Missing value imputation for predictive models on large and distributed data sources

Jing Shyr [1], Jane Chu[1], Sier Han[2]
[1]IBM SPSS Predictive Analytics, 233 S. Wacker Dr. 11th Fl., Chicago, IL 60606, USA
[2] IBM SPSS Predictive Analytics, 2&3 Floor, Building C, Outsourcing Park Phase I, No. 11 Jinye 1st Rd., High Tech Zone, Xi'an, China

Abstract

The paper proposes a method to impute missing values of predictors for the subsequent predictive models on large and distributed data sources using a Map-Reduce approach. Firstly, for each predictor that has missing values, imputation models based only on the target variable are built independently on different data sources and on different machines using the Map functions. During the step, validation samples are extracted randomly across all data sources and merged into one global validation sample along with the collection of imputation models using the Reduce function. Then all imputation models are evaluated based on the global validation sample in a distributed manner using another set of Map functions to select the top K models and form an ensemble model. Thirdly, the ensemble model is sent to each data source to impute missing values of predictors. Finally, the complete dataset can be used to build any models for prediction as well as discovery and interpretation of relationships between the target and a set of predictors.

Different types of imputation models are built based on whether the predictor and target are categorical or continuous. Since only the target variable is used, only basic statistics between the predictor and target variables, such as means, variances, covariance, counts, etc. need to be collected using a single data pass which is important for the large and distributed data sources.

Key words: missing value imputation, basic statistics, MapReduce.

1. Introduction

Predictive models are widely used and are often built on demographic, survey and other data that contain many missing values. When these missing values aren’t handled appropriately during model building, then the predictive model isn’t reliable and any decision based on it might result in losses for the company. Thus, the treatment of missing values is an important problem to consider carefully. In addition, the very large and distributed data sources that are now encountered in practice call for new approaches in order to handle them efficiently.

There are many existing methods addressing the imputation of missing data. Howell (2008) provides a detailed summary of those methods. We also include one patent in our discussion. The following is a brief summary of those methods.

(1) Mean imputation: Replace missing values of a continuous variable with its mean (computed based on all non-missing records).

(2) Regression imputation: Regress the variable that has missing values on all other variables then use the regression equation to impute missing values for that variable. Random errors can be added to the imputed values to overcome the problem in underestimating the variance in the imputed variable.

(3) Multiple imputation: Similar to the regression imputation, multiple imputation needs to build imputation models for a variable that has missing values on other variables. The imputation model is a linear or logistic regression model for a continuous or categorical variable that has missing values, respectively. The key idea of multiple imputation is to impute multiple complete datasets by its imputation process. Then an appropriate predictive model is built on each complete dataset and the results of the multiple predictive models are combined.

(4) Expectation maximization (EM): The EM algorithm is an iterative method: estimate the model parameters then impute the missing values; use the filled-in dataset to re-estimate the parameters then use the re-estimated parameter to impute missing values; and alternate these 2 steps until the process converges on stable estimates.

(5) Method proposed by Bhaskar and Sundararajan [2] (US 2009/0177598 A1): Impute missing values while building a predictive model. A population of solutions is created using the dataset with missing values, wherein each solution comprises parameters of the model and the missing values. Each of the solutions in a population is checked for fitness. After the fitness is checked, the solutions in a population are genetically evolved to establish a successive population of solutions. The process of evolving and checking fitness is continued until a stopping criterion is reached.

Mean imputation is simple and can be readily applied; however, it has been proven to provide extremely inaccurate results.

Regression imputation was, for a time, the best of the simple solutions. However, it gives inaccurate results if the data don't follow the assumptions of a linear regression model; for example, when the variable to be imputed is a categorical variable, the variable to be imputed has a non-linear relationship with other variables used to impute, and so on. Moreover, it usually regresses the variable with missing values on all other variables, so the imputation model building cannot be based on some basic descriptive statistics only.

