The Genetic Algorithm Incorporates with Rough Set Theory—An Application in Marketing

Wen-Yau Liang

***Abstract*—This paper proposes the constrained clustering technique combining Genetic Algorithm (GA) and Rough set theory. Based on the result of clustering, Apriori is then used to generate the associate rules in products and marketing people can recommend related products to the targeting segment. The experiments had showed that the proposed approach is better than the other clustering methods.**

***Index Terms*—Market segmentation, clustering, rough Set theory, genetic algorithm.**

I. Introduction

Clustering is a combinatorial optimization question if the scale of the question is increased. Researchers also indicated the clustering will become NP-Hard problems under particular functions and the number of clustering is greater than three [1]. Since self-organizing maps (SOM) was suggested by Kohonen in 1982, it has been applied to many studies because of its good performance [2]. Despite several excellent applications, SOM has some limitations hindering its performance. Especially, just like other clustering methods, such as K-means, Ant System. SOM has no mechanism to determine the number of clusters, initial weights and stopping conditions. And it will affect the qualities of clustering results [3]. Although, GA is believed to be effective on NP-complete global optimization problems and they can provide good near-optimal solutions in reasonable time [3], applying constraints into the process of clustering will promote the quality of clustering [4]. Therefore, in this paper, the Rough set theory is used to generate the constraints to improve the GA performance. In the Rough set method, we can extract the minimal attribute sets without deterioration of quality of approximation, and minimal length decision rules corresponding to lower or upper approximation [5]. In this study, a new constrained clustering technique combining Genetic Algorithm and Rough set theory is proposed. The Rough set is used to generate the number of clustering and rules. Then using these outputs, constrained GA Clustering method can decreases the running time and improve the quality of clustering. Afterward, Apriori algorithm is used to find out the associated itemsets for each group and these rules can be implemented as the marketing strategy.

Manuscript received March 13, 2013; revised June 17, 2013.

W. Y. Liang is with the Department of Information Management, National Changhua University of Education, Taiwan, ROC (e-mail: wyliang@cc.ncue.edu.tw).

II. Research Method

The main idea of the constrained clustering is applying constraints to the process of clustering to decrease the running time and improving the quality of clustering. Although literatures have shown GA as effective in NP-complete global optimization problems, applying constraints in GA could improve the efficiency. Therefore, we proposed the GA-Based Clustering method incorporating Rough set theory to solve the clustering problems. The Rough set is used to generate the number of clustering and rules. Then using these outputs, constrained GA Clustering method can improve the quality of clustering.

The GA incorporating Rough set theory is performed in the following three stages. The GA incorporated with Rough set theory is detailed next.

Stage 1: Predispose of Rough set: in this paper, the rule extraction algorithm (REA) developed by [6] is applied to extract the useful reduct rules, which includes four steps:

Step1: Create basic units and put into Database.

Step2: Calculate the lower and upper approximations for basic units.

Step3: Find the core and reduct of attributes.

Step4: Find the core and reduct of attributive values.

After above 4 steps, we can get rules and number of clusters 'm', then put m into GA-Based Clustering method.

Stage 2: GA-Based Clustering: the reduct rules generated in stage 1 are applied in this stage to increase the effectiveness of the GA evolution.

Step 1: Define the Parameters and select the rules

- 1.1 Collect all objects.
- 1.2 Set initial population size as *N*.
- 1.3 Set generation *k* to 1.
- 1.4 Set the number of chromosome (cluster) as m (obtained from the previous stage).
- 1.5 Set mutation rate, crossover rate and termination condition.
- 1.6 Select the rules.

Step 2: Initialization: in the initial step, a population of *N* chromosomes is generated. All the chromosomes are of length m. Each chromosome represents a subset of $\{O_1, O_2, \ldots, O_m\}$ .The O is defined as the initial center for each cluster.

