# Soft Set Based Clustering and Its Comparison on Categorical Data


Iwan Tri Riyadi Yanto
*Information System Department*
*Universitas Ahmad Dahlan*
Yogyakarta, Indonesia
yanto.itr@is.uad.ac.id

Cheah WaiShiang
*Faculty of Computer Science & IT,*
*Universiti Malaysia Sarawak*
Samarahan, Sarawak, Malaysia
wscheah@unimas.my

Rahmat Hidayat
*Information Technology Department*
*Politeknik Negeri Padang*
Padang, Indonesia
rahmat@pnp.ac.id

Rofiul Wahyudi
*Islamic Banking Department*
*Universitas Ahmad Dahlan*
Yogyakarta, Indonesia
rofiul.wahyudi@pbs.uad.ac.id

Suprihatin
*Information System Department*
*Universitas Ahmad Dahlan*
Yogyakarta, Indonesia
suprihatin@is.uad.ac.id

Ani Apriani
*Institute Teknologi Nasional*
*Yogyakarta*
Yogyakarta, Indonesia
aniapriani@itny.ac.id



***Abstract*—Categorical data clustering is problematic since it is difficult or complex to determine how comparable the data is. Several methods, most recently centroid-based strategies, have been developed to reduce the complexity of the similarity of categorical data. These methods nevertheless result in lengthy processing durations. Another method, soft set-based clustering (SSC), based on the probability function of multivariate multinomial distributions, is suggested in this article. Soft sets are used to represent the data, and each soft set has a probability for each object. The joint cluster distribution function determines the probability for each object after the multivariate multinomial distribution function. The connected cluster would receive the highest likelihood. Benchmark data sets from UCI machine learning are used to compare the performance of the approach to the baseline techniques. The outcomes demonstrate that the suggested strategy performed better in purity, rank index, and calculation time.**



***Keywords—Soft set; categorical data; multinomial distribution.***


## I. INTRODUCTION (*HEADING 1*)

The concept of a cluster differs amongst algorithms; therefore, choosing the best approach for a given task requires making some choices. The belongings of the clusters discovered by various algorithms have differed greatly, and understanding the variations between these various clusters among the various techniques depends on these belongings. Two definitions are the foundation for the partitioning or clustering technique used to categorize the data into different groups. To quantify the similarity of two data objects, you must first define similarity or distance. Second, the characteristics of the data can be used to construct a method for deciding whether the item will be in the same groupings or divided into distinct groups [1], [2].

Because some metrics can reveal structural features, the clustering of numerical data in continuous space has been actively investigated in recent years. However, due to the discrete nature of the attribute value in categorical data, it may be challenging to determine the structural information of the data right once [3].

Due to the lack of natural order, high dimensionality, and conversion of categorical to numerical data, categorical data are hurdles to the existing clustering approaches [4]. Distance calculations cannot be performed directly on a categorical characteristic because, unlike numerical data, categorical data comprises attributes without a natural order. As a result, many data mining experts are drawn to the demanding and promising task of clustering categorical data [5].

Another issue is the lack of an inherent distance measurement between objects in categorical data. Clustering categorical data using the clustering techniques created for managing numerical data is impossible. Therefore, not only numbers make categorical data difficult to organize. Tallies from a predefined number of trials, where each trial makes a single decision from a predetermined set of categories, are typically observed for categorical data. It is reasonable to assume that the categorical data follows the multinomial distribution and is, therefore, trial independent. Therefore, the parametric technique is more appropriate for categorical data. [6]. A locally independent product of a multinomial is a widely used parametric model for categorical data in latent class clustering.

Contrarily, categorical data have several values for their attributes, which can be depicted as multiple soft [1]. The multi-soft set used for multi-valued characteristics provides benefits when encoding categorical data without transforming it to binary values. This study suggests a method based on soft set theory for multinomial distribution-based clustering of categorical data.

