Soft Set Multivariate Distribution for Categorical Data Clustering

By IWAN TRI RIYADI

Soft Set Multivariate Distribution for Categorical Data Clustering

Iwan Tri Riyadi Yanto[1,4], Rohmat Saedudin[2], Mustafa Mat Deris[2,3], Norhalina Senan[4]

[1] Department of Information Systems, University of Ahmad Dahlan,
Yogyakarta, Indonesia
email: yanto.itr@is.uad.ac.id
[2]Department of Information Systems,
Telkom University,
Bandung, West Java, Indonesia, 40257
[3]Faculty of Applied Science and Technology,
[4]Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor
email: {mmustafa,halina}@uthm.edu.my

Abstract— **Clustering categorical data is more challenging than numerical data since there is no inherent distance measure between categorical objects. In the categorical data, a standard parametric model used in latent class clustering is independent product of multinomial distributions. Meanwhile, multi-valued attributes on the categorical data can be represented without binary values as a multi soft set. In this paper, we proposed a clustering technique based on soft set theory for categorical data via multinomial distribution. The data will be represented as a multi soft set which is every soft set have its probability to be a member of the cluster. The data with highest probability will be assigned as the member of the cluster. The experiment of the proposed technique is evaluated based on the Dunn index with respect to the number of clusters and response time. The experiment results show that the proposed technique has lowest response time with high stability as compared to baseline techniques.**

Keywords— **Clustering; Categorical data; Soft set ; Multivariate.**

I. INTRODUCTION

Clustering is a fundamental problem that frequently arises in a broad variety of fields such as pattern recognition, image processing, machine learning and statistics [1]–[4]. It can be defined as a process of partitioning a given data set of multiple variables into groups [5], [6]. The k-means algorithm is the most popular among clustering algorithms developed to date because of its effectiveness and efficiency in clustering large data sets [7]. However, k-means clustering algorithm fails to handle data sets with categorical variables because it can only minimize a numerical cost function. Thus, clustering categorical data is more challenging than numerical data since there is no inherent distance measure between categorical objects [8]. Clustering algorithms developed for managing numerical data cannot directly be used to cluster categorical data. To address this deficiency, several clustering algorithms have been developed to deal with categorical data. As a result, Huang [9] proposed the k-modes clustering method that removes the numeric-only limitation of the k-means algorithm. Since then major improvements have been made in k-modes algorithms including new dissimilarity measures to the k-modes clustering [10]–[12] and a fuzzy set based k-modes algorithm [13].To improve the efficiency of fuzzy k-modes, Kim et al. [10] proposed a technique called fuzzy centroids approach. However, almost all fuzzy categorical data clustering algorithms mentioned above represent data set in the binary values. Thus, the issue with the aforesaid approaches is that they tend to have high computational time and low clusters purity. This indicates that an approach that does not suffer from high computational time and low clusters purity is needed. In this paper, we propose the clustering

technique based on soft set theory for categorical data via multinomial distribution. The categorical data can be represented as multi-valued information system (multi soft set) [14]. It can be following random sample from multivariate multinomial distribution. For multivariate categorical data, a standard parametric model used in latent class clustering is a locally (i.e., within-clusters) independent product of multinomial distributions [15]. Moreover, the multi-valued information system can be represented the categorical data as a soft set [14] without representing in the binary values.

II. Related Works

Recently, fuzzy-based clustering has been widely focused by many scholars and some significant results have been achieved in the theoretical and practical aspects. Huang [9] proposed the k-modes clustering method that removes the numeric-only limitation of the k-means algorithm. Since then major improvements have been made in k-modes algorithms including new dissimilarity measures to the k-modes clustering and a fuzzy set based k-modes algorithm [6-8]. Lets μ be a membership function, y is a data and v is a centroid of cluster, the objective function of fuzzy k-mode is to minimize the function $H_m(\mu, v)$.

