Efficient Imputation Method for Missing Data Focusing on Local Space Formed by Hyper-Rectangle Descriptors

Do Gyun Kim and Jin Young Choi
Department of Industrial Engineering, Ajou University, Suwon, Korea
{rlaehrbs90, choijy}@ ajou.ac.kr

Keywords: Missing Data, Imputation, Hyper-Rectangle Descriptor, Local Space.

Abstract: In real world data set, there might be missing data due to various reasons. These missing values should be handled since most data analysis methods are assuming that data set is complete. Data deletion method can be simple alternative, but it is not suitable for data set with many missing values and may be lack of representativeness. Furthermore, existing data imputation methods are usually ignoring the importance of local space around missing values which may influence quality of imputed values. Based on these observations, we suggest an imputation method using Hyper-Rectangle Descriptor (HRD) which can focus on local space around missing values. We describe how data imputation can be carried out by using HRD, named HRD_impute, and validate the performance of proposed imputation method with a numerical experiment by comparing to imputation results without HRD. Also, as a future work, we depict some ideas for further development of our work.

1 INTRODUCTION

Some data might be missing during collection due to various reasons such as physical or logical errors. However, since most of data analysis techniques cannot be performed properly with missing data, handling missing data is very important in machine learning area. As an alternative, one can simply exclude data with missing parts and analyze the rest of fully collected data, which is called data deletion method (McKnight et al., 2007). This approach can perform well only if few data points are missing. However, in real world data, there can be many missing data points, and analysis results obtained from using only fully known data cannot represent the whole data set. Therefore, we need a data imputation method that replaces missing data with new values estimated from fully collected data, rather than excluding them. In this case, although imputed values are estimated from observed data, scalability of original data set can be preserved, and data analysis can be applied to it that is a complete data set.

Generally, in data imputation process, local space around missing data is important since behavior of missing data is more likely to follow data pattern in local space rather than whole feature space. However, although there are many researches about missing data imputation, there exist few approaches focusing on local space. Some imputation methods including k-Nearest Neighbors (k-NN) utilizes information of local space. However, they have their own limitations such as parameter selection and ambiguous standard definition of local space.

Based on these observations, we propose an efficient imputation method that can (i) define local space around missing data systematically and (ii) impute missing values by focusing on that local space. Specifically, we suggest HRD_impute as a missing data imputation method using Hyper-Rectangle Descriptor (HRD) that was originally developed to carry out one-class classification (Jeong et al., 2019). The basic idea of HRD is to divide feature space into Hyper-Rectangles (H-RTGLs), formed by geometric rules called intervals, and classify instances in H-RTGLs as target class. Therefore, H-RTGLs as HRD can be considered as a certain local space including some instances and can be used to overcome one of limitations of existing missing data imputation methods.

The rest of this paper is organized as follows. Section 2 describes the literature survey about existing missing data imputation methods. We suggest details of the suggested HRD-based imputation method in Section 3. Then, we validate

Kim, D. and Choi, J.
Efficient Imputation Method for Missing Data Focusing on Local Space Formed by Hyper-Rectangle Descriptors.
DOI: 10.5220/0007582104670472
In *Proceedings of the 8th International Conference on Operations Research and Enterprise Systems (ICORES 2019)*, pages 467-472
ISBN: 978-989-758-352-0

the performance of proposed method by a numerical experiment in Section 4. Finally, we conclude our work and pose some ideas for future works in Section 5.

2 RELATED WORKS

There exist many imputation methods to handle missing data these days. For example, all missing data can be replaced with a single value. One can also consider using basic machine learning techniques such as regression analysis, k-NN and decision tree, and so on.

Mean or median imputation is a representative single value imputation method, which imputes missing values using a mean or median of observations not missing (Little and Rubin, 2014). Such imputation method is easy to implement and can perform well if there are few missing values. However, imputation with single value such as mean or median is not suitable in most cases since it cannot reflect variance and distribution of data, and imputed values are lack of representativeness.

