Accredited Ranking SINTA 2
Decree of the Director General of Higher Education, Research and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online on: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 7 No. 4 (2023) 870 - 883 ISSN Media Electronic: 2580-0760

Antlion Optimizer Algorithm Modification for Initial Centroid Determination in K-means Algorithm

Nanang Lestio Wibowo[1], Moch Arief Soeleman[2], Ahmad Zainul Fanani[3]
[1,2,3]Master of Informatics Engineering, Faculty of Computer Science, Universitas Dian Nuswantoro
[1]na2nk.mlekiti@gmail.com, [2]arief22208@gmail.com, [3]a.zainul.fanani@dsn.dinus.ac.id

Abstract

Clustering is a grouping of data used in data mining processing. K-means is one of the popular clustering algorithms, easy to use and fast in clustering data. The K-means method groups data based on k distances and determines the initial centroid randomly as a reference for processing. Careless selection of centroids can result in poor clustering processes and local optima. One of the improvements in determining the initial centroid on the k-means method is to use the optimization method for determining the initial centroid. The modified Antlion Optimizer (ALO) method is used to improve poor clustering in the initial centroid determination and as an alternative to determining the initial centroid in the k-means method for better clustering results. The results of the research on the use of the proposed method for determining the initial centroid provide an increase in clustering compared to the usual k-means and k-means++ methods. This is evidenced by the evaluation of the Sum of Intra-Cluster distance (SICD) with UCI datasets, namely iris, wine, glass, ecoli and cancer in each method, the best SICD value was obtained in the proposed method. Then measuring the best SICD value for each method and datasets is measured by providing a ranking proving that the proposed method on the iris, wine, cancer datasets gets the first rank and on the ecoli and glass datasets the proposed method and the k-means++ method both get the first rank. From the average ranking value, the proposed method is ranked first which provides evidence that the proposed method can improve clustering results and can be an alternative method for determining the initial center of a cluster using the k-means method.

Keywords: clustering; initial center; centroid; antlion optimizer; sum of intra-distance clusters; k-means

1. Introduction

Data mining is a data processing technique that is used to find or get hidden patterns, in large data sets [1]. One of the important techniques in data mining is clustering which is a data processing technique using partitioning techniques [2]. Clustering technique is a data processing technique by grouping data objects (patterns, entities, events, units, observations) into a number of clusters [3]. Another meaning of the clustering technique is an observation or case, grouping data based on the similarity of the objects studied. A cluster is a collection of dissimilarity data to another or similarity data to other groups [2], [4].

K-means is one of the popular algorithms [5], simple, fast [6] and easy to use in partition-based clustering techniques and is most often used for grouping data [6]. The k-means algorithm was officially published in the MacQueen [7] methods and Forgy [8] but the basic k-means was proposed by Stuart Lloyd in 1958. This method groups data into *k* clusters based on the closest distance of the data to the cluster center. This method determines k cluster centers randomly to represent the initial k cluster centers. This k-means algorithm has advantages in speed, is very efficient [2] and easy to group data [7]. In addition, k-means also has a clustering process which is generally fast [9] and has linear space complexity.

However, k-means has problems in determining the initial center of the cluster (centroids) [10]. K-means is very sensitive to the initial centroid. Careless determination of the initial centroid will affect the quality of the resulting clusters which causes sub-optimal clustering results [11] and can produce poor cluster results [12]. In addition, each series of clustering processes for the same datasets can produce different outputs [3].

In the literature review conducted using the Systematic Literature Review (SLR), one of the handling methods for determining the initial centroid in the k-means method is using the k-means++ method. It is proven that there are 8 papers that use the k-means++ method as a comparison method of the proposed method. The k-

Accepted: 16-03-2023 | Received in revised: 01-06-2023 | Published: 12-08-2023

means++ algorithm proposed by Arthur [13] is one of the proposed improvement methods to deal with the problem of determining the initial centroid randomly in the k-means algorithm. This method is proposed to reduce the negative impact of the k-means algorithm which is highly dependent on the initial cluster center. This method selects the initial center c1 randomly from the datasets and then selects the next center ci by calculating the maximum distance from the selected point to other points in the datasets. Each data has the opportunity to become a cluster center so that each data is calculated for the opportunity value to be selected and the closest is the most appropriate. The randomized seeding technique formula in this method will produce a value that can be used as a determination provided that the farther the data value is, the higher the probability that the data value will become the next C value. Then the cluster center data values are used to be processed with the k-means algorithm. However, the results of the improvement resulting from the k-means++ method according to Z. Wu in his research are still easily trapped in the local optima [14] which causes the clustering results to be less than optimal.

Several studies have also been conducted to better determine the initial cluster center. One of the studies in the Systematic Literature Review (SLR) used in this study, obtained a study conducted by Celebi et al [9] which provides an overview and has explained an overview of the algorithm for determining the initial centroid. In their research, eight linear time complexity initialization methods compared by them. Including the k-means++ [13], MacQueen [7], Principal Component Analysis-Part (PCA-part) [15], Forgy [8], Bradley and Fayyad [16], Maximin [17], VarPart [15], and greedy k-means ++ [13]. Non-parametric statistical tests were used and performed for the experimental analysis. The analysis obtained in his research reveals that popular initialization algorithms such as Maximin (1985), Macqueen (1967) and Forgy (1965) give unsatisfactory results. This method only provides a better alternative based on comparatively measured computational complexity [9].

From the conclusions of the SLR research that was conducted in the literature review, many new methods were obtained in determining the initial centroid, either by combining methods, improving methods or applying methods from other fields. This is of course still a mystery to researchers in order to improve cluster performance from problems in the k-means algorithm, namely in determining the initial centroid. This is also proven by research on determining the initial centroid which is still being obtained and is still being studied by several researchers from year to year to get better clustering performance in the k-means algorithm.

Kumar & Reddy proposed a Robust Density Based Initialization (RDBI) [5] approach to determine initial seed points located in dense areas and avoid outliers as initial seeds. The algorithm starts by calculating the kd-trees of the data set, determining that leaf nodes are those that contain less than the specified minimum number of points.

Cabria & Gondra proposed The Mean-Shift-Based Initialization (MS) Potential K-means [18]. Unlike K-means, mean shift clustering does not depend on prior knowledge of the number of clusters, i.e., the value of k. The average shift also finds the basic mode probability density function (pdf) of the observations, which would be an excellent choice of initial cluster center for K-means. This method uses the most popular Parzen-window [19] approach to estimate the unknown pdf p(x). The given point function is centered at each point. By the way every point xi {x1, ..., xn}, where xi ∈ Rd, it locates a window or kernel that contributes to the pdf estimate.

Capó et al proposed a new iterative approach to the K-means problem that is based on recursive partitioning of datasets, since each partition is thinner than the previous one. We call this the Recursive Partition Based K-means (RPKM) approach [12]. The idea behind this algorithm is to approximate the K-means problem for the complete data set by recursively applying a weighted version of the K-means algorithm to an ever-increasing, but small, subset of datasets. In the first step of the RPKM, the data set is partitioned into a number of subsets each characterized by representation (centre of mass) and corresponding weight (cardinality). Finally, a weighted version of Lloyd's algorithm [20] is applied to representative sets. From one iteration to the next, a finer partition is constructed and the process is repeated using the optimal set of centroids obtained in the previous iteration as initialization. This iterative procedure is repeated until certain termination criteria are met.

Khanmohammadi et al proposed a combination of methods between KHM and OKM to get better clustering performance [21]. In dealing with the initial cluster center problem Khanmohammadi et al used the KHM method, proposed by B. Zhang et al [22] and B. Zhang [23] by minimizing the harmonic average of all data points from the cluster center. The harmonic average gives weight to each data point based on its proximity to each center. These weights are considered as the importance of each point in identifying clusters in the dataset. In other words, the KHM algorithm introduces bias (using weights) to shift the cluster center to more important data points according to several criteria.

Nidheesh et al proposed a method with a density-based approach for initializing the initial center of a K-Means cluster with the name Density K-means++ [6] which was inspired by the Density K-Means (DKM) method proposed by Lan [24] and the K-means method.

DOI: https://doi.org/10.29207/resti.v7i4.4997

means++ proposed by Arthur [13]. DKM++ finds a set of data points as the initial centroids of a dense region in feature space. It starts by calculating the distance matrix M (paired distance between data points). M is normalized min-max to make the distance between data points fall in the interval [0; 1]. The next main goal of this method is to calculate the local density of each data point. The values are then subjected to min-max normalization.