- 2.1 Start with chromosome number, *x*, set to 1; clustering number, *y*, set to 1
- 2.2 Randomly select an object which satisfies the selected rules and put into the population. If cannot find one, Stop.
- 2.3 *y*=*y*+1, if *y*<=*m*, go to 2.2.
- 2.4 *x*=*x*+1, if *x*<=*N*, *y* set to 1 and go to step 2.2, otherwise go to step 3.

DOI: 10.7763/IJEEEE.2013.V3.238

Step3: The Procedures for Generating Clusters: this step is adapted from [7] and the parameters, symbols are defined as follows:

Oi: Objects *i*

Ca: The number of each cluster.

Sa: The center of C_a.

When $S_a = (S_{a1}, S_{a2}, \cdots, S_{ap})$, we can use formula 1 to calculate all the central values of C_a, and $O_i \in C_a$, $|C_a|$ is the number of C_a:

$$S_{aq} = \frac{\sum_i O_{iq}}{|C_a|} \quad (1)$$

After the initial step, we begin to generate the clusters and the process as follows:

The objects in $\{O_1, O_2, \ldots, O_m\}$ are taken one by one and the distance between the taken object *O* and each *S* is calculated using the following distance function (Euclidean distance):

$$d(O_i, S_a) = \left(\sum_{q=1}^{p} O_{iq} - S_{aq} \right)^{1/2} \quad (2)$$

where $O_i = (O_{i1}, O_{i2}, \ldots, O_n)$ and $S_a = (S_{a1}, S_{a2}, \ldots, S_{ap})$.If $d(O_i, S_a) \le d(O_i, S_K)$ for all *k* so $1 \le k \le m$ and $k \ne a$, we can determine the O_i belong to C_a.After all objects have been put into clusters, we use the formula 1 for calculating the center of the clusters again.

- 3.1 Assign each object O_i , *i*=1,…,n to cluster *Ca*, *a*=1,…,m again. For all *K* ($1 \le k \le m$ and $k \ne a$), If $d(O_i, S_a) \le d(O_i, S_K)$ then O_i belongs to C_a.
- 3.2 If the objects in the clusters do not change anymore, the process of generating clusters terminates, and we use formula 3 to calculate the Fitness function *F*.

$$F = f \times \left(\frac{1}{SSW}\right)$$
$$SSW = \sum_{Oi \in Ca} d(Oi, Sa)^2 \quad (3)$$

It is expected to collect some objects to a cluster in clustering if there is only one object in one cluster $f = 0$, otherwise, $f = 1$.We have to judge the terminated condition after obtaining the fitness function, if the condition is coped with, stop, otherwise, go to step 4.

Step 4: Evolution process

- 4.1 Evaluate every chromosome for fitness.
- 4.2 If the termination condition is satisfied, go to end.
- 4.3 Use a suitable selection strategy to select and duplicate the new population.
- 4.4 Crossover: Crossover process is performed by the crossover rate.
- 4.5 Mutation: Mutation process is performed by the mutation rate.
- 4.6 Set generation *k*= *k*+1 and go to 3.1.

Stage 3: Apriori Algorithm: this stage determines the associated itemsets.

- 1.1 Set the minimum support.
- 1.2 Count the support of candidates C_k.
- 1.3 Check if find new large itemsets? If No, Go to Step 1.6.
- 1.4 Prune: delete from C_k all itemsets in which (k-1)-subsets are not in L_k-1, also does not delete any itemsets that could be in L_k.
- 1.5 Join: extend L_k-1 with each item in the database, and then delete itemsets in which the (k-1)-itemsets obtained by deleting the (k-1)th item is not in L_{k-1}. Go to Step 1.2.
- 1.6 Count the confidence.
- 1.7 Identify the associated itemsets.

