University of Massachusetts Medical School

eScholarship@UMMS

2016-06-01

Multiple Imputation based Clustering Validation (MIV) for Big Longitudinal Trial Data with Missing Values in eHealth

Zhaoyang Zhang
University of Massachusetts Medical School

Et al.

Let us know how access to this document benefits you.

Repository Citation

Multiple Imputation based Clustering Validation (MIV) for Big Longitudinal Trial Data with Missing Values in eHealth

Zhaoyang Zhang[1], **Hua Fang**[1], and **Honggang Wang**[2]

Zhaoyang Zhang: zhaoyang.zhang@umassmed.edu; Hua Fang: hua.fang@umassmed.edu; Honggang Wang: hwang1@umassd.edu

[1]Department of Quantitative Health Science, University of Massachusetts Medical School, Worcester, MA 01655, USA

[2]Department of Electrical and Computer Engineering, University of Massachusetts Dartmouth, North Dartmouth, MA 02747, USA

Abstract

Web-delivered trials are an important component in eHealth services. These trials, mostly behavior-based, generate big heterogeneous data that are longitudinal, high dimensional with missing values. Unsupervised learning methods have been widely applied in this area, however, validating the optimal number of clusters has been challenging. Built upon our multiple imputation (MI) based fuzzy clustering, MIfuzzy, we proposed a new multiple imputation based validation (MIV) framework and corresponding MIV algorithms for clustering big longitudinal eHealth data with missing values, more generally for fuzzy-logic based clustering methods. Specifically, we detect the optimal number of clusters by auto-searching and -synthesizing a suite of MI-based validation methods and indices, including conventional (bootstrap or cross-validation based) and emerging (modularity-based) validation indices for general clustering methods as well as the specific one (Xie and Beni) for fuzzy clustering. The MIV performance was demonstrated on a big longitudinal dataset from a real web-delivered trial and using simulation. The results indicate MI-based Xie and Beni index for fuzzy-clustering is more appropriate for detecting the optimal number of clusters for such complex data. The MIV concept and algorithms could be easily adapted to different types of clustering that could process big incomplete longitudinal trial data in eHealth services.

Keywords

Big data; Validation; Multiple imputation; Fuzzy clustering; Missing data; Longitudinal trial

Introduction

In eHealth services, web-delivered trials or interventions are in increasing demand due to their cost-effective potential in accessing a large population [1]. These trials commonly generate big, complex, heterogenous and high-dimensional longitudinal data with missing values. These data have the typical five "V" properties of big data [2]. Specifically, the

---

Correspondence to: Hua Fang, hua.fang@umassmed.edu.

Volume of such data is substantially large in terms of the number of participants and attributes, with which traditional clinical trials are incomparable; its Variety refers to different web-delivered components; its Velocity is undoubtly superior to traditional offline trials, because the data are recorded real-time; its Veracity is obvious because of its unstructured nature and messiness; and its Value would be substantial as long as its efficacy is clarified.

Our line of research focuses on multiple imputation based fuzzy clustering (MIfuzzy), as it fits better to longitudinal behavioral trial data than other methods based on our previous studies [3–5]. There is a paucity of literature in validating the clustering results from big longitudinal eHealth trial data with missing values and our line of research [3–6] attempts to fill this gap. Probabilistic clustering (e.g., Gaussian Mixture models [7]) and Hidden Markov Model-based Bayesian clustering [8], Neural networks models [9, 10] (e.g, Kohonen's Self Organizing Map, SOM), Hierarchical clustering [11], Partition-based clustering (e.g, K-means or Fuzzy C Means) are commonly used for clustering and demonstrated efficiently for specific data structure in other fields. However, these methods have at least one of these following disadvantages and are less appealing to big behavioral trial data which are typically high dimensional, heterogeneous, non-normal, longitudinal with missing values: Assumption of underlying statistical distributions (Gaussian) or prior distributions (Bayesian approach); (slow) convergence to a local maximum or no convergence at all especially for multi-modal distributions and large proportions of missing values with high-dimensional data and many clusters; unclear validation indices or procedures; inability to handle missing values or incorporate information about the shape and size of clusters; computational inefficiency; and their unknown utility in behavioral trial studies. With a pre-specified number of clusters, MI-Fuzzy was demonstrated to perform better than these methods in terms of its clustering accuracy and inconsistency rates using real trial data [3–5].

As aforementioned, missing data are common in longitudinal trial studies [3, 12, 13]. The performance of MI-Fuzzy was evaluated under these three mechanisms: Missing Completely at Random, Missing at Random (MAR) and Missing not at Random (NMAR). The preliminary results indicate that MIfuzzy is invariant to the three mechanisms and accounts for the clustering uncertainty in comparison to non- or single-imputed fuzzy clustering [14].

Built upon our multiple imputation (MI) based fuzzy clustering, MIfuzzy [4, 6, 15], we proposed MI-based validation framework (MIV) and corresponding MIV algorithms for clustering such big longitudinal web-delivered trial data with missing values. Briefly, MIfuzzy is a new trajectory pattern recognition method with a full integration and enhancement of multiple imputation theory for missing data [3, 16–23] and fuzzy logic theories [24–26]. Here, we focus on cluster validation and extend traditional validation of complete data to MI-based validation of incomplete big longitudinal data, especially for fuzzy-logic based clustering [27–29]. Unlike simple imputation such as mean, regression, and hot deck that cause bias and lose statistical precision, multiple-imputation accounts for imputation uncertainty [30–32].

To build the MIV, we will consider two clustering stability testing methods, cross-validation and bootstrapping; to adapt to fuzzy clustering, we will use Xie and Beni (XB), a widely-accepted fuzzy clustering validation index [33–35], and another newly emerging index, modularity [36, 37]. All four validation methods will be integrated with MI to demonstrate our proposed MIV framework.

