

# 21% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.





## Filtered from the Report

- ▶ Bibliography




## Exclusions

- ▶ 1 Excluded Source
- ▶ 14 Excluded Matches

## Match Groups

-  **49 Not Cited or Quoted 20%**  
Matches with neither in-text citation nor quotation marks
-  **2 Missing Quotations 0%**  
Matches that are still very similar to source material
-  **0 Missing Citation 0%**  
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**  
Matches with in-text citation present, but no quotation marks

## Top Sources

- 9%  Internet sources
- 18%  Publications
- 2%  Submitted works (Student Papers)

## Integrity Flags

### 0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

### Match Groups

- **49 Not Cited or Quoted 20%**  
Matches with neither in-text citation nor quotation marks
- **2 Missing Quotations 0%**  
Matches that are still very similar to source material
- **0 Missing Citation 0%**  
Matches that have quotation marks, but no in-text citation
- **0 Cited and Quoted 0%**  
Matches with in-text citation present, but no quotation marks

### Top Sources

- 9% Internet sources
- 18% Publications
- 2% Submitted works (Student Papers)

### Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Publication		
		Iwan Tri Riyadi Yanto, Cheah WaiShiang, Rahmat Hidayat, Rofiul Wahyudi, Suprih...	9%
2	Publication		
		Iwan Tri Riyadi Yanto, Mustafa Mat Deris, Norhalina Senan. "Chapter 2 PSS: New ...	2%
3	Internet		
		www.joiv.org	2%
4	Publication		
		Iwan Tri Riyadi Yanto, Ririn Setiyowati, Mustafa Mat Deris, Norhalina Senan. "Cha...	1%
5	Student papers		
		University of Malaya	1%
6	Internet		
		spectrum.library.concordia.ca	1%
7	Internet		
		insightsociety.org	1%
8	Publication		
		Miin-Shen Yang, Yu-Hsuan Chiang, Chiu-Chi Chen, Chien-Yo Lai. "A fuzzy k-partitio...	1%
9	Internet		
		xmmssc.irap.omp.eu	1%
10	Internet		
		www.actapress.com	1%

11	Publication	Ani Apriani, Iwan Tri Riyadi Yanto. "Clustering Coping Capacity Using HCSS", 2023 ...	0%
12	Internet	www.math.nagoya-u.ac.jp	0%
13	Publication	Yanto, Iwan Tri Riyadi, Maizatul Akmar Ismail, and Tutut Herawan. "A modified F...	0%
14	Internet	ro.ecu.edu.au	0%
15	Internet	lambda.gsfc.nasa.gov	0%
16	Internet	www.bayes.city.ac.uk	0%
17	Internet	www.coursehero.com	0%
18	Publication	Chien-Yo Lai, Miin-Shen Yang. "Entropy-type classification maximum likelihood al...	0%
19	Publication	Qinrong Feng, Fenfen Wang. "A discernibility matrix approach to fuzzy soft sets b...	0%

# Fuzzy Soft Set Clustering for Categorical Data

Iwan Tri Riyadi Yanto<sup>a</sup>, Ani Apriani<sup>b</sup>, Rofiul Wahyudi<sup>c</sup>, Cheah WaiShiang<sup>d</sup>, Suprihatin<sup>a</sup>, Rahmat Hidayat<sup>e</sup>

<sup>a</sup>Information System Department, Universitas Ahmad Dahlan, Yogyakarta, Indonesia

<sup>b</sup>Institute Teknologi Nasional Yogyakarta, Yogyakarta, Indonesia

<sup>c</sup>Islamic Banking Department, Universitas Ahmad Dahlan, Yogyakarta, Indonesia

<sup>d</sup>Faculty of Computer Science & IT, Universiti Malaysia Sarawak, Malaysia

<sup>e</sup>Information Technology Department, Politeknik Negeri Padang, Padang, Indonesia

Corresponding author: [yanto.itr@is.uad.ac.id](mailto:yanto.itr@is.uad.ac.id)

**Abstract**— Categorical data clustering is difficult because categorical data lacks natural order and can comprise groups of data only related to specific dimensions. Conventional clustering, such as k-means, cannot be openly used to categorical data. Numerous categorical data using clustering algorithms, for instance, fuzzy k-modes and their enhancements, have been developed to overcome this issue. However, these approaches continue to create clusters with low Purity and weak intra-similarity. Furthermore, transforming category attributes to binary values might be computationally costly. This research provides a categorical data with fuzzy clustering technique due to soft set theory and multinomial distribution. The experiment showed that the approach proposed signifies better performance in purity, rank index, and response times by up to 97.53%.

