



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

보건학 석사 학위논문

Feature Selection with Particle
Swarm Optimization Substantially
Improves the Accuracy of Missing
Data Imputation for a Large-scale
Data

BPSO 기반 변수 선택 기법으로 보정한 결측치 대체
알고리즘 개발

2021 년 02 월

서울대학교 보건대학원
보건학과 및 보건통계학전공
정수린

Feature Selection with Particle Swarm Optimization Substantially Improves the Accuracy of Missing Data Imputation for a Large-scale Data




BPSO 기반 변수 선택 기법으로 보정한 결측치 대체
알고리즘 개발

지도교수 원 성 호

이 논문을 보건학석사 학위논문으로 제출함
2020 년 11 월

서울대학교 보건대학원
보건학과 보건통계학전공
정 수 린

정수린의 석사 학위논문을 인준함
2021 년 02 월

위 원 장	김 호	
부 위 원 장	이우주	
위 원	원성호	

ABSTRACT

Feature Selection with Particle Swarm Optimization Substantially Improves the Accuracy of Missing Data Imputation for a Large-scale Data

Surin Jung

Department of Public Health
The Graduate School of Public Health
Seoul National University

Introduction

Missing data are common problem in large scale data setting. Handling missing data appropriately is crucial in data analysis. Missingness can be categorized into the missing completely at random(MCAR), (2) missing at random(MAR), and (3) missing not at random(MNAR)^{1,7}. Different types of missingness mechanism need different imputation strategy². Multiple Imputation - an approach for averaging the outcomes across multiple imputed data is more suitable than single imputation dealing with various missing mechanism^{2,7}. The missForest is one of the most prevalent multiple imputation method³. It is known that missForest has advantages over other imputation method in that it is applicable for mixed type data with non-linearity and interaction and does not require any distributional assumption of the given variables unlike MICE which assumes linearity between the variables^{3,4,5}. However, in a recent study, it is found out that missForest can produce a biased results for non-normal data^{6,8}. Additionally, missForest is computationally expensive⁴. Therefore, we developed missForest algorithm by combining BPSO based feature selection strategy.

Methods

Binary Particle Swarm Optimization(BPSO) is an evolutionary algorithm well-known for the global search ability and computational efficiency. Combining BPSO based feature selection step prior to impute missing values with missForest, imputation accuracy for continuous variables can be increased by pruning redundant variables.

Results

The missForest is one of the most prevalent missing data imputation method since it can be applied to mixed-type data and does not need distributional assumption. However, it turned out that missForest can produce a biased results for non-normal data. Thus, we improve the imputation accuracy of missForest by selecting important features using BPSO algorithm. BPSO is an evolutionary algorithm and also well-known for its global optimization and efficient computing. In this study, BPSO shows better imputation accuracy than missForest with respect to the continuous variables by feature selection prior to the imputation step.

Keywords: Feature selection, BPSO, missForest, Imputation, Missing
Student Number: 2019-22081

Table of Contents

I. Introduction	5
II. Methods	
1. Data Description	7
2. MissForest	8
3. Particle Swarm Optimization(PSO)	8
4. Binary Particle Swarm Optimization(BPSO)	10
5. Feature Selection with BPSO	11
6. Simulation Setting	12
III. Results	
1. Descriptive statistics of data	14
2. Imputation Accuracy Comparison between BPSOmf and missForest	14
IV. Discussion	15
V. Reference	17

List of Figures

Figure 1 Illustration of Particle Swarm Optimization	20
Figure 2 Basic terms and scheme of PSO	21
Figure 3 Update Process in PSO	22
Figure 4 Error Rate by data	23
Figure 5 Error Rate by Missing Rate	24
Figure 6 Error Rate by Missing Mechanism	25

Lists of Tables

Table 1. Psuedo code: Feature selection based on BPSO	26
Table 2. The Overall missing rate of 6 dataset	27
Table 3. Experimental design for BPSOmf vs. missForest	27
Table 4. Exclusion Criteria when inducing missing values	28
Table 5. The Number of Variables and Observations of the Datasets used in Simulation Study	29
Table 6 Overall Error rate	30
Table 7. Error Rate by data	30
Table 8. Error Rate by Missing Rate	31
Table 9. Error rate by Missing Mechanism	31
Table 10. Error rate by <i>ntree</i>	32

I . INTRODUCTION

Missing data are common problem in large scale data setting. Handling missing data appropriately is crucial in data analysis. Simply discarding any missing value or replacing it by mean/mode might lead a substantial amount of bias^{1,11,12}. Moreover, it might reduce statistical power. Therefore, a large number of imputation methods have been developed to deal with the problem especially, those based on machine learning techniques such as MICE, KNNI, missForest^{2,3,5,9,11,14}.

Missingness can be categorized into the following three type: (1) missing completely at random(MCAR), (2) missing at random(MAR), and (3) missing not at random(MNAR) Missing completely at random(MCAR) means that causes of missing are irrelevant with the observed or the missing data^{1,2,9}. Neither observed ones nor unobserved ones has a relationship with missing values. Missing at random(MAR) implies that the probability of missingness possibly depends on the observed ones^{1,2,9}. Finally, Missing Not at random (MNAR) is when data are neither MCAR nor MAR. In this case, causes of missing data not only depend on the observed ones but also missing ones^{1,2,9}. In summary, missing data happens with various reasons and keep in mind that especially MNAR type missing data are handled carefully^{1,2,7,9,12}.

As a result, different types of missing mechanism need different imputation strategy and there are various kinds of imputation methods have been developed so far^{7,9}. As for single imputation, for example, Hot-Deck Imputation simply imputes a missing value with a randomly selected similar value⁷. Another technique, Mean Substitution is replacing the missing value with the mean/mode of that variable⁷. Those kind of approaches, Single Imputation, however, is prone to potential bias and may result in severe distortion in statistical inference^{2,9}. Moreover, single imputation is not flexible enough to deal with MAR and MNAR. Therefore, Multiple Imputation - an approach for averaging the outcomes across multiple imputed data is more suitable to deal MAR of MNAR^{2,3,5,9}.

Multiple imputation is an general approach to deal with MAR or MNAR in that it allows the uncertainty about the missingness and average the multiple outcomes⁹. Basically, Multiple Imputation follows below three

steps.

1. **Imputation** – missing values in data are imputed and the imputed values are drawn m times from a distribution rather than just once. At the end of this step, there should be m completed datasets.
2. **Evaluation** – Each of the m datasets is evaluated. At the end of this step there should be m analyses.
3. **Pooling** – The m results are combined into one result by considering the distribution of the variable of concern.

