A Comparison of the Effects of Data Imputation Methods on Model Performance

Wooyoung Kim*[1], Wonwoong Cho*[1], Jangho Choi[2], Jiyong Kim[2], Cheonbok Park[1], Jaegul Choo[1]

1 Department of Computer Science, Korea University, Seoul, Republic of Korea

2 Electronic Telecommunications Research Institute (ETRI), Daejeon, Republic of Korea

{ddsksw01, tyflehd21, cb_park, jchoo}@korea.ac.kr, {janghochoi, kjy}@etri.re.kr

***Abstract*— Missing values cause critical problems on training a prediction model. Various missing data imputation methods have been introduced to settle down the problem. However, the imputation accuracy obtained by the methods is insufficient to validate performance of prediction models. Thus, in this study, we compare (1) imputation accuracy from various imputation methods as well as (2) the effects of imputation methods on prediction accuracy, investigating a relationship between imputation accuracy and prediction accuracy. For the comparison, we use water quality data composed of the latest actual observational multi-sensor data from Daecheong Lake. We conduct several experiments to compare seven imputation methods including a state of the art method, and their effects on three distinct prediction models. Through quantitative comparison and analysis, we proved that it is necessary to consider both imputation accuracy and model prediction accuracy when choosing an imputation method.**

***Keywords*— missing values, missing data, incomplete data, imputation methods, linear interpolation, mean imputation, knn imputation, svd imputation, randomforest imputation, mice imputation, amelia imputation, model performance**

I. INTRODUCTION

Missing values denote empty and unobserved values in an arbitrary dataset. They have been a cumbersome problem in diverse research areas, such as Biology [1] and Medicine [2] because it causes information loss followed by a model performance degradation. The simplest way to deal with missing values is to discard every instance of data that contain at least one missing value. However, it reduces the amount of data used in training a model and causes a time discontinuity in time series data.

To alleviate the problem, previous studies [3]-[8] have introduced statistical or mathematical methods of 'missing values imputation' [9]-[11], whose aim is to substitute each missing value while preserving characteristics of the data. Because there are various methods of missing values imputation, several studies [12]-[14] have reported a quantitative comparison between diverse imputation methods, providing a guideline to choose a proper method for a given task.

The high imputation accuracy, however, does not guarantee the high prediction model accuracy. To be concrete, due to the fact that there is no ground-truth for missing values in real data, most of the methods discard every instance containing missing values in the first place. Second, arbitrary values from the remaining data are removed. Third, the imputed data are used for evaluation on imputation accuracy. However, the generated data may contain different patterns from the original ones, so that the imputation accuracy may not be proportional to model prediction accuracy.

In this study, we quantitatively compare (1) imputation accuracy from diverse imputation methods as well as (2) their effects on the prediction accuracy. We demonstrate an importance of a prediction accuracy when choosing an imputation method. Specifically, our work covers major statistical imputation methods, namely linear interpolation imputation [3], mean imputation, k-nearest neighbor (KNN) imputation [4], singular value decomposition (SVD) imputation [5], random forest (RandomForest) imputation [6] and multiple imputation by chained equations (MICE) imputation [7], including a state-of-the-art imputation method, Amelia imputation [8]. To evaluate their effects on prediction models, we implemented three regression models for comparison: Linear regression, RandomForest regression and artificial neural net (ANN) regression.

II. METHODS

In this section, we describe previously developed and studied imputation methods and regression models that are used in this work.

A. Imputation Methods

There are various ways to deal with missing values in incomplete dataset. The simplest approach to solve this problem is to discard the observations that contain at least one missing value. However, this approach can cause a significant amount of information loss, which degrades the performance of the model. Hence, in general, when there are missing values in dataset, we apply imputation methods. Imputation is to

* These authors contributed equally.

replace missing values with plausible values. Through this work, we can convert an incomplete dataset into a completed dataset that does not contain missing values. The followings are prior studies related to imputation methods.

