Missing Data Imputation Methods in Classification Contexts

Juheng Zhang

Department of Operations and Information Systems, University of Massachusetts Lowell, Lowell, MA

***Abstract*—** *We examine different imputation methods that deal with missing data in classification contexts and compare the performance of the methods with an experiment study. We investigate the performance of the methods under the assumption that data are missing at random. We find that, as the number of missing holes in data increases, the imputation methods deteriorate and the misclassification rates of the imputation methods increase. We also examine the scenario where missing data are due to strategic behaviors of data providers. We find that imputation methods play an important role at deterring strategic behaviors of data providers and minimizing the misclassification rate.*

Keywords*— *missing data, imputation method, classification.

I. INTRODUCTION

Often in many empirical studies, data are missing due to various reasons. Missing data may be caused by negligence of data collectors, poor experiment designs or procedures, or even purposely hiding behaviors of data providers. The two general assumptions of missing data are: data missing at random and data missing strategically. Randomly missing data assumption assumes that the missing data of an attribute are not related to the values themselves nor the values of other attributes. For instance, in U.S. census data, a specific home address is missing, which is likely due to a random reason. As for strategically missing data assumption, the data are missing due to strategic reasons. For instance, an insurance applicant can purposely hide her/his smoking/drinking when apply for a health insurance in hope for a more likely result of approval. Another example is limited information disclosure in financial markets [5, 6]. Certain companies strategically hide information from investors. Missing data are a common problem in many research fields such as economics, marketing, health, statistics, psychology, and education.

Missing data can lead to a number of problems [8]. The high level of statistical power requires a large amount of data. When data are missing, sample size decreases dramatically if only observations with complete data are used. Empirical studies found that if two percent of data are missing randomly in a data set, then eighteen percent of the total data can be lost when observations having a missing value are removed. Missing data decreases statistical power.

In this study, we consider different imputation methods that either designed for randomly missing data or strategically missing data. We compare the performance of the imputation methods in classification contexts under the assumption of data missing at random. We also examine the imputation methods when data providers act strategically and data are hidden intentionally. In the following section, we overview related research works and briefly discuss different imputation methods.

II. LITERATURE REVIEW

In the statistics field, a few imputation methods such as the Average Method, the Similarity, and the Regression Method have gained widespread acceptance. These methods normally consider attributes having continuous values. They have become conventional methods for dealing with missing data, and we see adaption of these methods in different fields. We refer readers to survey papers [4, 9] for detailed discussion on these conventional methods, although we discuss some of these below. There are also some imputation techniques unique to classification problems. Several papers [2, 3, 7] summarize different linear discriminate methods for handling missing data and compare the performance of these methods. The simplest imputation method perhaps is the Average Method, also known as marginal mean imputation. The method was first mentioned in the study [11]. For each missing value on a given variable, the Average Method finds the estimated value by calculating the mean for those cases with observed data present on that variable. The Similarity Method finds an observation that is the most similar to the record with a missing value as measured by the values not missing, and uses the actual value in the most similar record to replace the missing value. Proponents of the Similarity Method argue that the method improves accuracy since it uses realistic values, and that the method also preserves the shape of the distribution. The underlying principal of the Similarity Method can be used for discrete variables, and this variant of the Similarity Method is called Hot-deck imputation, which has become popular in survey research. A data set is "hot" if it is currently being used for imputing a score. The Hot-deck imputation replaces a missing value with actual scores of a similar case. If there are several equally similar cases, then the method randomly chooses one of them. The Regression Method, also called the conditional mean

imputation method, is another statistical imputation method. The Regression Method uses a regression equation to calculate an estimate of the true value. Assume only one variable has missing values on some cases. Using the cases with complete data on other variables, the method regresses on all of the other variables and then uses the regression equation to generate the substitutes for the cases with missing data. The substitutes are predicted values for missing data. According to the study [1], the Regression Method generates predicted values that preserve deviations from the mean and the shape of the distribution. It does not attenuate correlations between variables as much as mean substitution.