G. Zhang et al proposed K-means Based on Density Canopy [10] which was inspired by the Canopy algorithm proposed by Andrew McCallum, Kamal Nigam and Lyle Ungar [25]. The Canopy Algorithm sets two distance thresholds $T1$ and $T2$, randomly selects the initial cluster center, and calculates the Euclidean distance between the sample and the initial center. Samples will be classified into appropriate clusters according to thresholds. Next, the clustering data set is divided into n clusters. The improvement made is in the process after the distance calculation is fulfilled, then the average value is taken from the results of the distance calculation, then the maximum value is taken, then it is repeated again to get the new centroid value, the division of the previous maximum value is divided by the maximum value obtained.

Wangchamhan et al proposed an improvement method from the League Championship Algorithm (LCA) method proposed by Kashan [26]. The LCA method works where the population of solutions evolves to the optimal solution. Each team (individual) in the league (team population) is a feasible solution to the problem being solved and consists of n players, which corresponds to a variable number. Once an artificial weekly league schedule is created, teams play against teams, each of which has a playing strength that matches their fitness value. According to the league schedule, the teams compete in pairs for S x (L - 1) weeks, where S is the number of seasons and one week is recorded as The result of play is to determine the winner or loser. To prepare for the next match, each team checks the results of the previous week's matches, and uses this information to form its new team. A productive team formation that is expected to replace the best in the team and it is selected with the guidance of the team formation with better playing strength. The proposed solution improvements are to improve global convergence and to help prevent getting stuck in a local LCA minimum. The proposed solution will replace constant parameters in LCA with chaotic mappings with ergodic disorder, disorder and stochastic properties. This new algorithm is called the Chaotic League Championship Algorithm (C-LCA) [27].

Therefore, from the description of the problem and some of the literature obtained in the SLR process in this topic area, it is still being carried out with the suggestion that several methods proposed in previous studies are still relevant for improving clustering results. This study proposes an optimization method for determining the initial centroid, namely by using the Antlion Optimizer (ALO) method which is modified on the number of iterations and adjusting the initial input to get a good initial cluster center so as to produce a more efficient cluster. well and improve clustering performance on the k-means algorithm.

The ALO method is a new population-based metaheuristic algorithm inspired by the hunting behavior of antlions [28]. When hunting for ants, the antlion will make a cone-shaped trap hole then stay silent at the bottom of the trap hole while waiting for ants that move randomly looking for food to be trapped in the trap hole. ALO has four parameters, namely lower limit (lb), upper limit (ub), the maximum number of iterations (max_iter) and the number of search agents (N). Parameters lb and ub are used to set the random motion limits of the ants. Mirjalili (2015) has tested the performance of the ALO algorithm using 3 test function groups, namely composite, multi-modal and uni-modal. The performance of the ALO method in this test is compared to the performance of two well-known metaheuristic algorithms, namely the Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) as well as the new metaheuristic algorithms, namely the Firefly Algorithm (FA), Cuckoo Search (CS), Bat Algorithm (BA), Flower Pollination Algorithm (FPA) and State of Master Search (SMS) composite test results show that compared to other optimization algorithms, the ALO algorithm has better performance, which means that to overcome various difficulties in a challenging search space, the operators of the algorithm ALO can balance exploitation and exploration precisely. The multi-modal test results show that compared to other optimization algorithms, the ALO algorithm has better performance, which means that the ALO algorithm has a high exploration level so that it avoids local optimum and can help explore promising search space areas so that it has the opportunity to reach a global optimum. The results of the uni-modal test also show that compared to other optimization algorithms, the ALO algorithm has better performance, which means that the ALO algorithm is able to quickly obtain the optimum solution because it has high exploitation [28].

From the superior ALO algorithm that has been tested by Mirjalili (2015), the ALO algorithm is selected and modified at the maximum number of iterations and changes to the initial input parameters are adjusted so as to provide an alternative choice for determining the initial centroid to be better in order to improve clustering results in the k-algorithm. means. Modification of the ALO method was made to minimize the number of iterations where an inaccurate number of iterations could eat up the computational process and result in inappropriate Elite results.

DOI: https://doi.org/10.29207/resti.v7i4.4997

Adjustments to the initial input parameters are also adjusted to the type of datasets used because each dataset also has a different number of attributes and characters.

2. Research Methods

2.1 Systematic Literature Review (SLR)

In this study, a review was conducted to get insights in determining the initial center of cluster in the k-means algorithm. The review method used in this study is using a Systematic Literature Review (SLR). This review aims to analyze and identify research trends, methods and datasets in the topic of determining the initial centroid on the k-means algorithm. SLR is a term used to refer to specific research or research methodology as well as development of analyzes carried out to collect and evaluate related research on a particular focus topic [29]. SLR is a literature search technique to identify, analyze, evaluate and interpret the results of research that has been carried out as a whole that is relevant to the topic area or research question with the aim of providing answers to specific research questions [29]. The review method, style, literature sources and formulation of questions in this study were inspired by Wahono [30]. Figure 1 represents the SLR stages.

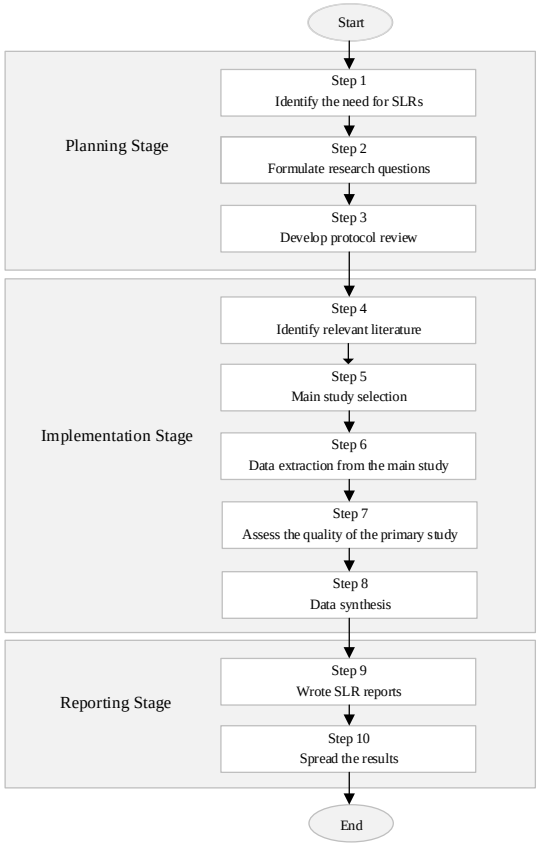

Figure 1. Stages of Systematic Literature Review (SLR)

In general, the SLR review is carried out in three stages, namely the planning stage, the implementation stage and finally the reporting stage (as shown in Figure 1). The first step is to identify the purpose of conducting a literature review, formulate research questions to be focused and consistent according to the research topic to be reviewed and develop a review protocol. The second step is identification of the literature, selection of main studies, data extraction from the main studies, assessing the quality of the main studies and data synthesis. And the third or final step is to write a report on the results of the SLR and publication.

Research Questions (RQ): RQ are structured so that the implementation of the review can be focused and consistent. Research questions were designed using Population, Intervention, Comparison, Outcomes, and Context (PICOC) criteria [30]. This RQ on SLR is a research question for SLR and is different from the research questions on the main research in the study. Table 1 shows the PICOC summary.

Table 1. PICOC Summary

Population	K-means clustering
Intervention	Method of determining the initial center of the cluster
Comparison	-
Outcomes	K-means performance
Context	Small and large datasets. Studies in academia and industry.

Research questions are made based on the needs of the chosen topic. Table 2 is the research questions in this literature review. RQ in this study consists of 5 questions that are relevant to the needs and are used as knowledge and references in the main research in research.

Table 2. Research Questions (RQ)

RQID	RQ	Motivation
RQ1	Which year of publication and journal most often publishes the topic of determining the initial centroid on k-means?	Identify the journals that publish the most frequently on the topic and determine the initial center of the k-means cluster.
RQ2	Who are the active researchers on the topic of determining the initial centroid on k-means?	Identify researchers who are active on the topic of determining the initial centroid on k-means.
RQ3	What datasets are most frequently used in the topic of determining the initial centroid on k-means?	Identify the most frequently used datasets in the topic of determining the initial centroid on k-means.
RQ4	What method does the researcher propose to determine the initial cluster center on k-means?	Identify the method proposed by the researcher for determining the initial centroid on k-means.
RQ5	What method is often used to determine the initial cluster center?	Identify the methods that are often used to determine the initial cluster center on k-means.

