Initializing the Fuzzy C-Means Cluster Center With Particle Swarm Optimization for Sentiment Clustering

Rimbun Siringoringo1,2, Jamaluddin Jamaluddin1
1 Department of Manajemen Informatika, Universitas Methodist Indonesia, Medan, Indonesia
2 rimbun.ringo@gmail.com

Abstract. Fuzzy C-Means (FCM) is one of the best-known clustering algorithms, however, FCM is significantly sensitive to the initial cluster center values and easily trapped in a local optimum. To overcome this problem, this study proposes and improved FCM with Particle Swarm Optimization (PSO) algorithm for high dimensional and unstructured sentiment clustering. PSO is applied for the determination of better cluster center initials. The results showed that FCM-PSO can provide better performance compared to the conventional FCM in terms of Rand Index, F-measure and Objective Function Values (OFV). The better OFV value indicates that FCM-PSO requires faster convergence time and better noise handling.

1. Introduction
Clustering aims to classify data, where the distribution of the data is unknown. The purpose of the clustering is to find the structure that is hidden in the data. Clustering can be either hard or fuzzy clustering. In the first category, a data point belongs to only one cluster exclusively and hence, it is called hard or crisp clustering [1]. In the fuzzy clustering, a data point may belong to more than one cluster depending on the distance criteria and hence, it is called soft clustering [2].

Fuzzy C-Means (FCM) is one of the fuzzy clustering. In the fuzzy clustering, the data elements can belong to more than one cluster, and each element is associated with a set of membership levels. In many situations, FCM gives better performance than hard clustering [3]. However, FCM have a limitation in case of sensitivity to the initial cluster center [4]. This sensitivity can produce a very complicated process iteration and trapped in local optimum conditions [5]. The most popular approach to handling the initial cluster center is a random way of generating random numbers [6]. The approach has two major drawbacks. First, it takes a long time in the solution search process. Second, given the same input, the algorithm not always return the same output from one test to the next [7].

Previous studies has established that optimization techniques such as Genetic Algorithms (GAs) [8], and Ant Colony Optimization (ACO) [9], have succeeded in overcoming FCM's weaknesses. These studies also showed that GAs and ACO could obtain more efficient algorithm and better clustering performance. PSO is simpler method than the Genetic Algorithm and is easy to implement because PSO does not have many procedures such as selection, mutation or crossover [10]. PSO has been proven to successfully improve the performance of machine learning methods. PSO successfully combined with Support Vector Machine (SVM) on bone age testing using an age dataset. The results achieved are a more effective SVM-PSO model [11]. PSO has successfully designed the number of layers, the number of neurons and the number of biases in Artificial Neural Networks [12], the combination of PSO with...
SOM (Self Organizing Map) has produced a new method, namely SOSwarm [13]. The combination of PSO with these machine learning methods results in significant performance improvements.

This study focused on the application of Particle Swarm Optimization (PSO) to find the best solution to the problem of the sensitivity of the initial cluster center. In this study, PSO (Particle Swarm Optimization) is applied to the initialization of the FCM cluster center. This study builds an FCM-PSO clustering model to evaluate and analyze customer sentiment towards online shopping products. Optimization of cluster centers by PSO is expected to improve FCM performance on clustering tasks.

2. Fuzzy C-Means Clustering Algorithm
FCM is one of the clustering algorithms that adapts the concept of fuzzy sets. FCM is the most widely used fuzzy clustering algorithm. FCM is a data grouping technique where the existence or location of a data in a cluster depends on the degree of membership of the data. The value of membership degree is in the range of 0 and 1. The higher the value of the membership level, the greater the similarity between the data with the existing group [5]. FCM allows each data point to be owned by more than one group or cluster depending on the membership value of the data. If it is assumed that we want to group as many as n data objects into c groups or clusters, then the objective function can be applied to FCM using equation (1) [4]

\[ (U,V) = \sum_{i=1}^{c} \sum_{k=1}^{n} \mu_{ik}^m d_{ik}^2 \]  

\( U = [\mu_{ik}] \) is a membership degree matrix with dimensions \( c \times n \), \( \mu_{ik} \) is the degree of membership between a data \( k \) to group \( i \). Membership degree values are in the range 0 and 1. The higher the value of \( \mu_{ik} \), the greater the ownership of data \( k \) against the group \( i \). \( d_{ik} = \|x_k - v_i\| \) is the euclidean distance between \( k \) and the center of the cluster \( i \), \( m \) is a fuzzy index which has a value in the range \( 1, \infty \).

The FCM algorithm can be applied with the following steps [5].