Clustering stability has been used in recent years to help select the number of clusters [38–40]. It measures the robustness against the randomness of clustering results. The core idea is based on the intuition that a good clustering will produce a stable result that does not vary from one sample to another. The clustering stability method can be used in both distance based and non-distance based clustering methods, such as model based clustering [41–43] and spectrum clustering [44–46]. Bootstrap and cross-validation are two common clustering stability testing methods. Bootstrap is a statistical technique to assign measures of accuracy, such as bias, variance and confidence intervals, to sample estimates [47–49]. Bootstrap is used when the sampling size is small or impossible to draw repeated samples from the population of interest. In such cases, bootstrap can be used to approximate the sampling distribution of a statistic [50–52]. Cross-validation can be used in clustering algorithms to estimate its predictive strength [53–56]. In cross-validation, the data is split to two or more partitions. Some partitions are used for training the model parameters, and the others, namely the validation (testing) set, are used to measure the performance of the model.

Two types of cross-validation can be distinguished, exhaustive and non-exhaustive: The first one includes leave-p-out and leave-one-out cross-valuation; the latter does not compute all ways of splitting the original data. The non-exhaustive cross-validation contains k-fold cross-validation, holdout and repeated random sub-sampling validation [57, 58]. The holdout method is the simplest among cross-validation methods, with which the data set is only separated into one training and one testing set. Although computationally efficient, the evaluation may be significantly different depending on how the division of the dataset is made between the training and testing sets. The *k*-fold cross validation improves and generalizes the holdout method by dividing a dataset into *k* subsets, where the variance of the resulting estimate is reduced as *k* is increased. A variant of this method is called repeated random sub-sampling validation, also known as Monte Carlo cross-validation to randomly divide the data into a test and training set *k* different times. Due to randomness, some data may never be selected while others may be selected more than once, resulting in potential overlapped validation subsets. The *k*-fold cross validation was used in this work to ensure that all data points are used for both training and validation, and each data point is used for validation exactly once. Modularity can measure the structure of networks or graphs [36, 37, 59], and can be used to cluster data by transforming the data points into a graph with their similarities [60]. Thus, modularity can be used to determine the number of clusters in data analyses. Most importantly, for fuzzy clustering, Xie and Beni [33], this widely accepted validation fuzzy clustering index was incorporated into this MI validation framework.

Here, we propose MIV algorithms to auto-search, compare, synthesize and detect the optimal number of clusters for incomplete big longitudinal data based on MI-based clustering stability tests (MI-cross-validation and MI-bootstrapping), MI-XB, and MI-modularity. The rest of the paper is organized as follows: Section "Multiple-imputation-

based validation framework (MIV) for incomplete big web trial data in eHealth" presents MIV theoretical framework and algorithms; Section "Numerical analyses and simulation" performs numerical analyses using real and simulated incomplete big longitudinal data and simulation; and Section "Conclusion" concludes the paper. Table 1 lists notations used in this paper.

Multiple-imputation-based validation framework (MIV) for incomplete big web trial data in eHealth

Our MI-based validation framework (MIV) is designed to detect the optimal number of clusters from incomplete big longitudinal data in eHealth, using a suite of MI-based methods and indices, such as MI-based clustering stability (MI-S), MI-based XB index (MI-XB) and MI-based Modularity (MI-Q). The procedure of the proposed MIV platform is described in Fig. 1. Briefly, the MIV is an auto-iterative validation procedure where the MI-based index is calculated for a set of cluster numbers on each imputed dataset, incorporating the idea of the multiple imputation theory to minimize the "uncertainty" in selecting the optimal number of clusters for incomplete data sets.

MI-based clustering stability for incomplete big web trial data in eHealth

For incomplete big longitudinal web trial data, rather than single imputation, we incorporate Multiple Imputations (MI) to impute missing values to reduce imputation uncertainty [30–32]. In the imputation step, Markov chain Monte Carlo (MCMC) was used to estimate the missing values. The expectation-maximization (EM) algorithm was first applied to find the maximum likelihood estimates of the parameters for the distribution of incomplete big web trial data, then Markov chains were constructed such that the pseudo random samples were drawn from the limiting, or stationary distribution of the data to stabilize to a stationary distribution [17]. Specifically, denote g as different missing patterns, the maximized observed data log likelihood is expressed as,

$$\log L(\theta|Y_{obs}) = \sum_{g=1}^{G} \log L_g(\theta|Y_{obs}), \quad (1)$$

in which

$$\log L_g(\theta|Y_{obs}) = -\frac{n_g}{2}\log\left|\sum_g\right| - \frac{1}{2}\sum_{ig}(y_{ig}-\mu_g)'\sum_g^{-1}(y_{ig}-\mu_g), \quad (2)$$

where n_g is the number of observations in the g-th group, y_{ig} is a vector of observed values corresponding to observed variables, μ_g is the corresponding mean vector, and Σ_g is the associated covariance matrix. The EM algorithm was also used to find the posterior mode where the observed data posterior density is used instead of the observed data likelihood as it is guaranteed to be non-decreasing at each iteration. The logarithm of the observed data posterior density is calculated by

$$\log P(\theta|Y_{obs}) = L(\theta|Y_{obs}) + \log \pi(\theta), \quad (3)$$

in which

$$\begin{aligned}
\log \pi(\theta) &= -\frac{m+p+2}{2} \log |\textstyle\sum| - \tfrac{1}{2} \mathrm{tr} \textstyle\sum^{-1} M_0, \\
M_0 &= \Lambda^{-1} + \tau (\mu - \mu_0)(\mu - \mu_0)^T,
\end{aligned} \quad (4)$$

where $(\tau, m, \mu_0, \Lambda)$ are the parameters for the normal inverted-Wishart prior. When the prior information about the is unknown, we apply the Bayes' theorem with the prior,

$$\pi(\theta) \propto |\textstyle\sum|^{-\left(\frac{p+1}{2}\right)}, \quad (5)$$

which is the limiting form of the normal inverted-Wishart density as $\tau \to 0$, $m \to -1$ and $\Lambda^{-1} \to 0$. The prior distribution of μ_0 is assumed to be uniform and $\mu_0 \to 0$. This noninformative prior is also called jeffreys prior in [17].

