JOIV : Int. J. Inform. Visualization, 6(2) - June 2022 284-289

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Data Clustering for Identification of Building Conditions Using Hybrid Multivariate Multinominal Distribution Soft Set (MMDS) Method

Rohmat Saedudin [a], Iwan Tri Riyadi Yanto [b,e], Avon Budiono [a], , Sely Novita sari [c], Mustafa Mat Deris [d], Norhalina Senan [e]

[a] *Department of Information Systems, Telkom University, Bandung, West Java, Indonesia*
[b] *Department of Information Systems, Universitas Ahmad Dahlan, Yogyakarta, Indonesia*
[c] *Faculty of Civil Engineering and Planning, Institute Teknologi Nasional Yogyakarta, Indonesia*
[d] *Faculty of Applied Science and Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia*
[e] *Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia*

Corresponding author: [*]rdrohmat@telkomuniversity.ac.id

***Abstract*—Identifying building conditions for user safety is an urgent matter, especially in earthquake-prone areas. Clustering buildings according to their conditions in the categories of danger, vulnerable, normal, and safe is important information for residents and the government to take further action. This study introduces a new method, namely hybrid multivariate multinomial distribution with the softest (MMDS) in working on the process of clustering building conditions into the most appropriate category and comparable to the condition data presented in the building data set. Research using the MMDS method is very important to map the condition of existing buildings in an area supported by available data sets. The results of the measurements carried out can provide information related to the building index and were clustered based on the index value of the condition of the building. The dataset used in this study is data on school buildings in the West Java region. There are 286 school building data with four condition parameters: foundation, concrete reinforcement, easel pole, and roof. From existing data and defined condition parameters, buildings can be classified accurately and in proportion to the facts on the ground. This study also compared the proposed method, MMDS, with the baseline method, namely Fuzzy Centroid Clustering (FCC) and Fuzzy k-means Clustering (FKC). The results show that the proposed method is superior to the baseline method with a faster processing time.**

***Keywords*— Clustering; soft set; multivariate multinomial distribution.**

Manuscript received 10 Jan. 2022; revised 20 Mar. 2022; accepted 12 Apr. 2022. Date of publication 30 Jun. 2022.

I. INTRODUCTION

Building development is progressively widening, not only in big cities but also in rural areas. The environment conversion is needed to transform the area by introducing safety and energy-efficient buildings [1]. Identifying a building is important to investigate whether the building is safe or needs to be repaired or reconstructed [2]. Detailed seismic vulnerability evaluation is a technically complex procedure and can only be performed on a limited number of buildings [3]. Therefore, very important to use simpler procedures that can help to rapidly evaluate the vulnerability profile of different types of buildings so that the more complex evaluation procedures can be limited to the most critical buildings [4].

An incident like collapsed building should never have happened if there had been accurate data and information regarding the number and condition of school buildings in the area. The problem is that no accurate and precise data describes the existing conditions of all school buildings in the area. The first step that must be taken is to identify the number and condition of school buildings in the area.

The data can be processed to obtain information and knowledge from the results shown. An important process in this data processing is the grouping of data based on the kind of resemblance or similarities between data, namely the clustering process [5]. This process divides a large dataset into smaller classes called clusters. The accuracy of this clustering step could affect the information and knowledge that can be extracted from the data [6].

Clustering has been successfully applied for pattern recognition, segmentation, and statistics [7]–[11]. Previous studies by Zhou [12] and Irfan et al. [13] also show how multivariate datasets can be divided into small groups. Another technique that has been developed is the clustering technique to group large datasets, namely the k-means method [14]. The disadvantage of the k-means technique is that it cannot handle categorical data directly. This technique is suitable for numerical data. The problem is that categorical data is different from numerical data, where categorical objects do not have an inherent distance measure. As a result, categorical grouping data is more difficult than numerical grouping data [15].

Many clustering techniques have been proposed to overcome this problem of categorical grouping data, including to avoid the k-means constraint on data categorization, researchers use hard k-mode as a simple matching function [16]. Next, a new inequality measure is used to improve hard k-modes [17]–[19] and to create fuzzy k-modes [20]. Kim et al. [21] demonstrated how to increase the efficiency of the fuzzy k-mode by converting it to a fuzzy centroid. Yang et al. [22] proposed a highly effective clustering technique based on parametric data called fuzzy K-partitioning.

