Soft Set Multivariate Distribution for Categorical Data Clustering

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

[a] *Department of Information Systems, University of Ahmad Dahlan, Yogyakarta, Indonesia*
[b] *Department of Information Systems, Telkom University, Bandung, West Java, 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, Parit Raja, Batu Pahat, Johor, Malaysia*
[e] *Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, Johor, Malaysia*
Corresponding author: [*] yanto.itr@is.uad.ac.id

***Abstract*— Clustering is the process of breaking down a huge dataset into smaller groups. It has been used in some field studies including pattern recognition, segmentation, and statistics with remarkable success. Clustering is a technique for dividing multivariate datasets into groups. No inherent distance measure on data category makes clustering data more challenging than numerical data. Data category can be assumed following the data from a multinomial distribution. Thus, the standard model parametric model can be used in latent class clustering based on the independent product of multinomial distributions. Meanwhile, multi-valued attributes on the categorical data can be decomposed into the standard set on a multi soft set. In this paper, a clustering technique based on soft set theory is proposed for categorical data through a multinomial distribution. The data will be represented as a multi soft set which is every soft set has its probability of being a member of the cluster. The data with the highest probability will be assigned as the member of the cluster. The experiment of the proposed technique is evaluated based on the Dunn index with regard to the number of clusters and response time. The experiment results show that the proposed technique has the lowest response time with high stability compared to baseline techniques. This study recommends a maximum number of clusters in implementation on the real data.**

***Keywords*— Clustering; categorical data; soft set; multivariate.**

Manuscript received 6 Apr. 2021; revised 22 Jul. 2021; accepted 2 Aug. 2021. Date of publication 31 Oct. 2021.

I. INTRODUCTION

The process of dividing a large dataset into smaller classes is known as clustering. It has been successfully applied in fields such as pattern recognition, segmentation, and statistics [1]–[4]. Clustering which divides multivariate datasets into groups has been studied by Zhou [5] and Irfan *et al.* [6]. The k-means clustering technique is the most effective and efficient clustering techniques for large datasets [7]. Because k-objective mean's function is a numerical cost function, it cannot handle categorical data. Unlike numerical datasets, categorical objects do not have an inherent distance measure. As a result, clustering numerical data is easier than clustering categorical data [8].

The application of clustering techniques developed for numerical data cannot be carried out directly on categorical data. Numerous clustering techniques have been proposed to address this issue. As a result, hard k-mode employs a simple matching function to circumvent the k-means constraint on data categorization [9]. Then, a new dissimilarity measure is used to enhance the hard k-modes [10]–[12] and to create fuzzy k-modes [13]. Kim *et al.* [10] demonstrate how to increase the efficiency of fuzzy k-mode by converting it to the fuzzy centroid. Yang *et al.* [14] proposes a highly effective clustering technique based on parametric data called fuzzy K-partitioning. However, almost all of the algorithms mentioned previously require the data to be represented in binary values. Thus, the disadvantage of the aforementioned approaches is that they typically require a large amount of computational time and have a low degree of cluster purity. This indicates the need for a method that is not hampered by lengthy computations or low cluster purity. Furthermore, the number of clusters is determined in a data set, the use of the letter k to represent the quantity as in the k-means algorithm, is a problem that often occurs in data clustering and is different from the process of solving clustering problems. The selection of the correct k is frequently ambiguous, with interpretations varying according to the scale and shape of the data set's point distribution and the clustering resolution desired by the user. Additionally, the number of errors in clustering generated to points can be reduced by increasing k without penalty, where

each data point is considered as its own cluster (i.e., 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 for each data point is achieved when the value of k is optimal. Another method should be chosen if the appropriate value for k is not clear from prior knowledge of the data set properties [15]. Thus for the clustering optimization problem, the clustering stability in terms of the optimal number of solutions is a frequently used heuristic for determining the cluster size in various clustering applications [16], [17]–[19].

This article proposes a clustering technique via multinomial distribution for categorical data based on SST (soft set theory). Multi-valued information systems can be used to represent categorical data [20]. It might be a multivariate multinomial distribution used to take a random sample. A standard parametric model used in clustering latent classes for multivariate categorical data is the locally independent product of the multinomial distribution. (i.e., within clusters) [21]. Additionally, the multi-valued information system can be used to represent categorical data as a soft set [20] rather than as binary values. In the experiment, validation based on internal cluster analysis is used to determine the cluster's stability as the number of clusters increases.