Next, MCMC was used to impute the missing values by making pseudorandom draws from the probability distributions with parameters obtained by the EM algorithm. Information about known parameters can be expressed in the form of a posterior probability distribution by Bayesian inference,

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(q)d\theta}. \quad (6)$$

The entire joint posterior distribution of the known variables can be simulated and the posterior parameters of interest can be estimated.

Similar to the EM algorithm, the imputation algorithm has two steps, 1) I-step: make pseudorandom draws from the probability distribution for the missing values,

$$Y_{mis}^{(t+1)} \sim P(Y_{mis}|Y_{obs}, \theta^{(t)}), \quad (7)$$

and 2) P-step: update the parameters,

$$\theta^{(t+1)} \sim P(\theta|Y_{obs}, Y_{mis}^{(t+1)}). \quad (8)$$

If the parameter is multivariate normal, the I-step involves the independent simulation of random normal vectors for each row in the incomplete big dataset.

Assuming a normal distribution of the incomplete big data and Jeffreys prior, the parameter θ is updated at the P-step by

$$\begin{aligned}
\sum\nolimits^{(t+1)} | \mathbf{Y} \quad &\sim W^{-1}(N-1, (N-1)\mathbf{S}), \\
\mu^{(t+1)} | (\sum\nolimits^{(t+1)}, \mathbf{Y}) \quad &\sim N\left(\overline{y}, \tfrac{1}{n}\sum\nolimits^{(t+1)}\right),
\end{aligned} \quad (9)$$

where n is the number of observations, $\mathbf{Y}$ is completed data generated by previous I-step, $\overline{y}$ is the mean vector, and $(N-1)\mathbf{S} = \mathbf{Y}'\mathbf{Y} = \sum_i y_i y_i^T$.

To obtain multiply imputed datasets, Multiple Markov Chains were constructed, where the I- and P-steps were performed iteratively until the stationary distributions were reached. The initial portion of these Markov chain samples, called burn-in, were discarded, where the default was set as 200 according to literature [17, 61]. After the burn-in periods, the Markov Chains continue, as shown in Fig. 2, until additional I-steps were performed to obtain a complete dataset from the stationary distribution for each Markov chain, marked as X_i, i.e., the i-th imputation data.

A fuzzy clustering method ψ is applied to each imputed dataset X_i, $i = 1, 2, \ldots, M$, where M is the number of imputations, $\Psi_{i,k} = \psi(X_i, k)$, $i = 1, 2, \ldots, M$, $k = 1, 2, \ldots, K$, where ψ is a fuzzy clustering method that clusters the data X into k latent groups. K is the maximum number of clusters. For each k, M clustering outputs were obtained, and each case has M cluster memberships. We count how many times a case belongs to a cluster and the maximum count determines his final cluster membership. For the j-th case x_j, $1 \le j \le N$, c_u, $(u = 1, 2, \ldots, k)$ is the frequency the case belongs to the u-th cluster, thus $\sum_{u=1}^{k} c_u = M$. The final cluster membership of x_j, denoted by v_j is decided by $v_j = \arg\max_u c_u$.

If we have N Cases $\mathbf{O}^n = \{x_1, x_2, \ldots, x_n\}$, and each case has p features, $\Psi(X, k)$, $k = 1, 2, \ldots$ is the clustering method that can cluster the data X into k clusters, as defined above. Note that when $k = 1$, $\Psi(X^n, 1) \equiv 1$ for any data X.

Definition 1

The clustering distance between any two clustering method $\psi_1(x)$ and $\psi_2(x)$ is defined as [62],

$$D(\psi_1, \psi_2) = \Pr(I(\psi_1(X) = \psi_1(Y)) + I(\psi_2(X) = \psi_2(Y)) = 1), \quad (10)$$

where $I(\cdot)$ is an indicator function and X, Y are independently sampled from O.

Based on this definition, the clustering distance measures the disagreement between two clusters. It equals to the sum of $\Pr(\psi_1(x^0) = \psi_1(y^0), \psi_2(x^0) \neq \psi_2(y^0))$ and $\Pr(\psi_1(x^0) \neq \psi_1(y^0), \psi_2(x^0) = \psi_2(y^0))$.

Definition 2

The clustering stability of $\Psi(\cdot, k)$ is defined as,

$$s_k = 1 - E(D(\psi_1(X, k), \psi_2(Y, k))), \quad (11)$$

where $E(\cdot)$ is the expectation function, k, X and Y are the same as in Definition 1.

We proposed two MI-based bootstrap and cross-validation methods to assess the clustering stability. The procedure of MI-based stability validation is shown in Fig. 2. Briefly, multiple samples are generated by bootstrapping or permutation, then the stabilities are calculated for a range of number of clusters. Finally, the optimal number of clusters is identified at the largest stability value.

MI-based bootstrapping for incomplete big web trial data in eHealth—The MI-based clustering stability using bootstrap method for k clusters is expressed as

$$\text{MI-} S_{BS}(k) = \frac{1}{MB} \sum_{m=1}^{M} \sum_{b=1}^{B} D(\Psi(X_{mb1}, k), \Psi(X_{mb2}, k)), \quad (12)$$

where $D(\Psi(X_{mb1}, k), \Psi(X_{mb2}, k))$ is the clustering distance for clustering methods $\Psi(X_{mb1}, k)$, $\Psi(X_{mb2}, k)$, $k = 1, 2, \ldots, K$, which are based on the B independent bootstrap sampling pairs (X_{mb1}, X_{mb2}), $b = 1, 2, \ldots, B$ where each sample has N cases.

The maximum number of clusters K is set to be $K = \sqrt{N/2}$ in our numerical examples [4, 15]. However, this value may not fit all kinds of datasets. If $\hat{k} = K$, we need to increase the maximum number of clusters K and auto-search the location of the maximum stability value.

