Application of Particle Swarm Optimization (PSO) to Improve K-means Accuracy in Clustering Eligible Province to Receive Fish Seed Assistance in Java

Nur Syahrani Majdina¹, M Arief Soeleman², Catur Supriyanto³
¹(Student of Masters of Informatics Engineering, Dian Nuswantoro University, Semarang, Indonesia)
²³(Lecturer in Informatics Engineering, Dian Nuswantoro University, Semarang, Indonesia)

Abstract:
Background: The k-means algorithm is to divides the data into clusters based on the closest distance by using the Euclidean distance formula. K-means is often used because the resulting approach is easy to implement, but has the disadvantage that the central point depends on the choice of K, resulting in a decrease in the speed and quality of a cluster.

Methodology: The research method uses the Cross-Industry Standard Process for Data Mining (CRISP-DM) cycle using 6 phases. The purpose of this research is to automate the selection of numbering on K-means so as to further increase the speed and quality of a cluster.

Discussion: The algorithm used is the K-means algorithm and Particle Swarm Optimization (PSO) which is validated using the K-NN algorithm because if it is calculated using the confusion matrix it has an accuracy of 97.78% and has a Davies-Bouldin Index (DBI) value of 0.369877333 which is included in the high category when applied to data on the volume of fishery products on the island of Java.

Conclusion: The final result of this study is that it has succeeded in automating the selection of numbering on K-means so that the speed and quality of a cluster is to determine which provinces are entitled to receive assistance in the form of fish seeds to increase the volume of fishery products on the island of Java because of the results of the calculation of K-means clustering and Particle Swarm Optimization (PSO) which was tested using the K-NN classification which was calculated using the confusion matrix and the Davies-Bouldin Index (DBI) the accuracy value increased, there were 261 fishery production volume data that we're entitled to receive assistance in the form of fish seeds on the island of Java.

Key Word: Clustering; Fishery Production Volume; K-means Algorithm; PSO Algorithm.

Date of Submission: 04-01-2022
Date of Acceptance: 15-01-2022

I. Introduction

The clustering algorithm is to divides the data into clusters based on the shortest distance between the data and the central point in each cluster [1]. One of the clustering algorithms is K-means. K-means is to divide the data into clusters based on the closest distance using the Euclidean distance formula. K-means is often used because the computational approach is easy to implement [2]. This research will apply K-means.

The application of K-means has been widely carried out in the multimedia field, for example for handwriting recognition, face recognition, and image segmentation [8]. K-means is also applied in the field of image segmentation, for example for the development of software applications for the reconstruction of 3D models of the heart and liver [9]. In addition, K-means is also applied in the field of dimension reduction by integrating SVM for diabetes diagnosis [10], where K-means is used to look for patterns.

In the K-means algorithm, the resulting center point depends on the choice of K, resulting in a decrease in the speed and quality of a cluster [2]. In this study, the research problem will be raised on choice K.

The researcher tries to solve the problem of choice K using several methods, namely Cross-Validation [3], The Cluster Center and The Nearest Neighbor Cluster (VCN) [4], Gap Statistics [5], and Cluster Number Assisted K-means. (CNAK). [6]. This method can solve the problem with choice K, but it is not suitable for handling very large data sets, because it requires large storage space.

This study uses Particle Swarm Optimization (PSO), K-means, and Distance methods to overcome the problem of choosing a center point and streamline time when calculating the distance in each iteration, using a dataset from data.go.id, namely data on the volume of fishery production in Indonesia. Indonesia. Java Island.

The K-means algorithm is quite easy to implement [2] but has the disadvantage that the resulting center point depends on the choice of K, resulting in a decrease in the speed and quality of a cluster.
The Particle Swarm Optimization (PSO) algorithm is an optimization method that has been proven effective in solving the K problem [7]. In this study, Particle Swarm Optimization (PSO), K-means, and Distance will be applied to overcome the problem of K choice and save time in calculating the distance in each iteration by using a dataset from data.go.id, namely data on the volume of fishery production in Java Island.

