Cluster Ensemble and Its Applications in Gene Expression Analysis

Xiaohua Hu, Illhoi Yoo

College of Information Science and Technology
Drexel University, Philadelphia, PA 19104

thu@cis.drexel.edu

Abstract

Huge amount of gene expression data have been generated as a result of the human genomic project. Clustering has been used extensively in mining these gene expression data to find important genetic and biological information. Obtaining high quality clustering results is very challenging because of the inconsistency of the results of different clustering algorithms and noise in the gene expression data. Many clustering algorithms are available and different clustering algorithms may generate different clustering results due to their bias and assumptions. It is a challenging and daunting task for the genomic researchers to choose the best clustering algorithm and generate the best clustering results for their data sets. In this paper, we present a cluster ensemble framework for gene expression analysis to generate high quality and robust clustering results. In our framework, the clustering results of individual clustering algorithm are converted into a distance matrix, these distance matrices are combined and a weighted graph is constructed according to the combined matrix. Then a graph partitioning approach is used to cluster the graph to generate the final clusters. The experiment results indicate that cluster ensemble approach yields better clustering results than the single best clustering algorithm on both synthetic data set and yeast gene expression data set.

Keywords: cluster ensemble, gene expression analysis, graph partition

1 Introduction

Clustering is to group analogous elements in a data set in accordance with its similarity. Therefore "good" clustering means elements in each cluster are similar while elements from different clusters are dissimilar. Unlike classification, clustering does not require the class label information about data set because it is inherently a data-driven approach; that is why clustering plays a very important role as the initial step in data exploratory analysis. In gene expression analysis, normally not much prior knowledge is accumulated, so genomic researcher tend to apply clustering algorithms on the gene expression data sets to gain better understanding and insightful genetic and biological information. Clustering is one of most widely and frequently used data mining technologies in gene expression analysis (Azuaje 2002, Bellaachia, Portnoy, Chen and Elkahloun 2002, Ben-Dor and Yakhini 1999, Berrar, Dubitzky and Granzow 2002, Bloch and Arce 2002, Xing and Karp 2001, Zeng, Tang, Garcia-Frias and Gao 2002, Zhao and Karypis 2003).

Through the use of clustering algorithms on gene expression data can answer some challenging biological and genetic questions, such as identifying the functionality of genes, finding out what genes are co-regulated, distinguishing the important genes between abnormal tissue and normal tissues etc (Zhao and Karypis 2003). There are multiple clustering techniques that can be used to analyse gene expression data. Advantages and limitations may depend on factors such as data distribution, pre-processing procedures, number of genes etc. Choosing "the best" algorithm for a particular problem may represent a challenging task. Moreover, it is not uncommon to observe inconsistent results when different clustering methods are tested on a particular data set. K-Means, Self-Organizing Map (SOM), Hierarchical approaches, Fuzzy C-Means, etc, are very different in some cases (Jain, Murty and Flynn 1999). This is because clustering methods have their own bias and function criterion. For example, the popular K-means algorithm performs miserably in several situations where the data cannot be accurately characterized by a mixture of K Gaussian with identical covariance matrices. It is well known that no single clustering algorithm that performs best across various data sets and it is very challenging to choose the best clustering algorithm for gene expression analysis.

Design, performance evaluation, and application of clustering algorithm on gene expression data must take into account the data characteristics and randomness arising from both biological and experimental variability. Instead of focusing on developing single clustering algorithm only work for a narrow-range of data sets, in this paper, we take a different approach. We present a unified cluster ensemble framework to combine the clustering results from various clustering algorithms. In our approach, the clusters ensemble problem is converted to a graph partitioning problem. A distance matrix is first constructed based on the cluster results from each individual clustering algorithm; these distance matrices are combined to form a master distance matrix. Then a weighted graph is constructed from the master distance matrix and a graph-based partitioning algorithm is applied to the graph for the final clustering results. The cluster ensemble builds a robust clustering portfolio that can perform reasonable well over a wide range of data sets with little hand-tuning.

Our experiment results on both synthetic data sets as well as gene expression data sets indicate that the clustering quality of the ensemble approach significantly outperforms the best individual clustering algorithm.