Another category of imputation methods [12-14] assumes data are missing strategically by data providers who try to game the decision makers' decision rules. The imputation methods proposed in the studies [12, 13] include the D and DNeg methods. These methods were designed for classification problems. The decision maker may use the D or DNeg method to impute missing values and minimize misclassification rates when facing with strategic data providers. The DNeg method was to thwart negative data providers from gaining a positive classification when they intentionally hide information. The D method considered not only negative data providers but also positive ones. Using the D method, the decision maker can deter negative data provider's gaming behaviors and also incent positive data provider to reveal information. The D method is more conservative than the DNeg method in a sense that it considers both positive and negative data providers while the DNeg method is online for negative ones.

III. EXPERIMENT DESIGN

We compare eight methods, Average, Regression, Similarity, D, DNeg, AvgNeg, RegNeg, and SimNeg. The Average, Regression, Similarity, D, and DNeg are as what we discussed in the above section. The AvgNeg, RegNeg, and SimNeg are the revised version of the original Average, Regression, and Similarity method respectively, in which only negative training samples are used for imputing missing values. We first start with the case of randomly missing data, and then examine the strategically missing data. The parameters of experimental design are listed in Table 1.

TABLE 1.
SUMMARY OF EXPERIMENTAL DESIGNS

Treatment	Parameter
Replications	30
Dimensionality	3, 4, 5, 6, 7
Training set size	20, 100, 200, 1000
Testing set size	20, 100, 200, 1000
Randomly missing data percent	1%, 2%, 3%, 4%, 5%
Data providers' methods	D, Average, Regression, Similarity, DNeg, AvgNeg, RegNeg, SimNeg
Decision Maker's methods	D, Average, Regression, Similarity, DNeg, AvgNeg, RegNeg, SimNeg
Outcomes	TotMisc, PosMisc, NegMisc, Notclassified, TotStrMisc, PosStrMisc, NegStrMisc, StrPos, StrNeg

We use 30 replications for each case. The number of attributes ranges from 3 to 7. We use different training and testing set sizes, 20, 100, 200, and 1000. In the randomly missing data case, we consider various percent of missing holes, 1%, 2%, 3%, 4%, and 5%. In the strategically missing data case, data providers may use one of the eight methods to hide information: D, Average, Regression, Similarity, DNeg, AvgNeg, RegNeg, SimNeg. The decision maker chooses one of the eight methods to impute missing values. We use different misclassification measurements. TotMisc is the misclassification over all data providers, PosMisc is the misclassification rate over positive records, NegMisc is the misclassification rate over negative records, Notclassified is the rate that records not get classified. To stabilize the variance of the rates of misclassification in statistical tests [10], we map the performance measures to 2 × arc sin (sqrt (misclassification rate)).

IV. EMPIRICAL RESULTS

We first conduct an ANOVA analysis on the misclassification rate (TotMisc) for the randomly missing data case. We see that all experiment factors are significant, as well as all interaction effect of all factors at 0.0001 confidence level. The ANOVA analysis results are included in Table 2.

