ICIC Express Letters
Part B: Applications
Volume **12**, Number **11**, November **2021**

ICIC International ©2021 ISSN 2185-2766
pp. **995–1001**

CLUSTERING INDONESIAN PATIENTS WITH PERSONALITY DISORDERS USING FUZZY C-MEANS

Alifia Listu Samatha[1], Afiahayati[1,*] and Fuad Hamsyah[2]

[1]Department of Computer Science and Electronics
Faculty of Mathematics and Natural Sciences
Universitas Gadjah Mada
FMIPA UGM Sekip Utara, Bulaksumur, Sleman, Yogyakarta 55281, Indonesia
alifia.l@mail.ugm.ac.id; *Corresponding author: afia@ugm.ac.id

[2]Faculty of Psychology
Universitas Gadjah Mada
Jl. Humaniora, Bulaksumur, Yogyakarta 55281, Indonesia
fuadhamsyah@ugm.ac.id

Received January 2021; accepted April 2021

Abstract. *Essentially, every individual has the potential to have a personality disorder but still within an appropriate limit. However, there are those who have an uncontrollable trigger which harms an individual from the disorder. It is very important to map comorbidity or the occurrence of more than one disorder in the same individual on personality disorders with its tendencies. At worst, if professionals fail to map any of the diagnosis, the patient will have to suffer from Dissociative Identity Disorder (DID), which is a serious type of dissociation that causes a loss of sense of identity. This research aims to cluster personality disorder cases into multiple clusters. In this research, the Fuzzy C-Means algorithm is going to be used to do the clustering task, and then its performance will be validated. The result shows that the C-Means clustering was best to perform clustering on the personality disorder dataset with the combination of parameter $m = 3$ for calculating centroid and $m = 2$ for calculating membership function. The validation score was 0.9461 for cluster purity and 0.2007 for cluster entropy which indicates a good clustering.*

Keywords: Personality disorder, Fuzzy C-Means, Purity, Entropy

1. **Introduction.** Personality disorder is often described as a different perspective or mindset compared to people in general, often accompanied by unhealthy behaviours. People with personality disorders tend to have difficulty adapting to social life. Borderline Personality Disorder (BPD) is associated with suicidal behaviours and self-harm [1]. Up to 10% of BPD patients will die by suicide. Essentially, every individual has the potential to have a personality disorder but still within an appropriate limit. However, some have an uncontrollable trigger that harms an individual from the disorder.

Mental health professionals use the Diagnostic and Statistical Manual of Mental Disorders-IV (DSM-IV) published by the American Psychiatric Association as a guideline to diagnose patients, as it offers a common language and standard criteria for the classification of mental disorders. This manual had five sections/axes, where each axis provided distinct diagnostic information. Those five axes are composed of clinical disorders, personality disorders, general medical disorder, psycho-social and environmental factors, and global assessment of functioning respectively.

According to a clinical psychologist of Universitas Gadjah Mada, Mr. Fuad Hamsyah, S.Psi., M.Sc., patients may have more than one diagnosis for each axis. In general, some symptoms can occur in many disorders simultaneously. The occurrence of more than one

DOI: 10.24507/icicelb.12.11.995

disorder in the same individual, either at a single point in time or across the lifespan, is often referred to as comorbidity. The comorbid disorder is any distinct additional clinical entity that has existed or may occur during the clinical course of a patient who has the index disorder under study [2]. It is very important to map comorbidities on personality disorders with their tendencies. At worst, if professionals fail to map any of the diagnosis, the patient will have to suffer from Dissociative Identity Disorder (DID). DID is a serious type of dissociation, a mental process that causes a loss of connection in the thoughts, perceptions, emotions, behaviours, or sense of identity of a person.

Computer scientists have been working closely with psychologists and healthcare professionals to address more precise diagnosis using computational methods. Fuzzy logic can make the decision making more efficient and precise [3]. Sulistiani and Muludi showed that fuzzy logic is suitable to be used to diagnose stress level, depression, and mental disorder [4]. Sumathi and Poorna conducted research to predict mental health problems among children using the Fuzzy K-Means [5]. The result shows that the Fuzzy K-Means performed well compared to entropy regularization and Fuzzy K-Medoids. de la Fuerte-Tomás et al. used K-Means clustering to cluster the severity of bipolar disorder based on several characteristics [6]. They also validate it in the span of 3 years and prove that the K-Means model remains good for longitudinal validity. Casalino et al. used a variant of Fuzzy C-Means to predict bipolar disorder and produced a high cluster quality compared to supervised methods [7].

