International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. .., No... (Mon-20..)

http://dx.doi.org/10.12785/ijcds/XXXXXX

IWOKM-GA Hybrid Method To Improve Clustering Accuracy In Banking Data

Ni Luh Gede Pivin Suwirmayanti[1,2], I Ketut Gede Darma Putra[3], Made Sudarma[4], I Made Sukarsa[3] and Emy Setyaningsih[5]

[1]*Department of Computer Systems, Faculty of Informatics and Computer, Institut Teknologi dan Bisnis STIKOM Bali, Bali, Indonesia*
[2]*Faculty of Engineering, Udayana University, Bali, Indonesia*
[3]*Department of Information Technology, Faculty of Engineering, Udayana University, Bali, Indonesia*
[4]*Department of Electrical Engineering, Faculty of Engineering, Udayana University, Bali, Indonesia*
[5]*Department of Computer Systems Engineering, Universitas AKPRIND Indonesia, Yogyakarta, Indonesia*

Received Mon. 20, Revised Mon. 20, Accepted Mon. 20, Published Mon. 20

Abstract: Clustering is one of the critical approaches in data mining, which aims to divide and group data into groups that have similar characteristics. Some of the main problems in clustering are grouping with high-dimensional datasets that have many attributes, both numerical and categorical data types, high time consumption, calculation complexity, and overhead, which makes some algorithms in the clustering process less efficient. The clustering algorithm often used is K-Means, but the algorithm needs to improve in the computational method that is quite long. The results of grouping data on K-Means must be defined first, as well as noise or outliers, due to outliers in grouping results and difficulties in finding global solutions that can reduce the quality of clustering results on the K-Means algorithm. Therefore, this research is focused on developing the K-Means Algorithm to improve model performance as well as the quality of the resulting clusters by combining the K-Means (KM) method with Invasive Weed Optimization (IWO), and Genetic Algorithm (GA) called the Hybrid IWOKM-GA method to produce data clustering with close genetic diversity. The results showed that the Hybrid IWOKM-GA method managed to find the best clustering results with a Cost Function Value value of 2400.51, almost three times when compared to the K-Means model combined with GA, which has a computational time of 328.08 seconds

Keywords: Clustering, Genetic Algorithm, IWO, K-Means

1. Introduction

Data mining is one of the essential approaches in clustering methods, which aims to divide and categorize data into groups with similar characteristics. In today's significant data era, everything generates large amounts of data from various sources, such as smart home devices, mobile devices, or data from business and financial transactions. The application of data mining models can be seen in the health field and can have a significant impact [1], [2], besides that, it is applied in the world of education, such as to extract individual student learning achievements from course information [3]. Suppose the data containing many variables generated in this process is further processed using the proper method. In that case, it will help entrepreneurs or other actors organize data into the desired categories.

The application of data mining models can be seen in the health field and can have a significant impact; besides that, it is applied in the world of education, such as to extract individual student learning achievements from course information. Suppose the data containing many variables generated in this process is further processed using the proper method. In that case, it will help entrepreneurs or other actors organize data into the desired categories.

The application of data mining models can be seen in the health field and can have a significant impact; besides that, it is applied in the world of education, such as to extract individual student learning achievements from course information. Suppose the data containing many variables generated in this process is further processed using the proper method. In that case, it will help entrepreneurs or other actors organize data into the desired categories.

At this point, the categorization of text data should be essential for data mining activities [4], [5], [6]. In multidimensional data processing, clustering can be used to group and divide data effectively so that the data owned

E-mail address: pivin@stikom-bali.ac.id, suwirmayanti.2391011009@student.unud.ac.id

can be grouped according to similarities. Clustering is an unsupervised learning model on datasets that will be grouped based on Euclidean values that show parameters of similarity between one data and another [7], [8]. The processes in grouping, broadly speaking, are grouped into two processes: the hierarchical method and the partition method. In hierarchical methods, data is formed in a tree model called a dendrogram in the form of levels in grouping results. Because the results obtained are considered natural and complicated in the reading process, this process is often not used in grouping data that will be used for commercial purposes [4], [5], [6].

In another process, namely the partitional clustering process, data grouping is done by an iteration process. At first, the data is divided into several parts, and then the partitioning process is engineered in such a way and transformed into a new data cluster, approaching the Euclidean distance value. The process performed on clustering can be used as a function of the K-Means algorithm. The K-Means algorithm itself is an algorithm that is currently still popularly used for the clustering process. This algorithm has advantages because of its simplicity in the process of grouping data; among others, this process has low complexity, and the growth of this algorithm is linear [7]. On the other hand, K-Means has several disadvantages, including a reasonably long computing process, the data grouping results in K-Means must be defined first, and noise or outliers in the grouping results, which can reduce the quality of the clustering results [7], [9], [10], [11], [12].

Research is currently underway to address the weaknesses of the K-means algorithm. The focus is on developing the algorithm to enhance its performance and the quality of the resulting clusters [13], [14]. Improving cluster quality includes proposing a method to obtain the initial centroid value to improve class label validation [10], [15], [16]. Meanwhile, to overcome K-Means' inability to find global solutions, some researchers use the Invasive Weed Optimization (IWO) method [6], [7], [17], [18]. The selection of IWO was inspired by weed plants that can live scattered and then able to colonize the process on overgrown land. This algorithm approximates the distribution of random values at a wide point, the distribution of random values carried out based on normal functions for the centroid distribution.

IWO is often combined with other algorithms for optimization purposes, such as combining IWO with Gray Wolf Optimization (IWOGWO) [18]. IWOGWO aims to improve clustering algorithms' ability to find better solutions in complex search spaces. In addition, some researchers also propose merging IWO with K-Means (IWOKM) [6], [17], [19] which aims to improve the quality of the clustering algorithm by utilizing a more optimal initial centroid search process. Therefore, the IWOKM algorithm shows better results than ordinary models in IWO regarding accuracy and speed in generating convergence. In addition, IWO is also combined with other algorithms, such as Data Envelopment Analysis (DEA) and K-Means, which can improve process efficiency and shorter execution time [7]. The study used DEA to assess the efficiency of existing facilities, then combined it with the K-Means method to identify efficient site clusters, and finally utilized the IWO algorithm to optimize the location of new facilities to be placed. Although IWO can improve clustering algorithms, it performs poorly when applied to clustering problems for high-dimensional datasets. This is due to how invasive weeds spread in nature [20], which limits the algorithm's ability to handle high-dimensional search spaces. One algorithm that can be used to overcome these problems is the GA [10].

