

Cluster Validity Analysis of Alternative Results from Multi-Objective Optimization

Yimin Liu Tansel Özyer Reda Alhajj Ken Barker

Department of Computer Science
University of Calgary
Calgary, Alberta, Canada
{liuyi, ozyer, alhajj, barker}@cpsc.ucalgary.ca

Abstract

This paper investigates validity analysis of alternative clustering results obtained using the algorithm named Multi-objective K-Means Genetic Algorithm (MOKGA). The reported results are promising. MOKGA gives the optimal number of clusters as a solution set. The achieved clustering results are then analyzed and validated under several cluster validity techniques proposed in the literature. The optimal clusters are ranked for each validity index. The approach is tested by conducting experiments using three well-known data sets. The obtained results for each dataset are compared with those reported in the literature to demonstrate the applicability and effectiveness of the proposed approach.

Keywords: clustering, gene expression data, genetic algorithms, multi-objective optimization, validity analysis.

1. Introduction

Traditional clustering algorithms, in general, do not produce alternative solutions, and most of them do not lead to the optimal number of clusters in the dataset that they work on. For example, hierarchical clustering method can get the heuristic overview of a whole dataset, but it cannot relocate objects that may have been 'incorrectly' grouped at an early stage. It cannot tell the optimal number of clusters nor give the non-dominated set. *K*-means needs the number of clusters as a predefined parameter, and it may give local optimal solutions because it is a local search from a random initial partitioning. SOM has the same disadvantage in that it requires the number of clusters be given *a priori*. Clearly, a clustering algorithm is needed to get the global pareto optimal solution set required to give users the best overview of the whole dataset according to the number of clusters and their quality. Further, it is required to get clustering results with the optimal number of clusters.

The main contribution of this paper is a new clustering approach that considers multiple objectives in the process and its application for clustering microarray and other datasets; we have tested our approach on three data sets. The proposed approach has two components. 1) Multi-objective *K*-means Genetic Algorithm (MOKGA) based clustering approach, which delivers a pareto optimal clustering solution set without taking weight values into account. Otherwise, users need to consider several trials weighting with different values until a satisfactory result is obtained. 2) Cluster validity analysis employed to evaluate the obtained

candidate optimal number of clusters, by applying some of the well-known cluster validity techniques, namely Silhouette, *C* index, Dunn's index, DB index, SD index and S-Dbw index, to the clustering results obtained from MOKGA. It provides one or more options for the optimal number of clusters.

The applicability and effectiveness of the proposed clustering approach and clustering validity analysis process are demonstrated by conducting experiments using three datasets: namely Fig2data, cancer (NCI60), and Leukaemia data sets available at Genomics Department of Stanford University, UCI machine learning repository.

The balance of the paper is organized as follows. Section 2 is devoted to the development of the new clustering system MOKGA. Section 3 reports experimental results to illustrate the applicability, performance and effectiveness of the system. Section 4 discusses advantages of the proposed approach in comparison with other existing methods. Section 5 is summary and conclusions.

2. The Proposed Clustering Approach

The proposed clustering approach named Multi-Objective Genetic *K*-means algorithm (MOKGA) has been developed on the basis of the Fast Genetic *K*-means Algorithm (FGKA) [8] and the Niche Pareto Genetic Algorithm [5].

After running the multi-objective *K*-means genetic algorithm, the Pareto-optimal front giving the optimal number of clusters as a solution set can be obtained. The system then analyzes the clustering results found with respect to six cluster validity techniques proposed and well documented in the literature, namely Silhouette, *C* index, Dunn's index, SD index, DB index, and S-Dbw index.

MOKGA uses a list of parameters to drive the evaluation procedure as in other genetic types of algorithms: including population size (the number of chromosomes), t_{dom} (the number of comparison set) representing the assumed non-dominated set, crossover, mutation probability, and the number of iterations for the execution of the algorithm to obtain the result. Subgoals can be defined as fitness functions, and instead of scalarizing them to find the goal as the overall fitness function with the user defined weight values, it is expected that the system can 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.

