Heterogeneous Clustering Ensemble Method for Combining Different Cluster Results

Hye-Sung Yoon[1], Sun-Young Ahn[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, lovesy@ewhain.net, shlee@ewha.ac.kr
[2] Seoul National University Biomedical Informatics (SNUBI),
Seoul National University College of Medicine, Seoul 110-799, Korea
csb@medigate.net, juhan@snu.ac.kr

Abstract. Biological data set sizes have been growing rapidly with the technological advances that have occurred in bioinformatics. Data mining techniques have been used extensively as approaches to detect interesting patterns in large databases. In bioinformatics, clustering algorithm technique for data mining can be applied to find underlying genetic and biological interactions, without considering prior information from datasets. However, many clustering algorithms are practically available, and different clustering algorithms may generate dissimilar clustering results due to bio-data characteristics and experimental assumptions. In this paper, we propose a novel heterogeneous clustering ensemble scheme that uses a genetic algorithm to generate high quality and robust clustering results with characteristics of bio-data. The proposed method combines results of various clustering algorithms and crossover operation of genetic algorithm, and is founded on the concept of using the evolutionary processes to select the most commonly-inherited characteristics. Our framework proved to be available on real data set and the optimal clustering results generated by means of our proposed method are detailed in this paper. Experimental results demonstrate that the proposed method yields better clustering results than applying a single best clustering algorithm.

1 Introduction

Bioinformatics is a combined interdisciplinary subject focused on the use of computational techniques to assist the understanding and organization of information associated with biological macromolecules. Genome sequencing projects and high-throughput technologies, like microarray experimental data, have resulted in a tremendous amount of information-rich data [4], [6].

Data mining techniques have been used extensively as approaches to uncover interesting patterns from large databases [1]. Of these, clustering analysis is one of the most important approaches, because it groups elements in a data set in terms of their similarities and does not require class label information. Genomic

J. Li et al. (Eds.): BioDM 2006, LNBI 3916, pp. 82–92, 2006.

researchers are willing to apply clustering algorithms to gain better genetic understanding and biological information in the bio-data, because most bio-data are associated with insufficient prior knowledge. However, clustering techniques can be applied to analyze bio-data with their different characteristics. The challenge selecting the best algorithm, because variety clustering methods often lead to inconsistent results due to their own methodological bias and varying function criteria [12], [13]. In this paper, we describe a novel approach that digresses from using a single clustering algorithm for bio-data analysis.

The clustering ensemble problem recently has been introduced that partitions a set of objects without accessing its original features. This process demonstrated usefulness in improving the scalability and reliability of cluster results [5]. Rather than merely selecting a winning partition, we want to show that combining the clustering results of different clustering algorithms yields a better clustering solution than selecting results from a single clustering process alone. We also show a new heterogeneous clustering ensemble (HCE) method based on a genetic algorithm (GA) that combines different clustering results from diverse clustering algorithms. The use of GA is a probabilistic search approach that is founded on the concepts of evolutionary processes. Hence, we used GA approach to further improve clustering results in a HCE problem.

The paper is organized as follows. The prior clustering ensemble methods are reviewed in Section 2, along with a description of combined methods, a review of the importance of clustering results and a presentation of reasons to consider applying GA. Section 3 explains the proposed HCE method based on GA for bio-data applications. Section 4 reviews significant experimental results obtained by applying the proposed method. Finally, section 5 contains concluding remarks and future research ideas.

2 Related Works and Background

Generating high quality cluster results is a challenging problem in bio-data analysis because of the inherent noise that exists in experimental data and the inconsistency that exists among the different clustering algorithms. In the past, clustering analysis often has repeated execution of a clustering procedure, followed by selection of an individual solution that maximizes a user-defined criterion [2]. However, recent research has shown that combining of clustering results often yields better results.

Clustering ensemble techniques have recently been successfully applied to increase the accuracy and stability of classification in data mining [3], [10]. That being said, it remains difficult to say which clustering result is best because the same algorithm can lead to different results as a result of various repetitions and random initialization. The goal of cluster ensemble methods is to combine the results of multiple clustering algorithms to obtain higher-quality and more robust cluster results [8], [9]. One of the major issues of clustering ensemble is how to combine different clustering results. Previous studies regulated clustering results from clustering algorithms into the same number of clusters [13].