The problem is that all baseline methods require data to be represented in binary values so it requires a long computation time and a low level of cluster purity. Several techniques and methods have been developed to assist the data clustering process, such as Fuzzy Logic, Fuzzy Centroid Clustering, and Fuzzy K-Means Clustering (FCC). It is just that the accuracy of the results and response time are still relatively unsatisfactory.

To overcome these two problems, a method is needed that is not constrained by the calculation's length and the cluster's low purity. Furthermore, ambiguity often occurs in determining the number of clusters in a data set by using the letter k to represent the quantity as in the k-means algorithm. The number of errors in clustering generated to a point can be reduced by increasing k indefinitely, where each data point is considered as its cluster (that is, when the number of data points (n) is equal to k). Naturally, a balance between maximum data compression through a single cluster and maximum accuracy through cluster assignment to each data point is reached when the value of k is optimal.

Another method should be chosen if the appropriate value for k is unclear from prior knowledge of the data set properties [22]. So for clustering optimization problems, clustering stability in terms of the optimal number of solutions is a heuristic that is often used to determine cluster size in various clustering applications [23]–[26]. This study introduces a new Multivariate Multinominal Distribution Softset (MMDS) method to overcome the previous problem where clusters can be formed with high purity and better response time in the clustering process. Multi-valued information systems can be used to represent categorical data [27].

II. Material and Method

A. Data Building

The dataset used in this study is data on the condition of school buildings in one district in West Java. There were 2122 elementary school, junior high school, senior high school, and vocational school buildings recorded and collected. The data successfully processed and used in this study were 286 data of 303 data collected. Four condition parameters are used in the dataset: Foundation, Concrete Reinforcement, Easel Pole, and Roof [4]. The details of the part of the data set used can be seen in Table 1.

TABLE I
DATA SET OF SCHOLL BUILDING CONDITION

Building ID	Foundation	Concrete Reinforcement	Easel Pole	Roof	Condition
1	2	4	3	3	**2**
2	1	3	2	3	**1**
3	3	4	3	2	**2**
4	4	4	3	3	**4**
5	4	4	4	3	**4**
6	4	4	4	4	**4**
7	3	3	3	3	**3**
8	3	3	2	2	**2**
9	3	4	3	3	**3**
10	1	2	2	2	**1**
11	3	3	3	4	**3**
12	2	2	2	3	**2**
13	4	3	4	4	**4**
14	4	4	4	4	**4**
15	3	3	2	3	**2**
16	3	2	4	NaN	**2**
17	3	2	3	3	**2**
18	4	4	4	4	**4**
19	4	4	4	4	**4**
20	NaN	2	4	3	**2**

The data in Table 1 shows the data on the buildings' condition. Column 1 shows the identity number of the building, columns 2 to 5 are condition parameters, and column 6 is the decision or result of building conditions. Values 1 – 4 in the condition and outcome parameter columns represent 1 (Not Safe), 2 (Unsafe), 3 (Normal), 4 (Secure). The NaN value is a condition where the data value in the condition parameter does not exist/is incomplete. In the table, it can be seen that buildings with ID 16 and ID 20 have incomplete condition data. The existing dataset shows the classification of existing buildings categorized in 4 conditions, namely safe, normal, vulnerable, and dangerous. The grouping process carried out using both the baseline and proposed methods refer to the conditions of this building dataset.

B. Data collection

This study uses primary data collected directly from basic education data of the Ministry of Education and Culture of the Republic of Indonesia, and directly observed in the field. The survey and observations were carried out in Tasikmalaya, West Java. The field survey was carried out by looking directly at the existing buildings and then adapting them into a simple building assessment method. The basic form of the building includes the core parts of a building that can represent the structural strength of a building. The core variables observed were 11 parts consisting of 40 basic

building standard components. Of the eleven sections observed, four condition parameters are used as a reference for the robustness of a building to withstand the load, namely sloof (reinforced iron construction), columns, floor beams, and floor plates. The floor beams and floor slabs are connected to the foundation and hold the columns connected to the beams, trusses, and roof. A complete list of building variables is given in Table 2.

