Estimating Semi-Parametric Missing Values with Iterative Imputation

Shichao Zhang, Zhejiang Normal University and Zhongshan University, China

ABSTRACT

In this paper, the author designs an efficient method for imputing iteratively missing target values with semi-parametric kernel regression imputation, known as the semi-parametric iterative imputation algorithm (SIIA). While there is little prior knowledge on the datasets, the proposed iterative imputation method, which impute each missing value several times until the algorithms converges in each model, utilizes a substantially useful amount of information. Additionally, this information includes occurrences involving missing values as well as capturing the real dataset distribution easier than the parametric or nonparametric imputation techniques. Experimental results show that the author’s imputation methods outperform the existing methods in terms of imputation accuracy, in particular in the situation with high missing ratio.

Keywords: Algorithms, Dataset Distribution, Iterative Imputation Method, Missing Values, Semi-Parametric Iterative Imputation Algorithm (SIIA)

INTRODUCTION

The real data usually are incomplete as some instances may have missing values. In fact, many reasons can result in missing values, for instance, malfunction of equipments, erroneous from human imputation, and so on. Missing values is an unavoidable problem in the real world, and various methods for dealing with such issues have been developed in data mining and in statistics. For example, case deletion, learning with no handling with missing data and missing values imputation. In real application, the imputation method is a popular strategy comparing to other methods. Missing values imputation is to find an efficient way to "guess" the missing values (imputation) based on other information in datasets. One advantage of this approach is that missing values treatment is independent of the learning algorithm used. That allows users to select the most suitable imputation method for their applications.

Commonly used imputation methods for missing values include parametric regression imputation methods and non-parametric regression imputations. However, there are other relations within real world data, and both parametric imputation method and non-parametric imputation method are not adequate to capture the relations. That is, we know a part of relation

DOI: 10.4018/jdwm.2010070101

Copyright © 2010, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
between independent variables (condition attributes) and dependent variable (target attribute), e.g., we can regard this relation as parametric model, but we have no knowledge on the relation between other independent variables and dependent variable, e.g., we can take it as nonparametric model. However, combining these two parts, it is difficult for us to consider the compound relation with parametric model or nonparametric model. Moreover, the case is very general in real application. In this paper, we regard the relation containing two models as semi-parametric model or partial parametric model. In real application, semi-parametric model is natural than non-parametric model because users can always know some information but no all on the datasets, such some parameters in the datasets. To model this semi-parametric relation, in this paper, we design an efficient semi-parametric iterative imputation method (SIIMA) that takes into account the advantages of parametric models and pure non-parametric models so as to overcome their certain shortcomings for each single model.

In the left parts, we will first review the existing literatures for dealing with missing values. And then we design the iterative imputation methods which can impute missing values with kernel method or even in the dataset with high missing ratio. After that, we will demonstrate our proposed methods with all kinds of experiments. Finally, we will conclusion our works and put forward our future work.

RELATED WORK

There are at least three different ways of dealing with missing data based on Little and Rubin (2002): single imputation, multiple imputation, and iterative procedure.

Single imputation strategies provide a single estimate for each missing data value. Many methods for imputing missing values are single imputation methods, such as, C4.5 algorithm, kNN method, and so on. We can partition single imputation methods into parametric methods and nonparametric ones. The parametric regression imputation methods are superior if a dataset can be adequately modeled parametrically, or if users can correctly specify the parametric forms for the dataset. Non-parametric imputation (Qin et al., 2007) offers a nice alternative if users have no idea on the actual distribution of a dataset because the method can provide superior fits by capturing structure in datasets. While much work focus on modeling data by parametric or nonparametric approaches, in Engle et al. (1986) have studied the semi-parametric model. They model the electricity demand $y$ as the sum of a smooth function $g$ of monthly temperature $t$, and a linear function of $x_1$ and $x_2$, as well as 11 monthly dummy variables $x_{3,1}, \ldots, x_{3,11}$ to build a semi-parametric model firstly. In fact, semi-parametric model is more ordinary in real application than nonparametric model or parametric model because we always contain a little but no all information on our datasets, however, there are a little literatures, such as, Nikulin (2008), focusing on this issue because of the analysis complexity, in this paper, we introduce SIIMA algorithm to model the partial parametric model for filling up iteratively missing target values.

