Fast Hard Clustering Based on Soft Set Multinomial Distribution Function

Iwan Tri Riyadi Yanto[1,4](✉), Ririn Setiyowati[2], Mustafa Mat Deris[3], and Norhalina Senan[4]

[1] Department of Information Systems, University Ahmad Dahlan, Yogyakarta, Indonesia
yanto.itr@is.uad.ac.id

[2] Department of Mathematics, Universitas Sebelas Maret, Jalan Ir. Sutami 36A, Kentingan, Surakarta, Indonesia
ririnsetiyowati@staff.uns.ac.id

[3] Faculty of Applied Science and Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia
mmustafa@uthm.edu.my

[4] Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia
halina@uthm.edu.my

Abstract. Categorical data clustering is still an issue due to difficulties/complexities of measuring the similarity of data. Several approaches have been introduced and recently the centroid-based approaches were introduced to reduce the complexities of the similarity of categorical data. However, those techniques still produce high computational times. In this paper, we proposed a clustering technique based on soft set theory for categorical data via multinomial distribution called Hard Clustering using Soft Set based on Multinomial Distribution Function (HCSS). The data is represented as a multi soft set where every soft set have its probability to be a member of the clusters. Firstly, the corrected proof is shown mathematically. Then, the experiment is conducted to evaluate the processing times, purity and rand index using benchmarks datasets. The experiment results show that the proposed approach have improve the processing times up to 95.03% by not compromising the purity and rand index as compared with baseline techniques.

Keywords: Clustering · Categorical data · Multi soft set · Multinomial distribution function

List of Symbols and Abbreviations

S:	Information system/information Table
$S_{\{0,1\}}$:	System with value {0, 1}
U:	Universe
$\|U\|$:	Cardinality of U
u:	Object of U
A:	Set of Attribute/Variables

R. Ghazali et al. (Eds.): SCDM 2022, LNNS 457, pp. 3–13, 2022.
https://doi.org/10.1007/978-3-031-00828-3_1

a:	Subset of attribute
E:	Parameter in soft set
i:	Index i
j:	Index j
k:	Indek k
l:	Index l
e:	Subset of parameter
V:	Domain Value set
V_a:	Domain (values set) of variable a
f:	Information Function
F:	Maps parameter function
y:	Object
$P(U)$:	Power of Universe
(F, A):	Soft set
$F(a)$:	*Soft set of parameter a*
$C_{(F,E)}$:	Class soft set
P:	Probability
p_i:	Probability for each trial i
$f(x, a_k)$:	Probability mass function
n_i, N_i:	Number of Trial i
λ:	Probability of multinomial distribution
C_k:	Cluster k
K:	Number of clusters
z_{ik}:	Indicator function
$CML(z, \lambda)$:	Conditional maximum likelihood function
$MaximizeL_{CML}(z, \lambda)$:	Maximizing the log-likelihood function
$L_{CML}(z, \lambda, w_1, w_2)$:	Lagrange function
w_1:	Lagrange multiplier constrains 1
w_2:	Lagrange multiplier constrains 2
HCSS:	Hard Clustering using Soft Set based on Multinomial Distribution Function

1 Introduction

Clustering is the process of partitioning data sets from multiple variables into groups. The clustering problem often arises in the fields like image processing [1], pattern recognition [2], control system [3]. Until now, the most popular algorithm from various clustering algorithms that have been developed is k-means algorithm [2, 4, 5]. It produces efficiency and effectiveness in clustering with a large amount of data sets. However, k-means clustering algorithm unable to solve data sets that has categorical variables. The algorithm is only able to minimize a numerical cost function. Nevertheless, the k-means clustering algorithm was improved by Huang [4] into the k-modes clustering algorithm to eliminate the numeric-only limitation. Since then, the k-modes algorithms began to make major improvement such as the improvement of k-modes clustering using new dissimilarity measures [6–8] and k-modes algorithm based on fuzzy set [9, 10]. Another algorithms least sum of square based for non-parametric approach clustering has been discussed in [11–14].