***1) Linear interpolation imputation method*:** Linear interpolation is a mathematical method that uses linear polynomials to estimate unknown values within the range of known values. Missing values are estimated and imputed by the last known value before the missing value and the first known value after the missing value [3]. For example, if the value of January 1st is known as 1 and the value of January 3rd is known as 7, the missing value of January 2nd can be imputed as 4 through this method.

***2) Mean imputation method*:** This method simply calculates a mean value of each variable ignoring missing values. Missing values are replaced with the mean value of the corresponding variable. This method is comparatively easy to use but it does not perform well in almost all cases. For example, if the value of January 1st is known as 1, the value of January 3rd is known as 7, the value of January 4th is known as 10, then the missing value of January 2nd can be imputed as 6 through this method.

***3) KNN imputation method*:** As its name suggests, this algorithm assumes that a specific missing value can be predicted from the values of other similar observations, chosen based on other variables that do not include the variable occurred by the missing value [4]. The number of neighbors (K) is hyperparameter. Setting a low K leads to high variance and low bias and a high K leads to low variance and high bias. Therefore, hyperparameter K should be chosen carefully.

***4) SVD imputation method*:** Before applying the SVD algorithm, this method fills missing values using mean imputation method. Then, the method implements a low rank-K approximation of the completed dataset. Missing values are replaced with the estimated values. We repeat SVD imputation a certain number of times with the imputed value [5]. Low rank-K less than full-rank is a hyperparameter. Because low rank-K which is smaller than full rank is used for approximation and imputation, this approach is called as truncated SVD.

***5) RandomForest imputation method*:** This approach is based on tree-based ensemble model called RandomForest [6]. This method starts by replacing missing values with variable medians. Then, RandomForest algorithm is used to calculate proximity matrix with the completed dataset. Proximity is the closeness between pairs of observations. For example, if two observations are allocated to a same terminal node through one tree, their proximity is increased by one. After all trees are constructed, the proximity matrix is made and used for imputation. For continuous variables such as sensor data, missing value is imputed by a weighted average of non-missing observations.

***6) MICE imputation method*:** This method dose not implement imputation only once but multiple times [7]. Through this process, it creates multiple completed datasets and multiple predictions for each missing value using an imputation process with a random component. Then, it analyzes each completed dataset and combines the results. Because missing value is estimated and imputed by multiple predictions, this method can solve the problem of less standard errors of statistics based on imputed values than actual values.

7) Amelia imputation method: Although Amelia method is based on multiple imputation approaches, it is different from traditional multiple imputation methods. Amelia assumes that the data follow a multivariate normal distribution, so that it uses a joint modeling approach whereas MICE imputes data on variable by variable basis [8]. Amelia can reflect time trends in the process of imputation by incorporating polynomials of time into the model which makes this method powerful in the time series data. After Amelia version 2, Bootstrap-based EM algorithm is used to impute missing values in this method.

B. Regression Models

Regression models are widely used in prediction and forecasting. These models focus on finding relationship between a dependent variable and one or more independent variables. Unlike a classification model which predicts categorical labels, a regression model is about predicting a quantity. Therefore, if dependent variable is continuous quantity variable, regression models are employed. The performance of regression models can be evaluated by root mean square error (RMSE). We describe three popular models in the following.

***1) Linear regression model*:** This model is simple and widely applied approach in prediction task. Linear regression is a linear approach to find the relationship between a continuous quantity dependent variable and one or more independent variables. If only one independent variable is used for model building, it is called as simple linear regression, and if more than one independent variables are used in the model, it is called as multiple linear regression.

***2) Randomforest regression model*:** RandomForest is powerful ensemble model in both classification and regression problems. This model builds multiple decision trees and merges them. This model gives randomness to variable selection as well as observation selection. Through this process, the model can improve accuracy and stability. The model makes integrated prediction by calculating the average of the predictions across the trees.

