Lecture Notes in Networks and Systems 457

Rozaida Ghazali · Nazri Mohd Nawi · Mustafa Mat Deris · Jemal H. Abawajy · Nureize Arbaiy *Editors*

Recent Advances in Soft Computing and Data Mining

Proceedings of the Fifth International Conference on Soft Computing and Data Mining (SCDM 2022), May 30–31, 2022

Springer

Lecture Notes in Networks and Systems

Volume 457

Series Editor

Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Advisory Editors

Fernando Gomide, Department of Computer Engineering and Automation—DCA, School of Electrical and Computer Engineering—FEEC, University of Campinas—UNICAMP, São Paulo, Brazil

Okyay Kaynak, Department of Electrical and Electronic Engineering, Bogazici University, Istanbul, Turkey

Derong Liu, Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, USA
Institute of Automation, Chinese Academy of Sciences, Beijing, China

Witold Pedrycz, Department of Electrical and Computer Engineering, University of Alberta, Alberta, Canada
Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Marios M. Polycarpou, Department of Electrical and Computer Engineering, KIOS Research Center for Intelligent Systems and Networks, University of Cyprus, Nicosia, Cyprus

Imre J. Rudas, Óbuda University, Budapest, Hungary

Jun Wang, Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong

The series “Lecture Notes in Networks and Systems” publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

Indexed by SCOPUS, INSPEC, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

For proposals from Asia please contact Aninda Bose (aninda.bose@springer.com).

More information about this series at https://link.springer.com/bookseries/15179

Rozaida Ghazali · Nazri Mohd Nawi ·
Mustafa Mat Deris · Jemal H. Abawajy ·
Nureize Arbaiy
Editors

Recent Advances in Soft Computing and Data Mining

Proceedings of the Fifth International Conference on Soft Computing and Data Mining (SCDM 2022), May 30–31, 2022

Springer

SCDM2022

Editors
Rozaida Ghazali
Faculty of Computer Science
and Information Technology
Universiti Tun Hussein Onn Malaysia
Batu Pahat, Malaysia

Mustafa Mat Deris
Faculty of Computer Science
and Information Technology
Universiti Tun Hussein Onn Malaysia
Batu Pahat, Malaysia

Nureize Arbaiy
Faculty of Computer Science
and Information Technology
Universiti Tun Hussein Onn Malaysia
Batu Pahat, Malaysia

Nazri Mohd Nawi
Faculty of Computer Science
and Information Technology
Universiti Tun Hussein Onn Malaysia
Batu Pahat, Malaysia

Jemal H. Abawajy
School of Information Technology
Faculty of Science, Engineering
and Built Environment
Deakin University
Geelong, VIC, Australia

ISSN 2367-3370 ISSN 2367-3389 (electronic)
Lecture Notes in Networks and Systems
ISBN 978-3-031-00827-6 ISBN 978-3-031-00828-3 (eBook)
https://doi.org/10.1007/978-3-031-00828-3

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2022
This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Rapid advancements in data storage technology along with the increase in data accessibility have paved the way for data science to become one of the fastest-growing research and application fields. Data science revolves around gaining insights from data, using different tools, statistical models, and machine learning algorithms, with the goal to discover hidden patterns from the raw data. To take on competitors, organizations need to recruit more and more skilled data scientists to help them leverage data analytics. However, extracting useful information has proven extremely challenging. Our conventional mathematical and analytical methods still face difficulty in deciphering complex data systems. To tackle this, data mining, which supports a wide range of business intelligence applications, has opened up exciting opportunities for discovering patterns in various types of data. With the deployment of data and soft computing techniques to scour extensive databases, diverse unique and meaningful patterns can be found, which otherwise remain unknown. As a result, new theories, algorithms, and technologies are continually being developed to run advanced statistical interpretations. Additionally, soft computing techniques can handle imprecision, uncertainty, partial truth, and approximation to achieve tractability, robustness, and low solution cost. The techniques, individually or in an integrated manner, are turning out to be strong candidates for performing tasks in the area of data mining, business, decision support systems, supply chain management, medicine, financial systems, automotive systems and manufacturing, image processing, etc. It provides the challenge of transforming data into innovative solutions perceived as a new value by customers.

Following the success of our four previous SCDM conferences in 2014 until 2020, we were glad to continue this journey of achievements with our fifth international conference. This year, the SCDM 2022 was held in a virtual space on May 30–31, 2022. It allowed remote participants to access live, interactive networking opportunities, and content, no matter where they are located. We received 61 paper submissions from 14 countries around the world. The conference also approved one special session that is Emerging Trends in Intelligent Systems and Data Science. Each paper in regular submission and special session was screened by the proceeding's chair and carefully peer-reviewed by at least three experts from the program committee. Finally, only 39 papers with the highest quality and merit were accepted for oral presentation and publication in this volume proceeding, giving an acceptance rate of 64%.