GA has characteristics that can be used for optimization and increasing speed and accuracy. In terms of cluster optimization, GA can help overcome the problem of random initial centroid selection in K-Means and result in better and more optimal clustering [8]. Optimization can also be done using the Multi-Objective Genetic Algorithm (MOGA) with categorical data clustered using the K-Means algorithm to minimize intra-cluster distances and maximize distances between clusters to lower the clustering error rate [14]. GA ability in cluster determination is also seen in the SOMI-GANB hybrid model research that uses GA in the feature selection process to improve the performance of the Naïve Bayes algorithm. The method can group attribute values and achieve high accuracy. This research shows that using GA helps optimize the number of attribute clusters, improving accuracy and addressing problems that often occur when faced with incomplete data [21]. Like IWO, GA development based on Swarm Optimized Clustering in different topics aims to select high-dimensional data features that can improve system efficiency and accuracy [22].

Based on this background, this study proposes using a combination of IWOKM optimized using GA. The selection of IWO was inspired by weed plants that can live scattered and then able to colonize the process on overgrown land. This algorithm approximates the distribution of random values at a wide point, the distribution of random values performed based on normal functions for the centroid distribution. After the process of initialization and formation of cluster groups from each data, the role of GA in the proposed study is to maintain genetic diversity from one generation to another, ensure the best population crossover, and eliminate chromosomes. Suppose the process with GA has succeeded in obtaining clusters with the best genetic diversity. In that case, the iteration process at IWOKM will be carried out until a data cluster with the closest Euclidean distance is obtained. The superiority of the IWOKM GA algorithm will later be verified by conducting experiments on high-dimensional banking datasets.

This paper's composition continues with Part 2 describing the proposed model, followed by Part 3 containing the results of the analysis and discussion of the proposed model, and finally, Part 4 containing the conclusions of this paper.

2. Research Methods

Figure 1 shows the proposed model carried out by combining IWOKM with GA optimization. This study focuses on combining IWOKM with GA optimization to produce data groupings with close genetic diversity. It also seeks to optimize the computational time used in grouping data and the resulting diversity compared to conventional K-means methods.

In the evolution of data grouping. This scheme uses the combination to determine the computational time produced and the results of clustering the data obtained. The process begins with initializing variable components used in the GA process: generation, chromosomes, and population. Then, the process continues with the initialization of IWOKM, which determines the group of each data owned. After the previous method is carried out, the next step is the determination of the Fitness Counting Function, which determines the fitness value produced by each individual based on the parameters that have been initialized. The step continues with a crossover process to move from the best populations one and 2. After this, a mutation process is held to maintain the genetic diversity of one generation.

When a chromosome value from GA cannot be used, a process of elimination is carried out to obtain generation with close diversity. When the GA process has succeeded in producing a population with a generation with the best gene affinity, the process will be completed; otherwise, the IWOKM will be carried out.

The pseudocode of the proposed method is shown in Figure 1 is presented in Figure 2.

A. K-Means Clustering

K-Means Clustering is a method of grouping data with K objects as centroids on each cluster. The selection of centroids must be considered carefully, so that no data outliers and clustering can cover the data as a whole. The first step in the K-Means process is to select a centroid point as far away as possible from other centroids. Then, after the centroid determination process is complete, a distance calculation process will be carried out which causes the data to be grouped on one of the corresponding centroids [16]. The K-Means Clustering equation can be seen in Eq.(1).

$$j = \sum_{j=1}^{k} \sum_{i=1}^{n} \left| X_i^{(j)} - c_j \right|^2 \quad (1)$$

The K-Means Clustering flow can be seen in Figure 3. The algorithm accepts the number of clusters to group data into, and the dataset to be clustered as input values. Specifies the number of clusters. Each cluster formed will be calculated for its average value. The average of a cluster is the average of all records contained in that cluster. Allocate data into clusters randomly. Calculate the centroid / average of the data in each cluster Allocate each data to the nearest centroid/average Repeat the previous steps until stable clusters are formed and the K-Means procedure is complete. A stable cluster is formed when iterations or iterations of K-Means do not create a new cluster as the center of the cluster or the arithmetic mean values of all new clusters are equal to the old cluster.

B. Invasive weed algorithm (IWO)

IWO is an invasive algorithm inspired by the ability of weed plant colonization in host plants; the realization of the invasion process in IWO begins with random population initialization, and then the reproduction process is carried out depending on the fitness value, thus ensuring linear growth [7]. In IWO, a spatial dispersal process causes seeds to disperse with normal distribution. The following process is competitive exclusion, which is the selection process of developed seeds; seeds with low fitness value will not be able to develop, so they will naturally be eliminated. It is such a process that the optimal solution can be obtained quickly.

Swarm intelligence-based IWO is an optimization technique that can steer the search process through intra-group competition and collaboration. IWO algorithms are population-based global search strategies as opposed to evolutionary algorithms. The robust, aggressively growing crops that seriously endanger cultivated plants are referred to as weeds. Historical data spanning thousands of years suggests that weeds are winners: Herbicides have been used for more than 60 years, yet weeds are still prevalent everywhere; agriculture and human hands weeding first existed thousands of years ago, and weeds are still present everywhere. The potential of weeds to self-transform and survive is demonstrated by the rise in anti-herbicide weeds in recent years [17]. The IWO algorithm is presented in Figure 4.

1) Population Initialization Phase

The initialization phase is the initial phase of the IWO process. It establishes a population size (n) that represents the number of potential solutions. Initialize the initial population with random solutions in the search space. The initial center point, which is also a candidate for the initial solution determined at random, will be the mother grass in the IWO process, which will then produce child grass and spread in the search area.

