A Novel Framework for Discovering Robust Cluster 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. We propose a novel method, called heterogeneous clustering ensemble (HCE), to generate robust clustering results that combine multiple partitions (clusters) derived from various clustering algorithms. The proposed method combines partitions of various clustering algorithms by means of newly-proposed the selection and the crossover operation of the genetic algorithm (GA) during the evolutionary process.

1 Introduction

Data mining techniques have been used extensively as approaches to uncover interesting patterns within large databases. Of these, genomic researchers are willing to apply clustering algorithms to gain a better genetic understanding of and more biological information from bio-data, because there is insufficient prior knowledge of most bio-data. However, no single algorithm has emerged as the method of choice within the bio-data analysis community, because most of the proposed clustering algorithms largely are heuristically motivated, and the issues of determining the correct number of clusters and choosing a good clustering algorithm are not yet rigorously resolved [6].

Recent research shows that combining clustering merits often yields better results, and clustering ensemble techniques have been applied successfully to increase classification accuracy and stability in data mining [1][3][4]. Still, it remains difficult to say which clustering result is best, because the same algorithm can lead to different results as a result of repetition and random initialization. One of the major dilemmas associated with clustering ensembles is also how to combine different clustering results. Therefore, a new mechanism to combine the different numbers of cluster results is needed.

In this paper, we want to demonstrate a new heterogeneous clustering ensemble (HCE) method, based on a genetic algorithm (GA) that combines clustering results from diverse clustering algorithms, thereby resulting in clustering ensemble improvement. We focus on optimizing the information provided by a collection of different clustering results, by combining them into one final result, using a variety of proposed methods.

N. Lavrač, L. Todorovski, and K.P. Jantke (Eds.): DS 2006, LNAI 4265, pp. 373–377, 2006.
© Springer-Verlag Berlin Heidelberg 2006

The paper is organized as follows. Section 2 explains the proposed heterogeneous clustering ensemble method based on genetic algorithm. Section 3 describes the experimental data and significant experimental results obtained by using the newly-proposed method. Finally, Section 4 presents our conclusions.

2 Algorithm

This section explains our proposed heterogeneous clustering ensemble (HCE) method, which functions via the modified application of basic GA operations.

2.1 Genetic Algorithm (GA) for the Better Selection Problem

The following modified genetic operators, Reproduction and Crossover, were applied in this paper.

Reproduction. 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. We applied different types of clustering algorithms to a dataset and constructed a paired non-empty subset with two clusters, among all clustering results of clustering algorithms. For example, one clustering algorithm generates three clusters (1, 2, 3) and the other also generates three clusters (A, B, C) using different parameters. These six clusters are created as an initial population and are comprised of 30 paired non-empty subsets (Figure 1). This natural reproduction process uses the fitness function as a unique way to determine whether each chromosome will or will not survive. In this paper, we select a pair (subset) with the largest number of highly-overlapped elements among all paired subsets, which is the fitness function to select a pair for the next crossover operation.

For instance, suppose that bio-data with 10 elements, as shown in Figure 2, generate an initial population through the reproduction operation. If two subsets (1, 2) and (1, 3) are selected paired subsets from Figure 1, the first cluster (1, 2, 3) in {(1, 2, 3) (4, 5, 6) (7, 8, 9, 10)} is compared with the other clusters {(1, 2) (3, 4) (5, 6) (7, 8, 9, 10)}. That is, the first cluster (1, 2, 3) and the first cluster (1, 2) from the other cluster results have 2 values to the highly-overlapped than {(3, 4) (5, 6) (7, 8, 9, 10)} clusters, as shown in Figure 2. Moreover, the (4, 5, 6) cluster has a value of 2 to the highly-overlapped with the other cluster (5, 6) and the (7, 8, 9, 10) cluster has a representative value equal to 4, derived by comparing it to the other cluster (7, 8, 9, 10) of {(1, 2) (3, 4) (5, 6) (7, 8, 9, 10)}. This process adds the representative values of each cluster and selects a subset for the crossover operation, by comparing all population pairs. As shown as (A) and (B) in Figure 2, the subsets of (1, 2) and (1, 3) each have 17 and 15, and finally, the (1, 2) subset was selected with greater selection probability.

Crossover. The selected subset produces offspring from two parents, such that the offspring inherit as much meaningful parental information as possible. This operator process is based on [2] where the methodological ideas can be found. These procedures exchange the cluster traits from different cluster results and

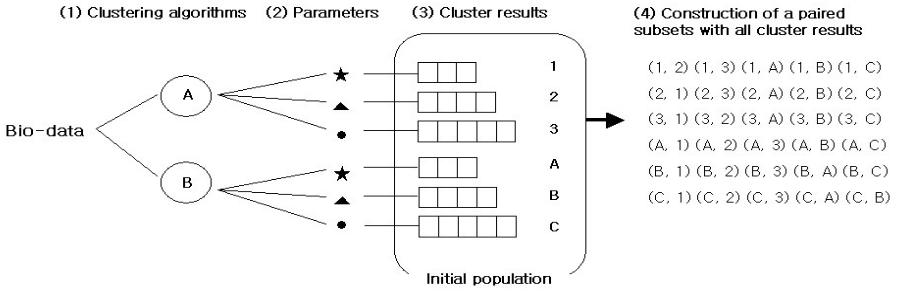
(1) Clustering algorithms
(2) Parameters
(3) Cluster results
(4) Construction of a paired subsets with all cluster results
Bio-data
A
B
1
2
3
A
B
C
Initial population
(1, 2) (1, 3) (1, A) (1, B) (1, C)
(2, 1) (2, 3) (2, A) (2, B) (2, C)
(3, 1) (3, 2) (3, A) (3, B) (3, C)
(A, 1) (A, 2) (A, 3) (A, B) (A, C)
(B, 1) (B, 2) (B, 3) (B, A) (B, C)
(C, 1) (C, 2) (C, 3) (C, A) (C, B)

