Integration of Fuzzy C-Means Clustering and TOPSIS (FCM-TOPSIS) with Silhouette Analysis for Multi Criteria Parameter Data

Victory Tyas P. Swindiarto
Department of Information Technology
Management
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
email: pambudi.17092@mhs.its.ac.id

Riyannarto Sarno
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
email: riyanarto@if.its.ac.id

Dian C. R. Novitasari
Department of Mathematics
UIN Sunan Ampel Surabaya
Surabaya, Indonesia
email: diancrini@uinsby.ac.id

Abstract— Rank theory is one effective method used as an evaluation, cause in the presence of these rankings there will be competition from all aspects factor parameters. TOPSIS is excellent method in ranking. It just takes a weight and make the decision matrix. Integration required by using the best membership value in a data cluster for weighting on TOPSIS. Because the process is used for multi-criteria data, the best membership value based on the results of cluster sub-criteria to get the weight of TOPSIS. A good weight based on a valid FCM structure, silhouette coefficient needed to analyze possible displacements to other clusters. This research conducted for company performance evaluation of PT XYZ based on all branch of company.

Keywords—FCM, fuzzy, clustering, TOPSIS, silhouette, evaluation, multicriteria

I. INTRODUCTION

Ranking is one of the effective methods used as an evaluation. This rating will create competition from all aspects of factor parameters. PT XYZ is one of the companies engaged in services related to ship and transportation services. In spurring the continuity of the national industry competitiveness through the cost of competitive logistics. PT XYZ always make various cost efficiency efforts to reduce logistics costs. With competitive logistics costs, it expected that the national port industry grow and compete, both regionally and internationally. Evaluation is need to improve company performance, especially financial performance, in accordance with the expectations of the stakeholders. These performance improvement efforts are implemented using sound corporate governance principles where stakeholders play an active role as a control function. Carry out analysis and evaluation of risk exposure level of all risks across all work units within the company, based on the Key Performance Indicator (KPI) assigned in each unit.

Differences in performance appraisal of PT XYZ's branches that have several criteria. To know the value of these criteria there are several sub criteria used as indicator. The number of criteria and sub criteria in the performance appraisal in each port with 17 branches of company and head office make the performance appraisal complicated so that an approach that can accommodate all the criteria and sub criteria are needed. The approach can use the Multi Criteria Decision Making (MCDM) method. Previously PT XYZ used a comparison analysis between branches of one company to another based on one factor compared to other factors. For example, compare only in terms of income (financial) only, but in terms of infrastructure, human resources and investment values are different.

Previous Fuzzy C-Means (FCM) related studies include in FCM, data distribution observations used to perform cluster validation analysis using either K-Means or Fuzzy C-Means Method [1], the FCM Method may also combined with AHP in decision-making predictions, predicting decision impact outsourcing locations abroad on the resilience of supply chain [2]. In other fields, FCM also used to incorrectly in nuclear technology [3] and FCM analysis based on stream data to analyze misclassified [4]. FCM method used as data clustering while for ranking using TOPSIS method. The integration of the Fuzzy AHP and TOPSIS methods also provides a robust decision support system for pattern derivation suitable for cultivation and cultivation which should take precedence over priority elements [5], also used in forecasting bankruptcy of a company [6]. Used in the maritime field as a support to see the ranking there is an event at sea [7], and the most used as an evaluation / risk assessment either at the company or at school and banking [8]. Besides being used as a clustering method, it is also used for forecasting such as forecasting in maritime weather [9], it shows that fuzzy is best used in all fields. In Multi Criteria Decision Making (MCDM) is also used in the optimal selection of solar farm sites by combining two methods of fuzzy and AHP [10]. Requires a cluster validity analysis created, can use many methods such as fuzzy index [11] or by using silhouette analysis [12].