2) Reproduction Phase

Based on availability, each seed grows or develops into a flowering plant according to its seed fitness. The ability of a plant population to produce seeds varies based on the best and lowest colony fitness. The number of seeds produced by each plant increases from the minimum value (S_{min}) to the maximum value (S_{max}). The linear growth of a plant can be determined by examining the seeds of the highest fitness individual in the colony

This phase encompasses a series of intricate stages, including the calculation of the fitness value of each cluster, the determination of the potential seed production of each

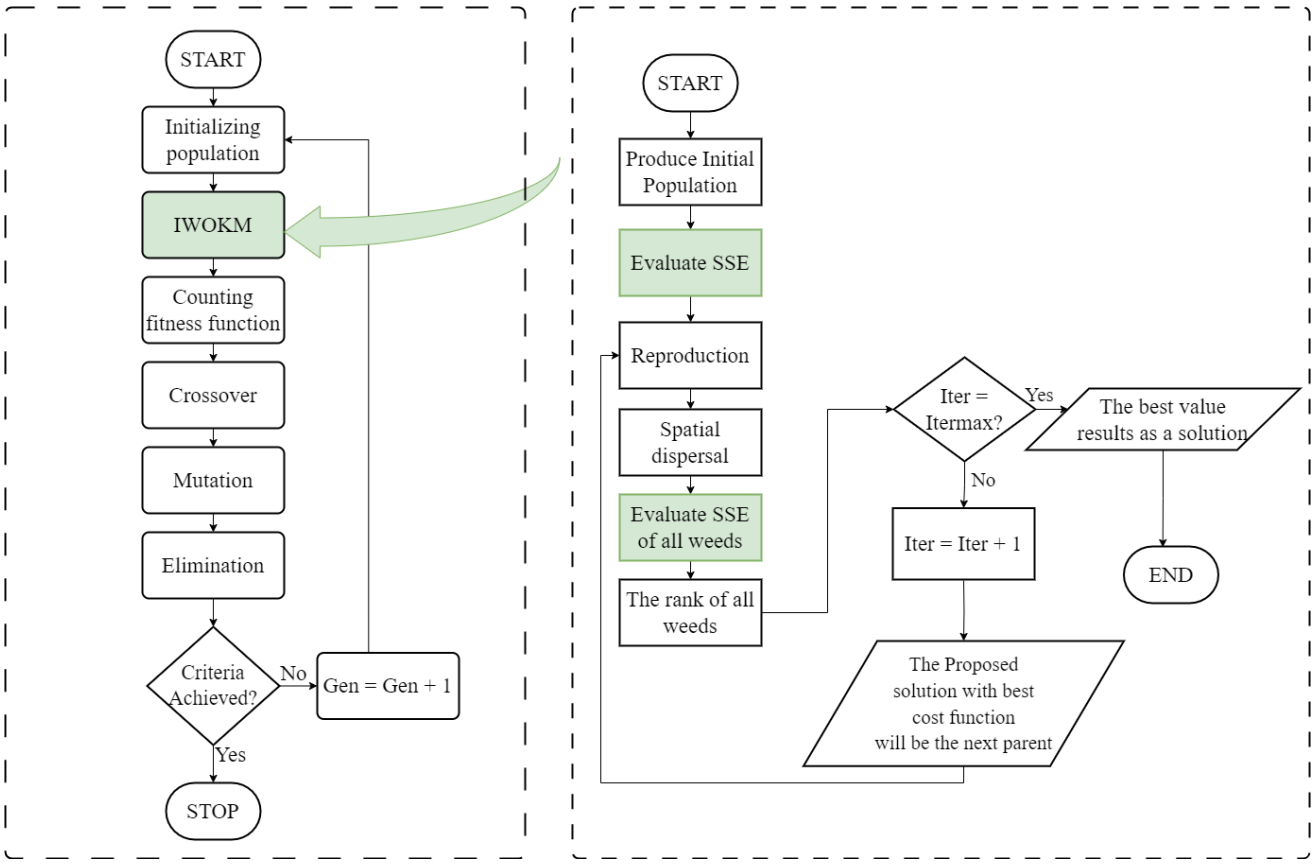


Figure 1. Proposed method

parent grass, and the final decision on the distribution of the offspring grass. The fitness function used to determine the potential seed production of each parent grass is the Mean Squared Error value of each cluster, a complex yet comprehensive measure.

A plant is allowed to produce seeds depending on its fitness value. The number of seedlings that each plant can produce increases linearly from the smallest possibility to the most significant possibility. The equation for determining the amount of child grass that will be made by each parent grass as conveyed by [7] is as shown in Eq.(2).

$$n_{seed} = \frac{F_i - F_{worst}}{F_{best} - F_{worst}} (S_{max} - S_{min}) + S_{min} \qquad (2)$$

Where:

- n_{seed} represents number of child grasses to be produced
- F_i indicates fitness value from grass to I
- F_{worst} signifies worst fitness score in grass colony
- F_{best} denotes best fitness value in grass colony
- S_{max} is maximum amount of daughter grass that the parent grass can produce
- S_{min} is minimum amount of daughter grass that the parent grass can produce

3) Spatial Dispersal Phase

The resulting seeds are randomly dispersed in a distribution with an average equal to zero and different variances in the search space so that they can approach the variance of the parent and grow into new plants. For standard deviation, σ, for the initial function on a predetermined basis. At every stage or iteration, reduced to the final value iteration. Calculation of the spatial dispersal phase using Eq.(3).

$$\sigma_{iter} = \left(\frac{iter_{max} - iter}{iter_{max}} \right)^n \left(\sigma_{iter} - \sigma_{final} \right) + \sigma_{final} \qquad (3)$$