In these studies, the methods used are highly related to the methods we use. The difference lies in the data used. We also want to investigate the performance of Fuzzy C-Means to diagnose personality disorders, as it is able to cluster one patient in more than one disorder category, producing more accurate mapping on the diagnosis. For better computational disorders mapping, the list of symptoms obtained from a mental hospital will be checked by a clinical psychologist directly. After that, the data will be clustered using the Fuzzy C-Means clustering algorithm. This Fuzzy C-Means method will later produce a list of patients with its percentage of each disorder. The Fuzzy C-Means method will later be validated using cluster purity and entropy. Also, the results of Fuzzy C-Means clustering will be verified directly by a clinical psychologist.

This paper is an extended version of the author's thesis [8]. The remainder of this paper is structured as follows: Section 2 details the dataset and the methods used; Section 3 details the experiment results; Section 4 details the discussion of the results; Section 5 provides conclusions and suggests future work.

2. Materials and Methods.

2.1. **Personality disorder dataset.** The personality disorder dataset consists of 130 cases, 68 symptoms, and 10 diagnoses from a doctor. This data was taken at Grhasia Mental Hospital, Yogyakarta, by Jones and Hartati [9]. Table 1 shows the contents of the dataset. Labels Y, S, L, N are a symbol of Yes (Y), Somewhat (S), Less (L), and No (N), which is an indication of a patient suffering from a symptom. Each case is represented by C1 through C130, while the symptoms are symbolized as S1 through S68. The overview of information regarding the symptoms can be seen briefly in Table 2. The complete contents of Table 1 and the extended description of symptoms of Table 2 can be accessed at http://ugm.id/PersonalDisorder.

2.2. **Methodology.** Figure 1 visualizes the system used in this study. After inputting patient data, the dataset will go through data processing and Fuzzy C-Means clustering. Afterwards, the resulting clusters will be validated using cluster purity and entropy. Finally, the cluster results will be verified by a clinical psychologist, Mr. Fuad Hamsyah, from the Faculty of Psychology, Universitas Gadjah Mada.

Table 1. Overview of dataset

	S1	**S2**	**...**	**S67**	**S68**	**Diagnosis**
C11	Y	S	...	N	N	Paranoid
C12	S	S	⋮	N	L	Paranoid
C13	L	Y	⋮	N	S	Paranoid

Table 2. Description of symptoms

	Clinical symptoms
S1	Excessive sensitivity to failure and rejection
S2	Tendency to keep holding grudges
⋮	⋮
S67	Anger that is strong and out of place or difficulty in controlling anger
S68	Paranoid ideas that are transient and associated with stress or severe dissociative symptoms

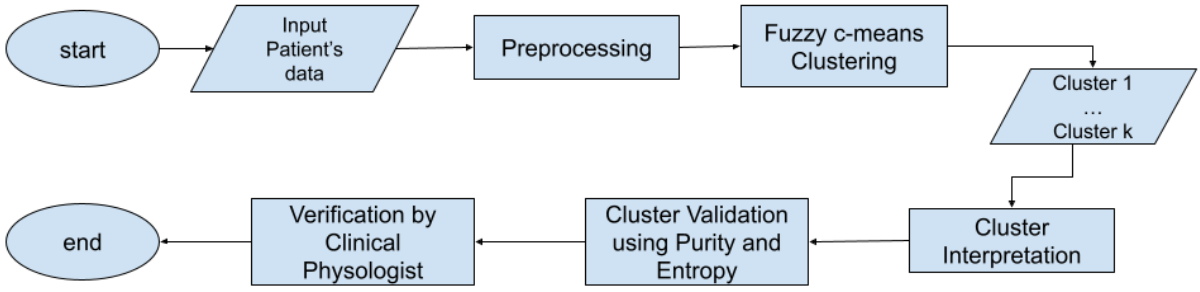

Figure 1. System flowchart

2.3. **Data preprocessing.** Data preprocessing will include missing value handling in the dataset. Moreover, ordinal variables of Y, S, L, N will be converted into arithmetic sequences between 0 and 1. Y (Yes) = 1, S (Somewhat) = 0.75, L (Less) = 0.25, and N (No) = 0. The diagnosis column is also converted to a number to facilitate the validation process.

