Integration Analysis of Diverse Genomic Data Using Multi-clustering Results

Hye-Sung Yoon[1], Sang-Ho Lee[1], Sung-Bum Cho[2], and Ju Han Kim[2]

[1] Ewha Womans University, Department of Computer Science and Engineering, Seoul 120-750, Korea
comet@ewhain.net, shlee@ewha.ac.kr
[2] Seoul National University Biomedical Informatics (SNUBI), Seoul National University College of Medicine, Seoul 110-799, Korea
csb1749@snu.ac.kr,juhan@snu.ac.kr

Abstract. In modern data mining applications, clustering algorithms are among the most important approaches, because these algorithms group elements in a dataset according to their similarities, and they do not require any class label information. In recent years, various methods for ensemble selection and clustering result combinations have been designed to optimize clustering results. Moreover, conducting data analysis using multiple sources, given the complexity of data objects, is a much more powerful method than evaluating each source separately. Therefore, a new paradigm is required that combines the genome-wide experimental results of multi-source datasets. However, multi-source data analysis is more difficult than single source data analysis. In this paper, we propose a new clustering ensemble approach for multi-source bio-data on complex objects. In addition, we present encouraging clustering results in a real bio-dataset examined using our proposed method.

1 Introduction

Recent data mining approaches employ multiple representations to achieve more general results that are based on a variety of aspects. The extraction of meaningful feature representations yields a variety of different views on the same set of data objects using various methods. Moreover, generating high-quality results is difficult, because of the inherent noise that exists in the application of data and the inconsistency that exists among different algorithms. Therefore, recent research has show that combining the merits of several clustering methods often yields better results than using one method alone, and that clustering ensemble techniques can be applied successfully to increase classification accuracy and stability in data mining [2][6][11].

Different clustering techniques create different errors on the same set of data objects, which means that we can arrive at an ensemble that makes more accurate decisions by combining clustering results [1]. For this purpose, diverse clustering results are grouped together into what is known as a *cluster ensemble*.

N. Maglaveras et al. (Eds.): ISBMDA 2006, LNBI 4345, pp. 37–48, 2006.

However, previous work has identified several problems in optimizing the performance of clustering ensembles. First, previous methods generally have fixed the result numbers from the applied clustering algorithms, thereby resulting in the same number of clustering results; and second, highly-overlapped clustering results often are generated, clusters that are assumed to indicate the final clustering result. These problems are fundamentally difficult, and cannot be solved to yield better results. Directly combining the same number of clustering results cannot generate a meaningful result, because of the inherent noise that exists in the data, and because of the inconsistency that exists between different clustering algorithms. It also remains difficult to say which clustering result is best, because the same algorithm can lead to different results, merely secondary to repetition and random initialization. Meanwhile, with respect to the latter ensemble combination, this method generates clustering results with the same parameters to all applied algorithms. Here too, it is difficult to say which clustering result is best; even though there are different numbers of clustering results, this is not considered a characteristic of the clustering algorithm or the applied data set.

Bioinformatics is a combined interdisciplinary subject that focuses on the use of computational techniques to assist in the understanding and organization of information associated with biological macromolecules. Bioinformatics not only deal with raw DNA sequences, but also with other various types of data, such as protein sequences, macromolecular structure data, genome data and gene expression data [9]. These various types of data provide researchers with the opportunity to predict phenomena that formerly had been considered unpredictable, and most of these data can be accessed freely on the internet. Among the features of bio-data, one is that the same variables can be used to generate different types of multi-source data through a variety of different experiments and under several different experimental conditions. These multi-source data are useful for understanding cellular function at the molecular level, and they also provide further insight into their biological relatedness by means of information from disparate types of genomic data.

This paper describes a machine learning approach to an information fusion method intended for combining and analyzing multi-source genomic data. Our proposed method involves a diversity-based clustering ensemble mechanism that identifies optimal clusters, using collaborative learning of an unsupervised clustering method, based on multi-source bio-data.