II. Related Work

Based on [1] the goal is to obtain information to determine policies and assist the community in utilizing information on the potential and drawbacks of the fisheries sector in each district or city to develop business. This study using fuzzy k-means using a dataset of fishery production volume in East Java. The result of the research is that there are 3 clusters, cluster 1 is 9 districts/cities in East Java which is included in the lowest cluster of all clusters. Cluster 2 is 3 districts or cities in East Java that are included in the best cluster. Cluster 3, which is 26 districts or cities in East Java, is ranked second compared to other clusters.

Based on [4] the purpose of this study is to improve the quality to predict the number of K-means clusters if the selection of the resulting center point depends on the choice of K, numbering is done using cross-validation using an iris dataset which is split into 10 parts. The results of the research are quite consistent and can increase the speed and quality to predict the number of K-means clusters but are not suitable to be applied to big data.

Based on [5] the goal is to improve the quality to predict the number of K-means clusters if the selection of the resulting center point depends on the choice of K, numbering is done using the gap statistic, using k-means clustering via the gap statistics, and using the UCI dataset, namely iris, breast cancer, wine, vowel, and glass. The result is that it can improve the quality for predicting the number of K-means clusters but is not suitable to be applied to big data.

Based on [7] the purpose of this study is to improve the quality to predict the number of K-means clusters if the selection of the resulting center point depends on the choice of K. In this study using the Cluster Number Assisted K-means (CNAK) algorithm using the UCI dataset, namely iris, breast cancer, wine, vowel, and glass. The result is that it can improve the quality to predict the number of K-means clusters but it is less effective because it takes a long time.

Based on [2] the purpose of this study is to improve the quality for predicting the central point of K-means if the selection of the resulting center point depends on the choice of K, on nonlinear partition clustering. This study using FAPSO (fuzzy adaptive particle swarm optimization), ACO, and K-means using a benchmark dataset. The result is that the FAPSO (fuzzy adaptive particle swarm optimization) algorithm, ACO, and K-means have better performance than PSO, ACO, simulated annealing (SA), PSO and SA, ACO and SA, PSO and ACO, genetic algorithm (GA), tabu search (TS), honey bee mating optimization (HBMO) and k-means.

III. Methodology

3.1 Research method

The research method is the Cross-Industry Standard Process for Data Mining (CRISP-DM) cycle using 6 phases, namely:

![CRISP-DM Fase phase](image)

Figure 3.1: CRISP-DM Fase phase
3.1.1 Business Understanding Phase
This study aims to apply the Particle Swarm Optimization (PSO), K-means and Distance methods. Where Particle Swarm Optimization (PSO) is used to automate the selection of numbering on K-means so as to further improve the quality of a cluster. K-means clustering is used to assist in grouping the data and Distances are used to streamline time when calculating the distance in each iteration, so as to increase the speed and quality of a cluster. If it is done manually, it takes a long time and is not necessarily accurate.

3.1.2 Data Comprehension Phase
The data used in this study are datasets from data.go.id, namely data on the volume of fishery production in Indonesia, which amounted to 76792 data, but to narrow the topic of discussion, based on the problem limitations, only fishery production volumes were used on the island of Java in 2012 so that only 315 were used.

3.1.3 Data Processing Phase
In this study, only 2 stages of data mining were used in the data processing phase, namely data cleaning and data selection. Data cleaning is used to remove empty data. The results of the data after cleaning:

![Figure 3.2: Data after cleaning](image)

After cleaning the data, the next process is data selection. This study using Particle Swarm Optimization (PSO). The following is the data after the selection process is carried out:

![Figure 3.3: Data after selection](image)

3.1.4 Modeling Phase
The modeling phase is carried out by choosing a suitable modeling technique to be applied or implemented in the dataset from data.go.id, namely data on the volume of fishery products on the island of Java, and adjusting the existing model rules which aim to optimize the results of this research. In this study, the most suitable algorithm used on the data is to automate the selection of the numbering on the K-means so as to further improve the quality of a cluster by using Particle Swarm Optimization (PSO), clustering using the K-means algorithm, and making time-efficient when calculating the distance to the data. each iteration using Distance.