By regression, one can obtain a mathematical model that describes relationship of input and output variables. Data imputation with regression is performed as follows: At first, a regression model is formed by using a feature with missing value as output variable and other features as input variables. Then, missing value of the feature can be computed by observed values of other features (Brown and Kros, 2003). Data imputation methods using regression can be categorized according to the method to formulate regression model. Regression model using Least-Squares (LS) is most common and basic (Raghunathan et al., 2001). Based on this, calculating LS sequentially or iteratively was also considered (Zhang et al., 2008; Shi et al., 2013). More complicated regression model such as Support Vector Regression (SVR) and nonlinear regression were also tackled (Aydilek and Arslan, 2013; Tang and Zhao, 2013). Clear disadvantage of regression-based imputation methods is that they cannot focus on local relationship.

Data imputation using k-NN utilizes information about the nearest neighbors of missing data. Specifically, k-NN imputation replaces missing values as follows. We select k nearest neighbors by considering features not in missing data. Then, missing value is estimated from features of observed nearest neighbors by using means or weighted means and so on (Chen and Shao, 2000). Choosing nearest neighbors sequentially or iteratively is one of the most common variations of k-NN-based imputation (Zhang, 2012; Kim et al., 2004). Huang and Zhu (2002) proposed imputation method based on pseudo-nearest neighbors, expected to follow the same Gaussian distribution. Christobel and Sivaprakasam (2013) devised class-wise k-NN that utilizes class information to choose nearest neighbors for labelled data set. García-Laencina et al., (2009) adopted mutual information as distance metrics for choosing nearest neighbor. Jonsson and Wohlin (2004) evaluated the performance of various k-NN-based imputation methods. k-NN-based imputation methods could partially utilize local relationship. However, there was no implicit standard about how to decide k.

Other methods suggested for missing data imputation including decision tree are as follows. To impute missing data, decision tree can be generated by rules to calculate missing value. C4.5 and CN2 are representative algorithms to build decision tree (Grzymala-Busse and Hu, 2000; Batista and Monard, 2003). Decision tree is useful to impute missing data by intuitive rules, but it may become too complex in case of using many features not scalable well. Also, it can be over fitted easily. Expectation-Maximization (EM) algorithm tries to impute missing data with new values likely to be missing, while maximizing likelihood function (Gold and Bentler, 2000). It can find imputation values systematically, but its performance depends on initial parameter setting. Furthermore, Bertsimas et al. (2017) considered missing data imputation as an optimization problem and proposed fast first-order methods to obtain high quality solutions for it.

3 SUGGESTION OF NEW IMPUTATION METHOD

3.1 Problem Definition and Solution Framework

We define notation and index necessary for describing missing data imputation problem as follows. We assume a whole data set $= \{\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_n\}$, containing n data points. Then, i-th instance $\mathrm{x}_i (i = 1,2, \dots, n)$ can be expressed as $\mathrm{x}_i = (x_{i1}, x_{i2}, \dots, x_{iq})$, where q is the total number of features. Also, there is an index set of missing entries $\mathcal{M} = \{(k, e) | e$ -th feature of k -th instance is missing$\}$. As a result, the objective of imputation problem is to substitute missing entry $x_{ke}, (k, e) \in \mathcal{M}$ by using fully collected data points.

Meanwhile, we carry out the imputation of missing data by using HRD_m, which is a method generating HRD based on interval merging. The framework of HRD_impute tackled in this research can be depicted as Figure 1. At first, we obtain HRD_m from data points except for missing ones and fit a linear regression model for each resulted HRD_m. Then, we identify HRD_m expected to include missing data and impute missing value from regression model fitted into HRD_m. Subsection 3.2 describes detailed procedure of missing data imputation method by using HRD_m.