The remainder of this paper is organized as follows. The application of multi-source data and clustering ensemble methods are reviewed in Section 2. Section 3 explains the proposed diversity-based clustering ensemble method, based upon genetic algorithm (GA). Section 4 describes experimental results generated by applying the proposed method, and compares these results with those generated using three other algorithms. Finally, concluding remarks and possibilities for future research are presented in Section 5.

2 Multi-source Data Analysis and Clustering Ensemble

Although the volume of data in molecular biology is growing at an exponential rates, the key features of this biological data are not so much their volume, but their diversity, heterogeneity and dispersion. Therefore, combining and analyzing different types of data is widely acknowledged in bioinformatics and genomics.

The objective of data integration analysis is to compile information from multiple data sources, so as to generate experimental results that better fit the users' goals. Also, multi-source data analysis provides and identifies correlations more accurately, using diverse independent attributes in gene classification, clustering, and regulatory networks. The collection of bio-data sources has the property that similar data can be contained in several sources, and represented in several different ways depending upon the source. However, this multi-source data analysis is useful in understanding cellular functions at the molecular level, and in providing further insight into the cells' biological relatedness. In [10], the problem of inferring gene functional classification from a heterogeneous dataset consisting of DNA microarray expression measurements and phylogenetic profiles from whole-genome sequence comparisons is considered; [10] also demonstrates that more important information can be extracted by means of using disparate types of data.

Many genomic researchers apply clustering algorithms to gain various genetic understandings of and biological information from bio-data. Clustering algorithms comprise a technique of unsupervised learning, whereby the task is to identify interesting patterns that exist within an inadequately-labeled bio-data set [14]. However, it remains difficult to say which clustering result is best, because the same algorithm can lead to many different results, as a result of repetition and random initialization. *Clustering ensemble* is a method that combines several runs of different clustering algorithms to achieve a common partition of the original dataset, aiming for consolidation of results from a portfolio of individual clustering results. This method also combines clustering results through several clustering algorithms, to generate a specific view of the data. Each clustering algorithm outputs a clustering result or label, comprised of group labels for some or all objects.

Generating high-quality clustering results is difficult, because of the inherent noise that exists in the experimental data and the different characteristics that exist among different clustering algorithms [4][5][7][8][12]. One of the major dilemmas associated with clustering ensembles is how to combine different clustering results [3]. Previous reports describing other methods have referred to the importance of ensemble algorithms, but the methods used fixed the cluster number from the clustering algorithms and ended up with the same number of clustering results [13]. However, directly combining the same number of clustering results cannot generate a meaningful result. In addition, highly-overlapped cluster results were assumed to indicate a final clustering result, but these investigators invariably searched for the optimal cluster number as well, and reapplied that cluster number, as a parameter, to all algorithms.

Several important factors must be considered when applying clustering ensemble methods.

(a) One must find a pertinent objective function when selecting the clustering results;
(b) One must use pertinent clustering algorithms to apply the ensemble;
(c) One must use an adequate fusion function to combine cluster outputs.

Diversity measures are designed to be objective functions for ensemble selection, but their performance is not convincing. Moreover, when *genetic algorithms* (GA) are used as a searching algorithm for ensemble selection, the evaluation of diversity measures may be very time consuming. To offset this problem, we now propose a method for selecting and combining cluster results. Our proposed method combines diversity measures from a multi-source dataset with the simple proposed method of GA operators, and thus allows for effective GA searching for ensemble selection.

In this paper, we assumed that our proposed method may outperform other methods in two ways. First, analysis of combined biological datasets should lead to a more understandable direction than experimental results derived from a single dataset. Second, the same variables can be used to make various types of multi-source data through different experiments and under several different experimental conditions. Therefore, we focus on optimizing the information provided by a collection of different clustering results, combining them into one final result from different data sources, using a variety of proposed methods.

3 Methods

In this section, the experimental data and experimental methods applied in this paper are explained, in detail.

