

A ROBUST MAXIMUM-ENTROPY FUZZY CLUSTERING APPROACH FOR IMPRECISE DATA

Fordellone Mario¹, Di Gennaro Piergiacomo¹, Nicolao Giovanni¹, Schiattarella Paola¹, Smimmo Annafrancesca¹, Simeon Vittorio¹, Signoriello Giuseppe¹, Chiodini Paolo¹

¹Unità di Statistica Medica,
Dipartimento di Salute Mentale e Fisica e Medicina Preventiva,
Università degli Studi della Campania "Luigi Vanvitelli"

Introduction

In conventional statistical data analysis, point data are typically analyzed, which means exact measurement results that consist of features of the reference sample. These values can either be directly observed as measurement outcomes (e.g., a person's systolic and/or diastolic blood pressure) or as counts of a category (e.g., the gender of that person). However, in many real-life applications, these measurement results are never precise, and there is always some degree of uncertainty associated with them.

The uncertainty of a measurement can be defined as the interval on the measurement scale within which the true value lies with a specified probability, considering all sources of error [1]. Quantifying this uncertainty can become a crucial issue in the statistical quality of data. In the medical field, it is standard practice for chemists and biologists to provide a statement of uncertainty alongside their estimated measurements, so that it can be taken into account during data analysis [1,2]. In other words, a measurement cannot be properly interpreted without knowledge of its uncertainty. In clinical practice, many rules and guidelines have been proposed to provide a general overview of the uncertainty concept in the measurement phase and to account for it in data interpretation [3]. For example, readers can refer to [4], which provides a review suggesting a rule-based approach for calculating measurement uncertainty, and [5], which offers a systematic review on uncertainty tolerance in health and healthcare-related outcomes. Following this research line, Fordellone et al., 2023 [6] proposed an entropy-based fuzzy clustering technique for interval-valued (EFC-ID). The novelty of this statistical approach was to consider the uncertainty of the data in the classification procedure using the standard deviation of data variables as a measure of the uncertainty. In this work, to mitigate the disruptive effects that potential outlier interval-valued data might have on the clustering process, we propose an enhanced robust version of the EFC-ID model, referred to as the rEFC-ID.

Objectives

The rEFC-ID method represents a significant advancement in clustering frameworks, offering enhanced resilience particularly suited for environments rife with data contamination or anomalies. By integrating robustness into the clustering process, it addresses challenges that traditional methods may struggle with. In practical terms, this means that rEFC-ID can effectively identify and mitigate the impact of outliers or corrupted data points, ensuring that the resulting clusters are more accurate and reliable.

Methods

This rEFC-ID is built upon the concept of "trimming", which involves the systematic exclusion of a certain proportion of the most extreme data points that could skew the clustering results. By incorporating the trimming approach, the rEFC-ID aims to enhance the stability and reliability of the clustering outcomes, ensuring that the core structure of the data is accurately captured without being unduly influenced by outliers. Then, the objective function to optimized became:

$$J_{\text{rEFC-ID}}(\mathbf{U}, \tilde{\mathbf{X}}, \mathbf{w}) = \sum_{i=1}^{n-\alpha n} \sum_{g=1}^k u_{ig} [w_c^2 d^2(\mathbf{c}_i - \mathbf{c}_g) + w_r^2 d^2(\mathbf{r}_i - \mathbf{r}_g)] + p \sum_{i=1}^{n-\alpha n} \sum_{g=1}^k u_{ig} \log(u_{ig}),$$

under the following constraints: $\sum_{g=1}^k u_{ig} = 1, u_{ig} \geq 0; w_c \geq w_r \geq 0, w_c + w_r = 1; 0 \leq \alpha \leq 1$. Where u_{ig} indicates the membership degree of the i -th unit in the g -th cluster; c_i and r_i are the centers and radii of the i -th unit, respectively; c_g and r_g are the “trimmed” data centroids of the centers and radii in the g -th cluster, respectively; $p \sum_{i=1}^{n-\alpha n} \sum_{g=1}^k u_{ig} \log(u_{ig})$ is the fuzzy entropy function; p is a weight factor, called degree of fuzzy entropy; α indicates the fraction of objects discarded in the clustering process and, thus, not considered in the optimization problem. Then, αn is the number of data objects to be trimmed from the dataset.

For evaluating the performance of the model, the Frobenius distance (Fdist) computed between the natural (generated) memberships matrix \mathbf{U} and the memberships matrix obtained by the model $\hat{\mathbf{U}}$, was used. The Frobenius distance has been then averaged over the 300 simulation runs and the obtained results were compared with the results obtained by EFC-ID. In particular, in the simulated scheme we have (i) the centers-radii scenario, where the centers and the radii of the interval-valued data generated have a group structure, and (ii) the centers scenario, where the radii of the interval-valued data are all randomly generated, while the centers of the data generated have a group structure. The percentage of outliers considered for both scenarios are 0%, 1%, 5%, and 10%. Each simulated dataset is composed of one hundred objects ($n = 100$) and two interval-valued variables ($J = 2$). For details on the simulation design, reader can refer to [6].

Results

Table 1 shows the compared results obtained by EFC-ID and rEFC-ID with respect two different scenarios and four different percentages of outliers.

Table 1 - EFC-ID and rEFC-ID clustering models' performance for different scenarios and different percentages of outliers for 300 simulation runs

		CENTERS / RADII		CENTERS	
	Outliers	Mean (F. dist)	St. Dev (F. dist)	Mean (F. dist)	St. Dev (F. dist)
<i>EFC-ID</i>	0%	0.000	0.00E+00	0.000	5.47E-05
	1%	0.009	0.04E+00	0.120	8.31E-05
	5%	1.653	0.42E+00	1.896	9.38E-05
	10%	2.344	0.51E+00	2.756	9.49E-05
<i>rEFC-ID</i>	Outliers	Mean (F. dist)	St. Dev (F. dist)	Mean (F. dist)	St. Dev (F. dist)
	0%	0.000	0.00E+00	0.000	5.47E-05
	1%	0.001	0.01E+00	0.091	4.11E-05
	5%	0.998	0.02E+00	1.317	4.28E-05
	10%	1.589	0.02E+00	1.912	5.04E-05

Conclusions

The simulation study highlighted a key finding: the EFC-ID model exhibits greater sensitivity to outliers when contrasted with the rEFC-ID method. In future research, we will investigate other possible robust clustering for interval-valued data, such as, for instance, the possibilistic clustering approach.

References

- [1] Fordellone, M.; Chiodini, P. Unsupervised Hierarchical Classification Approach for Imprecise Data in the Breast Cancer Detection. *Entropy* 2022, 24, 926.

- [2] Analytical Methods Committee. Uncertainty of measurement: Implications of its use in analytical science. *Analyst* 1995, 120, 2303–2308.
- [3] White, G.H.; Farrance, I. Uncertainty of measurement in quantitative medical testing: A laboratory implementation guide. *Clin. Biochem. Rev.* 2004, 25, S1.
- [4] Farrance, I.; Frenkel, R. Uncertainty of measurement: A review of the rules for calculating uncertainty components through functional relationships. *Clin. Biochem. Rev.* 2012, 33, 49.
- [5] Strout, T.D.; Hillen, M.; Gutheil, C.; et al. Tolerance of uncertainty: A systematic review of health and healthcare-related outcomes. *Patient Educ. Couns.* 2018, 101, 1518–1537.
- [6] Fordellone, M.; De Benedictis, I.; Bruzzese, D.; et al. A Maximum-Entropy Fuzzy Clustering Approach for Cancer Detection When Data Are Uncertain. *Appl. Sci.* 2023, 13, 2191.