**Keywords**— Function of multinomial distribution, clustering, categorial data, multi soft-set.

Manuscript received 15 Oct. 2020; revised 29 Jan. 2021; accepted 2 Feb. 2021. Date of publication 17 Feb. 2021.

International Journal on Informatics Visualization is licensed under a Creative Commons Attribution-Share Alike 4.0 International License.



## I. INTRODUCTION

Data clustering is categorizing data according to their similarity, which aims to create similar or difference data category [1], [2]. It is a partition of a given data set from multiple variables into groups, which is a vital step in exploratory data mining. It is useful for revealing the data becomes natural structure. Clustering has been used in various fields, including earth science, life science, social sciences, information sciences, medical sciences, policy, and decision-making. It is also applicable to the preliminary stages in other studies, such as bioinformatics, collaborative filtering, customer breakdown, data exploration and summarization, dynamic trend detection, information retrieval, market analysis, medical diagnostics, and text mining as well as analysis on multimedia, social network, and web [3]–[5]. Clustering can be in the form of hard and fuzzy, depending on how they assign data points to clusters. Each data point is assigned to exactly one cluster in hard clustering, but multiple clusters in fuzzy ones [6]. Fuzzy clustering is often natural and effective, especially when the data is not separable into distinct clusters [7].

Categorical data differs from numeric data in that it organizes it into categories rather than numerical values. Categorical data is normally applied in real-world, for instance,

medical data and retail purchase transactions. Categorical factors, for example, nationality, gender, occupation, level of education, marital status, and smoking status, for example, are included in medical data. Product classifications, consumer types, and locales play a role in retail purchase transactions. [8], [9]. Due to the absence of natural order, the possibility of subspace clusters, and the conversion of categorical to numeric data, data with categorical features pose certain hurdles to existing clustering methods.

Categorical data clustering algorithms have been developed and proposed with k-modes clustering approach that overcomes the k-means algorithm's numerical-only constraint [10]. Data clustering have been developed with new dissimilarity measures to the k-mode clustering [11]–[13] and a fuzzy set-based k-mode algorithm [14], [15]. Kim et al. [16] suggested to improve the efficiency of fuzzy k-modes with the fuzzy centroids approach. Another fuzzy approach for grouping document data based on a new construction of category data has been developed by Umayahara et al. [17]. Non-parametric techniques due to the least sum of squared errors within clusters are used in categorical data clustering and its variants [14], [18], [19]. This choice involves, in essence, the expectation that data will be structured into spherical clusters and that either precision or purity will be reached [20]–[23].

1 Miin-Shen et al. [22] presented a parametric technique based on the likelihood function of multivariate multinomial distributions with the Fuzzy k-partition (FkP) algorithm. FkP enhances the Grade of Membership (GoM) model for categorical data analysis [24]. However, almost all the data clustering techniques represent data sets in binary values. 2 Furthermore, the maximum parameter of the classification likelihood function in the same categories has the same probability value in the FkP method [25]. Although the GoM and FkP models are useful for categorical data clustering, the algorithms include sophisticated iteration computations that take a long time to complete. This implies the significance of techniques that without high calculation times and low clusters purity. 1

The converted values are arbitrary and appear to serve no use other than as a convenient label for a specific value. The reason for this is that each value in a categorical characteristic reflects a separate logical concept and, as such, cannot be meaningfully ordered or manipulated in the same way that numbers can [26]. In probability theory and statistics, categorical data is likely to follow the function of multivariate multinomial-distribution randomly. Categorical data, in contrast, has multi-valued attributes that represent a multi-soft set [27]. Using a multi-soft set for multi-valued attributes has the advantage of capturing categorical data without the necessity for conversion into binary values [28]. Therefore, this study proposes a new fuzzy clustering method based on multi soft sets.