3.1 Experimental Data

In this paper, the CAMDA 2006 conference dataset[1], was used as a source of multi-source data in order to test the application of the proposed method. This dataset is derived from the CDC (Center for Disease Control and Prevention) chronic fatigue syndrome (CFS) research group and contains microarray, proteomics, single nucleotide polymorphisms (SNPs), and clinical datasets. CFS is a condition that is diagnosed based upon classification criteria that are highly subjective, for the most part. The illness has no disease-specific diagnostic clinical signs or laboratory abnormalities, and it is unclear if CFS represents a single entity or a spectrum of many disorders. Prior analyses into CFS pathogenesis have not yielded further insights into the nature of this condition. One objective of the current study was to observe how our proposed method might deal with various experimental datasets on CFS, a condition for which both the clinical parameters and the pathogenesis of disease are unclear.

[1] http://www.camda.duke.edu/camda06/datasets

In our experiments, three data categories - microarray, proteomics and clinical - were used for application and verification. The first dataset, microarray data, is a single-channel experimental dataset that is comprised of 20,160 genes, using DNA from 177 patients. The second dataset, a proteomics dataset, was generated from three ProteinChip Array chemistries on the same samples (patients): Reversed Phase (*H50*), Metal Affinity Capture (*IMAC30*) and Weak Cation Exchange (*CM10*) to detect the maximal number of proteins. Among these several conditions of proteomics data, we applied four (2x2) different experimental conditions H50 and IMAC30 ProteinChip data under both high and low stringency conditions. Clinical data were used to validate the proposed method. We compared our method with three other clustering algorithms, using data from 64 patients who were common to both the microarray and proteomics datasets.

3.2 Diversity-Based Clustering Ensemble

Our diversity-based clustering ensemble approach is described as follows.

3.2.1 Generating Clustering Result Outputs

We first had to identify the optimal clustering algorithm for analysis of multi-source bio-data. However, we were faced with the inherent challenges due to the diverse features of multi-source data and the existence of many clustering algorithms. To counteract some of these concerns, we applied clustering algorithms with various characteristics to a given multi-source. We also constructed paired subsets with two clustering results that were composed of different numbers of clustering results from applied clustering algorithms. The next step was to select two parents as a couple, the couple with the largest number of highly-overlapped elements of the fitness function $F(t)$ to allow for crossover into the next GA operation. Continually, the previous process replaced two parents from the population to generate offspring after crossover, until an optimal subset was formed.

The following explains the order of the proposed method, by which we applied GA operators to a multi-source bio-data set.

3.2.2 Application of GA Operators

We propose new two GA operators, Selection and Crossover, in order to generate the optimal result.

- Method for ensemble selection

Once a suitable chromosome is chosen for analysis, it is necessary to create an initial population to serve as the starting point for the GA. The following explains in the order of the proposed selection method, with examples.

1. We construct paired subsets from two clustering results, out of all the possible clustering results for the population generation. Generating the initial population for the selection operator combines different clustering results, because multi-source bio-datasets can lead to different outputs.

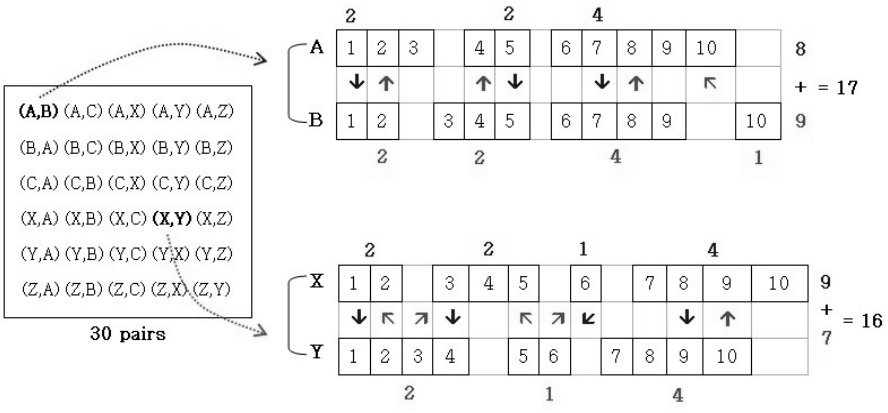

Fig. 1. Selection method for the evolutionary reproduction process

2. After generating the initial population, the next step involves selecting parents for recombination. We applied the *roulette wheel selection method* as our proposed crossover operation in this paper.

