Poster: Intelligent fuzzifier-based cluster validation for incomplete longitudinal digital trial data

Hieu Ngo†
University of Massachusetts Dartmouth
Dartmouth, Massachusetts, U.S.A
hngo1@umassd.edu

Hua Fang
University of Massachusetts Dartmouth
Dartmouth, Massachusetts, U.S.A
hfang2@umassd.edu

Honggang Wang
University of Massachusetts Dartmouth
Dartmouth, Massachusetts, U.S.A
hwang1@umassd.edu

ABSTRACT

Digital technology has huge potentials in transforming clinical trial research. One common issue in digital clinical trials for long-term behavioral treatments is incomplete longitudinal data, as subjects' behavior changes over time. In this paper, we aim to improve the fuzzy clustering accuracy and stability of digital clinical trials by intelligently searching for the optimal fuzzifier, which is the key to identify the optimal number of overlapped clusters for incomplete longitudinal data. Our findings showed that integrating optimal fuzzifier searching with cluster validation can streamline the clustering process, thus enabling the intelligent fuzzy clustering procedure.

KEYWORDS

Fuzzifier, Intelligent, Validation, Incomplete Longitudinal, Digital Trials

1 Introduction

The use of digital technology in clinical trial research has the potential not only to reduce the cost and time for research, but also enable research opportunities that are more patient-centered to quickly address important health issues such as substance abuse, mental health, chronic diseases, and cancers, etc. In digital clinical trial research, one of the emerging smart health areas, missing data is common, especially for longitudinal digital trials. Clustering incomplete longitudinal digital trial data is another challenge due to participants' complex behaviors over time during these trials. Fuzzy clustering uses the concept of membership into data partition to indicate the degree to which an object belongs to different clusters. To fully enable intelligent validation of fuzzy methods, such as MIFuzzy [1], one key component is to search for the optimal fuzzifier, also called the weighing exponent. In this paper, we propose an intelligent fuzzifier-based cluster validation method for incomplete longitudinal digital data.

CHASE' 22, November 17–19, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9476-5/22/11.
https://doi.org/10.1145/3551455.3559154

2 Related Work

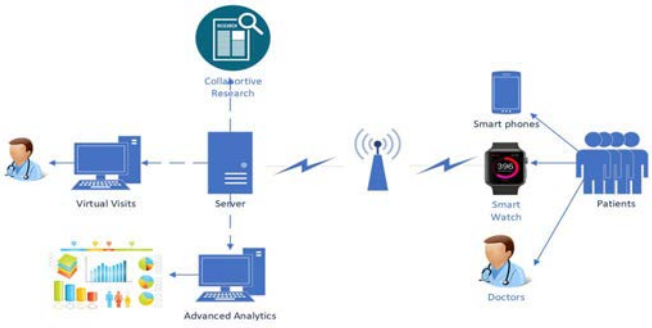

Figure 1: Potential of Digital Trials

MIFuzzy clustering is a non-parametric soft clustering method, used for semi-supervised or unsupervised learning [1]. Multiple Imputation (MI) seeks to deal with missing data that happens often in longitudinal digital trials, e.g., when people drop out of treatments, fail to respond to a survey, etc. To validate fuzzy clustering results, we need to use fuzzy validation indices. The Xie-Beni (XB) Index is a popular fuzzy cluster validity measure [2]. It measures the compactness of the fuzzy partition over the separation of the clusters.

Aside from validating indices, selecting an appropriate fuzzifier (m) is particularly crucial in identifying the optimal number of clusters. Studies explore this idea of obtaining the fuzzifier directly from the dataset [3,4,5]. In [3], the author theoretically proved and computed the fuzzifier by searching a global optimal solution. Similarly, study [4] learned a general function relation between the fuzzifier and the dataset properties: data dimension and size. In [5], the authors proposed a new integrated fuzzifier evaluation and selection to assess and select the fuzzifier.

This paper proposes an intelligent fuzzifier-based cluster validation for incomplete longitudinal digital trials. This method aims to intelligently (auto-) select the fuzzifier for

incomplete digital trial data to further reduce the complexity of applying such methods and make the validation more tractable.

3 Methodology

Algorithm 1 Intelligent Fuzzifier-based Cluster Validation for incomplete longitudinal digital trials

Input: incomplete data X, fuzzifier testing size P, imputed datasets number Q, cluster range R

Output: Optimal fuzzifier m and optimal number of cluster k

1: Multiple Imputation on dataset X to create Q MI sets MI_X_1 to MI_X_Q
2: **for** p = 1 to P **do**
3: **for** q = 1 to Q **do**
4: m = 1 + p*0.1 an data = MI_X_q;
5: **for** r = 2 to R **do**
6: $U_{ij} = MIFuzzy(MI_{X_q}, r\ clusters, fuzzifier\ m)$
7: **end for**
8: **end for**
9: Calculate k_p using MI_XB
10: **end for**
11: **return** m = m_p, $k = k_p$ such that $k_p = k_{p+1} = k_{p+2} = k_{p+3}$