Multiple imputation became commonly used to handle missing values over the last 15 years. However, the whole process is very time consuming even for moderately-sized datasets. The iterative nature of the imputation process can require many data passes to impute a single complete dataset, and those data passes are multiplied when multiple complete datasets are created. This tremendous computation cost makes it almost impossible for multiple imputation to handle very large and distributed data sources. In addition, although multiple imputation corrects one of regression imputation's drawbacks by using logistic regression for a categorical variable with missing values, it would need several data passes to obtain the solution just for one variable within one imputation because, unlike linear regression, logistic regression doesn't have a closed form solution and need an iterative process which each iteration means one data pass. Thus the existence of categorical variables with missing values would increase computation cost even more for using multiple imputation.

The EM algorithm along with multiple imputation are two most important methods of imputing missing data in the recent literature. Typically the EM algorithm is used under a multivariate normal model and missing values are imputed based on a regression model, so it has the same drawbacks of regression imputation. Moreover, it is an iterative process that requires many data passes, so it has similar drawbacks to multiple imputation. If the predictive model of interest is more complicated than a multivariate normal model, then the EM algorithm is a system of equations which has specific forms for specific applications. Thus applying it in practice often requires considerable skill to obtain the custom-made solutions for different applications.

The method proposed by Bhaskar et al. is an evolving process which usually needs many runs of populations of solutions to reach stopping criterion so many data passes are needed to check fitness. The computation cost would prevent it from handling very large and distributed data sources.

In summary, both the EM algorithm and the method by Bhaskar et al. need many data passes to estimate missing values as an integral part of the model building process, so applying them to different predictive models would need different ways of imputing missing values to incorporate with specific models. On the other hand, regression imputation and multiple imputation would impute missing values first then run whatever analysis is appropriate on the complete dataset (two-step process), but they usually use complicated linear or logistic models which regress the variable with missing values on all other variables such that they might not be able to handle large and distributed data sources. In addition, using the linear regression for continuous variable with missing values might give inaccurate results because linear regression models might not catch nonlinear relationship among variables.

Due to the deficiencies of the existing methods, we propose an efficient system of missing value imputation which has the advantages of the two-step process and the ability to model possible nonlinear relationships by building piecewise linear regression models between the variable to be imputed and the variable used to impute. Contrary to other two-step process methods, we use only the target variable for the subsequent model building to impute missing values in the predictors. Most importantly, our system can handle missing value imputation with reasonable accuracy for large and distributed data sources because it requires only one data pass to collect necessary statistics between the predictors with missing values and the target to build imputation models, which include piecewise linear regression models to be built on a list of bins that are arranged by the locations of the predictors. We call such a system the “target-based” (TB) method.

The rest of this paper is organized as follows: Section 2 will describe the missing value imputation process we propose in details. Section 3 will describe how different imputation models are built. Some examples from simulated will be given in Section 4 and a few concluding remarks are in Section 5.

2. Missing value imputation process

Our target-based method of missing value imputation process consists of three steps. The first step is to build imputation models and extract validation samples. The second step is to evaluate the imputation models based on a global validation sample and select the top K of them as an ensemble model. The third step is to impute missing values in data based on the ensemble model. The whole operation can be implemented in Hadoop with

MapReduce interface or other comparable systems in order to handle large and distributed data sources and computation in parallel.

In the first step, the data are distributed into Mappers. An imputation model between each predictor with missing values and a target variable is built in each Mapper independently. What kind of imputation model will be built depends on the measurement levels of predictor and target variables and the details will be given in Section 3. At the same time, a validation sample is randomly extracted from each Mapper independently. The Mappers will pass the validation samples and imputation models for all predictors to a single or multiple Reducers. The Reducers merge validation samples into a global validation sample and have a collection of N imputation models of each predictor with missing values.

In the second step, a global validation sample is scored by each Mapper to evaluate the accuracy of an imputation model. The Reducer selects the top K models out of N possible imputation models based on some accuracy measures as the final ensemble model for each predictor with missing values.

