IMPLEMENTATION OF GENETIC ALGORITHMS TO CLUSTER NEW STUDENTS INTO THEIR CLASSES

Zainudin Zukhri[1] and Khairuddin Omar[2]
[1]Department of Informatics, Faculty of Industrial Technology, Islamic University of Indonesia
Kampus Terpadu UII Jl Kaliurang Km 14.5 Yogyakarta
[2]Department of System Management and Science, Faculty of Information Science and Technology,
National University of Malaysia, 43600 UKM Bangi
E-mail: [1]zainudin@fti.uii.ac.id, [2]ko@ftsm.ukm.my

ABSTRACT

Clustering new students into their classes at random probably make an educational problem, because the smartest student maybe clustered in a same class with the most stupid one. To avoid this problem, we can use sorting-score method which cluster new students based on their achievements. First, we sort the average of their scores, and then make the clusters(classes) base on its. This method is not so worse than the first one, but only the smartest class and the most stupid class that have a low gap. There are high gaps in the middle classes. This research tries to explore Genetic Algorithm (GA) to solve this problem. Experimental studies show that performance of GA is better than sorting-score method. The gap of intelligence in classes clustered by GA are relatively same each other. GA can reduce maximum gap in class that clustered by sorting-score method.

Keywords: *cluster, Genetic Algorithm, similarity, student.*

1. INTRODUCTION

Clustering new students is a new problem in education area, because universities register new students in a large number is just recently happened. It did not happened in many years ago. If clustering new students problem is neglected and it is not handled seriously, other educational problem can be happened. The other educational problem is resulted by classes that contain smart students and stupid students.

But this fact is often happened, because universities usually cluster new students into their classes by random, base on the sequence of registration time. A cluster of students who register early will be clustered into the first class, and then the next cluster of students who register at the next time will be clustered into the second class, and so on. Because of this method, the smartest student can be clustered in a same class with the most stupid student. This class has a high dissimilarity (gap). Of course, it can make an educational problem. Ideally, to avoid educational problem, the classes should contain similar students or students with low gap of intelligence.

A better method is sorting-score method that cluster the new students base on their achievements. Average of the students scores are sorted and then the students are clustered based on its. This method is not so worse than the first one, but only the smartest class and the most stupid class that have a good similarity. In additional, there are high gaps in the middle clusters.

2. RELATED WORK

There are too many researchers in general clustering area, but it is very difficult to find a researcher in clustering new students, because universities register new students in a large number is just recently happened. There are also too many researches in application of GA. Meanwhile in statistics area, there is a popular approach to solve clustering problem, that is agglomerative method [3]. But this method has few weaknesses, especially in clustering problem with very large objects [2].

GA is a computational abstraction of biological evolution that can be used to solve some optimization problems [4]. GA is not function optimizers, but can be adapted to work as such [6]. GA must be adapted to suit the problem, in particular the representation and operators need to be designed carefully [8].

Cole used GA to cluster any objects so that each cluster has high dissimilarities with other clusters and each cluster contains similar objects [2]. His idea about chromosomes representation and GA operators is very good to be used. But we cannot use all of his works to cluster new students into classes, because the dissimilarities between clusters (classes) are not important in clustering new students. In other word, we must define a special fitness function for clustering new students. Beside of that, chromosome representation in his works does not enough to represent classroom with their quota.

3. CLUSTERING NEW STUDENTS ANALYSIS

Recently universities have become accustomed to register new students in a large number, so that they have to cluster their students into some paralel classes in order to guarantee the education quality. If there are some paralel classes, of course, it needs some lecturers (teachers). It also

needed some classrooms. Practically only few universities which be able to provide facilities like classroom, lecturers in same capacities and same qualities. This condition can make an educational problem, how to make the standard services to the students in each class.

