Attractive Clustering Approach for Knowledge Discovery in Gene Expression Data

Mohammed Alshalalfa
Department of Computer Science
University of Calgary
Calgary, Alberta, Canada
msalshal@ucalgary.ca

Tansel Özyer
Department of Computer Engineering
TOBB University of Economic and Technology
Ankara, Turkey
ozyer@etu.edu.tr

Reda Alhajj
Department of Computer Science
University of Calgary
Calgary, Alberta, Canada
alhajj@ucalgary.ca
and
Department of Computer Science
Global University
Beirut, Lebanon

Abstract

One of the major drawbacks of the clustering algorithms described in the literature is requiring some parameters to guide the clustering process towards a certain solution which may not be necessary the most appropriate to the data in hand. This problem is mainly handled by the work described in this paper where the major contribution could be articulated as a parameter free clustering approach that leads to appropriate distribution of the given data instances into the most convenient clusters. This goal is realized in several steps. First, we apply multi-objective genetic algorithm to determine some alternative clustering solutions that constitute the pareto front. The result is pool of the clusters reported by all the solutions. Then, we determine the homogeneity of each cluster in the pool to keep the most homogeneous clusters. Finally, as a given data instance may belong to more than one cluster in the solution set we reduce this membership to the cluster in which the instance is closest to the centroid. Many applications like gene expression data analysis are in need for such parameter free approach because the correctness of the post processing is directly affected by the outcome form the clustering process. We demonstrate the applicability and effectiveness of the proposed clustering approach by conducting experiments using some benchmark gene expression data sets available at the Genomics Department of Stanford University.

Keywords: multi-objective genetic algorithm, clustering, knowledge discovery, gene expression data.

1 Motivation and Contribution

The central role of the DNA microarray technology in biological and biomedical domain allowed researchers to observe transcription levels of many thousands of genes. Information gathered by analyzing the genes at different levels (stages of the process) is used for finding the gene function, the reconstruction of the gene regulatory network, diagnosis of disease biomarkers and inference of medical treatment. Gene expression data analysis gives insight into critical issues related to several diseases. In other words, identifying differentially expressed genes is an interesting problem that has received considerable attention of researchers who already realized its scientific and social impacts; for instance, identifying differentially expressed genes would help in classifying cancerous cells [1, 2, 3, 4, 6].

There are two common trends to handle the problem, namely experimentally in the lab or computationally by applying machine learning, data mining and statistical techniques. While the former approach is costly and time consuming, the latter approaches are more attractive to analyze large volumes of data. Data mining and statistical techniques made it easier to interpret, understand, and extract the knowledge hidden within the microarray data and large data collections in general [11]. In other words, to analyze the expression level of all the genes in human cancer, microarray is used to study the gene expression of all the genes in "normal" and cancerous humans. The result of this gene expression study is a matrix of *m* genes and *n* samples, where samples represent either "normal" or cancerous humans; this matrix is called the microarray data. To help biologists and medical scientist developing effective analysis, different statistical and computing techniques are employed in the process; the main target is to reduce the space for better control and analysis. The goal to be achieved is the ability to distinguish between normal and cancer samples based on subset of features (genes) selected from the microarray data. This discovery gives insight into the cancer biomarker to be considered mostly in developing appropriate treatment and hence the outcome may shape the focus of the drug industry.

Several methods have already been proposed to extract the smallest number of biomarkers that can accurately classify different cancer samples from non-cancer samples. However, most of the proposed approaches ignore the fact that the microarray data is noisy; and thus, they have to deal with the data more carefully. We set two main objectives to handle the problem. First, we extract the smallest possible number of features, although the best number of features is a problem yet to be considered. Second, it is required to reduce the functional redundancy within the extracted genes. We argue that these two objectives could be satisfied by a powerful clustering approach. In this sense, several clustering approaches have been proposed so far [16], e.g., k-means, Fuzzy C-means and Self-Organizing Maps (SOM) have been used to cluster the samples into two or more classes depend on the number of available cancer samples. Unfortunately, the target clustering approach is yet to be developed. One such approach is the focus of the study described in this paper. Actually, clustering groups genes with similar expression profile. It ensures that we extract genes with different functionalities based on the hypothesis that genes with similar expression profile have similar functions. The extracted genes are considered as good representatives for the data.