TABLE II
THE LIST OF VARIABLE

No	Variable	Parameter Condition
1	Plan Drawing	
2	Floor plan	Foundation
3	House Foundation	
4	Sloof	
5	Column	
6	Wall	Concrete Reinforcement
7	Ring Back	
8	Reinforcement Details	
9	Connection	Easel Pole
10	Mountains	Roof
11	Stance	

The data collection process was carried out for three months (April - June 2021). The number of schools that were successfully visited was 303 schools out of a total of 2122 schools in Tasikmalaya. The data used as a dataset in this study is data on the condition of 286 school buildings (17 data incomplete was not used).

C. Analysis Technique

In this study, the clustering technique is used in analyzing the data to determine which buildings are in comparable repair conditions. Several basic techniques, including Fuzzy Centroid (PC) and Fuzzy k-mean Partition (FkP), are compared with the proposed multivariate multinomial distribution technique based on multiple soft sets (MMDS) [28], [29]. It uses MMDS to determine the highest probability and multi-soft sets decomposition to break down the data into many sets with comparable values [30], [31]. It can be defined as

$$\text{Maximize } L_{CML}(z,\lambda) = \sum_{i=1}^{|U|}\sum_{k=1}^{K} z_{ik}\sum_{j=1}^{|A|}\sum_{l=1}^{|a_j|}\ln\left(\lambda_{kjl}^{u_i}\right)^{|F,a_{jl}|} \quad (1)$$

Subject to

$$\sum_{k=1}^{K} z_{ik} = 1, \text{for } i = 1,2,\dots,|U|. \quad (2)$$

$$\sum_{l=1}^{|a_j|} \lambda_{kjl} = 1. \quad (3)$$

The maximization o the objective function $L_{CML}(z,\lambda)$ can be obtained by updating the equation as follows:

$$\lambda_{kjl} = \frac{\sum_{u_i\in(F,a_{jl})} z_{ik}(u_i)}{|U|} \quad (4)$$

$$z_{ik} = \begin{cases} 1 & if \quad \sum_{j=1}^{|A|}\ln\lambda_{kjl}^{u_i} = \max\limits_{1\le k'\le K}\sum_{j=1}^{|A|}\ln\lambda_{k'jl}^{u_i} \\ 0 & otherwise \end{cases} \quad (5)$$

where $U = \{u_1, u_2, \dots, U_n\}$ is a finite set of instances, $A = \{a_1, a_2, \dots a_m\}$ is a finite set of the attribute. $(F,E) = \left((F,a_1),(F,a_2),\cdots,\left(F,a_{|A|}\right)\right)$ can be defined as a multi-soft set over universe U as in [31], where $(F,a_1),\cdots,\left(F,a_{|A|}\right) \subseteq (F,A)$ and $\left(F,a_{j_1}\right),\cdots,\left(F,a_{j_{|a_j|}}\right) \subseteq \left(F,a_j\right)$.

The results that are measured and used as a comparison index in assessing the clustering method are the processing time and the purity level of the classes formed. The condition of the building is categorized into four categories in 4 condition indexes, namely very well, well, intermediate and bad. Details of the indexation of this condition can be seen in Table 3.

TABLE III
CONDITION INDEX SCALE.

Zone	Condition Index	Condition Description	Handling Measure	Building Categorization
1	86-100	Very Well	No immediate action is required.	Secure
2	70-85	Well	Preventive Maintenance	Normal
3	40-69	Intermediate	To determine the appropriate course of action, it is necessary to conduct an alternative economic analysis of improvements.	Unsafe
4	0-39	Bad	A thorough evaluation is required to determine the necessary repair, rehabilitation, and reconstruction actions, as well as to assess the safety.	Not Safe

Further comparative analysis was carried out on the response time required by the baseline method (FC and FkP) and the proposed method (MMDS). Process duration is calculated in seconds, and whether the proposed method outperforms the baseline method improves response time. Detailed results of this response time can be seen in Table 4.

The method with the fastest response time is the best and most suitable method for processing categorical data models. The results that are measured and used as a comparison index in assessing the clustering method are the processing time and the purity level of the classes formed. Building condition Based on the existing data, the building data clustering process was divided into 4 clusters according to the existing data. Buildings with any ID that were grouped/clustered; they were determined by the similarity of the values of the parameters in the dataset.