However, directly combining the same number of clustering results cannot generate a meaningful result. Therefore, a new mechanism to combine the different numbers of cluster results is needed to obtain better clustering results.

In this paper, we assume that effectively combining of clustering algorithms is an important method to improve cluster quality. We have focused on optimally exploiting the information provided by a collection of different clustering results by combining them into one final result, using a variety of methods. Applying GA is highly advantages for tasks requiring optimization and is highly effective in any situation in which many inputs (variables) interact to produce a large number of possible outputs (solutions) [8]. GA constitutes search method that also can be used both for solving problems and modeling evolutionary systems. Since it is heuristic, on one can know if the solution is totally accurate. However, most scientific problems are addressed via estimates, rather than assuming 100% accuracy.

Approach methods using GA can be classified broadly into two basic categories. The first category consists of generational GA that uses typical parameters such as roulette selection with elitism. This is a method by which the fittest potential parents are selected from a population; however, this does not guarantee that the fittest member proceeds to the next generation. The second method is the steady-state genetic algorithm that selects two individual parents by rank selection then combines them to produce one offspring, thereby replacing the worst characteristics (or traits) of a population with better characteristics. Unfortunately, the steady-state GA method has the potential of premature convergence, which occurs by quickly converging the solution set. The major difference between steady-state and generational GAs is that, for each parent of the population generated in the generational GA, there are two parents selected by means of the steady state method. Consequently, selection drifts appear twice as fast within a steady-state GA because this method first determines rank in the population and then every member receives fitness from as a result of this ranking.

Combining the strengths of the various methods counteracts the weaknesses of each system. Therefore, in this paper, we compromised with these two methods that first determined ranks of members and selected two parents for using crossover operation according to highly-overlapped objects.

3 Methods

In this section, the experimental data and methods applied in this paper are explained in detail. The overall experimental framework is illustrated in Figure 1.

3.1 Data

In this paper, CAMDA (Critical Assessment of Techniques for Microarray Data Analysis) 2006 conference data set (http://www.camda.duke.edu/camda06/datasets) were used in the current study as data set for the application of the proposed method. This data set is derived from the CDC (Center for Disease Control and Prevention) chronic fatigue syndrome (CFS) research group and contains

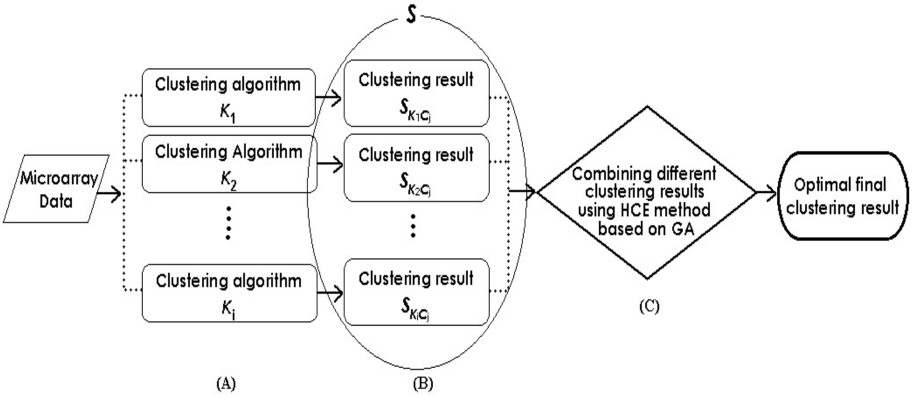
S
Microarray Data
Clustering algorithm K_1
Clustering Algorithm K_2
Clustering algorithm K_i
Clustering result $S_{K_1C_j}$
Clustering result $S_{K_2C_j}$
Clustering result $S_{K_iC_j}$
Combining different clustering results using HCE method based on GA
Optimal final clustering result
(A)
(B)
(C)

Fig. 1. Flowchart of the experimental method. (A) Apply the different types of algorithms. (B) Generate different clustering results by means of these algorithms. (C) Combine the different numbers of clustering results based on GA.

microarray, proteomics, single nucleotide polymorphism (SNP) and clinical data. In our experiments, two categories of data, microarray and clinical, were used for application and verification. The first microarray data set is a single-channel experimental data set that is composed of 20,160 genes using DNA from 177 patients. The second data set is classified 227 patients into three CFS patient subgroups (categorized by degree of clinical severity- least, middle and worst) from the CDC human subjects committee. Prior to analysis, we deal with missing values by assuming that the ratio of expression of given genes is greater than that of background intensity among microarray data, and we replaced the missing values by means of the k-nearest neighbor (kNN) method. In addition, we created a final experimental data set consisting of 19,592 genes from 169 patients; this was done by removing repeats and controls after transforming to a logarithmic ratio.