If universities still cluster their new student into the classes randomly, the educational problem will be happened. Surely, if the smartest student is clustered in a same class with the most stupid one, the lecturers will have such difficulties to decide the intelligence standard of the students. With random method, new students commonly cluster based on the sequence of registration time. As mention before, Students who register early will be clustered into the first class, and then if the class has full or quota of the classroom be reached, student will be clustered into the next class, and so on until all students have registered and got their class.

To avoid some problems that can be happened as a result of the random method, universities cluster their new student based on the ranking of academic score, such as subject scores, IQ, and so on. The academic score is sorted, and then a first cluster of students at top level will be clustered as the first class. The next cluster of students at next level will be clustered as the next class, and so on. With this method, the universities hope that there is a similar cluster of students in each class.

3.1 Problems in Clustering New Students

Clustering new student is not different from general clustering problem. It is a complex process, because one combination of objects that represented one cluster of the objects must be found from all possibilities of combination. Clustering new students into their classes is a part of clustering problem, even more complex than general clustering problem. The complexity is caused by quota of each classroom. It is a constrain of this problem. Because of the quota, number of students in a cluster or a classroom must equal or less than its quota. It is different from clustering other objects, like image, document, map, etc.

It is like in general clustering object, the important issue is how to determine the similarity between two objects, so that clusters can be formed from objects with a high similarity to each other [1]. To avoid educational problem, the classes should contain similar students or consist of students with low gap of intelligence.

It is difficult to find the solution of clustering new students problem in the very big solution space of clustering problem for large number of objects. The size of solution space or the number of ways of sorting n students into c classes is [7]

$$N(n,c) = \frac{1}{c!}\sum_{i=0}^{c}(-1)^i\binom{c}{i}(c-i)^n \qquad (1)$$

For example if there are only 25 students and 5 classes, the number of ways of sorting is N (25, 5) = 2,436,684,974,110,751. It is very large size, and if we try to find the solution by traditional methods, surely we only find a local optimum solution.

3.2 GA Approach

GA is a search algorithm based on the mechanism of natural selection and natural genetics. They combine survival of the fittest among string structures with a structured yet randomized information exchange to form a search algorithm with some of the innovative flair of human search. In every generation, a new set of artificial creatures (strings) is created using bits and pieces of the fittest of the old; an occasional new part is tried for good measure. While randomized, GAs are no simple random walk. They efficiently exploit historical information to speculate on new search points with expected improved performance [5].

3.2.1 Chromosome Representation

To cluster n new students into c classes with q_i quota of each class where $1 \le i \le c$, chromosome representation is designed as follows:

1. A chromosome divides into c sub chromosome. The i^{th} sub chromosome is represent i^{th} class. It consist q_i gen.
2. Each gen is an integer g where $1 \le g \le n$, j^{th} gen represents j^{th} student, so that gen is different each other in one chromosome.

The chromosome representation is shown as Figure 1.

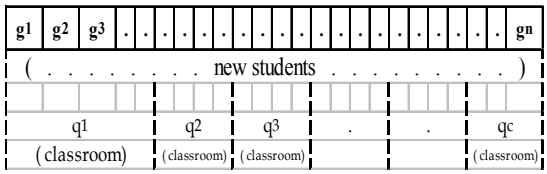

Figure 1. Chromosome Representation of clustering n new students into c classes with q_i quota of each class where $1 \le i \le c$.

3.2.2 Objective Function and Fitness Function

The objective function in clustering new students is minimization of the maximum intelligence gap in each classroom. In the clustering terminology, it is minimization of distance between the furthest objects in all clusters. Maximization of distance between the clusters does not considered in clustering new students.