In the third step, an imputation server includes the final ensemble model for each predictor with missing values, and a so-called “imputation strategy” is sent to all Mappers to impute missing values of each predictor. If the subsequent model building processes would also run computations in a distributed manner using the Map-Reduce, then there is no need to have a Reducer in the third step. However, if the complete data set should be exported, then the Reducer would be used to gather data together.

Then the complete datasets (for all possible predictors) can be used to build any models for prediction as well as discovery and interpretation of relationships between the target and a set of predictors.

The imputation strategy can be defined according to how the complete dataset (for all possible predictors) would be generated. One possible strategy is to impute missing values by the mean of K predicted values for the continuous predictor to be imputed from the final ensemble model and by the mode of K predicted values for the categorical predictor. The other possible strategy is to impute missing values by the predicted value from a randomly selected imputation model out of K models for the predictor to be imputed.

Figure 1 illustrates the target-based method of missing value imputation process in details.

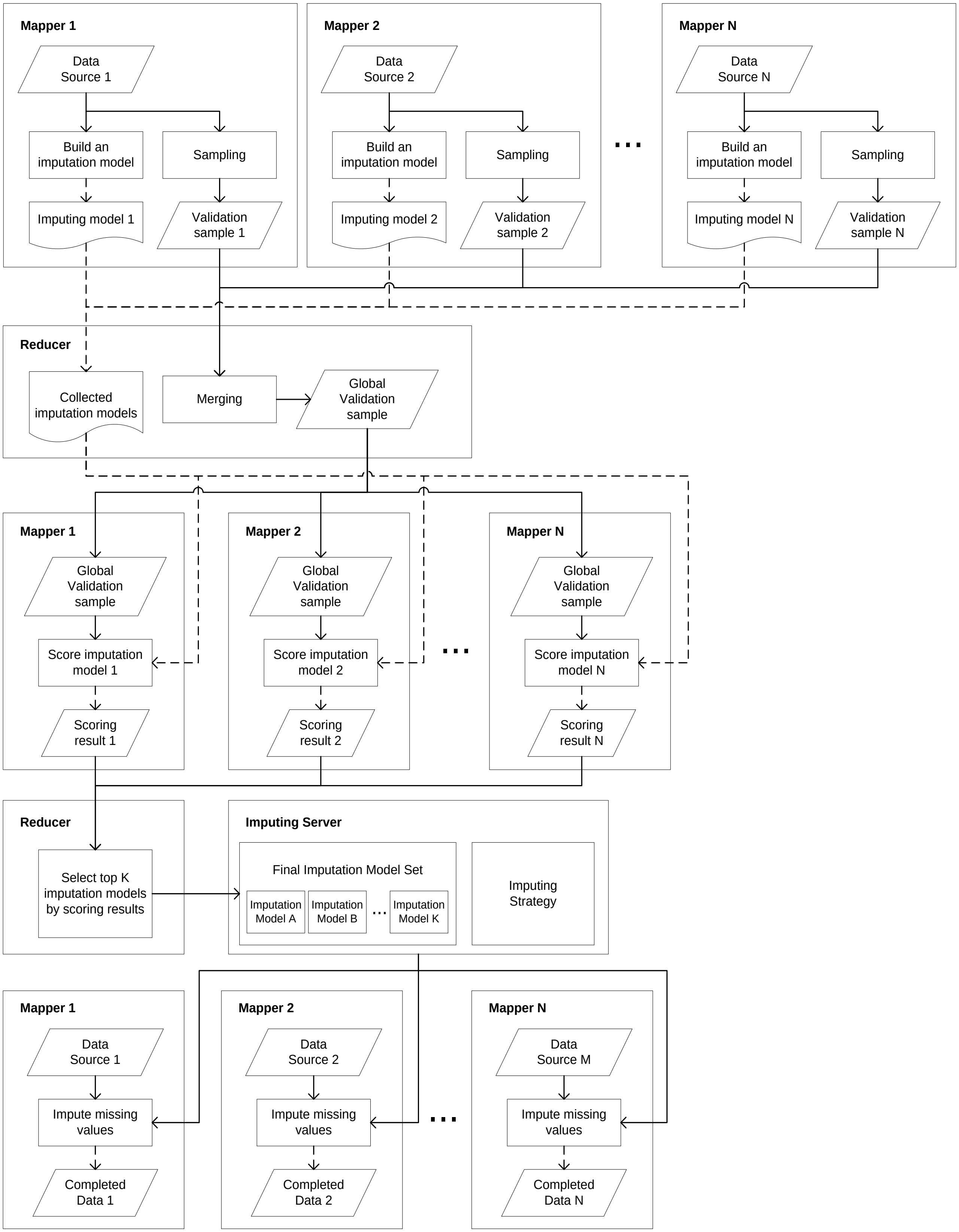

Fig 1 Map-Reduce framework of Distributed Missing Value Imputation

Based on the above description of the missing value imputation process, there are some features that can distinguish our method from other methods in the literature.

First, unlike regression imputation and multiple imputation methods, our system just uses the target to build imputation models in the process due to the following reasons:

(1) If one variable should be selected to make imputation models building feasible on large and distributed data sources , the target is the only variable related and relevant to all possible predictors with missing values.