Our approach presented in this paper has been designed to handle the clustering of a given set of instance without requiring any parameter be specified in advance; approaches like k-means require the number of clusters explicitly specified and approaches like DBSCAN are based on some parameters that implicitly simulate the number of clusters. A parameter free clustering approach is critical for many domains where finding the most appropriate clustering is directly reflected into the analysis of the results. One such domain is gene expression data analysis, which is the main concentration of the experiments conducted in this study. Our approach starts by applying a multi-objective k-means genetic algorithm (MOKGA) in order to determine several alternative clustering solutions without taking the weight values into account [24]. We run cluster validity analysis, namely Dunn index [9], Davies-Bouldin index [8], Silhouette index [18], C index [15], SD index [12] and S_Dbw index [13] on the alternative solutions to determine the number of compact clusters to have in the final solution. Then, we collapse all the alternative solutions obtained from MOKGA to form a common pool of clusters, where clusters coming from the same solution are disjoint and clusters from different solutions mostly overlap. Analyzing all the clusters in the pool at once gives equal opportunity to every cluster to show up in the final solution which should include the most compact clusters. This is more natural process than analyzing the alternative solutions themselves. In other words, we zoom into the details of each solution because some solutions may include more compact individual clusters than a single favored solution. At the end of this process, we will have a collection of compact clusters that mostly overlap. The overlap is eliminated by keeping each data instance only in the cluster where the data instance is closest to the centroid of the cluster. In case of objects that do not end up in any of the identified clusters, we first measure the distance between these objects and the centroids of the clusters in the final solution set. Then we have two choices, either to add an object to a cluster based on shortest distance and provided that it does not destroy the compactness of the cluster or to consider the object as outlier otherwise.

As gene expression data analysis is concerned, benefiting from the advantage of the proposed clustering approach, we use the gene closest to the centroid as reduced feature to represent the cluster. Thus, after the clustering is over and the most appropriate clustering is identified, the genes closest to centroids (one gene per cluster)

represent the whole data. The latter genes form valuable source of information for further analysis of the gene expression data to discover the biomarkers [2, 3, 4, 27]. In our previous work described in [2, 3, 4], we perform a kind of controlled multilevel (hierarchical) clustering to select some representative genes; one gene per cluster. However, the compact solution produced by the clustering approach described in this paper provides the opportunity to consider more appropriate biomarker genes. Finally, the applicability and effectiveness of the proposed approach has been tested using three benchmark data sets; the results are promising. Here it is worth mentioning that the proposed approach is capable of locating outliers, but this property is still to be validated by considering some other synthetic or real data sets with outliers. We have left this out as future study because none of the data sets used in the testing contains outliers.

The rest of the paper is organized as follows. Section 2 is an overview of multi-objective optimization. Section 3 describes the proposed approach. Section 4 discusses the experiments. Section 5 is summary and conclusions.

2 Overview of Multi-Objective Optimization

A multi-objective optimization problem has *n* decision variables, *k* objective functions, and *m* constraints. Objective functions and constraints are functions of the decision variables. The optimization goal may be described as follows:

$$\begin{aligned} &\max imize \setminus \min imize \quad y = f(x) = (f_1(x), f_2(x), ..., f_k(x)) \\ &subject \quad to \quad e(x) = (e_1(x), e_2(x), ..., e_m(x)) \le 0 \\ &where \qquad x = ((x_1, x_2, ..., x_n) \in X \\ &\qquad\quad y = ((y_1, y_2, ..., y_n) \in Y \end{aligned} \tag{1}$$

where x is the decision vector, y is the objective vector, X denotes the decision space, and Y is called the objective space. The constraints $e(x) \ge 0$ determine the set of feasible solutions [29].

Solutions to a multi-objective optimization method are mathematically expressed in terms of non-dominated or superior points. In a minimization problem, a vector $x^{(1)}$ is partially less than another vector $x^{(2)}$, denoted $x^{(1)} \prec x^{(2)}$, when no value of $x^{(2)}$ is less than $x^{(1)}$ and at least one value of $x^{(2)}$ is strictly greater than $x^{(1)}$. If $x^{(1)}$ is partially less than $x^{(2)}$, we say that $x^{(1)}$ dominates $x^{(2)}$ or the solution $x^{(2)}$ is inferior to $x^{(1)}$. Any vector which is not dominated by other vectors is said to be non-dominated or non-inferior. The optimal solutions to a multi-objective optimization problem are non-dominated solutions [21].