Just as there are multiple methods of single imputation, there are few methods of multiple imputation such as Multivariate Imputation by Chained Equation(MICE) KNN Imputation(KNNI) and missForest^{3,5,14}. Among those methods, The missForest is one of the most prevalent and commonly used imputation method^{3,4,8}. As the name suggest, missForest is an implementation of random forest based imputation algorithm. missForest regards missing data imputation as prediction problems using an random forest model trained on the observed part of the given data³.

The missForest algorithm can be described as follows^{3,4}. Firstly, for a variable with missing data, the missing values will be replaced by its mean or mode(mean for continuous variables and mode for categorical variables) Then, for each variable with missing values, random forest model on the observed part is grown and then the missing part will be predicted and replaced based on the RF model. This process repeats in an iterative process until a stopping criterion is met, or a maximum number of user–specified iterations is reached.

The missForest is one of the best and most widely used method since is has a lot of advantages^{4,6,8}. First of all, according to the original article, it is said that missForest is applicable for mixed type data with non–linearity and interaction. Also, it is known that missForest does not require any distributional assumption of the given variables unlike MICE which assumes linearity between the variables. Moreover, is gives an OOB error estimate for its predictions³. For there reasons, missForest has been known as a standard for non–parametric imputation methods.

However, in a recent study, it is found out that missForest can produce a biased results for non–normal data⁶. Moreover, when there are interactions between variables then the imputed variable can be highly

skewed⁶. Additionally, missForest is computationally expensive in that forest must be grown for each variable and the algorithm runs until it converge^{6,8}. In conclusion, it is controversial that missForest still performs the best when p is large and there are too many redundant variables. Thus, pruning irrelevant variables would be the key factor which can help to increase the missForest performance. Therefore, we developed missForest algorithm by combining BPSO based feature selection strategy. The examination of imputation accuracy of missForest with or without feature selection was done through survey data.

II. METHODS

1. Data Description

KoGES Ansan and Ansong study is a part of Korean genome and epidemiology study(KoGES) project. KoGES Ansan and Ansong study consists of men and women, lives in Ansan and Ansong, aged between 40~69 years at baseline. Comprehensive list of variables such as medical history, lifestyle, clinical examination and biospecimens(serum, plasma, urineand DNA) were collected since baseline recruitment in 2001–2001 up to 7th follow–up¹⁵. Out of the 10,030 baseline participants and 3,205 variables, 7th follow–up was conducted in 6318 participants and 1639 variables. All data used in comparative experiments composed of mixed–type variable.

Korea National Health and Nutrition Examination Survey(KNHANES) was collected since 1998 and the survey contains health and dietary, nutritional status of Koreans. The 8th follow–up(2019–2021) data is still being investigated and only the 7th follow–up(2016–2018) data was used in this study¹⁶. There are 8,150 participants and 799 variables, 8,127 participants and 857 variables, 7,992 participants and 785 variables each in 2016, 2017, 2018 data.

In this study, above all, it should be noted that the analysis presented here was to evaluate the newly developed imputation methods, and is not intend for definitive analysis of the data.

2. MissForest

The missForest is an implementation of random forest based imputation algorithm and regards missing data imputation as prediction problems using an random forest model trained on the observed part of the given data. The missForest algorithm can be described as follows³.

Let the data matrix $X = (X_1, X_2, \dots, X_p)$ to be a size of $n \times p$ matrix. In missForest, X is divided into four different part.

- (1) The observed part of variable X_i , denoted by $y_{obs}^{(s)}$
- (2) The missing part of variable X_i , denoted by $y_{mis}^{(s)}$
- (3) The variable other than X_i , with observation $i_{mis}^{(s)}$ denoted by $x_{mis}^{(s)}$
- (4) The variable other than X_i , with observation $i_{mis}^{(s)} / i_{obs}^{(s)}$ denoted by $x_{obs}^{(s)}$

Firstly, for a variable with missing data X , the missing values will be replaced by its mean or mode ; mean for continuous variables and mode for categorical variables. Then, for each variable with missing values, X_i build a random forest model on the observed part $y_{obs}^{(s)}$ and $x_{obs}^{(s)}$ is grown and then the missing part, $y_{mis}^{(s)}$ will be predicted and replaced based on the random forest model. Thess processes are repeated until a stopping criterion is met, or the maximum number of iterations is reached.

3. Particle Swarm Optimization(PSO)

The Particle Swarm Optimization(PSO) is an evolutionary algorithm proposed by Kennedy and Eberhart in 1995¹⁷. The PSO algorithm was inspired by the social behavior of bird flocking or fish schooling. For example, bird flocking has implicit rules which enable the group of birds to move simultaneously while dispersing suddenly and gathering again^{17,18}.

Before describing the detail of PSO algorithm, there are a few terms are

defined below and also example illustration presented in the Figure 1. In PSO, the *population* is a search space and the population can be consist of multiple candidate solutions, called *particles*. In PSO, each particle has its own *position* and *velocity*. The *position* and the *velocity* of each *particle* is iteratively updated in search space to move towards the best objective value.

First of all, PSO algorithm begins with the *population initialization* with random *particles*. The *position* of each *particle* is updated iteratively searching for the best solution. The current *position* of *particle* i is denoted by a vector of $x_i = (x_{i1}, x_{i2}, \dots, x_{id})$, where D is the dimension of the *population*. The *velocity* of the *particle* i is denoted by a vector $v_i = (v_{i1}, v_{i2}, \dots, v_{id})$.

During iteration, the $k+1^{th}$ *position* and the $k+1^{th}$ *velocity* is updated iteratively by evaluating the *fitness* of current k^{th} *position* and k^{th} *velocity*. The fitness of each *particle* is calculated by any suitable *fitness score function*. For example, Bayesian information criterion based on logistic regression could be the fitness function. In this study AUC, ROC, RMSE were used as a *fitness score function*. This process is repeated until a stopping criterion met, or a maximum number of iterations is reached.