To estimate the effectiveness of the proposed method, we analyzed from 118 patient data set, which includes identical partitions about a broad range of clinical severity and microarray data, and compared our multi-dimensional clustering technique with other single clustering approaches.

CFS is a syndrome that is diagnosed on the basis of classification criteria that, mostly, are highly subjective. The illness has no diagnostic clinical signs or laboratory abnormalities, and it is unclear if it represents a single entity or a spectrum of many. Prior analyses into CFS pathogenesis have not yield further insights into the nature of this condition [7], [11]. Our own previous attempts at analysis, to data, have not yielded further insights into CFS pathogenesis either. An objective of the current study was to observe how our multi-dimensional application method deals with a condition like CFS, in which both the clinical parameters and the pathogenesis of disease is unclear. Recall that we propose to combine the strengths of different clustering algorithms to offset the weaknesses of any single algorithm.

3.2 HCE Method Based on GA Operation

Based on the work presented in Section 2, we proposed a HCE method based on GA operations to achieve optimization between different types of algorithms, K_i, and different numbers of clustering results, C_j.

The proposed HCE method must be differentiated from previous ensemble approaches. First, previous methods referred to the importance of ensemble algorithms but they were methods that did not consider the characteristics of each algorithm and dataset. Therefore, the methods fixed clustering results with the same number of clustering algorithms. In addition, highly-overlapped clustering results were assumed to indicate the final clustering result among these C_j. It goes without saying that papers applying different numbers of cluster results existed, but these investigators invariably searched for the optimal cluster number as well and reapplied the cluster number to all algorithms as a parameter.

Algorithm. HCE method based on GA operation

```
Input :

(1) The data set of $N$ data points $D =\ X_1, X_2,.., X_N$
(2) A set of clustering algorithms $K_i$
      - $i$ : the number of clustering algorithms available for analysis
(3) The cluster numbers $C_j$
      - the $K_i$ generates different cluster numbers $C_j$ for the data set $D$
(4) The clustering result is $S= \{Sk_1c_j, Sk_2c_j,....., Sk_ic_j\}$
      - $Sk_ic_j$ are clustering results consisting of $C_j$ numbers of the $i^{th}$ algorithm

Output :
The optimal clustering result on the data set $D$

1. Run clustering algorithm $K_i$ on the $D$
2. Construct a disjoint non-empty subsets, $SM^{(g)}$ with only 2 elements, from
   the clustering result $S$
3. Iterate $n$ until convergence (permute the clustering result of the data every
   iteration) :
   3.1 Compute fitness $F$(t) to select two parents/subsets from $SM^{(g)}$
   3.2 Crossover two parents
         - compare between the first parent clusters and the second parent
           clusters
         - use the first parent to replace the cluster of the second parent, which
           has the largest number of highly-overlapped objects
         - repeat once by borrowing a cluster from the second parent
   3.3 Replace parents by offspring from $SM^{(g)}$
```

Second, prior ensemble methods generally selected one best algorithm among application algorithms and indicated clustering results using this one application. We wish to address both of these problems in this paper.

The premise of our proposed HCE method based on GA operation is as follows. Different types of clustering algorithms initially are applied to the data. We then generate optimal clustering result sets by means of multiple crossover repetitions based on GA, so as to generate different clustering results. GA is a probabilistic search approach that is founded on the concept of evolutionary processes [8] and applied to further improve clustering results in our method. Our proposed algorithm, HCE method based on GA operation, is outlined as follows. In the current experiment, we aim to find associations between patients. Therefore, the input data of this algorithm executed a vector for each gene base on patients (samples). The output shows similar patient clusters for CFS.