$$H_m(\mu, v) = \sum_{i=1}^{I} \sum_{k=1}^{K} \mu_{ik}^m d(y_i, v_k),$$

subject to
$\sum_{k=1}^{K} \mu_{ik} = 1$, for $i = 1,2, \dots, I$,

where $d(y_i, v_k) = \sum_{j=1}^{J} \delta(y_{ij}, v_{kj})$ is called the simple matching dissimilarity measure, $\delta(y_{ij}, v_{kj}) = 0$ if $y_{ij} = v_{kj}$ and $\delta(y_{ij}, v_{kj}) = 1$, if $y_{ij} \neq v_{kj}$. y is the categorical data value and v is the cluster centroid. where m is the fuzziness index. The update equations for hard k-modes are as follow:

$$\mu_{ik} = \begin{cases} 1 & if \quad d(y_i, v_k) = \min_{1 \le k' \le K} d(y_i, v_k') \\ 0 & \text{otherwise} \end{cases}$$

$$v_{kjl} = \begin{cases} 1 & if \quad \sum_{i=1}^{I} \mu_{ik} y_{ijl} = \max_{1 \le l' \le L} \sum_{i=1}^{I} \mu_{ik} y_{ijl}, \\ 0 & \text{otherwise} \end{cases}$$

Huang extended the hard k-mode to fuzzy k-modes [9]. Thus, the solution is given by update the equation as follows:

$$\mu_{ik} = 1 / \sum_{k'=1}^{K} \left[\frac{d(y_i, v_k)}{d(y_i, v_{k'})} \right]^{\frac{1}{m-1}}$$

$$v_{kjl} = \begin{cases} 1 & if \quad \sum_{i=1}^{I} \mu_{ik}^m y_{ijl} = \max_{1 \le l' \le L} \sum_{i=1}^{I} \mu_{ik}^m y_{ijl}, \\ 0 & \text{otherwise} \end{cases}$$

The use of hard centroids can give rise to the artifacts. For example, although the Fuzzy k-modes algorithm efficiently handles categorical data sets, it uses a hard centroid representation for categorical data in a cluster. The use of hard rejection of data can lead to misclassification in the region of doubt [12].

Kim et al. [17] improved the performance of fuzzy k-modes by changing hard centroids to fuzzy centroid with $\tilde{v}_{kj} = \left(\tilde{v}_{kj1}, \dots, \tilde{v}_{kjL_j} \right)$, for $k = 1,2, \dots K$ and $j = 1,2, \dots, J$, where $\tilde{v}_{kjl} \in [0,1]$ and $\sum_{i=1}^{L_j} \tilde{v}_{kjl} = 1$. The minimize objective function of fuzzy centroid is as follows:

$$H_m(\mu, v) = \sum_{i=1}^{I} \sum_{k=1}^{K} \mu_{ik}^m d(y_i, \tilde{v}_k),$$

subject to

$$\sum_{k=1}^{K} \mu_{ik} = 1, i = 1,2, \dots, I.$$

$$\sum_{l=1}^{L_j} \tilde{v}_{kjl} = 1.$$

The distance measure with the centroid updates equation which is given as following equation:

$$d(y_i, \tilde{v}_k) = \sum_{j=1}^{J} \delta(y_{ij}, \tilde{v}_{kj}) = \sum_{j=1}^{J} \sum_{l=1}^{L_j} (1 - y_{ijl}) \tilde{v}_{kjl},$$

$$\tilde{v}_{kjl} = \frac{\sum_{i=1}^{I} \mu_{ik}^m \cdot y_{ijl}}{\sum_{i=1}^{I} \mu_{ik}^m}.$$

The update equation of memberships can be obtained as follows:

$$\mu_{ik} = 1 / \sum_{k'=1}^{K} \left[\frac{d(y_i, \tilde{v}_k)}{d(y_i, \tilde{v}_{k'})} \right]^{\frac{1}{m-1}}.$$

Both of the Fuzzy k-modes with hard centroid and fuzzy centroid are non-parametric techniques. The algorithms use the dissimilarity functional based on the least total within cluster matching dissimilarity.