(2) Imputation models for all predictors with missing values can be built independently as they only depend on the target, so it is not necessary to build imputation models with all variables sequentially and iteratively which is the process used by multiple imputation and many data passes can be saved.
(3) While some information may be lost by not using all possible predictors to build the imputation models, our experiments indicate the accuracy results on the subsequent model building processes from our proposed process by using the target only are similar to or even better than those multiple imputation by using all other variables.

Thus only univariate and bivariate statistics between the target and a predictor with missing values are needed to build imputation models, regardless of their measurement levels, and those statistics can be computed for all predictors within each Mapper or data source independently.

Second, to catch possible nonlinear relationship between a predictor with missing values and the target if both of them are continuous, our target-based method builds piecewise linear regression models on a list of bins of the predictor for each Mapper.

Third, only a single data pass is needed to build the imputation models for all predictors with missing values as they are built on each Mapper independently. Such an approach runs computations in a distributed manner and can handle large and distributed data sources efficiently.

Fourth, using an ensemble of imputation models instead of only one imputation model to impute missing values would give more robust results for both imputation strategies because outliers would only affect few imputation models and those models are unlikely to be selected into the final ensemble model.

3. Imputation model building

According to the measurement levels of the predictor (X) and the target (Y), four types of imputation models for the predictor could be built and Figure 2 illustrates the imputation model building process in detail.