Roulette wheel selection - Simple reproduction allocates offspring using a roulette wheel, with slots that are sized according to fitness value. This is one method of choosing members from a population of chromosomes with a probability that is proportional to their fitness value. Parents are selected according to their fitness value. The better the fitness of the chromosome, the greater the probability that it will be selected.

3. In the initial population, we selected that pair had the higher fitness value; that is, two clustering results that form a pair with highly-overlapped elements. Suppose that bio-data containing 10 elements and a pair (A, B) with three and four clustering results are compared. The largest number of highly-overlapped elements is the representative cluster value. Specifically, the first cluster (1, 2, 3) of A is compared with the other clusters {(1, 2) (3, 4, 5) (6, 7, 8, 9) (10)} of B, as shown in Figure 1. The first cluster (1, 2, 3) from A and the first cluster (1, 2) from B have two values that are more highly-overlapped than the {(3, 4, 5) (6, 7, 8, 9) (10)} of B. Moreover, the (1, 2, 3, 4) cluster of Y has the same value as two of the highly-overlapped parents, with the other cluster being between the (1, 2) and (3, 4, 5) clusters of X. This process adds the representative values of each cluster and selects a final pair among 30 pairs population. As shown as (A, B) and (X, Y) in Figure 1, the representative values have 17 and 16, respectively. In this case, (A, B) pair has a greater probability of selection than the (X, Y) pair by having 17 value.

This is a process by which each chain is copied according to the values of the function which one wishes to optimize. It means that chains with greater fitness function values have a greater probability of contributing to the following generation, by creating offspring, than those with lesser fitness values. This operator is an artificial version of natural selection, wherein fitness is determined by the ability of individuals to survive.

The selection of a paired subset is executed whether each element in the clusters will survive or not, and this method is proposed as the crossover operator as follows.

■ Method for ensemble combination

Figure 2 shows the proposed crossover method.

1. During this phase, a pair produced by the selection phase initially is matched. For example, P_1 and P_2 are selected to two parents in the population.
2. Suppose that P_1 has three clustering results (C_{1_1}, C_{1_2}, and C_{1_3}) and P_2 has five clustering results (C_{2_1}, C_{2_2}, C_{2_3}, C_{2_4}, and C_{2_5}). First, we select the first cluster among the three clustering results from P_1 and see that it has more highly-overlapped traits than the other two clusters, when compared to clusters from P_2.
3. This process makes progress based upon all the clustering results of P_1. Moreover, if C_{1_1} and C_{2_3} of P_2 have the largest number of similarities, then we replace traits C_{1_1} and C_{2_3} via the following process. The C_{1_1} traits include 7, 27, 39, 58, 63, 65, 71 and 84, and C_{2_3} traits include 7, 27, 39, 58, 59, 65 and 85. In the replacement process, certain traits in C_{1_1} (63, 71, and 84) do not appear as overlapping traits in C_{2_3}. However, traits 63 and 84 in C_{1_1} do appear as traits in C_{1_2} and C_{1_3}, respectively. Consequently, traits 63 and 84 are removed, so that each trait only belongs to one cluster. The remaining trait in C_{1_1} (trait 71) is taken from C_{2_3}, so that it does not appear in any other cluster.
4. Finally, the new clustering solution is represented by the first offspring's possessing traits (C_{2_1}, C_{2_2}, *revised* C_{1_1}, C_{2_4}, and C_{2_5}).
5. This crossover operation is repeated once more, by selecting a cluster from P_2 to generate the second offspring. Two parents, P_1 and P_2, are replaced by new offspring in the final population.
6. After replacement, we again compute the fitness function in the new paired non-empty subsets to generate two clustering results; then we determine another pair of new candidates for the subsequent parent selection; and repeat the stages above.