Due to its relatively good performance, some improved versions of k-modes [15–17] have been proposed using more effective dissimilarity measurements to distinguish the importance of different attribute values. Furthermore, Kim et al. [18] proposed the use of fuzzy centroids approach to upgrade the efficiency of fuzzy k-modes. It has been improved by [19] to handle mix data numerical and categorical data based on genetic algorithm. Also, the fast clustering is still in concern currently especially in large dataset [3, 20, 21]. Another problem in categorical data is there are no inherent distance measure object to another object. The clustering algorithms developed for managing numerical data cannot directly be used to cluster categorical data [11]. Thus, the challenging of categorical data clustering is more than the numerical. Since categorical data is regularly watched as tallies coming about from a settled number of trials in which each trial comprises of making one determination from a prespecified set of categories. The categorical data can be assumed as from trial independent following the multinomial distribution. Thus, the parametric approach is more suitable for categorical data [22]. In [23] discussed some of parametric approach for categorical data clustering. However, almost all categorical data clustering techniques listed in [19] represent binary data sets. The problem with the aforementioned methods is that they have a long computation time and a low cluster purity.

On the other hand, categorical data have multi-valued attribute where it can be represented as a multi soft set [24]. The theory of soft set proposed by Molodtsov [25] is a new method for dealing with uncertainties in data. Some exiting clustering techniques based on soft set theory have been proposed in [26–28]. When compared to the theories of fuzzy set, probability, and interval mathematics, one of the key advantages of soft set theory is that it is free of the insufficiency of the parameterization tools. Whereas, the concept of multi-soft sets proposed by [24] is used for a multi-valued information systems to be applied to the categorical data without representing data in the binary values [24]. Thus, we would like to propose a Fast Hard Clustering based on Soft Set Multinomial Distribution Function to cluster the categorical data.

The rest of the paper is organized as follows Sect. 2 describes related works on information system, soft set, multinomial distribution. Section 3 constructs the mathematical modelling of the problem and proof the solution mathematically. Section 4 runs the computation experiment on data set. Finally, we conclude our work in Sect. 5.

2 Related Works

This section describes the basic of Information system, soft set theory and multinomial distribution.

2.1 Information System

Let's 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 $a \subset A$ and the information function is a total function as in Eq. (1) such that $f(u, a) \in V_a, \forall_{(u,a) \in U \times A}$.

$$f : U \times A \to V. \tag{1}$$

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 $S_{\{0,1\}}$.

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

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

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

2.2 Soft Set Theory

Soft set [25, 26] is a mathematical method for dealing with uncertainty via appropriate parametrization. Let U be an universe set, E be a set of parameters and $A \subset E$, F be the function that maps parameter A into the set of all subsets of the set U as shown in Eq. (4).

$$F : A \to P(U). \tag{4}$$

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

Consider to an information system definition, a soft set can be interpreted as a special type of information systems, termed a binary-valued information.

Proposition 1. Each Soft set (F, A) can be defined as $\mathrm{S}_{\{0,1\}}$.

Proof: Lets the set of universe U in (F, E) can be considered as the universe U, the set of parameters denoted by E where $A \subset E$. Next, the function of the information system, f is written as:

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

That is, when $u_i \in F\left(e_j\right)$, where $u_i \in U$ and $e_j \in E$, then $f\left(u_i, e_j\right) = 1$, otherwise $f\left(u_i, e_j\right) = 0$. To this, we have $V\left(h_i, e_j\right) = \{0, 1\}$. Therefore, for $A \subset E$, (F, A) can be represented as $\left(U, A, V_{\{0,1\}}, f\right)$. Thus, based on Definition 1, it can be defined as $\mathrm{S}_{\{0,1\}}$.

Definition 2. The value-class of the soft set denoted by $C_{(F,E)}$ are the class of all value sets of a soft set (F, E).

Based on Proposition 1, A Boolean-valued information system deals with the "standard" soft set. For a categorical value of information system denoted by $S = (U, A, V, f)$ with $V = \bigcup_{a \in A} V_a$ and V_a states the domain of attribute a. The domain V_a has categorical values or multi values. A decomposition can be constructed from S into $|A|$ number of Boolean-valued information system $S = \left(U, A, V_{\{0,1\}}, f\right)$. The decomposition of $A = \left\{a_1, a_2, \cdots, a_{|A|}\right\}$ into the disjoint-singleton attribute $\{a_1\}, \{a_2\}, \cdots, \left\{a_{|A|}\right\}$ is the basis of decomposition of $S = (U, A, V, f)$.

