HARD CLUSTERING TECHNIQUE BASED ON MULTI SOFT SET AND MULTINOMIAL DISTRIBUTION FUNCTION FOR CATEGORICAL DATA

IWAN TRI RIYADI YANTO

A thesis submitted in
fulfillment of the requirement for the award of the
Doctor of Philosophy in Information Technology

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

APRIL 2023

In the name of Allah, Most Gracious, Most Compassionate.

I praise and thank Allah.

Special thanks for my beloved father Ranto Sihman (Alm) and mother Sadiyem.

For dearest,
Ani Apriani, Raihana Isykarima Ananta, Ainun Mahya Ananta,
Muhamamad Farhan Al Fatih
(Wife, daughter, daughter, son)
For their love, support, enthusiasm, encouragement and motivation.

For my supervisor,
Prof. Dr. Mustafa Mat Deris
Dr. Norhalina Senan
For their incredible help, patience, understanding and support.

For all postgraduate members, fellow friends and ummah.

This thesis is dedicated to all of you.

ACKNOWLEDGEMENT

In the name of Allah, the Most Gracious, the Most Merciful. I praise and thank Allah for His blessings uncounted in my life and His willingness, I was able to complete this research successfully. This dissertation would not have been possible without the guidance, help, and support of many people who contributed and extended their valuable assistance in preparing and completing this research. I take this opportunity to express my profound sense of gratitude and respect to all those people.

First and foremost, I would like to express my sincere appreciation to my supervisor, Prof. Dr. Mustafa Mat Deris, and Dr. Norhalina Senan for their guidance and endless support from the initial to the final level, patience, and motivation, enthusiasm, and immense knowledge. Their feedback, editorial comments, and suggestions were also invaluable for writing this thesis, and I appreciate it.

In preparing this research, my gratitude is extended to Universiti Ahmad Dahlan (UAD) for supporting this research under the UAD scholarship. I would like to thank the Centre for Graduate Studies (CGS) of the Universiti Tun Hussein Onn Malaysia (UTHM) for their hospitality during my study. I also wish to thank the Faculty of Computer Science and Information Technology Postgraduate friends and staff who supported my research work in one way or another.

A special thanks to my beloved family for their continuous prayer, encouragement, love, support, patience, and care whenever I needed them during these challenging days. I dedicate this work to all of you. My overwhelming gratitude to all my friends who have been together with me; thanks for your love, care, concern, and support. Lastly, it is a pleasure to thank those who have helped directly or indirectly. Thank you.

ABSTRACT

Categorical data clustering is still an issue due the complexities of measuring the similarity of data. Unlike the numerical data, the categorical data contains the attributes which do not have any natural order. Distance measure-based technique such as *k*-mean cannot be executed straightforwardly on the categorical attribute. Fuzzy *k*-modes and its improvement likes Hard *k*-modes, Ng's *k*-modes, He's *k*-modes, Initialization *k*-modes, Fuzzy *k*-modes, Hard and Fuzzy Centroid were proposed to avoid the limitation of *k*-mean handling the categorical data. The Grade of Membership (GoM) and Fuzzy *k*-Partition (FkP) were proposed as a parametric-based to improve the Purity and accuracy. However, these clustering techniques still produce clusters with weak intra-similarity and low Purity . Moreover, converting categorical attributes into binary values makes complexities be high. On the other hand, categorical data have multivalued attribute that can be represented as a multi soft set and can be assumed following a random sample multivariate multinomial distribution. This study proposes a clustering technique based on soft set theory for categorical data via multinomial distribution function. The data is represented as multi soft set where every object in each soft set has probability. The probability of each object is calculated by the cluster joint distribution function following the multivariate multinomial distribution function. The experiment results show that the proposed technique has better performance cluster stability in term of Dunn Index. It has improved the error mean of the estimation parameters up to 24.29 % and 2.24%, reducing the complexity to 73.75% and processing times up to 92.96%, Rank Index up to 0.8850 and Purity 0.9197.

ABSTRAK

Pengelompokan data kategori masih menjadi isu kerana kerumitan dalam mengukur persamaan data. Tidak seperti data berangka, data kategori mengandungi atribut yang tidak mempunyai sebarang susunan semula jadi. Teknik berasaskan ukuran jarak seperti *k-mean* tidak boleh dilaksanakan secara langsung pada atribut kategori. *Fuzzy k-modes* dan penambahbaikannya seperti *Hard k-modes*, *Ng's k-modes, He's k-modes, Initialization k-modes, Fuzzy k-modes, Hard* dan *Fuzzy Centroid* dicadangkan untuk mengatasi kelemahan *k-mean* mengendalikan data kategori. *Grade of Membership* (GoM) dan *Fuzzy k-Partition* (FkP) dicadangkan sebagai teknik berasaskan parametrik untuk meningkatkan kemurnian dan ketepatan. Walau bagaimanapun, teknik pengelompokan ini masih menghasilkan kelompok dengan intra-kesamaan yang lemah dan kemurnian yang rendah. Lebih dari itu, penukaran atribut kategori ke dalam nilai binari membuat pengiraan lelaran menjadi kompleks. Sebaliknya, data kategori mempunyai atribut berbilang nilai di mana ia boleh diwakili sebagai set berbilang lembut dan boleh diandaikan berikutan taburan sampel rawak *multinomial multivariate*. Oleh itu, dalam kajian ini, teknik pengelompokan berdasarkan teori set lembut untuk data kategori melalui fungsi pengedaran multinomial dicadangkan. Data diwakili sebagai set berbilang lembut di mana setiap objek dalam setiap set lembut mempunyai kebarangkaliannya. Kebarangkalian setiap objek dikira oleh fungsi taburan bersama kelompok berikutan fungsi taburan multinomial multivariat. Hasil pengujian eksperimen menunjukkan bahawa teknik yang dicadangkan mempunyai kestabilan kelompok lebih baik dari segi indeks Dunn. Ia juga telah meningkatkan min ralat parameter anggaran sehingga 24.29% dan 2.24%, mengurangkan kerumitan komputeran sehingga 73.75% dan masa pemprosesan sehingga 92.96%, Rank Index sehingga 0.8850 dan kemurnian sehingga 0.9197.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

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

LIST OF APPENDICIES

LIST OF PUBLICATIONS

Journal :

(i) I. Tri, R. Yanto, A. Apriani, R. Hidayat, M. M. Deris , and N. Senan, "Fast Clustering Environment Impact using Multi Soft Set Based on Multivariate Distribution," JOIV Int. J. Informatics Vis., vol. 5, no. September, pp. 291–297, 2021.