The fitness score of each *particle* is recorded to update *pbest* and *gbest*. *pbest* is the best previous *position* obtained as a personal best and *gbest* is the best *position* obtained by the population so far. Updating the *pbest* and *gbest*, PSO searches for the optimal solution by updating the *velocity* and the *position* of each *particle* according to the following equations ①, ②.

$$v_{id}^{k+1} = wv_{id}^k + c_1r_1(pbest_{id} - x_{id}^k) + c_2r_2(gbest_{id} - x_{id}^k) \dots \textcircled{1}$$

$$x_{id}^{k+1} = x_{id}^k + v_{id}^{k+1} \dots \textcircled{2}$$

In equations ①, ②, k denotes the k th iteration in the search process and p denotes the p th dimension in the *population*. v represent the velocity, w is inertia weight, c_1 and c_2 are acceleration constants(learning

factor), r_1 and r_2 are random values distributed Uniform(0,1). $pbest_{id}$ and $gbest_{id}$ stands for each element of $pbest$ and $gbest$. Note that the w , c_1 , c_2 , r_1 , r_2 is the arbitrarily defined parameters where $pbest_{id}$ and $gbest_{id}$ should be obtained in the *fitness evaluation* step. Those update process scheme presented above is illustrated in Figure 3.

The equation ①, ② is quite straightforward and implies the two basic rules enables the PSO algorithm to find global and local optimum. Firstly, the new velocity in $k+1^{th}$ iteration, v_{id}^{k+1} is the linear combination of x_{id}^k , v_{id}^k , $pbest_{id}$ and $gbest_{id}$ weighted by some constants. This makes the update would be reflected in x_{id}^{k+1} and v_{id}^{k+1} be able to respond to fitness scores within the personal best but also commit to the global optimum^{17,18,19}.

Secondly, both x_{id}^{k+1} and v_{id}^{k+1} is affected by the previous values x_{id}^k , v_{id}^k . Basically, population should be robust to the drastic change. However, at the same time, the population also need to change its behavior when if it is worthy. This principle is well demonstrated in equation ①, ② where x_{id}^{k+1} and v_{id}^{k+1} are being kept updated in each iteration while the new values are generated from the previous one^{17,18,19}.

Thanks to those properties, PSO algorithm is good at seeking global and local optimum. Moreover, it has some computational advantages over other evolutionary algorithms such as GAs since it has fewer parameters and this allows the PSO to be easy to be implemented^{19,20,21}. These properties of PSO is competitive when the pool of candidate solution is very large with limited available resources¹⁸.

4. Binary Particle Swarm Optimization(BPSO)

To solve feature selection problem, the position in BPSO should follow the binary coding principle. A population of candidate solutions are encoded as a *particles* in search space. The candidate solutions, *positions* are encoded in the binary string where “1” represents that the feature is selected and “0” otherwise. Thus, in BPSO, the *velocity* means *the probability of the corresponding element in the position taking value 1*. So,

a sigmoid function $s(v_{id})$ is applied to transform v_{id} to the range of (0,1).

$$x_{id} = \begin{cases} 1, & \text{if } rand() < s(v_{id}) \\ 0, & \text{otherwise} \end{cases} \dots\dots\dots \textcircled{3}$$

where

$$s(v_{id}) = \frac{1}{1 + e^{-v_{id}}} \dots\dots\dots \textcircled{4}$$

5. Feature Selection with BPSO

In this study, feature selection based on BPSO was combined with missForest. Thus, firstly, when imputing i^{th} variable with missing values, select p' variables in the dataset based on BPSO. Then, impute i^{th} variable with selected p' variables with missForest.

Suppose that the $n \times p$ data matrix has the form of $X = (X_1, X_2, \dots, X_p)$ and there are total m variables having any missing. Variable selection with BPSO starts with the *Population Initialization*. In *population initialization* step, build simple regression models by p variables and assess those models with AIC to select first p' variables with low AIC.

The total number of possible subsets of p' variables is $2^{p'}$. Randomly select k *particles* with selected p' variables. The group of k particles is so-called *population* in BPSO and each p' variable is the *particle*. As mentioned above, *positions* are encoded in the binary string where 1 represents that the variable is selected and 0 otherwise and the *velocity* means *the probability of the corresponding element in the position taking value 1*.

After *population initialization* step is done, evaluate the fitness of each particle. In the *fitness evaluation step*, build a randomForest model having i^{th} variable as a response and k^{th} *particle* as explanatory variables. Then, the fitness scores of k^{th} *particles* can be ROC, AUC or RMSE depending on the type of i^{th} variable. The psuedo code of feature selection based on

BPSO is presented in Table 1.

In the simulation, the number of *Iteration* is set to 3 with the number of particles $p' = 20$ and search space dimension $D = \text{floor}(\sqrt{\# \text{ of variable}})$. with the weight inertia $w = 0.3$, the acceleration constants $c_1 = 0.3$ and $c_2 = 0.6$ and r_1 and r_2 from the $Uni(0,1)$. Once feature selection process has done, impute missing values in the dataset with the selected variable subset and missForest. Using missForest package in R, the maximum number of iterations is set to 3 with $n\text{tree} = 100$, $\text{replace} = FALSE$ and other options are set to default.

By combining feature selection step prior to impute missing values of large-scale data, the computation time would decrease since feature selection with BPSO is less demanding than other evolutionary based feature selection method such as GA. Moreover, by pruning redundant variables, imputation accuracy can be increased¹⁸.

6. Simulation Setting

Data used in this simulation study is not a complete dataset. Table 2 shows the overall missing rate of 6 data. The missing rate in each data distributed from 0.230 to 0.571. The reason why incomplete dataset were used is that synthesis data can't perfectly mimic the complicated, diverse and vague characteristics of real world dataset such as MAR and MNAR, class imbalance. For that reason, missing values were arbitrarily made within the observed part so that imputed values were respectively compared with ground truths.

After applying the two imputation method of BPSOmf and missForest, for continuous variables, NRMSE are calculated and for categorical(binary and multiclass) variable, PFC are calculated to quantify the error rate.

The main goal of this simulation study is comparing the performance and computational efficiency of BPSO + missForest and missForest. Three different experiments were carried out to assess the efficacy of the feature selection algorithm. For convenience, **BPSOmf** and is the abbreviation for **BPSO+missForest** from now on.

Firstly, in terms of missing data imputation of i^{th} variable, only the

selected features were used in BPSOmf setting. On the other hand, all features in the dataset were used in missForest setting.

Secondly, three percentages(30%, 50%, 70%) of missing values was induced from the given dataset to investigate the performance of BPSOmf with a different levels of missing rate, from low to high. Lastly, MCAR and MAR missing mechanism were used when generating missing values. In summary, all simulation settings can be represented as the Table 3. The Simulation was repeated 10 times for each experiment.

Before producing missingness with MCAR and MAR mechanism, the missing values were generated *only* for those variables satisfying the followig criteria in Table 4. As mentioned above, in the simulation study, *incomplete datasets* were used which possibly interrupt the imputation, for example, random forest model building, missing values prediction. Those criteria in Table 4 is the minimum restriction for filetering the candidate variables which will be used when inducing missing values.

