Comparing Accuracy of Imputation Methods for Incomplete Categorical Data

Hyung Won Shin[1] & So Young Sohn[2]

Abstract

Various kinds of estimation methods have been developed for imputation of categorical missing data. They include modal category method, logistic regression, and association rule. In this study, we propose two imputation methods (neural network fusion and voting fusion) that combine the results of individual imputation methods. A Monte-Carlo simulation is used to compare the performance of these methods. Five factors used to simulate the missing data are (1) true model for the data, (2) data size, (3) noise size (4) percentage of missing data, and (5) missing pattern. Overall, neural network fusion performed the best while voting fusion is better than the individual imputation methods, although it was inferior to the neural network fusion. Result of an additional real data analysis confirms the simulation result.

Keywords: Categorical Missing Data, Neural Network, Voting, Fusion

1. Introduction

The simplest way of handling missing data in statistical analysis is to ignore missing observations. It is well known that this procedure, the so-called complete case analysis, typically leads to inconsistent estimators and misleading statistical tests (Little and Rubin, 1987). Moreover, by discarding data, we throw away information, which is also unsatisfactory. One of alternatives is to try to determine these values by imputation. When done carefully, imputation leads to consistent estimators and valid tests. Many imputation

[1] Dept. of Computer Science & Industrial Systems Engineering, Yonsei University, Graduate student.

[2] Dept. of Computer Science & Industrial Systems Engineering, Yonsei University, Professor.

sohns@yonsei.ac.kr

This work was supported by grant No. R04-2002-000-20003-0 from Korea Science and Engineering Foundation.

methods have been developed. However, studies of imputation methods for incomplete categorical data were mostly based on empirical data. These results may not be applied to general purposes. Also, in the simulation based studies for continuous data, they generate the missing data using the simplest type of missing data pattern, MCAR (missing completely at random). Moreover, previous studies used only individual imputation methods not taking advantage of combined methods.

In this paper, we generate incomplete categorical data from MAR (missing at random) mechanism, and A Monte-Carlo simulation is used to compare the performance of these methods.

Organization of this paper is as follows. In section 2, we review the literature on three imputation methods and propose two imputation methods based on fusion. In section 3, we introduce the experimental design for Monte Carlo simulation. Simulation results are summarized in section 4. Finally in section 5, we discuss the implication of our results and suggest further study areas.

2. Imputation method of incomplete categorical data

In this section, we briefly introduce three single imputation methods chosen for our study such as modal category method, logistic regression, and association rule. Details regarding the methods used for imputation are as follows. Let x_i ($x_{i1},..,x_{ip},\cdots,x_{iP}$) be the ith observation of categorical non-missing variables x, and y_i be the corresponding categorical variable y with a total number of C levels. We assume that missing occurs in one variable, that is, x variables are fully observed whereas only N_{obs} of the y values are observed ($N_{obs}<N$).

2.1 Modal category method

This method can be described as follows: out of a total of N observations, if missing is occurred at y_i, then look for the observations that have the same value as the x_i variables among the rest N-1 observations, and then if the number of these observations is k, impute using the modal category of y variable among these k observations (Lee and Kim, 1997).

2.2 Logistic regression

Assume that the categories of the dependent variable y are coded $1,\cdots,c,\cdots,C$. In the C category model we have C-1 logit functions: one for each y=j versus y=1 ($j=2,\cdots c\cdots,C$). For the application of logistic regression to imputation, let us assume that missing occurs in dependent variable y, while independent variables are complete. The missing value is

imputed with the category of y associated with the highest probability.

2.3 Association rule

Out of a total of N observations, if missing is occurred at the y_i variable, then look for the observations that have the same values with one of the x_i variables (say x_{ip}), and then calculate the support value between x_{ip} and y_i. This process is continued for all combination of x variables that have the same values as x_i. Lastly we impute missing as category of y_i with the highest support value (Lee.1999; Ng and Lee,1998). If a tie happens in support value, the rule contains more variables is selected.