***3) ANN regression model*:** This model is based on deep learning. Because of the fact that this model can approximate non-linear relationships between a dependent variable and independent variables, it can solve more complex problems that cannot be solved by the traditional regression models.

III. EXPERIMENTS AND RESULTS

In fact, water quality data of Daecheong Lake also contain a lot of missing values that interfere with the prediction task. Even though, our goal is to predict the amounts of blue-green algae by chlorophyll-a variable to prevent negative effects of large amounts of blue-green algae, but we cannot use the data directly because of missing values. To select a proper imputation technique for our task, we tried to evaluate not only the performance of the missing values imputation method but also its effects on the prediction performance of the model.

Many studies of imputation methods focus on evaluating the performance of imputation methods themselves. For performance evaluation, they remove some data points and predict the missing values with several imputation methods. Then, they measure the difference between the actual values and imputed values.

However, this approach may not be appropriate in some cases. As verified in the following experiments, the performance of imputation method is not proportional to the performance of the prediction model with the imputation method. When selecting an imputation method, we should consider the purpose of data analysis then carefully choose one. In other words, imputation is just a means and a process, not a goal to achieve. Our goal is to fit our target model which can be used for a statistical testing task or classification task or regression task in a better way. Therefore, the criteria of selecting imputation method should be set to increase the performance of our target model, not to improve the performance of the imputation method itself. Furthermore, the influence of the imputation method may vary depending on the task and data. Hence, it is dangerous to choose a certain imputation method simply because it has been reported to have excellent performance in previous studies including our study, without a reliable experiment.

Dataset we used is about water quality complex sensor observations. This data is observed at Daecheong Lake in Daecheonghosu-ro Daedeok-gu Daejeon, in republic of Korea. The specific detail of the dataset can be found in Table 1.

TABLE 1. DATASET SPECIFICATION

Number of instances	Number of variables	Type of variables	Data collection
1705	9	Continuous Variables	Data collected on a daily basis from 2012 to2018

A. Experiments with Non-missing Dataset

The overall process of experiments is shown in Figure 1. Similar to the previous studies, these experiments were implemented by generating arbitrary missing values in the complete dataset, then applying several imputation methods to the incomplete dataset. Furthermore, based on the completed dataset made by the imputation method, we conducted several experiments to evaluate the performance of the model that predicts chlorophyll-a.

***1) Data pre-processing*:** Of the nine water quality variables, the total phosphorus variable was removed because it contained too many missing values as a matter of the sensor. In these experiments, as all instances including at least one missing value must be removed, we used this preprocessing step to prevent instances of the complete dataset from getting too small. After this process, all instances that contained at least one missing value were removed. Through these processes, we could construct a complete dataset. In addition, standardization was conducted to each variable to make the influence of each variable on the evaluation criteria (RMSE) the same.

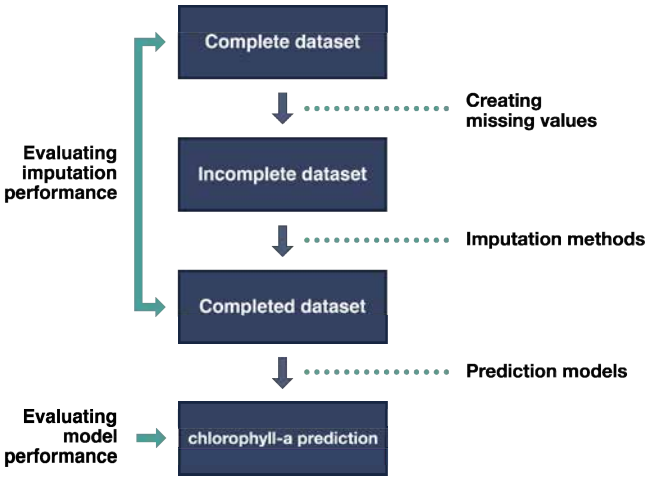

Figure 1. Experimental process

***2) Experimental design*:** This experiment was based on the complete dataset without missing values because all missing values were removed through the previous data pre-processing steps. We randomly assigned missing values according to various probability range and filled the missing values by applying several imputation methods. Then we measured the difference between the filled values and actual values. We also conducted another experiment to measure the performance of prediction model. Since we have the actual observations in hand, we could build the baseline models that predict chlorophyll-a variable using actual data. Based on these baseline models, we proceeded experiments to measure the performance of the prediction models using the completed datasets made by several imputation methods. The hyperparameters of imputation methods and the prediction models we used in these experiments are shown in Table 2 and Table 3.