## 19 2 II. RELATED WORKS

### 5 1 A. Information system

Let tuple  $S = (U, A, V, f)$ , where universe  $U$  is signified,  $A$  represents parameters,  $V$  embraces a value set of variable  $a \in A$ . Thus, function of information comprises overall function as equation (1) shows, for instance,  $f(u, a) \in V_a, \forall (u, a) \in U \times A$ .

$$1 f: U \times A \rightarrow V. \quad (1)$$

5 **Definition 1.** Assumed  $S = (U, A, V, f)$  as system of information. Assume that  $a \in A, V_a = \{0, 1\}$ , so,  $S$  contains a system of bivalued-information. Thus, the definition is  $S_{\{0,1\}}$ .

$$1 S_{\{0,1\}} = (U, A, V_{\{0,1\}}, f). \quad (2)$$

1 Definitely, representing each  $u \in U, f(u, a) \in \{0, 1\}$ , representing each  $a_i \in A$  and  $v \in V$ , the map  $a_i^v$  of  $U$  is  $a_i^v: U \rightarrow \{0, 1\}$ , as shown in equation below.

$$1 a_i^v = \begin{cases} 1 & f(u, a) = v \\ 0 & \text{otherwise} \end{cases}. \quad (3)$$

### 4 B. Theory of Soft-set

19 A soft-set encompasses a mathematical method to sort out the ambiguity through proper parametrization [25], [26]. Let  $U$  be a set of universe,  $E$  be a parameter set and  $A \subset E, F$  be works for charts  $A$  parameter into a set of completely subsets of set  $U$  as equation (4) presents.

$$4 F: A \rightarrow P(U). \quad (4)$$

Next, a  $(F, A)$  pair is labelled a soft-set on  $U. \forall a \in A, F(a)$  be measured as  $a$  estimated parts set of  $(F, A)$ .

Considering the information system designation, a soft-set can be understood as an information-system exceptional category, which is called binary-valued information.

**Proposition 1.** For each soft-set  $(F, A)$  could be classified as  $S_{\{0,1\}}$ .

**Proof:** Let a universe set  $U$  in  $(F, E)$  be counted as universe  $U$ , the parameter set represented by  $E$  where  $A \subset E$ . Then, information system function  $f$  is shown in following equation:

$$f = \begin{cases} 1, & u \in F(e) \\ 0, & u \notin F(e) \end{cases}. \quad (5)$$

For example, when  $u_i \in F(e_j)$ , where  $u_i \in U$  and  $e_j \in E$ , then  $f(u_i, e_j) = 1$ , then  $f(u_i, e_j) = 0$ . Thus, we have  $V(h_i, e_j) = \{0, 1\}$ . Hence, for  $A \subset E, (F, A)$  can be signified as  $(U, A, V_{\{0,1\}}, f)$ . So, based on Definition 1, it can be specified as  $S_{\{0,1\}}$ .

**Definition 2.** The soft-set value-class is represented by  $C_{(F,E)}$  are all value soft-set class  $(F, E)$ .

In proposition 1, it shows the Boolean-valued information system on the "standard" soft set. Representing an information-system categorical value of represented by  $S = (U, A, V, f)$  with  $V = \cup_{a \in A} V_a$  and  $V_a$  affirms the attribute  $a$  domain.  $V_a$  domain has multi-values. A breakdown can be structured from  $S$  into  $|A|$  number of Boolean-valued information system  $S = (U, A, V_{\{0,1\}}, f)$ ,  $A = \{a_1, a_2, \dots, a_{|A|}\}$  into the split-isolated attribute  $\{a_1\}, \{a_2\}, \dots, \{a_{|A|}\}$  due to  $S = (U, A, V, f)$ .

**Definition 3.** [31] Consider  $S = (U, A, V, f)$  is a system of categorical-valued information and a Boolean-valued information is denoted by  $S = (U, a_i, V_{a_i}, f), i = 1, 2, \dots, |A|$  in addition to

$$S = (U, A, V, f) = \begin{cases} S^1 = (U, a_1, V_{\{0,1\}}, f) \Leftrightarrow (F, a_1) \\ S^2 = (U, a_2, V_{\{0,1\}}, f) \Leftrightarrow (F, a_2) \\ \vdots \\ S^{|A|} = (U, a_{|A|}, V_{\{0,1\}}, f) \Leftrightarrow (F, a_{|A|}) \end{cases} = ((F, a_1), (F, a_2), \dots, (F, a_{|A|})). \quad (6)$$

Also, a multi-soft set on universe  $U$  represent a system of categorical-valued information  $S = (U, A, V, f)$ , which is represented as  $(F, E) = ((F, a_1), (F, a_2), \dots, (F, a_{|A|}))$ .