On behalf of SCDM 2022, we would like to express our highest gratitude to the conference organizer; Faculty of Computer Science & Information Technology, UTHM, and also to the Soft Computing & Data Mining research group, Steering Committee, Conference Chair, Program Committee Chair, Organizing Chairs, Special Session Chair, all Program and Reviewer Committee members for their valuable efforts in the review process that helped us to guarantee the highest quality of the selected papers for the conference.

We would also like to express our thanks to the keynote speakers, Prof. Dr Farid Meziane from the University of Derby, England; Dr Afnizanfaizal Abdullah from Aerodyne Group, Malaysia; and Prof. Dr Abdul Samad Hasan Basari from Universiti Tun Hussein Onn Malaysia. Our special thanks are also due to Dr Thomas Ditzinger for publishing the proceeding in Lecture Notes in Networks and Systems, Springer. We wish to thank the members of the organizing committee for their very substantial work, especially those who played essential roles.

Lastly, we would like to give the warmest of thanks to all the authors for their valuable input as well as all the participants for their enthusiastic engagement. We thank you for your time, service, and for making this conference as successful as it is.

Rozaida Ghazali
Nazri Mohd Nawi
Mustafa Mat Deris
Jemal H. Abawajy
Nureize Arbaiy

Conference Organization

Patron

Wahid Razzaly (Vice Chancellor)	Universiti Tun Hussein Onn Malaysia

Advisory Committee

Ajith Abraham	Machine Intelligence Research Labs, USA
Hamido Fujita	Iwate Prefectural University, Japan
Junzo Watada	Waseda University, Japan
Nikola Kasabov	KEDRI, Auckland University of Technology, New Zealand
Rajkumar Buyya	University of Melbourne, Australia
Witold Pedrycz	University of Alberta, Canada

Steering Committee

Mustafa Mat Deris	Universiti Tun Hussein Onn Malaysia
Jemal H. Abawajy	Deakin University, Australia
Nazri Mohd Nawi	Universiti Tun Hussein Onn Malaysia
Rozaida Ghazali	Universiti Tun Hussein Onn Malaysia

Chair

Nazri Mohd Nawi	SMC, Universiti Tun Hussein Onn Malaysia

Proceeding Chairs

Rozaida Ghazali	Universiti Tun Hussein Onn Malaysia
Nureize Arbaiy	Universiti Tun Hussein Onn Malaysia

Program Committee Chair

Mohd Norasri Ismail	Universiti Tun Hussein Onn Malaysia

Special Session Chair

Ezak Fadzrin Ahmad Shaubari	Universiti Tun Hussein Onn Malaysia

Organizing Committee

Hairulnizam Mahdin	Universiti Tun Hussein Onn Malaysia
Norhalina Senan	Universiti Tun Hussein Onn Malaysia
Sofia Najwa Ramli	Universiti Tun Hussein Onn Malaysia
Nurezayana Zainal	Universiti Tun Hussein Onn Malaysia
Noor Zuraidin Mohd Safar	Universiti Tun Hussein Onn Malaysia
Siti Hawa Ruslan	Universiti Tun Hussein Onn Malaysia
Norashid Hassan	Universiti Tun Hussein Onn Malaysia
Sahran Amzah	Universiti Tun Hussein Onn Malaysia

Program Committee

Abd Samad Hasan Basari	Universiti Tun Hussein Onn Malaysia
Ali Mohammadi	Isfahan University of Technology
Athraa Jasim Mohammed	University of Technology, Iraq
Bazeer Ahamed B.	University of Technology and Applied Sciences Al Musanna
Choon Sen Seah	Universiti Tunku Abdul Rahman
Chuah Chai Wen	Universiti Tun Hussein Onn Malaysia
Chuah Min Hooi	Universiti Sains Malaysia
Ezak Ahmad	Multimedia University, Malaysia
Fairouz Zendaoui	Ecole Nationale Supérieure d'Informatique
Fatima Zahra Fagroud	Hassan II University, Casablanca
Gede Pramudya	Universiti Tun Hussein Onn Malaysia
Kanaka Durga	Stanley College of Engineering and Technology for Women, India
Karrar Hameed Abdel Kareem	Al-Muthanna University
Khalil Ghathwan	University of Technology, Iraq
Mohammed Saeed Jawad	Universiti Tun Hussein Onn Malaysia
Mohd Amin Yunus	Universiti Tun Hussein Onn Malaysia
Mohd Fadzli Marhusin	Universiti Sains Islam Malaysia
Mohd Farhan Md Fudzee	Universiti Tun Hussein Onn Malaysia
Mohd Hafizul Afifi Abdullah	Universiti Teknologi Petronas
Mohd Najib Mohd Salleh	Universiti Tun Hussein Onn Malaysia