III. Proposed Technique

The proposed technique uses soft set to represent the data. The data which have the same value are decompose into multi soft set. Since the value of each soft set is the same, thus all members in each soft set have same probability to be assigned on the cluster. Thus, we will find the high probability of each instance with respect to all parameters on the data using multinomial distribution function.

Definition 1. Let U be an universe set, E be a set of parameters $A \subset E$, F is the function that mapping parameter A into the set of all subsets of the set U as

$$F: A \to P(U).$$

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-aproximate elements of (F, A).

Definition 2. Let $S = (U, A, V, f)$ be a categorical-valued information system , where $U = \{u_1, u_2, \dots, U_n\}$ is finite set of instance, $A = \{a_1, a_2, \dots a_m\}$ is finite set of attribute, V is values set of each attribute A , f is maping function $f: (U, A) \to V$ and $S = (U, a_i, V_{a_i}, f), i = 1,2, \cdots, |A|$ Boolean-valued information system, it can be decomposed to be multi-boolean information system as

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

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

Consider to the pair (F, A), assign to multi-soft set over U, representing a categorical-valued information system $S = (U, A, V, f)$, where $(F, a_1), \cdots, (F, a_{|A|}) \subseteq (F, A)$ and $\left(F, a_{j_1}\right), \cdots, \left(F, a_{j_{|a_j|}}\right) \subseteq (F, a_j)$. Lets say $\lambda_{kjl}^{u_i}$ be a probability of $u_i \in (F, a_{jl})$ into cluster $C_k, k = 1,2, \dots, K$, where $i = 1,2, \dots, |U|, j = 1,2, \dots, |A|$ and $l = 1,2, \dots,$ thus, the multivariate multinomial distribution of multi soft set can be defined as

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

Subject to

$$\sum_{k=1}^{K} z_{ik} = 1, for\ i = 1,2, \dots, |U|.$$

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

The maximization of the objective function $L_{CML}(z, \lambda)$ can be obtained by updating the equation as follows:

$$\lambda_{kjl} = \frac{\sum_{u_i \in (F, a_{jl})} z_{ik}(u_i)}{|U|}$$

$$z_{ik} = \begin{cases} 1 & if \quad \sum_{j=1}^{|A|} \ln \lambda_{kjl}^{u_i} = \max_{1 \le k' \le K} \sum_{j=1}^{|A|} \ln \lambda_{k'jl}^{u_i} \\ 0 & otherwise \end{cases}$$

IV. Results and Discussion

This section presents the validation of the proposed algorithms using some benchmarks datasets form UCI machine learning. The experiments are conducted on a PC with Intel i5-8400six core CPU 2.8 GHz and 2 GB RAM using MATLAB programming language. The experiments are conducted to compare the proposed technique with the baseline technique which are fuzzy centroid and fuzzy k-partition. Fuzzy centroid uses simple matching dissimilarity function to measure the distance of centroid. It is non parametric approach, but it is distance based which is to find the best centroid where the member cluster is determined by the closest centroid. Thus, it produces a spherical cluster. It can be low of purity. Meanwhile, the fuzzy k-partition is parametric approach which is depends on the likelihood function of multinomial distribution function. However, the data must be represented as binary variable. So, it tends to produce high computational time. We elaborate the three approaches through the UCI benchmark datasets as follow:

- Zoo data set which is comprised of 101 objects, where each data point represents information of an animal in terms of 18 categorical variables.
- Balloon dataset which contains 20 instances and 4 categorical variables.
- Monk dataset which contains 432 instances and 6 variables.
- Spect dataset which contains 187 instances and 922 variables.
- Breast dataset which contains 683 instances and 9 variables.