(ii) I. Tri, R. Yanto, R. Saedudin, S. Novita, M. M. Deris , and N. Senan, "Soft Set Multivariate Distribution for Categorical Data Clustering," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 11, no. 5, pp. 1841–1846, 2021.

Proceeding/book chapter :

(i) I. T. R. Yanto, R. Setiyowati, M. M. Deris, and N. Senan, "Fast Hard Clustering Based on Soft Set Multinomial Distribution Function," in Recent Advances in Soft Computing and Data Mining, 2022, pp. 3–13.

(ii) I. T. R. Yanto, M. M. Deris, and N. Senan, "PSS: New Parametric Based Clustering for Data Category," in Recent Advances in Soft Computing and Data Mining, 2022, pp. 14–24.

CHAPTER 1

INTRODUCTION

1.1 Research background

Data clustering is the task of identifying groups or clusters of data instances so that instances in the same cluster are similar to each other in comparison to those in different clusters (Goncalves & Lourenco, 2019; McLachlan, Rathnayake, & Lee, 2020). It can be defined as a process of partitioning a given data set of multiple variables into groups. Clustering also serves as an important step in exploratory data mining, where the natural affinity of the data instances at hand can be revealed and utilized. Numerous branches of research and engineering have used clustering, including earth science, life science, social science, information science, medical science, policy, and decision-making. Additionally, it is adaptable to the preliminary stages of other research areas and applications, including bioinformatics, collaborative filtering, customer segmentation, data exploration, data summarization, dynamic trend detection, information retrieval, market basket analysis, medical diagnostics, multimedia data analysis, social network analysis, text mining, and web analysis (Soppari & Chandra, 2020; Wu & Zhang, 2020; Thrun & Stier, 2021).

Based on the membership of the data item (belongingness) in a cluster, the methods are again classified into hard and fuzzy clustering methods (Mrudula & Reddy, 2019; Gupta & Das, 2022). Any clustering method which produces clusters such that each data item categorically belongs to a single cluster is called the hard clustering method. In other words, hard clustering is when each data point is uniquely assigned to one and only one cluster (Carvalho *et al.,* 2018; Khandaker, Hussain, & Ahmed, 2019; Vardhan, Sarmah, & Das, 2020). In fuzzy clustering, each data item

belong to different clusters with some membership in each cluster (Wu *et al.,* 2019; Pinheiro, Aloise, & Blanchard, 2020).

Starting with a data set as input, the clustering process groups related data points into clusters until all the data points are grouped. A similarity/distance metric is used to determine the data points similarity. The first clustering methods such as *k*-mean, fuzzy c-mean, focus on numerical data by using derived concepts from statistics and geometry. Since the real-world data in most cases involves categorical rather than numerical values due to changing needs and time, existing clustering techniques have been restricted to numeric data exclusively (Saxena & Singh, 2016). Given that some measurements can identify structural characteristics, the clustering of numerical data in continuous space has been thoroughly explored during the past few decades. In contrast, due to the attribute value in the discrete domain, it is challenging to determine the structural information of categorical data immediately (Zhu & Xu, 2018). One of the structural information of categorical data is the multinomial distribution function (Chattamvelli & Shanmugam, 2020).

Categorical data is different from numeric data in the sense that it groups the data into categories and not any numeric values. Numerous real-world applications regularly use categorical data, including medical data and retail purchase transactions. For instance, categorical variables like nationality, gender, occupation, level of education, marital status, and smoking status are included in medical data. Retail purchase transactions include product categories, customer kinds, and locations. (Zhu & Xu, 2018; Dinh, Huynh, & Sriboonchitta, 2021). Recently, the research on categorical data clustering has also gained much attention (Zhu & Xu, 2018) and applied in the real case study such as social science, business, marketing, and finance (Beck *et al.,* 2021; Herrero & Villar, 2021; Holden & Hampson, 2021; Kim 2022), healthcare and medical science (Mosia & Joubert, 2020; Coombes *et al.,* 2021), and computer science (Cheng, Wang, & Ma, 2019). In social science, intelligent crime analysis makes it possible to find areas with high rates of criminology and illegal activity through clustering (Phillips & Lee, 2011). In Chicago, more than 30,000 people's travel patterns are examined. The investigation looked at the people's innate daily activity patterns and the variety of their daily activities, identified clusters of individual behaviors, and revealed information about their socio-demographics (Jiang, Ferreira, & González, 2012). The goal of clustering in business, marketing, and finance is to group the top tourist locations according to the expansion of the key tourism

metrics (Claveria & Poluzzi, 2017) which permits identifying highly desirable destinations. In another application, customer turnover is a serious issue that might impact the telecom sector. In actuality, keeping current clients will have a less of an economic impact than recruiting new ones (Amin *et al.,* 2017). The clustering approach, also known as cluster alarms, is used in the healthcare and medical research disciplines to assess and locate disease zones. Finding these diseased areas enables effective resource allocation for health control and prevention (Paul & Hoque, 2010). In computer science, the clustering approach is used to examine data streams in web applications (online social networks, blogs, and wikis), allowing for target marketing and group segmentation for electronic commerce (Chen & He, 2016). Clustering in cyber security can discover abnormalities, intrusions, and harmful information, allowing for the classification of both legitimate and malicious network traffic (Husak *et al.,* 2019).

The data containing categorical attributes pose some challenges to the existing clustering methods due to the absence of natural order, existence of subspace clusters, and conversion of categorical to numeric data. The traditional similarity measures are based on the co-occurrence of attribute values. Some others like the Jaccard Coefficient and Cosine similarity can even define similarity by seeing whether two attribute values occur together for any data point. If the attributes are not naturally ordered like in categorical data, the similarity between data points cannot be measured through the existing measures. Categorical data, being high dimensional, fail to cluster data in all dimensions and are limited to a certain number of dimensions. The only possible approach initially for categorical data clustering was to convert it into equivalent numeric form. There are a number of categorical data clustering techniques that have been developed. Huang (1998) proposed the *k*-modes clustering method that removes the numeric-only limitation of the *k*-means algorithm. However, no one approach can produce the optimal results across all data sets such as Link-Based cluster Ensemble, ROCK: A Robust Clustering Algorithm, Top-Down Parameter Free Clustering (Kim, Lee, & Lee, 2004). To improve the efficiency of fuzzy *k*-modes, (Kim, Lee, & Lee, 2004) proposed a technique called fuzzy centroids technique. Its non-parametric techniques are based on clusters least sum of squared errors. This selection implies, in essence, the assumption of data organized into spherical clusters where it is make low Purity (Bryant & Williamson, 1978; Yang, Chiang, Chen, & Lai, 2008; Chatzis, 2011).

