A Performance of the Scattered Averaging Technique based on the Dataset for the Cluster Center Initialization

Arief Bramanto Wicaksono Putra, Achmad Fanany Omnita Gaffar, Bedi Supraptty, Mulyanto
Department of Information Technology, Politeknik Negeri Samarinda, East Kalimantan, Indonesia
Email: ariefbram@gmail.com; onnygaffar212@gmail.com; bedirheody@gmail.com; yanto1294@gmail.com

Received: 06 December 2020; Accepted: 18 February 2021; Published: 08 April 2021

Abstract: Clustering is one of the primary functions in data mining explorations and statistical data analysis which widely used in various fields. There are two types of the clustering algorithms which try to optimize certain objective function, i.e. the hierarchical and partitional clustering. This study focuses on the achievement of the best cluster results of the hard and soft clustering (K-Mean, FCM, and SOM clustering). The validation index called GOS (Global Optimum Solution) used to evaluate the cluster results. GOS index defined as a ratio of the distance variance within a cluster to the distance variance between clusters. The aim of this study is to produce the best GOS index through the use of proposed method called the scattered averaging technique based on datasets for the cluster center initialization. The cluster results of each algorithm are also compared to determine the best GOS index between them. By using the annual rainfall data as the dataset, the results of this study showed that the proposed method significantly improved K-Mean clustering ability to achieve the global optimum solution with a performance ratio of 69.05% of the total performance of the three algorithms. The next best clustering algorithm is SOM clustering (24.65%) followed by FCM clustering (6.30%). In addition, the results of this study also showed that the three clustering algorithms achieve their best global optimum solution at the number of even clusters.

Index Terms: Global Optimum Solution; K-Mean; FCM; SOM; Scattered averaging technique.

1. Introduction

Clustering is one of the primary functions in data mining explorations and statistical data analysis which widely used in various fields. The purpose of clustering is to group datasets into groups or clusters based on the relationships between them. There are various algorithms for clustering, depend on the relationship of between the data in datasets and how the clusters are formed efficiently. The types of the relationship between data commonly are the distance of between data, connectivity between data, the density of the data space, the interval or the distribution of certain statistics, and the other definition of their relationships. From the standpoint of how to achieve efficient cluster formation, clustering regarded as a multi-purpose optimization problem.

There are two types of the clustering algorithms which try to optimize certain objective function, i.e. the hierarchical and partitional clustering. Hierarchical clustering builds clusters using dendrogram based on agglomerative or divisive approaches. This algorithm has been used to measure stock market liquidity in [1-3]. Partitional clustering algorithms attempt to partition the datasets directly into a given number of clusters where each cluster characterized by a cluster representative or vector prototype (a collection of the cluster center called centroid). The most common inter-data relationship used in partitional clustering is the shortest distance. The most common optimization goal is until there is no change in cluster membership during the grouping process. The K-Mean Clustering is the most primitive, simple and effective method. The relationship of between its datasets and the center of its cluster is the shortest distance. This algorithm has been used to cluster the earthquake epicenter in [4], to mining the attitude of social network users in [5], and mapping of image and video in [6]. The uniform effects of this algorithm have studied in [7]. The Genetic Algorithm has also been used to optimize K-Mean for GPS data clustering in [8].

K-Mean clustering is classified as hard clustering because each data pattern occupies only one cluster. The cluster membership stated by 1 and vice versa expressed by 0. Unlike the hard clustering, each data pattern of soft clustering has a certain degree of membership on each cluster expressed in the interval [0...1]. One type of this algorithm is Fuzzy C-Mean (FCM). The FCM has been used to perform the meta-analysis of cell-specific transcriptomic data discovers versatile viral responsive genes in [9]. This algorithm also used to segment an image in [10, 11] and classify the mental task in [12]. Combined FCM and PCA have been studied to modify the multiple generalized regression neural network in [13]. In-depth analyses of the convergence of between the K-Mean and FCM algorithms have studied in [14].
Other clustering types are model-based clustering [15-18], density-based clustering [19, 20], and spatial clustering [21]. SOM (Self-Organizing Maps) is one of Artificial Neural Network architecture used for clustering that works with one type of machine learning that is unsupervised learning. SOM has been used for geospatial analysis of extreme weather events in [22] and reduce the complexity of terrestrial lidar data in [23]. The asymmetric K-Mean compare to the asymmetric K-Mean have studied in [24]. Several other studies have also been carried out related to the application of SOM clustering [25-28].