TABLE 2. HYPERPARAMETERS OF IMPUTATION METHODS

Imputation Method	Hyperparameter(s)
Linear Interpolation	Nothing
Mean Imputation	Nothing

KNN Imputation	Number of neighbors = 5
SVD Imputation	Low rank = 5
RandomForest Imputation	Number of trees = 100
MICE Imputation	Multiple times = 10
Amelia Imputation	Multiple times = 10, Polytime = 3

TABLE 3. HYPERPARAMETERS OF PREDICTION MODELS

Regression Model	Hyperparameter(s)
Linear Regression	Nothing
RandomForest Regression	Number of trees = 500
ANN(1) Regression	Hidden_dim=30, Hidden layers_num = 2, Activation_fun = RELU, Epoch=50, Learing_rate=0.05, Dropout_rate = 0.2, Batch_size = 30
ANN(2) Regression	Hidden_dim=60, Hidden layers_num = 2, Activation_fun = RELU, Epoch=80, Learing_rate=0.02, Dropout_rate = 0.2, Batch_size = 30

***3) Evaluation criteria*:** The evaluations were made by averaging 10 simulation results. For the model performance evaluation, each simulation was conducted with 4-fold cross validation [15]. We used RMSE, which is the most commonly used measure of performance in both imputation and prediction model in the similar studies [16]. It measures the difference between actual (true) values and predicted (imputed) values (See Eq. (1)). The lower the RMSE, the better the performance.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i}^{n}(Predicted_i - Actual_i)^2} \qquad (1)$$

***4) Experiment results*:** The detailed results of two types of experiments are shown in Table 4 and Table 5. Overall, imputation performance decreases as the percentage of missing values increases. Linear interpolation method shows the best performance regardless of the percentage of missing values. Except for this imputation method, the RandomForest imputation and Amelia imputation showed the excellent performance. Based on these results, we set a specific percentage of missing values and conducted an experiment to evaluate the model performance using the completed datasets, which are generated by the top three imputation methods. As shown in Table 5, linear interpolation method, which provided the best performance in the imputation error, showed worse performance in prediction accuracy, compared to those of Randomforest imputation and Amelia imputation. In this way, we found that the performance of imputation itself is not directly related to the performance of prediction model. We interpreted the linear interpolation imputation as having the unusually high performance in the case of imputation performance because of following reasons. As the daily variation of sensor data is very small, linear interpolation method is the most conservative and safest approach to predict the missing value by using information only from the immediately preceded observation and the immediately followed observation on the corresponding variable. For example, in a situation where daily fluctuations are not large, filling the missing January 2nd data with the average value of January 1st and January 3rd data could be a safer strategy to lower RMSE than filling the missing value using more observations and inter-variables information.

TABLE 4. IMPUTATION PERFORMANCE

Imputation Methods	RMSE according to Specific percentages of missing values			
	10%	20%	30%	70%
Linear Interpolation	**0.11434**	**0.13605**	**0.13984**	**0.21281**
Mean Imputation	0.76450	0.79850	0.77135	0.80628
KNN Imputation	0.31595	0.41208	0.52240	0.79405
SVD Imputation	0.64674	0.70929	0.65852	0.78589
RandomForest Imputation	**0.21287**	**0.24206**	**0.25319**	**0.62360**
MICE Imputation	0.52086	0.54354	0.6221	0.80734
Amelia Imputation	**0.14657**	**0.16876**	**0.19840**	**0.35514**