2.4. **Fuzzy C-Means clustering.** Fuzzy C-Means is a well-known algorithm for soft clustering, which is based on fuzzy logic. This algorithm allows each data to be in more than 1 cluster by using a degree of membership. Thus, the algorithm is appropriate to be used in medical cases as it can be track complications. The Fuzzy C-Means clustering algorithm is composed of the following steps.

Step I: Initialize collection of matrices W. W contains weight $w_{i,j}$ that represents the degree of membership of data x_i in cluster C_j. $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, k$, where n is the number of data and k is the number of clusters. The degree of membership $w_{i,j}$ is used to indicate the degree to which each data point belongs to a cluster.

$$W = w_{i,j} \in [0, 1]$$

Step II: For every cluster C_j, calculate the center vectors c_j using the weight using the equation below. m is the degree of importance that determines the influence of the weight, with its value ranged in $[0, \infty]$.

$$c_j = \frac{\sum_{i=1}^{n} w_{i,j}^m \cdot x_i}{\sum_{i=1}^{n} w_{i,j}^m}$$

Step III: Update every $w_{i,j}$ by minimizing the sum of squared error between each data point and the current center vectors.

$$w_{i,j} = \frac{\frac{1}{dist(x_i,c_j)^2}^{\frac{1}{m-1}}}{\sum_{q=1}^{k} \frac{1}{dist(x_i,c_q)^2}^{\frac{1}{m-1}}}$$

Step IV: Stop if the change in centroid is below a certain threshold; otherwise, return to Step II.

3. **Experiment Results.** This research was conducted to cluster personality disorders using the Fuzzy C-Means clustering algorithm, which has the advantage of clustering overlapping data. This advantage may help psychiatrists map the personality disorders that are suffered by the patients. In this study, Fuzzy C-Means clustering successfully clustered each case of the patient into 10 clusters of personality disorders. The algorithm also produced the membership degree of personality disorder for each case in each cluster. Table 3 shows that each cluster has its own membership degree (the full version of Table 3 can be accessed at http://ugm.id/PersonalDisorder). The membership degree of each case in each cluster will later determine the level of the tendency for each disorder.

Table 3. Overview of cluster results

	Doctor's diagnosis	**Cluster**	**0**	**1**	**...**	**8**	**9**
C1	Paranoid	8	0.0413	0.0742	...	0.5421	0.0616
C2	Paranoid	8	0.0174	0.0352	...	0.7946	0.0283
C3	Paranoid	1	0.0509	0.2332	...	0.1445	0.1380

The doctor's original diagnosis that already exists within the dataset is used to determine which personality disorder is included in a cluster. For example, most of cluster 8 has the diagnosis of paranoid. Therefore, cluster 8 is identified as a paranoid disorder (Table 3). Likewise, the schizoid diagnosis mostly exists in cluster 4. Therefore, cluster 4 is identified as a schizoid disorder. This method is used to interpret each cluster in order to identify their personality disorder. Table 4 shows the cluster interpretation. The 'Total cluster' indicates the number of cases that exists within a cluster of personality disorder according to implementation results. The 'Total member' shows how many cases for each personality disorder according to the doctor's diagnosis.

Table 5 shows the patient's top three tendencies towards personality disorder. Filtering results is done with the conditions that membership function is not prominent in only one disorder. For example, if a patient has a tendency of $> 40\%$ (0.4) in one disorder, then there is no need to consider other disorder, because the tendency in other disorders must

Table 4. Cluster interpretation

Cluster	**Personality disorder**	**Total cluster**	**Total member**
0	Dissocial	15	15
1	Avoidant	10	6
2	Anankastic	9	10
3	Borderline	20	20
4	Schizoid	10	10
5	Narcissistic	15	14
6	Passive-Aggressive	15	15
7	Histrionic	15	15
8	Paranoid	12	16
9	Dependent	9	9

Table 5. Overview of cluster results filtered version

Case	Doctor's diagnosis	Tendency of personality disorder					
		H	(%)	M	(%)	L	(%)
C18	Schizoid	S	25.97	A	11.81	D	11.38
C19	Schizoid	S	26.30	A	12.41	D	11.15
C21	Schizoid	S	26.31	A	11.98	D	11.52

be below 10% (0.1). Among 130 cases, 56 cases meet the requirements. Another 74 cases have a tendency of more than 40% (0.4) in only one disorder. High (H), Medium (M), and Low (L) indicate the tendency of each case towards each personality disorder.