A disadvantage of single imputation strategies is that they tend to artificially reduce the variability of characterizations of the imputed dataset. The alternatives are to fill in the missing values with multiple imputation methods (e.g., Multiple Imputation (MI); Golfin & Rizz, 2009) and iterative imputation methods (EM algorithm). In multivariate analysis, MI methods provide good estimations of the sample standard errors. However, data must be missing at random in order to generate a general-purpose imputation. In contrast, iterative approaches can be better developed for missing data since it can utilize all useful information including the instances with missing values (Pighin & Ironon, 2008). That can receive significant performance in the datasets with high missing ratio. The well-known of these methods is the Expectation-Maximization (EM) algorithm for
parametric model. Articles (Caruana, 2001; Pighin & Ieronutti, 2008) present an EM-style nonparametric iterative imputation model embedded with kNN algorithm to impute missing attribute values.

In this paper, at first, we impute missing values with general methods (e.g., mean method) in order to utilize substantial all the useful information in the datasets, then we impute each missing values iteratively based on kernel regression imputation method until the algorithm converges. In this process, given part information between independent variables and dependent variable, we present semi-parametric iterative imputation algorithm (SIIA) to deal with missing target values since second imputation. Different from the existing parametric methods and nonparametric methods, the proposed method can be easily applied to real application because we usually have a little supplement knowledge on our datasets. Different from single imputation method and multiple imputation method, our iterative method can utilize substantial all useful information including the information in the instances with missing values for improving imputation performance.

SIIA ALGORITHM

SIIA method imputes each missing value several times until the algorithm converges. In the first iteration, all complete instances are used to estimate missing values. The information in instances with missing values is used from the second iteration. This method is of benefit to capture the distribution of a dataset much better and easier than parametric/nonparametric imputation or existing single imputation methods.

Existing imputation methods usually impute missing values with the instances without missing values, such as, C4.5, kNN algorithm, and so on. Indeed, the information in the instances with missing values can play an important role to estimate missing values. For example, it can be applied to identify the neighbors of an instance with missing values in nearest neighbor (NN) imputation algorithms, or the class of the instance in clustering-based imputation algorithms. On the other hand, the datasets do not enough complete instances to impute missing values even if the datasets contain low missing ratio. For example, the missing ratio in UCI datasets (Blake & Merz, 1998) Bridge only is 5.56% which is a low missing ratio in real application because most datasets in industrial reach to 50% or above. However, there are 38 complete instances in the dataset with 108 instances and 6 class labels. If we impute the missing values with only 38 complete instances with any imputation algorithms, the imputation performance will generate easily bias due to less complete instances because the size in large sample should be beyond 30 in statistics. On the other way, there are 6 classes in this dataset in which the maximal number of complete instances for one class only contains 11 complete instances, the classification accuracy won't be well even if the most excellent classification algorithm is employed. Furthermore, many datasets in UCI present the same case as Bridge. Hence, to utilize the information in the instances with missing values is useful for building our imputation model especially in the case with high missing ratio.

In SIIA, we employ some existing methods which fit for statistical proof (such as, mean/mode method) in the first imputation for all missing values in the dataset. After all missing values have been imputed once, we re-impute all the missing values till the algorithm converges. Since the second imputation, imputation is condition on all available information, that is, if the missing values have been imputed prior to being used as observed values for the other missing values. Generally speaking, we denote missing value as $M^i_v, i = 1, \ldots, n$ (n is the number of missing values) corresponding to the imputed missing values denoted as $\hat{M}^i_v, i = 1, \ldots, n, j = 1, \ldots, t$ (j is the imputation time), all missing values $M^i_v$ are im-

Copyright © 2010, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
puted as \( \hat{MV}_i^j \) with the first imputation. Since the second imputation, the observed information will include \( \hat{MV}_i^{j-1}, i = 1, k-1, k+1, ..., n, j = 2, ..., t-1 \) while we want to impute a missing value \( \hat{MV}_i^j, k = i, j = 2, ..., t \), the imputation process will continue till algorithms reach to approximate convergence. Meanwhile, since the second imputation, we will employ kernel regression method for imputing missing values under nonparametric model or semiparametric model. The pseudo of algorithm SIIA is presented as follows:

**IMPUTATION AT FIRST ITERATION**

There exist many methods to fill in missing values in the first iteration of imputation including any single imputation methods, such as, C4.5 algorithm, kNN algorithm, and so on. In Qin et al. (2007), the authors compute mean (or the mode if the attribute is categorical) to impute missing values in the first time. They think that the method is a popular and feasible imputation method in data mining and statistics. Meanwhile, they also believe to impute with the mean (or mode) is valid if and only if the dataset is chosen from a population with a normal distribution. However, in real world ap-

---

**Algorithm 1.**

```plaintext
//The first iteration of imputation
FOR each \( MV_i^j \) in Y
    \( \hat{MV}_i^j = \text{mode} \left( S_i \text{ in Y}\right) \); // if Y is discrete variable
    \( \hat{MV}_i^j = \text{mean} \left( S_i \text{ in Y}\right) \); // if Y is continuous ones
END FOR
//t-th iteration of imputation (t>1)
t = 1;
REPEAT
    t + 1;
    FOR each missing value \( MV_i^j \) in Y
        Get \( \beta^t \) based on Equation (7)
        IF \( MV_i^j \) is current imputed missing value
            \( MV_i^t = \hat{MV}_i^{j-1}, p \in S_i, p = 1, ..., m, p \neq i \); // if Y is continuous variable
        ELSE
            \( MV_i^t = \begin{cases} 
                0 & \text{if } \hat{MV}_i^{j-1} < x_i, \\
                1 & \text{if } \hat{MV}_i^{j-1} \geq x_i
            \end{cases} \); \( - MV_i^t, p \in S_i, p = 1, ..., m, p \neq i \); // if Y is discrete variable
        END IF
    END FOR
    UNTIL //finishing iteration of imputation
    \( \frac{M_i}{V_i} \rightarrow 1 \), and \( \frac{V_i}{V_{i+1}} \leq \varepsilon \)
3.0 //output the imputation iteration times and imputation results
OUTPUT
    t; // t is the iterative times
    Completed dataset;
```
plication, we cannot know the real distribution of the dataset in advance. So running the extra iterations of imputation to improve imputation performance is reasonable based on the first iteration of imputation for dealing with the missing values. Caruana (2001) thinks the first step, which imputes each missing value with the mean/median values calculated from cases that are not missing that value, will cause cases missing many values to appear to be artificially close to each other. For example, if both cases being compared were missing values for attribute $x$, the distance along dimension $x$ will be 0 since they will be imputed with the same value. The paper demonstrates this subtlety is not critical for the proper behavior of the method, but does speed convergence on datasets that have many missing values. However, the method in Caruana (2001) is designed to impute missing attribute values rather than missing target values. In our paper, we will employ mean/mode method to impute missing values in the first imputation.

SUCCESSIVE ALGORITHM

In real application, we can know partial relation between condition attributes and target attribute. For example, letting us consider the sales of ice-cream in summer, condition attributes can be weather, sale place, and so on. We could conclude easily that the relation between weather and the sale of ice cream is linear because the sale of ice-cream increases at hotter weather, hence, part condition attribute (such as, weather) is related linear to target attribute and we can explain the relation as parametric model (such as, linear model in this example). However, it is impossible for us to really know the relation between the other factors (such as, sale place, some unthinkable reasons, or the others condition attributes) and the sale of ice cream but we know they must relate to the sale of ice-cream. Obviously, we can analyze the relation using non-parametric method under the assumption without any knowledge. Given combined these two relations, we can analyze them with single nonparametric mode or parametric model respectively, however, we cannot model the real relation between independent variables and dependent variable no matter that which one strategy is employed. So in this case, it is reasonable for us to build semi-parametric model to deal with this case.

In single semi-parametric imputation method (Nikulin, 2008), a general semi-parametric regression model is as follows:

$$Y_i = X_i^T \beta + g(T_i)$$

(1)

Where the $Y_i$'s are i.i.d (independent identically distributed) scalar response variables, the $X_i$'s are i.i.d $d$-dimensional random covariate vectors (parametric model), the $T_i$'s are i.i.d $d^*$-dimensional random covariate vectors, the function $g(T_i)$ is unknown (non-parametric model) on (Pearson, 2005; Pighin & Ieroncutti, 2008; Qin et al., 2007).