3.1.5 Evaluation Phase
The evaluation phase process in this study uses the davies-Bouldin index (DBI), to calculate cluster performance in the K-means algorithm and the confusion matrix to calculate the accuracy of the K-means algorithm and Particle Swarm Optimization (PSO).

3.1.6 Implementation Phase
The stages in an analysis process are the first step, which is to determine what software will be used to measure the performance of the K-means and Particle Swarm Optimization (PSO) algorithms, perform data cleaning, select data using Particle Swarm Optimization (PSO), and perform testing of previously obtained data.

3.2 Experiment
The proposed method is the use of k-means to determine which provinces are entitled to receive assistance in the form of fish seeds to increase the volume of fishery products on the island of Java. To increase the speed and quality of clustering volume of fishery production in Java, the Particle Swarm Optimization (PSO) algorithm is used to automate the selection of numbering on K-means so as to further improve the quality of a
cluster, and Distance is used to streamline time when calculating the distance in each iteration, So that it can increase the speed and quality of a cluster.

Data on the volume of fishery products on the island of Java
Pre-processing
- Data cleaning
- Data selection

Initialization of velocity and particle population
Initialize phbest and gbest fitness values

Particle evaluation
Particle speed update
Particle position update

Is the fitness value better than phbest?

Yes
Particle position update

No

Is the fitness value better than gbest?

Yes
Set as new gbest

No
Max Iteration

No
Application of gbest in the k-means algorithm

K-means

Performance of clustering accuracy of fishery production volume in Java

Figure 3.1 : Experiment Steps

IV. Discussion

Here are the results of manual calculations, for example, when looking for id 1 and the name of the province of West Java, it will display id 1, the name of the province of West Java, the name of catfish, volume 8475, selling value 140209383 and one table id 1, data 1 is volume 8475, data 2 namely the selling value of 140209383, included in cluster 2 which is included in the medium category.

a. K-means Clustering Testing

Taken from data on the volume of fishery production based on random data, which amounted to three data used to find clustering results using the K-means algorithm, namely volume and selling value, namely:

M1 = data on id 1 in West Java province, the name of catfish which has a volume of 8475 and a selling value of 140209383

M2 = data on id 2 in West Java province, the name of cendro fish which has a volume of 61 and a selling value of 596836

M3 = data on id 29 in the province of West Java, the name of red snapper which has a volume of 5940 and a selling value of 234812448
The centroid of each cluster randomly

<table>
<thead>
<tr>
<th>Cluster</th>
<th>X1</th>
<th>X2</th>
</tr>
</thead>
<tbody>
<tr>
<td>c1</td>
<td>100</td>
<td>400000</td>
</tr>
<tr>
<td>c2</td>
<td>8000</td>
<td>100000000</td>
</tr>
<tr>
<td>c3</td>
<td>5940</td>
<td>234812448</td>
</tr>
</tbody>
</table>

Enter the object to the nearest centroid, using the euclidian distance formula, where c1 is the low cluster, c2 is the medium cluster and c2 is the high cluster. Results of the 1st iteration:

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>c1</td>
<td>M2</td>
</tr>
<tr>
<td>c2</td>
<td>M1</td>
</tr>
<tr>
<td>c3</td>
<td>M3</td>
</tr>
</tbody>
</table>

The new center point is obtained:

<table>
<thead>
<tr>
<th>Cluster</th>
<th>X1</th>
<th>X2</th>
</tr>
</thead>
<tbody>
<tr>
<td>c1</td>
<td>61</td>
<td>596836</td>
</tr>
<tr>
<td>c2</td>
<td>8475</td>
<td>140209383</td>
</tr>
<tr>
<td>c3</td>
<td>5940</td>
<td>234812448</td>
</tr>
</tbody>
</table>

Performed calculations using the euclidian distance formula using the new center point. Then the final result of the grouping of the training data iteration-2 is obtained:

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>c1</td>
<td>M2</td>
</tr>
<tr>
<td>c2</td>
<td>M1</td>
</tr>
<tr>
<td>c3</td>
<td>M3</td>
</tr>
</tbody>
</table>

Since no data has changed, the calculation is stopped.

