Cluster Validity Index to Determine the Optimal Number Clusters of Fuzzy Clustering for Classify Customer Buying Behavior

I Dewa Made Widia(1), Sovia Rosalin(2), Salnan Ratih Asriningtias(3)

Vocational Education Program, Universitas Brawijaya, Indonesia

E-mail: (1)dewa_vokasi@ub.ac.id, (2)soviavokasi@ub.ac.id, (3)salnanratih@gmail.com

Received: 15 March 2021; Revised: 24 April 2021; Accepted: 5 May 2021

Abstract
One of the strategies in order to compete in Batik MSMEs is to look at the characteristics of the customer. To make it easier to see the characteristics of customer buying behavior, it is necessary to classify customers based on similarity of characteristics using fuzzy clustering. The number of clusters is one of the parameters that must be determined at the beginning of the fuzzy clustering method. In Fuzzy clustering to get best performance is not with increasing the number of clusters, but choose the right number cluster have effect in performance of fuzzy clustering. So to get optimal number cluster, we can measured the result of clustering in each number cluster using the cluster validity index. From several types of cluster validity index, NPC give the best value. Optimal number cluster that obtained by the validity index is 2 and this number cluster give classify result with small variance value.

Keywords: Cluster Validity Index; Fuzzy Clustering; Number Cluster;

INTRODUCTION
Surakarta is a city that has more than 1000 MSMEs (Micro, Small and Medium Enterprises) units with various types of businesses. The type of business of the MSMEs that has the largest production capacity is the Batik MSMEs (Soebagiyi & Wahyudi, 2008). The high competitiveness, Batik MSMEs is require to be more observant about the opportunities that exist especially to get loyal customers. One strategy that can be done is to look at customer buying behavior (Oke et al., 2016). There are several factors that influence customer buying behavior, including personal factors (Ramya, 2016) and historical factors of customer transactions (Hidayatullah et al., 2018). To make it easier to see the characteristics of customer buying behavior, it is necessary to group customers based on similarity of characteristics. The method that can be used in grouping is Fuzzy Clustering, because fuzzy clustering is able to group data only by looking at the similarity of characteristics (Redjeki, 2017).

The number of clusters is one of the parameters that must be determined at the beginning of the fuzzy clustering method. Based on (Wiharto & Suryani, 2019) in Fuzzy clustering to get best performance is not with increasing the number of clusters, but choose the right number cluster have effect in performance of fuzzy clustering. To get optimal number cluster results, the number of clusters must be optimal. To get optimal number clusters, we use the number of clusters with a minimum number of 2 clusters to the maximum number of clusters that can be obtained from the equation in (Feng et al., 2013) so that we can get the number of clusters with a good error value. However, an error does not necessarily give maximum cluster results. Cluster results will be good if the clusters have similarities. This can be measured using the cluster validity index (Rizman Zalik, 2010). Previous
research used (Xing & Li, 2019) three validity indices for optimal cluster numbers such as Partition Coefficient, Partition Entropy and Xie Bani. The results show Xie Bani has the best result. In this research, Fukuyama Sugeno, Improved Partition Entropy and Improved Partition Coefficient will be used compared with Xie Bani to determine the optimal number clusters. For each cluster value will be evaluated using the cluster validity index and the best is selected.

MATERIALS AND METHODS
Data that used in this research are the results of interviews from customers MSME. The data consisted of 120 respondents including customer personal data as shown in Table 1 and transaction history data as shown in Table 2. Data is text or numeric in a range. In order to be processed using fuzzy clustering, it must be converted into numerical data and normalized according to equation (Akanbi et al., 2015).

The parameters that must be determined at the beginning of the clustering process are weighting exponent (m = 2), maximum iteration (I = 1000), smallest expected error (ε = 0.00001), initial objective function (P0 = 0) and initial iteration (t = 1). The number of clusters has a value in range, from 2 to 11 and the optimal value will be taken with the cluster validity index.