In SIAM, for each iterative imputation, the value of $\delta$ will be computed firstly, then compute the value of $g(T_i)$ based on the $\delta$ based on Equation 3, we define:

$$Y_i = X_i^T \beta + g(T_i)$$

(2)

Where

$$Y_i = \begin{cases} Y_i, & \text{if } \delta_i = 0 \text{ or } i = 1, \ldots, r \smallskip Y_i^{t+1}, & \text{if } \delta_i = 1 \text{ or } i = r+1, \ldots, n \end{cases}$$

Where $t$ is the number of iterative imputation, $\delta$ based on the first imputation. From (2), we have:

$$Y_i - X_i^T \beta = g(T_i). \quad i = 1, \ldots, r$$
Assuming $\beta'$ is known in advance, we have a kernel estimator $\hat{g}(T'_i)$ for $g(T'_i)$ based on the completely observed data:

$$\hat{g}(T'_i) = \frac{1}{n} \sum_{j=1}^{n} \frac{K\left(\frac{T_i' - T_j'}{h}\right)(Y_i' - X'_i \hat{\beta}')}{\int K\left(\frac{T_i' - T_j'}{h}\right) + n^{-2}}, \quad i = 1, \ldots, r$$

(3)

Using $\hat{g}(T'_i)$ to replace $g(T'_i)$ in Eq. (3), we obtain:

$$Y_i' = X_i'^T \hat{\beta}' + \frac{1}{n} \sum_{j=1}^{n} \frac{K\left(\frac{T_i' - T_j'}{h}\right)(Y_i' - X'_i \hat{\beta}')}{\int K\left(\frac{T_i' - T_j'}{h}\right) + n^{-2}}, \quad i \in s_r.$$

(4)

Converting (4), we have

$$Z'_i \approx U_i'^T \hat{\beta}', \quad i \in s_r$$

(5)

Where

$$Z'_i = Y_i' - \frac{1}{n} \sum_{j=1}^{n} \frac{K\left(\frac{T_i' - T_j'}{h}\right)Y_j'}{\int K\left(\frac{T_i' - T_j'}{h}\right) + n^{-2}}, \quad i \in s_r,$$

$$U_i' = X_i' - \frac{1}{n} \sum_{j=1}^{n} \frac{K\left(\frac{T_i' - T_j'}{h}\right)X_j'}{\int K\left(\frac{T_i' - T_j'}{h}\right) + n^{-2}}, \quad i \in s_r.$$  

(6)

According to the theory of linear regression model, $\beta'$ is estimated by Eq. (7):

$$\hat{\beta}' = \left(\sum_{i=1}^{n} \delta_i U_i'^T \hat{U}_i \right)^{-1} \left(\sum_{i=1}^{n} \delta_i U_i' Z_i' \right).$$

(7)

Where $n$ is the sample size. Note that, for simplicity, the transform from Equation 4 to 5, we assume parametric model is linear for estimating the value of $\beta'$, however, in real application, parametric model can be linear model, nonlinear model, and so on.

Combining with (3), the final estimator for $\hat{g}(T'_i)$ is given by

$$\hat{g}(T'_i) = \frac{1}{n} \sum_{j=1}^{n} \frac{K\left(\frac{T_i' - T_j'}{h}\right)(Y_i' - X'_i \hat{\beta}')}{\int K\left(\frac{T_i' - T_j'}{h}\right) + n^{-2}}.$$

(8)

Hence, in algorithm SIIA, since the second imputation, we use Equation 2 to impute missing target values till the algorithm converges.

