Proceedings of the 5th International Conference on Statistics: Theory and Applications (ICSTA'23)
Brunel University, London, United Kingdom – August 03-05, 2023
Paper No. 115
DOI: 10.11159/icsta23.115

Fuzzy *C*-Medoids Clustering Of CUB Data to Handle Likert-Type Scales Uncertainty

Elena Barzizza[1], Nicolò Biasetton[1], Riccardo Ceccato[1], Marta Disegna[1], Alberto Molena[1,]
[1]Department of Management and Engineering, University of Padova
Stradella S. Nicola, 3, 36100 Vicenza, Italy
elena.barzizza@phd.unipd.it; nicolo.biasetton@phd.unipd.it; riccardo.ceccato@unipd.it; marta.disegna@unipd.it; alberto.molena1@phd.unipd.it

Abstract - In investigating customer satisfaction (CS) with products or services the most popular approach still relies on interviews or questionnaire to obtain consumer's opinions and responses are usually measured by means of Likert-type scales. However, Likert-type data are inherently imprecise and uncertain. Thus, to obtain reliable analysis using such data, an a-posteriori correction must be adopted. Fuzzification procedure is the most common a-posteriori way to deal with uncertainty of Likert-type data, but it requires the subjective definition of the membership function. To overcome this problem, while a cluster analysis is conducted using Likert-type data as segmentation variables, a new approach based on CUB model and Fuzzy *C*-Medoids Clustering of Mixed Data algorithm (FCMd-CUB) is theoretically and empirically presented in this paper. Advantages of the suggested method are discussed in the conclusion section.

***Keywords*:** Likert-type variables, Cluster analysis, CUB model, Mixed information.

1. Introduction

Feelings, emotions, and preferences are complex psychological processes of interest in several disciplines (e.g. marketing, economics, and business) and are the basic of information of any Customer Satisfaction (CS) analysis. These human feelings/sentiments are usually captured through surveys in which ordinal scales, such as Likert-type scales, are commonly adopted. Since their introduction [1], Likert-type scales become widely and commonly adopted both in industry and in academia since they are user-friendly, easy-to-develop and easy-to administer (see [2]). These scales are made up of a set of items, usually formulated in terms of linguistic expressions coded into natural numbers, characterised by a rank order. Despite their advantages, Likert-type scales can only return imprecise and vague information about the investigated respondents' feelings. Firstly, individuals are asked to convert their thoughts into a linguistic expression and then to a natural number and these conversions can be inaccurate, causing loss of information, imprecision, and uncertainty (see [3]). Secondly, the meaning of each linguistic expression of the scale can be subjectively interpreted by respondents due to their knowledge about the phenomenon investigated, their culture, nationality, personal experience and understanding of the question (see [4]). However, since CS analysis rely on Likert-type scales to capture human thinking and personal feelings, it is fundamental to understand how to handle the uncertainty embedded in these scales to obtain trustworthy and accurate analysis and thus correctly inform practitioners. For the best of our knowledge, both a-priori alternatives (e.g. simple visual analogue scales and fuzzy rating scales) and a-posteriori correction have been suggested in the literature. A-priori alternatives are less user-friendly and less popular, complex both to implement and to analyse. Furthermore, when secondary data collected through Likert-type scales are used in the analysis, the a-priori alternatives are not applicable. Among a-posteriori alternatives, two approaches emerge in the literature: the fuzzy sets theory-based approach; the CUB (Combination of discrete Uniform and shifted Binomial random variables) model-based approach. The fuzzy sets theory-based approach has been extensively used to consider the imprecision and vagueness inherent to both Likert-type variables and human thinking (see [5], [6], [7]). Based on this approach, Likert-type data are recorded into fuzzy numbers before further analysis. Differently, the CUB model-based approach (introduced by [8]) aims to model the final answer as a mixture of two internal aspects, feeling and uncertainty. While the fuzzification of the Likert-type scales has already been used as a method to handle uncertainty when Likert-type data are used as segmentation variables in a cluster analysis (see for instance [9]), the CUB model-based approach has been used for this task only marginally in [10]. However, while in [10] the estimates of the CUB

model with covariates have been used to derive the parameters of the fuzzy numbers' membership function to use in the fuzzification process, in this study we suggest using the estimates of the CUB model with covariates directly as segmentation variables of the clustering algorithm. Therefore, Likert-type data are not pre-transformed into fuzzy data before a cluster analysis while the uncertainty embedded in such data are still considered and modelled.