According to the results, many cases have an avoidant (A) and dependent (D) tendency. According to the DSM-5, it is important to distinguish avoidant personality disorder from similar personality disorders such as dependent (D), paranoid (P), schizoid (S), and schizotypal (St). However, all these disorders can also occur together. This is particularly likely for avoidant personality disorder and dependent personality disorder. Thus, if criteria for more than one personality disorder are met, all can be diagnosed.

4. Discussions.

4.1. **The parameter m on Fuzzy C-Means.** Fuzzy C-Means clustering is a well-recognized algorithm for clustering that allows us to construct a fuzzy data partition. The algorithm depends on a parameter m which corresponds to the degree of fuzziness of the solution. Parameter m should be more than 1. The initiation of the parameter $m = 1$ indicates the application of hard clustering. A typical choice for initiating parameter m for Fuzzy C-Means clustering implementation is $m = 2$. This research also began by implementing $m = 2$. Implementation with the parameter $m = 2$ alongside with $k = 10$ did not turn out well. The results of the clustering are not following the expected number of k initiations. The final membership function is also not well distributed. Many membership values fell in the range $0.999 < x < 1.000$. Trials of iteration changes have also been tried with multiple variations, but the results still fail to cluster into the number of expected k and generate a clear membership function.

After that, we increased the m parameter value for the centroid calculation into $m = 3$ while still using $m = 2$ to update the membership degrees. In the mathematical perspective, the Fuzzy C-Means clustering algorithm formula used only one m parameter value. However, research by Torra [10] shows that it is possible to use a combination of two different parameters so that all cases can belong to all clusters equally. By combining these two parameters, we successfully clustered the cases into the desired number of cluster and also produced clearer degree of membership.

4.2. **Cluster purity and entropy.** Purity is a measure of the extent to what degree a cluster represents a single class. Its calculation can be thought of as follows: for each cluster, count the number of data points from the most common class in the cluster. Afterwards, take the sum over all clusters and divide by the total number of data points. Bad clustering has purity values close to 0, while perfect clustering has a purity of 1. Entropy is a metric that measures the amount of disorder in a vector. The smallest possible value for entropy is 0, occurring when all symbols in a vector are similar. In other words, there is no disorder in the vector. The larger the value of entropy, the more disorder there is in the associated vector.

Table 6 shows the best cluster performance is achieved with the combination of parameter $m = 3$ for the centroid calculation and $m = 2$ for updating the membership degree. The purity score of 0.9461 from the results of validation shows that a combination of $m = 3$ and $m = 2$ succeeded in clustering the data. Likewise, the entropy cluster with a

TABLE 6. Performance comparison of parameter m

m for membership function	m for centroid	Cluster validation score	
$m = 2$	$m = 2$	**Purity**	0.2461
		Entropy	2.9060
	$m = 3$	**Purity**	0.9461
		Entropy	0.2007
	$m = 4$	**Purity**	0.8923
		Entropy	0.3444
	$m = 5$	**Purity**	0.8923
		Entropy	0.3444

value of 0.2007 shows that there is only a little disorder in the vector. The second-best validation score is achieved by a combination of $m = 4$ with $m = 2$ and $m = 5$ with $m = 2$, having the same results of cluster purity and entropy score. Lastly, the use of a single parameter $m = 2$ shows the worst results of the validation.

4.3. **Comparison with K-Means.** Both methods have the same clustering results, with the same performance comparison as tabulated in Table 7. What makes these two methods differ is the membership degree in the Fuzzy C-Means can cluster the dataset into more than one cluster.

