p-ISSN : 2774-9576
e-ISSN : 2747-2981

Teacher Placement Using K-Means Clustering and Genetic Algorithm

Haris Sriwindono[1], PH Prima Rosa[2], Kartono Pinaryanto[3]
[1,2,3]Informatics Department, Science and Technology Faculty, Sanata Dharma University[1,2,3]
Kampus III Paingan, Maguwoharjo, Sleman, Yogyakarta, phone: 62-21-883037[1,2,3]
Indonesia[1,2,3]
e-mail: haris@usd.ac.id[1]

To cite this document:
Sriwindono, H., Rosa, P. P., & Pinaryanto, K. . The Teacher Placement using K-Means Clustering and Genetic Algorithm. Conference Series, 4(1). Retrieved from https://adi-journal.org/index.php/conferenceseries/article/view/669

Hash:ABCY2It412TG6suXNHVvuWBuqgB5hdOAZmqjqPBPQNKFBPcrOHio96PDIU1qcuiR

Abstract

The problem of teacher placement in a school is a problem faced by Magelang Regency. The success of teacher placement is determined by the minimum total distance between the teacher and the school, with the aim that teacher performance is maintained. In computer science this problem is an NP-hard problem that takes a very long time to achieve optimal results when done with conventional methods. Another approach to solve this problem is to use heuristic algorithms, one of which is by using genetic algorithms. To further improve the performance of genetic algorithms, one way is to narrow the search space. In this study, the problem will be broken down first through a clustering process so that the search space becomes narrower, before being subjected to genetic algorithm processes. This study will cluster the original data, before being subjected to a genetic algorithm to solve the problem of teacher placement. The clustering method used is the K-Means clustering method while the Genetic Algorithm uses the Ordered Crossover (OX) operator and the Partial Shuffle Mutation (PSM) mutation operator. From this study, it was found that by performing K-Means clustering before the optimization process using the Genetic Algorithm turned out to get better results than without using clustering. The total distance without clustering is 11751 km while with clustering it 9259 km. Also the total running time to execute this program turned out to be much shorter (from the order of hours to the order of minutes).

Keywords : clustering, K-means, genetic, crossover, mutation

1. Introduction

Factors that influence teacher performance are teacher education level, teaching supervision, upgrading program, conducive atmosphere, facilities and infrastructure, teacher's physical and mental condition, principal's leadership style, welfare insurance, principal's managerial ability, training, and provision of incentives [1]. Teachers will have good performance if they are supported by good physical and mental conditions. A healthy teacher will be able to complete his tasks well. Therefore health factors must really be considered. Likewise the mental condition of the teacher, if his mental condition is good he will teach well too. Meanwhile, in many places in Indonesia, the distance from the teacher's residence to the school has never been used as a determinant of teacher placement. This condition affects the physical and mental health of teachers. It is also known that teacher performance also has a positive effect on student achievement [2]. The long distance from home to school will result in physical fatigue and possibly mental fatigue due to traffic conditions. Therefore the distance between home and school should be close or can be reached in a short time.

The placement of elementary school teachers in Magelang Regency, Indonesia so far has not taken into account the distance between home and school, so this condition can still be improved by rearranging the placement of teachers. The teacher placement arrangement is intended to reduce the total distance between all teachers to the school. This problem is an optimization problem in which a configuration of teacher placement must be formed so that the total distance between teachers and schools is obtained with this configuration to a minimum. If a deterministic algorithm is used, the optimal result will be obtained after trying all possible configurations, namely n! where n is the number of teachers. The number of teachers is very large, in this case there are 636 teachers, so it will take a very long time if it is resolved deterministically. So the deterministic optimization algorithm is not suitable to be used as a solution. In this study, one of the probabilistic algorithms will be used, namely the Genetic Algorithm, where there is no need to try all possible teacher configurations.