II. Materials and Method

A. Fuzzy Clustering for categorical data

Recently, numerous fuzy-based clustering techniques for handling data categories have been proposed due to their superior performance in both the theoretical and practical realms. Huang proposed the hard k-mode [9], which was improved by applying new dissimilarity measures to the k-modes clustering and using a fuzzy set-based k-modes algorithm [6]–[8]. Assuming that is a membership function, that y is data, and that v is the cluster's centroid, the objective of hard k-mode is to minimize the function $H_m(\mu, v)$.

$$H_m(\mu, v) = \sum_{i=1}^{I} \sum_{k=1}^{K} \mu_{ik}^m d(y_i, v_k), \quad (1)$$

subject to

$$\sum_{k=1}^{K} \mu_{ik} = 1, \text{for } i = 1,2, \dots, I, \quad (2)$$

where $d(y_i, v_k) = \sum_{j=1}^{J} \delta(y_{ij}, v_{kj})$ is called the simple matching dissimilarity measure, $\delta(y_{ij}, v_{kj}) = 0$ if $y_{ij} = v_{kj}$ and $\delta(y_{ij}, v_{kj}) = 1$, if $y_{ij} \neq v_{kj}$. y is the categorical data value and v is the cluster centroid. m is the fuzziness index. The solution of the objective on hard k-modes are written as:

$$\mu_{ik} = \begin{cases} 1 & if \quad d(y_i, v_k) = \min\limits_{1 \le k' \le K} d(y_i, v_k') \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

$$v_{kjl} = \begin{cases} 1 & if \quad \sum_{i=1}^{I} \mu_{ik} y_{ijl} = \max\limits_{1 \le l' \le L} \sum_{i=1}^{I} \mu_{ik} y_{ijl}' \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Hard k-modes were expanded to fuzzy k-modes as done by Huang [9]. Thus, the solution is given by update the equation as follows:

$$\mu_{ik} = 1 / \sum_{k'=1}^{K} \left[\frac{d(y_i, v_k)}{d(y_i, v_{k'})} \right]^{\frac{1}{m-1}} \quad (5)$$

$$v_{kjl} = \begin{cases} 1 & if \quad \sum_{i=1}^{I} \mu_{ik}^m y_{ijl} = \max\limits_{1 \le l' \le L} \sum_{i=1}^{I} \mu_{ik}^m y_{ijl}, \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

The show of fuzzy k-modes is improved by making hard centroids to fuzzy centroid by Kim et al. [22] where $\tilde{v}_{kj} = \left(\tilde{v}_{kj1}, \dots, \tilde{v}_{kjL_j} \right)$, for $k = 1,2, \dots K$ and $j = 1,2, \dots, J$, where $\tilde{v}_{kjl} \in [0,1]$ and $\sum_{i=1}^{L_j} \tilde{v}_{kjl} = 1$. Minimizing H_m (μ,v) is a fuzzy centroid objective function and is written as follows:

$$H_m(\mu, v) = \sum_{i=1}^{I} \sum_{k=1}^{K} \mu_{ik}^m d(y_i, \tilde{v}_k), \quad (7)$$

subject to

$$\sum_{k=1}^{K} \mu_{ik} = 1, i = 1,2, \dots, I. \quad (8)$$

$$\sum_{l=1}^{L_j} \tilde{v}_{kjl} = 1. \quad (9)$$

The update equation based on the distance measure with the centroid is given

$$d(y_i, \tilde{v}_k) = \sum_{j=1}^{J} \delta(y_{ij}, \tilde{v}_{kj}) = \sum_{j=1}^{J} \sum_{l=1}^{L_j} (1 - y_{ijl}) \tilde{v}_{kjl}, \quad (10)$$

$$\tilde{v}_{kjl} = \frac{\sum_{i=1}^{I} \mu_{ik}^m \cdot y_{ijl}}{\sum_{i=1}^{I} \mu_{ik}^m}. \quad (11)$$

The update result of membership equations can be written as:

$$\mu_{ik} = 1 / \sum_{k'=1}^{K} \left[\frac{d(y_i, \tilde{v}_k)}{d(y_i, \tilde{v}_{k'})} \right]^{\frac{1}{m-1}}. \quad (12)$$

Fuzzy k-modes with fuzzy and hard centroid are non-parametric techniques. The dissimilarity function used in this algorithm is based on the smallest total in the cluster match dissimilarity.