Initially, the *current generation* is assigned to zero. Each chromosome takes the *number of clusters* parameter within the range 1 to the maximum number of clusters given by the user. A population with the specified number of chromosomes is created randomly by using the method described by Rousseeuw [11], where data points are randomly assigned to each cluster at the beginning and the rest of the points are randomly assigned to clusters. By using this method, we can avoid generating illegal strings, which means some clusters do not have any pattern in the string.

Using the current population, the next population is generated and the generation number is incremented by 1. During the next generation, the current population performs the Pareto domination tournament to get rid of the worst solutions from the population. Crossover, mutation, and the k-means operator [8] are then performed to reorganize each object's assigned cluster number. Finally, we will have twice the number of individuals after the Pareto domination tournament. The ranking mechanism used by Zitzler in [2] is applied to satisfy the elitism and diversity preservation. This halves the number of individuals.

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 chromosome in the population. In the selection part, t_{dom} individuals are randomly picked from the population. Two chromosome candidates are randomly selected from the current population except those in the comparison set (*population size- t_{dom}*), and each of the candidates is compared against each individual in the comparison set t_{dom} . If one candidate has larger total within-cluster variation fitness and larger number of cluster values than all the chromosomes in the comparison set, then it is dominated by the comparison set and will be deleted from the population permanently. Otherwise, it resides in the population.

After the Pareto domination tournament, the dominated chromosome is deleted from the population. The next step is crossover: one point crossover is used in the employed multi-objective genetic clustering approach. An index into the chromosome is selected and all data beyond that point in the chromosome are swapped between the two parent chromosomes. The resulting chromosomes are the children.

Mutation is applied to the population in the next step by randomly changing the values in the chromosome according to probability distribution.

The *K*-means operator is applied last to reanalyze each chromosome gene's assigned cluster value. It calculates the cluster centre for each cluster and re-assigns each gene to the closest cluster to each instance in the gene. Hence, *K*-means operator is used to speed up the convergence process by replacing a_n by a_n' , for $n=1$ to N simultaneously, where a_n' is the closest to object X_n in Euclidean distance.

After all operators have been applied, twice the number of individuals remains. After having the Pareto dominated tournament, we cannot give an exact number equal to the initial population size because at each generation randomly picked candidates are selected for the survival test leading to

the deletion of one or both, in case dominated. To half the number of individuals, the ranking mechanism proposed by Zitzler [2] is employed: individuals obtained after crossover, mutation, and *K*-means operator are ranked; and the best individuals are picked for population of the next generation.

The approach picks the first l individuals by considering the elitism and diversity among $2l$ individuals. Pareto fronts are ranked. Basically, we find the Pareto-optimal front and remove individuals of the Pareto-optimal front from the $2l$ set and place them in the population to run in the next generation. In the remaining sets, we get the first Pareto-optimal front and put it in the population and so on. Since we try to get the first l individuals, the last Pareto-optimal front may have more individuals required to complete the number of individuals to l . We handle the diversity automatically. We rank them and reduce the objective dimension into one. We then sum the normalized value of the objective functions for each individual. These are sorted in increasing order and each individual's total difference from its individual pairs is calculated. The individuals are placed in population based on decreasing differences, and then we keep placing from the top as many individuals as we need to complete the number of individuals in the population to l . The reason for doing this is to take the crowding factor into account automatically so that individuals occurring closer to others are unlikely to be picked.

This method was also suggested as a solution for the elitism and diversity for improvement in NSGA-II. For example, in order to get 20 chromosomes from the population, we select 10 chromosomes from the Pareto front, delete them from the current population, then get 8 chromosomes from the Pareto front in the current population, delete them from the population. Suppose that we have 6 in the current population, we take 2 chromosomes that have the largest distance to their neighbours using the ranking method mentioned above. Finally, if the maximum number of generations is reached, or the Pareto front remains stable for 50 generations, then the process is terminated; otherwise the next generation is performed.