The clustering process was performed using a distance metric, such as the simple matching dissimilarity measure (Hamming distance) for categorical clustering. The distance often results in clusters with weak intrasimilarity; thus, lower either accuracy or Purity will be obtained (Hsu, Chen, & Su, 2007). Historically, one of the most intuitive methods used for categorical clustering was to convert the categorical attributes into binary values using 0 or 1 to indicate whether the corresponding modality is absent or present in the observation. Thus, this method was inefficient due to the high dimensionality of the data sets to be generated. The idea of co-occurrences between attributes and pairs defining a cluster and subspace algorithm locates clusters in various subspaces of the data set are the foundations of the numerous different clustering algorithms for categorical data that have been discovered to solve the problem from a different perspective. Finding a good partition in an unsupervised learning process requires clustering a lot of data, which is a challenging challenge attempting to minimize similarity between things in separate clusters while the maximizing similarity between objects in the same cluster without any previous information. The currently used category clustering method has numerous drawbacks due to temporal complexity; working with a lot of dimensions and data points can be challenging (Chen, & Lai, 2008). Yang *et al* (2008) proposed Fuzzy *k*-partititon (FkP) algorithm, a parametric technique based on the likelihood function of multivariate multinomial distributions. It improves the Grade of Membership (GoM) model for categorical data analysis proposed by Woodbury & Clive (1974). The FkP also can be referred as a fuzzy-based clustering algorithm for categorical data. However, almost all fuzzy categorical data clustering algorithms mentioned above represent data set in binary values. Moreover, in FkP algorithm, the maximized parameter of the classification likelihood function in the same categories always has the same probability value (Naouali *et al.,* 2020). Although the GoM and FkP model is a good analysis tool for categorical data clustering, the algorithms have complex iteration calculations with high computational time. This indicates that the techniques that does not suffer from high computational time and low clusters Purity is needed.

The converted values are arbitrary and seem to no use beyond using it as a convenient label for a particular value. The reason behind the same is that each value in a categorical attribute represents a separate logical concept and therefore can neither be meaningfully ordered nor can be manipulated the way numbers could be (Saxena & Singh, 2016). In probability theory and statistics, categorical data can be assumed

to follow the random multivariate multinomial distribution function (Traylor, 2017; Chattamvelli & Shanmugam, 2020). This distribution can be considered as a generalization of the binomial distribution with two categories. It gives the joint probability of occurrence of multiple events in n independent trials. For multivariate categorical data, a standard parametric model used in latent class clustering is a locally independent product of multinomial (Chattamvelli & Shanmugam, 2020). Examples include the outcomes of elections for multiple political parties, the spread of epidemics (or deaths caused by them) among various ethnic groups, the inclusion of stocks in investment portfolios, accidents involving various vehicle types (cars, buses, and trucks), claims received under various types of actuarial sciences, and the types of books (such as fiction, novels, stories, poems, and science) that patrons check out from the library. Bacteriologists use it to simulate microbe counts by type in randomly scattered colonies and by geologists for soil analysis. Similarly, civil engineers create buildings that can survive sporadic occurrences like earthquakes, floods, strong winds, tornadoes, and fires. These depend on the structure's location, of course. Multinomial distribution can be used to calculate the likelihood of structural damage from numerous causes if the relevant probabilities are known from earlier data. (Chattamvelli & Shanmugam, 2020).

On the other hand, categorical data have multi-valued attribute that can be represented as a multi soft set (Herawan, Deris, & Abawajy, 2010; Khan *et al.,* 2018; Pardasani, 2018). The theory of soft set proposed by Molodtsov (D. Molodtsov, 1999) is called elementary neighborhood systems that it is free from the inadequacy of the parameterization tools, like in the theories of fuzzy set, probability and interval mathematics. In recent years, research on soft set theory has been active, and great progress has been achieved, including the works of fundamental soft set theory (Akram, Adeel, & Alcantud, 2019; Liu *et al.,* 2019; Aziz-ul-Hakim *et al.,* 2021; Kar & Dutta, 2021), association rule (Feng *et al.,* 2016; Gupta & Rai, 2017; Feng *et al.,* 2020; Jia & Zhang, 2021), decision making (Manna, Basu, & Mondal, 2020; Tao *et al.,* 2020; J. Yang & Yao, 2020) and handling incomplete data (Wen, Chang, & Lai, 2020; Kong *et al.,* 2021). It is a soft computing tools for data mining that has much use in the fields of business, health, education, agriculture, and many more (Kottam & Paul, 2020). In clustering problem, soft set theory has been applied in education (Saedudin *et al.,* 2017), web mining (Sutoyo *et al.,* 2019), and the environment (Tri *et al.,* 2021).

1.2 Problem statement

The clustering algorithms developed for managing numerical data cannot directly be used to cluster categorical data (Wei *et al.,* 2019). Thus, the challenge of categorical data clustering is more than the numerical. There are various algorithms have been introduced for clustering categorical data. Kim *et al* (2004) proposed using the hard and fuzzy centroids technique to upgrade the efficiency of fuzzy *k*-modes. The use of simple matching dissimilarity distance obtains the weak intra-similarity (Hsu *et al.,* 2007) and make either accuracy or Purity will be low. Another problem in categorical data is that there is no inherent distance measure object to another object. Since categorical data is regularly watched as tallies coming about from a settled number of trials in which each trial makes one determination from a prespecified set of categories, the categorical data can be assumed to from trial independent following the multinomial distribution. Thus, the parametric technique is more suitable for categorical data (Morris, Raim, & Sellers, 2020). For categorical data, a standard parametric model used in latent class clustering is a locally independent product of multinomial. Moreover, the categorical data can be assumed following a random sample multivariate multinomial distribution.