**Step 1.** Initialize the parameters \( c, m \) and \( T \) (maximum number of iterations). Initialize \( \mu_{ik} \) randomly \( (i=1, ..., c \text{ and } k=1, ..., n) \) for object \( k \) to group \( i \).

**Step 2.** Determine the center of the cluster using the equation (2):

\[ v_i = \frac{\sum_{k=1}^{n} \mu_{ik}^m x_k}{\sum_{k=1}^{n} \mu_{ik}^m} \]  

**Step 3.** Determine the change in the membership function matrix using the equation (3):

\[ u_{ik} = \frac{1}{\sum_{l=1}^{c} \frac{d_{il}}{d_{ik}^m}} \]  

**Step 4.** Calculate \( J(U, V) \) using equation (2) and note the stop criteria.

**Step 5.** Check the stop condition

- If \( |J_t - J_{t+1}| < \xi \) or \( t > T \) stops
- Else \( t = t + 1; \) return to step 2

3. Particle Swarm Optimization
Particle Swarm Optimization (PSO) is a population-based stochastic optimization technique. PSO can be used to search for optimal solutions in a broad search space. In the process of finding the optimal location, the particles move at a certain velocity and always change position until they finally find the optimal location. A particle moves in the n-dimensional search space, the particle's vector position is \( X_i = (x_{i1}, ..., x_{in}) \) dan velocity vector \( V_i = (v_{i1}, ..., v_{in}) \). To find the optimum location, particles continuously renew position and velocity of use. Equation (4) is used to calculate the new velocity of each particle and equation (5) is used to update the position of each particle [14].

\[ V_i(t+1) = w V_i(t) + c_1 r_1 (P_i - X_i) + c_2 r_2 (P_g - X_i) \]  

\[ X_i(t+1) = X_i(t) + V_i(t+1) \]  

\( P_i \) is the best position for each particle and \( P_g \) is the best position for the swarm, \( r_1 \) and \( r_2 \) are random numbers with intervals \([0, 1]\), \( c_1 \) and \( c_2 \) are learning factors which represent a cognitive component and
a social component respectively. The parameter \( t \) is an iteration index, \( \omega \) is the inertia weight parameter used to balance global search capabilities and local search [15]. Usually, inertia weight is less than one.

4. Performance Measurement

The Rand Index (RI) is the most popular criteria for measuring clustering performance. The RI states the exact percentage of decisions produced by the clustering algorithm and how similar the clustering results to the actual data. F-measure is a measure that combines precision and recall values. Recall and precision values can sometimes have different weights or values. The size that shows the reciprocity between recall and precision is F-measure which is the harmonic mean weight on recall and precision.

\[
\text{Recall (R)} = \frac{TP}{TP+FN} \tag{6}
\]
\[
\text{Precision (P)} = \frac{TP}{TP+FP} \tag{7}
\]
\[
\text{Rand Index (RI)} = \frac{TP+TN}{TP+TN+FP+FN} \tag{8}
\]
\[
\text{F-Measure} = \frac{2PR}{P+R} \tag{9}
\]

True positive (TP) is the decision to place two similar data to the same cluster, true negative (TN) is the decision to place two data that are not similar to different clusters. There are two types of errors that can occur in clustering. False positives (FP) decisions place two data that are not similar to the same cluster. False negative (FN) decisions place two similar data into different clusters.

5. Fuzzy C Means algorithm based on PSO

**Step 1**: Initialize the PSO parameter consisting of the number of swarms, number of particles, constants \( C1, C2, W_{up} \) and \( W_{low} \). Initialize FCM parameters which consist of Maximum Iteration (MaxIter), Smallest Error (\( \epsilon \)), Number of clusters (c), Rank (w), objective function initial (\( p_o \)), initial iteration (t)

**Step 2**: Initialize the cluster center. In implementing FCM, the cluster center must be initialized first. The cluster center initialization is shown in Figure 1

![Figure 1. Schematic initialization scheme cluster](image)

**Step 3**: Generate initial particles randomly. The initial cluster center is generated in parallel as many as n particles. The particle generation scheme is presented in Figure 2

![Figure 2. Particle coding scheme](image)

**Step 4**: Each particle is evaluated based on a fitness value. The fitness used to calculate fitness or the level of goodness of an individual to survive. This function takes the parameters of an individual and produces the fitness value output of that individual. For each problem that will be resolved with the PSO algorithm must be defined as a fitness function. In this study, the fitness function is using equation (10)
\[ F(P) = \frac{1}{J_{FCM}} \quad (10) \]

\[ J_{FCM} = \sum_{i=1}^{c} \sum_{k=1}^{n} (\mu_{ik})^p (d_{ik})^2 \quad (11) \]

**Step 5**: Update particle position and velocity. The PSO procedure requires that the position and velocity of each particle be evaluated according to equations (4) and (5).