Mohd Norasri Ismail	Universiti Tun Hussein Onn Malaysia
Mohit Jain	NSIT, University of Delhi, India
Nayef Alduais	Universiti Tun Hussein Onn Malaysia
Noor Azah Samsudin	Universiti Tun Hussein Onn Malaysia
Noorhaniza Wahid	Universiti Tun Hussein Onn Malaysia
Nordiana Rahim	Universiti Tun Hussein Onn Malaysia
Norhalina Senan	Universiti Tun Hussein Onn Malaysia
Norhamreeza Abdul Hamid	Universiti Tun Hussein Onn Malaysia
Norhanifah Murli	Universiti Tun Hussein Onn Malaysia
Noureen Talpur	Universiti Tun Hussein Onn Malaysia
Nur Fatin Liyana Mohd Rosely	Universiti Teknologi Malaysia
Nur Ziadah Harun	Universiti Tun Hussein Onn Malaysia
Nureize Arbaiy	Universiti Tun Hussein Onn Malaysia
Nurezayana Zainal	Universiti Tun Hussein Onn Malaysia
Pradeep Kumar	Maulana Azad National Urdu University
Pramit Brata Chanda	Jadavpur University, India
Rabatul Aduni Sulaiman	Universiti Tun Hussein Onn Malaysia
Rachid Saadane	LETI, EHTP, Morocco
Rahayu Hamid	Universiti Tun Hussein Onn Malaysia
Rosmayati Mohemad	Faculty of Ocean Engineering Technology and Informatics
Rozaida Ghazali	Universiti Tun Hussein Onn Malaysia
Ruhaila Maskat	Universiti Teknologi Mara (UiTM), Malaysia
Salama A. Mostafa	Universiti Tun Hussein Onn Malaysia
Sasalak Tongkaw	Songkhla Rajabhat University
Shaik Rasool	Methodist College of Engineering and Technology
Suziyanti Marjudi	Universiti Tun Hussein Onn Malaysia
Syarulnaziah Anawar	Fakulti teknologi Maklumat dan Komunikasi
Szymon Lukasik	AGH University of Science and Technology
Uma N. Dulhare	Muffkham Jah College of Engineering and Technology
Umer Iqbal	Universiti Tun Hussein Onn Malaysia
Vittorio Cuculo	University of Milan
Waidah Ismail	Universiti Sains Islam Malaysia
Yana Mazwin Mohmad Hassim	Universiti Tun Hussein Onn Malaysia
Zubaile Abdullah	Universiti Tun Hussein Onn Malaysia

Special Session Committee

Emerging Trends in Intelligent Systems and Data Science

Muhammad Faheem Mushtaq	The Islamia University of Bahawalpur, Pakistan
Rizwan Majeed	The Islamia University of Bahawalpur, Pakistan
Urooj Akram	The Islamia University of Bahawalpur, Pakistan

Organizer

Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia

Contents

General Track

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

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
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 $f(u, a) \in V_a, \forall_{(u,a)\in U\times A}$.

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

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 $a_i^v : U \rightarrow \{0, 1\}$, 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).

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

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

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

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, $1 \le i \le m$, $\sum_{i=1}^{m} p_i = 1$. Then for each m-tuple of non-negative integers $(n_1, n_2, \ldots, n_m)$ with sum n.

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

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;

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

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_{kjl}^{i} 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

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

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.

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

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).