A common difficulty with the multi-objective optimization is the conflict between the objective functions. None of the feasible solutions allow optimal solutions for all the objectives. Pareto-optimal is the solution, which offers the least objective conflict. In traditional multi-objective optimization, objectives are combined to form one objective function. One of the traditional methods being used is weighting each objective and scalarizing the result. At the end of each run, pareto-optimal front may be obtained. But it actually represents one single point. However, the approach described in this paper favors producing all the alternative solutions along the pareto-front. Reporting the latter alternative solutions is the key step in our approach. In other words, the remaining steps of our approach depend on the outcome from MOKGA. The objectives to be optimized in the multi-objective process applied in this study are: maximizing cluster homogeneity, maximizing separateness between the clusters minimizing the number of clusters and minimizing the partitioning error.

3 The Proposed Approach

In this section, we describe the clustering approach that starts by applying the Multi-Objective Genetic K-means algorithm (MOKGA) to produce alternative solutions which are collapsed into one pool of clusters to be further analyzed. Although we tested our approach on gene expression data, it is a general purpose approach for clustering other datasets after modifying the fitness functions and changing the proximity values as distance or non-decreasing similarity function according to the requirements of the dataset to be clustered. Our interest in the gene expression data is because we have already produced a successful approach for identifying biomarker genes [2, 3, 4, 27], and we target to integrate the new approach into the validation process, i.e., to at least confirm whether the genes identified by the previous approach [2, 3, 4, 27] are in fact the most representative biomarkers.

Concerning our approach, after running MOKGA, we get the pareto-optimal front that gives the alternative solutions. Then, the system analyzes the clustering results by applying six of the cluster validity techniques proposed in the literature, namely Silhoutte, *C* index, Dunn's index, SD index, DB index and S_Dbw index. The favored number of clusters guides the process in selecting the most compact clusters from the pool.

The employed clustering approach MOKGA is basically the combination of the Fast Genetic *K*-means Algorithm (FGKA) [17] and Niched Pareto Genetic Algorithm [14].

MOKGA uses a list of parameters which has nothing to do with the clustering process; these parameters are particular to the process of the genetic algorithm: population size (number of chromosomes), t_dom (number of comparison set) representing the assumed non-dominated set, mutation probability and the number of iterations that the execution of the algorithm needs run in order to report the result.

Sub-goals can be defined as fitness functions; instead of scalarizing them to find the goal as the overall fitness function with the user defined weight values, we expect the system to find the set of best solutions, i.e., the pareto-optimal front. By using the specified formulas, at each generation, each chromosome in the population is evaluated and assigned a value for each fitness function.

The coding of our individual population is a chromosome of length *N* (number of data points). Each allele in the chromosome takes a value from {1, 2,…, *K*}, and represents a pattern. The value indicates which cluster the corresponding pattern belongs to.

1. Initially, assign the current generation to 0. A population with the specified number of chromosomes is created randomly by using the method in [18]: Data points are randomly assigned to each cluster at the beginning. By using this method, we can avoid generating illegal strings where some clusters do not have any pattern in the string.

2. Generate the next population and increment the current generation by 1.
   a. The first step in the construction of the next generation is the selection using pareto domination tournaments: In this step, two candidate items picked among (*population size- t_{dom}*) individuals participate in the pareto-domination tournament against the t_{dom} individuals for the survival of each in the population. In the selection part, t_{dom} individuals are randomly picked from the population. With two randomly selected chromosome candidates in (*population size- t_{dom}*), each of the candidates is compared against each individual in the comparison set, t_{dom}. A candidate that has larger total within-cluster variation fitness value and larger number of clusters than all of the chromosomes in the comparison set is said to be dominated by the comparison set already and will be deleted from the population permanently. Otherwise, it resides in the population.
   b. Some of our initial experiments demonstrated that one-point cross-over leads to earlier converging to the solution than multi-point attempts. So, in this study one-point crossover operator is applied on two randomly chosen chromosomes. The crossover operation is carried out on the population with crossover rate p_c. After the crossover, assigned cluster numbers for each gene are renumbered beginning from a_1 to a_n. For example, consider the following two chromosomes having 3 and 5 clusters, respectively:

      Number of clusters=3: 1 2 3 3 3;
      Number of clusters=5: 1 4 3 2 5,

   They need to have a crossover at the third location, we will get the new chromosomes: 1 2 3 2 5 and 1 4 3 3 3; the clusters in these chromosomes are renumbered as follows:

   Number of clusters=4: 1 2 3 2 4 (for 1 2 3 2 5)
   Number of clusters=3: 1 2 3 3 3 (for 1 4 3 3 3)

   The mutation operator on the current population is employed after the crossover. During the mutation, each gene value a_n is replaced by a_n', with respect to the probability distribution: for n=1 to N, simultaneously. a_n' is a cluster number randomly selected from {1, …, K} with the probability distribution {p_1, p_2,…, p_K} defined as:

$$p_i = \frac{1.5 \times d_{\max}(\overline{X_n}) - d(\overline{X_n}, \overline{C_K})}{\sum_{k=1}^{K} (1.5 \times d_{\max}(\overline{X_n}) - d(\overline{X_n}, \overline{C_k}))} \qquad (2)$$

   where i = (1, 2, …, K) and $d(X_n, C_k)$ denotes the Euclidean distance between pattern X_n and centroid C_k of the k^{th} cluster, $d_{\max}(X_n) = \max_k\{d(X_n, C_k)\}$, p_i represents the probability interval of a mutating gene assigned to cluster i. Using this method, the probability of changing gene value a_n to a cluster number k is greater if X_n is closer to the centroid of the k^{th} cluster G_k.
   c. Perform the k-means operator. The k-means operator is used to reanalyze each chromosome gene's assigned cluster value and it calculates the cluster center for each cluster and then it re-assigns each gene to the cluster closest to the instance in the gene. Hence, the *k*-means operator is used to speed up the convergence process by replacing a_n by a_n' for n=1, …, N, simultaneously, where a_n' is the closest to object X_n in Euclidean distance.

3. If the maximum number of generations is reached or the difference in fitness between two consecutive generations is smaller than a threshold, then exit, else go to step 2.

To guarantee homogeneity in partitioning the N data objects into K clusters one goal is to minimize the Total Within-Cluster Variation (*TWCV*),

$$TWCV = \sum_{n=1}^{N}\sum_{d=1}^{D} X_{nd}^{2} - \sum_{k=1}^{K}\frac{1}{Z_k}\sum_{d=1}^{D} SF_{kd}^{2} \qquad (3)$$

where $X_1, X_2, .. , X_N$ are the N objects, X_{nd} denotes feature d of object X_n (n = 1 to N), Z_k denotes the number of objects in cluster k, and SF_{kd} is the sum of the d-th features of all the objects in cluster k.

$$SF_{kd} = \sum_{\overline{X_n} \in G_k} X_{nd}, \quad (d = 1, 2, ...D). \qquad (4)$$

After getting the patero-front and deciding on the most appropriate number of clustering using validity analysis, the alternative solutions are collapsed into a pool of clusters. Then we compute the distance between each object and the centroid of every cluster to which the object belongs. As a result, every object survives only in the cluster that satisfies the minimum distance. At the end, objects that do not belong to any of the identified compact clusters are classified into two sets: some of them join the existing compact clusters if they are not destroying the compactness; the rest of the objects are classified as outliers. The conducted experiments did not report any outliers for the utilized three benchmark data sets. We will run the proposed approach on some other data sets that do report some outliers; this will give us better insight into the power of the proposed approach in identifying real outliers.

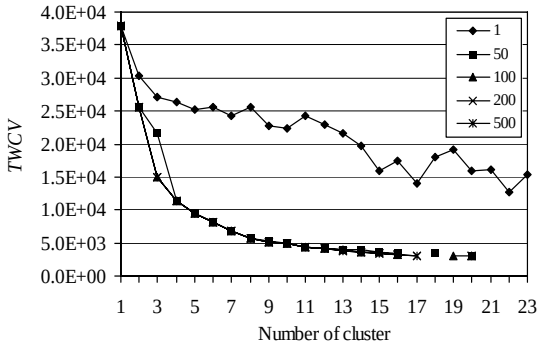

Figure 1 Pareto-fronts for Fig2data dataset

4 Experiments

To evaluate the performance and efficiency of the developed clustering approach, experiments were conducted on a computer with the following features: Pentium PC, 3.00 GHz CPU, 2 GB RAM and running Windows XP. The system was implemented using MS Visual C++. The running platform is Microsoft Visual Studio.NET 2003.