2. Methods

In this paper we suggest combining the CUB model with a fuzzy clustering algorithm for mixed data to capture both the uncertainty related to Likert-type scale data and the uncertainty associated to the assignment of a unit to each cluster. More precisely, CUB model with covariates is used to estimate two latent components of respondent's answer behaviour, i.e. feelings and uncertainty, that are then included as input information, along with the original respondents' answer, of a fuzzy clustering algorithms able to handle segmentation variables of different kind. In the following subsection, CUB model ([8]) and the Fuzzy *C*-Medoids Clustering of Mixed Data (FCMd-MD) model ([11]) are briefly introduced.

2.1. CUB models

Combination of discrete Uniform and shifted Binomial random variables (CUB) model has been firstly introduced by [8] to analyse and model Likert-type data responses. The assumption underlying this model is that individual responses to a Likert-type question is the result of the combination of two components named feeling and uncertainty. Specifically, the feeling component determines the level of respondent's agreement/ pleasantness towards the object investigated while the uncertainty component collects different non-measurable factors, e.g. respondent laziness, difficulties in understanding the question, ignorance of the topics, wording and length of the scale, that affect the final response. Thus, the lower the uncertainty, the more reliable the answers. The final respondent judgement for the evaluated object is the result of latent pairwise comparisons between the specific evaluation of the object and all the other possible evaluations (i.e. items of the Likert-type scale), and each comparison corresponds to a success (the object obtains a high score) or a failure (the object receives a low score). Therefore, the probability distribution of the response variable should be a shifted binomial. Usually, each final answer is characterized by a certain degree of feeling and a certain degree of uncertainty. However, when the respondent is completely uncertain towards a question (i.e. no feeling component), the probability of choosing each evaluation is the same and the uniform distribution can properly represent the decision-making process. Consequently, the probability distribution of the response variable for the CUB model is a mixture of a shifted binomial and a uniform distribution. Since both feelings and uncertainty can be affected by respondents' characteristics the generalization of the CUB model with covariates is considered. Note that feelings and uncertainty are not constrained to be affected by the same set of covariates, so if p covariates affect the uncertainty and q covariates affect the feelings, the CUB (p; q) model with covariates is formulated as follows:

$$\Pr(R_i = r|\boldsymbol{x}_i, \boldsymbol{w}_i) = \pi_i \left[\binom{m-1}{r-1}(1-\xi_i)^{r-1}{\xi_i}^{m-r}\right] + (1-\pi_i)\frac{i}{m} \tag{1}$$

$$\begin{cases} logit(\pi_i) = \log\left(\dfrac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i\boldsymbol{\beta} \\ logit(\xi_i) = \log\left(\dfrac{\xi_i}{1-\xi_i}\right) = \boldsymbol{w}_i\boldsymbol{\gamma} \end{cases} \Leftrightarrow \begin{cases} \pi_i = \dfrac{1}{1+e^{-\beta_0-\beta_1 x_{1i}-\cdots-\beta_p x_{pi}}} \\ \xi_i = \dfrac{1}{1+e^{-\gamma_0-\gamma_1 w_{1i}-\cdots-\gamma_q w_{qi}}} \end{cases} \tag{2}$$

where R is a random discrete variable (respondent evaluation) that can only takes values from 1 (worst evaluation) to m (best evaluation), $m>3$; R_i represents the answer of the i-th respondent, i=1,…,n; $(1-\pi)$, with $\pi \in (0; 1]$, measures the uncertainty in the responses and $(1-\xi)$, with $\xi \in [0, 1]$, measures the respondent's perception towards the object to be evaluated, i.e. feeling towards the object under investigation; $\boldsymbol{x}_i$ is the vector of p covariates affecting uncertainty of the i-th respondent; $\boldsymbol{w}_i$ is the vector of q covariates affecting the feelings of the i-th respondent; $\boldsymbol{\beta}$ is the vector of p+1 unknown coefficients to be estimated for the uncertainty; and $\boldsymbol{\gamma}$ is the vector of q+1 unknown coefficients to be estimated for the feeling. Model's parameters are estimated using the maximum likelihood function. An Expectation-Maximization (EM) algorithm provides a computational solution for calculating the maximum likelihood estimates and

the adequacy of the model can be measured through the significance of parameters, the increasing of log-likelihood and other similar methods. Operatively the CUB model algorithm initially set randomly the estimation of both π and ξ and then it iteratively updated them to fit the model to the observed frequency distribution of the Likert-type data. At each iteration the unknown vectors of parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are estimated and updated.