### C. Multinomial Distribution

A binomial distribution generalization comprises the multinomial distribution [32]. Let  $N_i$  be the number of category  $i$  in a individual experiment series using probability  $p_i$  for respectively experiment, where,  $1 \leq i \leq m, \sum_{i=1}^m p_i = 1$ . So, every  $m$ -tuple of non-negative integers  $(n_1, n_2, \dots, n_m)$  with sum  $n$ .

$$P(N_1 = n_1, N_2 = n_2, \dots, N_m = n_m) = \frac{n!}{n_1! n_2! \dots n_m!} p_1^{n_1} p_2^{n_2} \dots p_m^{n_m}. \quad (7)$$

**Example 1.** Assume ten balls in a basket entails some balls with two in red, three in green, and five in blue color. Four balls will be chosen from the basket with substitution. So, the probability of drawing two green and blue balls, respectively are as follows:

$$P(n_1 = 0, n_2 = 2, n_3 = 2) = \frac{4!}{0! 2! 2!} 0.2^0 0.3^2 0.5^2 = 0.135.$$

A multinomial distribution with parameter  $a_k = (a_k^{jl}, l = 1, \dots, m_j, j = 1, \dots, p)$  could be called as the probability mass function as shown in equation below:

$$f(x, a_k) = \prod_{j=1}^p \prod_{l=1}^{m_j} (a_k^{jl})^{x^{jl}}, \quad (8)$$

where  $\sum_{l=1}^{m_j} a_k^{jl} = 1$ . Variable of standard polytomous  $j(j = 1, \dots, p)$  have  $m_j$  categories, and  $m = \sum_{j=1}^p m_j$  denotes the total levels number.

### III. PROPOSED METHOD

#### A. Model Objective function

The categorical data clustering objective function and constraints are constructed using a function of multinomial distribution due to soft-set. The hypothesis of the function is how to find the weight of the object to be given to high probability cluster. The function of cluster joint distribution defines general model first by assuming that the data follows a certain function of distribution. The cluster intersection distribution function is supposed in Definition 4.

**Definition 4.** Suppose  $U$  includes a unsystematic sample-size  $|U|$  since division  $f(y, \lambda)$ . Partition  $U = \{u_1, u_2, \dots, u_{|U|}\}$  into  $K$  cluster  $C = \{c_1, c_2, \dots, c_K\}$  through value  $z_{ik}$  where  $z_{ik} = 1$  if  $u_i \in c_k$  and  $z_{ik} = 0$  if else. The function of cluster shared distribution of  $U$  due to cluster  $C$  could be described as  $\prod_{k=1}^K \prod_{u_i \in c_k} z_{ik} f_k(y, \lambda)$ .

Representing the data as multi soft set, assuming that the categorical data has multi-valued attributes following a multivariate multinomial distribution function can be defined as a **Multivariate Multinomial Distribution Function of Soft set** as in Definition 5.

**Definition 5.** Suppose  $(F, A)$  is a multi-soft set concluded  $U$  signifies a system of categorical-valued information  $S = (U, A, V, f)$ , with  $(F, a_1), \dots, (F, a_{|A|}) \subseteq (F, A)$  and  $(F, a_{j_1}), \dots, (F, a_{j_{|a_j|}}) \subseteq (F, a_j)$ . lets  $\lambda_{kjl}^i$  is a probability of  $u_i \in (F, a_{jl})$  into cluster  $C_k, k = 1, 2, \dots, K, i = 1, 2, \dots, |U|, j = 1, 2, \dots, |A|$  and  $l = 1, 2, \dots, |a_j|$ ; hence, the function of the multivariate multinomial distribution soft-set can be written as follows.

$$f_k(y, \lambda) = \prod_{j=1}^{|A|} \prod_{l=1}^{|a_j|} (\lambda_{kjl}^i)^{|F, a_{jl}|}, \text{ where } \sum_{l=1}^{|a_j|} \lambda_{kjl} = 1, \forall k, j.$$

From definition 5, function of multinomial distribution is substituted into function of cluster shared distribution in definition 4. So, it is obtained as a conditional maximum likelihood function.

$$CML(z, \lambda) = \prod_{k=1}^K \prod_{i=1}^{|U|} z_{ik} \prod_{j=1}^{|A|} \prod_{l=1}^{|a_j|} (\lambda_{kjl}^i)^{|F, a_{jl}|}. \quad (9)$$

where  $\sum_{l=1}^{|a_j|} \lambda_{kjl} = 1$  and  $\sum_{k=1}^K z_{ik} = 1, z_{ik} \in \{0, 1\}$  for  $i = 1, 2, \dots, |U|$ .