Yang *et al* (2008) proposed Fuzzy *k*-partititon (FkP) algorithm, a parametric technique based on the likelihood function of multivariate multinomial distributions. It improves the Grade of Membership (GoM) model for categorical data analysis proposed by Woodbury & Clive (1974). However, the algorithms is still need complex iteration calculations with high computational time. It is caused the FkP and GoM need to convert the categorical data into binary data. On the other hand, categorical data have multi-valued attribute that can be represented as a multi soft . The multi soft set used for multi-valued attribute has advantages in representing the categorical data without the need to be converted into binary values (Herawan, Deris, & Abawajy, 2010; Khan *et al.,* 2018). Thus, this study propose a clustering technique based on soft set theory for categorical data via multinomial distribution.

1.3 Research question

Having discussed the problem statement, the research questions that guided this research are as follows:

(i) How to implement the multinomial distribution function and soft set for categorical data clustering problem?

(ii) How do the multinomial distribution function and soft set work for categorical data clustering problem?

(iii) How is the multinomial distribution function on soft set performance compared to the baseline techniques?

1.4 Research objective

This research embarks on the following objectives:

(i) To propose a clustering technique for categorical data using multi soft set and multinomial distribution function in :
   a. reducing the computational complexity and time response, and
   b. improving the Purity of cluster.

(ii) To evaluate proposed technique performance in term of computational complexity, Mean Square Error of estimation parameter, Rank Index, Purity , Dunn Index, stability of creating number of cluster, and time response.

1.5 Scope of study

The scope of this research falls within the hard portioning type of clustering for categorical data using soft set and multinomial distribution function. The data is splitting into multi soft set and then the probability is calculated using a multinomial distribution function. The proposed technique will be validated by comparing results with the baseline techniques such as Hard Centroid, Fuzzy Centroid, GoM model and Fuzzy *k*-Partition with the performance measurement criteria of computational complexity, Mean Square Error of estimation parameter, Rank Index, Purity , Dunn Index, stability of creating number of cluster, and time response. The artificial data set, benchmarks data set and primary data set are used to validate the techniques. The

artificial data set is generated randomly from the mixture distribution function. The benchmarks data set taken from the University of California Irvine Machine Learning Repository (UCI) consists of seven data involving zoo, soybean, balloon, tic-tac-toe, monk, spect and car data sets. The primary data set is Rapid Visual Screening (RVS) data set collected at Kulonprogo, Yogyakarta.

1.6 Thesis outline

This research thesis comprises of six chapters including Introduction and Conclusion chapters. The followings are synopsis of each chapter.

Chapter 1: *Introduction*. Apart from providing an outline of the thesis, this chapter contains an overview of the research background, problem to be solved, objectives to be achieved, scope, aim, and outcome of the study.

Chapter 2: *Literature Review.* This chapter explains the basic of information system, Soft Set Theory, multinomial distribution function, reviews some of the work on categorical data clustering techniques and several indexes to measure the goodness of cluster results.

Chapter 3: *Research Methodology.* This chapter discusses the research methodology used to carry out the study systematically. The research phase involves literature review, mathematical modelling, model solving, data collection and performance analysis. The three data sources used for experiment are artificial data, secondary and primary data. The experiment is conducted measurement on computational complexity, estimation parameter, cluster analysis (Rank Index, Purity , Dunn Index) and time response.

Chapter 4: *Hard Clustering using Soft Set based on Multinomial Distribution Function*. This chapter explains the proposed technique. The mathematical modeling of the proposed technique is presented by assuming the categorical data following multivariate multinomial distribution function where each attribute decomposes the object into multi soft set. The objective function is solved by Lagrange multiplier to find the optimum solution. The manual calculation is given as an example of how the technique work.

Chapter 5: *Results and Analysis.* The performance analysis consists of measurements : computational complexity, estimation parameter, cluster analysis and time response.

The computational complexity is analyzed mathematically. The estimation parameter estimates the parameter of the artificial data. The cluster analysis consists of internal and external using Purity , Dunn Index, and Rank Index.

Chapter 6: *Conclusion and Future work*. The contributions of the proposed technique are summarized, and the recommendations are given for further continuation of works.

REFERENCES

Akram, M., Adeel, A., & Alcantud, J. C. R. (2019). Group decision-making methods based on hesitant N-soft sets. *Expert Systems with Applications*, *115*, 95–105. https://doi.org/https://doi.org/10.1016/j.eswa.2018.07.060

Alruwaili, M., Siddiqi, M. H., & Javed, M. A. (2020). A robust clustering algorithm using spatial fuzzy C-means for brain MR images. *Egyptian Informatics Journal*, *21*(1), 51–66. https://doi.org/https://doi.org/10.1016/j.eij.2019.10.005

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Alawfi, K., Hussain, A., & Huang, K. (2017). Customer churn prediction in the telecommunication sector using a rough set approach. *Neurocomputing*, *237*, 242–254. https://doi.org/https://doi.org/10.1016/j.neucom.2016.12.009

Aziz-ul-Hakim, Khan, H., Ahmad, I., & Khan, A. (2021). Fuzzy bipolar soft semiprime ideals in ordered semigroups. *Heliyon*, *7*(4), e06618. https://doi.org/https://doi.org/10.1016/j.heliyon.2021.e06618

Beck, A.-K., Berti, S., Czernochowski, D., & Lachmann, T. (2021). Do categorical representations modulate early automatic visual processing? A visual mismatch-negativity study. *Biological Psychology*, *163*, 108139. https://doi.org/https://doi.org/10.1016/j.biopsycho.2021.108139

Beraldi, P., Boccia, M., & Sterle, C. (2019). Special issue on: Optimization methods for decision making: advances and applications. *Soft Computing*, *23*(9), 2849–2852. https://doi.org/10.1007/s00500-019-03945-0

Bryant, P., & Williamson, J. A. (1978). Asymptotic behaviour of classification maximum likelihood estimates. *Biometrika* , *65*(2), 273–281. https://doi.org/10.1093/biomet/65.2.273

Cao, F., Liang, J., & Bai, L. (2009). A new initialization method for categorical data clustering. *Expert Systems with Applications*, *36*(7), 10223–10228. https://doi.org/https://doi.org/10.1016/j.eswa.2009.01.060

Chattamvelli, R., & Shanmugam, R. (2020). Multinomial Distribution BT - Discrete Distributions in Engineering and the Applied Sciences. In R. Chattamvelli & R.