## II. LITERATURE REVIEW

### *A. Related Works of Clustering*

Observations within a data collection are divided into various groups using the clustering technique known as partitional clustering. The related work on categorical data clustering is given in Table 1. The first extension of *k*-mean to avoid the numerical limitation problem for categorical data clustering is hard k-modes. It takes advantage of a straightforward comparison of matching and centroid dissimilarities. In order to minimize the cost function, it employs a frequency-based update in each iteration and computes the cluster centroids in each iteration by substituting modes for the means. Using simple matching dissimilarity distance obtains the weak intra-similarity [7] and makes either accuracy or purity will be low. In order to recognize this problem, the relative frequency of the parameters in the distance metric was added to modify the straightforward matching dissimilarity measure called Ng's and He's distance [8], [9]. The centroids are updated in the original k-modes using only the information contained within the cluster. As a result, weak clustering results may be obtained when the between-cluster separation is not considered. Fuzzy k-modes is the generalized form of the k-modes that includes fuzzy sets in the clustering procedure. To do so, fuzzy k-modes are added to the hard k-mode [10].

Despite efficiently handling categorical data sets, the Fuzzy k-modes technique employs a hard centroid representation for categorical data in a cluster. Hard rejection of data might result in misclassification in the area of uncertainty [11]. The initialization issue with the centroids, the final local solution found, and the need to change an extra control parameter for the membership fuzziness are all problems the fuzzy k-modes share with the k-modes. To avoid this limitation, Kim et al. [12] revealed that converting hard centroids to fuzzy centroids enhanced the performance of fuzzy k-modes.

TABLE I

THE CLUSTERING TECHNIQUE FOR CATEGORICAL DATA

| Existing Approach | Technique | Objective function | Solution |
|---|---|---|---|
| distanced based | Hard $k$-modes | $H_m(\mu, v) = \sum_{i=1}^{I} \sum_{k=1}^{K} \mu_{ik}^{m} d(y_i, v_k)$<br>$d(y_i, v_k) = \sum_{j=1}^{J} \delta(y_{ij}, v_{kj})$,<br>$\delta(y_{ij}, v_{kj}) = \begin{cases} 0 & if \quad y_{ij} = v_{kj} \\ 1 & if \quad y_{ij} \neq v_{kj} \end{cases}$. | $\mu_{ik} = \begin{cases} 1 & if \quad d(y_i, v_k) = \min_{1 \le k' \le K} d(y_i, v'_k) \\ 0 & \text{otherwise} \end{cases}$<br>$v_{kjl} = \begin{cases} 1 & if \quad \sum_{i=1}^{I} \mu_{ik}^{m} y_{ijl} = \max_{1 \le l' \le L} \sum_{i=1}^{I} \mu_{ik}^{m} y_{ijl} \\ 0 & \text{otherwise} \end{cases}$ |
| | Fuzzy k-modes | | $\mu_{ik} = 1 / \sum_{k'=1}^{K} \left[ \frac{d(y_i, v_k)}{d(y_i, v_{k'})} \right]^{\frac{1}{m-1}}$<br>$v_{kjl} = \begin{cases} 1 & if \quad \sum_{i=1}^{I} \mu_{ik}^{m} y_{ijl} = \max_{1 \le l' \le L} \sum_{i=1}^{I} \mu_{ik}^{m} y_{ijl} \\ 0 & \text{otherwise} \end{cases}$ |
| | Hard Centroid | $d(y_i, \tilde{v}_k) == \sum_{j=1}^{J} \sum_{l=1}^{L_j} (1 - y_{ijl}) \tilde{v}_{kjl}$, | $\mu_{ik} = \begin{cases} 1 & if \quad d(y_i, v_k) = \min_{1 \le k' \le K} d(y_i, v'_k) \\ 0 & \text{otherwise} \end{cases}$<br>$\tilde{v}_{kjl} = \frac{\sum_{i=1}^{I} \mu_{ik}^{m} \cdot y_{ijl}}{\sum_{i=1}^{I} \mu_{ik}^{m}}$ |
| | Fuzzy Centroid | $d(y_i, \tilde{v}_k) = \sum_{j=1}^{J} \sum_{l=1}^{L_j} (1 - y_{ijl}) \tilde{v}_{kjl}$, | $\mu_{ik} = 1 / \sum_{k'=1}^{K} \left[ \frac{d(y_i, \tilde{v}_k)}{d(y_i, \tilde{v}_{k'})} \right]^{\frac{1}{m-1}}$<br>$\tilde{v}_{kjl} = \frac{\sum_{i=1}^{I} \mu_{ik}^{m} \cdot y_{ijl}}{\sum_{i=1}^{I} \mu_{ik}^{m}}$ |
| Parametric based | GoM | $L_{GoM}(g, \lambda) = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{l=1}^{L_j} \ln \left( \sum_{k=1}^{K} g_{ik} . \lambda_{kjl} \right)^{y_{ijl}}$ | $g_{ik}^{(t+1)} = \frac{\sum_{j=1}^{J} \sum_{l=1}^{L_j} y_{ijl} \frac{g_{ik}^{(t)} . \lambda_{kjl}^{(t)}}{\sum_{k'}^{K} g_{ik'}^{(t)} . \lambda_{k'jl}^{(t)}}}{\sum_{j=1}^{J} \sum_{l=1}^{L_j} y_{ijl}}$,<br>$\lambda_{kjl}^{(t+1)} = \frac{\sum_{i=1}^{I} y_{ijl} \frac{g_{ik}^{(t+1)} . \lambda_{kjl}^{(t)}}{\sum_{k'}^{K} g_{ik'}^{(t)} . \lambda_{k'jl}^{(t)}}}{\sum_{l=1}^{L_j} \sum_{i=1}^{I} y_{ijl} \frac{g_{ik}^{(t+1)} . \lambda_{kjl}^{(t)}}{\sum_{k'}^{K} g_{ik'}^{(t)} . \lambda_{k'jl}^{(t)}}}$ |
| | Fuzzy k-partition | $J_m(\mu, \lambda) = \sum_{i=1}^{I} \sum_{k=1}^{K} \mu_{ik}^{m} \sum_{j=1}^{J} \sum_{l=1}^{l_j} ln \left( \lambda_{kjl} \right)^{y_{ijl}}$ | $\lambda_{kjl} = \frac{\sum_{i=1}^{I} \mu_{ik}^{m} \cdot y_{ijl}}{\sum_{i=1}^{I} \mu_{ik}^{m}}$<br>$\mu_{ik} = \left[ \sum_{s=1}^{K} \left( \frac{\sum_{j=1}^{J} \sum_{l=1}^{L_j} ln \left( \lambda_{kjl} \right)^{y_{ij}}}{\sum_{j=1}^{J} \sum_{l=1}^{L_j} ln \left( \lambda_{sjl} \right)^{y_{ij}}} \right)^{\frac{1}{m-1}} \right]^{-1}$ |

