 63–76

[16] Datta S and Mukhopadhyay S 2013 An in silico identification of human promoters: a soft computing based approach. *Current Bioinformatics* 8(3): 362–368

[17] Bhattacharya A and De R K 2008 Divisive correlation clustering algorithm (DCCA) for grouping of genes: detecting varying patterns in expression profiles. *Bioinformatics* 24(11):1359–1366.

[18] Bhattacharya A and De R K 2009 Bi-correlation clustering algorithm for determining a set of co-regulated genes. *Bioinformatics* 25(21):2795–2801

[19] Bhattacharya A and De R K 2010 Average correlation clustering algorithm (ACCA) for grouping of co-regulated genes with similar pattern of variation in their expression values. *Journal of Biomedical Informatics* 43:560–568

[20] Turner H, Bailey T and Krzanowski W 2005 Improved biclustering of microarray data demonstrated through systematic performance tests. *Computational Statistics and Data Analysis* 48(2):235–254.

[21] Santamaria R, Quintales L and Theron R 2007 Methods to bicluster validation and comparison in microarray data. In: *Proceedings of 8th International Conference Intelligent Data Engineering and Automated Learning* 780–789

[22] Filippone M, Masulli F and Rovetta S 2008 Stability and performances in biclustering algırithms. In: *Proceedings of the International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics* 91–101

[23] Ayadi W, Elloumi M and Hao J-K 2012 Bicfinder: a biclustering algorithm for microarray data analysis. *Knowledge and Information Systems* 30(2):341–358

[24] Saber H B and Elloumi M 2015 A novel biclustering algorithm of binary microarray data: BiBincons and Bibinalter. *BioData Mining* 38:1–14

[25] Eren K, Deveci M, Küçüktunc O and Çatalyürek Ü V 2013 A comparative analysis of biclustering algorithms for gene expression data. *Brief Bioinformatics* 14(3):279–292

[26] Monmarche N, Slimane N and Venturini G 1999 *AntClass: discovery of clusters in in numerical data by an hybridization of an ant colony with the Kmeans algorithm*. Internal Report, Universite de Tours

[27] Monmarche N, Slimane N and Venturini G 1999 On improving clustering in numerical databases with artificial ants. *Lecture Notes in Computer Science* 1674: 626–635

[28] Chandrashekar G and Sahin F 2014 A survey on feature selection methods. *Computers and Electrical Engineering* 40: 16–28

[29] Glaab E 2011 *Analysing functional genomics data using novel ensemble, consensus and data fusion techniques.* Unpublished PhD Thesis, University of Nottingham, Nottingham, UK

[30] Loennstedt I and Speed T P 2002 Replicated microarray data. *Statistica Sinica* 12: 31–46

[31] Symth G K 2004 Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* 3(1): 1–25

[32] Boulesteix A and Strimmer K 2007 Partial least squares: a versatile tool for the analysis of high dimensional genomic data. *Briefings in Bioinformatics* 8: 32–44

[33] Breiman L 2001 Random forests. *Machine Learning* 45(1): 5–32

[34] Tusher V, Tibshirani R and Chu G 2001 Significance analysis of microarrays applied to ioinizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* 98: 5116–5121

[35] Hall M A 1999 *Correlation-based feature selection for machine learning*. Unpublished PhD Thesis, University of Waikato, Hamilton, New Zealand

[36] Daxin J, Tang C and Zhang A 2004 Cluster analysis for gene expression data: a survey. *IEEE Transactions on Knowledge and Data Engineering* 16(11): 1370-1386

[37] Xu R and Wunsch D 2005 Survey of clustering algorithms. *IEEE Transactions on Neural Networks* 16(3): 654–678

[38] Han J and Kamber M 2006 Data mining concepts and techniques. San Francisco, Morgan Kaufmann

[39] Jain A K 2010 Data clustering: 50 years beyond k-means. *Pattern Recognition Letters* 31: 651–656

[40] Kaufman L and Rousseeuw P J 1990 *Finding groups in data: an introduction to cluster analysis*. New Jersey, John Wiley & Sons

[41] Park H S and Jun C H 2009 A simple and fast algorithm for k-medoids clustering. *Expert Systems with Applications* 36. 3336–3341

[42] Aggarwal C C and Reddy C K 2013 *Data clustering: algorithms and applications*, San Francisco, CRC

[43] Johnson R A and Wichern D W 2007 *Applied multivariate statistical analysis*. New Jersey, Prentice Hall

[44] Herrero J, Valencia A, Dopazo J 2005 A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics* 17:126–136

[45] Chipman H and Tibschirani R 2006 Hybrid hierarchical clustering with applications to microarray data. *Biostatistics* 7(3): 286–301

[46] Onan A 2013 *A study of hybrid evolutionary algorithms for cluster analysis*. Unpublished Master thesis, Ege University, Izmir, Turkey

[47] Onan A, Bulut H and Korukoğlu S 2017 An improved ant algorithm with LDA-based representation for text document clustering. *Journal of Information Science* 43(2): 275-292

[48] Chandra E and Anuradha VP 2011 A survey on clustering algorithms for data in spatial database management systems. *International Journal of Computer Applications* 24(9): 19–26