Based on the problem described there is a system for evaluation of company performance based on multi criteria using integration between Fuzzy C-Means Clustering method with TOPSIS and perform Silhouette analysis to see the strength of FCM structure. Where, a valid structure of cluster is required to analyze the presence or absence of data transfer from one cluster to another.
II. PRELIMINARIES

A. Multi Criteria Decision Making

Decision-making is based on very important principles of important factors called criteria. Criteria in a large company are not only based on one or two factors, but can be more than one factor and commonly called multi criteria. Two methods are used for multi criteria, Multi Criteria Decision Making (MCDM) and Multi Attribute Decision Making (MADM) [13].

MCDM methods include Weighted Product Method (WP), Technique for Order by Similarity to Ideal Solution Method (TOPSIS), Multi-Objective Optimization on the basis of Ratio, Profile Matching Method (PM), Simple Additive Weighting Method (SAW), Simple Multi Attribute Rating Technique Method (SMART), PROMETHEE, Analytic Hierarchy Process Method (AHP), and others.

B. Fuzzy C-Means Clustering (FCM)

FCM is a method of clustering data determined by the degree of membership. It used as clustering and classification methods. Here is the fuzzy c-means algorithm.

1. Insert data of all branches of PT XYZ based on the criteria and sub-criterion, in the form of matrix of m x n, where m is the line and n is the column.

2. Determine the number of clusters (c), Power (w), Maximum Iteration, Tolerance Error (E), Function initial objective (P0 = 0) Initial iteration (t = 1).

3. Generate a random number \( \mu_{ik} \), where: 
   \( i = 1, 2, ..., m \)
   \( k = 1, 2, ..., c \)
   as elements of the initial partition matrix called U. Then calculate the number of each column:
   \[
   Q_j = \sum_{k=1}^{c} \mu_{ik} \]
   Where: \( j = 1, 2, ..., n \)
   Calculate:
   \[
   \mu_{ik} = \frac{x_{ik}}{Q_j} 
   \]
   Calculate the center of the k-cluster:
   \[
   v_{kj} = \frac{\sum_{i=1}^{m} (\mu_{ik}) w_i r_{ij}}{\sum_{i=1}^{m} (\mu_{ik})} 
   \]
   Where k = 1, 2, ..., n and j = 1, 2, ..., m.

4. Compute the objective function on the t-iteration,
   \[
   P_t = \sum_{i=1}^{m} \sum_{k=1}^{c} (|x_{ij} - v_{jk}|^w) (\mu_{ik})^w 
   \]

5. Calculate the change in the partition matrix U by using the formula 5.
   \[
   F_t = \frac{|\sum_{k=1}^{c} (\mu_{ik})^w r_{ik} - \sum_{k=1}^{c} (\mu_{ik})^{w-1} r_{ik}|}{\sum_{k=1}^{c} (\mu_{ik})^{w-1}} 
   \]
   Where: \( i = 1, 2, ..., m \) and \( k = 1, 2, ..., c \)

6. Check the stop condition:
   If \( (Pt - Pt_{-1}) < E \) or \( T > \) Maximum Iteration then stop, if not meet the condition then repeat step 4.

C. Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS)

TOPSIS uses the concept that the best alternative not only has the shortest distance from the ideal positive solution, but also the longest distance from the negative solution [14]. The TOPSIS method is one of the MCDM models used in the assessment or selection of several criteria as solution alternatives as ranking. In the TOPSIS method there is no limit on the number of attributes and alternatives used, so it can be more efficiently used to solve the problem on data that has a quantitative attribute. TOPSIS algorithm steps:

1. Create a matrix of data. Suppose the matrix (D) contains the criteria, alternatives, and weighted vectors (W) as follows.

   \[
   D = \begin{bmatrix}
   D_{11} & D_{12} & \cdots & D_{1n} \\
   D_{21} & D_{22} & \cdots & D_{2n} \\
   \vdots & \vdots & \ddots & \vdots \\
   D_{m1} & D_{m2} & \cdots & D_{mn} 
   \end{bmatrix}
   \]

   \[
   W = [w_1, w_2, \ldots, w_n]
   \]

   Where, \( C_1 = C_n = \) Criteria
   \( A_j = A_{kn} = \) Alternative
   \( W = \) Weight Vector.