DOI: https://doi.org/10.29207/resti.v7i4.4997

From the studies that will be analyzed, publications conducted between 2017 and 2020, the source of publication and the year of publication in the topic of determining the initial centroid are used to answer RQ1. Then researchers who are most active or often publish on the topic of determining the initial centroid to answer RQ2. Furthermore, the use of datasets to answer RQ3. Then the method proposed in determining the initial centroid is used to answer RQ4. And finally, the most frequently used method for determining the initial centroid is used to answer RQ5.

Search Strategy: The search process is in the four SLR stages as shown in Figure 1 consisting of several processes including selecting a digital library and setting keywords. Before starting the search, it is necessary to determine or select the appropriate database to find relevant journals. The digital libraries that took the study were ieeexplore.ieee.org (IEEE eXplore) and sciencedirect.com (ScienceDirect).

The search step uses the developed keywords, namely the first to use PICOC to identify search terms, especially from populations and interventions. Step two use research questions to identify search terms. The third step is to use relevant keywords, abstracts and titles to identify search terms. The fourth step in the search term is identifying alternative spellings, synonyms and antonyms. The fifth step is thorough keyword determination using the identification of the boolean AND and OR search terms.

The keywords used in the search are initial* AND (Centroid OR Center OR Seed OR Cluster Center) AND K-means. The search string adjustment is conditional and performed, because the string adjustment directly increases the already irrelevant list of irrelevant studies, so the original search string is kept. In any existing database publication search engine, the search string will need to be adjusted to meet specific requirements. Based on the title, keywords and abstract the database was searched. Searches are restricted to years of publication between 2017-2020. Only research article publication criteria and journal publications with Q scores in the Q1 and Q2 categories were taken.

Study Selection: The process of searching for and selecting the selected studies at each stage is shown in Figure 2. In the study selection, there is stage 5 of the SLR which is carried out in two steps, namely the first is the study exceptions which are selected based on the abstract and title, and the second the study exceptions are selected based on the full text. The study selection used was only research articles and publication journals with Q categories Q1 and Q2, while books and proceedings were not used in the study selection.

It can be seen in Figure 2, the sciencedirect digital library search engine, 137 papers were found, while IEEE Explorer found 171 papers, which means a total of 308. Then excluding papers only in titles and abstracts, a total of 51 papers were obtained. Then exclude studies based on the full text and get 20 papers that are in accordance with the main research topic of the selected studies.

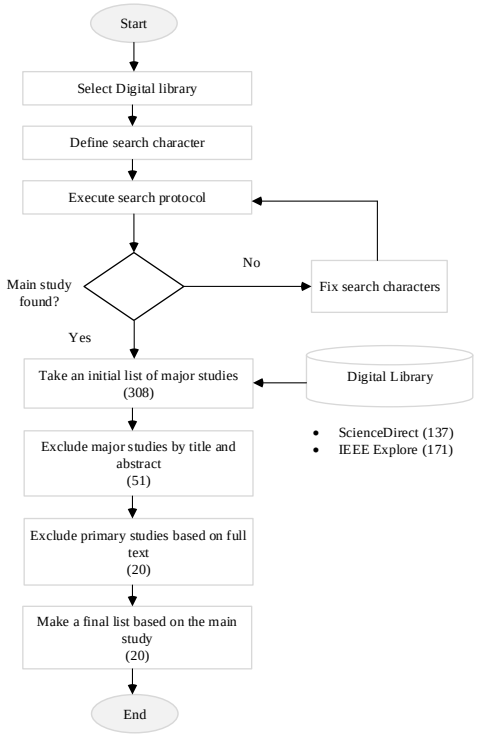

Figure 2. Selected Study Selection Search

Table 3 is a detail of the criteria taken and not taken (Inclusion and Exclusion) for the purpose of evaluating the research.

Table 3. Criteria of Inclusion and Exclusion

Research criteria taken (Inclusion)	Research that discusses determining the initial centroid on the k-means method. The research taken was only the type of paper research articles with Q1 and Q2 publication journal values. Research that provides suggestions for improvement methods for determining the initial centroid. For duplicate research, the most complete and up-to-date data will be taken.
Research criteria that are not taken (Exclusion)	Research outside of academia and industry. Research is not written in English. Books and proceedings. Q3 and Q4 publication journal values.

Data Extraction: Data extraction is designed to collect data from selected studies that have been defined as needed to answer research questions that have been determined at the research question stage. Properties are identified through RQ and analysis which will be introduced. The four properties used to answer the RQ are shown in Table 4.

DOI: https://doi.org/10.29207/resti.v7i4.4997

Table 4. Data Extraction Properties for Research Questions

Property	Research Question
Year of Publication and Journal of Publication	RQ1
Active Researcher	RQ2
Datasets Used	RQ3
Proposed Method	RQ4
The Method that is often Used	RQ5

Main Study Quality Score and Data Synthesis: The quality assessment of the selected studies was used to help provide a theoretical view of the synthesis findings and provide strength of the conclusions. The purpose of data synthesis is to provide insights and collect evidence from the main studies selected to answer the research question.

Threats to Validity: In this review aims to analyze studies on determining the initial center of a cluster in the k-means algorithm. This review was unaware of any bias in study selection. The search for all paper titles published in journals is not based on manual reading. This means that some of the initial cluster center determination papers on the k-means algorithm from the journal publication process in this review may be excluded. Because the literature review was only carried out on search engines sciencedirect.com and ieeexplore.ieee.org and only research articles and Q categories Q1 and Q2 in journal publications were taken for the review process.

Results and Analysis of the SLR are from Significant Journal Publications, Active Researcher, Datasets Used, The Method that is Often Used

Significant Journal Publications: From the studies selected in this study in this literature, there are 20 defined studies that discuss the central topic of cluster initial determination in the k-means algorithm. Many studies have been conducted regarding the determination of initial cluster centers over the years. However, in this study, a literature review was taken from January 2017 to June 2020 (at the time of the search) to obtain the latest research on the topic of determining the initial centroid. Figure 3 shows the number of paper publications per year from 2017 to 2020.

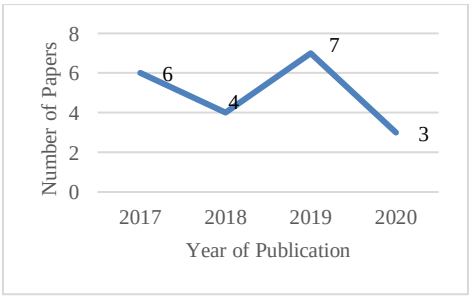

Figure 3. Number of Publications

In Figure 3, it can be seen shows that there are still many researchers who are conducting research on clustering problems in the topic of determining the initial centroid in the k-means algorithm. This was obtained from 2017 as many as 6 papers, then in 2018 there were 4 papers, experiencing a decrease from the previous year. However, in 2019 the research on the topic of determining the initial centroid increased again to 7 papers and in 2020 this research was in number 3 papers at the time this review was conducted, namely in June 2020. Figure 4 is a paper published in each journal publication which shows that this research topic is still very relevant today, because from several journal publications there is still research on the topic of determining the initial centroid in the k-means algorithm with a significant number of papers.

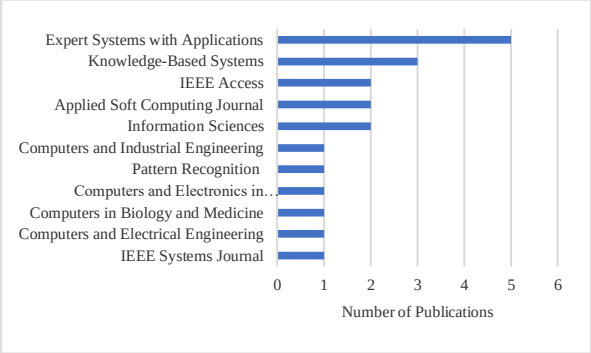

Figure 4. Number of Journal Publications

Active Researcher: From the studies selected in this study, namely as many as 20 papers, researchers were identified and investigated who contributed and were active in the topic of determining the initial centroid in the k-means algorithm. Figure 5 shows researchers who are active and contribute to this topic. Of the 20 main researchers, it was found that there were no researchers who conducted research on this topic in more than one study. This indicates that there are still many researchers who are still doing research on the topic of centroid determination.

Figure 5. Active Researcher and Number of Studies as Main Researcher

Datasets Used: In this literature review, 20 studies were selected and assigned for analysis. Figure 6 shows the percentage of the number of calculated datasets used from 2017 to 2020. 75% of public datasets in determining the initial center of the k-means algorithm cluster are used by researchers, while 25% of private datasets are used by researchers. Most of the public datasets used by researchers are from the University of

DOI: https://doi.org/10.29207/resti.v7i4.4997

California Irvine (UCI) Machine Learning repository and are freely distributed and can be used for further research. Meanwhile, private datasets belonging to individuals or private companies are not distributed as public datasets.