The validation index measure how well the general goal of clustering which is consists of two parameters, namely the similarity within clusters and separability between clusters. The clustering results can be evaluated by using SC (Silhouette coefficient) [29-31], PBM (Pakhira-Bandyopadhyay-Maulik) [32], SSE (Sum of Squared Errors), and other techniques. SC allows for the evaluation of a given assignment for a particular observation. PBM measure the goodness of clustering on different partitions of a given dataset and describe a cluster validity index of a cluster solution. SSE is the most popular evaluation of distance which only needs to consider the cohesion of clusters to evaluate the cluster quality [33]. WB index is one of the clustering results evaluation based on SSW (sum of squares within a cluster) used to measure the similarity within a cluster and SSB SSW (sum of squares between clusters) used to measure the separability between clusters [34].

Various optimization techniques related to clustering activities with different aims and objectives have also been studied as in [35-38]. Optimization of cluster center initialization is one approach that has been widely used to obtain a good validation index measurement of clustering activities. Several studies related to this have been carried out in [20, 39-42].

This study focuses on the achievement of the best cluster results of the hard and soft clustering (K-Mean, FCM, and SOM clustering). The validation index called GOS (Global Optimum Solution) used to evaluate the cluster results. GOS index defined as a ratio of the distance variance within a cluster to the distance variance between clusters. The aim of this study is to produce the best GOS index through the use of the proposed method called the scattered averaging technique based on datasets for the cluster center initialization. The cluster results of each algorithm are also compared to determine the best GOS index between them.

2. Materials and Methods

A. The scattered averaging technique

One of the simplest techniques to measure pattern the dataset to be grouped can be mapped into the N-dimensional Cartesian coordinates, where N is the number of data attributes. The clustering process performed on N-dimensional Cartesian coordinates that act as the data space. The scattered averaging technique is a proposed technique for cluster central initialization where the initial cluster center (called the initial centroid) placed scattered in the data space based on its dataset. This technique is expected to ensure the achievement of global optimum solutions regardless of how many iterations the process requires. Suppose there are 6 data has two attributes (X and Y) as follows:

<table>
<thead>
<tr>
<th>X</th>
<th>6</th>
<th>3</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>1</th>
<th>8</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>1</td>
<td>5</td>
<td>8</td>
<td>7</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>7</td>
</tr>
</tbody>
</table>

Firstly, the dataset is sorted ascending by all its attributes (by X, then by Y) as follows:

<table>
<thead>
<tr>
<th>X</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>1</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>3</td>
<td>8</td>
<td>5</td>
<td>7</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>7</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

If the dataset will be grouped into two clusters, then the initial cluster center as follows:

<table>
<thead>
<tr>
<th>X</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>3</td>
<td>8</td>
<td>5</td>
<td>7</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>7</td>
<td>4</td>
</tr>
</tbody>
</table>

If the dataset will be grouped into three clusters then the initial cluster center as follows:

<table>
<thead>
<tr>
<th>X</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>3</td>
<td>8</td>
<td>5</td>
<td>7</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>7</td>
<td>4</td>
</tr>
</tbody>
</table>

Copyright © 2021 MECS

I.J. Modern Education and Computer Science, 2021, 2, 40-50
The illustrations are shown in Fig. 1.

![Illustration of the cluster center initialization using the scattered averaging technique](image1)

If $D$ is a dataset, $M$ is the number of data, and $k$ is the number of clusters desired then the cluster center initialization using the scattered averaging technique is expressed by:

$$kk = \text{floor} \left( \frac{M}{k} \right)$$

$$c \left( i: \right) = D \left( \cdot, \text{floor} \left( \frac{kk}{2} \right) + (i - 1) \ast kk \right)$$

where $c \left( i: \right)$ is the $i$th cluster center initialization, and ($\cdot$) is the index notation of the number of data attributes. This technique will be applied to the $K$-Mean algorithm, $FCM$, and $SOM$ clustering to know its performance through measurement of $GOS$ index of cluster result of each algorithm.