Consider the extension to allow the indicator functions  $z_{ik} = z_k(y_i)$  to be functions  $\mu_{ik} = \mu_k(y_i)$  assuming values in the interval  $[0, 1]$  such that  $\sum_{k=1}^K \mu_{ik} = 1$  for all  $i = 1, \dots, |U|$ . In this case  $\mu = \{\mu_1, \mu_2, \dots, \mu_K\}$  is called a Fuzzy partition of  $U$  that had been used for fuzzy clustering. Now, the CML procedure can be extended to be likelihood CML as in (27).

$$\begin{aligned} \text{Maximize } L_{CML}(\mu, \lambda) &= \sum_{k=1}^K \sum_{i=1}^{|U|} \mu_{ik} \prod_{j=1}^{|A|} \prod_{l=1}^{|a_j|} (\lambda_{kjl}^i)^{|F, a_{jl}|} \\ &= \sum_{k=1}^K \sum_{i=1}^{|U|} \mu_{ik} \sum_{j=1}^{|A|} \sum_{l=1}^{|a_j|} \ln(\lambda_{kjl}^i)^{|F, a_{jl}|}. \end{aligned} \quad (10)$$

Subject to

$$\sum_{k=1}^K \mu_{ik} = 1, \mu_{ik} \in [0, 1] \text{ for } i = 1, 2, \dots, |U|.$$

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

The solution of the objective function (27) can be obtained by changing into the unconstrained problem by adding lagrange multiplier i.e  $w_1, w_2$ . The Lagrangian of  $L_{CML}$  should be as equation below.

$$\begin{aligned} L_{CML}(\mu, \lambda, w_1, w_2) &= \sum_{i=1}^{|U|} \sum_{k=1}^K \mu_{ik}^m \sum_{j=1}^{|A|} \sum_{l=1}^{|a_j|} \ln(\lambda_{kjl}^i)^{|F, a_{jl}|} \\ &\quad - w_1 \left( \sum_{k=1}^K \mu_{ik}^m - 1 \right) - w_2 \left( \sum_{l=1}^{|a_j|} \lambda_{kjl} - 1 \right) \end{aligned} \quad (11)$$

The first derivative of the lagrangian  $L_{CML}$  is taken regarding the  $z_{ik}, \lambda_{kjl}, w_1, w_2$  and set to 0 with the following equation.

$$\frac{\partial L_{CML}}{\partial \mu_{ik}} = m \mu_{ik}^{m-1} \sum_{j=1}^{|A|} \sum_{l=1}^{|a_j|} \ln(\lambda_{kjl}^i)^{|F, a_{jl}|} - w_1 = 0,$$

$$\frac{\partial L_{CML}}{\partial \lambda_{kjl}} = \frac{\sum_{i=1}^{|U|} \mu_{ik}^m |F, a_{jl}|}{\lambda_{kjl}} - w_2 = 0,$$

$$\frac{\partial L_{CML}}{\partial w_1} = - \left( \sum_{k=1}^K \mu_{ik}^m - 1 \right) = 0,$$

$$\frac{\partial L_{CML}}{\partial w_2} = - \left( \sum_{l=1}^{|a_j|} \lambda_{kjl} - 1 \right) = 0.$$



We can obtain  $w_1$  and  $w_2$  from (18) to (21) and then substitute them back into Eqs. (18) and (19). Thus, the solution are obtained as follows:

$$\lambda_{kjl} = \frac{\sum_{u_i \in (F, a_{jl})} z_{ik}}{\sum_{l=1}^{|a_j|} \sum_{u_i \in (F, a_{jl})} z_{ik}} \quad (12)$$

$$\mu_{ik} = \left[ \sum_{s=1}^K \left[ \frac{\sum_{j=1}^{|A|} \sum_{l=1}^{|a_j|} \ln(\lambda_{kjl}^i)^{|F, a_{jl}|}}{\sum_{j=1}^{|A|} \sum_{l=1}^{|a_j|} \ln(\lambda_{sjl}^i)^{|F, a_{jl}|}} \right]^{m-1} \right]^{-1}$$

The algorithm of the proposed technique can be described in Figure 1. The algorithm starts by decomposing the data into multi-soft sets and computing the initial membership randomly. Then, the probability and membership are updated until one of the three stopping criteria is fulfilled. The three stopping criteria are cost function has been convergent ( $|L_{cml}^{it}(z, \lambda) - L_{cml}^{it-1}(z, \lambda)| < \epsilon_1$ ), membership function has been convergent ( $\|z_{ik}^{it} - z_{ik}^{it-1}\| < \epsilon_2$ ) and the iteration has reached to maximum given iteration (M), where  $\epsilon_1, \epsilon_2$  are the small positive tolerance.