2.2. Clustering for mixed data

Since the segmentation variables of the clustering algorithm are of different kinds (the original data are ordinal data, while both feeling and uncertainty components are continuous variables), a clustering algorithm of mixed data must be adopted (for a detailed review on this kind of clustering algorithm see [11]). In this study we suggest using the Fuzzy *C*-Medoids Clustering of Mixed Data (FCMd-MD) model developed by [11], where the input data are the observed data together with the estimated information derived from the CUB model, i.e. estimates of the individual feeling and uncertainty per question. Thus, in the following we will shortly address to the suggested model as the Fuzzy *C*-Medoids Clustering of CUB Data (FCMd-CUB). Following [11], the FCMd-CUB objective function to be minimised is as follows:

$$\begin{cases} \min: \sum_{i=1}^{n}\sum_{c=1}^{C} u_{ic}^{m} d_{ic}^{2} = \sum_{i=1}^{n}\sum_{c=1}^{C} u_{ic}^{m}\left\{d_1^2(\mathbf{y}_i, \tilde{\mathbf{y}}_c) + \left[w_\xi^2 d_2^2(\boldsymbol{\xi}_i, \tilde{\boldsymbol{\xi}}_c) + w_\pi^2 d_2^2(\boldsymbol{\pi}_i, \tilde{\boldsymbol{\pi}}_c)\right]\right\} \\ (s.t.)\sum_{c=1}^{C} u_{ic} = 1, u_{ic} \geq 0\,;\; w_\xi + w_\pi = 1;\; w_\xi \geq 0\,; w_\pi \geq 0 \end{cases} \quad (3)$$

where u_{ic} indicates the membership degree of the *i*-th unit to the *c*-th cluster; $m > 1$ is a weighting exponent that controls the fuzziness of the final partition; $d_{ic}^2 = d_1^2(\mathbf{y}_i, \tilde{\mathbf{y}}_c) + \left[w_\xi^2 d_2^2(\boldsymbol{\xi}_i, \tilde{\boldsymbol{\xi}}_c) + w_\pi^2 d_2^2(\boldsymbol{\pi}_i, \tilde{\boldsymbol{\pi}}_c)\right]$ is the overall weighted squared distance, computed using the Gower approach (Gower, 1971), between the *i*-th unit and the *c*-th medoid; $\boldsymbol{Y}$ is the matrix of original Likert-type data; $\boldsymbol{\xi}$ is the matrix of estimated feeling; $\boldsymbol{\pi}$ is the matrix of estimated uncertainty; d_1^2 and d_2^2 are suitable distance measures for ordinal and continuous variables, respectively; w_ξ^2 and w_π^2 are the weights of feeling and uncertainty, respectively, in the calculation of the final distance; $\tilde{\mathbf{y}}_c, \tilde{\boldsymbol{\xi}}_c, \tilde{\boldsymbol{\pi}}_c$ are the vectors of values observed for the *c*-th medoid on the three components, i.e original data, feeling and uncertainty. The sum of the weights is constrained to be unitary. Note that the weight of the original Likert-type data is set equals to 1. For comparison reasons, the three distances have been normalised to vary in the range [0, 1] before the computation of the overall weighted squared distance. Finally, it is important to note that the weights (w_ξ, w_π) are automatically estimated within the clustering algorithm by solving a Lagrangian optimisation problem with two constraints, one for the membership degrees and one for the weights (for more details on the Lagragian problem see [11]).

The FCMd-CUB allows to discover homogeneous groups of units based on mixed data while measuring the relevance of feeling and uncertainty components towards the final partition. In fact, the weighting system used to compute the distance matrix between each pair of respondents allows to rank feeling and uncertainty based on their relevance in the computation of the final partition while fixing the final evaluation as the most important one (its weight is constant and equal to 1). In other words, this algorithm can determine whether and how much feeling and uncertainty are important in the identification of the final partition. Moreover, the adoption of a fuzzy clustering approach allows relaxing the constraint of exclusiveness of each unit to a cluster, describing more realistically the hidden relationship among units.

3. Case study

The suggested method is than applied to a real case study gathered from the tourism sector. Data are drawn from the annual inbound survey of International Tourism in Italy (source Banca d'Italia) and in this study a sample of 3.127 foreign tourists who spent at least one night in the Venice municipality in 2017 have been analysed. Respondents were asked to report their level of satisfaction with 9 different aspects of the destination: friendliness of local people, accommodation, food and beverages, prices, landscapes, quality and variety of products sold, information, safety and the overall satisfaction with the destination. The level of satisfaction has been collected using a 10-point Likert-type scale from 1 = "Very unsatisfied" to 10 = "Very satisfied". Together with satisfaction, socio-demographic characteristics and trip characteristics information were asked through the survey. The sample is made up by people who travel for tourism, holiday or fun and whose holiday is a cultural one. The clustering analysis was conducted using the level of satisfaction with the different aspects as

segmentation variables. The feeling and uncertainty have been estimated per respondent and per question by means of Eqs. (1) – (2) using socio-demographic and trip characteristics as covariates.