Several types of cluster validity index methods used in this research are Normalized Partition Coefficient (NPC) (Li, 2011), Normalized Partition Entropy (NPE) (Chen, 2012), Xie Bani Index (XB) (Singh et al., 2018) and Fukuyama Sugeno Index (FS) (Sengupta et al., 2011). The algorithm process of classify customer buying behavior using fuzzy clustering with optimal number cluster is as follows:

⇒ Step 1: Choose Cmin (number cluster minimum) and Cmax (number cluster maximum).
⇒ Step 2: Initialize Cmin cluster center (V)
⇒ Step 3: For cluster Cmin to Cmax
⇒ Step 4: Apply the Fuzzy Clustering algorithm to generate membership matrix (U) and the cluster center (V).
⇒ Step 5: Test for convergence, if convergence go to step 6 but if not convergence go back to step three (3).
⇒ Step 6: Compute a cluster validity index
⇒ Step 7: Choose the cluster number that has the optimal cluster validity index
⇒ Step 8: Test the cluster result with cluster variance.

RESULT AND DISCUSSION
Optimal number of clusters is found by implementing four different cluster validity index like Normalized Partition Coefficient (NPC), Normalized Partition Entropy (NPE), Xie Bani

| Parameter | Transaction Data | Conversion | Quantity |
|-----------|------------------|------------|----------|
| Price     | < 200.000        | 1          | 14       |
|           | 200.000 - 500.000| 2          | 80       |
|           | 500.000 - 1,000.000| 3 | 17       |
|           | > 1,000.000      | 4          | 9        |
| Number of | < 3              | 1          | 96       |
| Transactions | 3-5             | 2          | 16       |
|           | > 5              | 3          | 8        |
| Total Amount of Purchases | < 500.000 | 1 | 89       |
|           | 500.000 - 1,000.000| 2 | 21       |
|           | > 1,000.000      | 3          | 10       |
| Type of Day | Holiday         | 1          | 46       |
|            | Weekday          | 2          | 16       |
|            | Weekend          | 3          | 58       |
| Purchase Considerations | Taste | 1 | 33       |
|            | Trend            | 2          | 23       |
|            | Quality          | 3          | 27       |
|            | Price            | 4          | 13       |
|            | New Product      | 5          | 24       |

Table 1. Customer Transaction Data
Index (XB) and Fukuyama Sugeno Index (FS). As shown in Fig. 1, the minimum value of NPC VI is achieved at \( c = 2 \) with the value of cluster validity index at -0.933.

![Fig. 1 Values of NPC VI](image)

In Fig. 2, the minimum value of NPE VI is achieved at 0 except in \( c=3 \) and \( c=5 \). Fig. 3, the minimum value of NPC XB is achieved at \( c = 2 \) with the value of cluster validity index at 0 and in Fig. 4, the minimum value of NPC FS is achieved at \( c = 2 \) with the value of cluster validity index at 0.006.

![Fig. 2 Values of NPE VI](image)

![Fig. 3 Values of XB VI](image)

![Fig. 4 Values of FS VI](image)

For all different of validity indexes, give a result that 2 is the optimal number of clusters.

As shown in Fig. 5, from four different types of cluster validity index NPC gives the smallest value, but each of these have same optimal number cluster at 2. So, the number cluster that we use to classify customer buying behavior is 2. The results of classify using the fuzzy clustering method for 2 clusters can be seen in Table 3. A cluster is said to be good if it has a small variance within cluster (Vw) and variance (SV) and has a large variance between cluster (Vb).

![Fig. 5 Values of CVI](image)

As shown in Fig. 6 the smallest variance value for variance within cluster was achieved in 2 clusters. In Fig 8, same goes to variance cluster, smallest variance value was achieved in 2. As shown in Fig. 7 the largest variance value for variance between cluster was achieved in 2 clusters. So, this shows that 2 is the optimal number of clusters for classification.