The techniques mentioned above are non-parametric. The algorithms employ the least total within-cluster matching dissimilarity-based dissimilarity functional where obtaining weak intra-similarity makes either accuracy or purity will be low [13]. The Fuzzy k-partition model was proposed by Yang et al. [11]. It is one of the parametric techniques for categorical data clustering based on the likelihood function of the multivariate multinomial distribution. Another parametric-based technique is the GoM model. Each object can be a limited membership of the ultimate profile.

Given the multinomial distribution, the categorical data may be regarded as trial independent. This is because categorical data is frequently observed as tallies coming from a predefined number of trials, where each trial entails selecting one option from a predetermined set of categories. As a result, the parametric technique works better with categorical data [6].

### B. Multinomial Distribution

The multinomial distribution is a broadening of the binomial distribution [14]. Let $N_i$ represent the quantity of outcomes in $i$ category in a series of standalone trials with probability $p_i$ for outcomes in $i$ category for each trial, , $1 \leq i \leq m$, where $\sum_{i=1}^{m} p_i = 1$. Any $m$-tuple of non-negative integers $(n_1, n_2, \ldots, n_m)$ is with sum $n$

$$P(N_1 = n_1, N_2 = n_2, \ldots, N_m = n_m) = \frac{n!}{n_1! n_2! \ldots n_m!} p_1^{n_1} p_2^{n_2} \ldots p_m^{n_m}. \quad (1)$$

The probability mass function for a multinomial distribution with the parameter parameter $a_k = (a_k^{jl}, l = 1, \ldots, m_j, j = 1, \ldots, p)$ is as follows:

$$f(x, a_k) = \prod_{j=1}^{p} \prod_{l=1}^{m_j} \left(a_k^{jl}\right)^{x^{jl}}, \quad (2)$$

where $\sum_{l=1}^{m_j} a_k^{jl} = 1$. $m_j$ categories make up the generic polytomous variable $j (j = 1, \ldots, p)$, where $m = \sum_{j=1}^{p} m_j$ denotes the total number of levels.