The first stage of the algorithm is applying different types of clustering algorithms to the input data. From that result, we construct $SM^{(g)}$, a disjointed non-empty subset as a pair with only two elements from clustering results, S, of different clustering algorithms. The third stage is the GA application stage of the HCE method. We selected two parents as a couple, which has the largest number of highly-overlapped objects to fitness function $F(\mathrm{t})$ for crossover operation within the population $SM^{(g)}$. In clustering analysis, the objective of the crossover operation is to produce offspring from two parents such that the offspring inherit as much meaningful parental information as possible. That is, the clustering results convey important information and we need to find a way to effectively transmit meaningful information from parents onto their offspring. However, most traditional crossover operators were designed to deal with objects traits rather than clusters traits.

Hence, we present a novel crossover operation using data gleaned from multiple clustering processes, so as to exchange meaningful information among clusters efficiently and effectively. Our selection and use of fitness operation is elaborated in Section 3.3. In the 3.2 stage, the prior process is repeated by replacing two parents of the population to generate offspring after the crossover operation until an optimal $SM^{(g)}$ is formed.

The reason we used GA is that it allows for selection of more reliable clustering results and better extraction of optimal clustering. The algorithm replaces different clustering results by allowing the fitness function to identify similar cluster data subsets, dependent on the degree of influence that data should has on optimal final clustering. This fitness operation provides prior conditions by which to select two parents among clustering results from different algorithms.

3.3 Crossover Operation

We applied three clustering algorithms. To implement the initial population and comparison with existing algorithms, we applied k-means, hierarchical methods and principle component analysis (PCA) based clustering algorithm. The more complementary clustering algorithms also can be added without any changes to the architecture of the proposed framework. Thus far, we generated a population

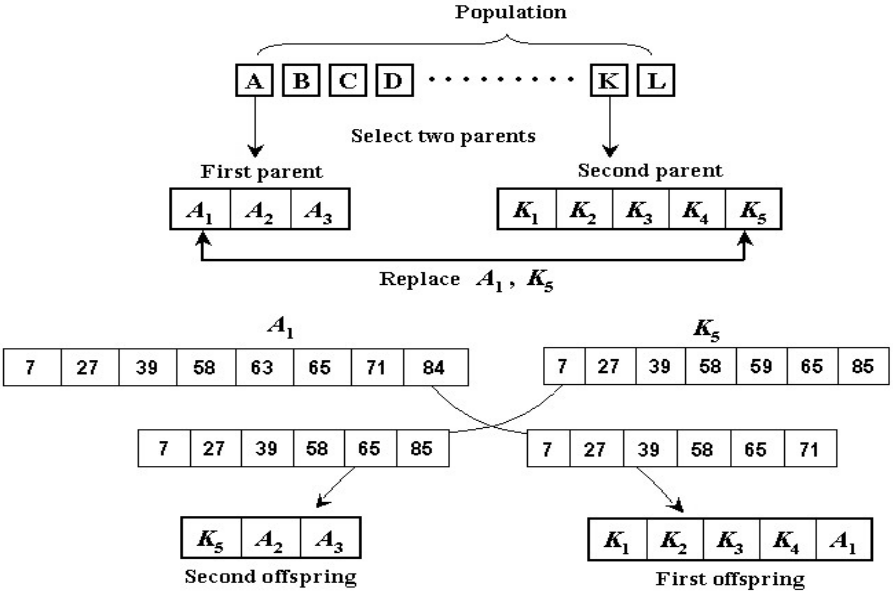

Fig. 2. Crossover operation to exchange the clustering results

totaling nine parents. That is, we created three different clustering results via the iteration and change of clusters k (3, 4 and 5) using k-means. Subsequently, the remaining two clustering algorithms were applied to yield three different clustering results.

We took the nine total different clustering results, generated by means of three clustering algorithms, and combined them with our proposed method to generate a final cluster results. We first computed the fitness function, which selects two parents, and briefly composed disjoint non-empty subsets with only two elements among nine different clustering results. For example, we can use 36 disjoint subsets with two clustering results as a pair if we have nine different clustering results. The pair with highly-overlapped objects then generates the selection of two parents during the crossover process stage.