Definition 3. [24] Suppose that $S = (U, A, V, f)$ is a categorical-valued information system anda Boolean-valued information system is expressed by $S = \left(U, a_i, V_{a_i}, f\right)$, $i = 1, 2, \cdots |A|$ with

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

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

2.3 Multinomial Distribution

A generalization of the binomial distribution is the multinomial distribution [29]. Lets N_i be the number of results in category i in a series of independent trials a with probability p_i for each trial, where, $1 \leq i \leq m$, $\sum_{i=1}^{m} p_i = 1$. Then for each m-tuple of non-negative integers $(n_1, n_2, \ldots, n_m)$ with sum n.

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

Example 1. Suppose, there are 10 balls in a basket consists 2 red balls, 3 green balls and 5 blue balls. From the basket, 4 balls will be selected, with replacement. Then, the probability of drawling 2 green balls and 2 blue balls is

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

A multinomial distribution with parameter $a_k = (a_k^{jl}, l = 1, \ldots, m_j, j = 1, \ldots, p)$ can be described as the probability mass function as follows;

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

where $\sum_{i-1}^{m_j} a_k^{jl} = 1$. The generic polytomous variable $j (j = 1, \ldots, p)$ consist of categories m_j, and $m = \sum_{j=1}^{p} m_j$ indicates the total number of levels.

3 Hard Clustering Using Soft Set Based on Multinomial Distribution Function (HCSS)

Assume that U is a random sample size $|U|$ from distribution $f(y, \lambda)$. A partition $U = \{u_1, u_2, \ldots, u_{|U|}\}$ into K cluster $C = \{c_1, c_2, \ldots, c_K\}$ by indicator z_{ik} where $z_{ik} = 1$ if $u_i \in c_k$ and $z_{ik} = 0$ if otherwise. Then, the cluster joint distribution function of U based on cluster C can be defined as $\prod_{k=1}^{K} \prod_{u_i \in c_k} z_{ik} f_k(y, \lambda)$.

To the pair (F, A), select it to multi-soft set over U which represents a categorical-valued information system $S = (U, A, V, f)$, with $(F, a_1), \cdots, \left(F, a_{|A|}\right) \subseteq (F, A)$ and $\left(F, a_{j1}\right), \cdots \left(F, a_{j|a_j|}\right) \subseteq \left(F, a_j\right)$. Suppose that λ^i_{kjl} is a probability of $u_i \in \left(F, a_{jl}\right)$ into cluster C_k, $k = 1, 2, \ldots, K$, $i = 1, 2, \ldots, |U|$, $j = 1, 2, \ldots, |A|$ and $l = 1, 2, \ldots, \left|a_j\right|$, thus, the MMD of multi soft set can be written as

$$f_k(y, \lambda) = \prod_{j=1}^{|A|} \prod_{l=1}^{|a_j|} \left(\lambda^i_{kjl}\right)^{\left|F, a_{j_l}\right|}, \textit{ where} \sum_{l=1}^{|a_j|} \lambda_{kjl} = 1, \forall k, j. \tag{9}$$

Thus, the objective function of the clustering is to find the highest probability (λ) of the conditional maximum likelihood function as in (10) to assign the U to cluster C.

$$CML(z, \lambda) = \prod_{k=1}^{K} \prod_{i=1}^{|U|} z_{ik} \prod_{j=1}^{|A|} \prod_{l=1}^{|a_j|} \left(\lambda^i_{kjl}\right)^{\left|F, a_{j_l}\right|}. \tag{10}$$

where

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

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

Equation (10) is equivalent to maximizing the log-likelihood as in (11).

$$\begin{aligned} \textit{Maximize} L_{CML}(z, \lambda) &= \sum_{k=1}^{K} \sum_{i=1}^{|U|} z_{ik} \prod_{j=1}^{|A|} \prod_{l=1}^{|a_j|} \left(\lambda^i_{kjl}\right)^{\left|F, a_{j_l}\right|} \\ &= \sum_{k=1}^{K} \sum_{i=1}^{|U|} z_{ik} \sum_{j=1}^{|A|} \sum_{l=1}^{|a_j|} \ln\left(\lambda^i_{kjl}\right)^{\left|F, a_{j_l}\right|}. \end{aligned} \tag{11}$$