2.4 Voting fusion

We use a majority voting based on the several results of individual imputation methods.

2.5 Neural network fusion

In order to utilize a neural network for imputation methods, let us assume that missing variable is a dependent variable y, while independent variables are complete. Then divide incomplete data into evaluation and test data sets. Evaluation data set consists of complete data, and test data set consists of missing data. Neural networks fusion consists of two phases: learning and operation. In the first phase, evaluation data set is partitioned into training and validation data set. The y values of the validation data set are assumed to be missing. Then three imputation methods are applied to impute that missing values. The predicted results of individual methods are used as the input for the neural network, while actual y is used as the output of the neural network. That is, the relationship between the prediction results of individual methods and actual y is learned. For constructing the accurate neural networks, cross-validation technique should be use. In the second phase, using test data set, the predicted results of individual methods are used as the input for the trained neural networks which is built in phase 1, the results of neural networks is considered as final results.

3. Design of experiment

In this section, we introduce the factors and levels used for Monte Carlo simulation. Five factors used in the simulation are (1) true model for the data, (2) data size, (3) noise size, (4) percentage of missing data, and (5) missing pattern. Details regarding the levels used for each factor are as follows.

3.1 True model for the data

We use three categorical input variables, each having four categories uniformly

distributed, and a binary output variable y. Each categorical input variable x_{ip} is transformed to several binary dummy variables. Samples of data were simulated under the logistic activation function between y and two types of combination of x.

That is, the probability that observation i belongs to the second category (j=2) of y based on the function k is as follows:

$$P_k(\mathrm{y}=2\mid d_i)=\frac{e^{[f_{2,k}(d_i)+\xi_i]}}{1+e^{[f_{2,k}(d_i)+\xi_i]}},\qquad (\mathrm{k}=1,2) \tag{1}$$

where $f_{2,1}(d_i)=0.5d_{i11}-0.5d_{i12}-0.23d_{i13}+0.28d_{i14}+0.1d_{i21}-0.72d_{i22}-0.24d_{i23}$
$+0.57d_{i24}-0.31d_{i31}-0.4d_{i32}+0.52d_{i33}-0.44d_{i34}$;

$$f_{2,2}(d_i)=\sin(d_{i11}-d_{i12})^2-0.5\sqrt{d_{i13}+3d_{i14}}\cos\left(\frac{d_{i21}}{d_{i22}+1}\right)-(d_{i33}+0.5d_{i34})\sin(2d_{i23}+2d_{i24}+d_{i31}+d_{i32})$$

; and random noise ξ_i is generated from a normal distribution with mean 0 and variance is set to adjust the noise level. Using P_i of (3), the output variable y_i is generated with two categories from the following binary distribution:

$$y_i=\begin{cases}2 & P_k(y=2\mid d_i)\\ 1 & 1-P_k(y=2\mid d_i)\end{cases}$$

3.2 Noise size

One might expect more difficulty in successfully imputing observation from data with a higher level of random noise, but it is not clear how large this effect is relative to the other factors under investigation. In this study, we used two levels of noise (ξ) using normal distribution which are Normal$(0,0.2^2)$ and Normal$(0,0.8^2)$.

3.3 Data size

We considered two levels of relative data size to be 5 and 100 times the number of parameters involved in input variables, respectively: 60 and 1200

3.4 Percentage of missing data

We use two levels of missing rate: 5% and 50%. That is, the probability that observation i is missing (M=2) in y variable based on the function s is as follows:

$$\mathrm{Q_s(M=2\mid d_i)}=\frac{e^{[g_{2,s}(d_i)+\varepsilon_i]}}{1+e^{[g_{2,s}(d_i)+\varepsilon_i]}},\qquad (s=1,2) \tag{2}$$

where random noise ε_i is generated from random normal distribution with mean 0 and variance is set to adjust the missing rate. According to the coefficient used in function $g_{2,s}(d_i)$, if the expected value in $[g_{2,s}(d_i)+\varepsilon_i]$ is close to 0, expected value of Q_s would be 0.5. Then the missing rate for M_i would be 50%.