Where σ_{iter} is the standard deviation value at the current time, $iter_{max}$ is the maximum number of iterations, n is the nonlinear modulation index, the initial standard deviation value, and the final standard deviation value.

```
1. Initialize parameters:
      -Population size (N)
      - Maximum number of generations (G)
      - Maximum number of seeds (Smax)
      - Minimum number of seeds (Smin)
      - Standard deviation (σ)
      - Number of clusters (k)
      - Crossover and mutation rates for GA
      - Maximal Iteration (I)
      - Population Min and Max for IWO (PopMin n PopMax)
2. Generate initial population (P) of weeds:
   FOR each weed in P:
       Initialize weed's position with random values (centroids for k clusters)
   ENDFOR
3. Evaluate the fitness of each weed using a fitness function (e.g., sum of squared errors in
   clustering).
   FOR each weed in P:
       Apply K-Means clustering using weed's position as initial centroids.
       Calculate SSE for the clustering result.
       Set weed's fitness as negative SSE (since lower SSE indicates better fitness).
   ENDFOR
4. WHILE Iteration < I DO:
      - Reproduction:
        FOR each weed in P:
            Calculate the number of seeds based on weed's fitness:
            num_seeds = Smin + ((Smax - Smin) * (weed_fitness - worst_fitness) / (best_fitness -
             worst_fitness))
                FOR each seed:
                    Create a new position for the seed by adding Gaussian noise to the weed's position
                ENDFOR
        ENDFOR
      - Combine parent weeds and their seeds into a new population P_new.
      - Evaluate the fitness of each individual in P_new.
         Sort P_new based on fitness.
      - Select the top N individuals to form the new population P.
   END WHILE

5. Return the best solution found (the weed with the highest fitness).
6. Genetic Algorithm Operations:
        FOR each pair of individuals in P:
             Perform crossover with a certain probability to create offspring.
             Perform mutation on the offspring with a certain probability.
        ENDFOR
        Evaluate the fitness of the offspring.
        Combine the offspring with the current population P.
        Sort the combined population based on fitness.
        Select the top N individuals to form the new population P.
   Function fitness_function(weed_position):
   - Apply K-Means clustering using weed_position as initial centroids.
   - Calculate the clustering cost (e.g., sum of squared distances from points to their respective
     cluster centroids).
   - Return the negative clustering cost as fitness (since lower cost means better fitness).
Function crossover(parent1, parent2):
   - Implement crossover operation to generate offspring from parent1 and parent2.
Function mutation(individual):
   - Implement mutation operation to introduce variations in the individual's position.
```

Figure 2. Pseudocode of the Proposed method

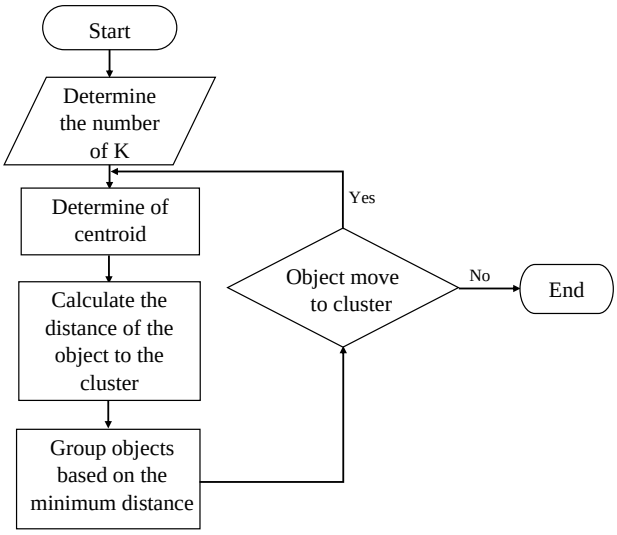


Figure 3. K-means clustering flow

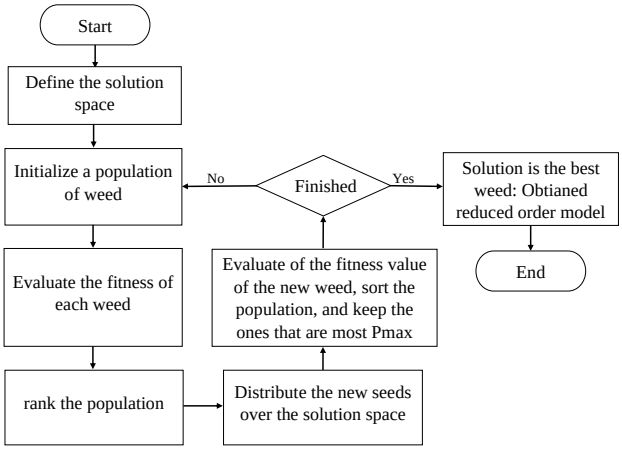


Figure 4. Invasive Weed Optimization (IWO) algorithm flow

4) Competitive Exclusion Phase

Reproduction continues to reach the maximum number of plants (P_{max}). Only plants with higher fitness can survive and produce seeds, and the rest are eliminated. The above steps continue to be carried out until maximum results are achieved in the number of plants and plants with the highest fitness. This is the shortest way to reach the optimal solution.

C. Genetic Algorithm (GA)

GA are part of the metaheuristic algorithms of evolutionary computational processes that adapt the laws of biology, namely Darwin's theory of the evolution of living things, which, in theory, is considered an adaptive heuristic search process that focuses on evolution and selection in producing generations with close diversity [14], [23]. This algorithm is regarded as one of the best methods for optimizing complex data based on mutations and natural selection. This optimization combines the best individuals in each cluster to produce a higher probability of better fitness outcomes than proposed before. The processes on GA utilize crossover, mutation, and elimination processes to create generations with close diversity that are still considered the best individuals. This algorithm has its advantages in helping K-Means grouping in the case of local minima because it has the advantage of optimizing cluster centers [23]. The GA algorithm is presented in Figure 5.

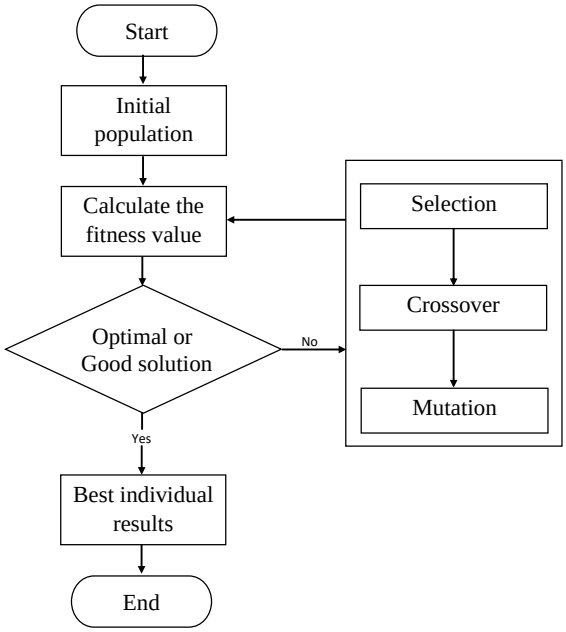


Figure 5. Genetic algorithm flow

The steps of the GA flow according to Figure 5 include:

1) Generating the initial population: To find an initial solution, this method involves randomly generating the initial population.
2) To find an initial solution, this first population is produced at random.
3) Fitness evaluation: This procedure involves figuring out each chromosome's fitness value and then assessing each population until the stop criteria are satisfied.
4) An Individual performance is assessed using a specific function as a basis for evaluation.
5) Individuals with low fitness values would have perished during the course of nature's evolution.
   a) Selection: The process of choosing who will be chosen for crossovering is known as the selection process.
   b) Crossover: Increasing string variety in a population is the goal of the crossover process.
   c) Mutation: The process of altering the value of one or more chromosomal genes is known as mutation.