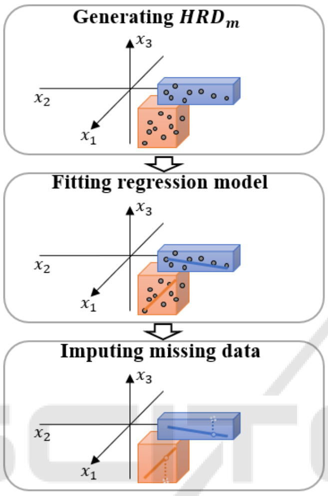

Figure 1: Framework of imputation USING HRD_m.

3.2 Detailed Procedure of HRD_impute using HRD_m

At first, we generate HRD_m from fully collected data points as follows (Jeong et al., 2018). The first step to construct HRD_m is to generate intervals for each feature r, which is main component of generating hyper-rectangle. An interval is calculated from a set of projection points of all instances into feature r. Specifically, we define the projection point of instance x_i into feature r as

$$proj_r(\mathrm{x}_i) = x_{ir}, \forall i. \tag{1}$$

By using (1), we can obtain the set of projection points Y_r containing $n_r(\leq n)$ points as

$$Y_r = \{y_r^1, y_r^2, \dots, y_r^{n_r}\}, \tag{2}$$

where y_r^h is h-th smallest value in Y_r satisfying $y_r^1 \leq y_r^2 \leq \cdots \leq y_r^{n_r}$. For each projection point y_r^h, we define an interval $itvl_r(y_r^h)$ as

$$itvl_r(y_r^h) = \left[y_r^h - \frac{l_r(y_r^h)}{2}, y_r^h + \frac{l_r(y_r^h)}{2}\right], \forall h, \tag{3}$$

where $l_r(y_r^h)$ is interval length calculated by the number of projection points with the same value and some parameters. Since intervals $itvl_r(y_r^h)$ are generated considering all projection points y_r^h, there can be overlapped intervals. For example, Figure 2 is an example of interval generation in two-dimensional feature space with five instances $\{\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3, \mathrm{x}_4, \mathrm{x}_5\}$. All instances were projected to each feature, and intervals were generated from each of projection points.

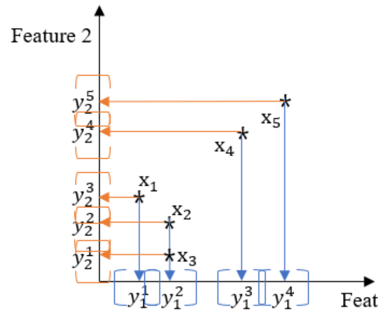

Figure 2: An example of interval generation for HRD_m.

While calculating intervals from projection points, some intervals may be overlapped. These overlapped intervals are merged, which is resulted in a set of disjoint intervals containing n_r^m disjoint intervals as $M_r = \{ITVL_r^1, ITVL_r^2, \dots, ITVL_r^{n_r^m}\}$. For example, there are overlapped intervals in Figure 2 such as $itvl_1(y_1^1)$ and $itvl_1(y_1^2)$. After applying merging operation to all of these overlapped intervals, resulting disjoint intervals in above-mentioned example were depicted in Figure 3.

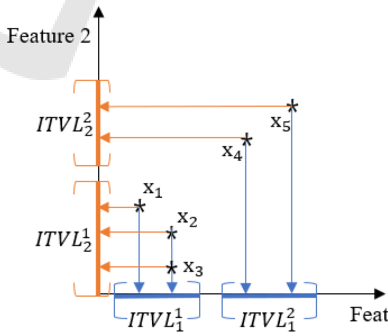

Figure 3: Disjoint intervals resulted from merging.