The rest of the paper is organized as follows. In Section 2, we give a brief overview of the various clustering approaches and summarize the related work. We present our cluster ensemble algorithm in details and experimental tests in Section 3. We conclude with our future plan and discussion in Section 4.

2 Related Work

Classification ensemble approaches such as bagging and boosting have been proved very popular and effective in supervised learning to improve the learning accuracy (Dietterich 2001, Hu 2001). Following the same philosophy, the goal of cluster ensemble is to combine the clustering results of multiple clustering algorithms to obtain better quality and robust clustering results. Generating high quality clustering result is very challenging in gene expression analysis because of the noise in the experimental data and the inconsistency among the different clustering algorithms. Even though many clustering algorithms have been developed (Han and Kamber 2001, Hartigan 1975, Jain, Murty and Flynn 1999), not much work is done in cluster ensemble in data mining and machine learning literature compared with classification ensemble method. Zeng et al. (Zeng, Tang, Garcia-Frias and Gao 2002) proposed an adaptive meta-clustering approach for combining different clustering results. In their research, they converted the individual cluster results into a distance matrix and then combine the distance matrix and apply a hierarchical clustering to recluster the combined distance matrix. Strethl et al. (Strehl and Ghosh 2002) proposed a hypergraph-partitioned approach to combine different clustering results. Each cluster in an individual clustering algorithm is treated as a hyperedge. This crisp hypergraph lost much useful information, and it is not suitable for ambiguous and noisy environment.

It is very hard to find the optimal way of combining clusters. It is considered that this is a natural phenomenon because each object has various characteristics and a group of various objects can be partitioned in several ways based on the many peculiarities. For example, consider people in a university. We can group them into so many groups based on gender, nationality, position (faculty, staff, student – we can also group faculty and student in fine grade; for faculty group there are full professor, associate professor, assistant professor, etc and for students there are graduate student and undergraduate student), etc. It is hard to say which clustering result is the best.

We observe even though various clustering algorithms present different types of knowledge concerning the clustering criterion, most clustering criteria in various algorithms are compensative rather than competitive in gene expression analysis. We believe that an effective combination of several clustering algorithms is an important way to improve the clustering quality, but cluster ensemble is different from the classification ensemble. Some of the major issues of cluster ensembles addressed in the proposed research are how to combine different clustering results and how to ensure symmetrical and unbiased consensus with regard to all the component partitions. The main difficulties are: (1) the quality of a clustering combination algorithm cannot be evaluated as precisely as a combining classifier, and (2) various clustering algorithms always produce results with large differences due to different clustering criteria. Directly combining the clustering results with integration rules such as product, sum and majority vote cannot generate a good meaningful result. A new mechanism to combine the different cluster results is needed to obtain better clustering results.

We propose a graph-based meta-clustering approach to extract the information from results of different clustering techniques, so a better interpretation of the data distribution can be obtained. A distance matrix is constructed to represent the statistical information of each cluster produced by various clustering techniques. Our method incorporates multiple cluster-based distance matrices into a weighted graph. A graph based clustering algorithm is used to cluster the graph for the final clustering results.

3 Cluster Ensemble Based on Similarity-Graph

The motivations for developing cluster ensembles are to improve the quality and robustness of results. There are two reasons for this: (1) the results of clustering are easily corrupted by the addition of noise, which is very common in gene expression analysis as the experimental measurement may not be very accurate or error may be introduced by the data transformation, (2) the clustering results of different clustering methods can vary significantly in the same data set, that indicates that there could be a great potential for improvement when using an ensemble for the purpose of improving clustering quality.

The purpose of cluster ensemble is to build a robust clustering portfolio that can perform as good as if not better than the single best clustering algorithm across a wide-range of data sets. Different clustering algorithm may take a different approach. For example, K-means is to group the data set so that the total Mean Square Error to the center of each cluster is minimum while graph-based partitioning clustering is to partition the graph into K parts based on the minimum edge weight cuts. Thus a cluster ensemble can be used to generate many cluster results using various clustering algorithms and then integrate them using a consensus function to yield stable results.