Our proposed crossover operation exchanges the clustering traits from different clustering results and traits with highly-overlapped and meaningful information inherited by the offspring, until we ultimately achieve an optimal clustering result.

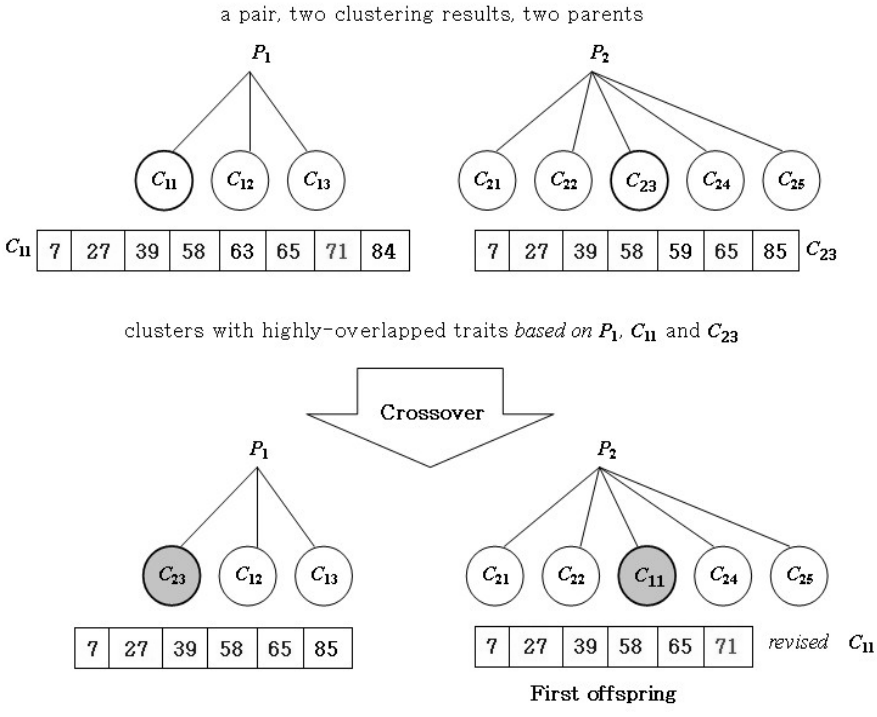

Fig. 2. Crossover operation for generating first offspring, based upon the parent P_1

4 Experimental Results

In this paper, the CLUSTER analysis tool[2] was used to generate clustering outputs from various clustering algorithms. Our experimental environment was conducted on Pentium 4 PC with 2.8G Hz CPU and 1GB. The proposed method was implemented using JAVA 1.4.2 language. The CLUSTER analysis tool performs a variety of types of clustering algorithms: hierarchical clustering, self-organizing maps (SOMs), k-means clustering, and principal component analysis. Of these, we applied hierarchical clustering, self-organizing maps, and k-means clustering algorithms, and compared the results generated using CLUSTER to those of our proposed method.

To generate clustering results using three applied algorithms, we set parameters as in Table 1.

For data analysis and validity testing, we selected 44 patients in common between the clinical, microarray and proteomics datasets.

Table 2 lists the comparisons between our method and the other clustering algorithms created by the parameter change using the H50 low and IMAC30 high-proteomics dataset. This demonstrates that the results generated using a

[2] http://rana.lbl.gov/EisenSoftware.htm

Table 1. Parameters applied to the clustering algorithms of the CLUSTER tool

Algorithms	**Parameters**
Hierarchical	All linkage clustering, based on arrays
SOMs	Ydim: 5,7,9 and 200–2,000 iterations, based on arrays
k-means	max cycles: 100 and k=3,4,5, based on arrays