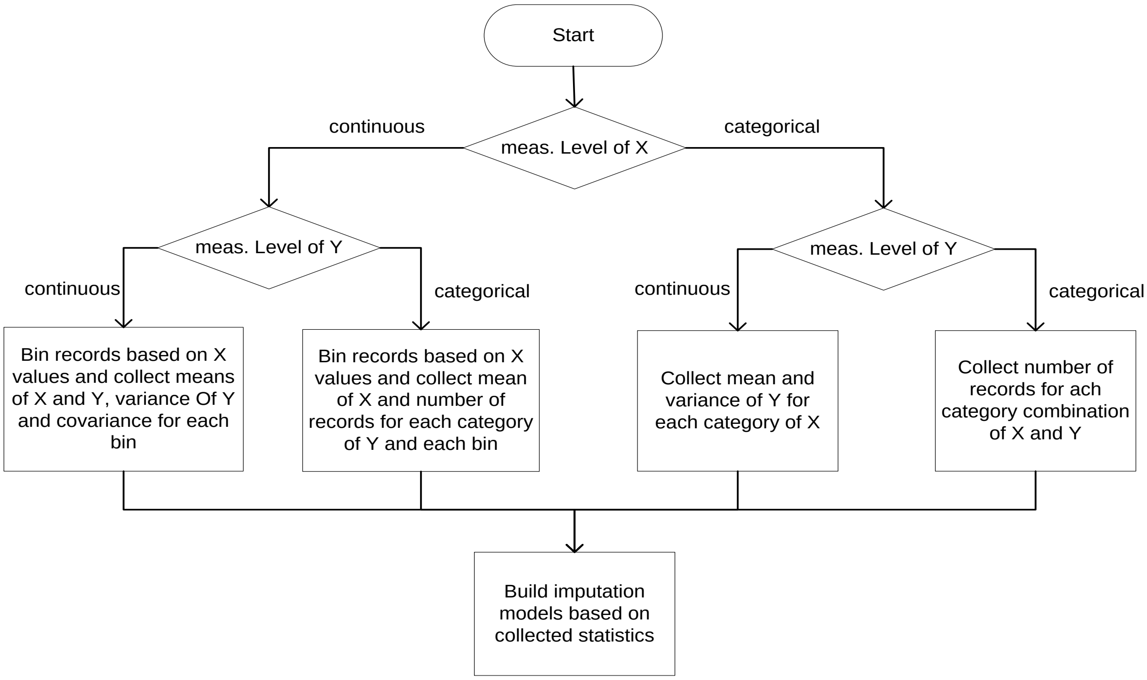
Start
meas. Level of X
continuous
categorical
meas. Level of Y
continuous
categorical
meas. Level of Y
continuous
categorical
Bin records based on X values and collect means of X and Y, variance Of Y and covariance for each bin
Bin records based on X values and collect mean of X and number of records for each category of Y and each bin
Collect mean and variance of Y for each category of X
Collect number of records for ach category combination of X and Y
Build imputation models based on collected statistics

Fig 2. Imputation model building process

3.1. Continuous predictor with continuous target

If the predictor and target are all continuous, then the piecewise linear regression models of the predictor on the target will be built as imputation models because they might capture a possible non-linear relationship between the predictor and target. Using Figure 3 as an example, suppose the true relation between X and Y is a parabola (the green curve) but a linear regression (the blue dotted line) is built to impute a missing value x_k then the imputed value based on the target value of y_k is $x_k^{''}$. On the other hand, if two piecewise linear regressions (the red dotted lines) are built then the imputed value is $x_k^{'}$ which is more accurate than $x_k^{''}$.This is another feature which can distinguish our method from other methods in the literature.

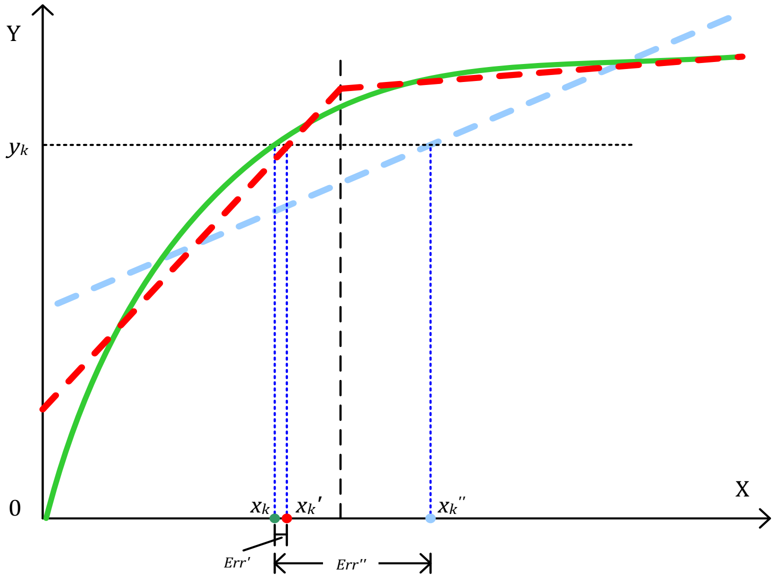

Fig 3. Comparison of missing value imputation between one linear regression and two piecewise linear regressions

To build piecewise linear regression models as imputation models, the predictor values as well as the corresponding target values are assigned into a list of bins based on the locations of the predictor. The approach is to sort predictor values and assign them to different bins.