3. Experimental Results

To evaluate the performance and efficiency of the proposed system consisting of the MOKGA clustering approach and conduct cluster validity analysis on the obtained alternative results, experiments were conducted on a computer with the following features: Pentium ®4 with 2.00 GHz CPU, 512 MB RAM and running Windows XP. The system was implemented using MS Visual C++. The running platform is Microsoft Visual Studio.NET 2003.

Three widely gene expression datasets, namely Fig2data, cancer (NCI60), and Leukaemia have been used to test the performance and accuracy of the system. Fig2data data is used for clustering genes, while cancer (NCI60) and Leukaemia data sets are used for group cell samples.

3.1 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 ranging from 15 minutes to 24 hours [1, 6].

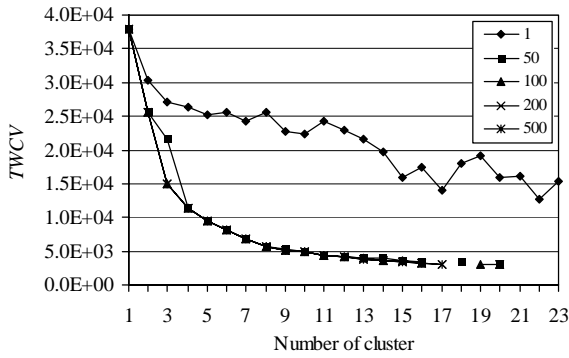


Figure 3.1 Pareto-fronts for Fig2data dataset

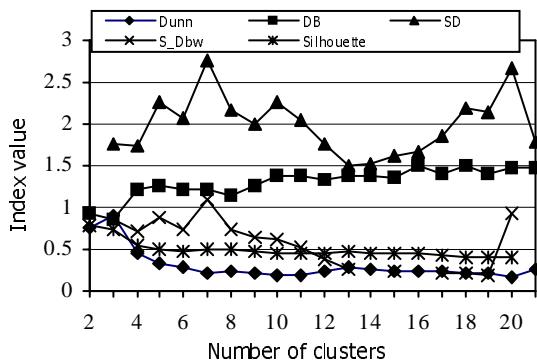


Figure 3.2 Fig2data dataset cluster validity results using Dunn, DB, SD, S_Dbw and Silhouette indexes



Figure 3.3 Fig2data dataset cluster validity results using C index

Lu *et al* [8] applied the Fast Genetic K-means Algorithm to Fig2data. They selected as their parameter setting: mutation probability = 0.01, population size = 50, and generation = 100. As a result, they obtained fast clustering process.

In our tests, MOKGA has been applied to Fig2data dataset. Experiments were conducted 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 corresponding experimental results are reported

in Figure 3.1. They also show how the system converges to a Pareto optimal front.

Figure 3.2 and Figure 3.3 report validity results and reflect comparisons with the studies described elsewhere [6, 8]. The study by Iyer *et al* [6] show that the optimal number of clusters for Fig2data is 10. Consistently, results in this paper indicate that it ranks among the best ones for C index, and the number of 10 clusters is among the best for other indices as well. According to Maria *et al* [4], SD, S_Dbw, DB, Silhouette, and Dunn indices cannot handle properly arbitrarily shaped clusters, so they do not always give satisfactory results.