III. RESULTS AND DISCUSSION

A. External Validity

This study used a ranking index to validate the external strategic performance. External validity is done by comparing the ranking index calculation using an external class with the cluster formed by the procedure. The building dataset was

divided into four categories based on the percentage of building damage determined through the results of recording and inspection on the existing form, namely the secure percentage > 85%, normal percentage 70-85%, unsafe percentage 40-69 percent, and not safe percentage < 40%, as shown in Table 2. The percentage value is obtained from the ratio of all data points to forty (simple number of building components) multiplied by one hundred percent. To get the proportion of the basic building using a simple building evaluation technique.

Column 1 in Table 3 shows the zone numbering, column 2 is the index of the condition of the building, column 3 is a description of the condition of the building, column 4 is what action should be taken regarding the condition of the building, and column 5 is the building categories. If the condition of the building is very well, then no immediate action is required, and the building is in the secure category. If the condition of the building is well, then the actions taken are only monitoring and prevention where the building is in the normal safe category. Conditions that must be seriously considered are intermediate conditions where the action is required to determine the appropriate course of action. It is necessary to conduct an alternative economic analysis of improvements. The building is in the unsafe category. Finally, the most severe is the bad/dangerous condition, so action is required. A thorough evaluation is required to determine the necessary repair, rehabilitation, and reconstruction actions and assess the safety and whether the building is in the not-safe category.

The experiment is repeated twenty times for each technique on a PC equipped with an Intel i5-8400 six-core processor running at 2.8 GHz and 8 GB RAM and the MATLAB programming language. Averages are used to calculate the rank index and time response. The results of the index rank calculation can be seen in Table 4.

TABLE IV
TIME RESPONSES

Indicator	FC	FKP	MMDS	Improvement
Rank Index (%)	68.89	69.53	71.07	3 %
Time Response (second)	0.0432	0.2884	0.0186	93.54%

The index value is obtained from the dataset's average rank index calculation of 303 buildings (See Fig. 1).

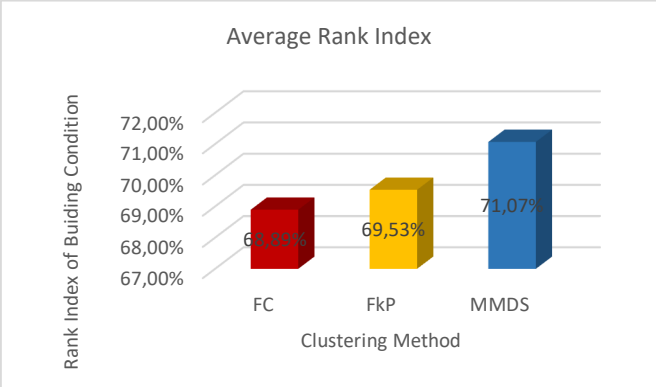

Fig. 1 The Average Rank Index

The ranking index calculation is based on the results of building clustering from three different methods: Fuzzy Centroid, Fuzzy k-mean Partition, and Multivariate Multinominal Distribution Softset. The results of the rank index with dataset clustering by Fuzzy Centroin, the index value is 68.8947. The rank index result using the clustering dataset by Fuzzy k-means partition is 69.5326, and the rank index result using the clustering dataset by the softest multivariate multinominal distribution is 71.067.

The results of the response time measurements can be seen in Table 4 as well. The clustering process for building datasets using the baseline method, namely Fuzzy Centroid, requires a response time of 0.0432 seconds, and the result of the clustering response time of the Fuzzy k-means Partition method is 0.2884. Clustering using the proposed method of multivariate multinominal distribution soft set (MMDS), the recorded response time is 0.0186. The response time generated by MMDS improves 93.54% over the response time by Fuzzy k-means Partition.

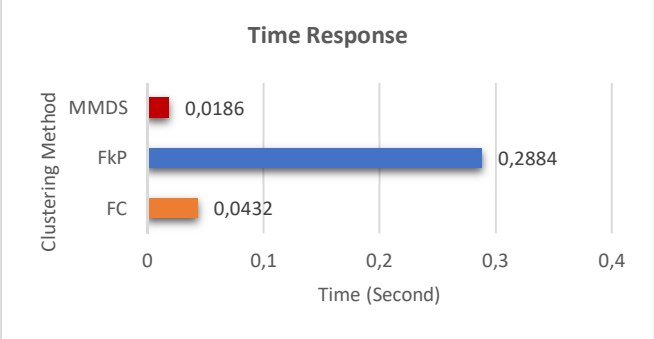

Fig. 2 Time Response of Clustering Dataset

B. *Internal validity based on Number of Clusters*

Internal validity is carried out by looking at the stability of the clustering results of the three compared methods: two baseline methods, fuzzy centroid and fuzzy k-means partition, and one proposed method, namely the multivariate multinomial distribution softest. Clustering stability can be seen by observing the stability of the number of clusters created against the increase in the number of clusters themselves. This section describes the stability performance of the three clustering methods used in this research.