III. Experimental Analysis

In this paper, a prototyping system is created in Java for test and validation. System execution environment: (1) CPU is Intel Q8400 2.66GHz; (2) RAM is 2 GB. The GA parameters such as population size, generation, crossover rate, mutation rate, chromosome length and termination condition all can be set through the system. Now an online 3C retailer would like to implement the proposed approach. The operation process of the example is illustrated as follows:

Stage 1: Predispose of Rough set: After we have completed the process of predisposing of the Rough set and get four rules and the number of clustering (4). This information will be applied into the GA-Based Clustering process.

Stage 2: GA-Based Clustering: The evolutionary process is executed until the termination condition is satisfied. Finally, at generation 17, all population reach the same value, and the fitness equal one another, we called it converged. The final result is obtained.

Stage 3: Apriori Algorithm: Cluster 3 (92 records) is used to demonstrate the processes, the other clusters can be referred. Only two association rules have satisfied the minimum confidence, That is, the system will recommend to customers.

To prove the effectiveness of the proposed approach, we compare it with other clustering techniques: GA based clustering, K-means and Two-Step Method. We evaluate the results of clustering by two indexes. The first index is Root-Mean-Square Standard Deviation (RMSSTD), which is used to evaluate differences in a cluster. The researchers evaluate differences between clusters by the second index, R-Square. Therefore, for the value of RMSSTD, the smaller the better, and the nearer the value of R-Square is to 1, the better. To decrease the effects of setting parameters on results, we run every experiment 1000 times, obtain the total amount with clustered results, and then calculate the averaged value. The results show that the clustered indexes of the Constrained Clustering GA method are clearly better than the other methods. The experiments demonstrated improvements in clustering accuracy/execution time/flexibility that could help businesses to effectively query market segmentation.

Business thus can deftly respond to rapid changes in marketing and recognize the characters of customers belonging to different clusters.

IV. Conclusion

In this paper, a novel approach was proposed for clustering that is systemic and complete. It shows great promise in the emerging demand for clustering from complicated datasets. Also, the rules generated by the Rough set and applied to different GA parts, the constrained GA is believed to be effective dealing with clustering issues. The results show the proposed approach leads to a better quality of clustering than other methods.

Acknowledgment

This work was partially supported by funding from the National Science Council of the Republic of China (Grant #: NSC 99-2410-H-018-016-MY3).

References

[1] J. W. Welch, "Algorithmic complexity: three NP-hard problems in computational statistics," *Journal of Statistical Computation and Simulation*, vol. 15, no. 1, pp. 17-25, 1983.

[2] E. A. Fernandez and M. Balzarini, "Improving cluster visualization in self-organizing maps: Application in gene expression data analysis," *Computers in Biology and Medicine*, vol. 37, no. 12, pp. 1677-1689, 2007.

[3] K. J. Kim and H. Ahn, "A recommender system using GA K-means clustering in an online shopping market," *Expert Systems with Applications*, vol. 34, no. 2, pp. 1200-1209, 2008.

[4] I. Davidson, M. Ester, and S. S. Ravi, "Efficient Incremental Constrained Clustering," in *Proc. the 13th SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 240-249, 2007.

[5] M. Inuiguchi and T. Miyajima, "Rough set based rule induction from two decision tables," *European Journal of Operational Research*, vol. 181, no. 3, pp. 1540-1553, 2007.

[6] C. C. Huang and T. L. Tseng, "Rough set approach to case-based reasoning application," *Expert Systems with Applications*, vol. 26, no. 3, pp. 369-385, 2004.

[7] R. J. Kuo, K. Chang, and S. Y. Chien, "Integration of Self-Organizing Feature Maps and Genetic-Algorithm-Based Clustering Method for Market Segmentation," *Journal of Organizational Computing and Electronic Commerce*, vol. 14, no. 1, pp. 43-60, 2004.

Wen-Yau Liang is a professor of Information Management at National Changhua University of Education. He received his Ph.D. from the University of Iowa. His research interests are object-oriented design, artificial intelligence, intelligent agent and electronic commerce. He has published papers in journals sponsored by various societies.