**B. The Global Optimum Solution (GOS)**

A centroid is the cluster center. The best $GOS$ achieved if the distance variance within a cluster as minimum as possible and the distance variance between clusters as maximum as possible. The $GOS$ illustrated in Fig. 2. Suppose the data as follows:

$$X = [x_1, x_2, \ldots, x_M]$$

where $M$ is the number of data. The measurement of data distribution to the average data is expressed by:

$$Var \left[ X \right] = E \left[ (X - \mu_x)^2 \right] = \frac{1}{n} \sum_{i=1}^{n} (X - \mu_x)^2$$

where $Var \left[ X \right]$ is the variance of $X$ and $\mu_x$ is the average of $X$. 

![Illustration of the global optimum solution](image2)

(a). Not a global optimum solution

(b). A global optimum solution
If \( \text{Var(\text{member})} \) represents the average of the distance variance between the cluster members against its centroid and \( \text{Var(\text{centroid})} \) represents the distance variance of between their centroids, then the global optimum solution is reached if it satisfies:

\[
F = \min (\text{Var(\text{member})}) \text{ AND } \max (\text{Var(\text{centroid})})
\]  

The GOS index is expressed by:

\[
\text{GOS} = \frac{\text{Var(\text{centroid})}}{\text{Var(\text{member})}}
\]

The formula shows that the smaller the \( \text{Var(\text{member})} \), the higher the similarity between cluster members in a cluster. Besides, the larger the \( \text{Var(\text{centroid})} \), the higher the separation between clusters. Finally, the higher the GOS index, the better the clustering results.

C. K-Mean clustering

K-Mean clustering is an iterative process that partitions a dataset into a number of \( K \) clusters. K-Mean clustering is classified as hard clustering because the cluster membership stated by 1 and vice versa expressed by 0. The initial cluster center (called centroid) used as the starting point of the partition based on the shortest distance between the cluster center against each data in the dataset. The iteration process is done continuously until no change of cluster membership in each cluster. The distance function used is Euclidean distance [43-46]. The K-Mean clustering algorithm is shown in Fig. 3.

![Algorithm of K-Mean clustering](image)

Fig. 3. The algorithm of K-Mean clustering

D. FCM clustering

In principle, Fuzzy C-Mean (FCM) is similar to K-Mean clustering. FCM clustering is classified as soft clustering because of the cluster membership of each data expressed by the degree of membership in the range \([0 ... 1]\) [46]. FCM clustering output is the most optimized partition of the centroid.

If \( X = \{ x_1, x_2, ..., x_n \} \) is the dataset, \( V = \{ v_1, v_2, ..., v_s \} \) is the set of centroid, and \( U = \{ u_{i1}, u_{i2}, ..., u_{is} \} \) is the degree of dataset membership in each centroid then the centroid partition optimization is obtained by minimizing the objective function expressed by:
A Performance of the Scattered Averaging Technique based on the Dataset for the Cluster Center Initialization

\[ J(U, V) = \sum_{j=1}^{M} \sum_{i=1}^{K} (u_{ij})^q \cdot (d_{ij})^2 \]  

(5)

where \( u_{ij} \) is the degree of membership of \( x_j \) in \( i^{th} \) cluster, and \( d_{ij} \) is the Euclidean distance of between \( x_j \) and \( v_i \), expressed by:

\[ (d_{ij})^2 = \|x_j - v_i\|^2 \]  

(6)

The degree of membership of each data in each cluster is expressed by:

\[ u_{ij} = \left( \frac{1}{\sum_{k=1}^{K} \left( \frac{1}{d_{ik}} \right)^{2/(q-1)}} \right)^{1/(q-1)} \]  

(7)

where \( q \) is the fuzzification parameter (fuzzyfier, \( q > 1 \)). The generation of new centroids using the COA (Center of Area) formula expressed by:

\[ V_j = \frac{\sum_{j=1}^{M} (u_{ij})^q \cdot x_j}{\sum_{j=1}^{M} (u_{ij})^q} \]  

(8)

The iteration process terminated if it satisfies:

\[ |V - \bar{V}| < \varepsilon \]  

(9)

where \( \varepsilon \) is the termination criteria within the range \([0 \ldots 1]\). The FCM clustering algorithm is shown in Fig. 4.