B. Soft Set Multivariate Distribution for Categorical Data Clustering

Presentation of data in finite tables is often used (later referred to as an information system), columns and rows are labeled variables and objects of interest, and the entries with variable values. A formal definition of an information system written in the form $S = (U, A, V, f)$ is as a quadruple (four-tuple) with U and A are a non-empty finite set of objects and a non-empty finite set of variables, respectively. While, $V = \bigcup_{a \in A} V_a$ with V_a is the domain (value set) of variable a, the total function is written $f : U \times A \to V$ such that $f(u, a) \in V_a$, $\forall (u, a) \in U \times A$, which is called information function.

The proposed technique represents the data using a soft set. The same-valued data are decomposed into multiple soft sets. Because each soft set has the same value, all members of that soft set have the same probability of being assigned to the

cluster. Thus, we will use the multinomial distribution function to determine the high probability of each instance with respect to all parameters in the data.

Definition 1. Let U and E be a universe set and parameter set, respectively, $A \subset E$, and F is the function that mapping parameter A into the set of all subsets of the set U as

$$F: A \to P(U).$$

The pair of (F, A) is called a soft set over U. While, the set of a-approximate elements of (F, A) is written as $F(a)$, $\forall_{a \in A}$.

Definition 2. Let$S = (U, A, V, f)$ be a categorical-valued information system, where $U = \{u_1, u_2, \dots, U_n\}$ is finite set of instance, $A = \{a_1, a_2, \dots a_m\}$ is a finite set of the attribute, V is values set of each attribute A, f is mapping function $f: (U, A) \to V$ and $S = \left(U, a_i, V_{a_i}, f\right), i = 1,2, \cdots, |A|$ Boolean-valued information system, it can be decomposed to be multi-boolean information system as

$$S = (U, A, V, f) = \begin{cases} S^1 = \left(U, a_1, V_{\{0,1\}}, f\right) \Leftrightarrow (F, a_1) \\ S^2 = \left(U, a_2, V_{\{0,1\}}, f\right) \Leftrightarrow (F, a_2) \\ \vdots \\ S^{|A|} = \left(U, a_{|A|}, V_{\{0,1\}}, f\right) \Leftrightarrow (F, a_{|A|}) \end{cases} = \left((F, a_1), (F, a_2), \cdots, \left(F, a_{|A|}\right)\right) \tag{13}$$

Then, the information system with categorical value $S = (U, A, V, f)$ which is represented by a multi-universe soft set of U which is defined in the form $(F, E) = \left((F, a_1), (F, a_2), \cdots, \left(F, a_{|A|}\right)\right)$.

An information system with categorical value $S = (U, A, V, f)$ is represented by considering pairs (F, A) and assigning them to the multisoft set over U, where $(F, a_1), \cdots, \left(F, a_{|A|}\right) \subseteq (F, A)$ and $\left(F, a_{j_1}\right), \cdots, \left(F, a_{j_{|a_j|}}\right) \subseteq \left(F, a_j\right)$. Lets say $\lambda_{kjl}^{u_i}$ be a probability of $u_i \in \left(F, a_{jl}\right)$ into cluster $C_k, k = 1,2, \dots, K,$ where $i = 1,2, \dots, |U|, j = 1,2, \dots, |A|$ and $l = 1,2, \dots, \left|a_j\right|$, thus, the multivariate multinomial distribution of multi soft set can be defined as

$$Maximize\ L_{CML}(z, \lambda) = \sum_{i=1}^{|U|} \sum_{k=1}^{K} z_{ik} \sum_{j=1}^{|A|} \sum_{l=1}^{|a_j|} \ln\left(\lambda_{kjl}^{u_i}\right)^{\left|F, a_{j_l}\right|} \tag{14}$$

Subject to

$$\sum_{k=1}^{K} z_{ik} = 1, for\ i = 1,2, \dots, |U|. \tag{15}$$

$$\sum_{l=1}^{|a_j|} \lambda_{kjl} = 1. \tag{16}$$

The following equations

$$\lambda_{kjl} = \frac{\sum_{u_i \in \left(F, a_{jl}\right)} z_{ik}(u_i)}{|U|} \tag{17}$$

$$z_{ik} = \begin{cases} 1 & if \quad \sum_{j=1}^{|A|} \ln \lambda_{kjl}^{u_i} = \max_{1 \le k' \le K} \sum_{j=1}^{|A|} \ln \lambda_{k'jl}^{u_i} \\ 0 & otherwise \end{cases} \tag{18}$$

are updated to get the maximum objective function $L_{CML}(z, \lambda)$.