In this section we discuss our novel cluster ensemble approach to combine the clustering results from various clustering algorithms. We present a two-phase clustering combination strategy. At the first step, various clustering algorithms are run against the same data sets to generate clustering results. At the second step, these clustering results are combined by an auto-associative additive system based on the distance matrix of graph clustering. The diagram below summarizes our approach.

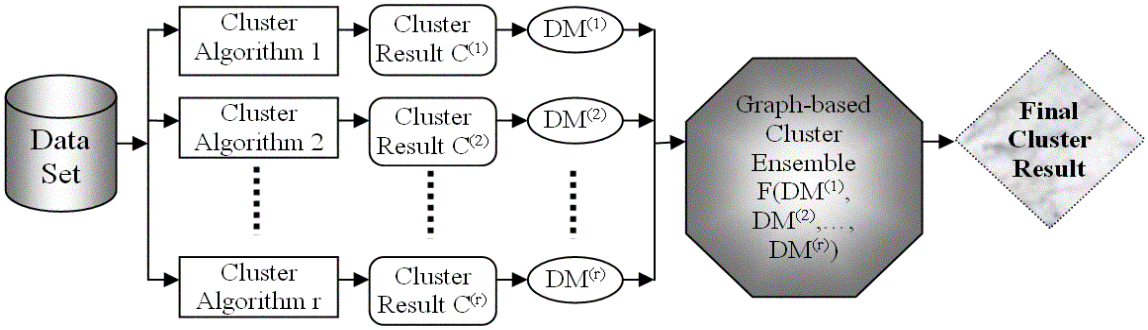

Figure 1 Cluster Ensemble Architecture

In our approach, a distance matrix is first constructed based on the cluster results from each individual clustering algorithm; these distance matrices are combined to form a master distance matrix. Then a weighted graph is constructed from the master distance matrix and a graph-based partitioning algorithm is applied to the graph for the final clustering results. Graph-based clustering uses various kinds of geometric structure or graphs for analyzing data. Different graphs reflect various local structure or inherent visual characteristic in the data set. Clustering divides the graph into connected components by identifying and deleting inconsistent edges, and each subgraph consisting of connected components refers to a cluster.

In the subsections below, we explain our cluster ensemble step by step. We first discuss the cluster validation indices, which help answer the tough question of how many clusters in the data sets, brief describe the clustering methods integrated in our framework, and then we explain the cluster ensemble mechanism and clustering result evaluation.

3.1 Clustering Methods

Many clustering algorithms have been developed from computer science and other disciplines such as data mining, machine learning, pattern recognition and statistics, to name a few. Clustering algorithms can be roughly classified into hierarchical methods and non-hierarchical methods. Non-hierarchical method can also be divided into four categories; partitioning methods, density-based methods, grid-based methods, and model-based methods (Han and Kamber 2001, Jain, Murty and Flynn 1999).

Hierarchical methods proceed successively by either merging the smaller clusters into large clusters or splitting the larger clusters. The methods yield a dendrogram or a tree of clusters representing how the clusters are related. Partitioning methods generate initial k clusters and improve the clusters by iteratively reassigning elements among k clusters. The number of "k" and iteration is user input. K-means and K-medoids (Partitioning Around Medoids (PAM) and Clustering LARge Applications (CLARA)) (Jain, Murty and Flynn 1999) belong to this category. Self-Organizing Map (SOM) as a model-based method was developed for better speech recognition by Teuvo Kohonen in the early 1980s (Kohonen 2000). Fuzzy C-means as one of fuzzy clustering methods has been developed by Bezdek (Bezdek 1981, Bezdek and Pal 1992) by generalizing Dunn's idea (Dunn 1974).

In our experiment we integrated three clustering algorithms; K-means, Self-Organizing Map (SOM), and Fuzzy C-means as our initial implementation and more complementary clustering algorithms can be added without any changes to the architecture of the ensemble framework.

3.2 Clustering Ensemble Algorithm

Based on multi-objective programming: a simple strategy of designing clustering ensemble algorithm is based on multi-objective programming that seeks a solution to satisfy multiple clustering criteria. Multi-objective programming can be transformed to single objective programming by a weighting method, which is employed in our algorithm.

Our algorithm to ensemble clusters is.