Next step is to get conjunction of these disjoint intervals. Conjunction of intervals can be obtained by cartesian product of intervals generated from all features. However, some conjunction of intervals may not include any instance since intervals are generated feature by feature. For example, conjunction of $ITVL_1^1$ and $ITVL_2^2$ does not include any instance. Thus, interval conjunction should be defined by considering instances. Specifically, interval conjunction $\pi(\mathrm{x}_i)$ including instance x_i is expressed as

$$\pi(\mathrm{x}_i) = ITVL_1^{t_1} \wedge ITVL_2^{t_2} \wedge \dots \wedge ITVL_q^{t_q}, \forall i, \qquad (4)$$

where $proj_r(\mathrm{x}_i) \in ITVL_r^{t_r}, \forall r$. From the information of disjoint intervals in Figure 3, for example, two interval conjunctions can be drawn as Figure 4.

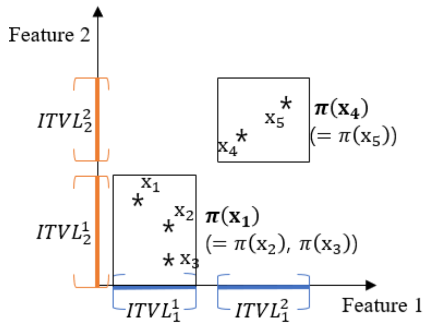

Figure 4: Interval conjunctions obtained from instances.

Merging-based H-RTGLs, MbH, are constructed by adjusting volume of each interval conjunction. However, some $MbHs$ may be overlapped even if they are formulated from disjoint intervals. We recommend Jeong et al., (2018) for revealing detailed adjusting procedure or other contents including roles of parameters.

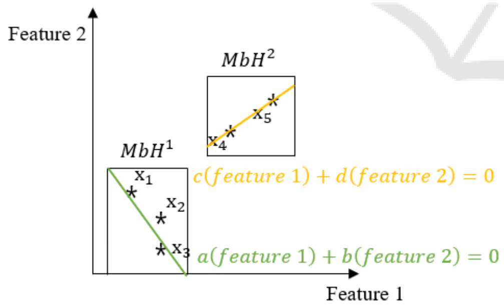

Figure 5: Two regression models fitted into two MbHs.

After such $MbHs$ are formulated, instances can be classified by HRD_ms. In other words, local spaces expected to include each instance with missing feature are identified, and imputation of missing data can be carried out by considering such relationship. We fit into a linear regression model for each MbH, and the regression model is calculated considering only instances belonging to MbH. Since there exist two $MbHs$ in above-mentioned example, two linear regression models can be fitted into as depicted in Figure 5.

Then, missing data can be imputed by using resulted linear regression model. Suppose that there are two more instances with missing entries x_6 and x_7, which are not used to formulate MbH due to missing entries. In addition, the index set of missing entries $\mathcal{M}$ is given as $\mathcal{M} = \{(6,2),(7,1)\}$, which means feature 2 of x_6 and feature 1 of x_7 are missing. If known values of x_6 (x_{61}) and x_7 (x_{72}) are belonging to respective MbH, imputed values should be calculated from the corresponding linear regression model. Figure 6 shows the imputation procedure for the given example. Missing values of instance $\mathrm{x}_6(x_{62})$ and instance $\mathrm{x}_7(x_{71})$ are imputed from the corresponding regression models. If there exist two or more candidates of $MbHs$, one MbH is randomly selected. The probability of each MbH to be selected is given by the number of instances belonging to MbH. If there is no MbH to be selected, the nearest MbH is used instead. In case of two or more missing features exits in one instance, candidate MbH is selected at first, and then random value on regression model corresponding to MbH is chosen as an imputation value.

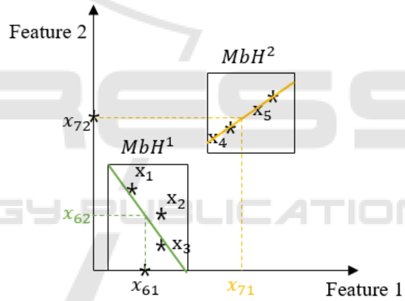

Figure 6: Imputed values for missing data.