We describe our algorithms in Algorithm 1. This algorithm uses MI_XB, which finds the optimal number of clusters across multiple imputed datasets. For each imputed dataset and each fuzzifier, MI_XB find the number of clusters k_{pq} such that:

$$MI\_XB(k_{pq}) = \min\{x \in XB_{pqr} : x \text{ is a local minimum of } XB_{pqr}\} \quad (1)$$

And the number of cluster k_p such that:

$$k_p = \widehat{k_{pq}} \quad (2)$$

Where XB is calculated as:

$$XB = \frac{\sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}^2 \left\|x_i - c_j\right\|^2}{n\, min_{1\le i\ne j\le c}\left\|c_i - c_j\right\|^2} \quad (3)$$

In these equation, XB_{pq} is the set of XB_{pq1} to XB_{pqR}; k_{pq} is the set of k_{p1} to k_{pQ}; $\widehat{k_{pq}}$ is the mode of k_{pq}; u_{ij} is the membership value of i^{th} point in j^{th} cluster; n is the number of points in the set; and c is the cluster number.

4 Analysis

For analyses, we use the incomplete dataset from a longitudinal dietary trial dataset [6]. The dataset contains 240 subjects and we used 4 variables that describe subjects' diet quality over time with a missingness of approximately 18%. This data is then imputed into 10 complete datasets MI_X_1 to MI_X_{10}, where each dataset contains 240 observations and 4 variables. The MIFuzzy algorithm is then applied to each of the 10 datasets, with the fuzzifier *m* ranging from 1.1 to 3. This results in 10x20 partition tables U_{ij}. With the partition tables U_{ij}, we use the MI_XB scores to determine the optimal number of clusters, which is 5 clusters. Finally, using the table of the majority voting results, we auto-identified the optimal m to be $m = 2.1$. Figure 2 shows the Fuzzy Sammon mapping of the clustering results to support the 5-cluster validation, thus fulfilling a streamlined fuzzy cluster validation.

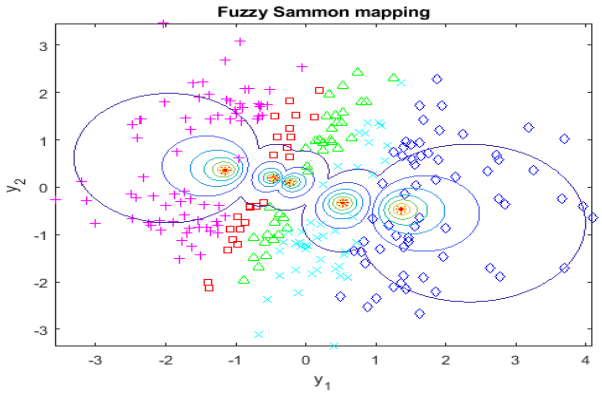

Figure 2: Example of MI_XB Scores for the optimal number of clusters at 5 clusters

5 Conclusion

In recent years, digital trials are becoming popular with the advanced development of information technology. We aim to improve the accuracy and stability of fuzzy clustering, which is useful in digital clinical trials. Specifically, we proposed to streamline a fuzzy cluster validation procedure for incomplete digital trials based on intelligent optimal fuzzifier search and fuzzy cluster validation index. The results show that this approach can find the adaptive fuzzifier for the real trial dataset and facilitate the identification of the optimal number of clusters. In future research, more real and synthetic digital trial data will be used to further test multiple imputation-based fuzzy clustering validation for incomplete digital trial data.

ACKNOWLEDGMENTS

This work is partly supported by NIH grant RO1DA033323-01A1, R01DK129432, 1R56DK114514, and NSFIIS-III 2140729 to Dr. Fang.

REFERENCES

[1] Fang, Hua. "MIFuzzy clustering for incomplete longitudinal data in smart health." Smart Health 1 (2017): 50-65. DOI:10.1016/j.smhl.2017.04.002
[2] Xie, Xuanli Lisa, and Gerardo Beni. "A validity measure for fuzzy clustering." IEEE Transactions on pattern analysis and machine intelligence 13.8 (1991): 841-847.
[3] Yu, Jian, Qiansheng Cheng, and Houkuan Huang. "Analysis of the weighting exponent in the FCM." IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 34.1 (2004): 634-639. Conference Short Name:WOODSTOCK'18
[4] Schwämmle, Veit, and Ole Nørregaard Jensen. "A simple and fast method to determine the parameters for fuzzy c–means cluster analysis." Bioinformatics 26.22 (2010): 2841-2848.
[5] Wang, Chanpaul Jin. "A new integrated fuzzifier evaluation and selection (nifes) algorithm for fuzzy clustering." Journal of Applied Mathematics and Physics 3.07 (2015): 802.
[6] Ma, Yunsheng, et al. "Single-component versus multicomponent dietary goals for the metabolic syndrome: a randomized trial." Annals of internal medicine 162 (4), 248-257