TABLE 5. MODEL PERFORMANCE

Baseline + Imputation Methods	RMSE Percentage of missing values = 20%			
	Linear Regression	Random Forest Regression	ANN(1) Regression	ANN(2) Regression
Baseline (Original Dataset)	0.47904	0.26152	0.24211	0.23871
Linear Interpolation	0.53896	0.26686	0.27485	0.27312
Random Forest Imputation	**0.50472**	**0.27723**	**0.26343**	**0.25964**
Amelia Imputation	**0.49808**	**0.25178**	**0.26984**	**0.27112**

B. Experiment with Missing Dataset

Figure 2 shows the overall process of the experiment. Unlike the previous experiments, this experiment was based on the incomplete dataset that contains the actual missing values, not the arbitrarily created missing values. We could not proceed with the experiment to measure imputation performance because it was impossible to find out the actual values of the missing values in these experimental settings. Therefore, we conducted a model performance evaluation of the regression task to predict the amounts of chlorophyll-a.

***1) Data pre-processing*:** First, all the instances where chlorophyll-a which is a dependent variable that was missing, were removed. In addition, unlike previous experiments, we used the total phosphorus variables containing a lot of missing values to maintain characteristics of the original dataset.

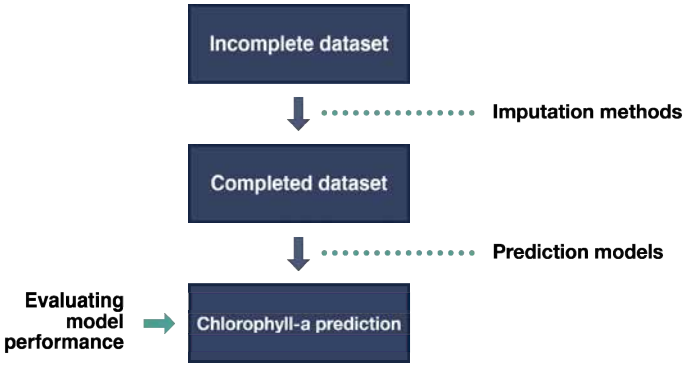

Figure 2. Experimental process

***2) Experimental design*:** This experiment was based on the incomplete dataset with original missing values. We filled the missing values by applying several imputation methods which covered six imputation methods except for the mean imputation method which had shown significantly poor performance. Then, we measured the performance of prediction models which proceeded with chlorophyll-a prediction task.

***3) Evaluation criteria*:** In this experiment, only model performance evaluations were conducted. Evaluation criteria was applied in the same way as the previous experiments.

***4) Experiment results*:** The detailed results of experiment are shown in Table 6. Similar to the previous experiments, three previously well-performed methods (linear Interpolation imputation, RandomForest imputation and Amelia imputation) performed well in this experiment which measured the model performance in the real dataset. Meanwhile, as in previous experiments, RandomForest imputation and Amelia imputation showed better performance than linear interpolation imputation. Once again, we verified that the linear interpolation imputation method, which had been the best method for imputation performance, could not be the best method for model performance.

TABLE 6. MODEL PERFORMANCE

Imputation Methods	RMSE			
	Linear Regression	Random Forest Regression	ANN(1) Regression	ANN(2) Regression
Linear Interpolation	**4.23829**	**0.72830**	**0.33489**	**0.31027**
KNN Imputation	4.25922	0.71846	0.33837	0.32471
SVD Imputation	4.24513	0.73399	0.33359	0.31605
RandomForest Imputation	**4.21133**	**0.71053**	**0.32453**	**0.30663**
MICE Imputation	4.24423	0.72874	0.33198	0.31817
Amelia Imputation	**4.21795**	**0.71774**	**0.33238**	**0.30818**

IV. VISUALIZATION OF RESULTS

In the following figures, we will name linear interpolation imputation as LI, RandomForest imputation as RF, linear regression as Linear, RandomForest regression as Random and baseline as BASE. Figures 3, 4, and 5 are visualizations of the results of Tables 4, 5, and 6, respectively. Figure 6, 7, 8, and 9 are specific graphs for each regression model showed in Figure 5.