Subject to

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

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

Proposition: Lets (F, A) be a soft set over U which represents a categorical-valued information system with $(F, a_1), \cdots, \left(F, a_{|A|}\right) \subseteq (F, A)$ and $\left(F, a_{j_1}\right), \cdots, \left(F, a_{j|a_j|}\right) \subseteq \left(F, a_j\right)$. Suppose $(F, a_1), \cdots, \left(F, a_{|A|}\right) \subseteq (F, A)$ and $\left(F, a_{j_1}\right), \cdots, \left(F, a_{j|a_j|}\right) \subseteq \left(F, a_j\right)$ be a multi soft set of U. Then z_{ik} and λ_{kjl} are local maximum for $L_{CML}(z, \lambda)$ if only if

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

$$z_{ik} = \begin{cases} 1 & if \sum\limits_{j=1}^{|A|}\ln\left(\lambda^{i}_{kjl}\right) = \max\limits_{1 \le k' \le K}\sum\limits_{j=1}^{|A|}\ln\left(\lambda^{i}_{kjl}\right) \\ 0 & \qquad\qquad otherwise \end{cases}. \tag{13}$$

Proof. The maximizing problem in Eq. (11) is equivalent to the Lagrangian function of L_{CML} as in (14).

$$L_{CML}(z, \lambda, w_1, w_2) = \sum_{i=1}^{|U|}\sum_{k=1}^{K} z_{ik}\sum_{j=1}^{|A|}\sum_{l=1}^{|a_j|}\ln\left(\lambda^{\mathrm{i}}_{kjl}\right)^{\left|F, a_{j_l}\right|} - w_1\left(\sum_{k=1}^{K} z_{ik} - 1\right) - w_2\left(\sum_{l=1}^{|a_j|}\lambda_{kjl} - 1\right) \tag{14}$$

By take the first derivative of the Lagrangian L_{CML} with respect to the z_{ik}, λ_{kjl}, w_1, w_2 and set to be 0. The equation system obtained can be defined as follows

$$\frac{\partial L_{CML}}{\partial z_{ik}} = \sum_{j=1}^{|A|}\sum_{l=1}^{|a_j|}\ln\left(\lambda^{\mathrm{i}}_{kjl}\right)^{\left|F, a_{j_l}\right|} - w_1 = 0, \tag{15}$$

$$\frac{\partial L_{CML}}{\partial \lambda_{kjl}} = \frac{\sum_{i=1}^{|U|} z_{ik}\left|F, a_{j_l}\right|}{\lambda_{kjl}} - w_2 = 0, \tag{16}$$

$$\frac{\partial L_{CML}}{\partial w_1} = -\left(\sum_{k=1}^{K} z_{ik} - 1\right) = 0, \tag{17}$$

$$\frac{\partial L_{CML}}{\partial w_2} = -\left(\sum_{l=1}^{|a_j|}\lambda_{kjl} - 1\right) = 0. \tag{18}$$

From (16)

$$w_2 = \frac{\sum_{i=1}^{|U|} z_{ik}\left|F, a_{j_l}\right|}{\lambda_{kjl}} \tag{19}$$

$$\lambda_{kjl} = \frac{\sum_{i=1}^{|U|} z_{ik}\left|F, a_{j_l}\right|}{w_2}$$

Substitute to (18)

$$\begin{aligned} \sum_{l=1}^{|a_j|} \lambda_{kjl} &= \sum_{l=1}^{|a_j|} \frac{\sum_{i=1}^{|U|} z_{ik}\left|F, a_{j_l}\right|}{w_2} \\ 1 &= \frac{\sum_{l=1}^{|a_j|} \sum_{i=1}^{|U|} z_{ik}\left|F, a_{j_l}\right|}{w_2} \\ w_2 &= \sum_{l=1}^{|a_j|} \sum_{i=1}^{|U|} z_{ik}\left|F, a_{j_l}\right| \end{aligned} \tag{20}$$

Substitute to (16), then

$$\begin{aligned} \sum_{l=1}^{|a_j|} \sum_{i=1}^{|U|} z_{ik}\left|F, a_{j_l}\right| &= \frac{\sum_{i=1}^{|U|} z_{ik}\left|F, a_{j_l}\right|}{\lambda_{kjl}}; \\ \lambda_{kjl} &= \frac{\sum_{i=1}^{|U|} z_{ik}\left|F, a_{j_l}\right|}{\sum_{l=1}^{|a_j|} \sum_{i=1}^{|U|} z_{ik}\left|F, a_{j_l}\right|} \end{aligned} \tag{21}$$