The technique analyzes categorical data and decomposes it into multiple soft sets using the relationship defined in Definitions 1 and 2. This decomposition results in the formation of an approximation space composed of equivalence classes. As a result, categorical data does not need to be converted to binary values. Additionally, each member of the soft set has a similar probability. Thus, the multinomial distribution function can be used to determine the likelihood that each data point will be assigned to a cluster. Thus, Figure 1 illustrates the proposed algorithm's pseudo-code.

Soft Set Multivariate Distribution based Algorithm Input: Categorical data set Output: Clusters
Begin 1. Decompose the data into multi soft set. 2. Compute the random initial z_{ik} 3. Repeat a. Update λ_{kjl} b. Update z_{ik} 4. Until μ_{ik} estimates to be stabil ($\| L_{cml}^{it}(z, \lambda) - L_{cml}^{it-1}(z, \lambda)\| < tol$ or $\| z_{ik}^{it} - z_{ik}^{it-1} \| < tol$) or given maximum number of iteration. End

Fig. 1 Proposed algorithm

III. Results and Discussion

This section validates the proposed algorithms using benchmark datasets from the Irvine's machine learning department, University of California. The experiments are carried out using the MATLAB programming language on a PC equipped with an Intel i5-8400 six-core CPU running at 2.8 GHz and 8 GB RAM. The experiments are designed to compare the proposed technique to two industry-standard techniques: fuzzy centroid and fuzzy k-partitioning. Fuzzy centroid measures the distance to the centroid using a simple matching dissimilarity function. It is a non-parametric approach, but it is based on distance, which means that the closest centroid determines the best centroid. As a result, a spherical cluster is formed. It may have a low purity level. Meanwhile, the fuzzy k-partition is a parametric approach dependent on the multinomial distribution function's likelihood function. The data, on the other hand, must be represented as a binary variable. As a result, it frequently results in lengthy computations. We illustrate the three approaches using UCI benchmark datasets.:

- Zoo data set, 101 objects, 18 attributes.
- Balloon dataset, 20 object and 4 attributes.
- Monk dataset, 432 object and 6 attributes.

- Spect dataset, 187 object and 922 attributes.
- Breast dataset, 683 object and 9 attributes.

All comparisons in this section are made using the Dunn index in terms of cluster size and response time. The experiments are set up for all algorithms with a cluster size of 2 to 100 or a maximum number of instances. Each technique is repeated twenty times. If the data is clustered into a number of cluster sets or the maximum number of instances, the technique is called divergent; if the data is clustered into a single cluster, the technique is called convergent to 1. Obviously, when convergent approaches 1, all data is clustered into a single cluster. The experiment's results are summarized in Tables 1 and 2 in terms of cluster number and cluster stability, respectively.

TABLE I
RESPONSE TIMES FOR DIFFERENT DATASETS

Data set	Response Times			Improvement (%)
	FC	FkP	Proposed	
Zoo	0.8732	0.2617	0.0236	90.98
Balloon	0.6914	1.2404	0.0273	97.80
Monk	0.9206	0.3754	0.0253	93.26
Spect	0.5662	0.4645	0.0995	78.58
Average	0.7629	0.5855	0.0439	92.50

According to Table 1, the proposed approach outperforms the baseline technique in terms of time consumption. The proposed approach has the potential to significantly reduce response time by as much as 92.50 percent on average.

The proposed technique is more stable than the established techniques. It is demonstrated using a balloon data set with a size of 20 objects. Thus, the cluster setting ranges from two to twenty (maximum number of instances). Fuzzy centroid-created clusters in accordance with the number of clusters specified; for example, if the number of clusters is set to 5, fuzzy centroid will create five clusters, indicating that it is likely to diverge. Meanwhile, the FkP technique converges to a single cluster when clusters are greater than 10. Additionally, the proposed approach is more consistent in the 9-10 range. The Dunn index for stability and the number of clusters created on the balloon data set are illustrated in Fig. 2 and Fig. 3, while the zoo data set is illustrated in Fig. 4 and Fig. 5. It demonstrates that the proposed technique is more stable than the baseline techniques. Table 2 summarizes the stability with respect to the cluster count.