E. SOM clustering

SOM is one type of ANN architecture classified as unsupervised learning. SOM has a more complex geometric grouping capability because of the non-linear nature of its neuronal functions. The goal of SOM is to map the input data pattern to the N-dimensional grid of the neuron. The SOM architecture consists of an input layer, an output layer, and an intra-layer unit connecting the input and output layer. Each neuron of the input layer directly connected to each neuron of the output layer. The input layer has a neuron of \( M \) vector training. The output layer has a neuron of \( K \) clusters. The SOM architecture is shown in Fig. 5.
In principle, the intra-layer is a layer that represents the neighboring distance function between the input layer weights with the output layer weights that map data from the input layer into the cluster membership in the output layer [45]. The neighboring distance function is expressed by:

\[ D = ndist(P, IW) = \|IW - P\| \]  

(10)

\[ ndist = D_{P, IW} \]

(10)

Fig. 5. The SOM architecture

\[ P \] is the dataset, \( IW \) is a weighted intra-layer matrix as the centroid of each cluster, and \( D \) is the distance matrix. Winning neurons are the \( Iw_i \) output neuron weights that have the shortest distance \( P \), expressed by:

\[ C(P) = \arg \min \|Iw_i - P\|, i = 1...K \]  

(11)

If there is still a change in cluster membership, then the intra-layer weighted is updated using the following formula:

\[ Iw_i(t + 1) = Iw_i(t) - \eta \{ P_i - Iw_i(t) \} \]  

(12)

where \( \eta \) is learning rate within the range \([0.1 \ldots 0.9]\). In this case, the SOM clustering algorithm can achieve its best GOS index by setting the appropriate learning rate value. The SOM clustering algorithm is shown in Fig. 6.

Fig. 6. The algorithm of SOM clustering
F. Dataset

This study uses annual rainfall data from the weather station of Samarinda city, East Kalimantan, Indonesia in period 2006-2016 (BPS-Statistic of Samarinda Municipality, https://samarindakota.bps.go.id). The dataset is shown in Table 1. The data to be clustered is the annual rainfall pattern in each month during the period. The data attributes used are min, max, and standard deviation expressed by:

\[ x(i)_{\text{min}} = \min (X(i, 2006 : 2016)) \]
\[ x(i)_{\text{max}} = \max (X(i, 2006 : 2016)) \]
\[ x(i)_{\text{stdev}} = \sqrt{\frac{\sum (x(i) - \overline{x(i)})^2}{(N - 1)}} \]

where \( x(i) \) is the annual rainfall in the \( i^{th} \) month, \( \overline{x(i)} \) is the average of the annual rainfall in the \( i^{th} \) month, and \( N \) is the number of year in period 2006-2016. Dataset by its attributes shown in Table 2.

Table 1. The Dataset of Annual Rainfall from The Weather Station of Samarinda City East Kalimantan Indonesia in the Period 2006-2016