The rest of this section is dedicated to report the results obtained for three highly cited benchmark data sets, namely Fig2Data, cancer (NCI60) and leukemia. The target is to find the most natural clustering for each of these datasets. This will allow us in a latter step (not covered in this paper) to highlight genes that mostly act as disease biomarkers.

4.1 Fig2data Dataset

Fig2data dataset is the time course of serum stimulation of primary human fibroblasts. It contains the expression data for 517 genes of which expression changed substantially in response to serum. Each gene has 19 expressions reflecting the response ranging from 15 minutes to 24 hours.

In this experiment, first MOKGA has been applied to Fig2data dataset with the following parameters: population size = 150, t_{dom} (number of comparison set = 10) and crossover = 0.8, mutation = 0.005, gene mutation rate= 0.005, and threshold = 0.0001, which is applied to check if the population stops evolution after 50 generations and if the process needs to be stopped. The range of [1,25] was picked to find the optimal number of clusters. The convergence towards the pareto-front is reported in Figure 1. Then validity analysis was applied to report best number of clusters. The literature reported that the optimal number of clusters for Fig2data is 10. Consistently, results in this paper indicate that it ranks in the best ones for C index, and 10 clusters is also among the best for the other indices. Actually, SD, S_Dbw, DB, Silhouette, and Dunn indices cannot handle properly arbitrarily shaped clusters, so they do not always give satisfactory results. This is justify our choice to apply majority voting to decide on the best number of clusters. Finally, this result injected the rest of the steps leading to the most compact clusters.

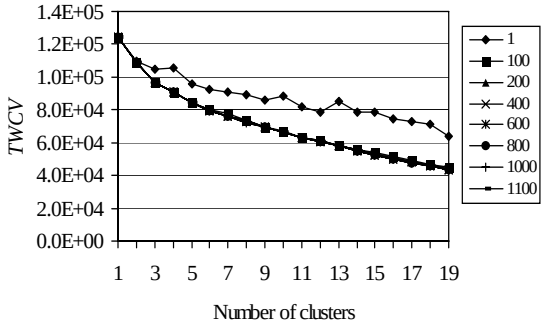

Figure 2 Pareto-fronts for Cancer dataset

4.2 Cancer (NCI60) dataset

NCI60 is a gene expression database for the molecular pharmacology of cancer. It contains 728 genes and 60 cell lines derived from cancers of colorectal, renal, ovarian, breast, prostate, lung, and central nervous system origin, leukaemia and melanoma. Growth inhibition is assessed from changes in total cellular protein after 48 hours of drug treatment using a sulphorhodamine B assay. The patterns of drug activity across the cell lines provide information on mechanisms of drug action, resistance, and modulation.

In our tests, MOKGA has been run for the Cancer dataset with the following parameters: population size = 100, t_{dom} (number of comparison set = 10) and crossover = 0.8, mutation = 0.005, gene mutation rate= 0.005, and threshold = 0.0001 which is used to check whether the population stops evolution for 50 generations or the process needs to be stopped. The range of [1, 20] was picked to find the optimal number of clusters.

Changes in the Pareto-optimal front after running the algorithm are displayed in Figure 2. The validity analysis on the produced alternative solutions reported 15 as the best number of clusters for the cancer (NCI60) dataset; note that this value also ranks the sixth for DB index, fifth for SD index and the fifth for C index. These are consistent with the results reported in the literature. The number of clusters reported by some indices is not good because the results from the validity indices are highly dependent on the shape of the clusters.

4.3 Leukaemia dataset

The third microarray dataset used in this paper is the Leukemia dataset, which has 38 acute leukemia samples and 50 genes. The purposes of the testing include clustering cell samples into groups and finding subclasses in the dataset.