TABLE 2.
TWO-WAY ANALYSIS FOR DEPENDENT VARIABLE: MISCLASSIFICATION IN RANDOM CASE

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F	R-Sqe	Coeff Var	Root MSE	Rate Mean
Model	214	3798.52	17.75	1013.14	<.0001	0.53	44.46	0.13	0.30
Error	191785	3360.07	0.02						
Corrected Total	191999	7158.59							
Source	DF	Anova SS	Mean Square	F Value	Pr>F				
Main effect									
n	4	98.43	24.61	1404.59	<.0001				
ℓ	3	4.71	1.57	89.63	<.0001				
ℓ'	3	187.58	62.53	3568.79	<.0001				
Ram	4	1546.32	386.58	22065.10	<.0001				
Mp	7	1691.19	241.60	13789.90	<.0001				
Two-way Interaction effect									
$n*\ell$	12	2.93	0.24	13.94	<.0001				
$n*\ell'$	12	3.15	0.26	14.97	<.0001				
$n*Ram$	16	13.56	0.85	48.37	<.0001				
$n*Mp$	28	91.22	3.26	185.95	<.0001				
$\ell*\ell'$	9	2.13	0.24	13.49	<.0001				
$\ell*Ram$	12	1.43	0.12	6.78	<.0001				
$\ell*Mp$	21	47.44	2.26	128.94	<.0001				
$\ell'*Ram$	12	5.45	0.45	25.92	<.0001				
$\ell'*Mp$	21	5.57	0.27	15.13	<.0001				
$Ram*Mp$	28	96.68	3.45	197.08	<.0001				

Table 2 shows that the percent of randomly missing data in datasets is a significant factor of misclassification rate. Next, we study the impact of percent of randomly missing data on performance measurements in details. The results of the misclassification rates for different percents of randomly missing data are provided in Table 3.

TABLE 3.
STATISTICS OF RANDOMLY MISSING HOLES

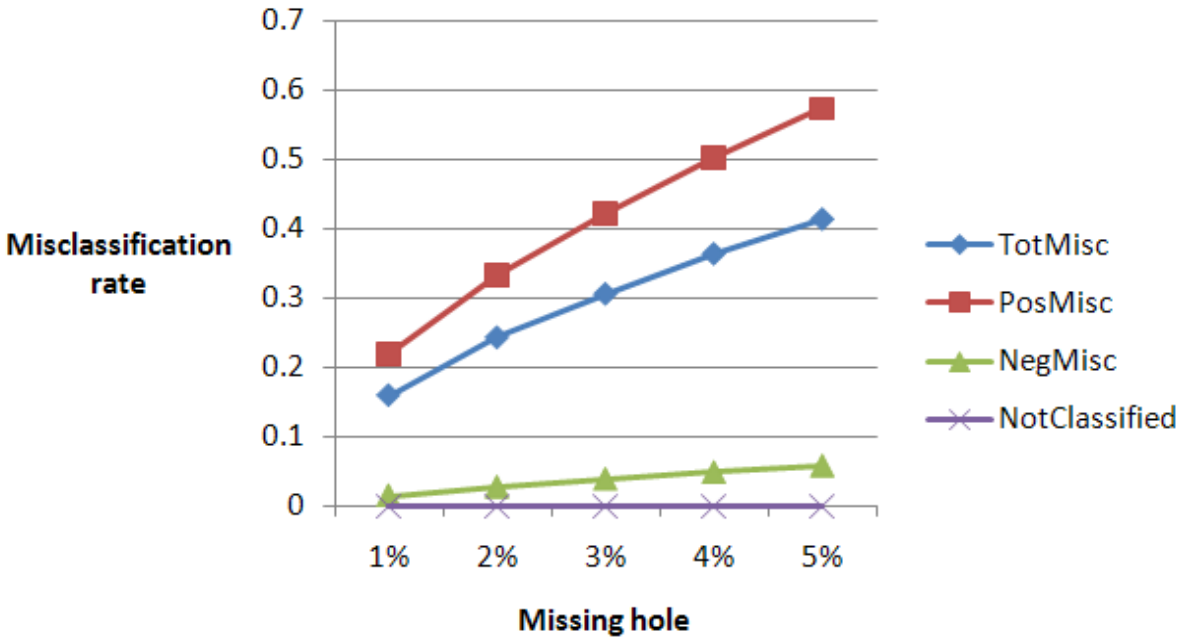
Missing Percent	TotMisc	PosMisc	NegMisc	NotClassified
1%	0.15957	0.21862	0.01514	0.00000
2%	0.24331	0.33372	0.02756	0.00006
3%	0.30635	0.42139	0.03891	0.00009
4%	0.36456	0.50299	0.04877	0.00026
5%	0.41469	0.57368	0.05807	0.00033

As shown in Table 3, the misclassification rate over positive records is higher than that over negative records. In addition, all of four misclassification rates, TotMisc, PosMisc, NegMisc, and NotClassified, increase as the percent increases. The results in percentage format are provided in Table 4. In percent format, the misclassification rate is 0.64% when 1% data are missing and increases to 4.24% when the percent of missing holes increases to 5%.

