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

Zainudin Zukhri¹, Khairuddin Omar²

¹Department of Informatics, Faculty of Industrial Technology, Islamic University of Indonesia
Kampus Terpadu UII Jln. Kaliurang Km 14.5 Yogyakarta

²Department 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 ≤ i ≤ c is presented in Algorithm 2.

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 ≤ i ≤ 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 ≤ g ≤ 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.
Then the objective function is

\[ f(x) = \frac{1}{h(x)+1} \]  

(2)

Then the objective function is

\[ h(x) = \min \left\{ \sum_{i=1}^{c} d_i(x_a|x_b) \right\}, \]

where \( a \neq b, 1 \leq a \leq q_1, \) and \( 1 \leq b \leq q_2. \)

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

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

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 shown 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>
<thead>
<tr>
<th>x1</th>
<th>x2</th>
<th>x3</th>
<th>\ldots</th>
<th>x20</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 1. 2-dimensional data

<table>
<thead>
<tr>
<th>x1</th>
<th>x2</th>
<th>x3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 2. Comparison between GA and AMs

<table>
<thead>
<tr>
<th>class</th>
<th>maximum gap</th>
<th>GA</th>
<th>SLM</th>
<th>CML</th>
<th>CM</th>
<th>AM</th>
<th>WM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>46.04</td>
<td>23.43</td>
<td>19.92</td>
<td>36.62</td>
<td>21.02</td>
<td>58.69</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>47.89</td>
<td>28.02</td>
<td>26.93</td>
<td>33.42</td>
<td>24.35</td>
<td>46.69</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>52.20</td>
<td>24.35</td>
<td>24.74</td>
<td>51.04</td>
<td>30.41</td>
<td>39.60</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>46.62</td>
<td>43.83</td>
<td>47.38</td>
<td>35.61</td>
<td>30.41</td>
<td>36.80</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>48.92</td>
<td>51.40</td>
<td>54.82</td>
<td>55.44</td>
<td>55.44</td>
<td>55.44</td>
<td></td>
</tr>
</tbody>
</table>

average: 48.33

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 chromosome 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

\[
200! \gg \frac{1}{5!} \sum_{i=0}^{5} (-1)^i \binom{5}{i}(5-i)^{200} \quad (4)
\]

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