Input: Categorical data set, tolerance given ( $\epsilon_1, \epsilon_2$ ), number of iterations  $M$

Output: Clusters

Begin

Spoil the data into the multi-soft set.

Run the random initial  $z_{ik}$

Update  $\lambda_{kjl}$

Update  $\mu_{ik}$

Repeat 3 and 4 until ( $|L_{cml}^{it}(z, \lambda) - L_{cml}^{it-1}(z, \lambda)| < \epsilon_1$  or  $\|z_{ik}^{it} - z_{ik}^{it-1}\| < \epsilon_2$ ) or iteration =M.

End

Figure 1. Soft-set on Function of Multinomial Distribution for Fuzzy Soft Set Algorithm

### B. Experimental Results

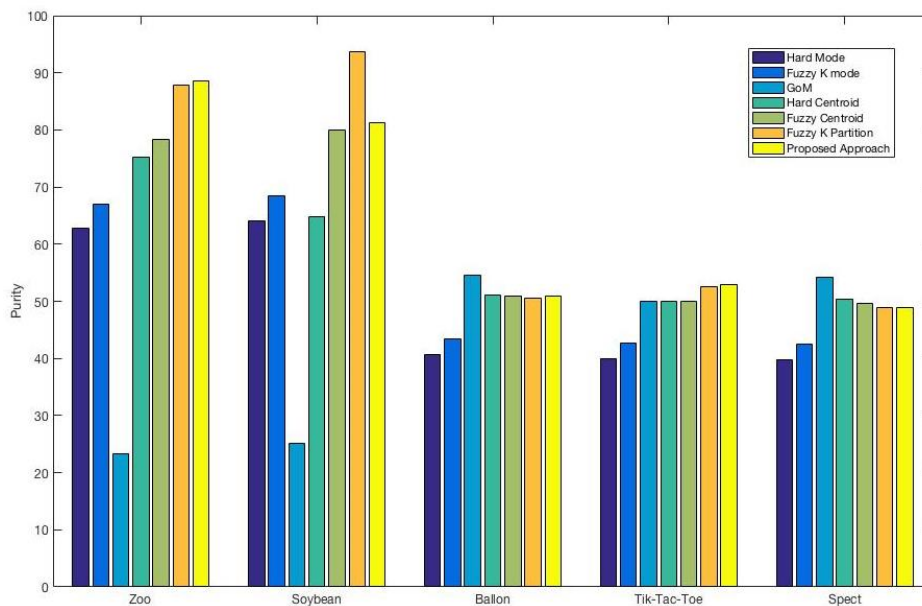


Figure 2. The results of the Cluster Purity experiment

MATLAB 9.0.0.341360 (R2016a) version was used in examining the working of cluster purity, rank index, and estimation time of the Proposed Method and other categorical data clustering. The experiment was conducted chronologically using an Intel Core i5 computer, 8GB Memory, and supported by Mac Operating System. Seven categorical dataset from UCI Machine Learning Repository was experimented [33], i.e. Zoo, Spect, Soybean, Tic-tac-toe, and Balloons as shown in Table 1 below. Table 1 defines the dataset name, attributes, and instances.

Table 1. Dataset Tested for Experiments Purpose

No	Dataset	#Attributes	#Instances
1	Zoo	18	101
2	Soybean	35	47
3	Balloons	4	20
4	Tic-tac-toe	9	958
5	Spect	922	187

Figure 2 explains the comparison results of the seven algorithms in terms of cluster purity implemented on the five datasets used. Based on Figure 2, it can be shown that the proposed technique can be said “comparable” to baseline techniques. Among the five data set, the proposed technique surpasses the Purity for five data sets (Zoo, Soybean, Tic-tac-toe, Monk, and Car) than the Hard Mode, Fuzzy K mode, Hard Centroid, Fuzzy Centroid, and GoM, respectively. Although, Fuzzy K partition has better purity on the soybean dataset, and GoM has better purity on the Ballon and Spect datasets. However, the proposed technique outperforms the baseline techniques in almost all datasets used. Moreover, Figure 3 shows the rank index results. It is illustrated that the proposed technique outperformed the baseline techniques while implementing the clustering problems. It can show that the proposed approach's rank index value achieves the highest value compared to the baseline techniques.