TABLE 7. Performance comparison Fuzzy C-Means and K-Means

	Fuzzy C-Means	**K-Means**
Purity	0.9461	0.9461
Entropy	0.2007	0.2007

4.4. **Result verification by professional.** Verification is done after the results of clustering personality disorder are obtained. The results were verified by a clinical psychologist, Mr. Fuad Hamsyah, S.Psi., M.Sc. from Faculty of Psychology, Universitas Gadjah Mada, along with the symptoms that the patient has in each case. He also stated that the tendency for the occurrence of more than one personality disorders in the cases on the results of this study most likely indeed occurs in patients. Thus, the Fuzzy C-Means clustering method successfully clustered patients to more than one personality disorder. However, there are few cases on the clustering results that are contradictive to the knowledge of psychology, for example, the schizoid (S) case, where the results suggest that there is a dependent (D) tendency. In fact, schizoid (S) does not care about the surroundings, while dependent (D) cares about the surroundings. This happens due to several factors. The first factor is the difference in the way of diagnosis. Doctors diagnose patients qualitatively or more by theory. Meanwhile, this study is conducted quantitatively or using computation and mathematical calculations. The second factor is the human error that the doctors made in diagnosing patients. The last factor is that the diagnosis in this study only uses data in terms of one axis and not multi-axial. Indeed, the results cannot be directly used as the final result of a patient's diagnosis. However, this research can be used as the first step in research related to computational methods that can help doctors diagnose a patient with personality disorders.

5. **Conclusions.** In conclusion, the Fuzzy C-Means algorithm is capable of clustering personality disorder. The Fuzzy C-Means also performs better in clustering when using the combination of m parameter compared to using only a single parameter m. The combination between parameter $m = 3$ for calculating centroid and $m = 2$ for calculating membership function produced the best performance for this task. Out of the 130 cases,

74 cases had a high tendency to be included in just 1 disorder, with a percentage of more than 50% (0.5) and causing a very low percentage for other disorders. The rest 56 cases have the potential tendencies of having more than 1 disorder. The tendencies that most patients have are towards avoidant (A) and dependent (D) disorder, as these two disorders are related one to another. Because there are some contradictions to the knowledge of psychology, the results cannot be directly used as the final result of a patient's diagnosis. However, this research can be used as the first step in using computational methods to help doctors diagnose a patient with personality disorders. The results of clustering using C-Means clustering are the same as the results of clustering using K-Means. The difference is that C-Means clustering has membership degree in all clusters to see the potential tendency of a case in each existing cluster. The use of the degree of membership is very important in this study considering that several personality disorders can occur in one single case. For further work, we suggest using a combination of C-Means clustering with classification algorithms such as random forest as ensemble learning for better prediction performance.

REFERENCES

[1] J. Paris, Suicidality in borderline personality disorder, *Medicina*, vol.55, no.6, p.223, 2019.

[2] A. R. Feinstein, The pre-therapeutic classification of co-morbidity in chronic disease, *Journal of Chronic Diseases*, vol.23, no.7, pp.455-468, 1970.

[3] Herman and N. Surantha, Smart hydroculture control system based on IoT and fuzzy logic, *International Journal of Innovative Computing, Information and Control*, vol.16, no.1, pp.207-221, 2020.

[4] H. Sulistiani and K. Muludi, Implementation of various artificial intelligence approach for prediction and recommendation of personality disorder patient, *Journal of Physics: Conference Series*, DOI: 10.1088/1742-6596/1751/1/012040, 2021.

[5] M. R. Sumathi and B. Poorna, Fuzzy clustering based Bayesian framework to predict mental health problems among children, *ICTACT Journal on Soft Computing*, vol.7, no.3, 2017.

[6] L. de la Fuente-Tomás, P. Sierra, M. Sánchez-Autet, B. Arranz, A. García-Blanco, G. Safont and M. P. García-Portilla, A clinical staging model for bipolar disorder: Longitudinal approach, *Translational Psychiatry*, vol.10, pp.1-9, 2020.

[7] G. Casalino, M. Dominiak, F. Galetta and K. Kaczmarek-Majer, Incremental semi-supervised fuzzy c-means for bipolar disorder episode prediction, *2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, pp.1-8, 2020.

[8] A. L. Samatha, *Clustering Patients with Personality Disorders Using C-Means Clustering Algorithm*, Bachelor Thesis, Universitas Gadjah Mada, Yogyakarta, Indonesia, 2020.

[9] A. H. S. Jones and S. Hartati, *Case Base Reasoning for the Diagnosis of Personality Disorders Using Bayesian Probabilistic (Case Base Reasoning Untuk Diagnosis Gangguan Kepribadian Dengan Memanfaatkan Probabilistic Bayesian)*, Master Thesis, Universitas Gadjah Mada, Yogyakarta, Indonesia, 2016.

[10] V. Torra, On the selection of m for fuzzy C-means, *2015 Conference of the International Fuzzy Systems Association and the European Society for Fuzzy Logic and Technology (IFSA-EUSFLAT-15)*, Asturias, Spain, pp.1571-1577, 2015.