In this research, this problem will be solved by using Genetic Algorithm and K-Means Clustering. In previous studies, the same problem was solved using only genetic algorithms and applied to all available data so that it requires a long running time [3] [4]. Clustering here is used to divide the initial data into several clusters so that the amount of data in each cluster becomes smaller, and Genetic Algorithm is used on the data in each of these clusters. Thus, it can be expected to increase the speed of the process or the running time of the program. In addition, it will also be reviewed whether the shortest distance obtained is better than without clustering.

2. Research Method

The data in this study were obtained from DISDIKPORA Magelang Regency, where the number of teachers was 636, the number of schools was 106 and each school was assumed to have 6 classes or 6 study groups. Teacher data consists of teacher ID, teacher name and home position coordinates (longitude, latitude), while school data consists of school ID, school name and school position coordinates (longitude, latitude). The data on the distance from the teacher to the school is calculated based on the Haversine formula.

$$a = \sin^2(\Delta\varphi/2) + \cos\varphi 1 \cdot \cos\varphi 2 \cdot \sin^2(\Delta\lambda/2)$$
$$c = 2 \cdot \text{atan2}(\sqrt{a}, \sqrt{(1-a)})$$
$$d = R \cdot c$$

φ is latitude, λ is longitude, R is earth's radius (mean radius = 6,371km);
note that angles need to be in radians to pass to trig functions!

So this case is a problem to place teachers into study groups in schools so that the total distance is minimal.

This research was conducted in the following order of the main steps

1. Input teacher, school and distance data
2. Perform k clustering of school data with K-Means

```
3. Perform clustering for teacher data according to school clustering
4. For i=1 to k do
    Perform optimal configuration for cluster-I with Genetic Algorithm
5. Next i
6. Add up the total distance of all clusters
```

2.1 K-Means Clustering

K-Means Clustering is a data mining method to perform the modeling process without supervision (unsupervised) and is a method for grouping data with a partition system. K-Means Clustering is a non-hierarchical cluster analysis method that seeks to partition existing objects into one or more clusters or groups of objects based on their characteristics, so that objects with the same characteristics are grouped in the same cluster and objects with similar characteristics. different groups are grouped into other clusters [5]. In other words, the K-Means Clustering method aims to minimize the objective function set in the clustering process by minimizing variations between data in one cluster and maximizing variation with data in other clusters. Thus, clusters will be found in the data, with the number of clusters represented by the variable K where K is the desired number of clusters. This algorithm accepts input in the form of data without group labels or clusters. In this learning algorithm, the input data will be grouped automatically without first knowing the target group or cluster. The inputs received are data or objects and the desired number of clusters (K). This algorithm will group data or objects into K groups. In each cluster there is a central point (centroid) that represents the cluster.

Data clustering using the K-Means Clustering method is generally carried out with the following basic algorithm [6]

1. Determine the number of clusters
2. Allocate data into clusters randomly
3. Calculate the centroid/average of the data in each cluster
4. Allocate each data to the nearest centroid/average
5. Go to Step 3, if there is still data that moves clusters or if there is a change in the centroid value, some are above the specified threshold value or if the change in the value of the objective function used is above the specified threshold value

2.2 Genetic Algorithm

Genetic algorithm is a heuristic optimization inspired by natural selection and genetics. This algorithm was developed by Holland [7] and Goldberg [8]. In genetic algorithms, the population of potential solutions is referred to as a set of chromosomes. Chromosomes or individuals evolve successively from generation to generation using a set of genetic operators, namely selection, crossover and mutation. New individual reproduction processes are created through the application of these operators. A large number of operators have been developed to improve the performance of the genetic algorithm, because the performance of this genetic algorithm depends, among other things, on the ability of the operator. The selection operator is used to select the chromosomes to be combined to become new individuals. The crossover operator is used to fuse genetic information between chromosomes to explore the search space. Meanwhile, mutation operators are used to maintain a population of adequate chromosomal diversity and avoid premature convergence or prevent local optimum.