### C. Soft Set

Information system can be defined as a tuple $S = (U, A, V, f)$, where $U$ represents the universe of objects, $A$ be a set of variables or parameters, $V$ is a domain (values set) of variable $a \subset A$ where the information function is a total function as in equation (3)

A tuple $S = (U, A, V, f)$ can be used to represent an information system, with $U$ standing for the universe of objects, A for a set of variables or parameters, and V for the domain (set of values) of a $a \subset A$ variable, and the information function being a total function as in equation (3) such that $f(u, a) \in V_a$, $\forall_{(u,a) \in U \times A}$.

$$f: U \times A \to V. \quad (3)$$

**Definition 1.** Given $S = (U, A, V, f)$ as an information system. Suppose that $a \in A, V_a = \{0,1\}$, then $S$ is a bivalued information system, and can be defined as $S_{\{0,1\}}$.

As an information system, consider $S = (U, A, V, f)$. If an $a \in A, V_a = \{0,1\}$, then S is a bivalued information system and is defined as $S_{\{0,1\}}$.

$$S_{\{0,1\}} = (U, A, V_{\{0,1\}}, f). \quad (4)$$

Definitely, for each $u \in U, f(u, a) \in \{0,1\}$, for each $a_i \in A$ and $v \in V$, the map $a_i^v$ of $U$ is $a_i^v: U \to \{0,1\}$, such that

$$a_i^v = \begin{cases} 1 & f(u, a) = v \\ 0 & otherwise \end{cases}. \quad (5)$$

When handling uncertainty with the proper parametrization, an information system can be described as a soft set [25], [26]. Let $U$ be a universe set, $E$ be a set of parameters and $A \subset E$. According to equation (4), F is the function that maps the parameter A into the set of all subsets of the set $U$.

$$F: A \to P(U). \quad (6)$$

Therefore, the $(F, A)$ pair is named a soft set over $U$. $\forall_{a \in A}, F(a)$ is considered as the set of $a$-approximate elements of $(F, A)$.

When it comes to the notion of an information system, a soft set can be thought of as a specific category of information systems known as binary-valued information.

**Proposition 1**. All soft sets $(F, A)$ can be defined as $S_{\{0,1\}}$.

**Proof**: Let us say that the universe $U$ in $(F, E)$ can be represented by the set of universes $U$, where $E$ stands for the set of parameters and $A \subset E$. Next, the information system's function, $f$, is expressed as follows:

$$f = \begin{cases} 1, & u \in F(e) \\ 0, & u \notin F(e) \end{cases}. \quad (7)$$

For example, once $u_i \in F(e_j)$, $u_i \in U$ and $e_j \in E$, then $f(u_i, e_j) = 1$, then $f(u_i, e_j) = 0$. We have $V(h_i, e_j) = \{0,1\}$. Therefore, for $A \subset E$, $(F, A)$ can be represented as $(U, A, V_{\{0,1\}}, f)$. Thus, Definition 1 can be defined as $S_{\{0,1\}}$.

**Definition 2.** All value sets of a soft set $(F, E)$ fall within the value-class of the soft set indicated by $C_{(F,E)}$.

A Boolean-valued information system works with the "standard" soft set based on Proposition 1. For an information system's categorical value defined by $S = (U, A, V, f)$, with $V = \bigcup_{a \in A} V_a$ and $V_a$ state the attribute's domain, respectively. Multi-valued or categorical values are available in the domain $V_a$. It can be decomposed from $S$ into $|A|$ number of Boolean-valued information systems using the formula $S = (U, A, V_{\{0,1\}}, f)$. The decomposition of $A = \{a_1, a_2, \cdots, a_{|A|}\}$ into the disjoint-singleton attribute $\{a_1\}, \{a_2\}, \cdots, \{a_{|A|}\}$ is the basis of the decomposition of $S = (U, A, V, f)$.