Then, for a given z, all the inner sums of quantity $\sum_{i=1}^{|U|} \sum_{k=1}^{K} z_{ik} \sum_{j=1}^{|A|} \sum_{l=1}^{|a_j|} \ln\left(\lambda_{kjl}\right)^{\left|F, a_{j_l}\right|}$ are non negative and independent. Maximizing the quantity is equivalent to maximizing the each inner sum. For $1 < k < K$ the inner sum the quantity as

$$\begin{aligned} &\sum_{i=1}^{|U|} z_{ik} \sum_{j=1}^{|A|} \sum_{l=1}^{|a_j|} \ln\left(\lambda_{kjl}\right)^{\left|F, a_{j_l}\right|} \\ \Leftrightarrow &\sum_{i=1}^{|U|} z_{ik} \left(\sum_{j=1}^{|A|} \sum_{l=1}^{|a_j|} \ln\left(\lambda_{kjl}\right)^{\left|F, a_{j_l}\right|} \right) \end{aligned} \tag{22}$$

for $1 < i < |U|$, z_{ik} is fix and non negative and for each $i = 1, 2, \ldots, |U|$, $\left|F, a_{j_l}\right| = 1$ if $u_1 \in \left(F, a_{j_l}\right)$ and $\left|F, a_{j_l}\right| = 0$ if $u_1 \notin \left(F, a_{j_l}\right)$, it follows that $\sum_{i=1}^{|U|} z_{ik}\left|F, a_{j_l}\right| = \sum_{u_i \in \left(F, a_{j_l}\right)} z_{ik}, \forall_{u_i} \in U, i = 1, 2, \ldots, |U|$. Thus,

$$\lambda_{kjl} = \frac{\sum_{u_i \in \left(F, a_{j_l}\right)} z_{ik}}{\sum_{l=1}^{|a_j|} \sum_{u_i \in \left(F, a_{j_l}\right)} z_{ik}} \tag{23}$$

and inner sum $\sum_{i=1}^{|U|} \sum_{k=1}^{K} z_{ik} \sum_{j=1}^{|A|} \sum_{l=1}^{|a_j|} \ln\left(\lambda_{kjl}\right)^{\left|F, a_{j_l}\right|}$ maximize iff each term $\sum_{j=1}^{|A|} \sum_{l=1}^{|a_j|} \ln\left(\lambda_{kjl}\right)^{\left|F, a_{j_l}\right|} = \sum_{j=1}^{|A|} \ln\left(\lambda_{kjl}^{i}\right), \forall_{u_i} \in U, i = 1, 2, \ldots, |U|, l = 1, 2 \ldots, |a_j|$ is maximize. Thus,

$$z_{ik} = \begin{cases} 1 & \mathit{if} \ \sum_{j=1}^{|A|} \ln\left(\lambda_{kjl}^{i}\right) = \max_{1 \le k' \le K} \sum_{j=1}^{|A|} \ln\left(\lambda_{kjl}^{i}\right) \\ 0 & \mathit{otherwise} \end{cases} \tag{24}$$

4 Computational Run on UCI Datasets

In the experiment, MATLAB version 9.0.0.341360 (R2016a) is used to determine the performance in terms of cluster purity, rand index and computational time of the HCSS and other two fuzzy k-based approaches. They are executed sequentially on the specifications of a computer with an Intel Core i5, the total main memory is 8GB, and the operating system is Mac OS High Sierra. The Experiment will be conducted on four categorical datasets obtained from the UCI Machine Learning Repository [30], namely Zoo, Spect, Monk and Car. The all techniques are run by 100 differences initial membership function randomly for each datasets. The average in term of cluster purity, Rank Index and Computational Time is captured in Fig. 1. It shows that the HCSS technique is able to maintain the cluster purity and Rank index compared by the FC and FkP. Nevertheless, The result of computation time indicates that HCSS overcome FC and FkP technique. In detail, FC and FkP respectively consume approximately 0.7017 s and 0.4615 s of execution time to Process four dataset in average. In contrast, PSS technique requires only approximately 0.031 s of execution time in average for four dataset. It clearly shows a improvement of execution time by 95.03% as in Table 1. Thus. the HCSS is superior in terms of computational time with able to maintenance the rank index and purity comparing to the baselines.