2. Make a decision matrix (\( r_{ij} \)) which is normalized by using the formula in equation 6, where \( i=1,2,\ldots,m \) and \( j=1,2,\ldots,n \).
   \[
   r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{j=1}^{n} x_{ij}^2}} 
   \]

3. Make a decision matrix (\( y_{ij} \)) which is normally weighted by the equation 7 where \( i = 1,2,\ldots,m \) and \( j = 1,2,\ldots,n \).
   \[
   y_{ij} = w_i y_{ij} 
   \]

4. Determine the positive ideal solution matrix (\( A^+ \)) as in the following 8 equation where: \( j = 1,2,\ldots,n \).
   \[
   A^+ = (y_{1j}^+, y_{2j}^+, \ldots, y_{nj}^+) 
   \]

5. Determines the ideal solution matrix (\( A^- \)) as in the following equation 9 where \( j = 1,2,\ldots,n \).
   \[
   A^- = (y_{1j}^-, y_{2j}^-, \ldots, y_{nj}^-) 
   \]

6. Determine the distance between the value of each alternative with the positive ideal solution matrix (\( D^+ \)) by the formula in the following equation 10.
where \( i = 1, 2, \ldots, m \)

7. Determine the distance between the value of each alternative with the negative ideal solution matrix \( \mathbf{D}_{i}^{-} \) by the formula in the following equation 11, where \( i = 1, 2, \ldots, m \)

\[
\mathbf{D}_{i}^{-} = \sqrt{\sum_{j=1}^{n} (x_{ij}^+ - y_{ij})^2}
\]

where \( i = 1, 2, \ldots, m \)

8. Determine the preference value for each alternative \( \mathbf{P}_{i} \) by the formula in the following equation 12 where \( i = 1, 2, \ldots, m \).

\[
\mathbf{P}_{i} = \frac{\mathbf{D}_{i}^{+} - \mathbf{D}_{i}^{-}}{2}\sqrt{a_{i}^{+} + a_{i}^{-}}
\]

D. Silhouette Coefficient

The Silhouette method used to find cluster strength or identify the quality of the cluster, and it can provide visual quality results of objects in each cluster [15]. Use information according to the number of groups in the data set (cluster). For each object denoted by the group to which it belongs, with the following 13 equations.

\[
\mathbf{a}(i) = \frac{1}{|A| - 1} \sum_{j \neq i} \mathbf{d}(j, i)
\]

\( \mathbf{a}(i) \) is the average difference of object \( i \) to all other objects in A. Observe the second cluster B different from A according to the formula 14.

\[
\mathbf{d}(i, B) = \frac{1}{|B| - 1} \sum_{j \in B} \mathbf{d}(i, j)
\]

\( \mathbf{d}(i, B) \) is the average difference of object \( i \) to all other objects in B (clusters other than cluster A), cluster B is not equal to A. After calculating \( \mathbf{d}(i, B) \) for all B, and take the value smallest.

\[
\mathbf{b}(i) = \min_{B \neq A} \mathbf{d}(i, B)
\]

Cluster B that reaches the minimum \( \mathbf{d}(i, B) = \mathbf{b}(i) \) is called the neighbor of the object \( i \). The value of Silhouette, \( \mathbf{S}(i) \) in object \( i \) is defined as in equation 16.

\[
\mathbf{S}(i) = \frac{\mathbf{b}(i) - \mathbf{a}(i)}{\mathbf{b}(i) + \mathbf{a}(i)}
\]

\( \mathbf{S}(i) \): the value of the silhouette range between -1 to 1.

III. METHODOLOGY

In this study, in general the process carried out is outline in flowchart figure 1 follows.

