COMPARATIVE EVALUATION OF GENETIC ALGORITHM AND MODIFICATION OF AGGLOMERATIVE METHOD FOR ALLOCATING NEW STUDENTS

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

ABSTRACT

Allocating new students into their classes is a clustering problem, that is how to cluster new students into their classes so that each class contains students in the number that less than or equals to its capacity and has minimum gap of intelligence. It needs a suitable method to avoid an educational problem. This paper describes the comparison of Genetic Algorithm (GA) and Modification of Agglomerative Methods (AM) for solving this problem. To determine which method is better then the other, the software of each method which can cluster n students with m attributes into c classes are evaluated by two-dimensional random data consists of 200 students. Then we compare the results. Comparison of GA and AM for clustering the data sets shows that although the GA cluster the data successfully, the method provides no advantages over AM. Intelligence gap of students in each class clustered by GA almost same each other, but the average of this value is greater than by AM. Meanwhile, the intelligence gap of student clustered by AM depend on the clustering sequence. This GA performance may be is caused by unsuitable GA approach, both chromosome representation and GA operators in this research. Better GA approach may enhance the effectiveness of the GA searching.

Keywords: *Agglomerative Method, cluster, Genetic Algorithm, student.*

1. INTRODUCTION

Allocating new students into their classes is a part of clustering problem[8], that is how to distribute students to classes so that each class consists of students with intelligence level as similar as possible, and the number of students in each class must not exceed the capacity. In other words, the classes should contain students with low gap of intelligence. It is an important matter, because it very difficult to give good education service for students in large number whose high diversity of achievements[9] or high variation of skills[10]. With the students allocated to the groups, discriminating policies to these groups can be implemented easily[6].

This problem usually is ignored and new students only allocated into their classes at randomly. It can make an educational problem. To avoid this problem, new students must be clustered with a suitable method. For a while, there is sorting-score method (SSM) which clusters the new students based on their achievements. This method is not so worse than the first one, but only the smartest class and the weakest class that have a good similarity. There are high gaps in the middle class.

Now, there are many clustering methods have been developed to be used in wide area. But the difference in principle of student clustering, make difficulty to use them directly, and we should modify them. As mentioned by Jain that clustering is a subjective process; the same set of data items often needs to be partitioned differently for different applications. This subjectivity makes the process of clustering difficult. This is because a single algorithm or approach is not adequate to solve every clustering problem[3]. In clustering new students, the number of objects (students) in each cluster (class) cannot be determined based on the result of clustering process, but it is determined before clustering process. In addition, dissimilarity between each class can be ignored in clustering new students. Hence implementation of clustering methods needs modification.

This paper discussed about clustering techniques for new student allocation problem with statistical approach and GA approach. The main inherent idea is to compare those clustering techniques to determine which clustering technique is better based on maximum gap of intelligence in the classes. The better technique must minimize this value.

2. RELATED WORK

Students allocation problem can be viewed as a type of constrained multi-dimensional bin packing problem, with students being ”items” to be packed and the classes being ”bins” [11]. If the objective is to minimize the number of classes, this view can be applied. Because of the objective is to minimize the gap of intelligence in each class, student allocation

problem should be viewed as clustering problem rather than a bin packing problem.

Susanto et al. have used Fuzzy C-Means algorithm (FCM) for solving this problem[8]. In their experiment, they cluster students of certain subject base on their score of prerequisite subjects. It is a good work, but it has not shown the advantage of FCM yet, because it only involve 20 students.

AM, the most popular statistical approach for clustering problem, cannot be applied to solve this problem directly and it should be modified. Experimental study shows that AM generates classes with maximum intelligence gap growing proportional with the clustering sequence[12].

Cole has used GA for solving general clustering problem[1]. He used GA to cluster any objects so that each cluster has high dissimilarities with other clusters and each cluster contains similar objects. His idea about chromosome 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. Beside of that, chromosome representation in his works does not enough to represent classroom with its capacity. However, it inspires us to modify his work for solving new students allocation problem[13].

3. RESEARCH METHODOLOGY

We assume that attributes of new students are their scores of admission test that represented as integer numbers between 0 and 100. We make two different approach to solve the student allocation problem, GA approach and AM approach. To determine the advantages and the disadvantages of each approach, we develop them as software which can cluster n students with m attributes (dimensions) into c classes and evaluate them with a same data. We generate a two-dimensional random data to do it. Finally we compare the results.

4. AM APPROACH

There are five popular AM[1], those are Single Linkage Method (SLM), Complete Linkage Method (CLM), Centroid Method (CM), Average Method (AVM), and Ward's Method (WM). Methods differ in how the distance between clusters is calculated. AM presented in Algorithm 1 for grouping n objects must be modified to group n students into c classes with its quotas.

Algorithm 1. Agglomerative Methods.

1. Begin with n clusters, each containing one object.
2. Calculate the Euclidean distance between each pair of clusters. These distances are usually stored in a symmetric distance matrix.
3. Merge the two clusters with the minimum distance.
4. Update the symmetric distance matrix.
5. Repeat Steps 3 and 4 until a single cluster remains.