Table 1. Comparison results in term of time responses

	Zoo	Monk	Spect	Car	Average
FC	0.8732	0.9206	0.7037	0.7037	0.7017
FkP	0.2617	0.3754	0.4645	0.0099	0.4615
HCSS	0.0236	0.0253	0.0995	0.0107	0.0310
Improvement					95.03%

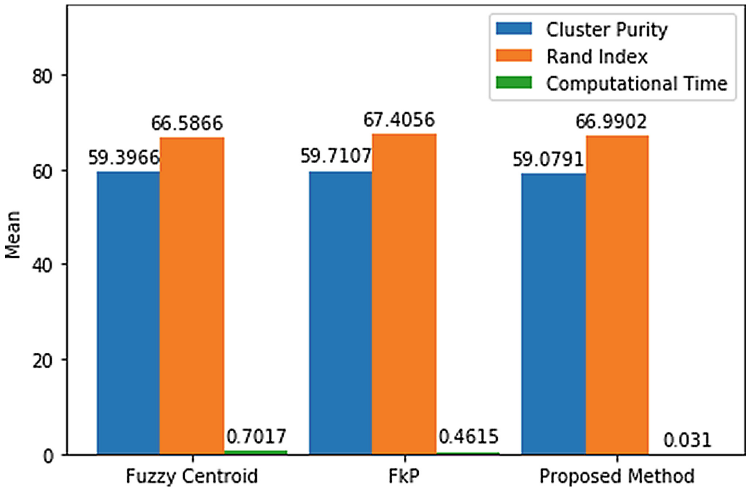

Fig. 1. Mean results of cluster purity, rand index, and computational time

5 Conclusion

The problem of fuzzy-based categorical data clustering can be overcome by several algorithms. However, these algorithms do not provide higher clusters purity and lower response times. Thus, the hard categorical data clustering based on soft set via multinomial distribution is proposed. The data is decomposed based soft set to become a multi soft set and multivariate multinomial distribution is used for clustering the data. Comparative analysis of the proposed algorithm called HCSS and two baseline algorithms with respect to purity, rand index and response time have been done. The results show that the proposed approach out performs the existing approaches in terms of lower response times up 95.03% by not compromising the purity and rand index. In the future work, we will extend the proposed approach based on fuzzy to increase the performance of the technique.

References

1. Arora, J., Tushir, M.: An enhanced spatial intuitionistic fuzzy c-means clustering for image segmentation. Procedia Comput. Sci. **167**, 646–655 (2020)
2. Chen, L., Wang, K., Wu, M., Pedrycz, W., Hirota, K.: K-means clustering-based kernel canonical correlation analysis for multimodal emotion recognition. IFAC-PapersOnLine **53**(2), 10250–10254 (2020)
3. Singh, S., Srivastava, S.: Review of clustering techniques in control system. Procedia Comput. Sci. **173**, 272–280 (2020)
4. Sinaga, K.P., Yang, M.: Unsupervised k-means clustering algorithm. IEEE Access **8**, 80716–80727 (2020)
5. Joshi, R., Prasad, R., Mewada, P., Saurabh, P.: Modified LDA approach for cluster based gene classification using k-mean method. Procedia Comput. Sci. **171**, 2493–2500 (2020)
6. Ng, M.K., Li, M.J., Huang, J.Z., He, Z.: On the impact of dissimilarity measure in k-modes clustering algorithm. IEEE Trans. Pattern Anal. Mach. Intell. **29**(3), 503–507 (2007)
7. San, O.M., Van-Nam, H., Nakamori, Y.: An alternative extension of the k-means algorithm for clustering categorical data. Int. J. Appl. **14**(2), 241–247 (2004)
8. He, Z., Deng, S., Xu, X.: Improving k-modes algorithm considering frequencies of attribute values in mode. In: Hao, Y., et al. (eds.) CIS 2005. LNCS (LNAI), vol. 3801, pp. 157–162. Springer, Heidelberg (2005). https://doi.org/10.1007/11596448_23
9. Huang, M.K.N.: A fuzzy k-modes algorithm for clustering categorical data. IEEE Trans. Fuzzy Syst. **7**(4), 446–452 (1999). https://doi.org/10.1109/91.784206
10. Wei, M.W.M., Xuedong, H.X.H., Zhibo, C.Z.C., Haiyan, Z.H.Z., Chunling, W.C.W.: Multi-agent reinforcement learning based on bidding. In: 2009 First International Conference on Information Science and Engineering (ICISE), vol. 20, no. 3 (2009)
11. Wei, W., Liang, J., Guo, X., Song, P., Sun, Y.: Hierarchical division clustering framework for categorical data. Neurocomputing **341**, 118–134 (2019)
12. Saha, I., Sarkar, J.P., Maulik, U.: Integrated rough fuzzy clustering for categorical data analysis. Fuzzy Sets Syst. **361**, 1–32 (2019)
13. Xiao, Y., Huang, C., Huang, J., Kaku, I., Xu, Y.: Optimal mathematical programming and variable neighborhood search for k-modes categorical data clustering. Pattern Recog. **90**, 183–195 (2019)
14. Zhu, S., Xu, L.: Many-objective fuzzy centroids clustering algorithm for categorical data. Expert Syst. Appl. **96**, 230–248 (2018)