4 A NUMERICAL EXPERIMENT

4.1 Experimental Design

To validate the performance of proposed missing data imputation method, we committed a numerical experiment by using real world dataset from UCI machine learning repository. We considered three datasets named Iris, Wine, E.coli. The number of instances and features (n, q) in each dataset is $(150,4)$, $(173,13)$, $(284,7)$, respectively. Missing entries are randomly generated in dataset at Missing at Completely Random (MCAR), and we used three missing percentage from 10% to 20%. We implemented two versions of HRD_impute with different parameter configurations, represented by HRD_impute_l and HRD_impute_s. The former approach generates MbH s from longer intervals, which make size of MbH to increase. In other words, HRD_impute_s separates whole feature space densely with many small MbH s. Also, we considered regression model fitted without HRD_m as control group. To evaluate imputation performance, we calculated Mean Absolute Percentage Error (MAPE) defined as

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{A_i - F_i}{A_i}\right|, \qquad (5)$$

where A_i is the actual value, and F_i is the forecasted value.

4.2 Experimental Results

Tables 1 to 3 summarize imputation performance of HRD_impute and control group in three datasets. 10 iterations were committed with the same missing percentage in dataset.

Table 1: Experimental result from Iris data.

	Missing percentage		
	10%	15%	20%
	Avg. MAPE (standard deviation)		
HRD_impute_l (δ, ρ)= (8, 0.5)	9.3 (1.2)	13.9 (1.4)	18.2 (1.3)
HRD_impute_s (δ, ρ)= (2, 0.2)	8.2 (0.9)	11.3 (1.0)	17.4 (1.2)
Regression	13.2 (1.8)	19.4 (2.5)	26.8 (2.6)

Table 2: Experimental result from Wine data.

	Missing percentage		
	10%	15%	20%
	Avg. MAPE (standard deviation)		
HRD_impute_l (δ, ρ)= (8, 0.5)	10.2 (1.3)	13.3 (0.9)	18.7 (2.9)
HRD_impute_s (δ, ρ)= (2, 0.2)	11.9 (0.8)	13.8 (1.4)	20.2 (2.3)
Regression	16.2 (1.3)	19.3 (2.1)	28.3 (3.2)

Table 3: Experimental result from e.coli data.

	Missing percentage		
	10%	15%	20%
	Avg. MAPE (standard deviation)		
HRD_impute_l (δ, ρ)= (8, 0.5)	13.5 (1.6)	17.9 (1.3)	21.4 (1.9)
HRD_impute_s (δ, ρ)= (2, 0.2)	13.7 (1.8)	16.8 (2.1)	20.4 (2.2)
Regression	15.9 (1.1)	20.3 (2.2)	27.9 (3.1)

As a result, imputation performance of HRD_impute was better than simple regression. This means that utilizing information of local space and local relationship can improve imputation performance of missing data. Regarding comparison of HRD_impute_l and HRD_impute_s , slight dominance of HRD_impute_s was observed in Iris dataset, while imputation performance of HRD_impute_l was a little bit better in Wine dataset. From these results, we can infer that dominance of different HRD_impute methods might depend on dataset.

5 CONCLUSIONS

In this paper, we proposed a new imputation method for missing data that can replace missing values by focusing local space having high potential to include missing data. Especially, HRD_impute proposed in this paper enabled local spaces to be identified systematically. HRD_impute was implemented by segmenting feature space into H-RTGLs and fitting regression models, which was a basis for imputation of missing values. As a result, missing values could be imputed by utilizing information of local space by using HRD_impute .

Even if performance of HRD_impute was validated through a numerical experiment, there are still plenty of further works to consider. Most of all, result of HRD_impute should be compared to other imputation methods rather than simple regression. Also, imputation performance may be improved by tuning parameters of HRD_m , since generation of HRD is sensitive to these parameters. Thus, thorough research of parameter tuning can be considered. We also plan to apply HRD_impute for large or complex dataset to verify scalability or generality of it. Moreover, HRD_m and imputation with regression can be substituted by other methods that have the same role. Detailed policies for multiple missing attributes in one instance can be another future research area.

ACKNOWLEDGEMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2017R1A2B4009841).

REFERENCES

Aydilek, I. B., & Arslan, A. (2013). A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. *Information Sciences*, 233, 25-35.

Batista, G. E., & Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied artificial intelligence*, 17(5-6), 519-533.

Bertsimas, D., Pawlowski, C., & Zhuo, Y. D. (2017). From predictive methods to missing data imputation: An optimization approach. *The Journal of Machine Learning Research*, 18(1), 7133-7171.

Brown, M. L., & Kros, J. F. (2003). Data mining and the impact of missing data. *Industrial Management & Data Systems*, 103(8), 611-621.

Chen, J., & Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official statistics*, 16(2), 113.

Christobel, Y. A., & Sivaprakasam, P. (2013). A New Classwise k Nearest Neighbor (CKNN) method for the classification of diabetes dataset. *International Journal of Engineering and Advanced Technology*, 2(3), 396-400.

García-Laencina, P. J., Sancho-Gómez, J. L., Figueiras-Vidal, A. R., & Verleysen, M. (2009). K nearest neighbours with mutual information for simultaneous classification and missing data imputation. *Neurocomputing*, 72(7-9), 1483-1493.

Gold, M. S., & Bentler, P. M. (2000). Treatments of missing data: A Monte Carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization. *Structural Equation Modeling*, 7(3), 319-355.

Grzymala-Busse, J. W., & Hu, M. (2000, October). A comparison of several approaches to missing attribute values in data mining. *In International Conference on Rough Sets and Current Trends in Computing*. 378-385

Huang, X., & Zhu, Q. (2002). A pseudo-nearest-neighbor approach for missing data recovery on Gaussian random data sets. *Pattern Recognition Letters*, 23(13), 1613-1622.

Jeong, I., Kim, D. G., Choi, J. Y., & Ko, J. (2019). Geometric one-class classifiers using hyper-rectangles for knowledge extraction. *Expert Systems with Applications*, 117, 112-124.

Jonsson, P., & Wohlin, C. (2004, September). An evaluation of k-nearest neighbour imputation using likert data. *In Software Metrics, 2004. Proceedings. 10th International Symposium on.* 108-118

Kim, K. Y., Kim, B. J., & Yi, G. S. (2004). Reuse of imputed data in microarray analysis increases imputation efficiency. *BMC bioinformatics*, 5(1), 160.

Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons.

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction*. Guilford Press.

Shi, F., Zhang, D., Chen, J., & Karimi, H. R. (2013). Missing value estimation for microarray data by Bayesian principal component analysis and iterative local least squares. *Mathematical Problems in Engineering 2013*, 1-5.

Tang, N. S., & Zhao, P. Y. (2013). Empirical likelihood-based inference in nonlinear regression models with missing responses at random. *Statistics*, 47(6), 1141-1159.

Templ, M., Kowarik, A., & Filzmoser, P. (2011). Iterative stepwise regression imputation using standard and robust methods. *Computational Statistics & Data Analysis*, 55(10), 2793-2806.

Trivellore E Raghunathan, James M Lepkowski, John Van Hoewyk, and Peter Solenberger. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27(1):85-96, 2001

Tutz, G., & Ramzan, S. (2015). Improved methods for the imputation of missing data by nearest neighbor methods. *Computational Statistics & Data Analysis*, 90, 84-99.

Zhang, S. (2012). Nearest neighbor selection for iteratively kNN imputation. *Journal of Systems and Software*, 85(11), 2541-2552.

Zhang, X., Song, X., Wang, H., & Zhang, H. (2008). Sequential local least squares imputation estimating missing value of microarray data. *Computers in biology and medicine*, 38(10), 1112-1120.