Shanmugam (Eds.) (pp. 179–188). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-02425-2_10

Chatzis, S. P. (2011). A fuzzy c-means-type algorithm for clustering of data with mixed numeric and categorical attributes employing a probabilistic dissimilarity functional. *Expert Systems with Applications*, *38*(7), 8684–8689. https://doi.org/http://dx.doi.org/10.1016/j.eswa.2011.01.074

Chen, J.-Y., & He, H.-H. (2016). A fast density-based data stream clustering algorithm with cluster centers self-determined for mixed data. *Information Sciences*, *345*, 271–293. https://doi.org/https://doi.org/10.1016/j.ins.2016.01.071

Cheng, L., Wang, Y., & Ma, X. (2019). A Neural Probabilistic outlier detection method for categorical data. *Neurocomputing*, *365*, 325–335. https://doi.org/https://doi.org/10.1016/j.neucom.2019.07.069

Claveria, O., & Poluzzi, A. (2017). Positioning and clustering of the world's top tourist destinations by means of dimensionality reduction techniques for categorical data. *Journal of Destination Marketing & Management*, *6*(1), 22–32. https://doi.org/https://doi.org/10.1016/j.jdmm.2016.01.008

Coombes, C. E., Liu, X., Abrams, Z. B., Coombes, K. R., & Brock, G. (2021). Simulation-derived best practices for clustering clinical data. *Journal of Biomedical Informatics*, *118*, 103788. https://doi.org/https://doi.org/10.1016/j.jbi.2021.103788

de Carvalho, F. de A. T., Simões, E. C., Santana, L. V. C., & Ferreira, M. R. P. (2018). Gaussian kernel c-means hard clustering algorithms with automated computation of the width hyper-parameters. *Pattern Recognition*, *79*, 370–386. https://doi.org/https://doi.org/10.1016/j.patcog.2018.02.018

Dinh, D. T., Huynh, V. N., & Sriboonchitta, S. (2021). Clustering mixed numerical and categorical data with missing values. *Information Sciences*, *571*, 418–442. https://doi.org/10.1016/j.ins.2021.04.076

Dunn, J. C. (1974). Well-Separated Clusters and Optimal Fuzzy Partitions. *Journal of Cybernetics*, *4*(1), 95–104. https://doi.org/10.1080/01969727408546059

Feng, F., Cho, J., Pedrycz, W., Fujita, H., & Herawan, T. (2016). Soft set based association rule mining. *Knowledge-Based Systems*, *111*, 268–282. https://doi.org/https://doi.org/10.1016/j.knosys.2016.08.020

Feng, F., Wang, Q., Yager, R. R., R. Alcantud, J. C., & Zhang, L. (2020). Maximal association analysis using logical formulas over soft sets. *Expert Systems with*

Applications, *159*, 113557. https://doi.org/https://doi.org/10.1016/j.eswa.2020.113557

Ghosh, D., Singh, A., Shukla, K. K., & Manchanda, K. (2019). Extended Karush-Kuhn-Tucker condition for constrained interval optimization problems and its application in support vector machines. *Information Sciences*, *504*, 276–292. https://doi.org/https://doi.org/10.1016/j.ins.2019.07.017

Gibson, D., Kleinberg, J., & Raghavan, P. (2000). Clustering categorical data: an approach based on dynamical systems. *The VLDB Journal The International Journal on Very Large Data Bases*, *8*(3–4), 222–236. https://doi.org/10.1007/s007780050005

Gonçalves, G. M., & Lourenço, L. L. (2019). Mathematical formulations for the K clusters with fixed cardinality problem. *Computers & Industrial Engineering*, *135*, 593–600. https://doi.org/https://doi.org/10.1016/j.cie.2019.06.028

Gupta, A., & Das, S. (2022). On efficient model selection for sparse hard and fuzzy center-based clustering algorithms. *Information Sciences*, *590*, 29–44. https://doi.org/https://doi.org/10.1016/j.ins.2021.12.070

Gupta, U., & Rai, J. (2017). An Efficient Soft Set Approach for Association Rule Mining Using Constraints, *7*(6), 13097–13100.

Hadibarata, T., & Rubiyatno, R. (2019). Active learning strategies in environmental engineering course: A case study in curtin university Malaysia. *Jurnal Pendidikan IPA Indonesia*. https://doi.org/10.15294/jpii.v8i4.19169

He, Z., Deng, S., & Xu, X. (2005). Improving K-Modes Algorithm Considering Frequencies of Attribute Values in Mode BT - Computational Intelligence and Security. In Y. Hao, J. Liu, Y. Wang, Y. Cheung, H. Yin, L. Jiao, … Y.-C. Jiao (Eds.) (pp. 157–162). Berlin, Heidelberg: Springer Berlin Heidelberg.

Herawan, T., & Deris, M. M. (2009). On Multi soft Sets Construction in Information Systems BT - Emerging Intelligent Computing Technology and Applications. With Aspects of Artificial Intelligence. In D.-S. Huang, K.-H. Jo, H.-H. Lee, H.-J. Kang, & V. Bevilacqua (Eds.) (pp. 101–110). Berlin, Heidelberg: Springer Berlin Heidelberg.

Herawan, T., Deris, M. M., & Abawajy, J. H. (2010). Matrices Representation of Multi Soft-Sets and Its Application. In D. Taniar, O. Gervasi, B. Murgante, E. Pardede, & B. O. Apduhan (Eds.), *Computational Science and Its Applications -- ICCSA 2010: International Conference, Fukuoka, Japan, March 23-26, 2010,*

Proceedings, Part III (pp. 201–214). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-12179-1_19

Herrero, C., & Villar, A. (2021). Dealing with categorical data in a multidimensional context: The multidimensional balanced worth. *Social Science Research*, *96*, 102561. https://doi.org/https://doi.org/10.1016/j.ssresearch.2021.102561

Holden, M. P., & Hampson, E. (2021). Endogenous variation in estradiol in women affects the weighting of metric and categorical information in spatial location memory. *Hormones and Behavior*, *128*, 104909. https://doi.org/https://doi.org/10.1016/j.yhbeh.2020.104909

Hsu, C.-C., Chen, C.-L., & Su, Y.-W. (2007). Hierarchical clustering of mixed data based on distance hierarchy. *Information Sciences*, *177*(20), 4474–4492. https://doi.org/https://doi.org/10.1016/j.ins.2007.05.003

Huang, M. K. N. (1999). A fuzzy k-modes algorithm for clustering categorical data - Fuzzy Systems, IEEE Transactions on. *IEEE Trans. Fuzzy Syst.*, *7*(4), 446–452.