4. Conclusion

In CS analysis, Likert-type scales are frequently used to gather personal feelings and evaluations about a post-experience or post-use. However, information obtained using such scales is uncertain and imprecise, as described in Section 1. Therefore, Likert-type data need to be pre-processed before being used in further analysis, such as cluster analysis. This study aims to present an alternative way to a-posteriori handle the uncertainty and vagueness naturally embedded in Likert-type data when a cluster analysis is conducted. While in the literature the most common pre-processing operation is the fuzzification, in this study we suggest modelling the Likert-type data uncertainty by means the CUB model with covariates. Furthermore, the estimates of the CUB model parameters, i.e. feelings and uncertainty, are used, together with the original respondents' answer to the Likert-type questions, as input variables of the clustering algorithm. Although any clustering algorithm for mixed data can be adopted, we suggest applying the FCMd-MD to obtain a more realistic multivariate description of the units. The main advantages in using the suggested a-posteriori correction are the possibility to model individual uncertainty using further features collected through the questionnaire and the removal of the elicitation problem, i.e. the necessity to define the membership function of the fuzzy numbers since no fuzzy recoded is involved. Furthermore, the FCMd-MD model allows the identification of the most important components in the identification of the final partition. From a business point of view, the FCMd-CUB is of particular interest since it allows to obtain more reliable clusters with the possibility to effectively direct marketing and managerial resources to specific customers.

Acknowledgements

BIRD 2022 project titled "Fuzzy theory in Unsupervised Machine Learning algorithm and Sentiment Analysis"

Special thanks to Prof. Pierpaolo D'Urso and to Prof. Luigi Salmaso for the supervision on the development of the present paper.

References

[1] R. Likert, "A technique for the measurement of attitudes" *Archives of psychology*, vol. 22, no.140, pp. 5-55, 1932.

[2] M.Á. Gil, M.A. Lubiano, S. De la Rosa de Sàa, B. Sinova, "Analyzing data from a fuzzy rating scale-based questionnaire: a case study", Psicothema, vol. 27, no. 2, pp. 182-191, 2015.

[3] P. D'Urso, "Fuzzy clustering of fuzzy data", *Advances in fuzzy clustering and its applications*, J. Valente de Oliveira and W. Pedrycz, Ed. John Wiley & Sons, 2007, pp. 155-192.

[4] E. Davidov, B. Meuleman, J. Cieciuch, P. Schmidt, J. Billiet, "Measurement equivalence in cross-national research", *Annual review of sociology*, vol. 40, pp. 55-75, 2014.

[5] R. Coppi, and P. D'urso, "Fuzzy k-means clustering models for triangular fuzzy time trajectories", *Statistical Methods and Applications*, vol. 11, pp. 21–40, 2002.

[6] W.-L. Hung and M.-S. Yang, "Fuzzy clustering on lr-type fuzzy numbers with an application in Taiwanese tea evaluation". *Fuzzy sets and systems*, vol. 150, pp. 561-577, 2005.

[7] H.-Y. Hu, Y.-C. Lee, and T.-M. Yen, "Service quality gaps analysis based on fuzzy linguistic servqual with a case study in hospital out-patient services". *The TQM Journal*, Vol. 22 No. 5, pp. 499-515, 2010.

[8] D. Piccolo, "On the moments of a mixture of uniform and shifted binomial random variables" *Quaderni di Statistica*, vol 5, pp. 85-104, 2003.

[9] M. Disegna, P. D'Urso, and R. Massari, "Analysing cluster evolution using repeated cross-sectional ordinal data", *Tourism Management*, vol. 69, pp. 524–536, 2018.

[10] N. Biasetton, M. Disegna, E. Barzizza, L. Salmaso, "A new adaptive membership function with CUB uncertainty with application to cluster analysis of Likert-type data" *Expert Systems with Applications*, vol. 213, pp. 118893, 2023.

[11] P. D'Urso and R. Massari, "Fuzzy clustering of mixed data" *Information Sciences*, vol. 505, pp. 513–534, 2019.