The results of comparing the performance stability of the dataset clustering process from the three methods can be seen in Figure 3. Based on Figure 3, the results show that the Fuzzy Centroid and Multivariate Multinominal Distribution Softset are more stable than the Fuzzy k-means partition.

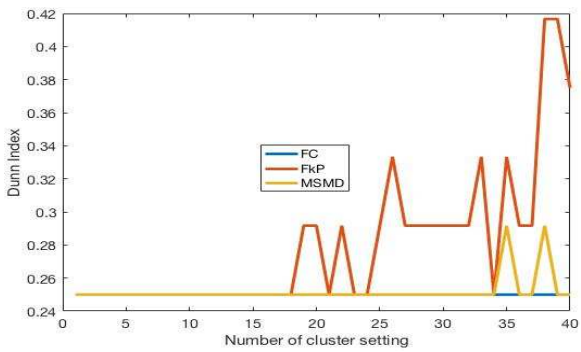

Fig. 3 The Dunn Index

The dunn index graph for Fuzzy Centroid looks very stable where the value of the Dunn index on the number of clusters from 2 to 40 looks the same with no change, which has a value of 0.25. The Dunn index value for the Fuzzy k-means Partition method shows less stability, whereas the Dunn index value for the 18th, 20th, 22nd and so on changes. The result of the proposed method, the soft set multivariate multinomial distribution, shows better stability than the baseline method of Fuzzy k-means Partition. The results of the Dunn index from the MMDS method are also shown in Figure 4.

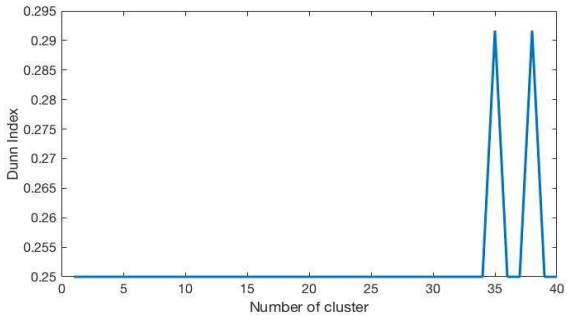

Fig. 4 The Dunn index of the data using the proposed approach

Figure 4 shows the dunn index on the data using the proposed multivariate multinominal distribution softest method based on the number of clusters given. It appears that clusters 2 to 34 have the same Dunn index value of 0.25 and slightly increase in the number of clusters of 35, and so on.

C. *Implementation on dataset*

Based on the rank index values, the technique has good performance. Then, figure 3 illustrates the Dunn index of the baseline and the proposed technique concerning the increasing number of clusters. Figure 4 is a subfigure on the Dunn index in 2- 40 clusters. It can be seen the dun index of MMD is stable in a variety of a number of clusters. Thus, any number of clusters can be selected based on user necessity. To divide the data into several levels of impact, therefore the cluster is determined to be 4 clusters (category of the level building). This condition is also under the data set obtained, where the building condition is categorized into four conditions: dangerous, vulnerable, normal, and safe. The results of clustering the data set divided into four clusters can be seen in Table 5.

TABLE V
THE CLUSTERING RESULTS OF RVS DATASET (BUILDING) WITH 4 CLUSTERS

Cluster	Condition Index	Level	Number of members			
			North Area	South Area	West Area	East Area
C 1	86 -100	High	29	11	20	22
C 2	0 -39	Low	7	32	15	18
C 3	40 - 69	Moderate	19	23	20	25
C 4	70 - 85	Medium	17	10	9	9

It can be seen in Table 5 that the distribution of the number of buildings in the area is divided into north, south, west, and east areas. The total number of cluster members is the same as the number of datasets processed—namely, 286. The clusters formed are in the high category with 82 members, moderate with 87 members, medium with 45 members, and low with 72 members. The area with the dominant number of low cluster members is the southern area, while the area with the least number of low cluster members is the northern area. This data shows that buildings in the southern area need more attention regarding their unsafe condition.