**Definition 3.** According to Herawan and Deris [17], the following equation describes the relationship between a categorical-valued information system, $S = (U, A, V, f)$, and a Boolean-valued information system, $S = (U, a_i, V_{a_i}, f), i = 1, 2, \cdots, |A|$.

$$S = (U, A, V, f) = \begin{cases} S^1 = (U, a_1, V_{\{0,1\}}, f) \Leftrightarrow (F, a_1) \\ S^2 = (U, a_2, V_{\{0,1\}}, f) \Leftrightarrow (F, a_2) \\ \vdots \\ S^{|A|} = (U, a_{|A|}, V_{\{0,1\}}, f) \Leftrightarrow (F, a_{|A|}) \end{cases} = \left((F, a_1), (F, a_2), \cdots, (F, a_{|A|})\right). \quad (8)$$

Furthermore, a multi soft set over universe $U$ representing a categorical-valued information system $S=(U,A,V,f)$ is expressed as $(F,E)=\left((F,a_1),(F,a_2),\cdots,(F,a_{|A|})\right)$.

## III. PROPOSED APPROACH

As shown in Figure 1 below, the objective function of the clustering problem is created under the assumptions of a well-defined notion of similarity, or distance, between data objects and a method for determining whether a collection of objects is a homogenous cluster. A clear understanding of the similarity or distance between data elements forms the first presumption. The multinomial distribution function is used in this approach as the parametric measure because the distance measure for categorical data will only generate weak similarity. The second assumption is a method for determining if a collection of objects is a homogenous cluster.

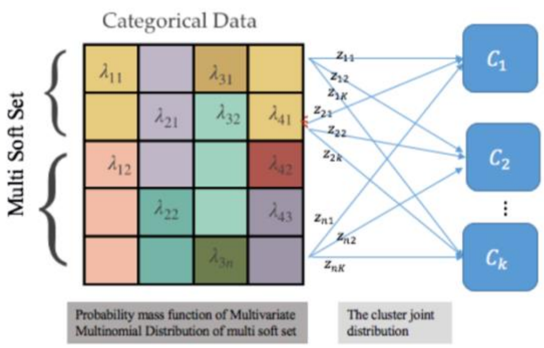


Figure 1. The illustration of Proposed Approach

Categorical data can be represented as an attribute-values system in a multi-valued information system. The multi-valued information system U shall be. The pair $(F,A)$, selected to multi-soft set over $U$ which represents a categorical-valued information system $S=(U,A,V,f)$, with $(F,a_1),\cdots,(F,a_{|A|})\subseteq(F,A)$ and $\left(F,a_{j_1}\right),\cdots,\left(F,a_{j_{|a_j|}}\right)\subseteq(F,a_j)$. Assume that $\lambda_{kjl}^{i}$represents the likelihood that $u_i\in(F,a_{jl})$ will enter the cluster $C_k, k=1,2,\dots,K$, with $i=1,2,\dots,|U|, j=1,2,\dots,|A|$ and $l=1,2,\dots,|a_j|$. Assume that $U$ is a sample of size $|U|$ drawn at random from the distribution $f(y,\lambda)$. An indicator $z_{ik}$divides a partition $U=\{u_1,u_2,\dots,u_{|U|}\}$ into a $K$ cluster $C=\{c_1,c_2,\dots,c_K\}$, where $z_{ik}=1$ if $u_i\in c_k$ and $z_{ik}=0$ otherwise. The cluster joint distribution function of $U$ based on cluster $C$ may then be written as $\prod_{k=1}^{K}\prod_{u_i\in c_k} z_{ik} f_k(y,\lambda)$. When the Multinomial Multivariate Distribution function of Multi Soft Set is substituted into the Cluster Joint Distribution function, the objective is to determine the highest probability ( $\lambda$ ). $L_{CML}(z,\lambda)$ can be used to specify the conditional objective function as a result.

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

Subject to

$$\sum_{k=1}^{K} z_{ik}=1, \mathrm{z_{ik}}\in\{0,1\}\, for\ i=1,2,\dots,|U|.$$

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

The objective function $L_{CML}(z,\lambda)$ is the constrained optimization problem. Using the Lagrange multiplier, the solution can be reduced to a single function of a constrained optimization problem. The necessary condition can then be implemented by obtaining the function's first derivative and setting it to 0. It is a set of nonlinear equations. Proposition 1 provides evidence that the model process's resolution.