To induce MCAR missingness in the data, only for the observed part of data, randomly select 50 variables and specified percentage(30%, 50%, 70%) of values of the selected variables were replaced by missing values.

For MAR, the following procedures were used. Firstly, filter the variables with the exclusion criteria presented in Table 5. Then, among the candidate variables, select 25 pairs of variables and missings were assigned by any of two variables depending on the observed value of another variable.

In detail, let X_j be the one of the selected variables and be the n dimensional vector, $X_j = (X_{1j}, X_{2j}, \dots, X_{nj})$. Likewise, the another variable X_k has the same structure with $X_k = (X_{1k}, X_{2k}, \dots, X_{1k})$. Each coordinate of X_j was made missing according to the tail behavior of a X_k , where $k \neq j$. The probability of selecting coordinate $X_{i,j}$ was

$$P\{\text{selecting } X_{i,j} | B_j\} \propto \begin{cases} F(X_{i,k}) & \text{if } B_j = 1, \\ 1 - F(X_{i,k}) & \text{if } B_j = 0 \end{cases} \quad \text{-- } \textcircled{5}$$

where $F(x) = (1 + \exp(-\mathbf{M}x))^{-1}$ and \mathbf{M} be the *median* (X_k) and B_j were *i.i.d* symmetric 0-1 bernoulli random variables. This process was repeated

until the variable X_j hit the predefined missing rate(30%, 50%, 70%).

III. Results

1. Descriptive statistics of data

6 real, survey data sets were used in total. The Table 6 shows the total number of observations and variables in the dataset used in simulation study. The feature types of data used in simulation were categorized into binary(1,0), multi-class and continuous types.

2. Imputation Accuracy Comparison between BPSOmf and missForest

The error rate of imputation result are compared to assess the performance of BPSOmf and MissForest. Assume that $X = (X_1, X_2, \dots, X_p)$ to be a $n \times p$ -dimensional data matrix. Then, X_{true} is the complete data matrix and X_{impu} is the imputed data matrix. For continuous variables, the error rate is calculated by the NRMSE. For categorical variable, PFC are calculated to quantify the error rate. NRMSE and PFC can be defined as the equations below.

$$NRMSE = \frac{\sqrt{mean(X_{true} - X_{impu})^2}}{var(X_{true})} \quad \text{-----} \quad \textcircled{6}$$

$$PFC = \frac{\sum_{i=1}^n (I_{X_{true} \neq X_{impu}})}{\# \text{ of missing values}} \quad \text{-----} \quad \textcircled{7}$$

First of all, the overall of error rate was compared between BPSOmf and missForest. In this study, the error rate of imputation result was compared

in terms of the PFC for the categorical variables and the NRMSE for the continuous variables. Overall, BPSOmf tends to show the lower NRMSE(0.547) and higher PFC(0.189) compared to those of missForest. In other words, BPSOmf shows the better performance with respect to the continuous data.

In addition, the error rate comparison was done with 5 factors being nested such as Missing Rates(30%, 50%, 70%), the Missing Mechanisms(MCAR, MAR), 6 datasets and the number of trees to grow in each forest when missForest imputation is being done. Those results are presented in the Figure 4, 5, 6 and Table 6, 7, 8, 9, 10 respectively.

To sum up, for the Missing Rate(30%, 50%, 70%) and the Missing Mechanism(MCAR, MAR) factors, BPSOmf shows the relatively lower NRMSE and higher PFC for all levels. Comparison study using the 6 datasets have the similar tendency.

what noticeable is the comparison of PFC between BPSOmf and missForest with different levels of *ntree*. The *ntree* is the number of trees to grow in each forest. As the number of trees increases, the PFC values for both BPSOmf and missForest decrease. However, for missForest, when the number of tree is not big enough(10, 30, 50), the PFC value is higher than those of BPSO in the same level of *ntree*. Moreover, PFC values(0.196) of BPSOmf with *ntree* = 50 is lower than the PFC values(0.197) of missForest with *ntree* = 70. Generally, it is known that the number of trees to grow in random forest should be at least 100. If *ntree* value is smaller than 100, it is likely to have the underfitted results. Thus, PFC in missForest with small *ntrees* tends to have a higher error rate than those of BPSOmf with same setting. This might implies that feature selection based on BPSO helps to missForest to perform better with not enough number of *ntrees*.

IV. Discussion

In this study, we proposed to BPSOmf to improve the performance of missForest algorithm by pruning unnecessary variables.

Overall, the BPSOmf algorithm shows the good performance especially for

the continuous variable while the missForest is better to deal with categorical variables. Although BPSOmf shows the better performance for continuous variables and it would be better to use BPSOmf when the imputation target data are mainly consist of continuous variables. However, there are several limits about this study.

First one is that the number of variables used in simulation setting is too small. For MCAR setting, 50 variables were used per each data and for MAR, only 25 variables were used. The result would be more reliable if using more data. Besides that, the actual dimension of data used in the simulation study is not really large enough to evaluate the performance of BPSOmf and missForest. The dimension of datasets is doable enough to perform missForest imputation so there was no sign of the curse of dimensionality.

Lastly, for *population initialization* in the feature selection with BPSO, filtering the candidate p' variables is performed based on the AIC value derived by a simple regression. Possibly, that could be the main reason why BPSOmf showed better Imputation accuracy with respect to the continuous variables.

Reference

- 1 Rubin, D. B. (1976). "Inference and missing data." *Biometrika* 63(3): 581–592.
- 2 Rubin, D. B. (1996). "Multiple imputation after 18+ years." *Journal of the American statistical Association* 91(434): 473–489.
- 3 Stekhoven, D. J. and P. Bühlmann (2012). "MissForest—non-parametric missing value imputation for mixed-type data." *Bioinformatics* 28(1): 112–118.
- 4 Tang, F. and H. Ishwaran (2017). "Random Forest Missing Data Algorithms." *Stat Anal Data Min* 10(6): 363–377.
- 5 Buuren, S. v. and K. Groothuis-Oudshoorn (2010). "mice: Multivariate imputation by chained equations in R." *Journal of statistical software*: 1–68.
- 6 Hong, S. and H. S. Lynn (2020). "Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction." *BMC Med Res Methodol* 20(1): 199.
- 7 Van Buuren, S. (2018). *Flexible imputation of missing data*, CRC press.
- 8 Shah, A. D., et al. (2014). "Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study." *American journal of epidemiology* 179(6): 764–774.