TABLE II
STABILITY COMPARISON BASED ON NUMBER OF CLUSTERS

Data	**Size of data**	**Number of clusters created**		
		FC	**FkP**	**Proposed**
Balloon	(20,4)	Divergent	Convergent to 1	Convergent to 9-10
Breast	(683,9)	Convergent ke 4	Divergent	Convergent to 80-85
Monk	(432,6)	Divergent	Divergent	Convergent to 70
Spect	(187,22)	Divergent	Convergent to 1	Convergent to 45-49
Zoo	(101,16)	Convergent 59	Convergent to 1	Convergent to 25-29

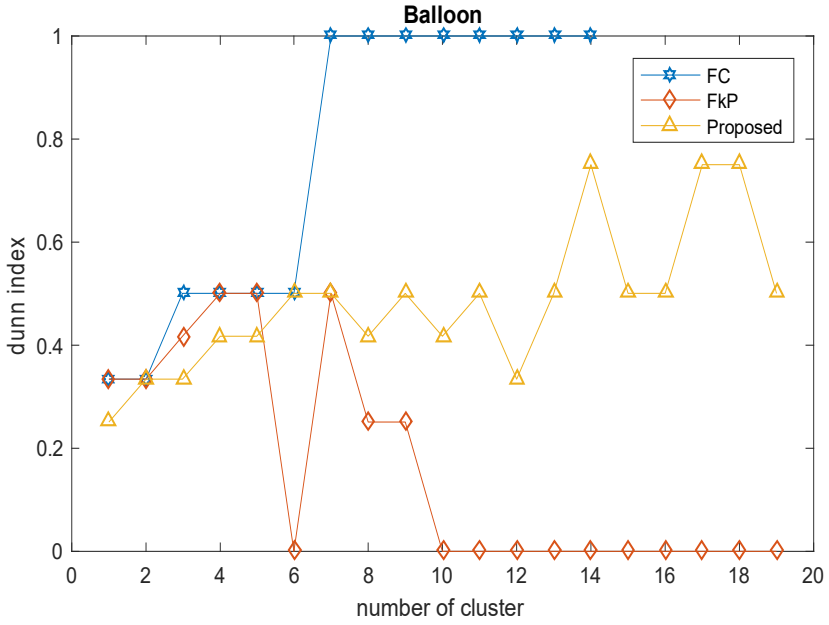

Fig. 2 The Dunn Index of Balloon Data Set

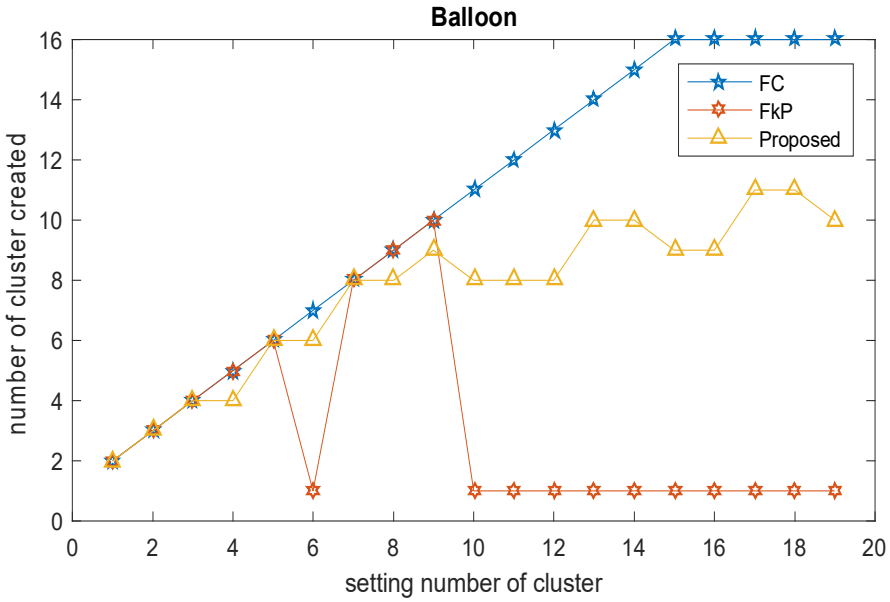

Fig. 3 The number of cluster created by given maximum number cluster setting of balloon dataset