6) The objective of the procedure is to reach the stop criteria, which are utilized to end the GA process. The outcome is the best answer that GA could find.

3. Result and Discussion

This section contains the results of the IWOKM-GA experiment using the German Credit Dataset from Hamburg University, Germany. The first step in the experiment is initializing its parameters, i.e., maximum generation values and total population, to get the best cost function. After the value of the cost function is obtained by generating variations in its parameters, the computational time to produce this value is also recorded, taking into account the value of the best cost function at the lowest computational time.

Each variation of the Test Class has a Cost Function value and results in its Computational time. The cost function value here will be a mean squared error (MSE) value to measure the quality of prediction results when applying the IWOKM-GA method. Measure the value of effectiveness later by measuring computational time at various stages of the iteration process that is needed and can be viewed thoroughly. It can be concluded that the best value of the cost function by considering the Computing Time, which is a cost function with the maximum generation size and total population, are both filled by the value of 5 with a score of 2441.91. The smaller the cost function value, the better the model, but because the distance between the cost function value is not too considerable between the smallest cost function and the selected model and the consideration of the computation time obtained, the cost function with a value of 2441.91 was chosen as the best result.

Figure 6 shows the Test Results I carried out with a variation in the number of iterations as much as 5, with the best cost from the range of 2300 – 2600. This test uses MaxG5_PopSize5 variation, which means that the maximum generation that can be produced is filled by a value of 5, and the population size is filled by a value of 5. MaxG5_PopSize10 means the maximum generation that can be produced is filled by a value of 5, and the population size is filled by 10. MaxG5_PopSize20 means the maximum generation that can be produced is filled by a value of 5, and the population size is filled by a value of 20. Based on Figure 6, the highest scores obtained use the MaxG5_PopSize20 variation.

The second stage is carried out with a variation in the number of iterations by ten, shown in Figure 7. This parameter goes hand in hand with other predefined parameters. This test uses MaxG10_PopSize5 variation, which means that the maximum generation that can be produced is filled by a value of 10, and the population size is filled by a value of 5. MaxG10_PopSize10 means the maximum generation that can be produced is filled by a value of 10, and the population size is filled by 10. MaxG10_PopSize20 means the maximum generation that can be produced is filled by

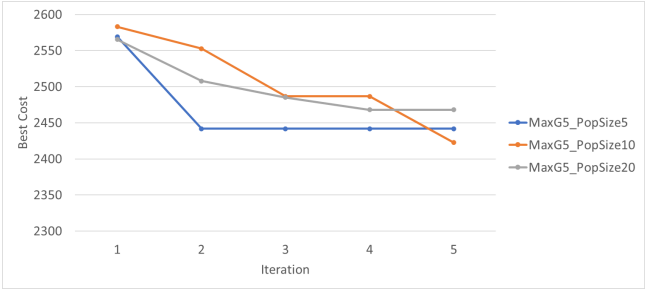


Figure 6. Trial results I

a value of 10, and the population size is filled by a value of 20. Based on Figure 7, the highest scores obtained use the MaxG10_PopSize20 variation.

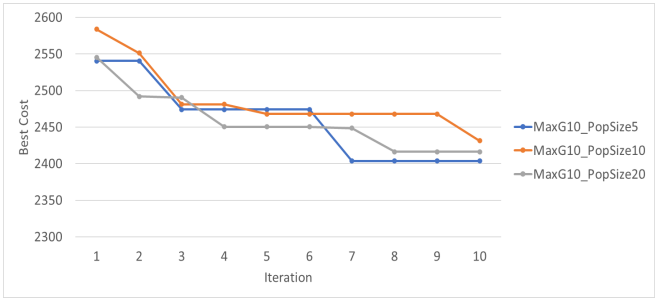


Figure 7. Trial results 2

The third stage is carried out with a variation in the number of iterations by 20, shown in Figure 8.

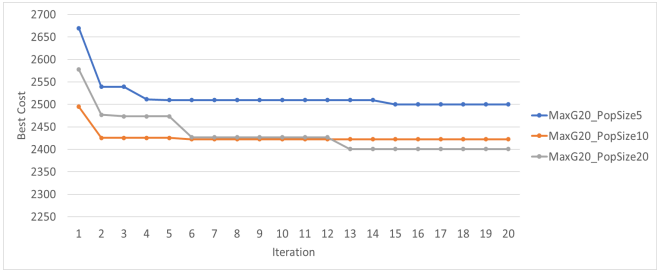


Figure 8. Trial results 3

This parameter goes hand in hand with other predefined parameters. This test uses MaxG20_PopSize5 variation, which means that the maximum generation that can be produced is filled by a value of 20, and the population size is filled by a value of 5. MaxG20_PopSize10 means the maximum generation that can be produced is filled by a value of 20, and the population size is filled by 10. MaxG20_PopSize20 means the maximum generation that can be produced is filled by a value of 20, and the population size is filled by a value of 20. The cost function

value distribution is at 2400 – 2500, with the lowest value at 2400.51 and the highest at 2499.85. K-Means Results Combined with GA This section describes the results of clustering experiments using K-Means optimized by GA. The following data are presented as a result of experiments with two optimized parameters: the maximum generation value and population size.

Based on the Table I presented as a result of this experiment, the maximum number of generations and population, both filled with a value of 20 in order to get the best cost function result, taking into account the computational time generated. This result was chosen as a result of the cost function values being too far apart when compared to the model with the fastest computation time.