![Flowchart of the Rank Process](image)

Figure 1: Flowchart of the Rank Process

Prior to the completion stage of the research, researchers identify on the research topics, namely analysis of performance evaluation of PT XYZ Branch and literature study to support the research. Then Data collection obtained at PT XYZ Head Office. The data obtained are data of company evaluation criteria based on Key Performance Indicator (KPI) that has been established in each unit and recapitulation and identify the data is needed in evaluating the performance of all branches of PT XYZ. The completion stage of this performance evaluation research is presented in Figure 1.
IV. RESULT AND ANALYSIS

PT XYZ Company has 17 Branches, there are more than 10 main criteria and more than 200 sub criteria used as Key Performance Indicator (KPI) as evaluation material of company per semester every year. In this study using five (5) main criteria with 12 sub criteria that exist . The first criterion called traffic, this traffic measures the movement of services both in and out of the port. In traffic criteria, there are five (5) sub criteria such as ship traffic, container traffic, cargo traffic, passenger traffic and animal traffic. The second criterion based on exploitation, the criteria of exploitation based on two sub criteria, business income and business operating. On the third main criteria is investment, investment is an activity company conducted to gain profit, on investment criteria based on two sub criteria that is budget and budget realization. The fourth main criterion is Cash flow, as well as investment criteria relating to the company's financial profit. The last criterion is Human Resources (HR), with two sub criteria that is budget plan and Budget Plan Realization. The entire data can seen in table 1, as Management Report PT XYZ in 2017.

To perform performance evaluation done by ranking using TOPSIS, in TOPSIS there is weighting to make decision matrix. The weight value in TOPSIS obtained from the highest value in each cluster based on the sub-parameters. Thus, the steps taken in clustering for each parameter (sub criteria) into several clusters first. To get a good cluster it needs analysis to valid or not. The method of evaluation or clustering validation used in this system is the silhouette coefficient method, to test the quality of the resulting cluster. This method is a cluster validation method that combines cohesion and separation methods. To calculate silhouette coefficient value required distance between data by using Euclidean distance formula. The first test is to see the results of silhouette on some cluster results on the parameters.

<table>
<thead>
<tr>
<th>No</th>
<th>Branch</th>
<th>Traffic</th>
<th>Exploitation</th>
<th>Investment</th>
<th>Cash Flow</th>
<th>HR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>0</td>
<td>196159</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>B</td>
<td>6053</td>
<td>141421</td>
<td>845142</td>
<td>209263</td>
<td>9545</td>
</tr>
<tr>
<td>3</td>
<td>C</td>
<td>8286</td>
<td>202676</td>
<td>28576</td>
<td>37250</td>
<td>2435</td>
</tr>
<tr>
<td>4</td>
<td>D</td>
<td>458</td>
<td>5459</td>
<td>261013</td>
<td>38235</td>
<td>3433</td>
</tr>
<tr>
<td>5</td>
<td>E</td>
<td>2392</td>
<td>0</td>
<td>5258829</td>
<td>96839</td>
<td>905</td>
</tr>
<tr>
<td>6</td>
<td>F</td>
<td>221</td>
<td>0</td>
<td>316377</td>
<td>4048</td>
<td>224</td>
</tr>
<tr>
<td>7</td>
<td>G</td>
<td>2194</td>
<td>6816</td>
<td>6835177</td>
<td>26726</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>H</td>
<td>551</td>
<td>20200</td>
<td>546635</td>
<td>76117</td>
<td>5316</td>
</tr>
<tr>
<td>9</td>
<td>I</td>
<td>1563</td>
<td>50557</td>
<td>172803</td>
<td>89891</td>
<td>24418</td>
</tr>
<tr>
<td>10</td>
<td>J</td>
<td>297</td>
<td>8795</td>
<td>17655</td>
<td>50423</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>K</td>
<td>473</td>
<td>11847</td>
<td>496578</td>
<td>20366</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>L</td>
<td>812</td>
<td>29010</td>
<td>473522</td>
<td>58491</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>M</td>
<td>779</td>
<td>0</td>
<td>11300957</td>
<td>3999</td>
<td>8874913000</td>
</tr>
<tr>
<td>14</td>
<td>N</td>
<td>1750</td>
<td>0</td>
<td>2230268</td>
<td>121008</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>O</td>
<td>596</td>
<td>0</td>
<td>1097294</td>
<td>2781</td>
<td>293</td>
</tr>
<tr>
<td>16</td>
<td>P</td>
<td>1793</td>
<td>4550</td>
<td>557094</td>
<td>407795</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>Q</td>
<td>584</td>
<td>0</td>
<td>14664</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Clustering test is done on the 7th parameter that is based on operating business sub criteria, tested by dividing into several clusters. The results of the experimental silhouette are determined by C (Cluster), for C = 6 in Fig. 3, C = 5 in Fig. 4, C = 4 in Fig. 5 and C = 3 is shown in Fig. 6. The results of the calculated coefficient silhouette values may vary from -1 to 1. As in Figure 3, for center cluster 6, it can be described all the values of silhouette is a positive value, there are even 4 data that is worth 1 which means the object is already in the right cluster. There is very small value in cluster 1 that is 0.0455, according to the determination Kaufman and Rousseeuw in Table 2 states that if the value is less than 0.25 it concluded no structures, which means that the data should not be on the cluster. So, the FCM structure for C = 6 is still less strong in structure. Then, more experiments on the number of other clusters should done.
Table 2. Kaufman dan Rousseeuw Silhouette Value Table