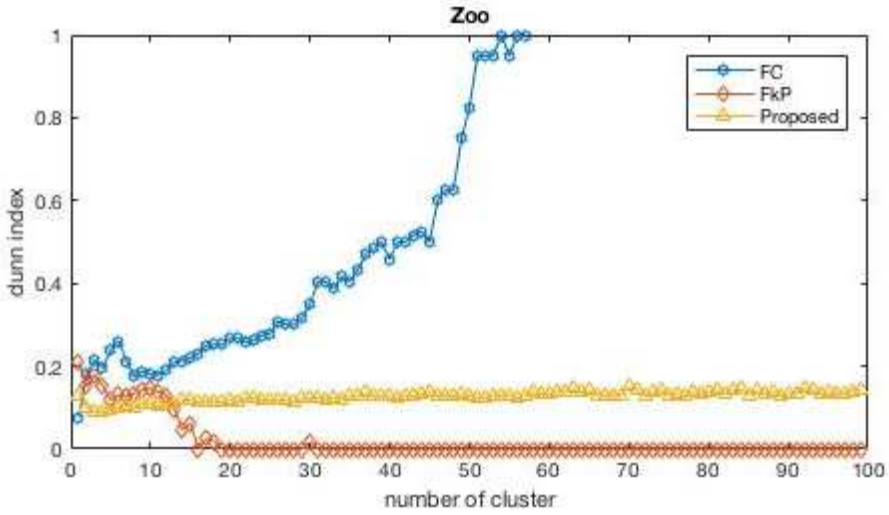

Fig. 4 The Dunn Index of Zoo Data Set

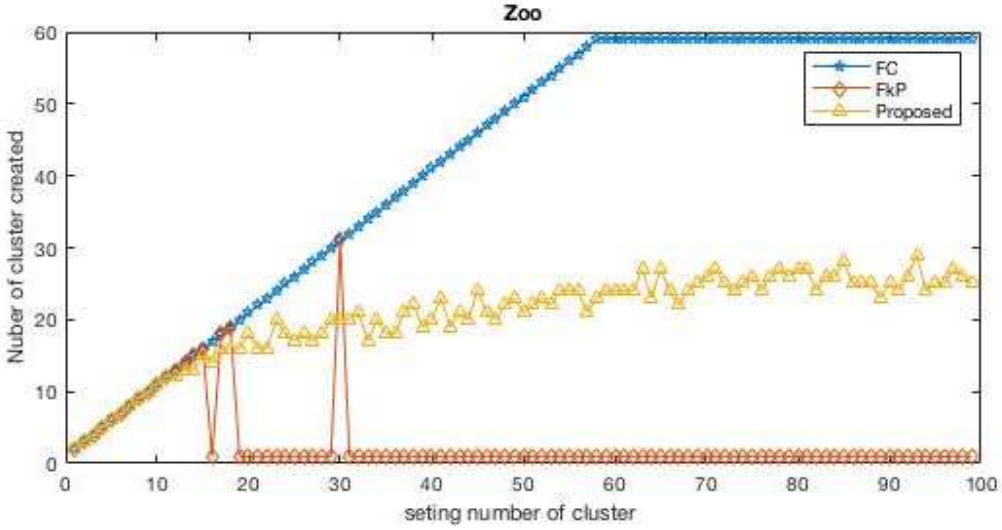

Fig. 5 The number of cluster created by given maximum number cluster setting of zoo dataset

IV. Conclusions

In this paper, the technique of soft set for categorical data clustering via multinomial distribution function is presented. For the experimental investigation, the benchmark datasets taken from UCI are used to compare the proposed technique with the existing techniques in terms of Dunn index and computational time with respect to a number of clusters. Based on the experiments that have been carried out in the five benchmark datasets, the proposed technique gives the result that the technique has a better stability number of clusters and achieves lower computational time. It can recommend a maximum number of clusters in implementation on the real data.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] Z. Huang and M. K. Ng, "A fuzzy k-modes algorithm for clustering categorical data," IEEE Trans. Fuzzy Syst., vol. 7, no. 4, pp. 446–452, 1999.

[14] 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.

[15] A. Karim, C. Loqman, and J. Boumhidi, "Determining the number of clusters using neural network and max stable set problem," Procedia Comput. Sci., vol. 127, pp. 16–25, 2018.

[16] S. Ben-David, D. Pál, and H. Simon, Stability of k-Means Clustering. 2007.

[17] 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.

[18] D. G. L. Allegretti, "Stability conditions, cluster varieties, and Riemann-Hilbert problems from surfaces," Adv. Math. (N. Y)., vol. 380, p. 107610, Mar. 2021.

[19] 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.

[20] 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.

[21] D. S. Morris, A. M. Raim, and K. F. Sellers, "A Conway–Maxwell-multinomial distribution for flexible modeling of clustered categorical data," J. Multivar. Anal., vol. 179, p. 104651, 2020.

[22] 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.