Assume that each student g has m attribute ($X_1 ... X_m$) and clustering is based on those attributes. If distance between two objects is defined as Euclidean distance as follows:

$$d(g_1, g_2) = \left[\sum_{k=1}^{m}\left(X_{k,1} - X_{k,2}\right)^2\right]^{\frac{1}{2}} \qquad (2)$$

then the objective function is

$$h(x) = \text{Min}\left\{ \sum_{i=1}^{c} d_i(g_a, g_b) \right\}, \quad (3)$$

where $a \neq b$, $1 \leq a \leq q_i$, and $1 \leq b \leq q_i$.

and the fitness function is

$$f(x) = \frac{1}{h(x)+1} \quad (4)$$

3.2.3 Operators

If quota of the classroom or number of object in the cluster is not considered, the chromosome representation is almost similar with permutation representation by Cole [2], but our model does not need special character as separator. Our model is similar with Zukhri's model in application of GA to solve Assignment Problem (AP) [9]. Hence we can use operators in application of GA to solve AP. We use Rollet Wheel Selection, Order Crossover and Reciprocal Exchange Mutation.

4. PERFORMANCE EVALUATION

In this section, we are going to show the performance of GA to cluster new students into their classes based on random data as shown in Table 1. The 200 students in Table 1 will be cluster by GA into 5 classes. The implementation of GA is using Delphi 5.0.

Table 1. 2-dimensional data generated by Microsoft Excel

i	*test1*	*test2*
1	79	73
2	98	92
…	…	…
…	…	…
200	58	85

We compared the performance of GA with sorting-score method. The experiment show that performance of GA is better than sorting-score method. The performance comparison is shown in Table 2. The best performance of GA is reached with parameters as follows:

- Population size = 300
- Probability of cross over = 75%
- Probability of mutation = 1%

Table 2. Comparison between GA and sorting-score method

i^{th} class	*maximum gap*	
	GA	*Sorting-score method*
1	46.044	31,890
2	47.885	46,690
3	52.202	60,142
4	46.615	51,662
5	48.918	36,878

As show in Table 2, maximum gap of intelligence in 3rd class by sorting-score method is 60,142. It is greather than the maximum gap by GA, 52,202. The classes that clustered by sorting-score method have greater potential to make educational problem.

5. CONCLUSION

The conclusions of this research are :

1. GA can cluster new students into their classes. GA reduce maximum gap of intelligence in class that clustered by sorting-score method.
2. The gap of intelligence in classes clustered by GA are relatively same each other.
3. We recommend that the next researches should try to increase the performance of GA by other parameters, operators or chromosome representations.

REFERENCES

[1] Bakar, Z. A., Deris, M.M., and Alhadi, A.C. (2005). Performance Analysis of Partitional and Incremental Clustering, *Proceeding of Seminar Nasional Aplikasi Teknologi Informasi* (*SNATI 2005*) - *Jurusan Teknik Informatika Fakultas Teknologi Industri UII*, G1.

[2] Cole, R.M. (1998). *Clustering with Genetic Algorithms*. Master Thesis University of Western Australia.

[3] Everitt, B.S., Landau, S. and Leese, M. (2001). *Cluster Analysis*. London: Heinemann Educational Books Ltd.

[4] Gen, M. & Cheng, R. (1999). *Genetic Algorithms and Engineering Optimization*. Canada: John Wiley & Sons, Inc.

[5] Goldberg, D.E. (1989). *Genetic Algorithms in Search, Optimization & Machine Learning*. Canada: Addison-Wesley Publishing Company.

[6] Jong, K.A.D & Whitley, L.D (editor). 1993. *Genetic algorithms are not function optimizers. Foundations of Genetic Algorithms 2*. California: Morgan Kaufmann.

[7] Liu, C.L. (1968). *Introduction to Combinatorial Mathematics*. California: McGraw Hill.

[8] Michalewicz, Z. (1998). *Genetic Algorithms + Data Structures = Evolution Programs*. Third revised and extended edition. New York: Springer-Verlag.

[9] Zukhri, Z. (2004). Penyelesaian Masalah Penugasan dengan Algoritma Genetika. *Proceeding of Seminar Nasional Aplikasi Teknologi Informasi* (*SNATI 2004*) - *Jurusan Teknik Informatika Fakultas Teknologi Industri UII*, J51.

ISBN: 1907-5022