Figure 2 explains a novel crossover approach. If we directly apply crossover operations to the ensemble problem, it may be inherited without considering clustering structures of parents, thereby eventually producing less optimal offspring [9]. For example, A and $K$are two selected parents in the initial population (Fig. 2). One parent has three clustering results (A_1, A_2, and A_3) and the other five clustering results (K_1, K_2, K_3, K_4, and K_5). First, we select one cluster, say cluster A_1, from the first parent and see that it has more highly-overlapped traits than the other two clusters (A_2 and A_3) when compared to clusters of the second parent, K. Then, we use A_1 to replace a cluster from the second parent, say K_5, which has the largest number of similarities to A_1 (objects 7, 27, 39, 58, 63, 65, 71 and 84). With replacement, those objects in A_1 (objects 63, 71 and 84) do not appear as overlapping objects in K_5, for example. However, object 63 and 84 in

A_1 appear as objects in K_2 and K_4, respectively. Consequently, objects 63 and 84 are removed so that each object belongs only to one cluster. The remaining objects in A_1 (object 71) are taken from K_5 until these objects do not appear in any other cluster. Finally, the new clustering solution is represented by the first offspring possessing traits K_1, K_2, K_3, K_4 and revised A_1. This crossover operation is repeated once by selecting a cluster from the second parent to generate the second offspring. Figure 2 shows the third stage of the proposed algorithm. Two parents are replaced by new offspring in the population in the final stage. After the replacement, we again compute fitness with the disjoint non-empty subsets using only two elements; then determine a pair of new candidates for the following parent selection; and finally repeat the stages above.

These procedures exchanges cluster traits of different clustering results and objects with highly-overlapped and meaningful information being inherited by offspring until finally we achieve an optimal final clustering result. Hence, we believe that the crossover operation we propose is a stable approach because of the invariable population of subsets and the process of combining highly-overlapped objects.

4 Experimental Results

The clinical data set from CAMDA is classified into three cluster groups: least, middle, and worst (most symptomatic) for CFS. In this paper, the AVADIS analysis tool (http://avadis.strandgenomics.com) was applied to different clustering algorithms and several parameters of the AVADIS analysis tool were applied to generate several clustering results. We also compared the results generated using AVADIS to those of our proposed method.

For data analysis and validity testing, we used 118 patients who were in common between the clinical data and microarray data sets. Table 1 represents the true classified clusters of the clinical data set.

Table 1. Classified clusters of the clinical data set. L, M and W mean least symptomatic, moderately symptomatic and most symptomatic patients number for CFS, respectively.

L	M	W	Total
42	51	25	**118**

Using the proposed algorithm, we discovered a final optimal result was composed of four clusters (cluster set # in Table 2) that have the largest number of fitness values among 36 disjoint subsets by means of 10,000 crossover operation repetitions. Using different fitness operations, it goes without saying that different cluster results may be captured. Four cluster results, those generated using three clustering algorithms and our proposed method, are compared.

Table 2 lists the comparisons between four clusters created by our method and four clusters of three clustering algorithms created by the parameter change.

Table 2. Clustering results comparison of the three clustering algorithms and HCE method based on GA

Microarray data set for CFS		**Clustering Results**	
Method	**Cluster set #**	**Algorithms**	**True clusters**
KM	Cluster 1	M	W
	Cluster 2	M	W
	Cluster 3	L	W
	Cluster 4	**L**	**L**
HC	Cluster 1	L	M
	Cluster 2	**L**	**L**
	Cluster 3	M	W
	Cluster 4	L	W
PCA	**Cluster 1**	**M**	**M**
	Cluster 2	L	W
	Cluster 3	M	W
	Cluster 4	**M**	**M**
HCE	**Cluster 1**	**L**	**L/M**
	Cluster 2	**M**	**M**
	Cluster 3	**M**	**M/W**
	Cluster 4	**L**	**L**

KM, HC, PCA and HCE mean *k*-means, hierarchical clustering, PCA-based clustering and our proposed method, respectively.

This demonstrate that the results using a clustering algorithm when we have no previously defined clusters, are not consistent with the classified three symptomatic of the clinical data set than the proposed method. To validity testing, we chose to the representative symptomatic among the largest number of similarities. The similar representative value between the proposed method and three different algorithms are written to bold characters. However, we discover that our HCE method mostly agrees with the clusters classified by the clinical data. Here, L/M and M/W are said to be clustering in the same ratio as the number of patients classified as least/middle and middle/worst.