The modification of AM to group n students into c classes whose q_i quota of i^{th} class where $1 \leq i \leq c$ is presented in Algorithm 2.

Algorithm 2. Modification of Agglomerative Methods.

1. Begin with n clusters, each containing one student.
2. Calculate the distance between each pair of clusters. These distances are usually stored in a symmetric distance matrix.
3. Merge the two clusters with the minimum distance as a cluster. If the cluster contains q_i students, collect the students as i^{th} class. Put this class out from the distance matrix.
4. Update the distance matrix.
5. Repeat Steps 3 and 4 until all students grouped into their classes.

5. GA APPROACH

GA is a computational abstraction of biological evolution that can be used to solve some optimization problems[2]. GA is not function optimizers, but can be adapted to work as such [4]. GA must be adapted to suit the problem, in particular the representation and operators need to be designed carefully[7].

We modify one of Cole's model for chromosome representation[1], that is permutation representation. But we define a special fitness function for clustering new students so that GA can generate classes that contain students with intelligence level as similar as possible[13]. We ignore the dissimilarity of intelligence between each class. We use Roulette Wheel Selection, Order Crossover and Reciprocal Exchange Mutation as GA operators. See [2] for the details of algorithm and these operators.

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

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

This representation is shown as Figure 1.

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.

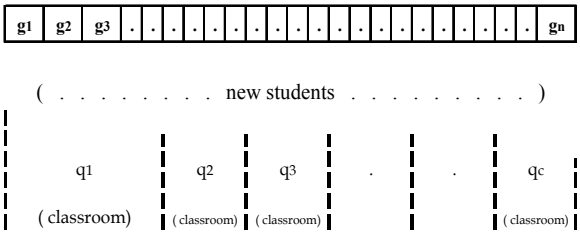

Figure 1. Chromosome Representation of clustering n new students into c classes with q_i quota of each class where $1 \le i \le c$.

Assume that each student g has m attribute (x_1, x_2, .. , x_m) and clustering is based on their attributes. If distance between two objects is defined as Euclidean distance as follows:

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

Then the objective function is

$$h(x) = Min\left\{ \sum_{i=1}^{c} d_i(g_a, g_b) \right\}, \quad (2)$$

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

And the fitness function is

$$f(x) = \frac{1}{h(x)+1} \quad (3)$$

6. EXPERIMENTAL RESULTS

In this section, we are going to show the comparison of GA and AMs to cluster new students into their classes based on random data as shown in Table 1. The 200 students in Table 1 will be clustered into five classes. The implementation of each approach is using Delphi 5.0.

We compared the performance of GA with AMs. Table 2 shows the performance comparison between GA and AMs. The best performance of GA is reached with population size equals to three hundred, cross over probability equals to seventy five percent, mutation probability equals to one percent and number of generations equal to two hundred.

As show in Table 2, maximum intelligence gap of classes generated by GA is more flatten than by AM, but the average of maximum intelligence gap is greater than by AM. It means that AM is relatively better than GA, but distribution of maximum the intelligence gap generated by AM which not flatten in all classes means that AM also provides no advantage over SSM.

Table 1. 2-dimensional data

i	*x1*	*x2*	*i*	*x1*	*x2*	*i*	*x1*	*x2*	*i*	*x1*	*x2*	*i*	*x1*	*x2*
1	79	73	41	99	64	81	98	77	121	78	60	161	80	69
2	98	92	42	92	98	82	82	97	122	81	89	162	75	91
3	82	90	43	64	86	83	76	95	123	98	58	163	65	99
4	59	61	44	64	57	84	93	91	124	58	70	164	73	65
5	92	70	45	87	91	85	58	99	125	58	68	165	71	77
6	87	78	46	75	75	86	94	88	126	98	56	166	82	92
7	83	96	47	99	71	87	97	62	127	65	67	167	65	64
8	59	64	48	73	75	88	91	67	128	82	96	168	79	89
9	66	96	49	67	85	89	74	78	129	95	88	169	77	85
10	74	77	50	62	70	90	99	61	130	74	71	170	61	86
11	62	65	51	68	75	91	59	89	131	62	83	171	80	68
12	56	64	52	73	78	92	82	71	132	67	56	172	65	69
13	91	64	53	79	66	93	86	68	133	80	97	173	76	98
14	56	59	54	71	94	94	59	77	134	76	75	174	90	62
15	61	93	55	56	73	95	95	98	135	58	78	175	56	74
16	95	66	56	81	87	96	74	87	136	58	85	176	61	84
17	79	64	57	79	66	97	82	95	137	80	77	177	97	87
18	75	99	58	87	91	98	96	94	138	79	82	178	86	73
19	61	60	59	94	87	99	93	72	139	88	58	179	59	57
20	93	79	60	79	77	100	83	69	140	75	58	180	97	96
21	94	64	61	61	83	101	92	68	141	84	92	181	70	98
22	68	69	62	70	96	102	78	78	142	82	95	182	95	57
23	75	83	63	78	91	103	77	84	143	59	59	183	87	71
24	65	88	64	94	90	104	57	74	144	87	72	184	56	91
25	74	60	65	61	92	105	82	56	145	84	64	185	61	64
26	61	74	66	87	99	106	90	83	146	78	90	186	55	79
27	57	100	67	56	78	107	91	74	147	72	98	187	94	57
28	71	99	68	84	68	108	82	67	148	61	67	188	76	91
29	76	70	69	90	59	109	81	77	149	56	79	189	76	56
30	66	80	70	68	84	110	86	83	150	88	62	190	78	57
31	98	61	71	60	71	111	57	82	151	99	94	191	92	81
32	78	80	72	63	76	112	94	70	152	69	95	192	58	74
33	69	72	73	82	62	113	64	99	153	75	66	193	74	92
34	61	75	74	81	63	114	73	67	154	92	81	194	70	95
35	61	59	75	57	83	115	64	89	155	99	71	195	76	64
36	77	99	76	74	63	116	91	83	156	83	68	196	90	74
37	77	77	77	100	80	117	74	66	157	96	80	197	78	77
38	86	63	78	61	96	118	81	89	158	64	82	198	93	86
39	93	88	79	91	63	119	71	60	159	60	82	199	59	61
40	57	83	80	75	69	120	92	94	160	79	58	200	58	85