MI-based cross-validation for incomplete big web trial data in eHealth—The MI-based clustering stability using cross-validation for k clusters is expressed by,

$$\text{MI-} S_{CV}(t) = \frac{1}{MU} \sum_{m=1}^{M} \sum_{u=1}^{U} \sum_{i<j} V_{ij}^*(X_1^*, X_2^*, k), \quad (13)$$

where $V_{ij}^*(\cdot)$ is clustering similarity, which is equal to

$I\left(I\left(\psi_1^*(x_i^*) = \psi_1^*(x_j^*)\right) + I\left(\psi_2^*(x_i^*) = \psi_2^*(x_j^*)\right) = 1\right)$, U is the number of permutations, $\psi_1^*(x_i^*) = \Psi(X_1^*, k)$ and $\psi_2^*(x_i^*) = \Psi(X_2^*, k)$ are two clustering methods, $X^* = \{x_1^*, x_2^*, \ldots, x_n^*\}$ is a permutation on the m-th imputed dataset, $X_1^* = \{x_1^*, x_2^*, \ldots, x_c^*\}$, $X_2^* = \{x_{c+1}^*, x_{c+2}^*, \ldots, x_{2c}^*\}$ and $X_3^* = \{x_{2c+1}^*, x_{2c+2}^*, \ldots, x_{n-c}^*\}$ are the splits of X^*. Overall, the higher MI-S_{BS} and MI-S_{CV}, the better the clustering stability.

MI-based Xie and Beni (MI-XB) index for incomplete big web trial data in eHealth

The XB index has been used in fuzzy clustering validation since it was proposed in 1991 [33]. It is defined as the quotient between the means of the quadratic error and the minimum

of the minimal squared distance between the points and cluster centroids. The XB index can be calculated by,

$$XB = \frac{\sum_{i=1}^{N}\sum_{j=1}^{c} f_0{}^m \|x_i - v_j\|^2}{N \cdot \min_{i,k} \|x_i - v_k\|^2}, \quad (14)$$

in which x_i, $i = 1, 2, \ldots, N$ are the cases, N is the number of cases, c is the number of clusters, v_k, $k = 1, 2, \ldots, c$ are the cluster centroids and m is fuzziness. A smaller XB index value indicates a partition that all clusters are compact and separate to each other, which means a "better clustering. Thus, we find the optimal number of clusters by minimizing the XB indices over a set of number of cluster. The MI-based XB index is represented as,

$$MI\text{-} XB(k) = \frac{1}{M} \sum_{m=1}^{M} XB_{q,k}, \quad (15)$$

in which $XB_{q,k}$ is the XB index for clustering q-th imputed dataset for k clusters, and M is the number of imputations. The smaller the MI-XB, the better the clustering. The XB indices are calculated for a set of number of clusters and the optimal number of clusters is identified with the minimal XB value.

MI-based modularity for incomplete big web trial data in eHealth

In recent years, network-based validation approach has been used for clustering data, where the data vectors are treated as "nodes" in the graph and the similarities between two data vectors are defined as the "edges" between them. Suppose N vector nodes $\mathbf{n}_i$, $i = 1, 2, \ldots, N$ represent the N cases, the Gaussian radial basis function kernel (RBF) is used to calculate the similarities between these nodes. The similarity between nodes $\mathbf{n}_i$ and $\mathbf{n}_j$, $1 \leq i, j \leq N$ is defined as,

$$W(\mathbf{n}_i, \mathbf{n}_j) = \exp\left(\gamma \|\mathbf{n}_i - \mathbf{n}_j\|^2\right). \quad (16)$$

Note if $i = j$ the similarity between $\mathbf{n}_i$ and $\mathbf{n}_j$ is 1, which means that there is a self-loop in the graph. Here, the similarity means how a vector is similar to its neighbors not to itself, thus

$$W(\mathbf{n}_i, \mathbf{n}_j) = \begin{cases} \exp\left(\gamma \|\mathbf{n}_i - \mathbf{n}_j\|^2\right), & \text{if } i \neq j \\ 0, & \text{otherwise} \end{cases} \quad (17)$$

Modularity has been widely used in finding communities in network mining. The modularity Q for a weighted network is calculated by,

$$Q = \frac{1}{2e} \sum_{i,j} \left[W_{ij} - \frac{d_i d_j}{2e} \right] \delta\left(v_i, v_j\right), \quad (18)$$

in which d_i and d_j are nodes strength, $d_i = \sum_j W_{ij}$ and $d_j = \sum_i W_{ij}$, e is the total strength of the network, $e = \frac{1}{2}\sum_i d_i$. v_i and v_j are the cluster membership of the i-th and j-th nodes; $\delta(v_i, v_j) = 1$ only when $v_i = v_j$ and $\delta(v_i, v_j) = 0$, otherwise.

The MI-based Modularity (MI-Q) is calculated by,

$$\text{MI-} Q(k) = \frac{1}{M}\sum_q Q_{q,k}. \quad (19)$$

Note that if $\hat{k} = K$, we need to increase K and compare MI-Q to find the optimal number of clusters. The higher MI-Q, the better the clustering. The entirely procedure of the proposed MIV framework is shown in Algorithm 1.

Algorithm 1

The proposed MIV algorithm

Require: Incomplete dataset:X; Number of Imputation: M;

     Maximal number of clusters: K

Ensure: MI-based validation: MIV

1:   Expectation Maximization (EM):

2:

$$\log L(\theta|Y_{obs}) = \sum_{g=1}^{G} \log L_g(\theta|Y_{obs})$$

3:   Posterior: $\log P(\theta|Y_{obs}) = L(\theta|Y_{obs}) + \log \pi(\theta)$

4:   **for** i from 1 to M **do**

5:      **for** t from 1 to N_{bi} **do**

6:

        I-step: $Y_{mis}^{(t+1)} \sim P(Y_{mis}|Y_{obs}, \theta^{(t)})$

7:

        P-step: $\theta^{(t+1)} \sim P(\theta|Y_{obs}, Y_{mis}^{(t+1)})$

8:      **end for**

9:      Imputed X_{MIi} by an additional I-step

10:      **for** k from 1 to K **do**

11:        Clustering data X_{MIi} to k clusters:

12:      $IDX_{i,k} = MIfuzzy (X_{MIi}, k)$

13:      **end for**

14:   **end for**

15:   Calculate MI-based validity indices by Eqs. (12), (13), (15) and (19).