- 9 Carpenter, J. and M. Kenward (2012). Multiple imputation and its application, John Wiley & Sons.
- 10 Sterne, J. A., et al. (2009). "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls." *Bmj* 338.
- 11 Little, R. J. (1988). "A test of missing completely at random for multivariate data with missing values." *Journal of the American statistical Association* 83(404): 1198–1202.
- 12 Donders, A. R. T., et al. (2006). "A gentle introduction to imputation of missing values." *Journal of clinical epidemiology* 59(10): 1087–1091.
- 13 Shah, A. D., et al. (2014). "Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study." *American journal of epidemiology* 179(6): 764–774.
- 14 Kowarik, A. and M. Templ (2016). "Imputation with the R Package VIM." *Journal of statistical software* 74(7): 1–16.
- 15 Kim, Y., et al. (2016). "Cohort Profile: The Korean Genome and Epidemiology Study (KoGES) Consortium." *International Journal of Epidemiology* 46(2): e20–e20.
- 16 Kweon, S., et al. (2014). "Data resource profile: the Korea national health and nutrition examination survey (KNHANES)." *International Journal of Epidemiology* 43(1): 69–77.

- 17 Kennedy, J. and R. Eberhart (1995). Particle swarm optimization. Proceedings of ICNN'95–International Conference on Neural Networks, IEEE.
- 18 Xiong, L., et al. (2019). "Multi–feature fusion and selection method for an improved particle swarm optimization." Journal of Ambient Intelligence and Humanized Computing.
- 19 Chuang, L.–Y., et al. (2008). "Improved binary PSO for feature selection using gene expression data." Computational Biology and Chemistry 32(1): 29–38.
- 20 Reeves, C. and J. E. Rowe (2002). Genetic algorithms: principles and perspectives: a guide to GA theory, Springer Science & Business Media.
- 21 Chiesa, M., et al. (2020). "GARS: Genetic Algorithm for the identification of a Robust Subset of features in high–dimensional datasets." BMC bioinformatics 21(1): 54.

List of Figures

Figure 1 Illustration of Particle Swarm Optimization

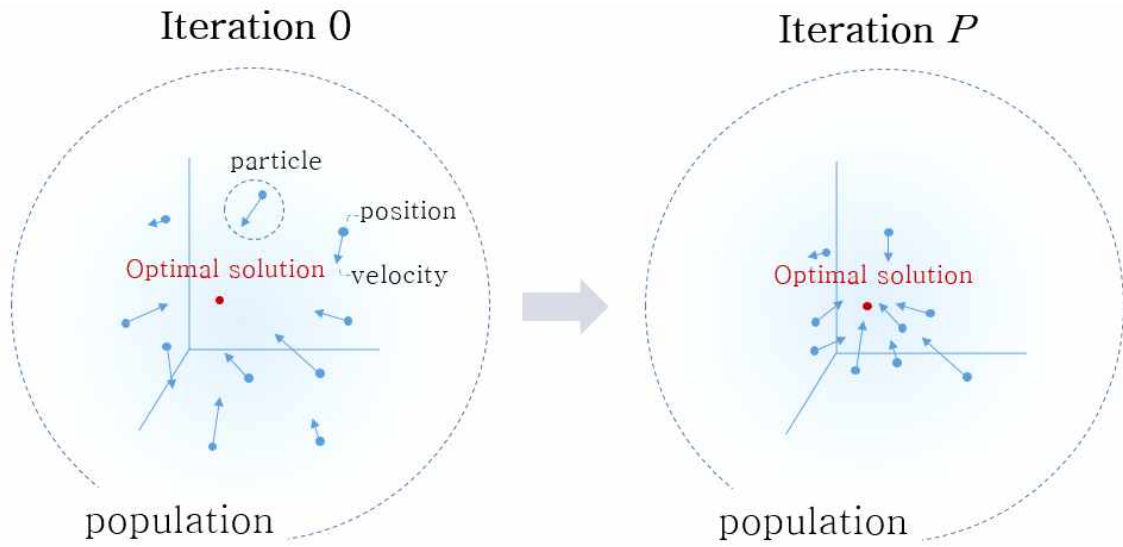
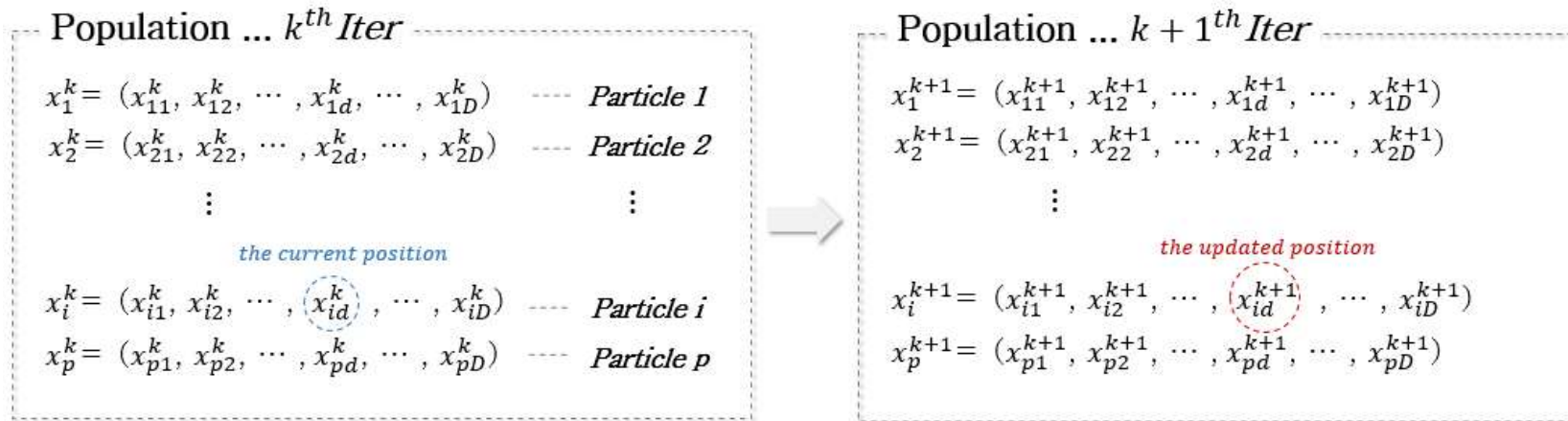
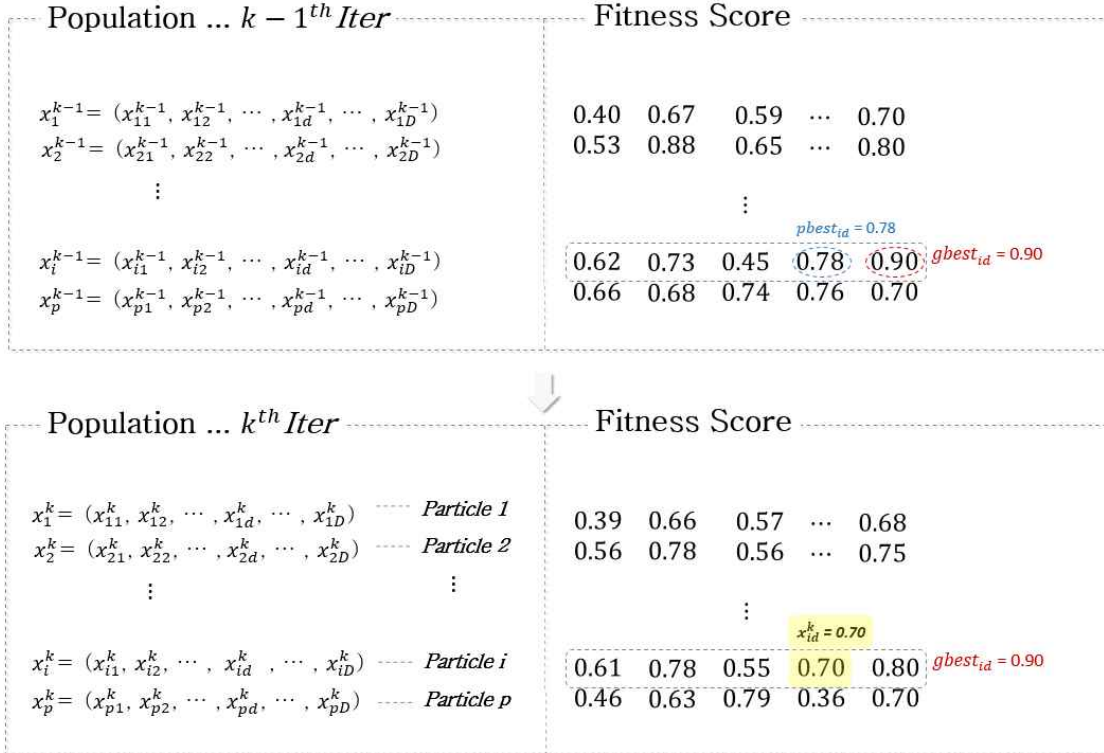