<table>
<thead>
<tr>
<th>Year</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td>228</td>
<td>207</td>
<td>215</td>
<td>207</td>
<td>307</td>
<td>185</td>
<td>24</td>
<td>98</td>
<td>108</td>
<td>70</td>
<td>191</td>
<td>110</td>
</tr>
<tr>
<td>2007</td>
<td>307</td>
<td>220</td>
<td>260</td>
<td>340</td>
<td>112</td>
<td>213</td>
<td>279</td>
<td>133</td>
<td>183</td>
<td>181</td>
<td>85</td>
<td>141</td>
</tr>
<tr>
<td>2008</td>
<td>143</td>
<td>194</td>
<td>211</td>
<td>259</td>
<td>51</td>
<td>205</td>
<td>333</td>
<td>149</td>
<td>153</td>
<td>208</td>
<td>501</td>
<td>350</td>
</tr>
<tr>
<td>2009</td>
<td>164</td>
<td>196</td>
<td>279</td>
<td>309</td>
<td>186</td>
<td>41</td>
<td>157</td>
<td>123</td>
<td>99</td>
<td>232</td>
<td>165</td>
<td>211</td>
</tr>
<tr>
<td>2010</td>
<td>148</td>
<td>162</td>
<td>157</td>
<td>164</td>
<td>223</td>
<td>320</td>
<td>259</td>
<td>144</td>
<td>202</td>
<td>235</td>
<td>207</td>
<td>224</td>
</tr>
<tr>
<td>2011</td>
<td>332</td>
<td>320</td>
<td>368</td>
<td>332</td>
<td>389</td>
<td>95</td>
<td>238</td>
<td>124</td>
<td>132</td>
<td>218</td>
<td>197</td>
<td>244</td>
</tr>
<tr>
<td>2012</td>
<td>330</td>
<td>206</td>
<td>257</td>
<td>371</td>
<td>128</td>
<td>172</td>
<td>147</td>
<td>140</td>
<td>110</td>
<td>116</td>
<td>228</td>
<td>220</td>
</tr>
<tr>
<td>2013</td>
<td>176</td>
<td>309</td>
<td>284</td>
<td>337</td>
<td>234</td>
<td>161</td>
<td>145</td>
<td>90</td>
<td>256</td>
<td>223</td>
<td>363</td>
<td>276</td>
</tr>
<tr>
<td>2014</td>
<td>219</td>
<td>197</td>
<td>216</td>
<td>126</td>
<td>210</td>
<td>190</td>
<td>121</td>
<td>50</td>
<td>82</td>
<td>111</td>
<td>300</td>
<td>449</td>
</tr>
<tr>
<td>2015</td>
<td>345</td>
<td>193</td>
<td>198</td>
<td>344</td>
<td>214</td>
<td>259</td>
<td>163</td>
<td>58</td>
<td>0</td>
<td>73</td>
<td>61</td>
<td>191</td>
</tr>
<tr>
<td>2016</td>
<td>159</td>
<td>99</td>
<td>318</td>
<td>369</td>
<td>225</td>
<td>202</td>
<td>163</td>
<td>99</td>
<td>226</td>
<td>175</td>
<td>292</td>
<td>357</td>
</tr>
</tbody>
</table>

Table 2. The Dataset by Its Attributes

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>143</td>
<td>99</td>
<td>157</td>
<td>126</td>
<td>51</td>
<td>41</td>
<td>24</td>
<td>58</td>
<td>0</td>
<td>70</td>
<td>61</td>
<td>110</td>
</tr>
<tr>
<td>Max</td>
<td>345</td>
<td>320</td>
<td>368</td>
<td>371</td>
<td>389</td>
<td>320</td>
<td>333</td>
<td>149</td>
<td>256</td>
<td>235</td>
<td>501</td>
<td>449</td>
</tr>
<tr>
<td>Stdev</td>
<td>82</td>
<td>61</td>
<td>62</td>
<td>85</td>
<td>92</td>
<td>74</td>
<td>93</td>
<td>29</td>
<td>73</td>
<td>64</td>
<td>125</td>
<td>100</td>
</tr>
</tbody>
</table>

3. Result and Discussions

In this study, the dataset grouped into 3 and 4 clusters using K-Means, FCM, and SOM clustering. By ignoring the number of iterative processes, the visualization of the optimum global solution of each algorithm is shown in Fig. 7, Fig. 8, and Fig. 9.

Table 3. The Results of the evaluation performance of All three clustering algorithms with cluster center initialization using the Scattered Averaging Technique

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Parameters</th>
<th>Three clusters</th>
<th>Four clusters</th>
<th>Five clusters</th>
<th>Six clusters</th>
<th>Average value</th>
<th>The performance ratio by total GOS Index (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-Mean</td>
<td>Var (centroid)</td>
<td>660.8964</td>
<td>709.9072</td>
<td>968.0315</td>
<td>1191.7707</td>
<td>890.3515</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Var (member)</td>
<td>1475.208</td>
<td>1433.436</td>
<td>644.617</td>
<td>135.959</td>
<td>276.9805</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GOS index</td>
<td>3.5244</td>
<td>4.9525</td>
<td>1.5571</td>
<td>8.7891</td>
<td>3.2145</td>
<td>69.05</td>
</tr>
<tr>
<td>FCM</td>
<td>Var (centroid)</td>
<td>110.8552</td>
<td>211.1042</td>
<td>453.4930</td>
<td>429.1057</td>
<td>301.1395</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Var (member)</td>
<td>147.9727</td>
<td>997.9576</td>
<td>1056.1882</td>
<td>570.3514</td>
<td>1026.1175</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GOS index</td>
<td>0.0749</td>
<td>0.2115</td>
<td>0.4294</td>
<td>0.7524</td>
<td>0.2935</td>
<td>6.30</td>
</tr>
<tr>
<td>SOM</td>
<td>Var (centroid)</td>
<td>1334.9529</td>
<td>1055.7393</td>
<td>755.8469</td>
<td>956.5152</td>
<td>1025.7636</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Var (member)</td>
<td>2059.3491</td>
<td>669.6715</td>
<td>426.1999</td>
<td>419.5858</td>
<td>893.7015</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GOS index</td>
<td>0.6482</td>
<td>1.5765</td>
<td>1.7735</td>
<td>2.2797</td>
<td>1.1478</td>
<td>24.65</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.6557</td>
</tr>
</tbody>
</table>