IV. CONCLUSION

Data collection will not be useful if the information and knowledge contained in it cannot be extracted. There have been many data mining methods developed. One of the important parts of the data mining process is the data classification or clustering process, where large data are grouped into smaller clusters based on similarities and homogeneity.

This research introduces the proposed method, namely the multivariate multinomial distribution soft set, and compares the results of the clustering process with two baseline methods. The results show that the proposed method provides a better response time and higher purity and stabilization of the number of clusters than the two comparison methods.

The clustering process of 286 building data sets used in this study formed 4 clusters: a cluster with a certain category of 82 buildings, a normal/moderate cluster of 87 buildings, a medium/unsafe cluster of 45 buildings, and a hazard/ not safe cluster of 72 buildings. The southern region is where hazard category buildings dominate the number of buildings.

REFERENCES

[1] A. Barbaresi, M. Bovo, and D. Torreggiani, "The dual influence of the envelope on the thermal performance of conditioned and unconditioned buildings," *Sustain. Cities Soc.*, vol. 61, p. 102298, 2020.

[2] E. Harirchian, K. Jadhav, K. Mohammad, S. E. A. Hosseini, and T. Lahmer, "A comparative study of MCDM methods integrated with rapid visual seismic vulnerability assessment of existing RC structures," *Appl. Sci.*, vol. 10, no. 18, 2020.

[3] M. M. Kassem, F. Mohamed Nazri, and E. Noroozinejad Farsangi, "The seismic vulnerability assessment methodologies: A state-of-the-art review," *Ain Shams Eng. J.*, vol. 11, no. 4, pp. 849–864, 2020.

[4] A. Darko, A. P. C. Chan, Y. Yang, and M. O. Tetteh, "Building information modeling (BIM)-based modular integrated construction risk management – Critical survey and future needs," *Comput. Ind.*, vol. 123, p. 103327, 2020.

[5] C. Wan, M. Ye, C. Yao, and C. Wu, "Brain MR image segmentation based on Gaussian filtering and improved FCM clustering algorithm," in *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 2017, pp. 1–5.

[6] R. R. Saedudin, S. B. Kasim, H. Mahdin, and M. A. Hasibuan, "Soft Set Approach for Clustering Graduated Dataset," in *International Conference on Soft Computing and Data Mining*, 2016, pp. 631–637.

[7] R. R. Saedudin, S. B. Kasim, H. Mahdin, and M. A. Hasibuan, "Soft Set Approach for Clustering Graduated Dataset BT - Recent Advances on Soft Computing and Data Mining," 2017, pp. 631–637.

[8] R. Shanker and M. Bhattacharya, "Brain Tumor Segmentation of Normal and Pathological Tissues Using K-mean Clustering with Fuzzy C-mean Clustering," in *VipIMAGE 2017*, 2018, pp. 286–296.

[9] E. Sutoyo, I. T. R. Yanto, R. R. Saedudin, and T. Herawan, "A soft set-based co-occurrence for clustering web user transactions," *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 15, no. 3, 2017.

[10] A. S. M. S. Hossain, "Customer segmentation using centroid based and density based clustering algorithms," in *2017 3rd International*