Algorithm 1: Cluster Ensemble Based on Similarity-Graph (**CESG**)

Input: (1) the data set $X=\{x_1,x_2,x_3,\ldots,x_n\}$, (2) the upper bound of the cluster number **k**, (3) edge threshold value δ (4) a set of clustering algorithms $\mathbf{C^{(q)}}$

Output: the final clustering result $C^{(opt)}$

Method:

Step 1: Run the individual clustering algorithm $\mathbf{C^{(q)}}$ multiple times on the same data set under different cluster numbers (clusters varies from 2 to k).

Step 2: Pick up the optimal number of clusters for each data set using three cluster validation indices (Silhouette index, Dunn index, and C index). If the number is not consistent, use voting strategy to choose the number with the majority as the number of the clusters.

Step 3: Construct a distance matrix (DM) for the clustering results for each clustering algorithm. (DM_{ij} represents the similarity of two data x_i and x_j points)

Step 4: Combine the distance matrixes by adding them into one master distance matrix (MDM)

Step 5: Construct a weighted graph based on the distance matrix. (There is an edge between data point x_i and x_j if the value MDM_{ij} of x_i and x_j is greater than

some threshold value δ, MDM_{ij} is also the weight of the edge link x_i and x_j)

Step 6: Cluster the graph into the optimal number of clusters based on the cluster number chosen at Step 2

End

In Step 3, there are so many ways to construct the distance matrix based on cluster results from individual clustering algorithm. We propose a solution based on statistical theory. Here we assume that our data set is in Gaussian distribution as in (Zhao and Karypis 2003).

Cluster-based distance matrix $DM^{(q)}$ for the clustering result $C^{(q)}$. $DM^{(q)}$ is a pair-wise distance matrix defined between two data points according to the clustering result. This distance is able to efficiently extract the statistical information from the obtained cluster structure. The matrix size is *nxn*. Since its size is independent of the clustering approach, it provides a way to align the different clusterings onto the same space even for some situations where the numbers of clusters are different for different clustering algorithms.

Assume that input data set are $X=\{x_1, x_2, x_3, \ldots x_n\}$, and the cluster algorithm generates m clusters for the data set X. Clustering result is $S=\{s_1, s_2, s_3, \ldots, s_m\}$, where s_j is the j^{th} clusters consisting of some data points in X. For example, $X=\{x_1, x_2, x_3, \ldots x_9\}$ and $S=\{s_1, s_2, s_3\}$, $s_1 = \{x_1, x_5\}$, $s_2=\{x_2, x_3, x_7\}$, $s_3=\{x_4, x_6, x_8, x_9\}$.

We assume that probability density function of s_j is given by $p(x_i|s_j)$, the posterior probability of cluster s_j given x_i can be expressed as:

$$P(s_j \mid x_i) = \frac{p(x_i \mid s_j) \times P(s_j)}{\sum_{k=1}^{m} p(x_i \mid s_k) \times P(s_k)} \quad , \quad \text{where}$$

$$P(x_i \mid s_j) = \frac{\exp[-\frac{1}{2}(x_i - \mu_j)^T \sum_j (x_i - \mu_j)]}{(2\pi)^{m/2} \| \sum_j \|^{1/2}},$$

m is the number of clusters. $\sum_j$ is a matrix of co-variances among attributes in cluster j, μ_j is the mean vector of the data points in the cluster s_j.

For example, to calculate $P(s_2|x_2)$, given the following elements in a cluster $s_2 =\{x_2, x_3, x_7\}$, (assume each x_i has three conditions),

	Att_1	Att_2	Att_3
x_2	x_{21}	x_{22}	x_{23}
x_3	x_{31}	x_{32}	x_{33}
x_7	x_{71}	x_{72}	x_{73}

$\sum_2$ is shown as

$$\sum_2 = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22}^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33}^2 \end{bmatrix}$$

where σ_{12} is covariance between Att_1 and Att_2 and σ_{11}^2 is variance of Att_1 in s_2

For each data point x_i, we calculate the corresponding probability vector $PX_i = \{P(s_1|x_i), P(s_2|x_i), \ldots P(s_m|x_i)\}$, where $\sum_{j=1,..m} P(s_j|x_i) =1$, the probability vectors form a probability space of dimension of *m*, with each dimension corresponding to one cluster. The probability space contains information from both the input data and the cluster results. So we believe the similarity of any two points PX_l and PX_m in the probability space is a good measurement to reflect the distance of the corresponding points x_l and x_m in the original space.