TABLE I. Comparison of experiment results

Parameter Variation Experiment	IWOKM		IWOKM+GA	
	Cost Function Value	Compute Time	Cost Function Value	Compute Time
MaxG5_PopSize5	2441.91	20.74	6825.13	0.12
MaxG5_PopSize10	2422.79	51.96	6684.24	0.25
MaxG5_PopSize20	2467.96	88.77	6512.86	0.48
MaxG10_PopSize5	2403.83	43.43	6634.39	0.25
MaxG10_PopSize10	2431.47	81.10	6509.82	0.45
MaxG10_PopSize20	2416.39	159.68	6515.44	0.95
MaxG20_PopSize5	2499.88	79.81	6709.58	0.46
MaxG20_PopSize10	2422.44	163.10	6478.72	0.89
MaxG20_PopSize10	**2400.51**	**328.08**	**6411.39**	**1.90**

Figure 9 shows the results of the IWOKMGA model have the lowest cost function results with a value of 2400.51, almost tripled or can be stated to decrease nearly 60% from a value of 6411.39 to a value of 2400.51, when compared to the K-Means model combined with the GA because of the combined advantages of IWO to model convergence and the ability of the GA to produce the best generation along with the process of iterating into convergence. On the other hand, this model is relatively slow for computational time, with almost ten times larger computational time compared to K-Means models combined with GA. Based on these results, the IWOKMGA model had a lower cost function value than K-Means combined with the GA.

Figure 9 visualized using a graph, shows the comparison of the overall value of the Cost Fungtion Value between IWOKMGA compared to KM+GA, where the cost function results show that IWOKMGA is able to have a lower Cost Fungtion value than KM+GA as a whole. So, from the Cost Function Value results, the performance of the IWOKMGA algorithm shows a significant decrease in the cost function when the parameters are varied. This indicates that IWOKMGA is able to adjust well to parameter changes and produce more optimal clustering than the KM+GA algorithm. In the iteration process, especially when the number of clusters (k) increases, IWOKMGA consistently provides the lowest cost function value. The performance of the KM+GA algorithm shows a decrease in the cost function, but the decrease is slower and better than the IWOKMGA algorithm.

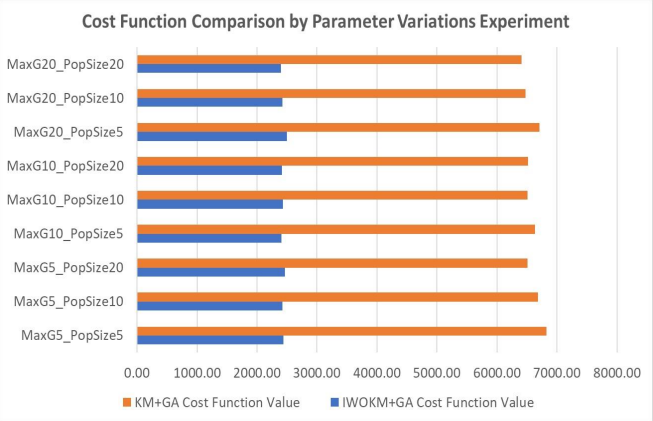


Figure 9. Cost function comparison

Figure 10 presents the performance comparison of the two methods, IWOKMGA and KM+GA. This experiment focuses on the variation of the MaxG5 parameter with the population sizes chosen: popsize5, popsize10, and popsize20. The results of each experiment show that the cost function value of KM+GA is always in the range of $[6512.86, 6825.13]$, which is much larger than the cost function of IWOKMGA with a value in the range of $[2422.79, 2467.96]$.

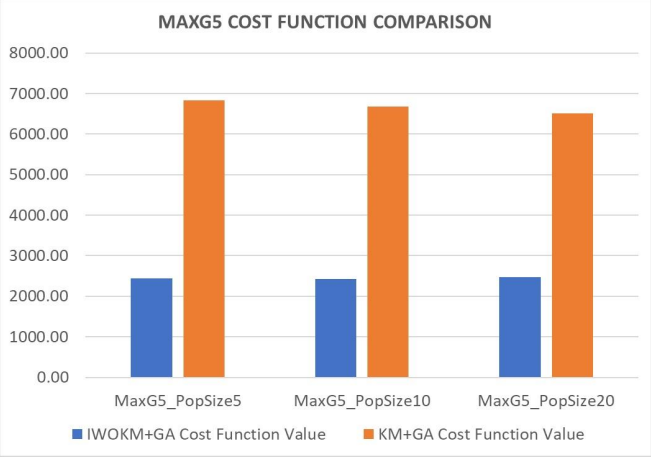


Figure 10. MaxG5 Cost function comparison

Figure 11 presents a comparison of the results of the MaxG10 parameter variation with the selected population size popsize5, popsize10, and popsize20. The results show that the cost function value of KM+GA has a minimum value of 6509.82 and a maximum value of 6634.39, where the minimum value of KM+GA is much greater than the cost function of IWOKMGA with a minimum value of 2403.83 and a maximum value of 2431.47.

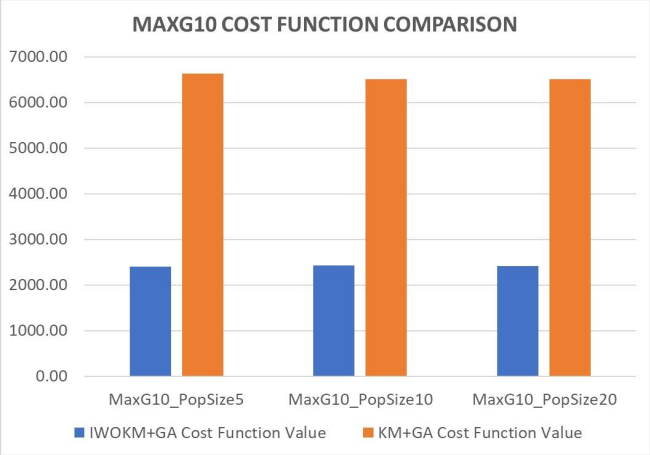


Figure 11. MaxG10 Cost function comparison

Figure 12 presents a comparison of the results of the MaxG10 parameter variation with the selected population size popsize5, popsize 10, and popsize20. The results show that the cost function value of KM+GA has a minimum value of 6509.82 and a maximum value of 6634.39, where the minimum value of KM+GA is much greater than the cost function of IWOKMGA with a minimum value of 2403.83 and a maximum value of 2431.47.