Conference on Electrical Information and Communication Technology (EICT), 2017, pp. 1–6.
[11] K. V Ahammed Muneer and K. Paul Joseph, "Performance Analysis of Combined k-mean and Fuzzy-c-mean Segmentation of MR Brain Images," in *Computational Vision and Bio Inspired Computing*, 2018, pp. 830–836.
[12] H. Zhou, "K-Means Clustering BT - Learn Data Mining Through Excel: A Step-by-Step Approach for Understanding Machine Learning Methods," H. Zhou, Ed. Berkeley, CA: Apress, 2020, pp. 35–47.
[13] S. Irfan, G. Dwivedi, and S. Ghosh, "Optimization of K-means clustering using genetic algorithm," in *2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN)*, 2017, pp. 156–161.
[14] B. K. D. Prasad, B. Choudhary, and B. Ankayarkanni., "Performance Evaluation Model using Unsupervised K-Means Clustering," in *2020 International Conference on Communication and Signal Processing (ICCSP)*, 2020, pp. 1456–1458.
[15] W. Wei, J. Liang, X. Guo, P. Song, and Y. Sun, "Hierarchical division clustering framework for categorical data," *Neurocomputing*, vol. 341, pp. 118–134, 2019.
[16] Z. Huang, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values," *Data Min. Knowl. Discov.*, vol. 2, no. 3, pp. 283–304, 1998.
[17] Y. Xiao, C. Huang, J. Huang, I. Kaku, and Y. Xu, "Optimal mathematical programming and variable neighborhood search for k-modes categorical data clustering," *Pattern Recognit.*, vol. 90, pp. 183–195, 2019.
[18] D. B. M. Maciel, G. J. A. Amaral, R. M. C. R. de Souza, and B. A. Pimentel, "Multivariate fuzzy k-modes algorithm," *Pattern Anal. Appl.*, vol. 20, no. 1, pp. 59–71, 2017.
[19] P. S. Bishnu and V. Bhattacherjee, "Software cost estimation based on modified K-Modes clustering Algorithm," *Nat. Comput.*, vol. 15, no. 3, pp. 415–422, 2016.
[20] M. K. N. Huang, "A fuzzy k-modes algorithm for clustering categorical data - Fuzzy Systems, IEEE Transactions on," *IEEE Trans. Fuzzy Syst.*, vol. 7, no. 4, pp. 446–452, 1999.
[21] D.-W. Kim, K. H. Lee, and D. Lee, "Fuzzy clustering of categorical data using fuzzy centroids," *Pattern Recognit. Lett.*, vol. 25, no. 11, pp. 1263–1271, Aug. 2004.
[22] M. S. Yang, Y. H. Chiang, C. C. Chen, and C. Y. Lai, "A fuzzy k-partitions model for categorical data and its comparison to the GoM model," *Fuzzy Sets Syst.*, vol. 159, no. 4, pp. 390–405, 2008.
[23] S. Ben-David, D. Pál, and H. Simon, *Stability of k-Means Clustering*. 2007.
[24] I. Landi, V. Mandelli, and M. V. Lombardo, "reval: a Python package to determine the best number of clusters with stability-based relative clustering validation," *arXiv*, vol. 2, no. 4. arXiv, p. 100228, 27-Aug-2020.
[25] D. G. L. Allegretti, "Stability conditions, cluster varieties, and Riemann-Hilbert problems from surfaces," *Adv. Math. (N. Y).*, vol. 380, p. 107610, Mar. 2021.
[26] E. Andreotti, D. Edelmann, N. Guglielmi, and C. Lubich, "Measuring the stability of spectral clustering," *Linear Algebra Appl.*, vol. 610, pp. 673–697, Feb. 2021.
[27] T. Herawan and M. M. Deris, "On Multi-soft Sets Construction in Information Systems BT - Emerging Intelligent Computing Technology and Applications. With Aspects of Artificial Intelligence," 2009, pp. 101–110.
[28] I. T. R. Yanto, R. Setiyowati, M. M. Deris, and N. Senan, "Fast Hard Clustering Based on Soft Set Multinomial Distribution Function BT - Recent Advances in Soft Computing and Data Mining," 2022, pp. 3–13.
[29] I. T. R. Yanto, M. M. Deris, and N. Senan, "PSS: New Parametric Based Clustering for Data Category BT - Recent Advances in Soft Computing and Data Mining," 2022, pp. 14–24.
[30] I. Tri, R. Yanto, R. Saedudin, S. Novita, M. Mat, and N. Senan, "Soft Set Multivariate Distribution for Categorical Data Clustering," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 11, no. 5, pp. 1841–1846, 2021.
[31] I. Tri, R. Yanto, A. Apriani, R. Hidayat, M. Mat, and N. Senan, "Fast Clustering Environment Impact using Multi Soft Set Based on Multivariate Distribution," *JOIV Int. J. Informatics Vis.*, vol. 5, no. September, pp. 291–297, 2021.
[32] T. Herawan, M. M. Deris, and J. H. Abawajy, "Matrices Representation of Multi Soft-Sets and Its Application," in *Computational Science and Its Applications -- ICCSA 2010: International Conference, Fukuoka, Japan, March 23-26, 2010, Proceedings, Part III*, D. Taniar, O. Gervasi, B. Murgante, E. Pardede, and B. O. Apduhan, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 201–214.