Then for any two points, x_l and x_m , in the data set, their distance is defined as the distance between PX_l and PX_m, namely, $\mathbf{DM^{(q)}}(x_l, x_m)$. Many different distance measures such as Euclidean distance, Mahalanobis distance or correlation distance can be used to calculate $DL(PX_l, PX_m)$.

We define the similarity of two points (x_i and x_j) in the data set as

$$DM_{ij}^{(q)} = 1 - \frac{\sum PX_i PX_j - \frac{\sum PX_i \sum PX_j}{N}}{\sqrt{\left(\sum PX_i^2 - \frac{(\sum PX_i)^2}{N}\right) \times \left(\sum PX_j^2 - \frac{(\sum PX_j)^2}{N}\right)}}$$

In step 6, a graph-based clustering algorithm is applied to the weighted graph for the final clustering result. Many graph-based partitioning algorithms can be used for this purpose, such METIS, HMETIS (Kayypis and Kumar 1998a). Clustering divides the graph into connected components by deleting edges based on some constraint such as minimum cuts, and each subgraph consisting of connected components refers to a cluster. In our experiment, we chose the graph partitioning-based algorithm METIS (Kayypis and Kumar 1998a, Kayypis and Kumar 1998b) because of its good performance and scalability.

3.3 Clustering Result Evaluation

To evaluate the quality of cluster is a non-trivial and often ill-posed task. Generally speaking, there are internal criteria and external criteria. Internal criteria formulate quality as a function of the given data and/or similarities. For example, the mean squared evaluation criterion (for k-means) and other measures of compactness are popular evaluation criteria. Measure can also be based on isolation such as the min-cut criterion, which uses the sum of edges weights across clusters (for graph portioning). When using internal criteria, clustering becomes an optimization problem, and a clusterer can evaluate its own performance and tune its results accordingly.

External criteria on the other hand impose quality by additional, external information not given to the clusteres, such as category labels. This is sometimes more appropriate since groupings are ultimately evaluated externally by humans. For example, when objects have already been categorized by an external source, i.e., when class labels are available, we can use information theoretical measure to quantify the match between the categorization and the clustering. In our cluster ensemble, external criteria fit very well with our architecture. We use the Minkowski Score (Ben-Hur and Guyon 2003) as our

cluster quality indicator. Below is our formula for the clustering quality evaluation.

A clustering solution for a set of n elements can be represented by an nxn matrix C where $C_{ij}=1$ iff x_i and x_j are in the same cluster according to the solution and $C_{ij}=0$ otherwise.

A measure of Minkowski Score (MS) between the clustering results $C^{(h)}$ from a particular clustering algorithm CA_h with a reference clustering T (or alternatively, the true clusters if the cluster information in the data set is known in advance) is defined as

$$MS(T, C^{(h)}) = \|T-C^{(h)}\|/\|T\|, \text{ where } \|T\| = \text{sqrt}(\sum_i\sum_j T_{ij})$$

The Minkowski score is the normalized distance between the two matrices. Hence a perfect solution will obtain a score zero, and the smaller the score, the better solution.

We abbreviate the set of cluster groupings from r different clustering algorithms as $\Psi = \{ C^{(q)} | q \in \{1,\ldots,r\}\}$. The average MS score of combined clustering result C with the Ψ is defined as

$$MS^{(ANMI)}(C,\Psi) = \frac{1}{r}\sum_{q=1}^{r} MS(C,C^{(q)})$$

3.4 Experimental Results

We conduct some experiment study on both data sets from the UCI machine learning repository (http://www.ics.uci.edu/~mlearn/MLRepository.html) and yeast gene data set (http://rana.lbl.gov/EisenData.htm).

3.4.1 UCI Data Sets

Iris data set, Pen digit data set, and Vowel data set were used for our experiment. The following table shows clustering result of the three clustering algorithms and cluster ensemble.