In general, the steps of the genetic algorithm are as follows [8]:

1. [Start] Generates an initial population of n chromosomes.
2. [Fitness] Calculates the fitness value f(x) of each chromosome in the population
   2.1. [New population] Create a new population by repeating the following steps until the new population is complete.
   2.2. [Selection] Selects two parental chromosomes from the population according to their fitness value (the better the fitness value, the more likely it is to be selected).
   2.3. [Crossover] With a certain probability of crossover (crossover) the selected parents to produce offspring. If there is no cross-breeding, then the new offspring are the original copies (duplicates) of their parents.

2.4. [Mutation] With certain mutation probability, mutation process is carried out on certain chromosomes in certain genes.
3. [Accepting] Placing a new breed in the population
4. [Replace] Use this new population for the next process.
5. [Test] If the final condition is reached then stop, and enter the best result in the population.
6. [Loop] to step 2

In this study, the roulette wheel method was used as the selection operator, while the crossover operator used the order crossover method, and the mutation process used the Partial Shuffle Mutation method.

Roulette Wheel Selection

Individual selection is the stage of selecting individuals in a population. Individuals who have a high probability value have a greater probability of being selected in the next process. In this research, the method used is roulette wheel selection. The roulette wheel method is often used in the selection process in genetic algorithms. To do this method, the value of the relative fitness value and the cumulative fitness value must be known first.
Here is the formula to find the relative fitness value of each chromosome:

$$P[i] = \frac{f[i]}{\Sigma f}$$

Where :

P = relative probability.
i = i-th chromosome.
f = fitness value.
f = Total fitness of all chromosomes.

After getting the relative fitness value, look for the cumulative fitness value with the formula:

$$C[i]=C[i-1]+P[i]$$

Where :

C = Cumulative probability.
i = i-th chromosome.
P = relative probability.

Then generate a random number Ri with {Ri ϵ R | 0 < Ri < 1, i = 1, 2, …. , N}. If Ri < Ci then the i-th chromosome is the parent. However, if C[i-1] < R[i] < C[i] choose the i-chromosome as the parent.
Where :

R = Random Number.
i = i-th chromosome.
C = Cumulative probability.

Ordered Crossover (OX)

Crossover is a mechanism for obtaining new offspring by involving two parent chromosomes, where the genes on the two parent chromosomes are 'crossed' to obtain new offspring that have a different gene arrangement from the parent. Various methods of crossover have been developed by researchers, including Single Point Crossover, N Point Crossover, Ordered Crossover, Partially Map Crossover, Cycle Crossover and others [9]. In this study, Ordered Crossover (OX) will be used. In this ordered crossover, two reference points are determined randomly on the parent chromosome. To generate a new chromosome the genes between the two reference points on the second chromosome are replaced with the genes from the first chromosome in their corresponding positions. Then the genes on the second chromosome are scanned circularly starting from the position after the second reference point. If there is a gene that has not yet appeared on the child's chromosome, then the gene is inserted in the position starting after the second reference point, and so on in a circular fashion until all gene positions on the child's chromosome are filled. The following is an example of implementing the OX operator:

Suppose there are two parent chromosomes, namely P1 and P2 with the following reference points:

P1: 2 1 5 4 | 7 8 9 3 | 6 10
P2: 1 5 4 6 | 10 2 8 7 | 3 9

Then the genes that lie between reference point 1 and reference point 2 from P1 will be copied to the child chromosome (O2) in the same position, the results are as follows:

O2 = _ _ _ _ | 7 8 9 3 | _ _

The genes in P2 are examined circularly, starting after the second reference point, whether the gene has appeared in O2 or not, if it has appeared it is ignored, if it has not appeared then it is stored first, then it will be inserted into the child's chromosomes sequentially and circularly.