In fact, the imputed values based on Equation 2 always is continuous values, and our SIIA algorithm can also impute discrete missing target attribute which is presented in pseudo of SIIA Algorithm 2.0. In our paper, we consider the case with two classes and the reader can extend our method to the case with multiple classes. In SIIA algorithm, instances are defined as belonging to class 0 if $\hat{M}V_i' < \chi$, and class 1 otherwise. The actual value of the class for each incomplete instance $x_i$ is denoted by $MC_i$. The new class assignment based on the imputed class is denoted by $\hat{MC}_i$ in t-th imputation to stress the dependence of the classification on . More specifically, the imputed value $\hat{m}_i(x) \in R$ is transformed into a (binary) class $MC_i' \in \{0, 1\}$ $\forall x_i \in D$ based on the
rule specified in SIIA algorithm. $\chi$ is specified by the user of the technique and in many applications is set so that $|\{x_i | MC_i^t = 1\}| = |\{x_i | MC_i = 1\}|$ (i.e., the number of class 1 instances before and after the application of our technique is the same). This rule for class assignment is the most natural choice and is the one primarily considered in our case studies, although the user of the technique may explore different choices near this preferred cutoff point. Hence, our proposed algorithm SIIA can also be used to impute discrete missing target values.

Finally, we can output the final imputation result after algorithm converged. Note that, the imputation times is $(t+1)$ rather than $(t+2)$ times even if the procedure is performed $(t+1)$ times and the first iteration is added. That is because the last imputation does not generate imputation result and only judge the fact whether the imputation reaches to convergence.

### ALGORITHM CONVERGENCE AND COMPLEXITY

An important practical issue concerning iterative imputation method is to determine at which point additional iterations have no meaningful effect on the imputed values, i.e., how to judge the convergence of the algorithm. Literatures Caruana (2001) and Pearson (2005) conclude that the average distance that the missing attribute values move in successive iterations drops to zero, that no missing values have changed and that the method has converged in nonparametric model. Here, we outline a strategy for the stopping criterion for our algorithms. With imputation times, assuming mean and variance of three successive imputations are $M_{i1}, M_{i2}, M_{i3}$, and $V_{i1}, V_{i2}, V_{i3}$, $(1 < l < t - 2)$ respectively.

If $\frac{M_{i1}}{M_{i+1}} > 1$, and $\frac{V_{i+1}}{V_{i2}} \leq \varepsilon$

That can be inferred that there is little change in imputations between the last and the former imputation, and the algorithm can be stopped for imputing without substantial impact on the resulting inferences. Different from the converged condition in existing algorithms, we summarize our stopping strategy using terminology such as ‘satisfying a convergence diagnostic’ rather than ‘achieving convergence’ to clarify that convergence is an elusive concept with iterative imputation.

While the complexity of the kernel method is $O(nm^2)$, where $n$ is the number of instances of the dataset, $m$ is the number of attributes, so the algorithm complexity of both NIIA and SIIA is $O(knm^2)$ ($k$ is the number of iteration times).

### EXPERIMENTAL ANALYSIS

In order to show the effectiveness of the proposed methods, extensive experiments are done on real dataset with VC++ programming by using a DELL Workstation PWS650 with 2G main memory, 2.6G CPU, and WINDOWS 2000. We compare the performance of SIIA with the existing methods in Nikulin (2008), parametric method as well as nonparametric method.

At first, we design different algorithms to impute target missing values, such as, our proposed algorithm SIIA, we also design three single imputation methods (stochastic semi-parametric imputation method for single imputation in Nikulin (2008), Semi for shorted, nonparametric imputation method in Qin et al. (2007) and Qin, Zhang, and Zhang (2010), Non for shorted, linear imputation method, Linear for shorted). We use RMSE to assess the predictive ability after the algorithm has converged for iterative imputation methods or the missing values are imputed for single imputation methods:
$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (e_i - \hat{e}_i)^2}$$

Where $e_i$ is the original attribute value; $\hat{e}_i$ is the estimated attribute value, and $m$ is the total number of predictions. The larger the value of the RMSE, the less accurate is the prediction.

We use the dataset in http://lib.stat.cmu.edu/DASL/Datafiles/USTemperatures.html to analyze the advantages of SIIA algorithm. The data give the normal average January minimum temperature in degrees Fahrenheit (Denoted as JanTemp) with the latitude (Lat) and longitude (Long) of 56 U.S. cities for each year from 1931 to 1960. Qin et al. in Nikulin (2008) presents there is an evident linear relationship between JanTemp and Lat, but the linear relationship between JanTemp and Long is not clearly. To apply our method to these real data, we denote the variables for JanTemp, Lat and Long to be Y, X and T respectively in the semi-parametric model. We suppose that Y, X and T satisfy the semi-parametric model. Because there exists linear relation as well nonparametric relation between independent variables and dependent variable in the real dataset, it is reasonable for us to design these algorithms to verify the advantages of our proposed algorithm in our experiments. On the other hand, there exist little real dataset containing partial relation between independent variables and dependent variable, the experiment only utilizes one dataset for imputing continuous missing target attribute in this paper.