**Step 6**: Test the FCM model to find the best particles. The best particles evaluated based on the fitness value in step 4 are applied to train the re-training data and calculate the fitness value.

**Step 7**: Update personal optimal fitness (pfit) and personal optimal position (pbest). So far all the particles in the initial (first) iteration have been evaluated, the first iteration produces the best particle (pbest).

**Step 8**: At this stage, it is evaluated whether all the iterations have been completed. If it has not been resumed in step 3 if it is finished then proceed in step 10.

**Step 9**: Each iteration has the best particles. At this stage, the best particles of all iterations will be evaluated to determine the best global optimal position (gbest) particles.

**Step 10**: If the gbest value in step 9 is the expected value, then the optimal cluster center initial is obtained. If the gbest value does not match the expected criteria then the new population will be re-generated by returning to step 3.

6. **Experiment testing and comparative analysis**

We collected customer product review data from Indonesian online shopping website. The collected data consisting of 627 positive reviews and 445 negative reviews as follows.

"kiriman cepat sampai, ….BARANG SESUAI.!!!"

**a. Pre-processing**

The collected customer product review data, because of the random and casual nature of reviewing, need to be pre-processed or filtered to remove unnecessary information. The pre-processing steps applied are Case Folding, Non-Alpha Numeric Removal, Stop Word Removal, and Stemming. The stop words list for the Indonesian Language consists of 760 words [16]. The Stemming algorithm that is applied is the stemming algorithm specifically for the Indonesian language, the Nazief-Andriani algorithm [17]

**b. String to word vector processing**

String to word vector processing based on TF-IDF algorithm. The results of the implementation of TF-IDF produce a data matrices with a dimension of 86 attributes x 1073 data. Because of the high dimensions of the data, need to be evaluated and filtered. We applied CFs algorithm. By applying CFs algorithm, produces a data matrices with 18 attributes.

Each particle in PSO represents the FCM cluster center. Particles are encoded with real numbers with a range of 0 to 5. For the number of clusters 2, each particle consists of 2 dimension matrices as follows. The following figure shows the initial particle coding form with a fitness value of 1.125.

| 1.84, 4.67, 0.41, 1.31, 1.07, 4.80, 2.91, 4.80, 3.78, 0.97, 4.47, 4.66, 0.08, 2.79, 3.86, 0.67, 3.68, 3.40 |
| 4.22, 1.97, 2.04, 1.21, 4.85, 4.83, 1.01, 0.61, 4.25, 0.33, 3.12, 4.52, 2.58, 1.87, 3.54, 2.58, 3.04, 4.72 |

**Figure 3. Initial particle**

| 0.01, 0.00, 0.00, 4.93, 0.00, 4.27, 0.00, 2.31, 0.00, 2.26, 0.00, 4.78, 4.77, 0.64, 0.00, 0.10, 4.91, 3.55 |
| 0.00, 3.44, 0.00, 2.69, 0.00, 0.00, 0.00, 0.00, 0.00, 0.02, 4.15, 0.90, 0.00, 0.00, 2.62, 4.71, 4.69, 0.67, 4.61 |

**Figure 4. Global best particle (GBest)**

Each particle is evaluated through its fitness value. The best fitness value search process is displayed on the figure 5. The smallest fitness value describes the center of the cluster which produces the smallest error.
The Global best particle (Gbest) is the final solution for determining the FCM cluster center. Based on figure 10, the best particle with a fitness value of 0.0821. Clustering using FCM requires good cluster center initials. Cluster center generation method was randomly replaced with PSO to generate cluster centers. The following table shows the average index of effectiveness of the clustering method tested in this study.

Table 1. Performance of FCM vs FCM-PSO

|                  | FCM-PSO | FCM  |
|------------------|---------|------|
| TP               | 609     | 221  |
| TN               | 272     | 430  |
| FP               | 18      | 406  |
| FN               | 174     | 16   |
| Recall (R)       | 0.778   | 0.932|
| Precision (P)    | 0.971   | 0.352|
| Rand Index (RI)  | 0.821   | 0.607|
| F-Measure        | 0.864   | 0.512|
| OFV              | 13070.68| 13136.61 |

Table 1 shows that the improvement of the FCM-PSO method shows better performance compared to the previous method. FCM-PSO has a RI value that is better than FCM. This indicates that FCM cluster center generation with PSO enhances FCM capabilities. It is also an indicator that FCM-PSO can work well on high-dimensional data. The FCM-PSO model also has a better F-Measure, which shows that FCM-PSO is more effective. From the OFV side, the FCM-PSO method has a better value. A large OFV value indicates that the method takes a long time to reach convergence. The picture shows that FCM has a longer time to reach convergence; with epsilon value $10^{-6}$ and iteration number by 100, FCM-PSO convergent at 26 iterations, while FCM converges at 32 iterations. FCM-PSO has better harmonics than FCM so FCM-PSO is better at handling data noise.