The results are as follows:
(3) has appeared (ignore)
(9) has appeared (ignore)
(1) has not appeared (ok)
(5) has not appeared (ok)
(4) has not appeared (ok)
(6) has not appeared (ok)
(10) has not appeared (ok)
(2) has not appeared (ok)

The gene sequences that have not yet appeared are 1, 5, 4, 6, 10, 2. Then these genes are inserted into the daughter chromosome (O2) starting from the position after the second reference point, in a circular manner as follows:

O2 = _ _ _ _ | 7 8 9 3 | _ _
O2 = 4 6 10 2 | 7 8 9 3 | 1 5

In the same way, another child is produced, namely O1. Finally we get two children as follows:

O1 = 5 4 9 3 | 10 2 8 7 | 6 1
O2 = 4 6 10 2 | 7 8 9 3 | 1 5

Partial Shuffle Mutation (PSM)

Mutation is another way how a population will get a new generation without involving cross-breeding between two parents, but only one parent. In one selected parent there will be a change in the composition of the genes that make up the chromosomes so that they turn into new chromosomes with different gene arrangements. Like crossovers, mutations also have various variants, including Insert Mutation, Creep Mutation, Uniform Mutation, Reversing Sequence Mutation, Interchanging Mutation, Flip Mutation, Partial Shuffle Mutation, Insert Mutation and others.

In this study, Partial Shuffle Mutation (PSM) will be used. This mutation, as the name implies, will partially rearrange the genes of a parent chromosome [10]. In the example below, for example, from the second to fifth genes of the parent chromosome, they will be rearranged randomly into new chromosomes.

Parent	**Child**
1 **2 3 4 5** 6	1 **5 3 2 4** 6

3. Findings

3.1 Problems found in research

When carrying out this research, a problem was found, namely that the number of teacher clusters must match the number of school clusters, meaning that the k value of the teacher cluster must be the same as the k value generated from school clustering. The number of school clusters found is 4, so k=4 is used to create a teacher cluster. The number of teachers for each cluster must be equal to the number of classes (6 x number of schools) in the cluster. So that the application of clustering for teachers is no longer looking for the value of k. The genetic algorithm will be applied in each cluster separately, so in each cluster there must be a certain number of schools or groups with the number of teachers.

3.2 How to solve problems found

To overcome the problems in the research above, the determination of teachers who enter certain clusters is carried out in stages starting from cluster 1 to cluster 4. After cluster 1 is found, the number of schools in cluster 1 for example S_1, it is found that the number of study groups is $6S_1$. This means that $6S_1$ teachers must be selected who will be placed in $6S_1$ classes or study groups. The way to determine which teachers are in cluster 1 is to use the distance between the position of the teacher's house and the centroid generated from the school clustering, then choose the $6S_1$ teacher closest to the centroid. Thus, the number of 6S1 groups is obtained and the number of teachers is $6S_1$ as well. Then the data (data of teachers and schools in cluster 1) will be arranged for placement using the Genetic Algorithm. Then in the same way it is determined the teachers who enter cluster 2 as many as $6S_2$. And so on until finally the distribution of teachers into clusters with the appropriate number is obtained. Thus the Genetic Algorithm can be applied in the four clusters. Each will produce an optimal configuration or arrangement of teacher placements, and when added together, the final result will be the total placement of all teachers with optimal conditions or a minimum total distance.

3.3 Research Implementation

The implementation is done using Python language and is run through the cloud facility from google at colab.research.google.com. Experiments were carried out by trying various combinations of input parameters, namely the maximum number of iterations and the probability of mutation compared to crossover. The results presented here are only a summary for the sake of brevity in the appearance of the results.

Table 1. Best fitness & Running time vs iteration

		iteration				
cluster No		50	40	30	20	10
1	best fitness	2261,172	2280,088	2263,909	2301,834	2301,834
	running time	166	133	100	96	33
2	best fitness	1701,424	1630,387	1687,102	1628,497	1707,996
	running time	66	52	40	26	12
3	best fitness	2852,490	2888,352	2897,002	2927,869	2916,105
	running time	167	132	99	66	32
4	best fitness	2526,333	2546,291	2496,736	2576,887	2576,887
	running time	93	75	56	37	19
Total	total fitness	**9341,418**	9345,117	9344,748	9435,087	9502,822
	total time	**492**	392	295	225	96

For the probability of mutation:crossover or pm = 1:10, the data is obtained as shown in table 1 above.