**Proposition 1**: Let $(F,A)$ remain a soft set on $U$ corresponding to a categorical-valued information system with $(F,a_1),\cdots,(F,a_{|A|})\subseteq(F,A)$ and $\left(F,a_{j_1}\right),\cdots,\left(F,a_{j_{|a_j|}}\right)\subseteq(F,a_j)$. Suppose $(F,a_1),\cdots,(F,a_{|A|})\subseteq(F,A)$ and $\left(F,a_{j_1}\right),\cdots,\left(F,a_{j_{|a_j|}}\right)\subseteq(F,a_j)$ remain a multi soft set of $U$. So $z_{ik}$ and $\lambda_{kjl}$ are local maximum for $L_{CML}(z,\lambda)$ if merely if meets.

$$\lambda_{kjl}=\frac{\sum_{u_i\in(F,a_{j_l})} z_{ik}}{\sum_{l=1}^{|a_j|}\sum_{u_i\in(F,a_{j_l})} z_{ik}},\ \text{and}$$

$$z_{ik}=\begin{cases}1 & if\ \ \sum_{j=1}^{|A|}\ln(\lambda_{kjl}^{i}) = \max_{1\le k'\le K}\sum_{j=1}^{|A|}\ln(\lambda_{kjl}^{i}).\\ 0 & otherwise\end{cases}$$

## IV. EXPERIMENTAL RESULTS

The clustering results are evaluated based on benchmarks dataset taken from UCI machine learning as shown in Table 2.

TABLE II
THE UCI DATASETS

| No | Data set Name | #Attributes | #Instances |
|---|---|---|---|
| 1 | Zoo | 17 | 101 |
| 2 | Soybean | 35 | 47 |
| 3 | Balloons | 4 | 20 |
| 4 | Tic-tac-toe | 9 | 958 |
| 5 | Monk | 6 | 432 |
| 6 | Spect | 22 | 187 |
| 7 | Car | 6 | 1728 |

The processing time, rank index, and cluster purity are used to gauge performance. Table 3 provides a summary of the Rank index. The suggested strategy outperforms the standard methods in the data set from Zoo and Monk. For the soybean and balloon data sets, FkP has a superior rank index, whereas HC has a better one for the Spect data set. Concerning the Tic-Tac-Toe and automobile datasets, GoM has a higher rank index. Therefore, it may be argued that the proposed technique is "comparable" to standard techniques. Moreover, Table 4 displays the purity results. When compared to the HC, FC, GoM, and FkP, the technique outperforms the purity for five of the seven data sets (Zoo, Soybean, Tic-Tac-Toe, Monk, and Car).

TABLE III
RANK INDEX

| | Zoo | Soybean | Tic | Monk | Spect | Car | Ballon |
|---|---|---|---|---|---|---|---|
| HC | 75.21 | 64.79 | 51.06 | 50.02 | 53.30 | 50.40 | 56.46 |
| FC | 78.38 | 80.01 | 50.82 | 49.93 | 50.61 | 49.64 | 62.05 |
| GoM | 23.30 | 25.06 | 54.66 | 50.03 | 50.57 | 54.24 | 49.47 |
| FkP | 87.89 | 93.66 | 50.63 | 52.55 | 50.57 | 48.88 | 82.82 |
| SSC | 88.50 | 81.30 | 50.82 | 52.87 | 50.91 | 48.93 | 65.00 |

TABLE IV PURITY

| | Zoo | Soybean | Tic | Monk | Spect | Car | Ballon |
|---|---|---|---|---|---|---|---|
| HC | 0.174 | 0.328 | 0.649 | 0.531 | 0.943 | 0.704 | 0.713 |
| FC | 0.667 | 0.749 | 0.651 | 0.529 | 0.943 | 0.698 | 0.781 |
| GoM | 0.058 | 0.090 | 0.326 | 0.263 | 0.459 | 0.175 | 0.300 |
| FkP | 0.668 | 0.759 | 0.651 | 0.594 | 0.927 | 0.699 | 0.892 |
| SSC | 0.674 | 0.782 | 0.651 | 0.597 | 0.919 | 0.705 | 0.666 |