Huang, Z. (1998). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*, *2*(3), 283–304. https://doi.org/10.1023/A:1009769707641

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218. https://doi.org/10.1007/BF01908075

Husák, M., Komárková, J., Bou-Harb, E., & Čeleda, P. (2019). Survey of Attack Projection, Prediction, and Forecasting in Cyber Security. *IEEE Communications Surveys & Tutorials*, *21*(1), 640–660. https://doi.org/10.1109/COMST.2018.2871866

Irsadi, A., Anggoro, S., Soeprobowati, T. R., Helmi, M., & Khair, A. S. E. (2019). Shoreline and mangrove analysis along semarang-demak, Indonesia for sustainable environmental management. *Jurnal Pendidikan IPA Indonesia*. https://doi.org/10.15294/jpii.v8i1.17892

Jia, X., & Zhang, D. (2021). Prediction of maritime logistics service risks applying soft set based association rule: An early warning model. *Reliability Engineering & System Safety*, *207*, 107339. https://doi.org/https://doi.org/10.1016/j.ress.2020.107339

Jiang, S., Ferreira, J., & González, M. C. (2012). Clustering daily patterns of human activities in the city. *Data Mining and Knowledge Discovery*, *25*(3), 478–510. https://doi.org/10.1007/s10618-012-0264-z

Kar, S., & Dutta, I. (2021). Soft ideals of soft ternary semigroups. *Heliyon*, *7*(6), e07330. https://doi.org/https://doi.org/10.1016/j.heliyon.2021.e07330

Karali, Y., Lyngdoh, B., & Behera, H. (2013). Hard and Fuzzy Clustering Algorithms Using Normal Distribution of Data Points: a Comparative Performance Analysis, *2*(10), 320–328.

Karthick, S., Yuvaraj, N., Rajakumari, P. A., & Raja, R. A. (2021). Ensemble Similarity Clustering Frame work for Categorical Dataset Clustering Using Swarm Intelligence BT - Intelligent Computing and Applications. In S. S. Dash, S. Das, & B. K. Panigrahi (Eds.) (pp. 549–557). Singapore: Springer Singapore.

Khan, M. S., Mujtaba, G., Al-garadi, M. A., Friday, N. H., Waqas, A., & Qasmi, F. R. (2018). Multi soft sets-based decision making using rank and fix valued attributes. In *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)* (pp. 1–11). https://doi.org/10.1109/ICOMET.2018.8346380

Khandaker, S. M., Hussain, A., & Ahmed, M. (2019). Effectiveness of Hard Clustering Algorithms for Securing Cyber Space BT - Smart Grid and Internet of Things. In A.-S. K. Pathan, Z. M. Fadlullah, & M. Guerroumi (Eds.) (pp. 113–120). Cham: Springer International Publishing.

Kim, D.-W., Lee, K. H., & Lee, D. (2004). Fuzzy clustering of categorical data using fuzzy centroids. *Pattern Recognition Letters*, *25*(11), 1263–1271. https://doi.org/10.1016/j.patrec.2004.04.004

Kim, S.-K. (2022). 3.12 - A New Paradigm for Analysis of Categorical Data in Psychology: The Correspondence Analysis Approach. In G. J. G. B. T.-C. C. P. (Second E. Asmundson (Ed.) (pp. 176–188). Oxford: Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-12-818697-8.00221-1

Kong, Z., Zhao, J., Wang, L., & Zhang, J. (2021). A new data filling approach based on probability analysis in incomplete soft sets. *Expert Systems with Applications*, *184*, 115358. https://doi.org/https://doi.org/10.1016/j.eswa.2021.115358

Kottam, S., & Paul, V. (2020). Soft Set Theory-A Novel Soft Computing Tool for Data Mining. *International Journal of Applied Engineering …*, *15*(11), 1052–1058. Retrieved from https://www.ripublication.com/ijaer20/ijaerv15n11_02.pdf

Kuo, R. J., Zheng, Y. R., & Nguyen, T. P. Q. (2021). Metaheuristic-based possibilistic fuzzy k-modes algorithms for categorical data clustering. *Information Sciences*, *557*, 1–15. https://doi.org/https://doi.org/10.1016/j.ins.2020.12.051

Liu, Y., Rodríguez, R. M., Alcantud, J. C. R., Qin, K., & Martínez, L. (2019). Hesitant linguistic expression soft sets: Application to group decision making. *Computers & Industrial Engineering*, *136*, 575–590. https://doi.org/https://doi.org/10.1016/j.cie.2019.07.040

Malefaki, S., & Iliopoulos, G. (2007). Simulating from a multinomial distribution with large number of categories. *Computational Statistics and Data Analysis*, *51*(12), 5471–5476. https://doi.org/10.1016/j.csda.2007.03.022

Manna, S., Basu, T. M., & Mondal, S. K. (2020). A soft set based VIKOR approach for some decision-making problems under complex neutrosophic environment. *Engineering Applications of Artificial Intelligence*, *89*, 103432. https://doi.org/https://doi.org/10.1016/j.engappai.2019.103432

McLachlan, G. J., Rathnayake, S. I., & Lee, S. X. (2020). 2.24 - Model-Based Clustering☆. In S. Brown, R. Tauler, & B. B. T.-C. C. (Second E. Walczak (Eds.) (pp. 509–529). Oxford: Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-12-409547-2.14649-9

Mehta, V., Bawa, S., & Singh, J. (2020). Analytical review of clustering techniques and proximity measures. *Artificial Intelligence Review*, *53*(8), 5995–6023. https://doi.org/10.1007/s10462-020-09840-7

Molodtsov, D. (1999). Soft set theory—First results. *Computers & Mathematics with Applications*, *37*(4–5), 19–31. https://doi.org/10.1016/S0898-1221(99)00056-5

Molodtsov, Dmitriy. (1999). Soft set theory—first results. *Computers & Mathematics with Applications*, *37*(4–5), 19–31.