Figure 2 Basic terms and scheme of PSO



- ▶ x_{id} : position
- ▶ v_{id} : velocity
- ▶ w : inertia weight
- ▶ $pbest_{id}$: the best previous position of x_{id}^k
- ▶ $gbest_{id}$: the best position obtained by population so far
- ▶ c_1, c_2 : acceleration constants
- ▶ r_1, r_2 : random value in $Uni(0,1)$

Figure 3 Update Process in PSO



- ▶ Update $x_{id}^{k+1} = x_{id}^k + v_{id}^{k+1}$
- ▶ $v_{id}^{k+1} = wv_{id}^k + c_1r_1(pb_{id} - x_{id}^k) + c_2r_2(g_{id} - x_{id}^k)$
- ▶ $v_{id}^{k+1} = wv_{id}^k + c_1r_1(pb_{id} - x_{id}^k) + c_2r_2(g_{id} - x_{id}^k)$
 $= wv_{1D}^1 + c_1r_1(0.78 - 0.7) + c_2r_2(0.9 - 0.7)$

Figure 4 Error Rate by data

BPSOmF vs. missForest by data

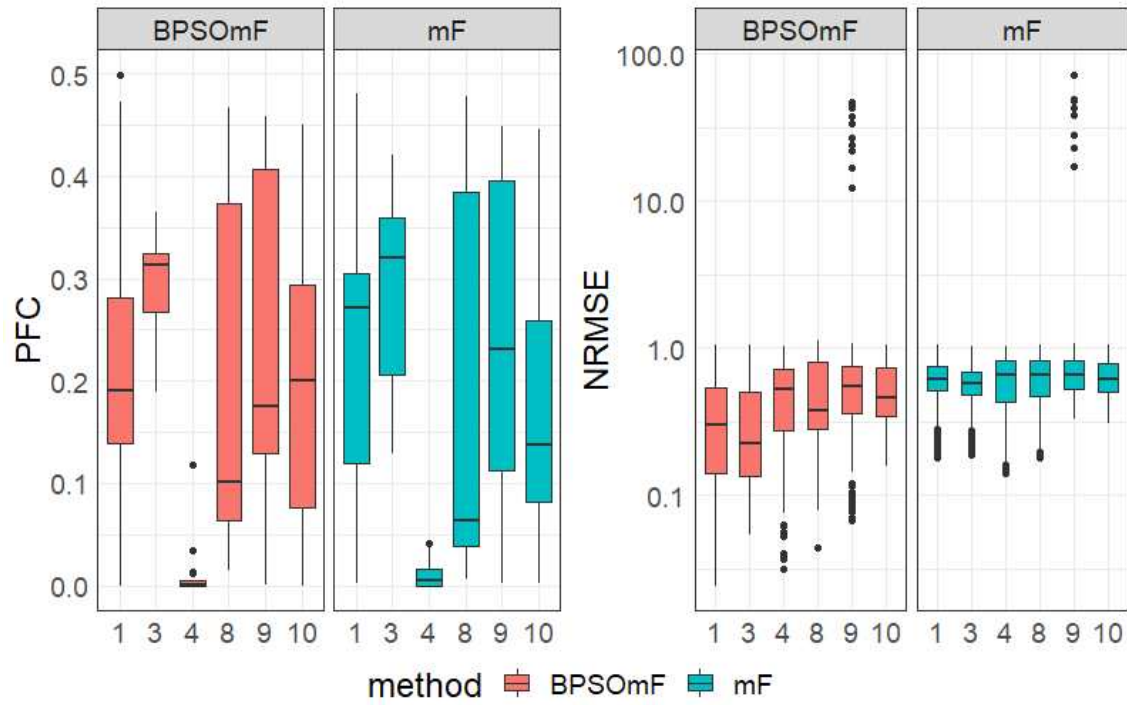


Figure 5 Error Rate by Missing Rate

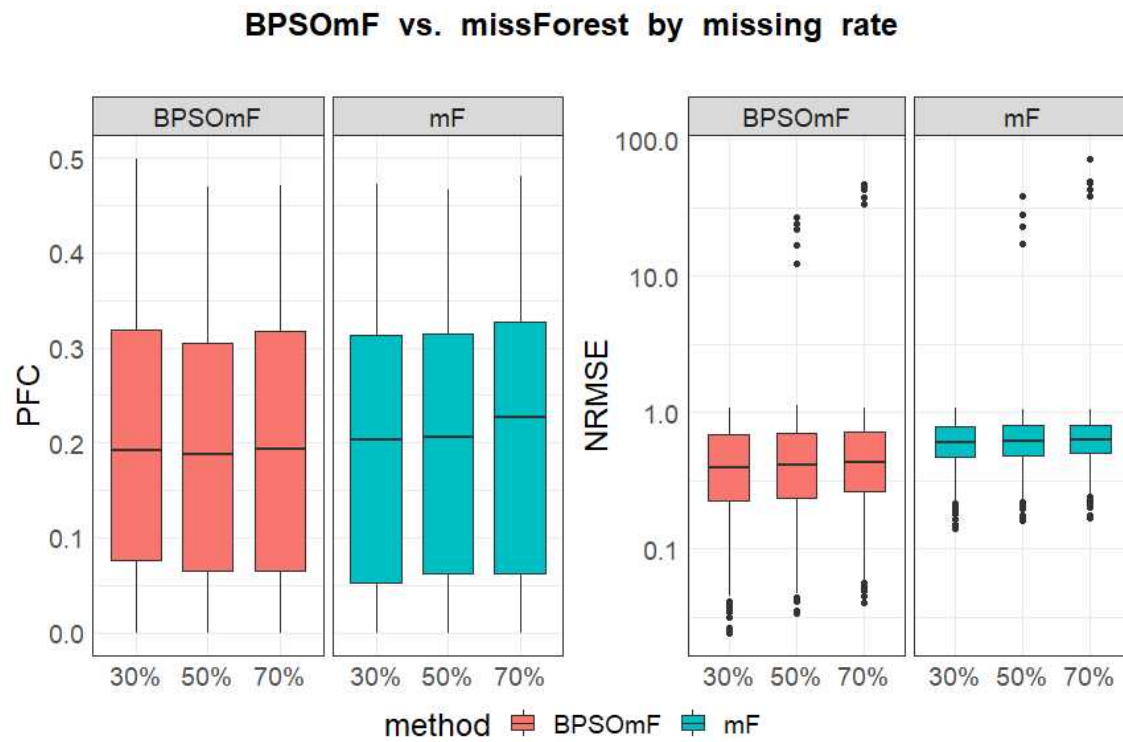
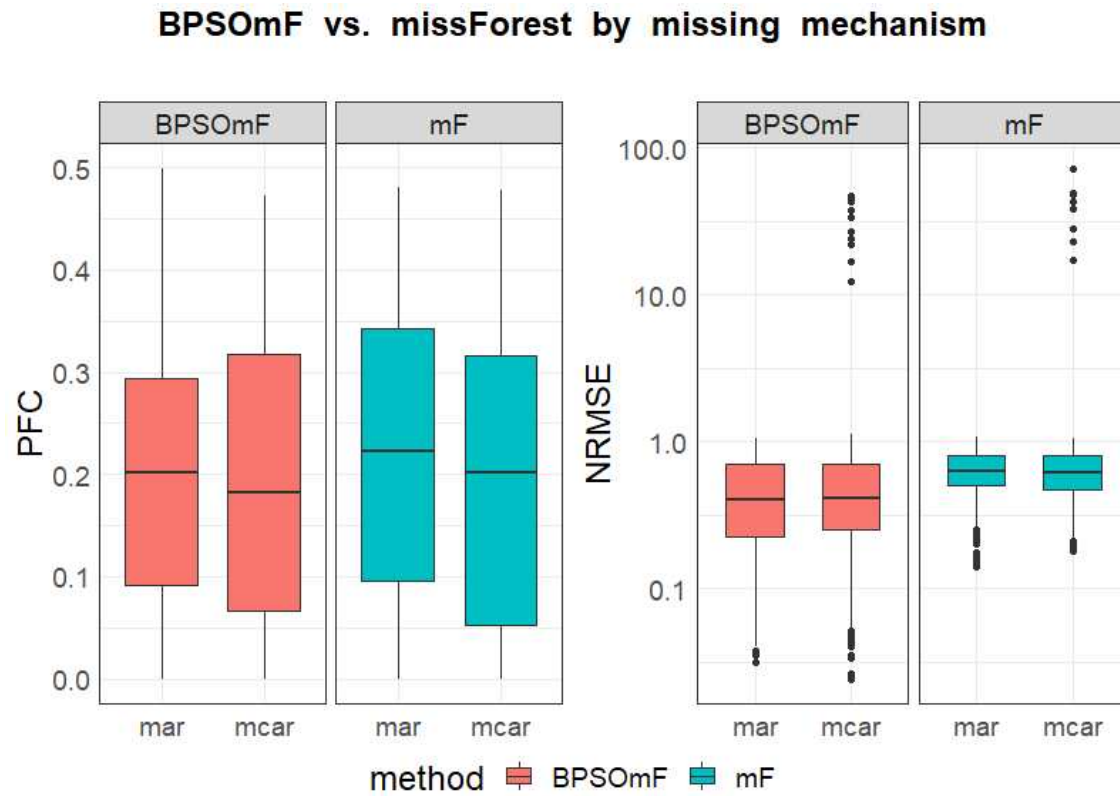


Figure 6 Error Rate by Missing Mechanism



List of Tables

Table 1. Psuedo code: Feature selection based on BPSO

The BPSO Algorithm for Feature Selection	
1	Begin
2	Population Initialization
3	While <i>maximum Iteration</i> do
4	evaluate fitness of each <i>particle</i>
5	for $i = 1$ to p' do
6	update the <i>pbest</i> of <i>particle i</i>
7	update the <i>gbest</i> of <i>particle i</i>
8	End
9	for $i = 1$ to p' do
10	for $d = 1$ to k do
11	update the <i>velocity</i> of <i>particle i</i>
12	update the <i>position</i> of <i>particle i</i>
13	End
14	End
15	Calculate the performance of the selected feature subset
16	Return the selected feature subset.
17	Return the performance value.
18	End

Table 2. The Overall missing rate of 6 dataset

Data	KoGES Baseline	KoGES 2nd	KoGES 3rd	KoGES 7th	KNHANES 2016, 7th
Missing Rate	0.230	0.421	0.421	0.516	0.554