![Fig. 6 Variance Within Cluster](image)
Optimal number cluster can determined by cluster validity. The result of optimal number cluster is 2 and can give best result for classify customer buying behavior. From several types of cluster validity index, NPC VI give the best value. That number cluster providing a small validity index value and the classification results also provide a small variance value. In classification, in order to obtain optimal results beside choose the right number of clusters, it must also be able to generate the optimal value of the cluster center. Because value of cluster center consist of random value. So, for future research, it is necessary to optimize the value of the cluster center.

REFERENCES

Akanbi, A. A., Amiri, I. S., & Elahe, F. (2015). A Machine-Learning Approach to Phishing Detection and Defense. Elsevier.

Chen, J. M. (2012). The Improved Partition Entropy Coefficient. International Conference on Multimedia and Signal Processing CMSP, 1–7.

Feng, X. B., Yao, F., Li, Z. G., & Yang, X. J. (2013). Improved Fuzzy C-means based on the optimal number of clusters. Applied Mechanics and Materials, 392, 803–807. https://doi.org/10.4028/www.scientific.net/AMM.392.803

Hidayatullah, D. P., Rokhmawati, R. I., & Perdanakusuma, A. R. (2018). Analisis Pemetaan Pelanggan Potensial Menggunakan Algoritma K-Means dan LRFM Model Untuk Mendukung Strategi Pengelolaan Pelanggan (Studi Pada Maninjau Center Kota Malang). Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer, 2(8), 2406–2415.

Li, C. S. (2011). The improved partition coefficient. International Conference on Advances in Engineering, 24, 534–538. https://doi.org/10.1016/j.proeng.2011.11.2691

Oke, A. O., Kamolshhotiros, P., Popoola, O. Y., Ajagbe, M. A., & Olujobi, O. J. (2016). Consumer behavior towards decision making and loyalty to particular brands. International Review of Management and Marketing, 6(4), 43–52.

Ramya, N. (2016). Factors affecting consumer buying behavior September 2016. International Journal of Applied Research, 2(10), 76–80.

Redjeki, S. (2017). Pemodelan Pengelompokkan Prestasi Dosen Menggunakan Metode Fuzzy C-Means. JIKO (Jurnal Informatika Dan
Rizman Zalik, K. (2010). Cluster validity index for estimation of fuzzy clusters of different sizes and densities. *Elsevier: Pattern Recognition, 43*(10), 3374–3390. https://doi.org/10.1016/j.patcog.2010.04.025

Sengupta, S., De, S., Konar, A., & Janarthanan, R. (2011). An improved fuzzy clustering method using modified Fukuyama-Sugeno cluster validity index. *2011 International Conference on Recent Trends in Information Systems, ReTIS 2011 - Proceedings*, 269–274. https://doi.org/10.1109/ReTIS.2011.6146880

Singh, M., Bhattacharjee, R., Sharma, N., & Verma, A. (2018). An improved xie-beni index for cluster validity measure. *4th International Conference on Image Information Processing, ICIIP 2017, 2018-Janua*, 95–99. https://doi.org/10.1109/ICIIP.2017.8313691

Soebagiyo, D., & Wahyudi, M. (2008). Analisis Kompetensi Produk Unggulan Daerah Pada Batik Tulis Dan Cap Solo Di Dati li Kota Surakarta. *Jurnal Ekonomi Pembangunan: Kajian Masalah Ekonomi Dan Pembangunan, 9*(2), 184. https://doi.org/10.23917/jep.v9i2.1026

Wiharto, W., & Suryani, E. (2019). The analysis effect of cluster numbers on fuzzy c-means algorithm for blood vessel segmentation of retinal fundus image. *2019 International Conference on Information and Communications Technology, ICOI-Act 2019*, 106–110. https://doi.org/10.1109/ICOI-Act46704.2019.8938583

Xing, R., & Li, C. (2019). Fuzzy c-means algorithm automatically determining optimal number of clusters. *Computers, Materials and Continua, 60*(2), 767–780. https://doi.org/10.32604/cmc.2019.04500