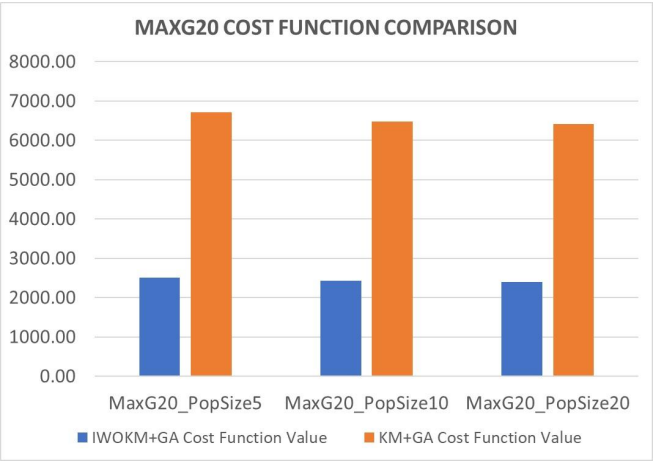


Figure 12. MaxG20 Cost function comparison

Figure 13 shows the results of the IWOKMGA model have the lowest cost function results with a value of 2400.51, almost tripled or can be stated to decrease nearly 60% from a value of 6411.39 to a value of 2400.51, when compared to the K-Means model combined with the GA because of the combined advantages of IWO to model convergence and the ability of the GA to produce the best generation along with the process of iterating into convergence. On the other hand, this model is relatively slow for computational time, with almost ten times larger computational time compared to K-Means models combined with GA. Based on these results, the IWOKMGA model had a lower cost function value than K-Means combined with the GA.

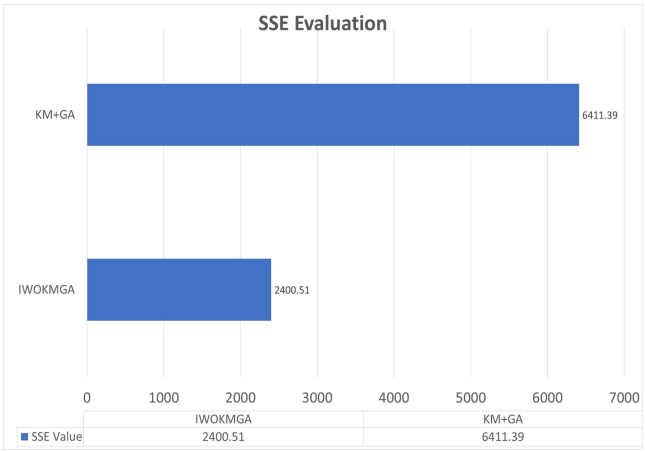


Figure 13. Cost function value comparison results

4. Conclusion

This research mainly focuses on combining Improved Weighted Optimized K-Means (IWOKM) clustering with Genetic Algorithm (GA). The dataset used is banking data, and the IWOKM combination algorithm is used to process large quantitative data sets using GA optimization methods. This combination results in a grouping of data with close genetic diversity. The research also seeks to optimize the computational time used in the data grouping process and improve the resulting diversity. The results showed promising outcomes where the GA optimization process helped IWOKM Clustering to find the best clustering results with a Value Cost Function of 2400.51, almost three times better compared to the K-Means model combined with GA, with a computational time of 328.08 seconds. This indicates that the IWOKMGA approach is not only effective in producing high-quality clusters but also efficient in terms of computation time.

Based on the promising results of this study, several areas for future research can be identified to further improve this method: Exploring hybrid models that combine IWOKMGA with machine learning or other optimization techniques to further improve clustering results and computational efficiency. This future work will help to further refine the IWOKMGA algorithm and explore its full potential, making it a powerful tool for clustering data sets across various application domains.

References

[1] S. A. Diwani and Z. O. Yonah, "Holistic diagnosis tool for prediction of benign and malignant breast cancer using data mining techniques," *International Journal of Computing and Digital Systems*, no. 1, pp. 417–432, apr 2021.

[2] A. Abed Mohammed, P. Sumari, and K. Attabi, "Hybrid K-means and Principal Component Analysis (PCA) for Diabetes Prediction," *International Journal of Computing and Digital Systems*, vol. 15, no. 1, pp. 1719–1728, jun 2024.

[3] F. Albalooshi, H. AlObaidy, and Ghan, "Mining Students Outcomes: An Empirical Study," *International Journal of Computing and Digital Systems*, vol. 8, no. 3, pp. 229–241, jul 2019.

[4] M.-C. Chiang, C.-W. Tsai, and C.-S. Yang, "A time-efficient pattern reduction algorithm for k-means clustering," *Information Sciences*, vol. 181, no. 4, pp. 716–731, feb 2011.

[5] T. Muthamilselvan and B. Balusamy, "Hybrid Approach for Data Classification in E-Health Cloud," *International Journal of Intelligent Engineering and Systems*, vol. 10, no. 3, pp. 75–84, jun 2017.

[6] C. Fan, T. Zhang, Z. Yang, and L. Wang, "A Text Clustering Algorithm Hybriding Invasive Weed Optimization with K-Means," in *2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom)*. IEEE, aug 2015, pp. 1333–1338.

[7] F. Faezy Razi, "A hybrid DEA-based K-means and invasive weed optimization for facility location problem," *Journal of Industrial Engineering International*, vol. 15, no. 3, pp. 499–511, sep 2019.

[8] M. A. Nemmich, F. Debbat, and M. Slimane, "A Data Clustering Approach Using Bees Algorithm with a Memory Scheme," in *Advances in Computing Systems and Applications*. Cham: Springer International Publishing, 2019, pp. 261–270.

[9] S. Katoch, S. S. Chauhan, and V. Kumar, "A review on genetic algorithm: past, present, and future," *Multimedia Tools and Applications*, vol. 80, no. 5, pp. 8091–8126, feb 2021.

[10] E. Setyaningsih, N. Hidayat, U. Lestari, and A. Septiarini, "Modification of K-Means and K-Mode Algorithms To Enhance the Performance of Clustering Student Learning Styles in the Learning Management System," *ICIC Express Letters*, vol. 17, no. 1, pp. 49–59, 2023.