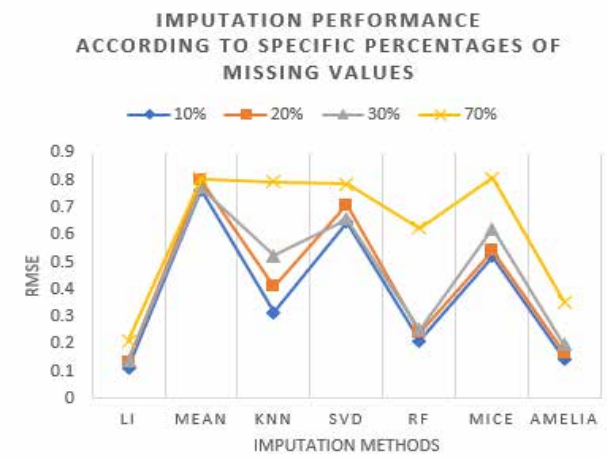

Figure 3. Imputation performance

Figure 4. Model performance

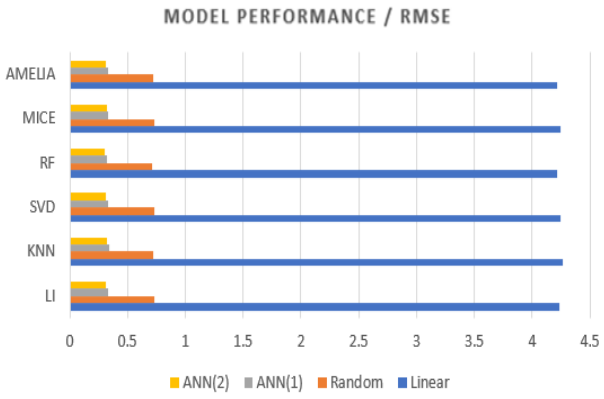

Figure 5. Model performance

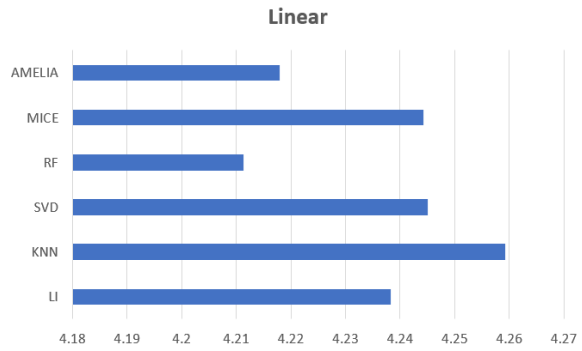

Figure 6. Model performance (Linear)

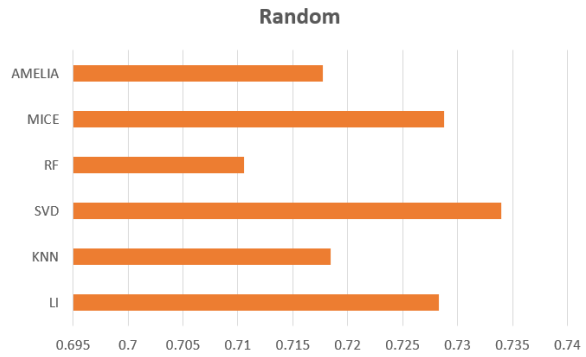

Figure 7. Model performance (Random)