In the proposed MIV algorithm, each imputed dataset is analyzed and the results of all imputed data are combined to obtain the validation for the incomplete data. The computation complexity of the MIV algorithm is $\mathcal{O}(rN\,dMK)$, in which r is missing rate, N is the number of cases, d is the dimensions, M is the number of imputation, and K is the maximal number of clusters.

Numerical analyses and simulation

Our MI based Validation (MIV) algorithms were first evaluated using the big data from a longitudinal web-delivered trial for smoking cessation (called QuitPrimo, see details in [63, 64]). Briefly, QuitPrimo study aims to evaluate an integrated informatics solution to increase access to web-delivered smoking cessation support. The trail includes 1320 cases with missing rate less than 8.4 %. The three intervention web trail components are 1) My Mail, 2) Online Community, and 3) Our Advice. As aforementioned, this big web trial data set is unstructured and formatted simply as time, e.g., each smoker has data like "27APR10:15:43:00". However, the primary values of big data come not from its raw form, but from its processing and analysis. Four clusters were identified using six monthly measures for each intervention component and web duration (total 19 attributes) in [63, 64].

Ten imputations ($M = 10$) are used according to [23]. Applying our MIV algorithm introduced in Section "Multiple-imputation-based validation framework (MIV) for incomplete big web trial data in eHealth," we auto-compute, search, and synthesize, the results for MI-based clustering stability, i.e., MI-Bootstrap and MI-Cross Validation, (MI-S_{BS} and MI-S_{CV}), as well as MI-XB and MI-Q.

Figure 3a displays the MI clustering stability indices, MI-S_{BS} and MI-S_{CV}, obtained by bootstrapping and cross-validation, respectively. The MI-S_{BS} shows the stability achieves the highest at 3 clusters, while the MI-S_{CV} indicates the 5 clusters. The minimal value of MI-XB in Fig. 3b clearly points to 4 clusters which is the correct optimal number of clusters. Figure 3c also indicates 2 clusters based on MI-based modularity (MI-Q). These results demonstrate that the stabilities and network-based validation methods may not be suitable for big longitudinal web trial data analyses.

Figure 4 shows the four identified behavioral trajectory patterns of this big web-delivered trial. The x-axis shows the time slots for the three web intervention components, My Advice, Our Advice, and Our Community; the y-axis displays individual IDs, and z-axis are the counts of each component. The colored trajectory layers represent the average engagement level for each cluster. In QuitPrimo data, $r = 0.084$, $N = 1320$, $d = 18$, $M = 10$, and $K = 10$, the running time of the proposed MIV algorithm is about 1 minute on our lab PC (i7-4770 double 3.4GHz CPU with 16G RAM).

Our simulation uses the joint zero-inflated Poisson (ZIP) and autoregressive (AR) model to simulate the QuitPrimo data [65]. We first train the joint model using the QuitPrimo data, to obtain the parameters which were used to simulate a bigger longitudinal web trial data with 10,000 cases and 54 dimensions (9 variables with 6 repeated meatures each). Then we evaluated our proposed MIV algorithms on the simulated data. Figure 5 again demonstrates that MI-XB (Fig. 5b) correctly identifies the 4 trajectory patterns while MI-S_{BS} and MI-S_{CV} (Fig. 5a) and MI-Q (Fig. 5c) did not. Our preliminary evaluation results [14] indicate that MIfuzzy is most robust to missing rates less than 20 %, although one empirical observational study showed that it could be robust to the missing rate up to 40 % where other included variables with missing values may be more or as informative as the variables without missingness for the subjects [14].

Conclusion

In eHealth services, big data from web-delivered longitudinal trials are complex. Determining the optimal number of clusters in such data is especially challenging. This paper built upon our MIfuzzy clustering designed a MI-based validation (MIV) framework and algorithms for big data processing, particularly for fuzzy clustering of big incomplete longitudinal web-delieved trial data. Although we included two conventional methods for testing clustering stability, bootstrap and cross-validation, they did not seem to add incremental value for detecting the optimal number of clusters. One major reason could be that the multiple imputation component in MIfuzzy already accounts for the imputation uncertainty to ensure the clustering stability using several complete imputed datasets. This concept is similar to the bootstrap and cross validation for stability tests, therefore this overlap decreases the incremental value of these conventional methods which are typically used for complete data sets. Another reason might be that the two methods were not specifically designed for or directly related to the fuzzy clustering which is widely accepted for biomedical data where clusters overlap or touch. Also the modularity validation index is widely accepted for network-based data, but appears not feasible for the structure of these big incomplete longitudinal web-delivered trial data in eHealth services. Consistently, we found multiple-imputation based XB index, specifically designed for fuzzy clustering, could facilitate detecting the optimal number of clusters for big incomplete longitudinal trial data, either from web-delivered or traditional clinical trials [4, 6, 15]. Different from the MI approach used for statistical analyses, MI based clustering only uses the imputation step, thus has no connection with the possible inconsistent analytical models for statistical inference. As our research indicates, it will especially contribute more to non-model-based clustering approaches, and could potentially improve clustering accuracy and computational efficiency for model-based clustering approaches. In future, embedding MIV algorithms into eHealth system could warrant the validity of identifying at-risk or abnormal patterns of patients, events, diagnoses or services using various unsupervised learning methods, and reduce the uncertainty in implementing pattern-derived adaptive trials or services.

