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 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%. There are many algorithms that can be used to solve the challenge of grouping fuzzy-based categorical data. However, these techniques do not always result in improved cluster purity or faster reaction times. As a solution, it is suggested to use hard categorical data clustering through multinomial distribution. This involves producing a multi-soft set by using a rotated based soft set, and then clustering the data using a multivariate multinomial distribution. The comparison of this innovative technique with the established baseline algorithms demonstrates that the suggested approach excels in terms of purity, rank index, and response times, achieving improvements of up to ninety-seven-point fifty three percent compared to existing methods.

Keywords—Function of multinomial distribution; clustering; categorical data; multi soft set.

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 diagnostis, 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 the 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].

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. 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.

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, in contrast, has multi-valued attributes that represent a multi-soft distribution randomly. Categorical data, in numbers can 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:

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

(5)

For example, when $u_i \in F(e)_j$, where $u_i \subset U$ and $e_j \in E$, then $f(u_i, e_j) = 1$, then $f(u_i, e_j) = 0$. Thus, we have $V_i(h, e_j) = \{0,1\}$. 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 $c$ as equation (4) presents.

$$C_{(F,E)} : \{(+, *, ) \} 
S_1 = \{(+, ), +% \}$$

(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 $(F, E) = \{(F, a_1), (F, a_2), \ldots, (F, a_\alpha)\}$.

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, $1 \leq i \leq m, \sum_{i=1}^{m} P_i = 1$.
1. So, every \( m \)-tuple of non-negative integers \((n_1, n_2, \ldots, n_m)\) with sum \( n \):

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

**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:

\[
P(n_1 = 0, n_2 = 2, n_3 = 3) = \frac{10!}{0!2!0!3!2!0!5!} 0.2^2 0.3^3 0.2^5 = 0.135.
\]

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

\[
f(x, a_k) = \prod_{j=1}^{p} \prod_{i=1}^{m} (a_k^{(l)})^{x_{i,j}},
\]

where \( \sum_{i=1}^{l} a_k^{(l)} = 1 \) and \( j = 1, \ldots, p \) have \( m \) categories, and \( m = \sum_{j=1}^{p} m_j \) denotes the total levels number.

**II. MATERIAL AND 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 to be in Definition 4.

**Definition 4.** Suppose \( U \) includes an unsystematic sample-size \( |U| \) since division \( f(y, \lambda) \). Partition \( U = \{u_1, u_2, \ldots, u_{|U|}\} \) into \( K \) cluster \( C = \{C_1, C_2, \ldots, C_K\} \) through value \( \lambda_{ik} \) where \( \lambda_{ik} = 1 \) if \( u_i \in C_k \) and \( \lambda_{ik} = 0 \) if else. The function of cluster shared distribution of \( U \) due to cluster \( C \) could be described as \( \prod_{i=1}^{K} \prod_{u \in C_k} \lambda_{ik} f(y, \lambda) \).

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 \((F, a_1), \ldots, (F, a_{|A|}) \subseteq (F, A) \) and \((F, a_1), \ldots, (F, a_{|A|}) \subseteq (F, A) \). Lets \( \lambda_{k,j} \) is a probability of \( u_i \in (F, a_j) \) into cluster \( C_k, k = 1, 2, \ldots, K, i = 1, 2, \ldots, |U|, j = 1, 2, \ldots, |A| \) and \( l = 1, 2, \ldots, |a_j| \); hence, the function of the multivariate multinomial distribution soft-set can be written as follows:

\[
f_k(y, \lambda) = \prod_{j=1}^{|A|} \prod_{i=1}^{|a_j|} (\lambda_{k,j})^{[F,a_j]}|_{\lambda_{k,j}} - \sum_{l=1} \lambda_{k,j} = 1, \forall k, j.
\]

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.

\[
CML(\lambda, \mu) = \prod_{k=1}^{K} \prod_{i=1}^{|U|} \prod_{j=1}^{|A|} (\lambda_{k,j})^{[F,a_j]} - \sum_{l=1} \lambda_{k,j} = 1, \forall i \in \{0,1\} for i = 1, 2, \ldots, |U|.
\]

Consider the extension to allow the indicator 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 \( i = 1, \ldots, I \). In this case \( \mu = \{\mu_1, \mu_2, \ldots, \mu_K\}\) 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)

\[
Maximize L_{CML}(\mu, \lambda) = \sum_{k=1}^{K} \sum_{i=1}^{|U|} \sum_{j=1}^{|A|} \ln(\lambda_{k,j})^{[F,a_j]}
\]

Subject to

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

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.

\[
L_{CML}(\mu, \lambda, w_1, w_2) = \sum_{k=1}^{K} \sum_{i=1}^{|U|} \sum_{j=1}^{|A|} \ln(\lambda_{k,j})^{[F,a_j]} - w_1 \left( \sum_{k=1}^{K} \mu_{ik} - 1 \right) - w_2 \left( \sum_{i=1}^{|U|} \lambda_{k,j} - 1 \right)
\]

The first derivative of the lagrangian L_{CML} is taken regarding the \( \lambda_{k,j}, \lambda_{k,j}, \lambda_{k,j}, w_2, \lambda_{k,j}, \lambda_{k,j}, w_2 \) and set to 0 with the following equation.

\[
\frac{\partial L_{CML}}{\partial \mu_{ik}} = m \mu_{ik}^{-1} \sum_{j=1}^{|A|} \ln(\lambda_{k,j})^{[F,a_j]} - w_1 = 0,
\]

\[
\frac{\partial L_{CML}}{\partial \lambda_{k,j}} = \sum_{k=1}^{K} \mu_{ik} m \lambda_{k,j}^{[F,a_j]} - w_2 = 0,
\]

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

\[
\frac{\partial L_{CML}}{\partial w_2} = -\left( \sum_{i=1}^{|U|} \lambda_{k,j} - 1 \right) = 0.
\]
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 is obtained as follows:

$$\lambda_{kj} = \frac{\sum_{i\in\mathcal{F}(\mathcal{A})} \psi_{kj}^{i} \mu_{ki}^{i}}{\sum_{i\in\mathcal{F}(\mathcal{A})} \mu_{ki}^{i}}$$  \hspace{1cm} (12)

$$\mu_{ik} = \left[ \frac{\sum_{j=1}^{n_{e}} \left( \frac{\psi_{kj}^{j}}{\psi_{kj}^{j}} \right) \mu_{ki}^{j}}{\sum_{j=1}^{n_{e}} \mu_{ki}^{j}} \right]^{1-1}$$

The algorithm of the proposed technique can be described in Figure 1.

![Figure 1](image1.png)

The algorithm starts by decomposing the data into multi-soft sets and computing the initial membership randomly.

The algorithm of the proposed technique can be described in Figure 1.

![Figure 1](image1.png)

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

The algorithm starts by decomposing the data into multi-soft sets and computing the initial membership randomly.

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 Balloon 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.
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/V1/2023. This support is gratefully acknowledged.

REFERENCES