IRIS data set(3 clusters)			
K-means	SOM	Fuzzy C-mean	**Cluster Ensemble**
0.599	0.770	0.599	**0.388**

The following table represents the element (or point) compositions of iris data set in the true clusters and the best clusters using cluster ensemble.

IRIS data set		Element composition	
		True clusters	**Best clusters using cluster ensemble**
Cluster1	Group1	**50**	**50**
	Group2	0	0
	Group3	0	0
Cluster2	Group1	0	0
	Group2	**50**	**48**
	Group3	0	2
Cluster3	Group1	0	0
	Group2	0	2
	Group3	**50**	**48**

The following tables show clustering results of the three clustering algorithms and cluster ensemble technology for pen digits data set and vowel data set.

Pen Digits data set(10 clusters)			
K-means	SOM	Fuzzy C-mean	**Cluster Ensemble**
0.962	1.164	0.998	**0.918**

Vowel data set (11 clusters)			
K-means	SOM	Fuzzy C-mean	**Cluster Ensemble**
1.215	1.332	1.503	**1.208**

3.4.2 Yeast gene data set

There are 6221 genes in the data sets but not every gene is classified into a certain function family. In our experiment we considered the genes in a function family as one cluster and created 6 data sets (cluster 2, 3, 4, 5, 6, 7). Table 1 shows 6 function families of yeast gene and how we construct the six data sets (C2, C3, C4, C5, C6, and C7) for our cluster ensemble comparison. For example, "C3" means the cluster set has 3 clusters (ATP synthesis, mitosis, and vacuolar protein targeting here)

Function Families	**# of genes**	**Cluster Sets**		
ATP synthesis	19	C3		
mitosis	19			
vacuolar protein targeting	19			
silencing	20	C5		
fatty acid metabolism	20			
meiosis	21			
phospholipid metabolism	21			
TCA cycle	22			
protein processing	27			C7
DNA repair	29	C4	C6	
protein folding	30			
nuclear protein targeting	31			
signaling	31			
major facilitator superfamily	32			
mRNA splicing	34			
chromatin structure	42	C2		
DNA replication	42			

Table 1 Some of Yeast gene function family

Table 2 shows the clustering results including cluster ensemble in Minkowski scores (MS) for each cluster set. As clearly indicated by the MS values of the clusters, the cluster ensemble method made significant improvement of quality of the clustering results over the individual clustering algorithm on all the six gene data sets. For example, the best individual clustering algorithm for C3 is K-means (***MS=0.890***), while the cluster ensemble has ***MS= 0.728***. For C5, the best individual clustering algorithm is SOM (***MS =1.241***) and the cluster ensemble reduced them to ***MS = 1.059***.

Cluster set #	**K-means**	**SOM**	**Fuzzy C-means**	**Cluster Ensemble**
C2	**0.902**	0.995	0.993	0.986
C3	0.890	0.931	0.941	**0.728**
C4	1.180	1.194	1.170	**1.071**
C5	1.207	1.241	1.229	**1.059**

C6	1.288	1.355	1.280	**1.192**
C7	1.326	1.301	1.284	**1.196**

Table 2 Clustering Results of Yeast gene data sets

4 Conclusion

In this paper we present a novel cluster ensemble approach for combining clustering results from multiple cluster algorithms. The experiment results on UCI machine learning data and gene expression data indicate that the cluster ensemble approach can generate better quality and robustness clusters compared with single best clustering algorithm.

Clustering ensemble is a new and very promising research area. There are a lot of open problems for future research. We plan to expand our ensemble approach to integrate feature selection for clustering very high dimensional data set and add some inference mechanism to automatically infer valid information from the clustering results and hope to report our findings in the future.

5 References

Azuaje, F. (2002): In Silico Approaches To Microarray-Based Disease Classification and Gene Function Discovery. *Annals of Medicine* **34**(4):299-305.

Bellaachia, A., Portnoy, D., Chen, Y. and Elkahloun, A. G. (2002): E-CAST: A Data Mining Algorithm For Gene Expression Data. *BIOKDD* 49-54.

Bloch, K. M. and Arce, G. R. (2002): Nonlinear Correlation For The Analysis Of Gene Expression Data. *Proceedings of the 2002 Workshop on Genomic Signal Processing and Statistics*, Raleigh, North Carolina.