Acknowledgments

References

1. Eysenbach G. Group C-E. Consort-ehealth: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res. 2011; 13(4)

2. Fang H, Zhang Z, Wang CJ, Daneshmand M, Wang C, Wang H. A survey of big data research. IEEE Netw. 2015; 29:6–9. [PubMed: 26504265]

3. Fang H, Espy KA, Rizzo ML, Stopp C, Wiebe SA, Stroup WW. Pattern recognition of longitudinal trial data with nonignorable missingness: An empirical case study. Int J Inf Technol Decis Mak. 2009; 8(03):491–513. [PubMed: 20336179]

4. Fang, H.; Dukic, V.; Pickett, KE.; Wakschlag, L.; Espy, KA. Detecting graded exposure effects: A report on an east boston pregnancy cohort. Nicotine & Tobacco Research; 2012. p. ntr272

5. Fang H, Zhang Z, Huang H, Wang Jingfang Huang. Validating patterns for longitudinal trial data. Section on Statistics in Epidemiology Joint Statistical Meeting, American Statistical Association. 2014

6. Zhang, Z.; Fang, H.; Wang, H. Visualization aided engagement pattern validation for big longitudinal web behavior intervention data. the 17th international Conference on E-health Networking, Application & Services. (IEEE Healthcom'15); 2015; Accepted

7. McLachlan, G.; Peel, D. Finite mixture models. Wiley; 2004.

8. Fran•ois O, Ancelet S, Guillot G. Bayesian clustering using hidden markov random fields in spatial population genetics. Genetics. 2006; 174(2):805–816.

9. Gan, G.; Ma, C.; Wu, J. Data clustering: theory, algorithms, and applications. Vol. 20. Siam; 2007.

10. Kubat, M. Neural networks: a comprehensive foundation by simon haykin, macmillan, 1994. 1999.

11. Bezdek, JC.; Keller, J.; Krisnapuram, R.; Pal, N. Fuzzy models and algorithms for pattern recognition and image processing. Vol. 4. Springer Science & Business Media; 2006.

12. Schafer, JL. Analysis of incomplete multivariate data. CRC press; 1997.

13. Little, RJ.; Rubin, DB. Statistical analysis with missing data. Wiley; 2014.

14. Zhang Z, Fang H. Multiple- vs non- or single-imputation based fuzzy clustering for incomplete longitudinal behavioral intervention data. Chase. 2016 Submitted.

15. Fang H, Johnson C, Stopp C, Espy KA. A new look at quantifying tobacco exposure during pregnancy using fuzzy clustering. Neurotoxicol Teratol. 2011; 33(1):155–165. [PubMed: 21256430]

16. Rubin, DB. Multiple imputation for nonresponse in surveys. Vol. 81. Wiley; 2004.

17. Schafer, JL. Analysis of incomplete multivariate data. CRC press; 1997.

18. Royston P. Multiple imputation of missing values. Stata J. 2004; 4:227–241.

19. Royston P. Multiple imputation of missing values: update of ice. Stata J. 2005; 5(4):527.

20. Little RJ. A test of missing completely at random for multivariate data with missing values. J Am Stat Assoc. 1988; 83(404):1198–1202.

21. Rubin DB. Inference and missing data. Biometrika. 1976; 63(3):581–592.

22. Rubin, DB. Multiple imputation for nonresponse in surveys. Vol. 81. Wiley; 2004.

23. Rubin DB. Multiple imputation after 18+ years. J Am Stat Assoc. 1996; 91(434):473–489.

24. Klir, G.; Yuan, B. Fuzzy sets and fuzzy logic. Vol. 4. Prentice Hall; New Jersey: 1995.

25. Zadeh LA. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Set Syst. 1997; 90(2):111–127.

26. Fang H, Rizzo ML, Wang H, Espy KA, Wang Z. A new nonlinear classifier with a penalized signed fuzzy measure using effective genetic algorithm. Pattern Recogn. 2010; 43(4):1393–1401.

27. Acock AC. Working with missing values. J Marriage Fam. 2005; 67(4):1012–1028.

28. Donders ART, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol. 2006; 59(10):1087–1091. [PubMed: 16980149]

29. Little RJ, Rubin DB. The analysis of social science data with missing values. Sociol Methods Res. 1989; 18(2–3):292–326.

30. Afifi A, Elashoff R. Missing observations in multivariate statistics i. review of the literature. J Am Stat Assoc. 1966; 61(315):595–604.

31. Buck SF. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. J R Stat Soc Ser B Methodol. 1960:302–306.

32. Marker DA, Judkins DR, Winglee M. Large-scale imputation for complex surveys. Survey Nonresponse. 2002:329–341.

33. Xie XL, Beni G. A validity measure for fuzzy clustering. IEEE Trans Pattern Anal Mach Intell. 1991; 13(8):841–847.

34. Kwon SH. Cluster validity index for fuzzy clustering. Electron Lett. 1998; 34(22):2176–2177.

35. Halkidi M, Batistakis Y, Vazirgiannis M. On clustering validation techniques. J Intell Inf Syst. 2001; 17(2–3):107–145.

36. Newman ME. Modularity and community structure in networks. Proc Natl Acad Sci. 2006; 103(23):8577–8582. [PubMed: 16723398]

37. Newman, M. Networks: an introduction. Oxford University Press; 2010.

38. Ben-Hur A, Elisseeff A, Guyon I. A stability based method for discovering structure in clustered data. Pac Symp Biocomput. 2001; 7:6–17.

39. Lange T, Roth V, Braun ML, Buhmann JM. Stability-based validation of clustering solutions. Neural Comput. 2004; 16(6):1299–1323. [PubMed: 15130251]

40. Ben-David, S.; Von Luxburg, U.; Pal, D. A sober look at stability of clustering. Proceedings of the Annual Conference on Computational Learning Theory; 2006;

41. Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc. 2002; 97(458):611–631.

42. Raftery AE, Dean N. Variable selection for model-based clustering. J Am Stat Assoc. 2006; 101(473):168–178.

43. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL. Model-based clustering and data transformations for gene expression data. Bioinformatics. 2001; 17(10):977–987. [PubMed: 11673243]

44. Ng AY, Jordan MI, Weiss Y, et al. On spectral clustering: Analysis and an algorithm. Adv Neural Inf Proces Syst. 2002; 2:849–856.

45. Von Luxburg U. A tutorial on spectral clustering. Stat Comput. 2007; 17(4):395–416.

46. Zelnik-Manor L, Perona P. Self-tuning spectral clustering. In. Advances in neural information processing systems. 2004:1601–1608.