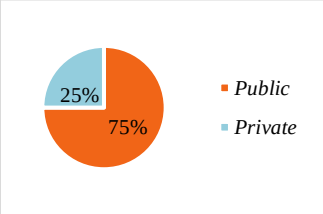

Figure 6. Distribution of Datasets

The distribution of 20 studies from publication between 2017 and 2020 from sources that have been analyzed shows that many public datasets have been used and more studies have been published on research on determining the initial cluster center on the k-means algorithm from 2017 to 2020. Use public datasets are relatively stable compared to the use of private datasets. A summary of annual publications is presented in Figure 7. The standard datasets used can make research verifiable, disprovable and repeatable [31].

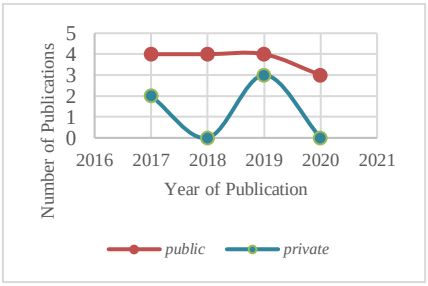

Figure 7. Private and Public Datasets Distribution

The public datasets that are most often used in research on the topic of determining the initial centroid in the k-means algorithm are presented in Figure 8 which is a dataset from UCI. Where it can be seen that the iris datasets are used by 8 papers, the wine datasets are used by 6 papers, the glass and ecoli datasets are used by 3 papers. While other datasets such as cancer, hayet-roth, cmc, sonar, abalone, musk, and thyroid were found in 2 papers each.

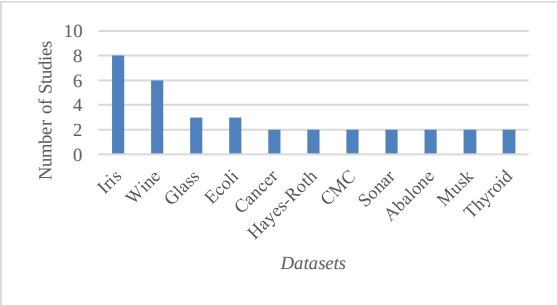

Figure 8. Frequently used Public Datasets

The Proposed Method: From the research that has been done in the literature review that was analyzed, a method was obtained to determine the initial cluster center value in the k-means algorithm. There are many methods or ways of determining the initial centroid value, some of which are improvements to existing methods, some use methods from other fields which are then applied in determining the initial centroid, or combining existing methods, both methods. or more to get the best performance in determining the initial centroid. Table 5 is a collection of methods proposed by researchers in 20 studies analyzed in determining the initial center of the k-means algorithm cluster. Each method has advantages and disadvantages in determining the initial centroid value with the datasets used and with the conditions at the time of the study. There are many new methods in the initial cluster center treatment on the k-means algorithm. Either by combining methods, improvement methods or methods of application from other fields. This is of course still a mystery to researchers in order to improve cluster performance from the problems that exist in the k-means algorithm, namely buying the initial center of the cluster.

Table 5. Proposed Method

No	Method
1	Robust Density Based Initialization(RDBI) [5]
2	The Mean-Shift-based Initialization (MS) Potential K-means [18]
3	Recursive Partition Based K-means (RPKM) [12]
4	Combine K-harmonic means and overlapping k-means algorithms (KHM-OKM) [21]
5	Density K-Means++ [6]
6	Chaotic League Championship Algorithm(C-LCA) and The hybrid of k-means and Chaotic League Championship Algorithm (KSC-LCA) [27]
7	K-means Based on Density Canopy [10]
8	LeaderRank Based [32]
9	Tri-Level K-means and The Bi-Layer K-means [33]
10	I-K-means−+ [34]
11	Self-Organizing-Center K-means [35]
12	Farthest Distance Cluster Center (FDCC) Initialization and K-means Based Co-clustering (kCC) [36]
13	Improved Density-Based Initial Cluster Centers Selection Algorithm [37]
14	Inward Intensified Exploration Firefly Algorithms (IIEFA) and Compound Intensified Exploration (CIEFA) [38]
15	Text Mining-Constrained Seed K-means [39]
16	Adaptive Clustering Number of K-means [40]
17	Multiview Registration Algorithm [41]
18	Improving Spherical K-means [42]
19	Quantum-Inspired Ant Lion Optimization [4]
20	Proposed k-Centroid Initialization Algorithm (PkCIA) [43]

The Method that is Often Used: Figure 9 is a method that is often used and as a comparison to measure the level of performance of the proposed algorithm from 20 studies that have been analyzed. The methods used as a comparison to measure performance that are most often used are the conventional k-means and k-means++ algorithms. In the 20 studies that have been analyzed, the conventional K-means algorithm is the main comparison to the algorithm proposed by the researcher, namely 14 papers include the conventional

DOI: https://doi.org/10.29207/resti.v7i4.4997

k-means algorithm as a comparison test tool for the proposed algorithm. Then the 8 papers include a comparison algorithm using the k-means++ algorithm. Furthermore, several comparison algorithms are used from the algorithms produced by previous researchers. The method most often used as a comparison will also be used as a comparison method in the main research.

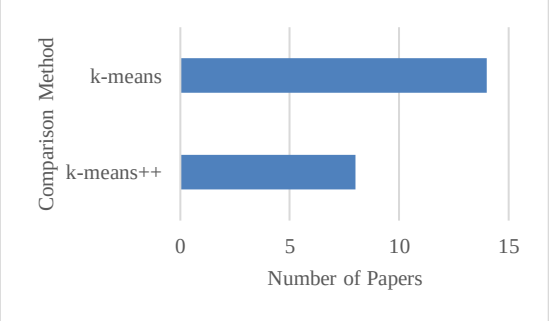

Figure 9. Frequently used Public Datasets

Analysis of selected studies suggests that research on determining the initial center of this cluster focuses on research trends, datasets, and methods. From the list of selected studies from 2017 to June 2020, 20 papers were obtained and determined, obtained from the search engines sciencedirect.com and ieeexplore.ieee.org. Of the 20 papers, there were 6 papers in 2017, then in 2018 there were 4 papers, experiencing a decrease from the previous year. However, in 2019 the research on the topic of determining the initial centroid increased again to 7 papers and in 2020 this research was in number 3 papers at the time this review was conducted. The publication that became the most publisher on the topic of determining the initial centroid of the research conducted was Expert Systems with Applications with an SJR value of 1.49 and Q1 category in Artificial Intelligence when research was carried out with 5 papers.

Of the total datasets distributed, 75% of researchers in determining the initial centroid used public datasets and 25% of researchers used private datasets. The most frequently used datasets are the UCI repository datasets. This shows that the UCI datasets can be recommendations and alternatives for using datasets requirements for clustering techniques in determining the initial centroid in the k-means algorithm.

From the research that has been done, there are many new methods in determining the initial centroid, either by combining methods, improving methods or applying methods from other fields. This is of course still a mystery to researchers in order to improve cluster performance from problems in the k-means algorithm, namely in determining the initial center of the cluster.

In the 20 studies that have been analyzed, the method that is often used as a comparison is the k-means and k-means++ algorithms where the conventional k-means algorithm is the main comparison to the algorithm proposed by the researcher, namely as many as 14 papers include the conventional k-means algorithm as a tool. comparison test of the proposed algorithm. Then the 8 papers include a comparison algorithm using the k-means++ algorithm.

2.2 K-means Algorithm

The k-means algorithm is an algorithm in the clustering technique which has a role in grouping data or partitioning data iteratively into several predetermined groups. K-means is one of the simplest unsupervised learning algorithms for solving clustering problems [2]. The k-means algorithm classifies objects that are similar to one group and dissimilar to another, so that objects in one cluster have high similarity compared to objects in other clusters. The k-means process flow begins by determining the number of clusters as much as k, then determining k cluster centers randomly. Furthermore, each data object will be grouped based on the closest distance to the cluster center, then the cluster center is updated based on the data points in each cluster. This stage is repeated until the convergent criteria are met or the centroid value does not change anymore. The similarity between one data and another is obtained by calculating the distance of each data from the cluster center. To get the value of the similarity measure, the euclidean distance formula is used with Formula 1.

$$D_{(i,j)} = \sqrt{(B_{1i} - B_{1j})^2 + (B_{2i} - B_{2j})^2 + \ldots\ldots + (B_{ki} - B_{kj})^2} \quad (1)$$

Where $D_{(i,j)}$ is the distance of the i data to the j cluster center, B_{ki} the i data on the k data attribute, and B_{kj} is the j center point on the k attribute. In the fourth stage of the process, each cluster representation is relocated to the center of the cluster with the arithmetic mean of each cluster. This is also what makes this method often referred to as the cluster centroid or cluster mean as the name suggests.