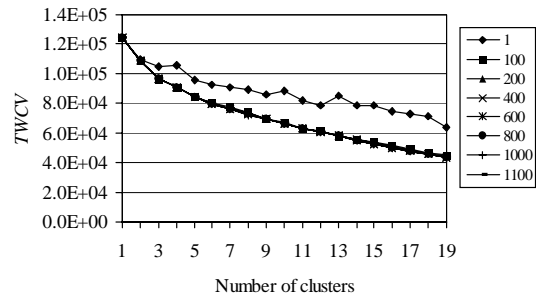


Figure 3.4 Pareto-fronts for Cancer dataset

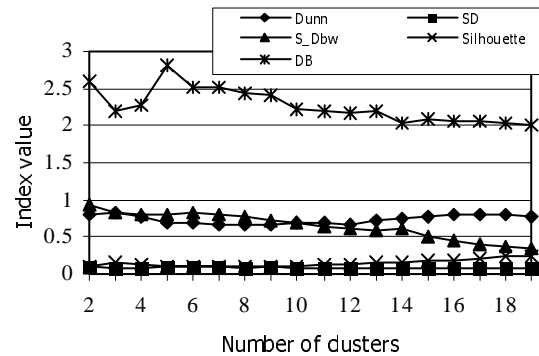


Figure 3.5 Cancer dataset cluster validity results using Dunn, DB, SD, S_Dbw and Silhouette indexes

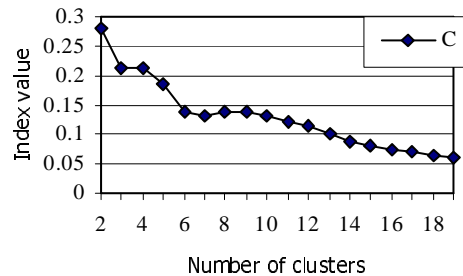


Figure 3.6 Cancer dataset cluster validity results using C index

3.2 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, leukaemias and melanomas. 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 [13].

The study by Scherf [13] uses an average-linkage algorithm and a metric based on the growth inhibitory activities of the 1,400 compounds for the cancer dataset. The authors observed 15 distinct branches at an average inter-cluster correlation coefficient of at least 0.3. In this method, the correlation parameter was used to control the clustering results. It might be hard to decide if it is an unsupervised clustering task.

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 if the population stops evolution for 50 generations and if 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 3.4. It demonstrates convergence to an optimal Pareto-optimal front.

Figures 3.5 and Figure 3.6 show the average results obtained. For the cancer (NCI60) dataset, we have 15 in the Pareto optimal front; 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 [13]. Having some indexes values not good demonstrates the fact that index values are highly dependent on the shape of the clusters.

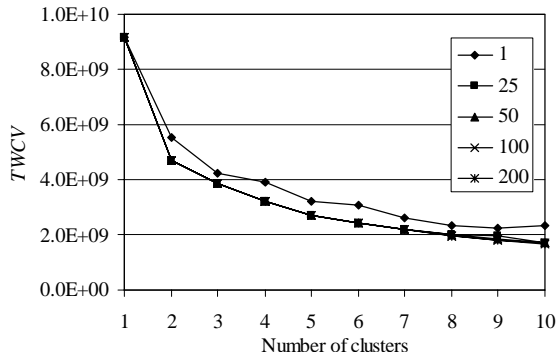


Figure 3.7. Pareto-fronts for Leukaemia dataset

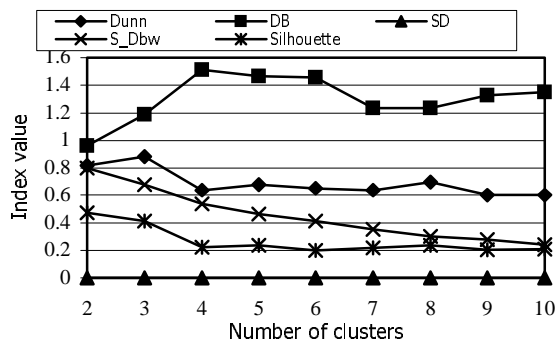


Figure 3.8 Leukemia dataset cluster validity results using Dunn, DB, SD, S_Dbw and Silhouette indexes

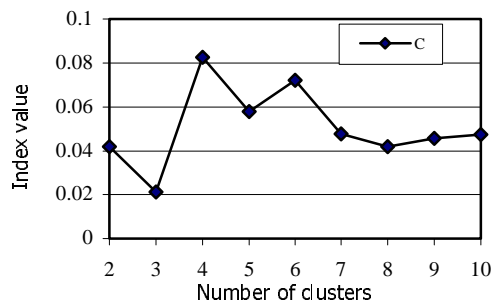


Figure 3.9 Leukemia dataset cluster validity results using C index

3.3 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.