Morris, D. S., Raim, A. M., & Sellers, K. F. (2020). A Conway–Maxwell-multinomial distribution for flexible modeling of clustered categorical data. *Journal of Multivariate Analysis*, *179*, 104651. https://doi.org/https://doi.org/10.1016/j.jmva.2020.104651

Mosia, S. J., & Joubert, A. (2020). Primary healthcare practice learning environment: A description of students' perspectives. *International Journal of Africa Nursing Sciences*, *13*, 100230. https://doi.org/https://doi.org/10.1016/j.ijans.2020.100230

Mrudula, K., & Reddy, E. K. (2017). Hard And Fuzzy Clustering Methods : A Comparative Study. *Ieee*, (April). Retrieved from https://www.researchgate.net/profile/Mrudula-K-2/publication/319526078_Hard_And_Fuzzy_Clustering_Methods_A_Comparat

ive_Study/links/5caed8ad299bf120975d7797/Hard-And-Fuzzy-Clustering-Methods-A-Comparative-Study.pdf

Mrudula, K., & Reddy, E. K. (2019). Hard And Fuzzy Clustering Methods : A Comparative Study Hard and Fuzzy Clustering Methods : A Comparative Study, (April).

Naouali, S., Ben Salem, S., & Chtourou, Z. (2020). *Clustering categorical data: A survey*. *International Journal of Information Technology and Decision Making* (Vol. 19). https://doi.org/10.1142/S0219622019300064

Ng, M. K., Li, M. J., Huang, J. Z., & He, Z. (2007). On the Impact of Dissimilarity Measure in k-Modes Clustering Algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(3), 503–507. https://doi.org/10.1109/TPAMI.2007.53

Nooraeni, R., Arsa, M. I., & Kusumo Projo, N. W. (2021). Fuzzy Centroid and Genetic Algorithms: Solutions for Numeric and Categorical Mixed Data Clustering. *Procedia Computer Science*, *179*, 677–684. https://doi.org/https://doi.org/10.1016/j.procs.2021.01.055

Oyelade, J., Isewon, I., Oladipupo, O., Emebo, O., Omogbadegun, Z., Aromolaran, O., … Olawole, O. (2019). Data Clustering: Algorithms and Its Applications. In *2019 19th International Conference on Computational Science and Its Applications (ICCSA)* (pp. 71–81). https://doi.org/10.1109/ICCSA.2019.000-1

Pantsar, M. (2021). Cognitive and Computational Complexity: Considerations from Mathematical Problem Solving. *Erkenntnis*, *86*(4), 961–997. https://doi.org/10.1007/s10670-019-00140-3

Pardasani, B. (2018). Multi Softset for Decision Making. *International Journal of Science and Research (IJSR)*, *7*(11), 55–56. https://doi.org/10.21275/ART20192036

Parmar, D., Wu, T., & Blackhurst, J. (2007). MMR: An algorithm for clustering categorical data using Rough Set Theory. *Data & Knowledge Engineering*, *63*(3), 879–893. https://doi.org/http://dx.doi.org/10.1016/j.datak.2007.05.005

Paul, R., & Hoque, A. S. M. L. (2010). Clustering medical data to predict the likelihood of diseases. In *2010 Fifth International Conference on Digital Information Management (ICDIM)* (pp. 44–49). https://doi.org/10.1109/ICDIM.2010.5664638

Perrone, D., Aiello, M. A., Pecce, M., & Rossi, F. (2015). Rapid visual screening for

seismic evaluation of RC hospital buildings. *Structures*, *3*, 57–70. https://doi.org/10.1016/j.istruc.2015.03.002

Phillips, P., & Lee, I. (2011). Crime analysis through spatial areal aggregated density patterns. *GeoInformatica*, *15*(1), 49–74. https://doi.org/10.1007/s10707-010-0116-1

Pinheiro, D. N., Aloise, D., & Blanchard, S. J. (2020). Convex fuzzy k-medoids clustering. *Fuzzy Sets and Systems*, *389*, 66–92. https://doi.org/https://doi.org/10.1016/j.fss.2020.01.001

Rahman, M. S., Shaikh, A. A., & Bhunia, A. K. (2020). Necessary and sufficient optimality conditions for non-linear unconstrained and constrained optimization problem with interval valued objective function. *Computers & Industrial Engineering*, *147*, 106634. https://doi.org/https://doi.org/10.1016/j.cie.2020.106634

Saedudin, R. R., Kasim, S. B., Mahdin, H., & Hasibuan, M. A. (2017). Soft Set Approach for Clustering Graduated Dataset BT - Recent Advances on Soft Computing and Data Mining. In T. Herawan, R. Ghazali, N. M. Nawi, & M. M. Deris (Eds.) (pp. 631–637). Cham: Springer International Publishing.

Saha, I., Sarkar, J. P., & Maulik, U. (2019). Integrated Rough Fuzzy Clustering for Categorical data Analysis. *Fuzzy Sets and Systems*, *361*, 1–32. https://doi.org/https://doi.org/10.1016/j.fss.2018.02.007

Sarmah, T., & Das, S. (2018). Earthquake Vulnerability Assessment for RCC Buildings of Guwahati City using Rapid Visual Screening. In *Procedia Engineering* (Vol. 212, pp. 214–221). Elsevier Ltd. https://doi.org/10.1016/j.proeng.2018.01.028

Saxena, A., & Singh, M. (2016). Using Categorical Attributes for Clustering. *International Journal of Scientific Engineering and Applied Science (IJSEAS)*, (2), 324–329.

Scott, A. J., & Symons, M. J. (1971). Clustering Methods Based on Likelihood Ratio Criteria. *Biometrics*, *27*(2), 387–397. https://doi.org/10.2307/2529003

Singh, S., & Srivastava, S. (2020). Review of Clustering Techniques in Control System: Review of Clustering Techniques in Control System. *Procedia Computer Science*, *173*, 272–280. https://doi.org/https://doi.org/10.1016/j.procs.2020.06.032

Soppari, K., & Chandra, N. S. (2020). Development of improved whale optimization-

based FCM clustering for image watermarking. *Computer Science Review*, *37*, 100287. https://doi.org/https://doi.org/10.1016/j.cosrev.2020.100287

Sutoyo, E., Yanto, I. T. R., Saadi, Y., Chiroma, H., Hamid, S., & Herawan, T. (2019). *A Framework for Clustering of Web Users Transaction Based on Soft Set Theory*. *Lecture Notes in Electrical Engineering* (Vol. 520). https://doi.org/10.1007/978-981-13-1799-6_32