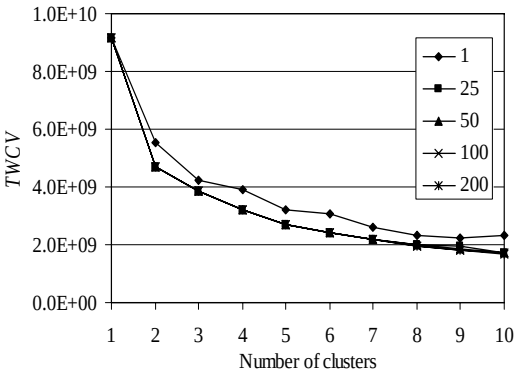

Figure 3 Pareto-fronts for Leukaemia dataset

The proposed genetic algorithm-based approach has been run for the Leukemia dataset with the following parameters: population size = 100, t_{dom} (number of comparison set = 10) and crossover = 0.8, mutation = 0.005, gene mutation rate = 0.005, and threshold = 0.01 which is used to check if the population stops evolution for 50 generations and if the process needs to be stopped. The range of [1, 10] was picked for finding the optimal number of clusters.

Changes in the Pareto-optimal front are displayed in Figure 3. The validity analysis results for the Leukaemia dataset are consistent with the literature where it is indicate that 2 (AML and ALL) is the best number of clusters; this two clusters as the best results has been concurrently reported by Dunn index, DB index, SD index, C index and Silhouette and 3 (AML, B-cell ALL and T-cell ALL) has been reported the second best. Y analyzing the results from the validity indices further, we discovered that S_Dbw is an exception; it is not suitable to test small datasets with fewer than 40 instances.

5 Summary and Conclusions

In this paper, we proposed a new clustering approach which depends on MOKGA as a multi-objective genetic algorithm based clustering approach. MOKGA is a combination of the Niched-Pareto optimal and fast k-means genetic algorithm. This way, we overcome the difficulty of determining the weight of each objective function taking part in the fitness. Otherwise, the user would have been expected to do many trials with different weighting of objectives as in traditional genetic algorithms. By using MOKGA, we aim at finding the pareto-optimal front so that the user will be able to see at once all possible alternative solutions identified by the system; then cluster validity index values are evaluated for each pareto-optimal front value which is the number of clusters value that is considered to be optimal. Then the solutions are all collapsed into a single pool of clusters which are individually evaluated to identify the most compact clusters to form the final solution. Comparing the clusters in final solution produced by the proposed clustering approach with the ones in the best clustering solution reported by the validity analysis, we realized that the former clusters are all compact and well separated while compactness of the latter clusters vary as well as their separateness. To validate the propose approach better, we still need to run more tests for data from different domains and with different characteristics. The outcome from this research project has interesting characteristics and it is very essential for several applications.

The user is no more in need for expertise in the domain of the data to be clustered because number of clusters is not needed but determined by the system. The process does not suffer from local minima kind of drawbacks because it leads to the most natural distribution of the data instances into the clusters leading to the most compact and separable clusters. On the other hand, the produced result will benefit another project in our group, namely the identification of biomarker genes. In a previous work, we tried multilevel clustering approach to identify best representative genes; however, we will have a more robust and consistent process by considering the new clusters. The outcome from this extension might be a good method to validate the previous results.

References:

[1] O. Abul, R. Alhajj, and F. Polat, "Powerful Approach for Effective Finding of Significantly Differentially Expressed Genes," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol.3, No.3, 2006, pp.220–231.

[2] M. Alshalalfa, R. Alhajj and J. Rokne, "Combining Singular Value Decomposition and t-test into Hybrid Approach for Significant Gene Extraction from Microarray Data," *Proceedings of IEEE BIBE*, Oct. 2008.

[3] M. Alshalalfa and R. Alhajj, "Motif Location Prediction by Divide and Conquer," *Proceedings of the International Conference on Bioinformatics Research and Development*, July 2008.

[4] M. Alshalalfa and, R. Alhajj, "Cancer Class Prediction: Two Stage Clustering Approach to Identify Informative Genes," *Intelligent Data Analysis*, (*in press*).

[5] Y. Barash & N. Friedman. Context-specific Bayesian clustering for gene expression data. *Proceedings of RECOMB*, 2001, pp.12-21.

[6] A. Ben-Dor, R. Shamir, & z. Yakhini. Clustering gene expression patterns. *Journal of Computational Biology* 1999.

[7] S-M. Chen, Y-J. Horng and C-H. Lee, Fuzzy information retrieval based on multi-relationship fuzzy concepy networks. *Fuzzy Sets and Systems*, Vol.140, No.1, November 2003, pp.183–205.

[8] D.L. Davies and D.W. Bouldin, A cluster separation measure. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, No.1, 1979, pp.224-227.