The study by Golub *et al* [3] uses Self-Organizing Maps (SOMs) to group the Leukemia dataset. In this approach, the user specifies the number of clusters to be identified. SOM finds an optimal set of centroids around which the data points appear to aggregate. It then partitions the data set with each centroid defining a cluster consisting of the data points nearest to it. Golub [3] got two clusters acute myeloid leukemia (AML) and acute lymphoblastic leukaemia (ALL), as well as the distinction between B-cell and T-cell ALL, which means that the optimal number of clusters is 2 or 3 (with subclasses).

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 3.7. It demonstrates how the system converges to an optimal Pareto-optimal front.

The Leukaemia dataset clustering results shown in Figure 3.8 and Figure 3.9 indicate the same conclusions reported in [3] by Golub *et al*. They also indicate that 2 (AML and ALL) is the best number of clusters after the validity analysis with Dunn index, DB index, SD index, and Silhouette and 3 (AML, B-cell ALL and T-cell ALL) is the second best. C index shows that 2 is the best cluster number and 3 is the second. It can be seen from Figure 3.8 that S_Dbw is an exception. SD index gives good values, but S_Dbw does not. This indicates that the inter-cluster density for number of clusters taken 2 and 3 is not high for the 38 samples. Experimental results in this paper also indicate that S_Dbw index is not suitable to test small datasets with fewer than 40 instances.

4. General Evaluation and Comparison

In this section, the MOKGA system is compared with other methods on basis of the results obtained for the same datasets. For instance, according to [6], Fig2data has 10

clusters. The proposed approach gave the same result using C index clustering validity method. Cancer data has 15 clusters according to the result in [13]. MOKGA produces the same result using the DB index. The optimal number of clusters of Leukemia dataset is 2 or 3. MOKGA reported the same results using Dunn, DB, SD, and Silhouette indexes.

Since MOKGA has been developed on the basis of Fast Genetic K-mean Algorithm (FGKA) [8] and Niche Pareto Genetic Algorithm (NPGA), MOKGA and FGKA share many features: both are evolutionary algorithms; they have the same mutation and K-mean operators; and they both use the Total Within-Cluster Variation (TWCV) for the fitness value evaluation.

According to the results, MOKGA and FGKA got similar TWCV values, MOKGA obviously need more generations to get the stable state, this might be because MOKGA is optimizing chromosomes with different number of clusters altogether.

MOKGA has some advantages over FGKA and GKA: it can find the Pareto optimal front, which allows us to get an overview of the entire clustering possibilities and to get the optimal clustering results in one run; it does not need the number of clusters as a parameter, which is very important because clustering is an unsupervised task, and we usually do not have any idea about the number of clusters before the clustering of gene expression data. These two issues are real concerns for FGKA, GKA and most of the other clustering algorithms.

Both MOKGA and K-means Algorithm minimize the overall within-cluster dispersion by iterative reallocation of cluster members. MOKGA has some advantages over K-means algorithm: it can find the Pareto optimal front; it does not need the number of clusters as a parameter; MOKGA can find global optimal solutions using mutation and crossover operators. MOKGA combines both the advantages of genetic algorithm and advantages of the K-means algorithm: by using GA operators it can get global optimal solutions, and by using K-means operators MOKGA can get solutions much faster.