Suppose that there are m bins $(a_1, a_2], (a_2, a_3], \ldots, (a_m, a_{m+1}]$, where $a_1 = -\infty$ and $a_{m+1} = \infty$, after collecting and merging bins then the following basic statistics in the i^{th} bin are collected:

- The number of records: $n^{(i)}$,
- Mean of Y: $\overline{Y}^{(i)}$,
- Mean of X: $\overline{X}^{(i)}$,
- Variance of Y: $S_{YY}^{(i)}$,
- Covariance of X and Y: $S_{XY}^{(i)}$.

Then the m piecewise linear regression models are built as follows:

$$X = \beta_0^{(i)} + \beta_1^{(i)} Y$$

where $\beta_1^{(i)} = S_{XY}^{(i)} / S_{YY}^{(i)}$ and $\beta_0^{(i)} = \overline{X}^{(i)} - \beta_1^{(i)} \overline{Y}^{(i)}, \; i = 1, \ldots, m.$

If the k^{th} record is a missing value in prediction X with a known target value, y_k, then piecewise linear regression models would be used to impute the missing value and an appropriate bin has to be selected first. If there is only one bin in which the condition

$\beta_0^{(i)} + \beta_1^{(i)} y_k \in (a_i, a_{i+1}]$ holds, then the linear regression model in the i^{th} bin will be used. However there may exist more than one bin that satisfies the above condition, say $\beta_0^{(i)} + \beta_1^{(i)} y_k \in (a_i, a_{i+1}]$ and $\beta_0^{(j)} + \beta_1^{(j)} y_k \in (a_j, a_{j+1}]$. Using Figure 4 as an example, a known target value y_k is applied to m models, then two scores $x_k^{'}$ in the i^{th} bin and $x_k^{''}$ in the j^{th} bin could be used to impute missing value of *X*. Under this circumstance, a random number from the binomial distribution, *B*(1, *p*), where $p = n^{(i)}/(n^{(i)} + n^{(j)})$, will be generated to determine which score would be used. If the random number is 1, then the missing value will be imputed by $x_k^{'}$ in the i^{th} bin, otherwise $x_k^{''}$ in the j^{th} bin is used. This method can be easily generalized to the situation of more than two bins by using a multinomial distribution.

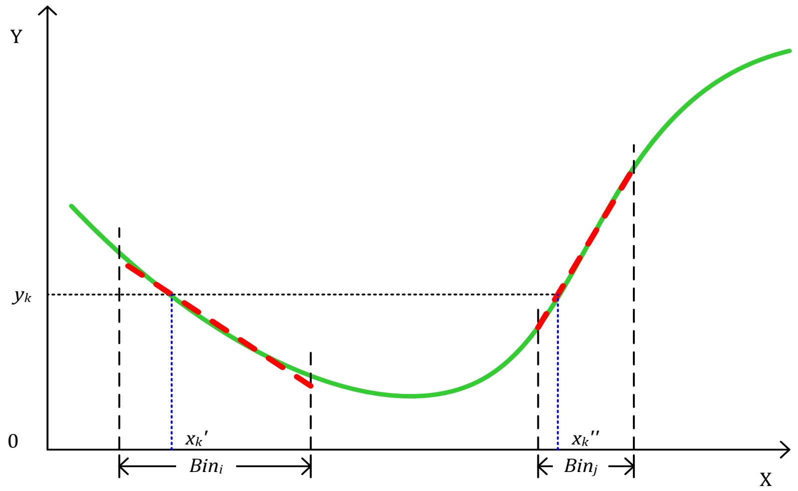

Fig 4. An example of how two values can be used to replace a missing value

3.2. Continuous predictor with categorical target

If the predictor is continuous but the target is categorical, then similar to Type I (continuous predictor with continuous target) the predictor values as well as the corresponding target categories are assigned into a list of bins based on the locations of the predictor.