Sutoyo, Edi, Yanto, I. T. R., Saadi, Y., Chiroma, H., Hamid, S., & Herawan, T. (2019). A Framework for Clustering of Web Users Transaction Based on Soft Set Theory (pp. 307–314). https://doi.org/10.1007/978-981-13-1799-6_32

Symons, M. J. (1981). Clustering Criteria and Multivariate Normal Mixtures. *Biometrics*, *37*(1), 35–43. https://doi.org/10.2307/2530520

Tao, Z., Shao, Z., Liu, J., Zhou, L., & Chen, H. (2020). Basic uncertain information soft set and its application to multi-criteria group decision making. *Engineering Applications of Artificial Intelligence*, *95*, 103871. https://doi.org/https://doi.org/10.1016/j.engappai.2020.103871

Thrun, M. C., & Stier, Q. (2021). Fundamental clustering algorithms suite. *SoftwareX*, *13*, 100642. https://doi.org/https://doi.org/10.1016/j.softx.2020.100642

Traylor, R. (2017). A Generalized Multinomial Distribution from Dependent Categorical Random Variables.

Tri, I., Yanto, R., Apriani, A., Hidayat, R., Mat, M., & Senan, N. (2021). Fast Clustering Environment Impact using Multi Soft Set Based on Multivariate Distribution. *JOIV : International Journal on Informatics Visualization*, *5*(September), 291–297. https://doi.org/http://dx.doi.org/10.30630/joiv.5.3.628

Vardhan, A., Sarmah, P., & Das, A. (2020). A Comprehensive Analysis of the Most Common Hard Clustering Algorithms BT - Inventive Computation Technologies. In S. Smys, R. Bestak, & Á. Rocha (Eds.) (pp. 48–58). Cham: Springer International Publishing.

Vazirani, U. (2000). Quantum computing and quantum complexity theory. In *2000 IEEE International Symposium on Circuits and Systems (ISCAS)* (Vol. 1, pp. 737–739 vol.1). https://doi.org/10.1109/ISCAS.2000.857201

Wei, W., Liang, J., Guo, X., Song, P., & Sun, Y. (2019). Hierarchical division clustering framework for categorical data. *Neurocomputing*, *341*, 118–134. https://doi.org/https://doi.org/10.1016/j.neucom.2019.02.043

Wen, Q., & Celebi, M. E. (2011). Hard versus fuzzy c-means clustering for color

quantization. *EURASIP Journal on Advances in Signal Processing*, *2011*(1), 1–12. https://doi.org/10.1186/1687-6180-2011-118

Wen, T.-C., Chang, K.-H., & Lai, H.-H. (2020). Integrating the 2-tuple linguistic representation and soft set to solve supplier selection problems with incomplete information. *Engineering Applications of Artificial Intelligence*, *87*, 103248. https://doi.org/https://doi.org/10.1016/j.engappai.2019.103248

Woodbury, M. A., & Clive, J. (1974). Clinical Pure Types as a Fuzzy Partition. *Journal of Cybernetics*, *4*(3), 111–121. https://doi.org/10.1080/01969727408621685

Wu, C., & Zhang, X. (2020). Total Bregman divergence-based fuzzy local information C-means clustering for robust image segmentation. *Applied Soft Computing*, *94*, 106468. https://doi.org/https://doi.org/10.1016/j.asoc.2020.106468

Wu, T., Zhou, Y., Xiao, Y., Needell, D., & Nie, F. (2019). Modified fuzzy clustering with segregated cluster centroids. *Neurocomputing*, *361*, 10–18. https://doi.org/https://doi.org/10.1016/j.neucom.2019.07.005

Yang, J., & Yao, Y. (2020). Semantics of soft sets and three-way decision with soft sets. *Knowledge-Based Systems*, *194*, 105538. https://doi.org/https://doi.org/10.1016/j.knosys.2020.105538

Yang, M.-S., Chiang, Y.-H., Chen, C.-C., & Lai, C.-Y. (2008). A fuzzy k-partitions model for categorical data and its comparison to the GoM model. *Fuzzy Sets and Systems*, *159*(4), 390–405. https://doi.org/https://doi.org/10.1016/j.fss.2007.08.012

Yang, M. S., Chiang, Y. H., Chen, C. C., & Lai, C. Y. (2008). A fuzzy k-partitions model for categorical data and its comparison to the GoM model. *Fuzzy Sets and Systems*, *159*(4), 390–405. https://doi.org/10.1016/j.fss.2007.08.012

Yanto, I.T.R., Rahman, A., & Saaadi, Y. (2017). Soft Maximal Association Rule for web user mining. In *Proceeding - 2016 2nd International Conference on Science in Information Technology, ICSITech 2016: Information Science for Green Society and Environment*. https://doi.org/10.1109/ICSITech.2016.7852659

Yanto, I T R, Rahman, A., & Saaadi, Y. (2016). Soft Maximal Association Rule for web user mining. In *2016 2nd International Conference on Science in Information Technology (ICSITech)* (pp. 339–343). https://doi.org/10.1109/ICSITech.2016.7852659

Yanto, Iwan Tri Riyadi, Ismail, M. A., & Herawan, T. (2016). A modified Fuzzy k-

Partition based on indiscernibility relation for categorical data clustering. *Engineering Applications of Artificial Intelligence*, *53*, 41–52. https://doi.org/10.1016/j.engappai.2016.01.026

Zhu, S., & Xu, L. (2018). Many-objective fuzzy centroids clustering algorithm for categorical data. *Expert Systems with Applications*, *96*, 230–248. https://doi.org/https://doi.org/10.1016/j.eswa.2017.12.013

VITA

The author was born on June 14, 1985, in the small city of Prambanan, Klaten, Central Java, Indonesia. He completed his primary and secondary schooling in the same city and moved to largest city of the country, Yogyakarta Indonesa in mathematic from Universitas Ahmad Dahlan (UAD), Yogyakarta, Indonesia. He graduated from Universitas Ahmad Dahlan in the year of 2009. He awarded Master in Information Technology (by Research) from Universiti Tun Hussein Onn Malaysia, Johor, Malaysia in 2011. His master research area includes Data Mining, KDD, and Real Analysis. Upon graduation, she worked as Lecturer in Universitas Ahmad Dahlan. In 2019, He decided to advance her educational level at Universiti Tun Hussein Onn Malaysia, where she was awarded-Doctor of Philosophy of Information Technology.