2.3 Antlion Optimizer (ALO)

The ALO algorithm is an algorithm that adopts mimicking the interaction between ants and antlions in a trap [44]. To model this interaction, the ants are required to move over the search space, and the antlions are allowed to hunt them [28]. Ants find food by moving around. The movement of these ants is modeled in Formula 2.

$$X(t) = [0, cumsum(2r(t1) - 1), cumsum(2r(tn) - 1)] \quad (2)$$

Where t is the step of the random number. Then n is the maximum value of the number of iterations. Then the *cumsum* is a calculation of the cumulative sum. And then $r(t)$ is a stochastic function. This stochastic function uses Formula 3.

$$r(t) = \begin{cases} 0, if\ rand \leq 0{,}5 \\ 1, if\ rand > 0{,}5 \end{cases} \quad (3)$$

DOI: https://doi.org/10.29207/resti.v7i4.4997

Where *rand* is a random number obtained with a uniform distribution in the interval [0,1] and *t* denotes a random walk step (iteration). Ant positions are stored and utilized during optimization in the M_{Ant} matrix using Formula 4.

$$M_{Ant} = \begin{bmatrix} S_{1,1} & S_{1,2} & \cdots & \cdots & S_{1,d} \\ S_{2,1} & S_{2,2} & \cdots & \cdots & S_{2,d} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ S_{n,1} & S_{n,2} & \cdots & \ldots & S_{n,d} \end{bmatrix} \quad (4)$$

Where $S_{i,j}$ shows the value of the *j-th* variable (dimension) of the *i-th* ant, *n* is the number of ants, and *d* is the number of variables, M_{Ant} is the matrix for storing the position of each ant. The fit function (objective) is used during optimization to evaluate each ant. The M_{OA} matrix is used to store the fitness value of all ants using Formula 5.

$$M_{OA} = \begin{bmatrix} f([S_{1,1} . S_{1,2} , \ldots , S_{1,d}]) \\ f([S_{2,1} . S_{2,2} , \ldots , S_{2,d}]) \\ \vdots \\ \vdots \\ f([S_{n,1} . S_{n,2} , \ldots , S_{n,d}]) \end{bmatrix} \quad (5)$$

Where M_{OA} is the matrix of saving the fitness of each ant, $S_{i,j}$ denotes the value of the *i-th* dimension of the ant, *n* is the number of ants, and *f* is the objective function. Apart from the ants, it is assumed that the antlions are also hiding somewhere in the search room. To store their positions and fitness values, a matrix with Formula 6 is used.

$$M_{Antlion} = \begin{bmatrix} QL_{1,1} & QL_{1,2} & \cdots & \cdots & QL_{1,d} \\ AL_{2,1} & AL_{2,2} & \cdots & \cdots & QL_{2,d} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ QL_{n,1} & QL_{n,2} & \cdots & \cdots & QL_{n,d} \end{bmatrix} \quad (6)$$

Where *d* is the number of variables (dimensions), *n* is the number of antlions, $QL_{i,j}$ denotes the value of the *j* from dimension of the *i*, and antlion, $M_{Antlion}$ is the matrix for storing the position of each antlion. For the fitness function of the antlion, it is entered into the matrix with Formula 7.

$$M_{OQL} = \begin{bmatrix} f([QL_{1,1}, QL_{1,2}, \ldots , QL_{1,d}]) \\ f([QL_{2,1}, AL_{2,2}, \ldots , QL_{2,d}]) \\ \vdots \\ \vdots \\ f([QL_{n,1}, QL_{n,2}, \ldots , QL_{n,d}]) \end{bmatrix} \quad (7)$$

Where *f* is the objective function, *n* is the number of antlions, $QL_{i,j}$ denotes the value of the *j* from dimension of the *i* antlion, and M_{OQL} is a matrix to store the fitness of each antlion.

Formula 1 is the basis of all random walks. Ants update their position by walking randomly at each optimization step. Because each search space has a limit (variable range), Formula 1 cannot be directly used to update ant positions. They are normalized by min-max normalization with Formula 8 to keep random walks in the search space.

$$X_i^t = \frac{(X_i^t - a_i)\ x\ (d_i - c_i^t)}{(d_i^t - a_t)} + C_i \quad (8)$$

Where a_i is the minimum value of the antlion running process, c_i is the minimum value of the iteration process, d_i^t is the maximum value of the iteration process. Ant traps in antlions are included in the mathematical model using Formula 9 [45].

$$c_i^t = Antlion_j^t + c^t, d_t = Antlion_j^t + d_t \quad (9)$$

Where d^t is the maximum value of the iteration process, c^t is the minimum value of the iteration process. The antlion hunt is modeled on a roulette wheel. The renewal model uses Formula 10.

$$c^t = \frac{c^t}{I}, d^t = \frac{d^t}{I}, I = 10^k . \frac{t}{R} \quad (10)$$

Where *R* is the maximum value of the iteration process, *t* is the iteration value that has been selected, *k* is the constant iteration value using calculations based on the formula *k = 2, if t > 0.1R, k=3, if t>0.5R, k=5 if t>0.9R, k=6 if t>0.95R*. The antlions rebuild their burrows after catching prey and to encourage the ants to sink into the sand. This property is used Formula 11.

$$Antlion_j^t = Ant_i^t, if\ f(Ant_i^t) > f(Antlion_j^t) \quad (11)$$

Where Ant_i^t denotes the position of the *i* ant in the *t* iteration, $Antlion_j^t$ denotes the position of the *j* antlion selected in the *t* iteration, and *t* denotes the current iteration.

Elitism is an important characteristic of evolutionary algorithms acquired at each stage of the optimization process that allows them to maintain the best solution. In a study conducted by Mirjalili (2015) the best antlion for each iteration obtained so far is stored and considered elite. This antlion can affect the movement of all ants during iterations because the elite are the most powerful antlions. Therefore, it is assumed that the antlion surrounded by each walking ant is randomly selected by the roulette wheel and elite simultaneously. The antlion process is surrounded by every ant that runs randomly using Formula 12.

$$Ant_i^t = \frac{R_A^t + R_E^t}{2} \quad (12)$$

Where Ant_i^t indicates the position indication of the ant in iteration to *t*, R_E^t is the random walk of the ant around the elite in iteration *t*, and R_A^t is a random walk describing the position of the ant around the antlion chosen by the roulette wheel in iteration *t*. Algorithm 1 is the pseudo code flow of the ALO algorithm stages.

Initialize the population and the first random permutation of antlions and ants is random
Assume the most optimal elite by finding the best antlion
while the final condition is not met
  for all ants

DOI: https://doi.org/10.29207/resti.v7i4.4997

```
    Use the Roulette Wheel to select the antlion
    With Formula (10), update c and d
    Generate random roads and normalize
          them using Formula (2) and (8)
    Using Formula (12) for update the position of the ants
  end for
  For all ants, calculate the fitness value
  If the suitable ant looks better than the antlion, use
     Formula (11) to replace the antlion
  If antlion is better than elite results, update elite
end while
Return elite
```

Algorithm 1. Antlion Optimizer (ALO) Algorithm Pseudo Code Flow

2.4 Proposed Method

In this main research, the proposed method is an optimization method for determining the initial centroid using the Antlion Optimizer (ALO) algorithm [28] which is modified in the iteration process and determines the initial input parameters. Figure 10 is a stage or flowchart of the proposed method. Modification of ALO in this study was carried out at the stage of setting the initial parameters and the number of iterations according to clustering needs. The initial input setting in this study is to determine the appropriate number of search agents based on the number of rows of data divided by the number of *K* classes in predetermined datasets. The selection of the appropriate objective function is based on the number of attribute matrices in the datasets used in the calculation process.