Suppose that there are m bins $(a_1, a_2], (a_2, a_3], \ldots, (a_m, a_{m+1}]$, where $a_1 = -\infty$ and $a_{m+1} = \infty$, after collecting and merging bins then the following basic statistics of X for the j^{th} category of Y, $j = 1, \ldots, J$, in the i^{th} bin are collected:

- The number of records: $n^{(i,j)}$,
- Mean of X : $\overline{X}^{(i,j)}$.

These statistics are displayed in the following Table 1.

Table 1. Basic statistics in each combination of predictor bin and target category

Y *X*	1	2	…	*J*

$(a_1, a_2]$	$n^{(1,1)}$, $\overline{X}^{(1,1)}$	$n^{(1,2)}$, $\overline{X}^{(1,2)}$	$\cdots$	$n^{(1,J)}$, $\overline{X}^{(1,J)}$
$(a_2, a_3]$	$n^{(2,1)}$, $\overline{X}^{(2,1)}$	$n^{(2,2)}$, $\overline{X}^{(2,2)}$	$\cdots$	$n^{(2,J)}$, $\overline{X}^{(2,J)}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$(a_m, a_{m+1}]$	$n^{(m,1)}$, $\overline{X}^{(m,1)}$	$n^{(m,2)}$, $\overline{X}^{(m,2)}$	$\cdots$	$n^{(m,J)}$, $\overline{X}^{(m,J)}$

The first and the last S bins would be discarded in the following mean computation, where S is a specified constant that does not vary with the size of the data. Then for each target category, the rest of the bins are merged and the corresponding mean of the predictor could be obtained as follows:

$$\overline{X}_{Y=j} = \frac{1}{\sum_{i=S+1}^{m-S} n^{(i,j)}} \sum_{i=S+1}^{m-S} n^{(i,j)} \overline{X}^{(i,j)}.$$

The conditional mean computed in this way would be more robust because the S smallest values and S largest values as potential outliers are excluded. We call such means "robust conditional means"

If the k^{th} record is a missing value in prediction X with a known target category, y_k, then the missing value will be imputed by the robust conditional mean of X conditional on $Y = y_k$.

3.3. Categorical predictor with continuous target

For a categorical predictor with continuous target, the first two moments of a target in each category of the predictor are collected to represent the target's distribution in the corresponding category. Suppose a categorical predictor X has $1, \ldots, J$ categories, then the following basic statistics of Y for the j^{th} category of X are collected:

- Mean of Y: $\overline{Y}^{(j)}$
- Variance of Y: $s_{YY}^{(j)}$

If the k^{th} record is a missing value in predictor X with a known target value, y_k, then the missing value will be imputed with a predictor category by judging which distribution the target value y_k is more likely to belong to, that is the missing value will be imputed as follows:

$$x_k = \arg\min_j \left\{ \left| \frac{y_k - \overline{Y}^{(j)}}{\sqrt{s_{YY}^{(j)}}} \right|, \ j = 1, \ldots, J \right\}.$$

3.4. Categorical predictor with categorical target

If both the predictor and target are categorical, then the contingency table is generated, that is the number of records in each category combination of predictor and target is collected. If the k^{th} record is a missing value in predictor X with a known target category, y_k, then the missing value will be imputed by the mode of X conditional on $Y = y_k$.

4. Some experiments

In this section, we present results of simulation study to assess the performance of our TB method.