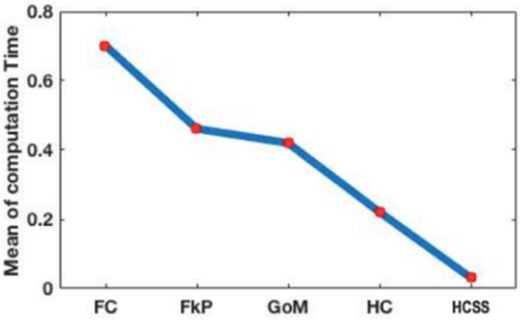


Figure 2. Time responses

Moreover, the proposed technique is superior in computational time in most data sets. In Figure 2, the average computation time is displayed. The technique effectively beats HC, FC, GoM, and FkP regarding calculation time for clustering issues. Specifically, processing the data set by HC, FC, GoM, and FkP takes around 0.2209, 0.7017, 0.4202, and 0.4614 seconds, respectively. In contrast, the execution time for the proposed method is only about 0.0309 seconds. As a result, the execution time was reduced on average by up to 0.4201 seconds. Therefore, the suggested technique outperformed HC, FC, GoM, and FkP in terms of purity, Rank Index, and computing time.

## V. CONCLUSION

Several strategies can solve the issue of centroid-based categorical data grouping. These methods do not, however, deliver purer clusters or quicker responses. This paper suggests a soft set-based clustering method. The data is clustered using a multivariate multinomial distribution after being deconstructed based on a soft set to produce a multi-soft set. A comparison of the suggested technique and the baseline algorithms has been conducted regarding purity, rank index, and response time. The outcomes demonstrate that the suggested approach outperforms the already-used approaches regarding faster response times (up to 92.96%), improved cluster stability, and Rank Index performance.

## ACKNOWLEDGMENT

This research is supported by the International Grant Scheme for Research from Ahmad Dahlan University No:04/RIA/LPPM-UAD/VI/2023. This support is gratefully acknowledged.

## REFERENCES

[1] G. M. Gonçalves and L. L. Lourenço, “Mathematical formulations for the K clusters with fixed cardinality problem,” Comput. Ind. Eng., vol. 135, pp. 593–600, 2019.

[2] G. J. McLachlan, S. I. Rathnayake, and S. X. Lee, “2.24 - Model-Based Clustering☆,” S. Brown, R. Tauler, and B. B. T.-C. C. (Second E. Walczak, Eds. Oxford: Elsevier, 2020, pp. 509–529.

[3] S. Zhu and L. Xu, “Many-objective fuzzy centroids clustering algorithm for categorical data,” Expert Syst. Appl., vol. 96, pp. 230–248, 2018.

[4] A. Saxena and M. Singh, “Using Categorical Attributes for Clustering,” Int. J. Sci. Eng. Appl. Sci., no. 2, pp. 324–329, 2016.

[5] T. P. Q. Nguyen and R. J. Kuo, “Partition-and-merge based fuzzy genetic clustering algorithm for categorical data,” Appl. Soft Comput. J., 2019.

[6] 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.

[7] C.-C. Hsu, C.-L. Chen, and Y.-W. Su, “Hierarchical clustering of mixed data based on distance hierarchy,” Inf. Sci. (Ny)., vol. 177, no. 20, pp. 4474–4492, 2007.

[8] Z. He, S. Deng, and X. Xu, “Improving K-Modes Algorithm Considering Frequencies of Attribute Values in Mode BT - Computational Intelligence and Security,” 2005, pp. 157–162.

[9] M. K. Ng, M. J. Li, J. Z. Huang, and Z. He, “On the Impact of Dissimilarity Measure in k-Modes Clustering Algorithm,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 3, pp. 503–507, 2007.

[10] 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.

[11] 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.

[12] 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.

[13] I. T. R. Yanto, M. A. Ismail, and T. Herawan, “A modified Fuzzy k-Partition based on indiscernibility relation for categorical data clustering,” Eng. Appl. Artif. Intell., vol. 53, pp. 41–52, 2016.

[14] S. Malefaki and G. Iliopoulos, “Simulating from a multinomial distribution with large number of categories,” Comput. Stat. Data Anal., vol. 51, no. 12, pp. 5471–5476, 2007.

[15] D. Molodtsov, “Soft set theory—first results,” Comput. Math. with Appl., vol. 37, no. 4–5, pp. 19–31, 1999.

[16] P. K. Maji, R. Biswas, and A. R. Roy, “Soft set theory,” Comput. Math. with Appl., vol. 45, no. 4–5, pp. 555–562, 2003.

[17] 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.