Based on comparing the results of the K-means clustering calculation which is calculated manually and the results of the rapidminer 8.1 tools, it is appropriate.

b. Analysis of Research Results

After getting the results of the K-means clustering calculation and Particle Swarm Optimization (PSO) to get the label, it will be validated using classification with the K-NN algorithm, then the analysis of the research results is carried out using a confusion matrix to find accuracy in the training data by using as many as test data. 315 data, namely:

Accuracy = \((TP + TN))/(Number\ of\ cases) \times 100\% = 308/315 \times 100\% = 97.78\%\

K-means algorithm and Particle Swarm Optimization (PSO) validated by classification using the Naïve Bayes algorithm:

Accuracy = \((TP + TN))/(Number\ of\ cases) \times 100\% = 304/315 \times 100\% = 96.50\%\

The K-means algorithm is validated by classification using the K-NN algorithm:

Accuracy = \((TP + TN))/(Number\ of\ cases) \times 100\% = 250/315 \times 100\% = 79.36\%\

The K-means algorithm is validated by classification using the Decision Tree algorithm:

Accuracy = \((TP + TN))/(Number\ of\ cases) \times 100\% = 240/315 \times 100\% = 76.19\%\

Based on the results of the above calculations, it can be concluded that clustering using the K-means algorithm and Particle Swarm Optimization (PSO) has high accuracy when applied to fishery production volume data.
c. K-means Evaluation
In this final project, the Davies-Bouldin Index (DBI) matrix is used. Taken from fishery production volume data based on three random data, which are used to find clustering results using the K-means and Particle Swarm Optimization (PSO) algorithm, namely the volume and selling value as follows:
M1 = data on id 1 in West Java province, the name of catfish which has a volume of 8475 and a selling value of 140209383
M2 = data on id 2 in West Java province, the name of cendro fish which has a volume of 61 and a selling value of 596836
M3 = data on id 29 in the province of West Java, the name of red snapper which has a volume of 5940 and a selling value of 234812448
The training data used will be divided into 3 clusters (c1, c2, and c3)

<table>
<thead>
<tr>
<th>Data ke-i</th>
<th>Fitur</th>
<th>Klaster</th>
<th>Centroid</th>
<th>Jarak ke centroid</th>
<th>SSW</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Volume</td>
<td>Nilai jual</td>
<td>Volume</td>
<td>Nilai jual</td>
<td>196836,0039</td>
</tr>
<tr>
<td>2</td>
<td>61</td>
<td>596836</td>
<td>2</td>
<td>8000</td>
<td>100000000</td>
</tr>
<tr>
<td>3</td>
<td>5940</td>
<td>234812448</td>
<td>3</td>
<td>5940</td>
<td>234812448</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data ke-i</th>
<th>SSB</th>
<th>Data ke-i</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>99600000,31</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>99600000,31</td>
<td>134812448</td>
</tr>
<tr>
<td>3</td>
<td>234812448,1</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data ke-i</th>
<th>R</th>
<th>Rmax</th>
<th>DBI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0,098777333</td>
</tr>
<tr>
<td>2</td>
<td>0,405685</td>
<td>0</td>
<td>0,405685</td>
</tr>
<tr>
<td>3</td>
<td>0,000840</td>
<td>0,298262</td>
<td>0,298262</td>
</tr>
</tbody>
</table>

Based on the results of the calculations above, it can be concluded that the cluster that uses the K-means algorithm and Particle Swarm Optimization (PSO) is included in the good cluster category because the SSW and DBI values are small, namely non-negative ≥ 0.

V. Conclusion
This research was successful because the results of the accuracy calculation using the confusion matrix were 97.78% which was included in the high category and the Davies-Bouldin Index (DBI) value was 0.369877333 which was included in the good cluster category.

References