3.5 Missing pattern

In order to understand the effect of the missing pattern on imputation method, we add two levels of noise ε_i to $g_{2,s}(d_i)$ in (4). Using Q_s in (4), we create M_i. If M_i is 2, then we set y variable in the ith observation as missing, and if M_i is 1 then leave y as it is.

3.6 Imputation method

The three individual methods (the modal category method, logistic regression, association rule) are compared to the two proposed fusion based imputation methods (neural network fusion and majority voting).

For performance evaluation, we replicate our 5×2^5 factorial design five times. Therefore we have a total of 800 experimental runs. The response is measured by imputation accuracy.

4. Results of Monte Carlo simulation

We use ANOVA (analysis of variance) to find significant factors at 5% level (see Table 1). In Table 1, we only present interaction effects used in our hypotheses.

Table 1. ANOVA for imputation accuracy

Source of variation	DF	SS	MS	F-value	P-value
F1×F6	4	2126.59	531.64	155.11	0.0001
F2×F3×F4×F5×F6	4	385.50	96.37	28.12	0.0001

F1: True model F2: Data size F3: Noise size F4: Percentage of missing F5: Type of missing F6: Imputation method

Duncan test is conducted at 5% significance level for multiple comparison of imputation methods according to data characteristics. As a result, when true model is linear, neural network fusion presented the highest performance while there were no significant differences among the rest of the four methods. When data size is small and percentage of missing is high, logistic regression showed low accuracy compared the other four methods. There is no significant difference between modal category method and association rule. However, if the number of categories in variable x and y are much larger, we expect these two methods have different imputation accuracy. Also neural network fusion and logistic regression presented higher performance than both modal category method and association rule when the missing pattern is strong. This result implies that imputing accuracy of modal category method and association rule decreases when there are insufficient non-missing data with the same value as x variables associated with the missing y.

Neural network fusion has the best performance when data size is large with large noise, low percentage of missing, and strong missing pattern. This can be explained that neural network fusion obtains combining effects when there are sufficient amount of data for neural network training and there are difficult conditions for imputation with large noise and strong missing pattern. Also, one can see that the performance of voting fusion decreases when the two of the three methods perform badly.

5. Conclusion

Results of Monte Carlo simulation indicated the following at 5 % significance level. We can conclude that neural network fusion has excellent performance in imputation of incomplete categorical data, while voting fusion is not as good as the neural network fusion but it shows a stable performance. Also, depending upon the data characteristics in missing occurrence, we could find out the suitable method among modal category model, logistic regression and association rule. In addition, we can see that modal category method and association rule have similarity performances.

We have a plan that comparison multiple imputation method with combination methods for classification, such as bagging and boosting, have been an active research area recently. We believe that combination methods for imputation could improve certain individual imputation methods. It is left for further study areas.

References

Lee, S.W. and Kim Y.J., 1997. Statistical Techniques for Treatment of Nonresponses in Public Health Categorical Data. J. of Korean Society of Health Statistics, 22(1) 114-132.

Little, J.A. and Rubin, D.B., 1987. Statistical Analysis with Missing Data. John Willey & Sons, New York.

Ng, V. and Lee, J., 1998. Quantitative association rules over incomplete data. IEEE International Conference on Systems, Man, and Cybernetics, 3 2821-2826.

Ragel, A., and Cremilleux, B., 1999. MVC-a preprocessing method to deal with missing values. Knowledge-Based Systems, 12(5) 285-291.

Rubin, D.B., 1987. Multiple imputation for nonresponse in surveys. Wiley, New York.

Schenker, N., Taylor, J., 1996. Partially parametric techniques for multiple imputation. Computational Statistics & Data Analysis, 22 425-446.