<table>
<thead>
<tr>
<th>Nilai Silhouette Coefficient</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7 &lt; SC &lt;= 1</td>
<td>Strong Structure</td>
</tr>
<tr>
<td>0.5 &lt; SC &lt;= 0.7</td>
<td>Medium Structure</td>
</tr>
<tr>
<td>0.25 &lt; SC &lt;= 0.5</td>
<td>Weak Structure</td>
</tr>
<tr>
<td>SC &lt;= 0.25</td>
<td>No structure</td>
</tr>
</tbody>
</table>

The results of Silhouette in Figures 5 and 4 with experiments $C = 5$ and $C = 4$, suggest that the clustered data also has a much negative value of 20% of the data. From the figure, it seen that in cluster one (1) also has a relatively small silhouette value of less than 3. Which means that the data should not be included in the structure of the cluster, the silhouette value contained in other clusters in the parameter is also an average from 4 to 6, if seen in table 2, enter to the medium structure. Even in Figure 4 other than cluster 1, cluster 5 also contains a negative silhouette value. So, clusters with $C = 4$ and $C = 5$, still do not have a strong structure, even the structure is very weak.
Strong structure, seen in Fig. 6, is when FCM with C = 3, 94% of the data has a silhouette value between 0.7 and 1. According to the Kaufman and Rousseeuw, Silhouette values in Table 2, FCM cluster results have a strong structure, if it classify as a strong structure, it will be difficult possibly transfer data to other clusters. Overall, the silhouette value for each experiment performed in Table 3.

For the system, conducted the test on each parameter. Then the selected membership value on the FCM used as a weight on TOPSIS process, the results obtained as shown in Figure 7 as follows.

![Table 7. Rank-Result FCM-TOPSIS all branch of PT XYZ](image)

From the calculation of TOPSIS with the weighting of FCM can be analyzed the performance of the company, the ranking of the branch based on the calculation is B, A, C, I, P, M, N, G, H, O, L, K, D, J, Q, F, and the last is F.

V. CONCLUSION

This research is an integration between clustering with ranking method, FCM and TOPSIS. Weighting by using the highest membership value on the clustering results of each criteria parameter. To analyze whether or not the clustering structure in FCM is using silhouette analysis. The result of silhouette analysis of each parameter criteria, obtained strong structure if cluster into C = 3 or C = 2. If the experiment in big C clusters, frequent overlapping or negative S (i) values. This causes there is still the possibility of moving data clustered to another cluster.

ACKNOWLEDGEMENT

Authors give a deep thank to Department of Information Technology Management Institut Teknologi Sepuluh Nopember, Department of Informatics Institut Teknologi Sepuluh Nopember and Department of Mathematics, UIN Sunan Ampel Surabaya for supporting this research.

REFERENCE