Note that the original data set is complete. We used all the 56 data and delete randomly 6, 14 or 23 Y values (Missing Rate is almost 10%, 20% or 40% respectively) and the repeated times are 1000.

Figures 1, 2 and 3 present the values of RMSE for our algorithm SIIA and single imputation method Linear, Non, and Semi at different missing ratio 10%, 20% and 40% respectively.

These results show SIIA can converge in semi-parametric model at any missing ratio. The higher missing ration, the more imputation times the algorithm SIIA is.

Comparing SIIA method with single imputation methods, in the first several imputation, the imputation results of SIIA are worse than single imputation methods at different missing ratio, for example, the results of RMSE in the first three times corresponding to algorithm Non, and the imputation times in the first 6 times, 6 times, and 7 times corresponding to algorithm Semi at missing ratio 10%, 20% and 40% respectively are better than SIIA. Since then, our method outperform single imputation methods, in particular, in the high missing ratio case (such as, 40%), our method presents significant profit than single imputation methods. Moreover, in each imputation time, the performance of our proposed method

\begin{figure}
\centering
\includegraphics[width=0.5\textwidth]{fig1.png}
\caption{Missing rate is at 10%}
\end{figure}
outperforms the result of parametric method (such as, linear algorithm). At the same time, both semi-parametric single imputation and nonparametric method are better than parametric method. That presents, parametric model can not present good imputation performance if there is no priori knowledge or a little information on the datasets. To our surprised, there is a variation in Figure 3, that is, the RMSE of SIIA in 9th /10th imputation is worse than the result in 8th /10th one. The experimental results have a little fluctuate is because there are only 56 instances in the dataset. However, in left iterative imputations, the result of successive imputation will be better than the former one. That assures the algorithm converges. That shows our algorithm SIIA can tune the imputation performance because the situation does not occur again until algorithm converges.

CONCLUSION AND FUTURE WORK

In this paper, for dealing with missing values with high missing ratio, we present an iterative imputation method (SIIA) for imputing missing target values. That is, SIIA algorithm can deal with the case in which we have a little knowledge with our datasets by building semi-parametric model with kernel regression method to impute iterative missing target values. Experimental results show our methods outperform than the existing single imputation method in term of RMSE at different missing ratio in different cases. In our future work, we will focus on the study how to more effective estimate the value of $\delta$ in semi-parametric model.

Figure 2. Missing rate is at 20%

![Graph showing RMSE vs Imputation Times for 20% missing rate]

Figure 3. Missing rate is at 30%

![Graph showing RMSE vs Imputation Times for 30% missing rate]
ACKNOWLEDGMENT

This work was supported in part by the Australian Research Council (ARC) under grant DP0985456, the Nature Science Foundation (NSF) of China under grant 90718020, the MOE Project of Key Research Institute of Humanities and Social Sciences at Universities (07JJD720044), the Guangxi NSF (Key) grants, and the Guangxi Colleges and Universities' Innovation Teams.

REFERENCES


Pearson, P. K. (2005). Mining imperfect data: dealing with contamination and incomplete records. SIAM.


Shichao Zhang is a distinguished professor and the director of Institute of Computer Software and Theory at Zhejiang Normal University, China. He holds a PhD degree in Computer Science from Deakin University, Australia. His research interests include data analysis and pattern discovery. He has published about 50 international journal papers and over 60 international conference papers. He is a CI for 10 competitive nation-level grants (3 China NSF grants, 2 China 863 Program grants, 1 China 973 Program grant, and 4 Australia large ARC grants). He is a senior member of the IEEE, a member of the ACM; and served as an associate editor for IEEE Transactions on Knowledge and Data Engineering, Knowledge and Information Systems, and IEEE Intelligent Informatics Bulletin.

Copyright © 2010, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.