All the comparisons made in this section are evaluated based on the Dunn index with respect to number of clusters and response time. The experiments are set for all techniques algorithm form 2 to 100 number of clusters or maximum number of instance. Each technique runs for 20 times. The technique called divergent if the data is clustered into number of cluster set or maximum number of instance, it means that the number of member of cluster is only one. Obviously, all data is clustered into one cluster when convergent approach to 1. Table 1 shows that the proposed technique can reduce the response times up to 92.50 in average. Meanwhile, Table 2 summarizes the stability with respect to the number of cluster. It shows that the proposed technique have better stability compared to the baseline techniques. An example to illustrate the Dunn index for stability and number of cluster created on the balloon data set is given in Fig. 1 and Fig. 2.

TABLE 1.

RESPONSE TIMES FOR DIFFERENT DATASETS

Data set	Response Times			Improvement (%)
	FC	FkP	Proposed	
Zoo	0.8732	0.2617	0.0236	90.98
Balloon	0.6914	1.2404	0.0273	97.80
Monk	0.9206	0.3754	0.0253	93.26
Spect	0.5662	0.4645	0.0995	78.58
Average	0.7629	0.5855	0.0439	92.50

TABLE 2.

STABILITY COMPARISON BASED ON NUMBER OF CLUSTERS

Data	Size of data	Number of cluster created		
		fc	fkp	proposed technique
Balloon	(20,4)	Divergent	Convergent to 1	Convergent to 9-10
Breast	(683,9)	Convergent ke 4	Divergent	Convergent to 80-85
Monk	(432,6)	Divergent	Divergent	Convergent to 70
Spect	(187,22)	Divergent	Convergent to 1	Convergent to 45-49
Zoo	(101,16)	Convergent 59	Convergent to 1	Convergent to 25-29

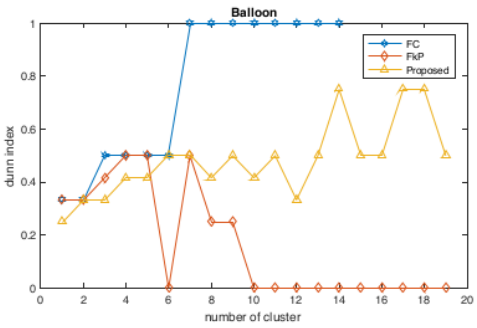

Fig. 1 The Dunn Index of Balloon Data Set

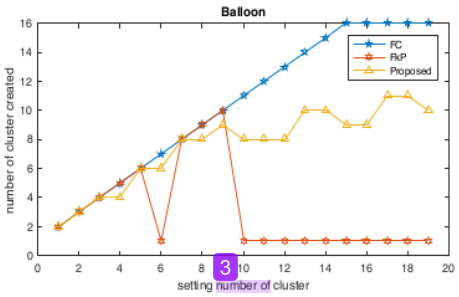

Fig. 2 The number of cluster created by given maximum number cluster setting

V. CONCLUSIONS

In this paper, the technique of soft set for categorical data clustering via multinomial distribution function is presented. For experimental investigation, the benchmarks datasets taken from UCI are used to compare the proposed technique with the existing techniques in terms of Dunn index and computational time respect to number of cluster. Based on the experiments that have been carried out in the five benchmark datasets, the results obtained show that the proposed technique achieves lower computational time and have better stability number of cluster. It can give recommendation of maximum number of cluster in implementation on the real data.

REFERENCES

[1] [1] C. Wan, M. Ye, C. Yao, and C. Wu, "Brain MR image segmentation based on Gaussian filtering and improved FCM clustering algorithm," in 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2017, pp. 1–5.

[2] [2] R. Shanker and M. Bhattacharya, "Brain Tumor Segmentation of Normal and Pathological Tissues Using K-mean Clustering with Fuzzy C-mean Clustering," in VipIMAGE 2017, 2018, pp. 286–296.

[3] [3] A. S. M. S. Hossain, "Customer segmentation using centroid based and density based clustering algorithms," in 2017 3rd International Conference on Electrical Information and Communication Technology (EICT), 2017, pp. 1–6.