7. Conclusions

FCM is an robust clustering algorithm, but this algorithm is sensitive to cluster centers. This study examines the effectiveness of the application of Particle Swarm Optimization on the optimization of clustering results by FCM. Particle Swarm Optimization is applied to determine better cluster center initials. This study is applied to high-dimensional data, namely online store product reviews. The results show that the PSO model can provide the initial FCM cluster center better. This is evident from the more optimal clustering performance after PSO implementation.
References

[1] J. Zhang and L. Shen, “An improved fuzzy c-means clustering algorithm based on shadowed sets and PSO,” Comput. Intell. Neurosci., vol. 2014, p. 22, 2014.

[2] Department of E&TC, SGDCOE (Jalgaon), Maharashtra, India, A. J. Patil, C. S. Patil, R. R. Karhe, and M. A. Aher, “Comparative Study of Different Clustering Algorithms,” Int. J. Adv. Res. Electr. Electron. Instrum. Eng., vol. 3, no. 7, pp. 10490–10497, Jul. 2014.

[3] Z. Cebeci and F. Yildiz, “Comparison of K-means and Fuzzy C-means algorithms on different cluster structures,” Agrár. Agric. Inform., vol. 6, no. 3, pp. 13–23, 2015.

[4] L. Ma, Y. Li, S. Fan, and R. Fan, “A Hybrid Method for Image Segmentation Based on Artificial Fish Swarm Algorithm and Fuzzy-Means Clustering,” Comput. Math. Methods Med., vol. 2015, 2015.

[5] A.-X. Ye and Y.-X. Jin, “A fuzzy c-means clustering algorithm based on improved quantum genetic algorithm,” simulation, vol. 9, no. 1, 2016.

[6] G. Vasanthy, “International Journal of Scientific Research in Computer Science, Engineering and Information Technology,” vol. 2, no. 2, p. 4.

[7] B. V. Kumar, G. R. Karpagam, and N. V. Rekha, “Performance analysis of deterministic centroid initialization method for partitional algorithms in image block clustering,” Indian J. Sci. Technol., vol. 8, no. 57, pp. 63–73, 2015.

[8] S. Wikaisuksakul, “A multi-objective genetic algorithm with fuzzy c-means for automatic data clustering,” Appl. Soft Comput., vol. 24, pp. 679–691, 2014.

[9] G. S. Raghtate and S. S. Salankar, “Modified fuzzy C means with optimized ant colony algorithm for image segmentation,” in Computational Intelligence and Communication Networks (CICN), 2015 International Conference on, 2015, pp. 1283–1288.

[10] T. Sun and M. Xu, “A swarm optimization genetic algorithm based on quantum-behaved particle swarm optimization,” Comput. Intell. Neurosci., vol. 2017, 2017.

[11] G. E. Güaraksin, H. Hakli, and H. Uğuz, “Support vector machines classification based on particle swarm optimization for bone age determination,” Appl. Soft Comput., vol. 24, pp. 597–602, 2014.

[12] B. A. Garro and R. A. Vázquez, “Designing artificial neural networks using particle swarm optimization algorithms,” Comput. Intell. Neurosci., vol. 2015, p. 61, 2015.

[13] M. O’Neill and A. Brabazon, “Self-organizing swarm (SOSwarm) for financial credit-risk assessment,” 2008, pp. 3087–3093.

[14] M. O. Olusanya, M. A. Arasonwan, and A. O. Adewumi, “Particle swarm optimization algorithm for optimizing assignment of blood in blood banking system,” Comput. Math. Methods Med., vol. 2015, 2015.

[15] Y. Zhang, S. Wang, and G. Ji, “A comprehensive survey on particle swarm optimization algorithm and its applications,” Math. Probl. Eng., vol. 2015, 2015.

[16] F. Z. Tala, “A study of stemming effects on information retrieval in Bahasa Indonesia,” Inst. Log. Lang. Comput. Univ. Van Amst. Neth., 2003.

[17] M. Adriani, J. Asian, B. Nazief, S. M. Tahaghoghi, and H. E. Williams, “Stemming Indonesian: A confix-stripping approach,” ACM Trans. Asian Lang. Inf. Process. TALIP, vol. 6, no. 4, pp. 1–33, 2007.