Table 3. Experimental design for BPSOmf vs. missForest

	Method	Missing Mechanism	Missing rate
Experiment 1	BPSOmf	MCAR	30%
Experiment 2	BPSOmf	MCAR	50%
Experiment 3	BPSOmf	MCAR	70%
Experiment 4	BPSOmf	MAR	30%
Experiment 5	BPSOmf	MAR	50%
Experiment 6	BPSOmf	MAR	70%
Experiment 7	missForest	MCAR	30%
Experiment 8	missForest	MCAR	50%
Experiment 9	missForest	MCAR	70%
Experiment 10	missForest	MAR	30%
Experiment 11	missForest	MAR	50%
Experiment 12	missForest	MAR	70%

Table 4. Exclusion Criteria when inducing missing values

1	Criteria	The variables having <i>only 1 level</i> of observations were excluded
	Purpose	it is pointless to generating missing values within the variable having only 1 level of observation.
2	Criteria	The categorical variables with a <i>class imbalance</i> were excluded.
	Purpose	those variables any of whose class frequency is less than 20% of the number of total observation.
3	Criteria	The variables with missing rate <i>greater than 75%</i> were excluded.
	Purpose	if missing induced within the variables with higher missing values, such problems mentioned above would possibly occur again.

Table 5. The Number of Variables and Observations of the Datasets used in Simulation Study

Data	The number of variables				The number of Observation
	Binary	Multiclass	Continuous	Total	
KoGES Baseline	471	528	369	1368	10030
KoGES 2nd	521	594	478	1593	7515
KoGES 3rd	423	230	322	975	6688
KoGES 7th	416	127	304	847	6318
KNHANES 2016, 7th	289	106	257	652	8150
KNHANES 2017, 7th	320	96	250	666	8127

Table 6 Overall Error rate

	NRMSE		PFC	
Error Rate	BPSOmf	missForest	BPSOmf	missForest
	0.547 (1.542)	0.688(1.619)	0.189(0.140)	0.171 (0.134)

Table 7. Error Rate by Data

Data	NRMSE		PFC	
	BPSOmf	missForest	BPSOmf	missForest
KoGES Baseline	0.502(0.247)	0.63(0.203)	0.184(0.127)	0.161 (0.157)
KoGES 2nd	0.337 (0.264)	0.58(0.185)	0.287(0.038)	0.261 (0.092)
KoGES 3rd	0.526(0.273)	0.616(0.231)	0.005 (0.01)	0.007(0.008)
KoGES 7th	0.497(0.302)	0.638(0.217)	0.212(0.164)	0.186 (0.175)
KNHANES 2016, 7th	0.916(3.723)	1.044(3.937)	0.23 (0.159)	0.237(0.143)
KNHANES 2017, 7th	0.52 (0.235)	0.63(0.187)	0.18(0.131)	0.161 (0.116)

Table 8. Error Rate by Missing Rate

Missing Rate	NRMSE		PFC	
	BPSOmf	missForest	BPSOmf	missForest
30%	0.474(0.281)	0.611(0.205)	0.21(0.135)	0.169(0.136)
50%	0.552(1.245)	0.694(1.42)	0.176(0.142)	0.168(0.133)
70%	0.615(2.351)	0.76(2.411)	0.18(0.144)	0.175(0.135)

Table 9. Error rate by Missing Mechanism

Missing Mechanism	NRMSE		PFC	
	BPSOmf	missForest	BPSOmf	missForest
MCAR	0.584(1.879)	0.718(1.978)	0.187(0.142)	0.166(0.136)
MAR	0.473(0.284)	0.628(0.198)	0.193(0.138)	0.181(0.131)

Table 10. Error rate by *ntree*

The number of Tree	NRMSE		PFC	
	BPSOmf	missForest	BPSOmf	missForest
10	0.538(1.322)	0.74(2.159)	0.206(0.148)	0.236(0.147)
30	0.521(1.484)	0.698(1.578)	0.2(0.141)	0.207(0.142)
50	0.517(1.283)	0.682(1.426)	0.196(0.143)	0.202(0.141)
70	0.511(1.285)	0.689(1.644)	0.198(0.141)	0.197(0.14)
100	0.547(1.542)	0.688(1.619)	0.189(0.14)	0.171(0.134)

초 록

BPSO 기반 변수 선택 기법으로 보정한 결측치 대체 알고리즘 개발

정수린

서울대학교 보건대학원
보건학과 보건통계학 전공

배경

데이터 내의 결측은 발생 원인에 따라 MCAR, MAR, MNAR로 나뉘며 이에 따라 결측 대체 방법도 달라진다. 많은 결측 대체 방법 중, missForest는 데이터에 대한 분포 가정을 필요로 하지 않으며 mixed-type 데이터에도 사용이 가능하기 때문에 다른 방법에 비해 큰 이점을 갖는다. 하지만 최근의 연구에 따르면 missForest를 이용한 결측 대체 결과에 편향이 발생할 수 있다는 것이 밝혀졌다. 또한 우수한 성능을 가지지만 데이터 차원이 커짐에 따라 계산량이 크게 증가한다는 단점 또한 존재한다. 이에 따라 본 연구에서는 BPSO를 기반으로 한 변수선택법으로 missForest를 보완하고자 한다.

방법

BPSO란 진화 연산(evolutionary algorithm) 중 하나로, 전역적인 최적화(global optimization) 기법과 효율적인 계산으로 잘 알려진 방법이다. 본 연구에서는 missForest로 결측치를 대체하기에 앞서, BPSO를 기반으로 한 변수 선택을 진행하는 방법을 통해 기존 missForest 방법보다 결측치 대체 정확도를 개선시키는 것을 목표로 한다.

결과

missForest는 mixed-type data에 사용가능하며, 특별한 분포가정이 필요하지 않고, 성능 또한 우수하기 때문에 널리 사용되는 결측 대체 방법 중 하나이다.

하지만 관측치 개수나 변수 개수가 증가함에 따라 계산량이 크게 증가하기 때문에 이를 보완하고자, BPSO를 기반으로 한 변수선택법으로 결측 대체에 사용된 변수들을 미리 선택한 후, missForest를 적용하였다. 본 연구에서는 missForest로 결측치를 대체하기에 앞서, BPSO를 기반으로 한 변수 선택을 진행함으로써 연속형 변수에 한하여 기존 missForest 방법보다 개선된 결과를 얻었다.

주요어: 변수 선택, BPSO, missForest, 결측

학번: 2019-22081