[4] [4] K. V Ahammed Muneer and K. Paul Joseph, "Performance Analysis of Combined k-mean and Fuzzy-c-mean Segmentation of MR Brain Images," in Computational Vision and Bio Inspired Computing, 2018, pp. 830–836.

[5] [5] H. Zhou, "K-Means Clustering BT - Learn Data Mining Through Excel: A Step-by-Step Approach for Understanding Machine Learning Methods," H. Zhou, Ed. Berkeley, CA: Apress, 2020, pp. 35–47.

[6] [6] S. Irfan, G. Dwivedi, and S. Ghosh, "Optimization of K-means clustering using genetic algorithm," in 2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN), 2017, pp. 156–161.

[7] [7] B. K. D. Prasad, B. Choudhary, and B. Ankayarkanni., "Performance Evaluation Model using Unsupervised K-Means Clustering," in 2020 International Conference on Communication and Signal Processing (ICCSP), 2020, pp. 1456–1458.

[8] [8] W. Wei, J. Liang, X. Guo, P. Song, and Y. Sun, "Hierarchical division clustering framework for categorical data," Neurocomputing, vol. 341, pp. 118–134, 2019.

[9] [9] Z. Huang, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values," Data Min. Knowl. Discov., vol. 2, no. 3, pp. 283–304, 1998.

[10] [10]Y. Xiao, C. Huang, J. Huang, I. Kaku, and Y. Xu, "Optimal mathematical programming and variable neighborhood search for k-modes categorical data clustering," Pattern Recognit., vol. 90, pp. 183–195, 2019.

[11] [11]D. B. M. Maciel, G. J. A. Amaral, R. M. C. R. de Souza, and B. A. Pimentel, "Multivariate fuzzy k-modes algorithm," Pattern Anal. Appl., vol. 20, no. 1, pp. 59–71, 2017.

[12] [12]P. S. Bishnu and V. Bhattacherjee, "Software cost estimation based on modified K-Modes clustering Algorithm," Nat. Comput., vol. 15, no. 3, pp. 415–422, 2016.

[13] [13]Z. Huang and M. K. Ng, "A fuzzy k-modes algorithm for clustering categorical data," IEEE Trans. Fuzzy Syst., vol. 7, no. 4, pp. 446–452, 1999.

[14] [14]T. Herawan and M. M. Deris, "On Multi-soft Sets Construction in Information Systems BT - Emerging Intelligent Computing Technology and Applications. With Aspects of Artificial Intelligence," 2009, pp. 101–110.

[15] [15]D. S. Morris, A. M. Raim, and K. F. Sellers, “A Conway–Maxwell-multinomial distribution for flexible modeling of clustered categorical data,” J. Multivar. Anal., vol. 179, p. 104651, 2020.
[16] [16]I. T. R. Yanto, M. A. Ismail, and T. Herawan, “A modified Fuzzy k-Partition based on indiscernibility relation for categorical data clustering,” Eng. Appl. Artif. Intell., vol. 53, pp. 41–52, Aug. 2016.
[17] [17]D.-W. Kim, K. H. Lee, and D. Lee, “Fuzzy clustering of categorical data using fuzzy centroids,” Pattern Recognit. Lett., vol. 25, no. 11, pp. 1263–1271, Aug. 2004.