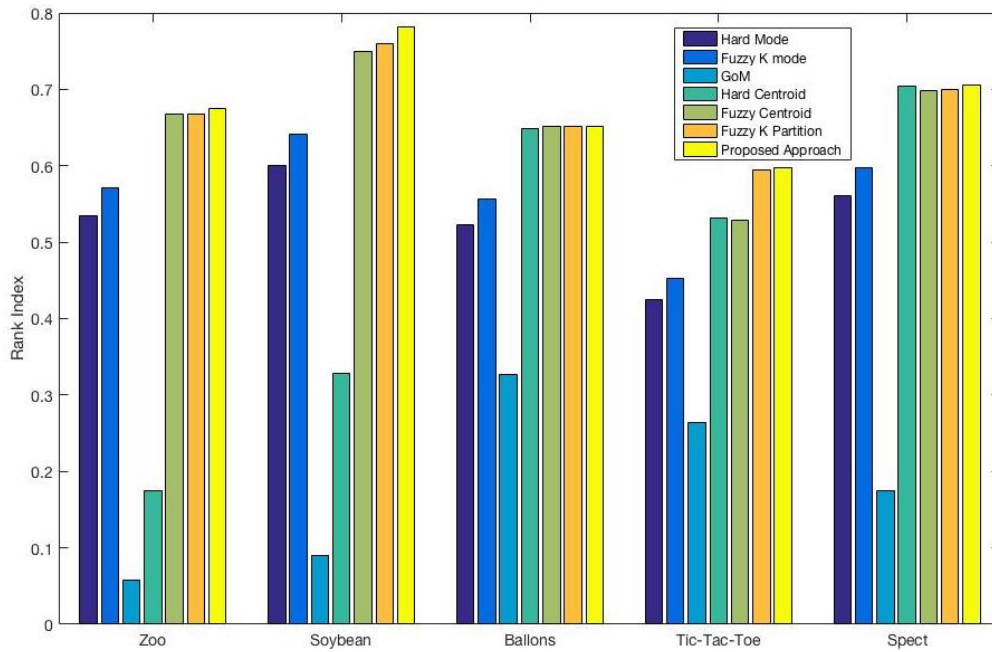


Figure 3. The results of The Rand Index experiment

Table 2. The results of comparison regarding time responses

Data Set	Hard Mode	Fuzzy K Mode	GOM	Hard Centroid	Fuzzy-Centroid	Fuzzy-K partition	Proposed Approach
Zoo	0.0388	0.0388	0.1193	0.4508	0.8732	0.2617	0.0236
Soybean	0.0327	0.0327	0.1425	0.6198	0.6534	0.7101	0.0199
Tic-tac-toe	0.0174	0.0174	0.3967	0.5333	0.5034	0.1682	0.0106
Monk	0.0416	0.0416	0.1018	0.1627	0.9206	0.3754	0.0253
Car	0.0176	0.0176	0.6279	0.8259	0.7037	0.0099	0.0107
Average	0.0296	0.0296	0.2776	0.5185	0.7309	0.3051	0.0180
Improvement	39.20%	39.20%	93.51%	96.52%	97.53%	94.09%	

Table 2 explains the results of comparison among the three algorithms regarding computational time implemented on the five datasets used. It indicates that the proposed technique successfully overcomes baseline techniques in terms of computation time for clustering problems. In detail, Hard Mode, Fuzzy K Mode, GOM, Hard Centroid, Fuzzy Centroid, and Fuzzy K partition consume approximately 0.0296, 0.0296, 0.2776, 0.5185, 0.7309, 0.3051 seconds of execution time of dataset processing in average, respectively. On the other hand, the proposed technique demands merely almost 0.0180 seconds of execution time on average. Thus, it indicates an average decrease of execution time of up to 97.53%. Therefore, the proposed technique is superior in computational time in most data sets. Meanwhile, the proposed technique worked better than baseline techniques regarding Purity, Rank Index, and computation time, respectively.

#### IV. CONCLUSION

Several algorithms can solve the challenge of fuzzy-based categorical data grouping. These techniques, however, do not give improved cluster purity or faster reaction times. As a result, hard categorical data clustering through multinomial distribution is suggested. To produce a multi-soft set, the data is rotted based soft set, and the data is clustered using a multivariate multinomial distribution. A comparison of the new technique and the baseline algorithms reveals that the suggested approach overtakes the current approaches regarding purity, rank index, and response times by up to 97.53%