Table 2. A comparison of the clustering algorithms

	H50.Low				
Cluster #	***k*-means**	**Hierarchical**	**SOMs**	**Our method**	**Actual value**
3	(**W**,L,L)	(L,L,L)	(L,**W**,L)	(**W**,**W**,L/**W**)	(W,W,W)
4	(**L**,L,**L**/W,W)	(**L**,**W**,**L**,**L**)	(**L**,L,**L**,**L**)	(**L**,**W**,**L**/W,**L**/W)	(L,W,L,L)
5	(**L**,**L**,L,**L**,**W**)	(**L**,W,L,**L**/W,L)	(W,**L**,**W**,**L**,L)	(**L**,**L**,**L**/**W**,**L**/W,**W**)	(L,L,L/W,L,W)

	IMAC30.High				
Cluster #	***k*-means**	**Hierarchical**	**SOMs**	**Our method**	**Actual value**
3	(**L**,**L**/W,**L**)	(**L**,W,**L**)	(W,**L**/W,**L**)	(**L**,**L**,**L**/W)	(L,L,L)
4	(L,W,**L**,W)	(L,L,W,L)	(W,**L**/**W**,**L**,L)	(W,L,**L**,**L**/**W**)	(L/W,L/W,L,L/W)
5	(**L**,**W**,**L**,W,**L**)	(W,L,**L**,W,**L**)	(**L**,L,**L**,**L**/W,**L**)	(**L**,M/**W**,**L**,**L**,**L**)	(L,W,L,L,L)

clustering algorithm when we have no previously defined clusters are no more consistent with the three clinical datasets than the proposed method. Specifically, the clinical dataset from CAMDA was classified into three cluster groups, based upon the overall severity of CFS symptoms- least symptoms (L), mid-level symptoms (M), and worst symptoms (W).

For validity testing, we chose to use those representative symptoms with the largest number of similarities. The representative values that are similar between the proposed method and the three different algorithms are written in bold characters. In Table 2, we discover that the results generated by our diversity-based clustering ensemble method more closely agree with the clusters classified using clinical data than the results produced by any of the other clustering algorithms. Here, L/M and M/W are found to cluster in the same ratio as the number of patients classified as least/middle and middle/worst.

Table 3 compares the cluster results for single source datasets (individual microarray and proteomics data) with the true classified clusters of the clinical dataset, using roulette wheel selection.

As shown in table 3, the proposed method demonstrates that five cluster results generate the best fitness value in paired clustering of various data sources. That is, this final cluster result number produced the most representative selected pair in the paired population. We chose the symptomatic class with the most representation and the largest number of similarities for validity testing.

Table 3. A comparison of the microarray and proteomics datasets

		Diversity-based Clustering Ensemble			Actual classification
Data set	**Cluster results#**	**Least (L)**	**Moderate (M)**	**Worst (W)**	**Representative value**
Microarray	1	2	3	0	L
	2	5	3	3	L/W
	3	6	1	6	W
	4	**4**	3	2	**L**
	5	2	1	3	M
	Total	**44 patients**			**44 patients**

		Diversity-based Clustering Ensemble			Actual classification
Data set	**Cluster results#**	**Least (L)**	**Moderate (M)**	**Worst (W)**	**Representative value**
Proteomics	1	**1**	**1**	0	**L/M**
	2	2	1	**3**	**W**
	3	5	2	5	W
	4	**6**	5	4	**L**
	5	**5**	2	2	**L**
	Total	**44 patients**			**44 patients**

Table 4. A comparison of the microarray and proteomics datasets

		Diversity-based Clustering Ensemble			Actual classification
Data set	**Cluster results#**	**Least (L)**	**Moderate (M)**	**Worst (W)**	**Representative value**
Multi-source data sets	1	**3**	2	**3**	**L/W**
	2	**5**	2	**5**	**L/W**
	3	**6**	5	4	**L**
	4	**5**	2	2	**L**
	Total	**44 patients**			**44 patients**

From these result tables, we found that the proteomics data yielded better experimental results than the microarray data, because the proteomics data more closely agrees with the clusters classified using the clinical data (comparison in bold typeface).

We also explain the experimental results of multi-source datasets. As more data sources are added to the experiment (combined microarray and proteomics data), the experimental results lead to better cluster solutions. As shown in Table 4, using the proposed method on four clusters produced the best fitness value among the generated paired subsets, and the four-cluster results were most comparable to the actual clinical data.