5. Summary and Conclusions

The MOKGA approach proposed in this paper has been developed on the basis of the Niche Pareto optimal and fast K-means genetic algorithm. By using MOKGA, it is aimed at finding the Pareto-optimal front sought to help the user to obtain several alternative solutions at once. Then, cluster validity index values are evaluated for each Pareto-optimal front value, which is considered the optimal number of clusters value. MOKGA overcomes the difficulty of determining the weight of each objective function taking part in the fitness when dealing with this multiple objectives problem. Otherwise, the user would have been expected to do many trials with different weighting of objectives as in traditional genetic algorithms. This method also gives the users an overview of different numbers of clusters, which may help them in finding subclasses and optimal number of clusters in a single run, whereas traditional methods like SOM, K-means, Hierarchical clustering algorithms and GCA can not find optimal number of clusters, or need it as a

prespecified parameter. MOKGA is less susceptible to the shape or continuity of the Pareto front. It can easily deal with discontinuous or concave Pareto fronts. These two issues are real concerns for mathematical programming techniques, like model-based approaches such as Bayesian method and mixed model-based clustering algorithms.

References

- [1] K. Chen, L. Liu, "Validating and Refining Clusters via Visual Rendering Gene Expression Data of the Genomic Resources," *Proc. of IEEE International Conference on Data Mining*, pp.501-504, 2003.
- [2] E. Zitzler, "Evolutionary algorithms for multiobjective optimization: Methods and applications," *Doctoral Thesis* ETH NO. 13398, Zurich: Swiss Federal Institute of Technology, 1999.
- [3] T. R. Golub, et al, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, 286, pp.531-537, 1999.
- [4] M. Halkidi, Y. Batistakis and M. Vazirgiannis, "Clustering Validity Checking Methods: Part II," *SIGMOD Record*, Vol.31, No.3, pp.19-27, 2002.
- [5] J. Horn, N. Nafpliotis and D. E. Goldberg, "A Niche Pareto Genetic Algorithm for Multiobjective Optimization," *Proc. of IEEE CEC*, Vol.1, pp.82-87, Piscataway, NJ. 1994.
- [6] V.R. Iyer, et al, "The transcriptional program in the response of human fibroblasts to serum," *Science*, 283(5398), pp.83-7, 1999.
- [7] Y. Liu, T. Özyer, R. Alhajj and K. Barker, "Multi-objective Genetic Algorithm based Clustering Approach and Its Application to Gene Expression Data," *Proc. of ADVIS*, Springer-Verlag, Oct. 2004.
- [8] Y. Lu, et al, "FGKA: A Fast Genetic K-means Clustering Algorithm," *Proc. of ACM Symposium on Applied Computing*, Cyprus, pp.162-163, 2004.
- [9] U. Möller, D. Radke, F. Thies, Testing the significance of clusters found in gene expression data. *Proc. of European Conference on Computational Biology*, Paris, pp.26-30, 2003.
- [10] T. Özyer, Y. Liu, R. Alhajj and K. Barker, "Validity Analysis of Clustering Obtained Using Multi-Objective Genetic Algorithm," *Proc. of ISDA*, Springer-Verlag, Hungary, Aug. 2004.
- [11] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of Comp App. Math*, Vol.20, pp.53-65, 1987.
- [12] E. H. Ruspini, "Numerical methods for fuzzy clustering," *Inform. Science*, vol.2, pp.319-350. 1970.
- [13] U. Scherf, et al, "A Gene Expression Database for the Molecular Pharmacology of Cancer," *Nat Genet*, Vol.24, pp.236-44, 2000.
- [14] B. Stein, S. Meyer and F. Wissbrock, "On Cluster Validity and the Information Need of Users," *Proc. of the International Conference on Artificial Intelligence and Applications*, Spain, Sep. 2003.
- [15] W. Shannon, R. Culverhouse J. Duncan, "Analyzing microarray data using cluster analysis," *Pharmacogenomics*, Vol.4, No.1, pp.41-52, 2003.