Soft Set Multivariate Distribution for Categorical Data Clustering

ORIGINALITY REPORT

20%

SIMILARITY INDEX

PRIMARY SOURCES

1	Wei Wei, Jiye Liang, Xinyao Guo, Peng Song, Yijun Sun. "Hierarchical division clustering framework for categorical data", Neurocomputing, 2019 Crossref	48 words — 2%
2	epdf.pub Internet	43 words — 2%
3	Lecture Notes in Computer Science, 2010. Crossref	32 words — 1%
4	Jiadong Ren. "A New Method of Software Security Checking Based on Similar Feature Tree", 2009 First International Conference on Information Science and Engineering, 12/2009 Crossref	32 words — 1%
5	fs.unm.edu Internet	30 words — 1%
6	Laura Anderlucci, Christian Hennig. "The Clustering of Categorical Data: A Comparison of a Model-based and a Distance-based Approach", Communications in Statistics - Theory and Methods, 2014 Crossref	25 words — 1%
7	nssdcftp.gsfc.nasa.gov Internet	24 words — 1%
8	Iwan Tri Riyadi Yanto, Rd Rohmat Saedudin, Saima Anwar Lashari, Haviluddin. "Chapter 25 A Numerical	24 words — 1%

Classification Technique Based on Fuzzy Soft Set Using Hamming Distance", Springer Science and Business Media LLC, 2018
Crossref

9 Dedy Hartama, Iwan Tri Riyadi Yanto, Muhammad Zarlis. "A soft set approach for fast clustering attribute selection", 2016 International Conference on Informatics and Computing (ICIC), 2016 — 20 words — 1%
Crossref

10 www.warse.org — 20 words — 1%
Internet

11 Rabiei Mamat, Ahmad Shukri Mohd Noor, Tutut Herawan, Mustafa Mat Deris. "Chapter 1 Cluster Validation Analysis on Attribute Relative of Soft-Set Theory", Springer Science and Business Media LLC, 2017 — 17 words — 1%
Crossref

12 Arkajyoti Saha, Swagatam Das. "Categorical fuzzy k-modes clustering with automated feature weight learning", Neurocomputing, 2015 — 14 words — 1%
Crossref

13 Zhi Kong. "Two Cases Based on Normal Parameter Reduction in Soft Sets", 2012 International Conference on Computer Science and Electronics Engineering, 03/2012 — 13 words — 1%
Crossref

14 Lecture Notes in Computer Science, 2009. — 13 words — 1%
Crossref

15 lambda.gsfc.nasa.gov — 10 words — < 1%
Internet

16 Tutut Herawan, Iwan Tri Riyadi Yanto, Mustafa Mat Deris. "Chapter 63 SMARViz: Soft Maximal Association Rules Visualization", Springer Science and Business Media LLC, 2009 — 10 words — < 1%
Crossref

17 "Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng- — 9 words — < 1%

2013)", Springer Science and Business Media LLC, 2014
Crossref

18	"Recent Developments in the Numerics of Nonlinear Hyperbolic Conservation Laws", Springer Science and Business Media LLC, 2013 Crossref	9 words — < 1%
19	Cao, F.. "A dissimilarity measure for the k-Modes clustering algorithm", Knowledge-Based Systems, 201202 Crossref	9 words — < 1%
20	www.asrojournal-sttal.ac.id Internet	9 words — < 1%
21	Yiyong Xiao, Changhao Huang, Jiaoying Huang, Ikou Kaku, Yuchun Xu. "Optimal mathematical programming and variable neighborhood search for k-modes categorical data clustering", Pattern Recognition, 2019 Crossref	9 words — < 1%
22	Constantinos Stylianou, Andreas S. Andreou. "A Hybrid Software Component Clustering and Retrieval Scheme Using an Entropy-Based Fuzzy k-Modes Algorithm", 19th IEEE International Conference on Tools with Artificial Intelligence(ICTAI 2007), 2007 Crossref	9 words — < 1%
23	www.jiyeliang.net Internet	8 words — < 1%
24	Sami Naouali, Semeh Ben Salem, Zied Chtourou. "Clustering Categorical Data: A Survey", International Journal of Information Technology & Decision Making, 2020 Crossref	8 words — < 1%
25	"Pattern Recognition and Machine Intelligence", Springer Science and Business Media LLC, 2013 Crossref	6 words — < 1%

EXCLUDE QUOTES	ON	EXCLUDE MATCHES	OFF
EXCLUDE BIBLIOGRAPHY	ON		