The number of iterations in the proposed method uses a comparison of the amount of data to be processed. The comparison of the data is using the number of rows of data divided by the number of *K* (the number of classes that have been determined). The number of comparisons is used as the number of iterations. Then the total number of datasets obtained is used to initialize ants after the initial input parameters are adjusted. Algorithm 2 is the flow of the proposed method pseudocode where the modification process is in the early stages of the ALO method.

```
Initiate SearchAgent based on the number of
  data divided by the number of classes
Initiate the maximum iteration based on the
  amount of data divided by the number of classes
Initialize the population and the first permutation of
  random antlions
Initialize population and first random permutation of
  ants based on datasets
Assume the most optimal elite by finding the best antlion using an
  objective function that is adjusted to the number of matrix datasets
while the final condition is not met
  for all ants
    Use the Roulette Wheel to select the antlion
    With Formula (10), update c and d
    Generate random roads and normalize them using
          Formula (2) and (8)
    Using Formula (12) for update the position of the ants
  end for
  For all ants, calculate the fitness value
  If the suitable ant looks better than the antlion, use
     Formula (11) to replace the antlion
  If antlion is better than elite results, update elite
end while
Return elite
```

Algorithm 2. Antlion Optimizer (ALO) Algorithm Modification Pseudocode Flow

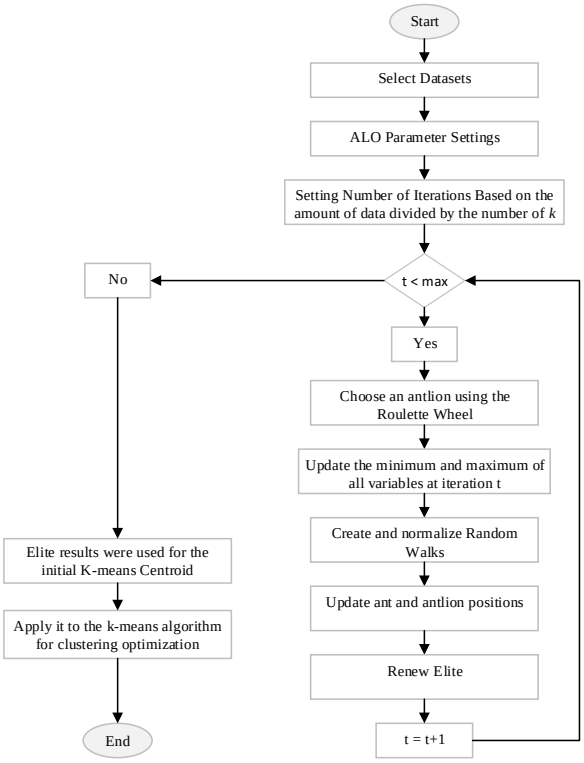

Figure 10. Flowchart Purpose Method

2.5 Datasets

In this study, the iris, wine, glass and ecoli datasets obtained in the SLR study were used. The datasets are obtained from UCI which can be downloaded for free. There are no missing values in these four datasets, so this data can be directly used for the clustering process. However, for the cancer datasets, there are missing values which contain the special character '?' in 16 of the 699 total rows of existing data. Then data processing was carried out as was done by Junwen in his research, namely by deleting data rows that contained missing values so that the data used became 683 [4]. Table 6 is a description of the datasets that are ready to be used for the clustering process.

Table 6. Description Datasets

No.	Datasets	Row Count	Class	Attribute	Missing Value
1	ecoli	336	8	7	no
2	wine	178	3	13	no
3	glass	214	6	9	no
4	iris	150	3	4	no
5	cancer	683	2	9	no

2.6 Evaluation

SICD or often called the Sum of Intra-Cluster Distance is used. SICD is a clustering evaluation method in

DOI: https://doi.org/10.29207/resti.v7i4.4997

which the process is to add up the distance between the centroids and each data point. The better the clustering results the shorter the distance [4]. In other words, the SICD value must be as minimal as possible to obtain optimal cluster quality. The SICD function was chosen and used for evaluation because of its speed and calculations which are easy to understand and easy to use to evaluate the resulting clustering process. The SICD function uses Formula 13. Where (d, o_j) shows the Euclidean distance between object *d* and the *j* center in the o_j grouping.

$$f = \sum_{j=1}^{k} \sum_{eec_j} dis(d, o_j)^2 \quad (13)$$

2.7 Statistic Test

Statistical tests are used to find out and obtain the performance of statistical analysis methods that have significant differences between the clustering carried out. The Friedman test was used in this study to compare the SICD values for each method. To find differences between groups for the dependent variable ordinal non-parametric test, Friedman's test was used. [46].

The first best SICD score will be given a rank of 1, then the second best score will be given a rank of 2, the third best score will be given a rank of 3 and if there are the same scores then a rating of the average will be given. Friedman's test uses Formula 14.

$$X_F^2 = \frac{12}{DM(M+1)} \sum_{j=1}^{M} R_j^2 - 3D(M+1) \quad (14)$$

Where *D* is the number of datasets, the total ranking of the models is R_j , the number of methods is *M*, the total ranking values per method are *j*. If the same value is obtained in giving a ranking value, then Formula 15 is used to divide the results from Formula 14.

$$1 - \frac{\sum T_i}{DM(M^2-1)} \quad (15)$$

Where $T_i = \sum(t_i^3 - t_i)$ and *t* is the number of the same value in a ranking observation in a group of data datasets. With the same value in giving a ranking in the Friedman test, the equation formula becomes as in Formula 16.

$$X_F^2 = \frac{\frac{12}{DM(M+1)} \sum_{j=1}^{M} R_j^2 - 3D(M+1)}{1 - \frac{\sum T_i}{DM(M^2-1)}} \quad (16)$$

To find out the significant difference between one method and the other tested methods, further tests need to be carried out. The follow-up test in this study uses the follow-up test with Nemenyi where this test is carried out if the H_0 hypothesis is not accepted [46]. In the Nemenyi test the clustering methods are compared. If the calculation of the average rank (AR) results in a difference (diff) value that is smaller than the critical value or Critical Difference (CD), then the comparison is said to be significantly different.

Calculation of critical value or Critical Difference using Formula 17 where $q_\propto$ is based on range statistics, D is the number of datasets and M is the sum of methods.

$$CD = q_\propto \sqrt{\frac{M(M+1)}{6D}} \quad (17)$$

2.8 Experiments

In this study experiments were carried out using auxiliary tools such as Matlab 2019, MS Excel, Notepad++, SPSS, R. Studio and Python. The specifications of the computer equipment used for processing are presented in Table 7.

Table 7. Specifications of the Computer Equipment Used

Processor	Intel(R) Core(TM) i3-2330M CPU
Memory	8GB (Dual Slot 4GB @2)
Storage	240 SSD (System)
Appearance	NVIDIA Geforce GT520M 5053MB
Operating	Windows 10 Professional 64bit

Experimental and testing stages are carried out by preparing datasets. Perform data preprocessing if there is data that does not match, such as empty crate or contains special characters. Each dataset is calculated using the proposed methods, the k-means method and the k-means++ method. Experimental experiments were carried out for each dataset and each method 10 times. Measuring the performance of the proposed methods, k-means method and k-means++ method on each dataset with the SICD function. Comparing the results of clustering performance of all methods with the best SICD function in each dataset. Non-parametric statistical test with the Friedman test and follow-up test with the Nemenyi post hoc.

3. Results and Discussions

In the experimental results, five datasets were tested 10 times using the proposed methods, k-means methods and k-means++ methods, then evaluated with SICD function. The number of clusters in the clustering process uses the number of classes that have been presented in Table 6. The number of classes presented in Table 6 is used as the number of clusters in the clustering process. The experimental results for the appearance of the worst, best and average SICD values are presented in Table 8.

Table 8. Summary of SICD results in 10 trials

Datasets	Measure	SICD		
		Proposed	K-means++	K-means
iris	Best	**96.66**	97.32	97.33
	Worst	97.35	124.18	124.18
	Average	97.191	105.162	102.477
wine	Best	**16380.75**	16555.68	16555.68
	Worst	16555.68	18436.95	18436.95
	Average	16538.187	17636.489	18060.696
glass	Best	**213.42**	**213.42**	215.87
	Worst	255.92	239.98	264.12

DOI: https://doi.org/10.29207/resti.v7i4.4997

Datasets	Measure	SICD		
		Proposed	K-means++	K-means
ecoli	Average	221.31	229.917	253.28
	Best	**62.38**	**62.38**	66.21
	Worst	66.51	68.79	74.72
cancer	Average	64.898	65.377	68.238
	Best	**2964.39**	2986.96	2986.96
	Worst	2986.96	2988.43	2986.96
	Average	2982.446	2988.283	2988.283

It can be seen that in Table 8 from the experiments conducted the appearance of the best value in bold for the iris datasets was 96.66 which was obtained in the proposed method. In the wine datasets the best occurrence of value is 16380.75 which is obtained in the proposed method. In the glass and ecoli datasets, the best occurrences of 213.41 and 62.38 were obtained in the proposed method and the k-means+ method. In the cancer datasets the best appearance of the value is 2964.39 which is obtained in the proposed method. Table 9 is the ranking results which can be seen from the average rank value and the sum of the overall results from the SICD scores, the proposed method ranks first, then the k-means++ method is in second position and k-means is in third position.