[11] S. Ahmed, "Data Mining for Non-Redundant Big Data Using dynamic KMEAN," *International Journal of Computing and Digital Systems*, vol. 14, no. 1, pp. 243–251, jul 2023.

[12] K. Dissanayake, M. G. M. Johar, and N. H. Ubeysekara, "Data mining techniques in disease classification: Descriptive bibliometric analysis and visualization of global publications," *International Journal of Computing and Digital Systems*, vol. 13, no. 1, pp. 289–301, 2023.

[13] I.-D. Borlea, R.-E. Precup, and A.-B. Borlea, "Improvement of K-means Cluster Quality by Post Processing Resulted Clusters," *Procedia Computer Science*, vol. 199, pp. 63–70, 2022.

[14] Y. Liu, J. C. Chai, X. Cui, W. Yan, N. Li, and L. Jin, "Multi-objective optimization of air dehumidification membrane module based on response surface method and genetic algorithm," *Energy Reports*, vol. 9, pp. 2201–2212, dec 2023.

[15] Y. Arkeman, N. A. Wahanani, and A. Kustiyo, "Clustering k-means optimization with multi-objective genetic algorithm," *International Journal of Electrical and Computer Sciences IJECS-IJENS*, vol. 12, no. 5, pp. 61–66, 2012.

[16] P. Kumar and A. Kanavalli, "A Similarity based K-Means Clustering Technique for Categorical Data in Data Mining Application," *International Journal of Intelligent Engineering and Systems*, vol. 14, no. 2, pp. 43–51, apr 2021.

[17] G. Pan, K. Li, A. Ouyang, X. Zhou, and Y. Xu, "A Hybrid Clustering Algorithm Combining Cloud Model IWO And K-Means," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 28, no. 06, p. 1450015, sep 2014.

[18] W. Abdulelah Qasim and B. Ahmed Mitras, "A Hybrid Algorithm Based on Invasive Weed Optimization Algorithm and Grey Wolf Optimization Algorithm," *International Journal of Artificial Intelligence Applications*, vol. 11, no. 1, pp. 31–44, jan 2020.

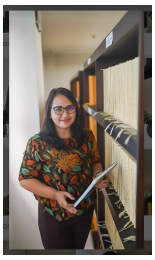
[19] N. L. G. P. Suwirmayanti, E. Setyaningsih, R. A. N. Diaz, and K. Budiarta, "Optimization of the K-Means Method for Clustering Banking Data Using the Hybrid Model of Invasive Weed Optimization and K-Means (Iwokm)," *ICIC Express Letters*, vol. 18, no. 4, pp. 413–422, 2024.

[20] W. T. Li, Y. Q. Hei, and X. W. Shi, "Synthesis of Conformal Phased Antenna Arrays With A Novel Multiobjective Invasive Weed Optimization Algorithm," *Frequenz*, vol. 72, no. 5-6, pp. 209–219, apr 2018.

[21] B. Khotimah, M. Miswanto, and H. Suprajitno, "Optimization of Feature Selection Using Genetic Algorithm in Naïve Bayes Classification for Incomplete Data," *International Journal of Intelligent Engineering and Systems*, vol. 13, no. 1, pp. 334–343, feb 2020.

[22] J. Keek, S. Loh, Y. Wong, X. Woo, and W. Lee, "Genetic Algorithms and Particle Swarm Optimization for Interference Minimization in Mobile Network Channel Assignment Problem," *International Journal of Intelligent Engineering and Systems*, vol. 14, no. 4, pp. 276–288, aug 2021.

[23] Y. Song, F. Wang, and X. Chen, "An improved genetic algorithm for numerical function optimization," *Applied Intelligence*, vol. 49, no. 5, pp. 1880–1902, may 2019.



Ni Luh Gede Pivin Suwirmayanti graduated with a Bachelor of Computers degree in the Computer Systems Study Program at the School of Information and Computer Management (STMIK) STIKOM Bali in 2010 and then continued her master's education at the Faculty of Engineering, Udayana University, Program Studying Information and Computer Systems Management where he obtained a Master's degree in 2014 and is currently in the process of studying a Doctoral degree at Udayana University. She has the status of Assistant Professor in the Computer Systems Study Program at the STIKOM Bali Institute of Technology and Business and teaches courses in Data Mining, Data Warehouse, System Analysis and Design and Object Oriented Modeling. Her research focus is in the fields of Data Mining and Machine Learning.

I Ketut Gede Darma Putra is a Professor at Udayana University since 2014 with a focus on Image Processing, Machine Learning and Data Science. Apart from being active as a teacher in Machine Learning, Image Processing and Artificial Intelligence courses at the Udayana University Engineering Doctoral Study Program, he is also active in various activities in professional organizations, community service, collaboration in the government sector, and the publication of scientific works in journals and conferences.

Made Sudarma is a professor of information technology science at the Electrical Engineering Study Program, Faculty of Engineering, Udayana University at Udayana University since 2019. His research includes internet and web applications, cloud computing, artificial intelligence, data warehousing and data mining, computer graphics and virtual reality, as the author of books and as a reviewer in international and national journals. In addition, he also completed vocational education (IPU., ASEAN Eng) and is active in academic activities, and also active as an Information Technology consultant in local government, private sector, and tourism..

I Made Sukarsa hold a Doctoral degree from Udayana University, Indonesia, in 2019. He also obtained his Master of Engineering degree from Gajah Mada University, Indonesia, in 2005. He received his S.T degree in informatics engineering from the Gajah Mada University, Indonesia, in 2000 and now he is a Associate Professor at the Department of Information Technology, Faculty of Engineering Udayana, Indonesia. Currently actively as a lecturer and conducting research on IT governance, dialog models on chatbot engines, datawarehouses and system integration..

Emy Setyaningsih is a lecturer in the Department of Computer Systems Engineering at AKPRIND University, Indonesia. She obtained a Bachelor's in Computer Science from IST AKPRIND Yogyakarta, Indonesia, in 1996. She has also earned M.Com. and Ph.D. in computer science from Universitas Gadjah Mada (UGM), Yogyakarta, Indonesia, in 2004 and 2019. Her research interests include cryptography, digital image processing, pattern recognition, and Artificial Intelligence. Scopus ID: 55943273200, Orcid: https://orcid.org/0000-0003-3254-3520.