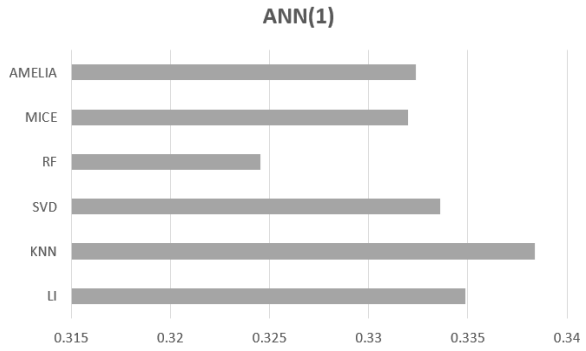

Figure 8. Model performance (ANN(1))

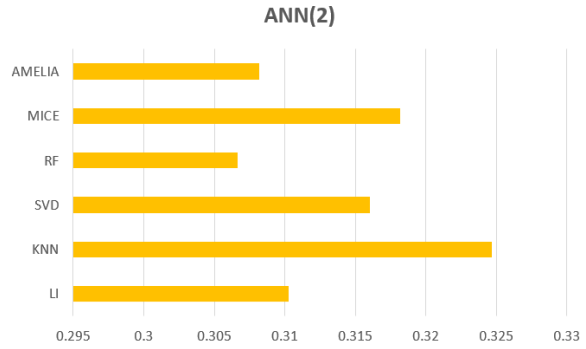

Figure 9. Model performance (ANN(2))

V. CONCLUSIONS

In this study, we conducted two types of experiments. One was to measure imputation performance and the other was to measure model performance. Both experiments had the same purpose to examine the effects of the imputation methods on our task. Through experiments in our paper, we verified that imputation performance itself may not be proportional to the model performance. So, we suggest that imputation methods should be evaluated not only by the imputation performance

but also the model performance, which is the most important thing in prediction task, after applying several imputation methods to the dataset. In the case of our study, our aim was to predict the amounts of the chlorophyll-a accurately. So, we chose the imputation methods which are RandomForest imputation method and Amelia imputation method, not Linear Interpolation imputation method which had shown the best imputation performance.

VI. FUTURE WORK

Further research about a deep learning based imputation approach [17] will be conducted. Furthermore, We will study the enhancement of model performance through data augmentation by using deep generative models such as VAE and GAN in the time series data [18], [19].

ACKNOWLEDGMENT

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.2018-0-00219, Space-time complex artificial intelligence blue-green algae prediction technology based on direct-readable water quality complex sensor and hyperspectral image)

REFERENCES

[1] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, et al. (2001), Missing value estimation methods for dna microarrays, *Bioinformatics* 17: 520-525.

[2] Lewis HD (2012), Missing data in clinical trials, *New England Journal of Medicine* 367: 2557-2558.

[3] Norazian, M. N., Ahmad Shukri, Y., Ramli, N. A., & Mustafa, A. A. (2007), Comparison of linear interpolation method and mean method to replace the missing values in environmental data set, *Proceeding of the International Conference on Sustainable Management (ICOSM)* 9–11 June. ISBN 9789834235826.

[4] Crookston, Nicholas L.; Finley, Andrew O. (2008), yaImpute: An R package for kNN imputation, *Journal of Statistical Software*. 23(10).16.

[5] Kurucz, M., Benczúr, A.A., Torma, B. (2007), Methods for large scale svd with missing values, In: *KDDCup* 2007.

[6] d D.J. Stekhoven, P. Buhlmann (2012), Missforest—non-parametric missing value imputation for mixed-type data, *Bioinformatics*, 28 (1), pp. 112-118.

[7] Buuren, S. van; Groothuis-Oudshoorn, K. (2010), *Journal of statistical software*, volume in press, pp. 1 – 68.

[8] Honaker, James, Anne Joseph, Gary King, Kenneth Scheve, and Naunihal Singh (1999), AMELIA: A Program for Missing Data. Department of Government, Harvard University.

[9] Little RJA, Rubin DB (2002), Statistical Analysis with Missing Data (2ndedn.), Wiley-Interscience.

[10] Rubin DB (1987), Multiple Imputation for Nonresponse in Survey, John Wiley and Sons, Inc.