Table 9. Ranking the Evaluation Results of the Best SICD Function

Datasets	K-means	K-means++	Proposed
iris	3	2	1
wine	2.5	2.5	1
glass	3	1.5	1.5
ecoli	3	1.5	1.5
cancer	2.5	2.5	1
Sum	14	10	6
Average	2.8	2	1.2

Figure 11 is a diagram of the results of a comparison between the methods of ranking. And it was found that among other methods the proposed method has a better ranking.

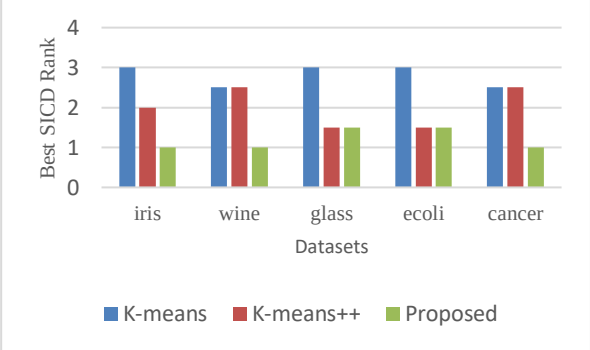

Figure 11. Method Performance Comparison Chart based on the Ranked SICD Function

To see differences between methods, a non-parametric statistical tests were performed using Friedman's test. Table 9 is the result of the Friedman Test with the help of SPSS tools where the H_0 hypothesis is if the proposed method has no significant difference to other comparison methods. Hypothesis H_1 is if there is a significant difference between the other comparison methods and the proposed method. The significance level uses $\propto$ = 5% (0.05). Table 10 is the result of the calculation of the Friedman test with SPSS tools.

Table 10. Friedman Test Calculation Results with SPSS Tools

Test Statistics	
Asymp. Sig.	0.018
Chi-Square	8
df	2
N	5
a. Friedman Test	

It can be seen in Table 10 that $\propto$ calculate 0.018 (Asymp. Sig.) < 0.05 $\propto$ significant 5%, then hypothesis H_0 is rejected, which means hypothesis H_1 is accepted with the conclusion that there is a difference between the other methods and the proposed method in handling the determination of the initial centroid in the k-means algorithm with the calculation of the SICD function and using five public datasets namely iris, wine, glass, ecoli and cancer.

Because there are significant differences in the Friedman test, to see significant differences between one method and another, a follow-up test is carried out using the Nemenyi post hoc. The calculated value of Critical Difference (CD) as a threshold parameter for determining differences between methods for the post hoc Nemenyi test obtained a CD value of 2.095958. Table 11 is a pairwise comparison of the Nemenyi test where the value strengthened in bold is a value that is greater than the CD value, namely 2.529 > 2.095958 (CD value) which means that it meets the elements of the calculation threshold for further calculations, namely to calculate the P-value to find out which method has a significant difference.

Table 11. Pairwise Comparison of the Nemenyi post hoc Test

	Proposed	K-means++	K-means
Proposed	0	1.264	**2.529**
K-means++	1.264	0	1.265
K-means	**2.529**	1.265	0

Table 12 is the calculation of the post hoc Nemenyi test P-value.

Table 12. P-value Nemenyi post hoc

	Proposed	K-means++	K-means
Proposed	1	0.41689	**0.03067**
K-means++	0.41689	1	0.41689
K-means	**0.03067**	0.41689	1

In Table 12, the P-value < 0.05 ($\propto$ value) is corroborated by bold, which means that there is a statistically significant difference between the proposed method to k-means method for the clustering method. While in the comparison of other methods there is no significant difference between one method to another method. The P-value obtained in the proposed method

DOI: https://doi.org/10.29207/resti.v7i4.4997

against the k-means method resulted in a value of 0.03067 <0.05 (∝ value) which means that the proposed method is able to provide better clustering quality and improve the performance of clustering results in the k-means algorithm and provide alternative determination cluster center.

4. Conclusion

K-means is one of the popular algorithms, simple, fast and easy to use for partition-based clustering techniques and is most often used for grouping data. However, k-means has several problems, one of which is determining the initial centroid. Careless determination of the initial centroid will affect the quality of the resulting clusters which causes less optimal clustering results and can produce poor cluster results. To fix the k-means problem, in this research an optimization method for determining the initial centroid is proposed using the Antlion Optimizer (ALO) method which is modified on the number of iterations and adjustments to the initial input.

Testing the proposed method uses five public datasets, namely glass, iris, wine, cancer and ecoli obtained from UCI Machine Learning which are measured and evaluated with the Sum of Intra-Cluster Distance (SICD) function. Then the proposed method is compared with the k-means method and the k-means++ method. Both of these methods were found in the Literature Systematic Review (SLR) conducted in this study on answers to questions RQ5 (a method often used as a comparison method).

The best SICD values for the iris, wine and cancer datasets were found in the proposed method while the glass and ecoli datasets were found in the proposed method and the k-means++ method. This indicates that the proposed method occupies the average value in the first position, then the k-means++ method in the second position and then the third position is occupied by the k-means method.

Non-parametric statistical tests were performed using the Friedman test was carried out to see whether there is a difference between the methods or not. The results of the Friedman test found that the P-value (Asymp. Sig) obtained was 0.018 which was less than 0.05 (∝ value) indicating that there was a significant difference between the methods.

To find out the differences between the methods, a follow-up test with the Nemenyi procedure was used by looking at the mean of rank parameter. From the pairwise comparison test between methods, one difference was obtained with a value above the critical difference threshold value of 2.095958 equal to the difference value of 2.529 found in the proposed method of the k-means method. Calculation of the P-value in the Nemenyi follow-up test obtained a P-value of 0.03067 less than 0.05 (∝ value), which means that there is a significant difference between the proposed method and the k-means method.

The results showed that the modified Antlion Optimizer (ALO) method in this study, namely the number of iterations and initial input adjustments, resulted in an increase in clustering performance on the k-means method as an optimization method for determining the initial centroid and providing contributions and answers to problems and objectives. In this research is to improve clustering performance on the k-means algorithm and provide alternative solutions for determining the initial center of the cluster and provide clustering performance solutions that further improve the k-means algorithm.

References

[1] J. Han and M. Kamber, "Data Mining Concepts and Techniques Second Edition," *San Fr. Morgan Kauffman*, 2001.

[2] D. T. Larose, "Discovering Knowledge in Data: In An Introduction to Data Mining," *Wiley-Interscience*, 2005.

[3] P. N. Tan, M. Steinbach, and V. Kumar, "Introduction to Data Mining, (First Edition)," *Addison-Wesley Longman Publ. Co*, 2005.

[4] J. Chen, X. Qi, L. Chen, F. Chen, and G. Cheng, "Quantum-inspired ant lion optimized hybrid k-means for cluster analysis and intrusion detection," *Knowledge-Based Syst.*, vol. 203, p. 106167, 2020, doi: 10.1016/j.knosys.2020.106167.

[5] K. M. Kumar and A. R. M. Reddy, "An efficient k-means clustering filtering algorithm using density based initial cluster centers," *Inf. Sci. (Ny).*, vol. 418–419, pp. 286–301, 2017, doi: 10.1016/j.ins.2017.07.036.

[6] N. Nidheesh, K. A. Abdul Nazeer, and P. M. Ameer, "An enhanced deterministic K-Means clustering algorithm for cancer subtype prediction from gene expression data," *Comput. Biol. Med.*, vol. 91, pp. 213–221, 2017, doi: 10.1016/j.compbiomed.2017.10.014.

[7] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proc. fifth Berkeley Symp. Math. Stat. Probab.*, vol. 1, no. 14, pp. 281–297, 1967.

[8] E. Forgy, "Cluster Analysis of Multivariate Data: Efficiency versus Interpretability of Classification," *Biometrics*, vol. 21, pp. 768–769, 1965.

[9] M. E. Celebi, H. A. Kingravi, and P. A. Vela, "A comparative study of efficient initialization methods for the k-means clustering algorithm," *Expert Syst. Appl.*, vol. 40, no. 1, pp. 200–210, 2013, doi: 10.1016/j.eswa.2012.07.021.

[10] G. Zhang, C. Zhang, and H. Zhang, "Improved K-means algorithm based on density Canopy," *Knowledge-Based Syst.*, vol. 145, pp. 289–297, 2018, doi: 10.1016/j.knosys.2018.01.031.

[11] Y. Zhou, R. Xie, T. Zhang, and J. Holguin-Veras, "Joint Distribution Center Location Problem for Restaurant Industry Based on Improved K-Means Algorithm with Penalty," *IEEE Access*, vol. 8, pp. 37746–37755, 2020, doi: 10.1109/ACCESS.2020.2975449.