Especially for SOM clustering, the global optimum solution is achieved by \( \eta=0.8 \). The performance evaluation results from the three clustering algorithms with cluster center initialization using the scattered averaging technique.
shown in Table 3. The dataset in this table has been grouped into 3, 4, 5, and 6 clusters. It was intended to review the results of clustering based on the number of odd and even clusters. Table 3 has shown that cluster center initialization using the scattered averaging technique based on the dataset used has improved K-Means clustering ability to achieve the global optimum solution with a performance ratio of 69.05% of the total performance of the three algorithms. This result is evidenced by the average GOS Index of the K-Means clustering (3.2145) is greater than the others (FCM = 0.2935, SOM = 1.1478). The next best clustering algorithm is SOM clustering (24.65%) followed by FCM clustering (6.30%). In addition, Table 3 also shown that the three clustering algorithms achieve their best global optimum solution at the number of even clusters.

![Fig. 7. The global optimum solution by K-Means clustering algorithm](image1)

![Fig. 8. The global optimum solution by FCM clustering algorithm](image2)

![Fig. 9. The global optimum solution by SOM clustering algorithm](image3)
4. Conclusions

This study has implemented the scattered averaging technique for cluster center initialization based on its dataset applied to K-Mean, FCM, and SOM clustering algorithms. Their performance measured by using the GOS index. The results of this study had showed that this technique significantly improves the ability of K-Mean clustering in achieving global optimum solutions compared to SOM and FCM clustering algorithms. In addition, those clustering algorithms achieve their best global optimum solution on even-numbered clusters.

Future work is how to apply the scattered averaging technique based on the data space boundary in the N-dimensional Cartesian coordinates applied to K-Mean, FCM, and SOM clustering algorithms.

Acknowledgment

The authors would like to express their heartfelt thanks to The Modern Computing Research Center, Politeknik Negeri Samarinda, for providing all their support.

References

A Performance of the Scattered Averaging Technique based on the Dataset for the Cluster Center Initialization


Authors’ Profiles

Implementation of the Naive Bayes Classifier Method for Potential Network Port Selection (MECS publisher with SCOPUS Indexing). Image transformation using fuzzy-based filtering for the texture class's measurement based on the distance of the feature and Magic cube puzzle approach for image encryption (SCOPUS indexing)

Areas of interest: Computer Vision, Computer Networks, Robotics & Artificial Intelligent

Achmad Fanany Onnilita Gaffar, Born in Surabaya, October 23, 1969. Completed undergraduate (S1) majoring in Electrical Engineering at Sepuluh Nopember Institute of Technology Surabaya in 1995. Completed postgraduate study of Electrical Engineering Department at Brawijaya University Malang in 2011. Beginning in 1998 working as a lecturer in the Department of Electrical Engineering, as a lecturer at the Department of Information Technology, State Polytechnic of Samarinda from 2012 until now. In the last 4 years, Scopus indexed articles have been produced 20 article in the form of journals and IEEE proceedings (https://www.scopus.com/author/detail.uri?authorId=57193740086)

Areas of interest: Robotics & Artificial Intelligent, Intelligent Control, Machine Learning, Computer Vision, Data Science


Mulyanto, received the bachelor degree of computer science from University of Indonesia 1999 and then received the master of computer science from Gadjah Mada University at 2016. He is a lecturer at Department of Information Technology, State Polytechnic of Samarinda. His current research interests include modelling and simulation, artificial intelligent and intelligent computing.