[11] Ibrahim JG, Chen MH, Lipsitz SR, Herring AH (2005), Missing-data methods for generalized linear models, *Journal of the American Statistical Association* 100: 332-346.

[12] Brock G, Shaffer J, Blakesley R, Lotz M, Tseng G (2008), Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes, *BMC Bioinformatics* 9: 1-12.

[13] Celton M, Malpertuy A, Lelandais G, Brevern A (2010), Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments, *BMC Genomics* 11: 1-16.

[14] Luengo J, Garca S, Herrera F (2012), On the choice of the best imputation methods for missing values considering three groups of classification methods, *Knowledge and Information Systems* 32: 77-108.

[15] Kohavi, R. (1995), A study of cross-validation and bootstrap for accuracy estimation and model selection, *In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* (pp. 1137–1143).

[16] Schmitt P, Mandel J, Guedj M. (2015), A Comparison of Six Methods for Missing Data Imputation, *J Biomet Biostat* 6: 224. doi:10.4172/2155-6180.1000224.

[17] Y. Duan, L. Yisheng, W. Kang, Y. Zhao. (2014), A deep learning based approach for traffic data imputation, *Intelligent Transportation Systems (ITSC) 2014 IEEE 17th International Conference on*, pp. 912-917.

[18] Nazabal, Alfredo, Olmos, Pablo, Ghahramani, Zoubin, Valera, Isabel. (2018), Handling Incomplete Heterogeneous Data using VAEs, arXiv:1807.03653.

[19] Dan Li, Dacheng Chen, Jonathan Goh, See-kiong Ng. (2018), Anomaly Detection with Generative Adversarial Networks for Multivariate Time Series, arXiv:1809.04758.

Wooyoung Kim received the B.S. degree in the Department of statistics from Korea University, Seoul, Korea in 2018. He is currently pursuing his M.S. degree in the Department of computer science from Korea University, Seoul, Korea. His current interests include machine learning, deep learning, data mining and recommender system.

Wonwoong Cho received the B.S. degree in the Departmen of Film, Television and Multimedia from SungKyunKwan University, Seoul, Korea in 2017. He is currently pusuing his M.S. degree in the Department of computer science from Korea University, Seoul, Korea. His research interests include interactive deep learning, computer vision, image translation and statistical model.

Jangho Choi received his B.S. degree in Computer Science & Business Administration in 2010 from the University of Southern California, Los Angeles, USA. He received his MS in Computer Science from the Korea Advanced Institute of Science and Technology, Daejeon, Rep. of Korea, in 2013. He joined SW & Contents Research Laboratory at ETRI as a researcher in 2013. His research interests include the machine learning, deep learning, big data analysis, and artificial intelligence.

Jiyong Kim received the M.S. degree from Seoul National University, Korea, in 1997. He is a principal researcher at Electronics and Telecommunications Research Institute(ETRI). His research topics include big data analysis, machine learning and IoT. Currently he is in charge of the project, “Space-time complex analysis technology for blue-green algae prediction”.

Cheonbok Park received the B.S. degree in the Department of Computer Science and Engineering from Korea University, Seoul, Korea in 2018. He is currently pursuing his M.S. degree in the Department of Computer Science and Engineering from Korea University, Seoul, Korea. His current interests include machine learning, deep learning, interpretable machine learning and natural language processing.

Jaegul Choo is currently an assistant professor in the Dept. of Computer Science and Engineering at Korea University. He received M.S in the School of Electrical and Computer Engineering at Georgia Tech in 2009 and Ph.D in the School of Computational Science and Engineering at Georgia Tech in 2013, advised by Prof. Haesun Park. From 2011 to 2014, he has been a research scientist at Georgia Tech. During the summer in 2009 and 2010, he worked at National Visualization and Analytics Center (NVAC) in Pacific Northwest National Laboratory. He earned his B.S in the Dept. of Electrical and Computer Engineering at Seoul National University.