Here, multi-source datasets using our proposed method mostly agree with the clusters classified by clinical data. In addition, the cluster results using a data

source are no more consistent with the three symptomatic classes (L, M, and W) of the clinical dataset than the multi-source dataset generated by our proposed method. Therefore, we can say that our proposed method yields better cluster results than applying clustering algorithms to a single data source.

5 Conclusion and Discussion

We proposed a diversity-based clustering ensemble approach to generate optimal clusters on multi-source bio-data, by designing and applying new operators of the GA. We initially considered the problems inherent in combining different clustering results, by considering multi-source bio-data characteristics and the analysis of different clustering results. We also considered characteristics that present optimal cluster results from different clusters and different clustering algorithms. The experimental results show that a combined clustering approach using multi-source bio-data can generate better cluster results than those obtained using just one data source. In addition, combining clustering results from different clustering algorithms can produce better end-result clusters than the single clustering results generated using a single clustering algorithm. We need not remove elements for preprocessing, nor fix the same number of clusters during the first application step, because the GA approach is a stochastic search method that has been successfully applied in many search, optimization, and machine learning problems.

The experimental methods introduced in this paper suggest several avenues that can be taken for future research. One direction would be to identify other bio-information based on genes, as opposed to patients, in multi-source datasets. Our experimental datasets were consistent in that the rows of genes and columns of patients reflected the same level of CFS disease. We applied the columns data based on patients. Another direction, since three biological data types were used for multi-source analysis, would be to include multiple biological data types in order to discover optimal cluster results and then to again apply our proposed method. Another important task would be to develop a more theoretically and experimentally-justified verification system of multi-source data than we currently have.

References

1. Alexander, P.T., Behrouz, M-B., Anil, K.J., William, F.P.: Adaptive clustering ensembles. Proceedings of the International Conference on Pattern Recognition, **1** (2004) 272–275
2. Alexander, S., Joydeep, G.: Cluster ensembles-A knowledge reuse framework for combining partitionings. Journal of Machine Learning, **3** (2002) 583–617
3. Ana, L.N. Fred., Anil, K.J.: Combining multiple clusterings using evidence accumulation. IEEE Transactions on Pattern Analysis and Machine Intelligence, **27** (2005) 835–850
4. Duda, R.O., Hart., P.E., Stork, D.G.: "Pattern classification", seconded. Wiley, (2001)

5. Everitt, B.: "Cluster analysis. John Wiley and Sons", Inc., (1993)
6. Greene, D., Tsymbal, A., Bolshakova, N., Cunningham, P.: Ensemble clustering in medical diagnostics. Proceedings of the 17th IEEE Symposium on Computer-Based Medical Systems, (2004) 576–581
7. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: A review. ACM Computing Surveys, **31** (1999)
8. Kaufman, L., Rosseeuw, P.J.: "Finding groups in data: An introduction to cluster analysis", John Wiley and Sons, Inc., (1990)
9. Larray, T.H.Yu., Fu-lai, C., Stephen, C.F.: Using emerging pattern based projected clustering and gene expression data for cancer detection. Proceedings of the Asia-Pacific Bioinformatics Conference, **29** (2004) 75–87
10. Pavlidis, P., Weston, J., Cai, J., Grundy, W.N.: Learning gene functional classifications from multiple data types. Journal of Computational Biology, **9** (2002) 401–411
11. Qiu, P., Wang, Z. J., Liu, K.J.: Ensemble dependence model for classification and prediction of cancer and normal gene expression data. Bioinformatics and Bioengineering, (2004) 251–258
12. Theodoridis, S., Koutroumbas, K.: "Pattern recognition", Academic Press (1999)
13. Xiaohua, H., Illhoi, Y.: Cluster ensemble and its applications in gene expression. Proceedings of the Asia-Pacific Bioinformatics Conference, **29** (2004) 297–302
14. Zhou, Z.-H., Tang, W.: Clustering ensemble. Knowledge-Based Systems, (2006)