Fig. 1. Initial generation of the population

Elements 10 : 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
(A) Subset (1, 2)
[{(1, 2, 3) (4, 5, 6) (7, 8, 9, 10)} , {(1, 2) (3, 4) (5, 6) (7, 8, 9,10)}]
$\Rightarrow$ [{(2, 2, 4) , (2, 1, 2, 4)}] $\Rightarrow$ [8 , 9] $\Rightarrow$ **17**
(B) Subset (1, 3)
[{(1, 2, 3) (4, 5, 6) (7, 8, 9, 10)} , {(1, 2) (3, 6) (4, 5) (7, 8) (9, 10)}]
$\Rightarrow$ [{(2, 2, 2) , (2, 1, 2, 2, 2)}] $\Rightarrow$ [6 , 9] $\Rightarrow$ 15

Fig. 2. Selection method for the evolutionary reproduction process

elements with highly-overlapped and meaningful information being inherited by the offspring, until finally we achieve an optimal final cluster result.

2.2 Heterogeneous Clustering Ensemble (HCE)

Our proposed HCE method based on the GA operation is as follows.

Algorithm. Heterogeneous Clustering Ensemble (HCE)

Input :

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

Output : The optimal cluster results on the dataset D

```
1. Run clustering algorithms $K_i$ on $D$
2. Construct a paired non-empty subset, $SM^{(g)}$, from the cluster results, $S$
3. Iterate $n$ until convergence:
   3.1 Compute the fitness function $F(t)$ and select two parents from $SM^{(g)}$
   3.2 Crossover two parents
   3.3 Replace $SM^{(g)}$ parent subsets by newly-created offspring
```

In the present experiment, we aim to identify associations between patients. Therefore, the input data of our algorithm is executed to a vector for each gene, based on patients (samples). The end result demonstrates similar patient clusters. The first stage presented in our algorithm applies different clustering algorithms to the input data. From that result, we construct, $SM^{(g)}$, a paired subset with only two elements from the cluster results, S, of different clustering algorithms. The third stage is the GA application stage within the HCE algorithm. We select two parents as a couple, with the largest number of highly-overlapped elements to fitness function $F(t)$ for crossover manipulation within the population $SM^{(g)}$.

3 Application

This section explains the experimental data and our experimental results.

3.1 The Datasets

We used the published CAMDA 2006 conference dataset. This dataset contains microarray, proteomics, single nucleotide polymorphisms (SNPs), and clinical data for chronic fatigue syndrome (CFS) [5]. In our experiments, both microarray and clinical, were used for application and verification [2].

3.2 Experimental Results

We applied the AVADIS analysis tool to generate cluster results from different clustering algorithms. We also compared the results that were generated using AVADIS to those generated by our proposed method [2]. Table 1 lists the comparisons between the four clusters created by our method, and the four clusters of three different clustering algorithms created by the parameter change. The applied different clustering algorithms - KM, HC, PCA and HCE - indicate k-Means, Hierarchical Clustering, Principal Component Analysis, and our proposed method, respectively.

For validity testing, we used the two categories of clinical datasets: clinical assessment data to determine whether CFS is a single or heterogeneous illness; and actual classified clinical data about the symptomatic groups. Thereafter, we compared the final four clusters that resulted by means of our proposed method with the four cluster results derived using the other clustering algorithms. As shown Table 1, we chose a representative symptomatic group from every cluster results of the two categories data. The most similar representative values between the two categories are written in **bold** typeface. As a result, we found that our HCE method mostly agrees with the clusters classified by the two categories of clinical data. Here, L/M and M/W are said to cluster in the same ratio as the number of patients classified symptomatically as least/moderate and moderate/worst.

Table 1. Cluster results comparison of three clustering algorithms and the HCE method

Algorithms	**KM**				**HC**				**PCA**				**HCE**			
Cluster results #	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
Category 1	L	M	L	**L/M**	**L/M**	**L**	M	L	**M**	L	M	**M**	**L**	**M**	**M**	**L**
Category 2	W	W	W	**L**	**L**	**L**	W	W	**M**	W	W	**M**	**L**	**M**	**L/M**	**L**

4 Conclusions

Experimental results using a real dataset prove that our method can search effectively for possible solutions and improve the effectiveness of clusters. Combining different clustering algorithms by considering bio-data characteristics and analysis of cluster results, also can overcome the instability inherent in clustering algorithm problems. And our proposed HCE method improves its performance as the number of iterations increase. Moreover, we need not remove elements for preprocessing, nor fix the number of clustering during the first application step, as required by existing clustering algorithms, because the GA is rapidly executed. Therefore, it can extract more reliable results than other clustering algorithms.

References

1. 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
2. Hye-Sung, Y., Sun-Young, A., Sang-Ho, L., Lee, Sung-Bum, C., Ju Han, K.: Heterogeneous clustering ensemble Method for combining different cluster results. Proceedings of Data Mining for Biomedical Applications, PAKDD Workshop BioDM, **LNBI 3916** (2006) 82–92
3. 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)
4. 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
5. 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)
6. Xiaohua, H.: Integration of cluster ensemble and text summarization for gene. Proceedings of the 4th IEEE Symposium on Bioinformatics and Bioengineering, (2004) 251–258