In the simulation, the binary target Y, which takes two values 0 and 1, follows a logistic regression model:

$$\log\left(\frac{\mathrm{p}(Y=0)}{\mathrm{p}(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 \mathrm{X}_3 + 0.2 \times \mathrm{I}(X_9 = 0) + 0.27 \times I(X_9 = 1) + 0.5 \times I(X_{10} = 0) + 0.3 \times I(X_{10} = 2)$$

where predictors $X_i, i = 1, \cdots, 3$, follow the standard normal distribution, N(0,1), independently, and X_9 and X_{10} follow the binomial distribution B(3,0.5), independently. The parameters $\beta_i, i = 0, \cdots, 3,$ follow a normal distribution N(0, 0.01), and the notation $\mathrm{p}(\cdot)$ and $I(\cdot)$ denote a probability function and an indicator function, respectively.

To compare the performance of our TB method to multiple imputation (MI) method, we add 5 variables $X_i, i = 4, \cdots, 8$ which also follow the standard normal distribution independently but they are not involved in the data generation process of the target Y.

The data is generated with 10 million records, where 7,000,000 records are used for training and 3,000,000 records for testing. For the training dataset, three types of datasets with missing values are designed:

Missing Completely At Random (MCAR): Each of variables X_1, X_2, X_3, X_9 and X_{10} has 10% missing values randomly.

Missing At Random 1 (MAR1) Each of variables X_1, X_2, X_3, X_9 and X_{10} has 20% missing values randomly when $Y = 0$, and has 10% missing values randomly when $Y = 1$.

Missing At Random 2 (MAR2) Each of variables X_1 and X_9 has 20% missing values randomly when $X_4 < -0.8$, and each of variables X_2 and X_{10} has 20% missing values randomly when $X_5 > 0.8$.

For each dataset with missing, the TB method and MI method are used to impute missing values, respectively. To be comparable, just one completed dataset is obtained instead of multiple completed datasets when MI method is used. The environment of experiments is Windows XP with 2 2.53GHz CPU and 3GB of RAM, and the software used in this section is IBM SPSS Statistics 21, so the experiments are done on one single PC which is equivalent to one Mapper and there is no need to extract the validation sample to select the final ensemble model.

We use performance of the missing value imputation process and goodness of fit of the subsequent model building as two criteria to do comparison in both TB and MI methods.

For performance criterion, the elapsed time values of missing value imputation (in minutes) are displayed in Table 2. It is obvious that the TB method is much faster than MI method. Please note that the elapsed time of the TB method should be less than the results in the table 2 if we use multiple Mappers.

Table 2. The elapsed time (in minutes)

Missing Type Method	**MCAR**	**MAR1**	**MAR2**
TB	1	1	1
MI	240	192	189

For goodness of fit criterion, the logistic regression models are built based on two imputed train datasets by the TB and MI methods, then the overall classification accuracy values based on two imputed training datasets as well as the testing data are displayed in Table 3. We can see that the TB method is a little better than MI method on the imputed training datasets, but is a little worse than MI method on the testing dataset. Comparing with MI method, our TB method trades a bit accuracy (for testing data) for huge time saving for the big data.

Table 3. Classification accuracy of logistic regression models

Method	**Data**	**MCAR**	**MAR1**	**MAR2**
TB	Imputed training data	0.810938	0.823813	0.800245
	Testing data	0.792619	0.787838	0.795341
MI	Imputed training data	0.795456	0.795472	0.795600
	Testing data	0.795971	0.795947	0.796121

5. Conclusion

The proposed target-based method for missing value imputation is to impute missing values in predictors for the subsequent predictive models on large and distributed data source using Map-Reduce. Different types of imputation models are built based on whether the predictor and target are categorical or continuous. All these imputing models are built just based on some bivariate statistics between the target and predictor which has missing values. Some experiments show that the method is extremely fast with reasonable accuracy for different types of missing values.

References

Bhaskar, T. and Sundararajan, R. G., “Method for building predictive models with incomplete data,” US 2009/0177598 A1.

Howell, D.C. (2008). The analysis of missing data. In Outhwaite, W. & Turner, S. (eds.) *Handbook of Social Science Methodology*. London: Sage.

IBM Inc. (2012), “Multiple Imputation Algorithms,” in *IBM SPSS Statistics 21 Algorithms*, Chicago, IL, 597–602.