15. Liu, C., et al.: A moving shape-based robust fuzzy k-modes clustering algorithm for electricity profiles. Electr. Power Syst. Res. **187**, 106425 (2020)
16. Golzari Oskouei, A., Balafar, M.A., Motamed, C.: FKMAWCW: categorical fuzzy k-modes clustering with automated attribute-weight and cluster-weight learning. Chaos, Solitons Fractals **153**, 111494 (2021)
17. Kuo, R.J., Zheng, Y.R., Nguyen, T.P.Q.: Metaheuristic-based possibilistic fuzzy k-modes algorithms for categorical data clustering. Inf. Sci. (Ny) **557**, 1–15 (2021)
18. Kim, D.-W., Lee, K.H., Lee, D.: Fuzzy clustering of categorical data using fuzzy centroids. Pattern Recogn. Lett. **25**(11), 1263–1271 (2004)
19. Nooraeni, R., Arsa, M.I., Kusumo Projo, N.W.: Fuzzy centroid and genetic algorithms: solutions for numeric and categorical mixed data clustering. Procedia Comput. Sci. **179**(2020), 677–684 (2021)
20. Schubert, E., Rousseeuw, P.J.: Fast and eager k-medoids clustering: O(k) runtime improvement of the PAM, CLARA, and CLARANS algorithms. Inf. Syst. **101**, 101804 (2021)
21. Leopold, N., Rose, O.: UNIC: A fast nonparametric clustering. Pattern Recogn. **100**, 107117 (2020)
22. Morris, D.S., Raim, A.M., Sellers, K.F.: A conway–maxwell-multinomial distribution for flexible modeling of clustered categorical data. J. Multivar. Anal. **179**, 104651 (2020)
23. Yang, M.S., Chiang, Y.H., Chen, C.C., Lai, C.Y.: A fuzzy k-partitions model for categorical data and its comparison to the GoM model. Fuzzy Sets Syst. **159**(4), 390–405 (2008)
24. Herawan, T., Deris, M.M.: On multi-soft sets construction in information systems. In: Huang, D.-S., Jo, K.-H., Lee, H.-H., Kang, H.-J., Bevilacqua, V. (eds.) ICIC 2009. LNCS (LNAI), vol. 5755, pp. 101–110. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04020-7_12
25. Molodtsov, D.: Soft set theory—first results. Comput. Math. Appl. **37**(4–5), 19–31 (1999)
26. Hartama, D., Yanto, I.T.R., Zarlis, M.: A soft set approach for fast clustering attribute selection. In: 2016 International Conference on Informatics and Computing (ICIC), pp. 12–15 (2016)
27. Jacob, D.W., Yanto, I.T.R., Md Fudzee, M.F., Salamat, M.A.: Maximum attribute relative approach of soft set theory in selecting cluster attribute of electronic government data set. In: Ghazali, R., Deris, M.M., Nawi, N.M., Abawajy, J.H. (eds.) SCDM 2018. AISC, vol. 700, pp. 473–484. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-72550-5_45
28. Sutoyo, E., Yanto, I.T.R., Saadi, Y., Chiroma, H., Hamid, S., Herawan, T.: A framework for clustering of web users transaction based on soft set theory. In: Abawajy, J.H., Othman, M., Ghazali, R., Deris, M.M., Mahdin, H., Herawan, T. (eds.) Proceedings of the International Conference on Data Engineering 2015 (DaEng-2015), pp. 307–314. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-1799-6_32
29. Malefaki, S., Iliopoulos, G.: Simulating from a multinomial distribution with large number of categories. Comput. Stat. Data Anal. **51**(12), 5471–5476 (2007)
30. Dheeru, D., Karra Taniskidou, E.: UCI Machine Learning Repository (2017)