Boley, D., Gini, M., Gross, R., Han, E., Hastings, K., Karypis, G., Kumar, V., Mobasher, B. and Moore, J. (1999): Partitioning-Based Clustering For Web Document Categorization. *Decision Support Systems* **27**:329-341.

Davies, D. and Bouldin, D. (1979): A Cluster Separation Measure. *IEEE Transactions on Patter Recognition and Machine Intelligence* **1**(2): 224-227.

Hu, X. (2001): Using Rough Sets Theory And Database Operations To Construct A Good Ensemble Of Classifiers For Data Mining Applications, *IEEE ICDM,* 233-240.

Jain, A. K., Murty, M. N. and Flynn, P. J. (1999): Data Clustering: A Review. *ACM Computing Surveys* 31(3): 264-323.

Richard, M.D. and Lippmann, R.P. (1991): Neural Network Classifiers Estimate Bayesian A Posteriori Probabilities. *Neural Computation* 3(4):461-483.

Xing, E. and Karp, R. (2001): CLIFF: Clustering of High-Dimensional Microarray Data Via Iterative Feature Filtering Using Normalized Cuts. *Bioinformatics* **17**: 306-315.

Zeng, Y., Tang, J., Garcia-Frias, J. and Gao, G.R. (2002): An Adaptive Meta-Clustering Approach: Combining The Information From Different Clustering Results, *CSB2002 IEEE Computer Society Bioinformatics Conference Proceedings* 276-287.

Dietterich, T. G. (2000): Ensemble Methods in Machine Learning. *First International Workshop on Multiple Classifier Systems, Lecture Notes in Computer Science* 1-15.

Ben-Hur, A. and Guyon, I. (2003): Detecting stable clusters using principal component analysis. In Methods in Molecular Biology, M.J. Brownstein and A. Kohodursky (eds.) Humana press, 159-182.

Zhao, Y. and Karypis, G. (2003): Clustering in Life Sciences, In *Functional Genomics: Methods and Protocols.* 183-218. Khodursky, A. and Brownstein M. (eds). Humana Press.

Berrar, D., Dubitzky, W. and Granzow M. (2002): *A Practical Approach To Microarray Data Analysis.* Kluwer Academic Publishers, 2002.

Bezdek, J. C. and Pal, S.K. (1992): *Fuzzy Models For Pattern Recognition: Methods That Search for Structures in Data*. New York, IEEE Press.

Bezdek, J. C. (1981): *Pattern Recognition With Fuzzy Objective Function Algorithms*. New York, Plenum Press.

Han, J. and Kamber, M. (2001): *Data Mining: Concepts And Techniques*. San Francisco, Morgan Kaufmann Publishers.

Hartigan, J. A. (1975): *Clustering Algorithms*. New York, John Wiley & Sons, Inc.

Kohonen, T. (2000): *Self-Organizing Maps*. New York, Springer.

Ben-Dor, A., Shamir, R., and Yakhini, Z. (1999): Clustering gene expression patterns. *Journal of Computational Biology* **6**: 281-297.

Dunn, J. (1974): Well Separated Clusters And Optimal Fuzzy Partitions. *Journal of Cybernetics* **4**: 95-104.

Goodman L. and Kruskal, W. (1954): Measures of Associations For Cross-Classifications. *J. Am. Stat. Assoc.* **49**: 732-764.

Hubert, L. and Schultz, J. (1976): Quadratic Assignment As A General Data-Analysis Strategy. *British Journal of Mathematical and Statistical Psychology* **29**: 190-241.

Kayypis, G. and Kumar, V. (1998): A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs. *SIAM Journal on Scientific Computing* **20**(1): 359-392.

Kayypis, G. and Kumar, V. (1998): Multilevel k-way Partitioning Scheme for Irregular Graphs. Journal of Parallel and Distributed Computing **48**(1): 96-129.

Strehl, A. and Ghosh, J. (2002): Cluster Ensembles - A Knowledge Reuse Framework For Combining Multiple Partitions. *Journal on Machine Learning Research (JMLR)* **3**:583-617.