[12] M. Capó, A. Pérez, and J. A. Lozano, "An efficient approximation to the K-means clustering for massive data," *Knowledge-Based Syst.*, vol. 117, pp. 56–69, 2017, doi: 10.1016/j.knosys.2016.06.031.

[13] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," *Proc. Annu. ACM-SIAM Symp. Discret. Algorithms*, vol. 07–09, pp. 1027–1035, 2007.

[14] Z. Wu and Z. Wu, "An Enhanced Regularized k-Means Type Clustering Algorithm with Adaptive Weights," *IEEE Access*, vol. 8, pp. 31171–31179, 2020, doi: 10.1109/ACCESS.2020.2972333.

DOI: https://doi.org/10.29207/resti.v7i4.4997

[15] T. Su and J. G. Dy, "In search of deterministic methods for initializing K-means and Gaussian mixture clustering," *Intell. Data Anal.*, vol. 11, no. 4, pp. 319–338, 2007, doi: 10.3233/ida-2007-11402.

[16] P. S. Bradley and U. M. Fayyad, "Refining Initial Points for K-Means Clustering," *Proc. 15th int. conf. Mach. Learn.*, vol. 54, no. 6, pp. 91–99, 1998, doi: 10.7567/JJAP.54.061701.

[17] T. F. Gonzalez, "Clustering to minimize the maximum intercluster distance," *Theor. Comput. Sci.*, vol. 38, no. C, pp. 293–306, 1985, doi: 10.1016/0304-3975(85)90224-5.

[18] I. Cabria and I. Gondra, "Potential-K-Means for Load Balancing and Cost Minimization in Mobile Recycling Network," *IEEE Syst. J.*, vol. 11, no. 1, pp. 242–249, 2017, doi: 10.1109/JSYST.2014.2363156.

[19] E. Parzen, "On the Estimation of Probability Density Functions and Mode," *Ann. Math. Stat.*, vol. 33, pp. 1065–1076, 1962.

[20] S. P. Lloyd, "Least Squares Quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, 1982, doi: 10.1109/TIT.1982.1056489.

[21] S. Khanmohammadi, N. Adibeig, and S. Shanehbandy, "An improved overlapping k-means clustering method for medical applications," *Expert Syst. Appl.*, vol. 67, pp. 12–18, 2017, doi: 10.1016/j.eswa.2016.09.025.

[22] B. Zhang, M. Hsu, and U. Dayal, "K-Harmonic means - A data clustering algorithm," *HP Lab. Tech. Rep.*, no. 124, p. Hewlett-Packard Labs, 1999.

[23] B. Zhang, "Generalized K-Harmonic Means," *Tech. Rep.*, p. Hewlett-Packard Laboratoris, 2000.

[24] X. Lan, Q. Li, and Y. Zheng, "Density K-means: A new algorithm for centers initialization for K-means," *Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS*, vol. 2015-Novem, pp. 958–961, 2015, doi: 10.1109/ICSESS.2015.7339213.

[25] A. McCallum, K. Nigam, and L. H. Ungar, "Efficient clustering of high-dimensional data sets with application to reference matching," *Proceeding Sixth ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 169–178, 2000, doi: 10.1145/347090.347123.

[26] A. H. Kashan, "League Championship Algorithm: A new algorithm for numerical function optimization," *SoCPaR 2009 - Soft Comput. Pattern Recognit.*, pp. 43–48, 2009, doi: 10.1109/SoCPaR.2009.21.

[27] T. Wangchamhan, S. Chiewchanwattana, and K. Sunat, "Efficient algorithms based on the k-means and Chaotic League Championship Algorithm for numeric, categorical, and mixed-type data clustering," *Expert Syst. Appl.*, vol. 90, pp. 146–167, 2017, doi: 10.1016/j.eswa.2017.08.004.

[28] S. Mirjalili, "The ant lion optimizer," *Adv. Eng. Softw.*, vol. 83, pp. 80–98, 2015, doi: 10.1016/j.advengsoft.2015.01.010.

[29] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," *EBSE Tech. Rep. Version 2.3, EBSE-2007*, 2007.

[30] R. S. Wahono, "A Systematic Literature Review of Software Defect Prediction: Research Trends, Datasets, Methods and Frameworks," *J. Softw. Eng.*, vol. 1, no. 1, pp. 1–16, 2015, doi: 2356-3974.

[31] C. Catal, "Software fault prediction: A literature review and current trends," *Expert Syst. Appl.*, vol. 38, no. 4, pp. 4626–4636, 2011, doi: 10.1016/j.eswa.2010.10.024.

[32] S. Kant, T. Mahara, V. Kumar Jain, D. Kumar Jain, and A. K. Sangaiah, "LeaderRank based k-means clustering initialization method for collaborative filtering," *Comput. Electr. Eng.*, vol. 69, pp. 598–609, 2018, doi: 10.1016/j.compeleceng.2017.12.001.

[33] S. S. Yu, S. W. Chu, C. M. Wang, Y. K. Chan, and T. C. Chang, "Two improved k-means algorithms," *Appl. Soft Comput. J.*, vol. 68, pp. 747–755, 2018, doi: 10.1016/j.asoc.2017.08.032.

[34] H. Ismkhan, "I-k-means−+: An iterative clustering algorithm based on an enhanced version of the k-means," *Pattern Recognit.*, vol. 79, pp. 402–413, 2018, doi: 10.1016/j.patcog.2018.02.015.

[35] S. Zhang and S. Ge, "User Power Interaction Behavior Clustering Analysis That is Based on the Self-Organizing-Center K-Means Algorithm," *IEEE Access*, vol. 7, pp. 175879–175888, 2019, doi: 10.1109/ACCESS.2019.2957922.

[36] S. F. Hussain and M. Haris, "A k-means based co-clustering (kCC) algorithm for sparse, high dimensional data," *Expert Syst. Appl.*, vol. 118, pp. 20–34, 2019, doi: 10.1016/j.eswa.2018.09.006.

[37] Y. Gu, K. Li, Z. Guo, and Y. Wang, "Semi-supervised k-means ddos detection method using hybrid feature selection algorithm," *IEEE Access*, vol. 7, pp. 64351–64365, 2019, doi: 10.1109/ACCESS.2019.2917532.

[38] H. Xie *et al.*, "Improving K-means clustering with enhanced Firefly Algorithms," *Appl. Soft Comput. J.*, vol. 84, p. 105763, 2019, doi: 10.1016/j.asoc.2019.105763.

[39] J. Chen, M. Tian, X. Qi, W. Wang, and Y. Liu, "A solution to reconstruct cross-cut shredded text documents based on constrained seed K-means algorithm and ant colony algorithm," *Expert Syst. Appl.*, vol. 127, pp. 35–46, 2019, doi: 10.1016/j.eswa.2019.02.039.

[40] K. Tian, J. Li, J. Zeng, A. Evans, and L. Zhang, "Segmentation of tomato leaf images based on adaptive clustering number of K-means algorithm," *Comput. Electron. Agric.*, vol. 165, no. March, p. 104962, 2019, doi: 10.1016/j.compag.2019.104962.

[41] J. Zhu, Z. Jiang, G. D. Evangelidis, C. Zhang, S. Pang, and Z. Li, "Efficient registration of multi-view point sets by K-means clustering," *Inf. Sci. (Ny).*, vol. 488, pp. 205–218, 2019, doi: 10.1016/j.ins.2019.03.024.

[42] H. Kim, H. K. Kim, and S. Cho, "Improving spherical k-means for document clustering: Fast initialization, sparse centroid projection, and efficient cluster labeling," *Expert Syst. Appl.*, vol. 150, p. 113288, 2020, doi: 10.1016/j.eswa.2020.113288.

[43] S. Manochandar, M. Punniyamoorthy, and R. K. Jeyachitra, "Development of new seed with modified validity measures for k-means clustering," *Comput. Ind. Eng.*, vol. 141, p. 106290, 2020, doi: 10.1016/j.cie.2020.106290.

[44] E. Umamaheswari, S. Ganesan, M. Abirami, and S. Subramanian, "Cost Effective Integrated Maintenance Scheduling in Power Systems using Ant Lion Optimizer," *Energy Procedia*, vol. 117, pp. 501–508, 2017, doi: 10.1016/j.egypro.2017.05.176.

[45] M. Jain, V. Singh, and A. Rani, "A novel nature-inspired algorithm for optimization: Squirrel search algorithm," *Swarm Evol. Comput.*, vol. 44, no. November 2017, pp. 148–175, 2019, doi: 10.1016/j.swevo.2018.02.013.

[46] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.

DOI: https://doi.org/10.29207/resti.v7i4.4997