#### 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] K. Soppari and N. S. Chandra, "Development of improved whale optimization-based FCM clustering for image watermarking," *Comput. Sci. Rev.*, vol. 37, p. 100287, 2020.
- [4] C. Wu and X. Zhang, "Total Bregman divergence-based fuzzy local information C-means clustering for robust image segmentation," *Appl. Soft Comput.*, vol. 94, p. 106468, 2020.
- [5] M. C. Thrun and Q. Stier, "Fundamental clustering algorithms suite," *SoftwareX*, vol. 13, p. 100642, 2021.
- [6] Y. Karali, B. Lyngdoh, and H. Behera, "Hard and Fuzzy Clustering Algorithms Using Normal Distribution of Data Points: a Comparative Performance Analysis," vol. 2, no. 10, pp. 320–328, 2013.
- [7] K. Mrudula and E. K. Reddy, "Hard And Fuzzy Clustering Methods : A Comparative Study Hard and Fuzzy Clustering Methods : A Comparative Study," no. April, 2019.
- [8] S. Zhu and L. Xu, "Many-objective fuzzy centroids clustering algorithm for categorical data," *Expert Syst. Appl.*, vol. 96, pp. 230–248, 2018.
- [9] D. T. Dinh, V. N. Huynh, and S. Sriboonchitta, "Clustering mixed numerical and categorical data with missing values," *Inf. Sci. (Ny)*, vol. 571, pp. 418–442, 2021.
- [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] 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.
- [12] 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.
- [13] O. M. San, H. Van-Nam, and Y. Nakamori, "An alternative extension of the k-means algorithm for clustering categorical data," *Int. J. Appl. ...*, vol. 14, no. 2, pp. 241–247, 2004.
- [14] 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.
- [15] M. W. M. Wei, H. X. H. Xuedong, C. Z. C. Zhibo, Z. H. Z. Haiyan, and W. C. W. Chunling, "Multi-Agent Reinforcement Learning Based on Bidding," *Inf. Sci. Eng. (ICISE), 2009 1st Int. Conf.*, vol. 20, no. 3, 2009.
- [16] 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.
- [17] Y. N. K. Umayahara, S. Miyamoto, "FORMULATIONS OF FUZZY CLUSTERING FOR CATEGORICAL DATA Kazutaka Umayahara," *Inf. Control*, vol. 1, no. 1, pp. 83–94, 2005.
- [18] D. Parmar, T. Wu, and J. Blackhurst, "MMR: An algorithm for clustering categorical data using Rough Set Theory," *Data Knowl. Eng.*, vol. 63, no. 3, pp. 879–893, Dec. 2007.
- [19] S. Wu, A. W.-C. Liew, H. Yan, and M. Yang, "Cluster Analysis of Gene Expression Data Based on Self-Splitting and Merging Competitive Learning," *IEEE Trans. Inf. Technol. Biomed.*, vol. 8, no. 1, pp. 5–15, Mar. 2004.
- [20] 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.
- [21] P. BRYANT and J. A. WILLIAMSON, "Asymptotic behaviour of classification maximum likelihood estimates," *Biometrika*, vol. 65, no. 2, pp. 273–281, Aug. 1978.
- [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. P. Chatzis, "A fuzzy c-means-type algorithm for clustering of data with mixed numeric and categorical attributes employing a probabilistic dissimilarity functional," *Expert Syst. Appl.*, vol. 38, no. 7, pp. 8684–8689, Jul. 2011.
- [24] M. A. Woodbury and J. Clive, "Clinical Pure Types as a Fuzzy Partition," *J. Cybern.*, vol. 4, no. 3, pp. 111–121, Jan. 1974.
- [25] S. Naouali, S. Ben Salem, and Z. Chtourou, *Clustering categorical data: A survey*, vol. 19, no. 1. 2020.
- [26] A. Saxena and M. Singh, "Using Categorical Attributes for Clustering," *Int. J. Sci. Eng. Appl. Sci.*, no. 2, pp. 324–329, 2016.
- [27] B. Pardasani, "Multi Softset for Decision Making," *Int. J. Sci. Res.*, vol. 7, no. 11, pp. 55–56, 2018.
- [28] M. S. Khan, G. Mujtaba, M. A. Al-garadi, N. H. Friday, A. Waqas, and F. R. Qasmi, "Multi-soft sets-based decision making using rank and fix valued attributes," in *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, 2018, pp. 1–11.
- [29] D. Molodtsov, "Soft set theory—first results," *Comput. Math. with Appl.*, vol. 37, no. 4–5, pp. 19–31, 1999.
- [30] 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.
- [31] 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.
- [32] 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.
- [33] D. Dheeru and E. Karra Taniskidou, "{UCI} Machine Learning Repository." 2017.