[9] J. Dunn, Well separated clusters and optimal fuzzy partitions. *J. Cybernetics*, Vol.4, 1974, pp.95–104.

[10] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caliguiri, C.D. Bloomfield, and E.S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring: *Science, 286*, 1999, pp.531–537.

[11] J. Grabmeier, et al, Techniques of Cluster Algorithms in Data Mining, Kluwer Academic Publishers, *Data Mining and Knowledge Discovery*, Vol.6, 2003, pp.303–360.

[12] M. Halkidi, M. Vazirgiannis and I. Batistakis, Quality scheme assessment in the clustering process, *Proceedings of PKDD*, Lyon, France, 2000.

[13] M. Halkidi, M. Vazirgiannis, Clustering Validity Assessment: Finding the optimal partitioning of a data set, *Proceedings of IEEE ICDM*, California, November 2001.

[14] J. Horn, N. Nafpliotis, and D. E. Goldberg, A niched pareto genetic algorithm for multiobjective optimization. *Proceedings of IEEE Conference on Evolutionary Computation, IEEE World Congress on Computational Computation*, Vol.1, Piscataway, NJ. , pp.82–87, 1994.

[15] L. Hubert and J. Schultz Quadratic assignment as a general data-analysis strategy. British J*ournal of Mathematical and Statistical Psychology*, Vol.29, pp.190-241, 1976.

[16] A. K. Jain, et al, Data Clustering: A Review. *ACM Surveys*, Vol.31, No.3, 1999.

[17] Y. Lu, S. Lu, F. Fotouhi, Y. Deng, and S. Brown, FGKA: A Fast Genetic K-means Clustering Algorithm, *Proceedings of ACM Symposium on Applied Computing*, pp.162-163, Nicosia, Cyprus, 2004.

[18] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comp App. Math*, Vol.20, 1987, pp.53–65.

[19] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Academic Press, 1998.

[20] Gene Expression Data of the Genomic Resources, University of Stanford (Available at: http://genome-www.stanford.edu/serum/data.html).

[21] K. Tamura, et al, Necessary and Sufficient Conditions for Local and Global Non-Dominated Solutions in Decision Problems with Multi-objectives. *Journal of Optimization Theory and Applications*, 27, 509-523, 1979.

[22] V.R. Iyer, M. B. Eisen, D. T. Ross, G. Schuler, T. Moore, J. C. F. Lee, J. M. Trent, L. M. Staudt, J. Hudson Jr., M. S. Boguski, D. Lashkari, D. Shalon, D. Botstein, P. O. Brown, The transcriptional program in the response of human fibroblasts to serum: *Science*, 283(5398):83-7, 1999.

[23] M. Khabbaz, K. Kianmehr, M. Alshalalfa and, R. Alhajj, "Comparing the Power of Apriori and FP-Growth for Building Adaptable Fuzzy Classifier," *International Journal of Data Warehousing and Mining*, (in press).

[24] T. Özyer and R. Alhajj, "Parallel Clustering of High Dimensional Data by Integrating Multi-Objective Genetic Algorithm with Divide and Conquer," *Applied Intelligence*, (in press)

[25] U. Scherf, D.T. Ross, M. Waltham, L.H. Smith, J.K. Lee, L. Tanabe, K.W. Kohn, W.C. Reinhold, T.G. Myers, D.T. Andrews, D.A. Scudiero, M.B. Eisen, E.A. Sausville, Y. Pommier, D. Botstein, P.O. Brown and J.N.Weinstein. A Gene Expression Database for the Molecular Pharmacology of Cancer: *Nat Genet* 24, 2000, pp.236–44.

[26] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, and T.R. Golub. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Nat'l Acad Sci USA*, 96, 1999, pp.2907–2912.

[27] M. Tan, M. Alshalalfa, F. Polat and R. Alhajj, "Combining Multiple Types of Biological Data in Constraint-Based Learning of Gene Regulatory Networks," *Proceedings of IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, Sep. 2008.

[28] K.Y. Yeung, C. Fraley, A.Murua, A.E. Raftery, and W.L. Ruzzo, Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17, 2001, pp.977–987.

[29] E. Zitzler, Evolutionary algorithms for multiobjective optimization: Methods and applications. *Doctoral thesis ETH NO. 13398, Zurich: Swiss Federal Institute of Technology (ETH)*, Aachen, Germany: Shaker Verlag, 1999.