The best result, which is 9341.418 km, is obtained when iterations are carried out 50 times but must be paid for with the longest running time, which is 492 seconds. It can be seen in table 1 that if the number of iterations is changed (increased) then the change in the fitness value is not significant.

p-ISSN : 2774-9576
e-ISSN : 2747-2981

For the probability of mutation:crossover or pm = 1:15 the data is obtained as shown in table 2 below:

Table 2. Best fitness & Running time vs iteration

		iteration			
Cluster No		100	90	60	30
1	best fitness	2264,240	2256,937	2283,655	2273,099
	running time	442	400	260	130
2	best fitness	1671,048	1646,682	1628,497	1768,552
	running time	421	157	84	54
3	best fitness	2854,428	2870,372	2888,352	2892,777
	running time	438	400	260	129
4	best fitness	2517,935	2521,385	2459,422	2509,574
	running time	256	226	147	74
Total	total fitness	9307,651	9295,376	**9259,926**	9444,001
	total time	1557	1183	**751**	387

The best result is 9259.926 km obtained when iteration is done 50 times with a running time of 751 seconds. It is also seen that by changing the number of iterations, the change in the fitness value is not significant. But if you look at the two tables of observations above, it appears that by changing the probability of mutation:crossing from 1:10 to 1:15, there is a change in the value significant fitness.

If the process is carried out without clustering, purely using the Genetic Algorithm so that the process is carried out on all data (636 teachers) then a graph is obtained as shown in Figure 1 below, for the probability of mutation: interbreeding or pm = 1:10

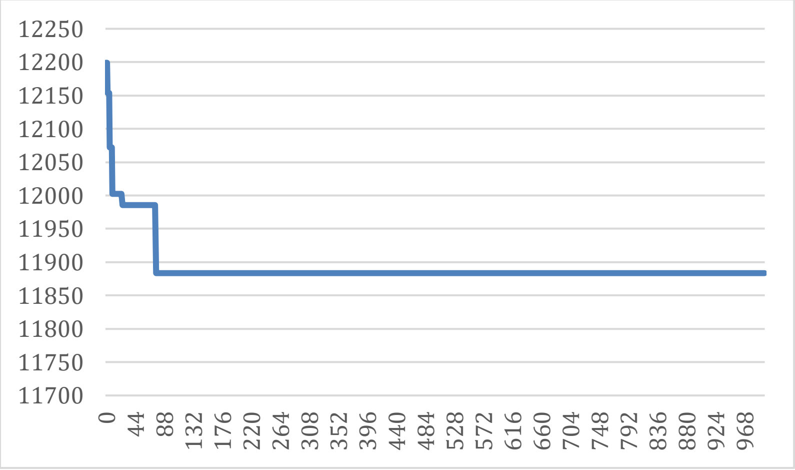

Figure 1. Iteration vs Fitness in pure GA with pm=1:10

Here the best results are obtained at the 76th iteration with fitness = 11883 with mutation probability:crossover or pm = 1: 10. And the recorded running time is very long, namely 5 hours 23 minutes 11 seconds, with a maximum iteration of 1000 times.

For the same experiment, without clustering, purely using only genetic algorithms, it was carried out on all data (636 groups) and pm = 1:20 obtained as follows:

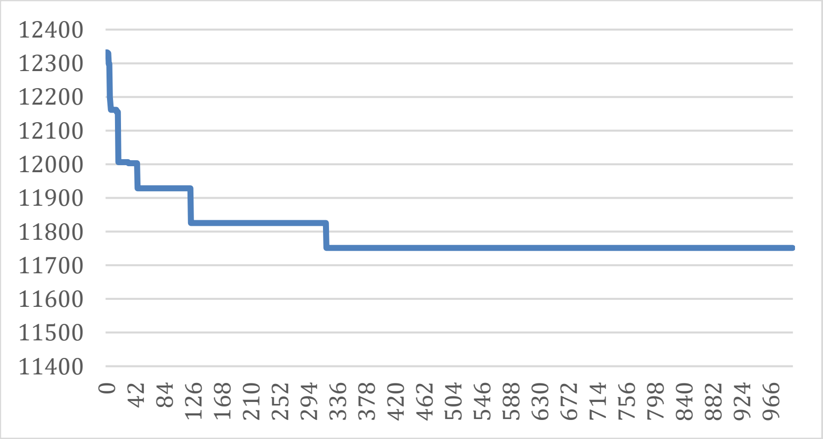

Figure 2. Iteration vs Fitness in pure GA with pm=1:20

Here the best results are obtained at the 321st iteration with fitness = 11751 with mutation probability: crossover probability = 1: 20. And the running time recorded is very long, namely 5 hours 12 minutes 31 seconds, with a maximum iteration of 1000 times.

In general, it can be seen that there is a very significant decrease in running time when using the clustering process first. This is very reasonable because the amount of data that is processed using genetic algorithms becomes less.

4. Conclusion

From the results of this study, as has been discussed, the following conclusions can be drawn:

a. It is better to solve the problem of teacher placement if clustering is carried out first on the initial data, then optimization is carried out using Genetic Algorithms. It is shown that by using clustering, the process can be carried out faster (in the order of minutes) than without clustering (in the order of hours). In addition, the total distance obtained by using clustering is shorter than without clustering.

b. The higher the iteration (at max = 100 iterations), the better the fitness results, but this does not apply to max = 1000 iterations. There is a certain iteration point where the results obtained are already convergent and better results cannot be obtained. Thus, to determine the maximum iteration, an experiment should be carried out first so that the optimal maximum iteration number can be obtained.

c. Significant changes/improvements in fitness occur more frequently at the time of mutation than during crossover. This reinforces the theory that mutations can release the search trap at the local optimum.

References

[1] Burhanudin, 2005. Kinerja Guru. *Andi Offset. Yogyakarta*.

[2] Nyakundi, G.M., 2018. Influence of Teacher Performance on Learning Achievement in Public Secondary Schools in Kisii County, Kenya. *International Journal on Education 10 (2):21*.

[3] Sriwindono, H., Rosa, P.H.P, Polina, A.M., Nugroho, R.A. 2017. The Model of Elementary Teacher School Placement in Magelang District by Using Genetic Algorithm. *Proceeding of Computer System and Artificial Intelligence (CSAI) International Conference*. Asociation for Computing Machinery.

[4] Rosa, P.H.P, Sriwindono, H., Nugroho, R.A., Polina, A.M. and Pinaryanto, K.. 2020. Comparison of Crossover and Mutation Operators to Solve Teachers Placement Problem by Using Genetic Algorithm. *Journal of Physics: Conference Series; Volume 1566; No 1.*

[5] Gothai, E., Balasubramanie, P. 2012. An Efficient Way for Clustering using Alternative Decision Tree. *American Journal of Applied Science, 9, 531-534*.

[6] Han, J., Kamber, M. 2012. Data Mining: Concepts and Techniques (4th ed.). San Francisco: *Morgan Kaufmann Publishers.*

[7] Holland, J.H. 1992. Adaptation in Natural and Artificial Systems. *MIT Press*

[8] Goldberg, D. 1989. Genetic Algorithms: in Search, Optimization, and Machine Learning. *Addison Wesley.*

[9] Kora, P and Yadlapalli, P. 2017. Crossover Operators in Genetic Algorithms: A Review. *International Journal of Computer Applications (0975 – 8887) Volume 162 – No 10,*

[10] Othman, A., Abouchabaka, J and Tajani, C. 2012. Hybridizing PSM and RSM Operator for Solving NP-Complete Problems: Application to Travelling Salesman Problem. *International Journal of Computer Science Issues.*