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

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

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

$$z_{ik} = \begin{cases} 1 & if \ \sum_{j=1}^{|A|} \ln\left(\lambda_{kjl}^{i}\right) = \max_{1 \le k' \le K} \sum_{j=1}^{|A|} \ln\left(\lambda_{kjl}^{i}\right) \\ 0 & otherwise \end{cases}. \tag{13}$$

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

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

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

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

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

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

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

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,

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

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%

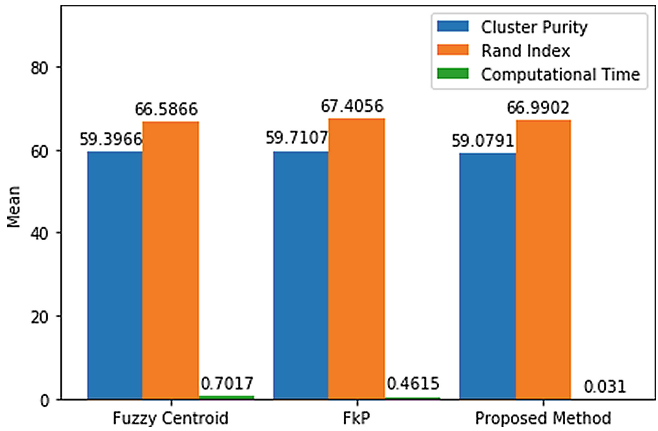

Fig. 1. Mean results of cluster purity, rand index, and computational time

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)

PSS: New Parametric Based Clustering for Data Category

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

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

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

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

Abstract. This paper proposes a new clustering technique for handling a categorical data called Parametric Soft set (PSS). It bases on statistical distribution namely multinomial multivariate function. The probability of the data category with binary value can be calculated by binomial distribution. Its generalization called multinomial distribution function for data category with multivariate values. Firstly, the data is represented as multi soft set where every object in each soft set has its probability. The probability of each object is calculated by cluster joint distribution function following the multivariate multinomial distribution function. The highest probability will be assigned to the related cluster. The first experiment is conducted to estimate the parameter of the data drawn from random multivariate mixtures distribution. While the second experiment is evaluated the processing times, purity and rand index using benchmarks datasets. The experiment results show that the proposed approach has improved the processing times up to 92.96%. It also has better performance in term of purity and rand index and error mean of the estimation parameters.

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

1 Introduction

There are two definitions assumed on the partitioning process or clustering process to group the data into several classes. First, well-defined notion of similarity or distance between data objects is needed to measure the resemblance the object. Second, the process to decide the object will be in the same groups or separate into differences group can be developed based on the characteristic of the data [1, 2]. In practice, it called unsupervised learning or clustering process.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
R. Ghazali et al. (Eds.): SCDM 2022, LNNS 457, pp. 14–24, 2022.
https://doi.org/10.1007/978-3-031-00828-3_2

There are so many clustering techniques developed because of many various similarity or distance measure in mathematics and many model which can be used to labeling the object such as [3–6]. It makes the notion of clusters cannot be precisely defined and create some various model of clustering i.e. centroid, density, distribution, connectivity, graph-based, neural models, etc. [7]. The clustering technique can be categorized into three types. i.e. pairwise distance cluster, target on optimizing by given merit function and statistical modeling [8]. Only pairwise distances between clustered objects are used in the first type. This is because a tractable mathematical representation for objects is not necessary, these approaches have a wide range of applications. However, due to the quadratic computational complexity of calculating all the pairwise distances, they do not scale well with big data sets. Linkage clustering [9–11] and spectra clustering [12] are two examples. The second type is concerned with optimizing a certain merit function. The merit function represents the widely held idea that good clustering requires objects in the same cluster to be similar, while objects in other clusters should be as diverse as possible. The similarity metric and criterion for evaluating the overall quality of clustering differ amongst algorithms. K-means and k-centroid are two terms that are included in this type. The third type is based on statistical analysis [8]. Each cluster is distinguished by a fundamental parametric distribution (known as a component), such as the multivariate Gaussian for continuous data, the Poisson distribution for discrete data, multinomial distribution for multi values data.

The differences of typical of the data requires careful consideration to determine the similarity or distance measure [2]. In practice, there are various types of data that are used to implement the clustering algorithm, such as numeric, and categorical. Unlike the numerical data, the categorical data contains the attributes which do not have any natural order, so distance measure cannot be executed straightforwardly on categorical attribute [13]. Data category can be assumed following the random multivariate multinomial distribution function [14]. Other hand, categorical data have multi-valued attribute where it can be represented as a multi soft set [15]. Thus, this paper proposes the parametric clustering approach based on soft set theory. The data is decomposed to be a multi soft set respect to all attributes where the probability every soft set in each attribute is calculated using multinomial distribution function. Each object on attributes has different values of probability respect to the cluster. The object with high probability will be assign into the related cluster.

The rest of the paper is organized as follows: Sect. 2 describes related works on information system, soft set, multinomial distribution. Section 3 describes the proposed approach based on soft set multinomial distribution function. Section 4 describes the experiment results on the estimation parameter. Finally, we conclude our work in Sect. 5.