The proposed algorithm shows that four clusters have the best fitness in disjoint non-empty subsets with two elements, and we compared them to different clustering results with four clusters. However, the clustering results of our proposed algorithm also outperformed three and five cluster results of the remaining clustering results, even though their fitness is not the best.

5 Conclusions and Future Work

Since a huge amount of gene expression data is produced by microarray experiments, a clustering technique that combines similar samples can be highly effective. The combined cluster results can find better clustering results than those obtained when using single cluster results alone.

In this paper, we considered characteristics of bio-data and clustering algorithms to present optimal clustering results by combining different types of clustering algorithms. Additionally, we proposed a HCE approach to generate optimal clusters, by newly-designing and applying the crossover operation of the genetic algorithm. The proposed method appears useful for understanding clustering results by combining several clustering algorithms for a related bio-data set.

Experiments with real microarray data show that this method can search for possible solutions effectively and improve the effectiveness of cluster analysis using crossover operations, which generate clusters of highly-overlapped traits.

We also observed that the proposed HCE method increases performance as more repetitions are added. We need not remove objects for preprocessing and fix the same cluster numbers to the first application step because the genetic algorithm is rapidly executed. Therefore, it can extract more reliable results than other clustering algorithms. In addition, clustering algorithm is an unsupervised learning method that appears useful in identifying experimental results in the absence of prior knowledge. Thus, combining different clustering algorithms by considering bio-data characteristics and analysis of clustering results also can overcome the instability inherent in clustering algorithm problems.

The experimental methods introduced in this paper suggest several avenues for future research. One direction would be to optimize cluster results by combining different bio-data sources in multi-source bio-data sets. Another would be applying different clustering algorithms under the assumption of no prior knowledge, since only one data source is used for the fitness operation. Therefore, we plan to design a proper fitness operation and novel analysis method for analysis of combined multi-source bio-data. Lastly, another important task would be to develop a theoretically and experimentally justified verification system to handle disparate data.

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. Banerjee, A., Krumpelman, C., Basu, S., Mooney, R., Ghosh, J.: Model-based overlapping clustering. Proceedings of the International Conference on Knowledge Discovery and Data Mining, (2005) 532–537
3. 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
4. Jaewoo, K., Jiong, Y., Wanhong, X., Pankaj, C.: Integrating heterogeneous microarray data sources using correlation signatures. Proceedings of the Data Integration in the Life Sciences , **LNBI 3615** (2005) 105–120
5. Jouve, P. E., Nicoloyannis, N.: A new method for combining partitions, applications for distributed clustering. Proceedings of the International Workshop on Parallel and Distributed Machine Learning and Data Mining, (2003)
6. Kasturi, J., Acharya, R.: Clustering of diverse genomic data using information fusion. Bioinformatics, **21** (2005) 423–429

7. Kenneth, J. R., Suzanne, D. V., Ellen, B., William C. R.: The economic impact of chronic fatigue syndrome. Cost Effectiveness and Resource Allocation, **2** (2004)
8. Liu, J. J., Cutler, G., Li, W., Pan, Z., Peng, S., Hoey, T., Chen, L., Ling X. B.: Multiclass cancer classification and biomarker discovery using GA-based algorithms. Bioinformatics, **21** (2005) 2691–2697
9. Patrick, C.H. Ma., Keith, C.C. Chan.: Discovering clusters in gene expression data using evolutionary approach. Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence, (2003) 459–466
10. Qiu, P., Wang, Z.J., Liu, K.J.: Ensemble dependence model for classification and prediction of cancer and normal gene expression data. Bioinformatics, **21** (2005) 3114–3121
11. Whistler, T., Unger, E. R., Nisenbaum, R., Vernon, S. D.: Integration of gene expression, clinical, and epidemiologic data to characterize Chronic Fatigue Syndrome. Journal of Translational Medicine, **1** (2003)
12. Xiaohua, H.: Integration of cluster ensemble and text summarization for gene. Proceedings of the 4th IEEE Symposium on Bioinformatics and Bioengineering, (2004) 251–258
13. Xiaohua, H., Illhoi, Y.: Cluster ensemble and its applications in gene expression. Proceedings of the Asia-Pacific Bioinformatics Conference, **29** (2004) 297–302