47. Efron B. Bootstrap methods: another look at the jackknife. Ann Stat. 1979:1–26.

48. Efron, B.; Tibshirani, RJ. An introduction to the bootstrap. CRC Press; 1994.

49. Varian H. Bootstrap tutorial. Math J. 2005; 9(4):768–775.

50. Davison, AC. Bootstrap methods and their application. Vol. 1. Cambridge University Press; 1997.

51. Beran R. Prepivoting test statistics: a bootstrap view of asymptotic refinements. J Am Stat Assoc. 1988; 83(403):687–697.

52. Bickel PJ, Freedman DA. Some asymptotic theory for the bootstrap. Ann Stat. 1981:1196–1217.

53. Shao J. Linear model selection by cross-validation. J Am Stat Assoc. 1993; 88(422):486–494.

54. Zhang P. Model selection via multifold cross validation. Ann Stat. 1993:299–313.

55. Yang Y. Comparing learning methods for classification. Stat Sin. 2006; 16(2):635.

56. Tibshirani R, Walther G. Cluster validation by prediction strength. J Comput Graph Stat. 2005; 14(3):511–528.

57. Kohavi R, et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In. Ijcai. 1995; 14:1137–1145.

58. Refaeilzadeh, P.; Tang, L.; Liu, H. Encyclopedia of database systems. Springer; 2009. Cross-validation; p. 532-538.

59. Leicht EA, Newman ME. Community structure in directed networks. Phys Rev Lett. 2008; 100(11):118703. [PubMed: 18517839]

60. Von Luxburg U. A tutorial on spectral clustering. Stat Comput. 2007; 17(4):395–416.

61. Sas, I. Sas/stat® 9.2 user's guide. SAS Institute Inc; Cary: 2008.

62. Wang J. Consistent selection of the number of clusters via crossvalidation. Biometrika. 2010; 97(4):893–904.

63. Houston TK, Sadasivam RS, Ford DE, Richman J, Ray MN, Allison JJ. The quit-primo provider-patient internet-delivered smoking cessation referral intervention: a cluster-randomized comparative effectiveness trial: study protocol. Implement Sci. 2010; 5:87. [PubMed: 21080972]

64. Houston TK, Sadasivam RS, Allison JJ, Ash AS, Ray MN, English TM, Hogan TP, Ford DE. Evaluating the quit-primo clinical practice eportal to increase smoker engagement with online cessation interventions: a national hybrid type 2 implementation study. Implement Sci. 2015; 10(1):154. [PubMed: 26525410]

65. Zhang Z, Fang H, Wang H. A new mi-based visualization aided validation index for trajectory pattern recognition of big longitudinal web trial data. IEEE ACCESS. 2015 accepted.
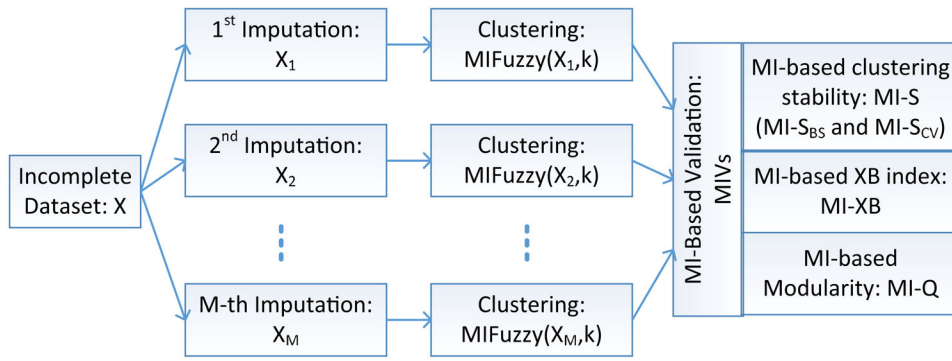
Fig. 1.
The proposed MIV platform for big web trial data in eHealth
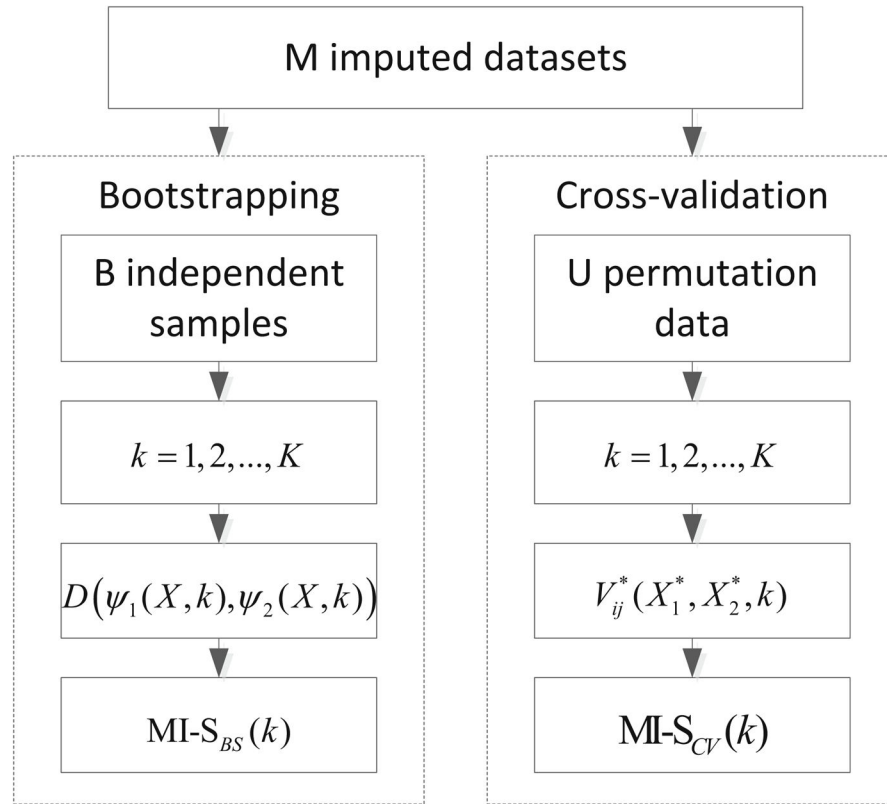
Fig. 2.
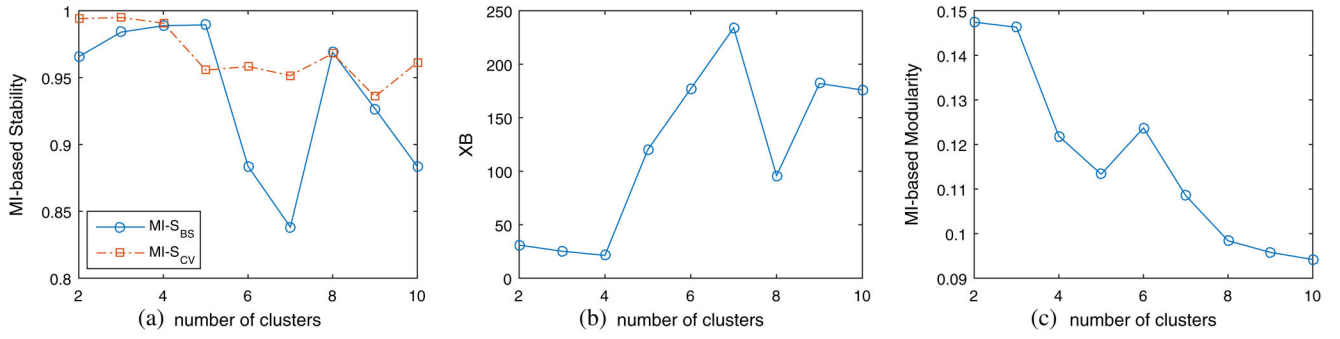Illustrative procedure of MI-based stability algorithm