Table 2. Comparison between GA and AMs

class	*maximum gap*					
	GA	*SLM*	*CLM*	*CM*	*AM*	*WM*
1	46.04	23.43	19.92	36.62	21.02	58.69
2	47.89	28.02	26.93	33.42	24.35	46.69
3	52.20	24.35	24.74	51.04	30.41	39.60
4	46.62	43.83	47.38	35.61	30.41	36.80
5	48.92	51.40	54.82	55.44	55.44	55.44
average	48.33	34.21	34.76	43.88	32.33	46.12

This result must be given an attention, because GA can reach a good achievement in optimization of many other areas. The result reached by GA in this experiment shows that we use unsuitable approach. It is very sensible because chromosom representation in this research make the searching space more wide than the real problem. The searching space of GA in this research depends

on the wide of chromosome, that is the number of students, and does not depend on the number of classes at all. For 200 students, the searching space is factorial of 200, it is much greater than the total way to cluster 200 students into five classes [5] or

$$200! >> \frac{1}{5!}\sum_{i=0}^{5}(-1)^i\binom{5}{i}(5-i)^{200} \qquad (4)$$

Hence a better result probably can be reached, if we can change the chromosome representation that can reduce the searching space. But it is also means changing the GA operators.

7. CONCLUSION

Comparison of GA with AM for clustering the new students shows that although the GA clusters the data successfully, the method provides no advantages over AM. Intelligence gap of students in each class clustered by GA almost same each other, but the average of this value is greater than by AM. Meanwhile, the intelligence gap of student clustered by AM depend on the clustering sequence. This GA performance may be caused by unsuitable GA approach, both chromosome representation and GA operators in this research. Better GA approach may enhance the effectiveness of the GA search.

REFERENCES

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

[2] Gen, M. & Cheng, R. *Genetic Algorithms and Engineering Design*. Canada: John Wiley & Sons, Inc., 1997.

[3] Jain, AK., Murty, MN., & Flynn, PJ. *Data Clustering: a Review*. ACM Computing Surveys, Vol. 31 No. 3, pp. 264-323, 1999.

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

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

[6] Ma, Y., Liu,B., Wong, CK., Yu, PS., & Lee, SM. *Targeting the right students using data mining*. Proceedings of the 6 th ACM SIGKDD international conference on Knowledge discovery and data mining KDD, pp. 457 - 464, 2000. Michalewicz, Z. *Genetic Algorithms + Data Structures = Evolution Programs*. Third revised and extended edition. New York: Springer-Verlag, 1998

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

[8] Susanto, S., Suharto, I., & Sukapto, P. *Using Fuzzy Clustering Algorithm for Allocation of Students*. Transaction on Engineering and Technology Education, Vol. 01 No. 02, pp. 245-248, 2002.

[9] Vanderhart, PG. *Why do some schools group by ability?* American Journal of Economics and Sociology, Vol.65 No. 02, 2006.

[10] Wiedemann, T. *A Virtual Textbook for Modeling and Simulation*. Proceedings of the 2000 Winter Simulation Conference, pp. 1660 - 1665, 2000.

[11] Wright, M. *Experiments with a Plateau-Rich Solution Space*. Proceedings of the 4th Metaheuristics International Conference, pp. 317-320, 2001.

[12] Zukhri, Z., & Omar, K. *Modification of Agglomerative Methods to Cluster New Students into Their Classes*. Proceedings of the 1st International Conference on Mathematics and Statistics, pp. 996-1013, 2006.

[13] Zukhri, Z., & Omar, K. *Implementation of Genetic Algorithms to Cluster New Students into Their Classes*. Proceedings of Application of Information Technology National Seminar, pp. A101 - A104, 2006.