Fig. 3.
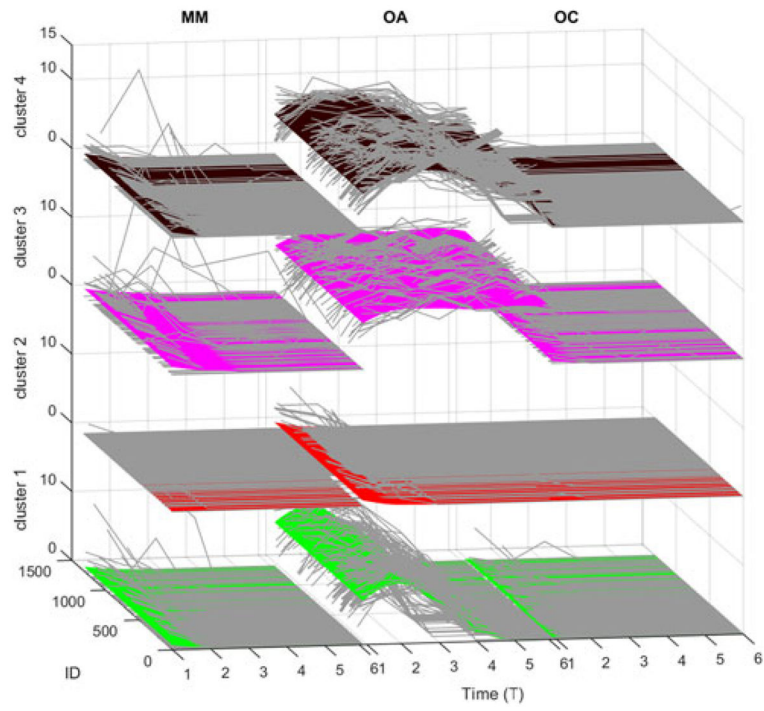MI-based validation indices for a big web-delivered trial dataset (QuitPrimo)

Fig. 4.
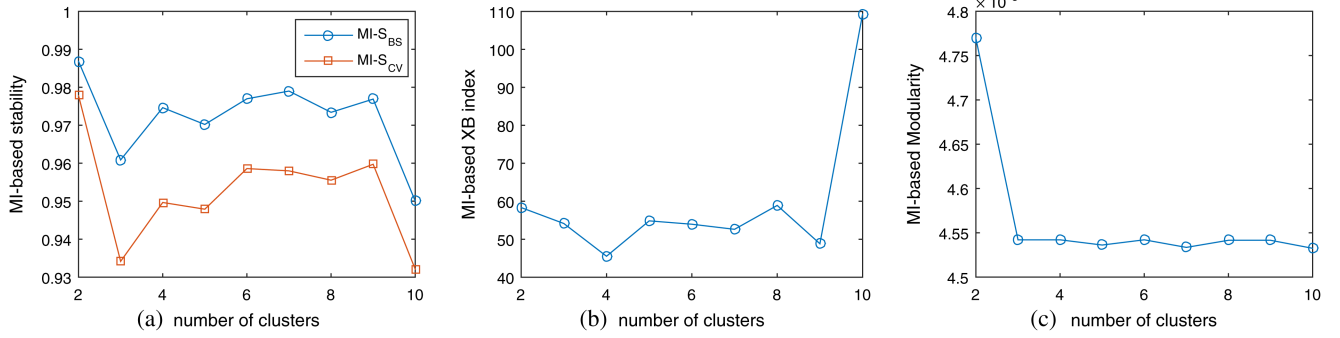The identified big longitudinal trajectory clusters of QuitPrimo data

Fig. 5.
MI-based validation indices for simulated big web-delivered trial dataset

Table 1

Notations

Symbol	Description
O	Observations
ψ	Clustering method
X, Y	Independent sampling from observations
$\Psi(X, k)$	Results of clustering data X into k clusters
$D(\psi_1, \psi_2)$	clustering distance
$V_{ij}^*(X_1^*, X_2^*, k)$	clustering similarity
M	Number of imputations
r	Missing rate
d	Dimensions
N	Number of observations
N_{bi}	Number of burn-in iteration
K	Maximal number of clusters
B	number of bootstrapping
U	number of permutation
s	Clustering stability
MI-S	MI-based stability
MI-S$_{BS}$	MI-based stability using bootstrapping
MI-S$_{CV}$	MI-based stability using cross-validation
MI-XB	MI-based Xie and Beni index
MI-Q	MI-based modularity index