TABLE 4.
STATISTICS OF RANDOMLY MISSING HOLES IN PERCENT

Missing Percent	TotMisc	PosMisc	NegMisc	NotClassified
1%	0.640%	1.190%	0.010%	0.00%
2%	1.470%	2.760%	0.020%	0.00%
3%	2.330%	4.370%	0.040%	0.00%
4%	3.290%	6.190%	0.060%	0.00%
5%	4.240%	8.000%	0.080%	0.00%

We plot the trend of misclassification rate with the increase in the missing percent in Fig 1. The top line is for the misclassification rate over positive records, and the TotMisc is the average over positive and negative misclassification rates. The non-classified records stay as zero when the percent of missing holes increases.



In the random case, data are missing randomly, that is the information is not hidden strategically. A principal still can choose one of eight methods to impute missing information. Next, we consider the case where data are missing strategically. We simulate the case where agents select a method in determining which attributes to hide and simultaneously a decision maker chooses from those eight methods to impute estimates for missing data of all agents. Similarly, we conduct an ANOVA test to examine the effect of factors and interaction effect on the misclassification rates and provide the results in Table 5. The hiding strategies of data providers is denoted as Ma . As shown in Table 5, the hiding strategies of data providers are significant , and all other experiment factors are significant at the 0.0001 level.

TABLE 5.
TWO-WAY ANALYSIS FOR DEPENDENT VARIABLE: MISCLASSIFICATION RATE IN THE STRATEGICALLY MISSING CASE

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F	R-Sq	Coeff Var	Root MSE	Rate Mean
Model	271	103371.3	381.4	15298.5	<.0001	0.931	10.579	0.158	1.493
Error	306928	7652.747	0.0249						
Corrected Total	307199	111024.1							
Main effect									
n	4	78.763	19.691	789.74	<.0001				
ℓ	3	66.17	22.057	884.62	<.0001				
ℓ'	3	4.944	1.648	66.1	<.0001				
Ma	7	2751.8	393.114	15766.6	<.0001				
Mp	7	42912.157	6130.308	245868	<.0001				
Two-way Interaction effect									
$n*\ell$	12	19.43	1.619	64.94	<.0001				
$n*\ell'$	12	1.015	0.085	3.39	<.0001				
$n*Ma$	28	430.728	15.383	616.97	<.0001				
$n*Mp$	28	1277.847	45.637	1830.37	<.0001				
$\ell*\ell'$	9	0.56	0.062	2.5	0.0075				
$\ell*Ma$	21	68.872	3.28	131.54	<.0001				
$\ell*Mp$	21	156.013	7.429	297.96	<.0001				
$\ell'*Ma$	21	1.672	0.08	3.19	<.0001				
$\ell'*Mp$	21	13.682	0.652	26.13	<.0001				
$Ma*Mp$	49	25245.6	515.218	20663.8	<.0001				

We conduct Tukey's range tests for the decision maker's methods. The summary of these test results can be found in Tables 6. Table 6 shows that when the decision maker uses the D or DNeg method, NegMisc is the lowest and PosMisc is the highest. More specifically, we see that NegMisc is 0 for the D method and 0.013 for the DNeg method (or 0.004% in terms of actual rate before the mapping), which are lower than 0.402 (or 3.98%) for Similarity, 0.324 (or 2.6%) for RegNeg, 0.322 (or 2.58%) for Regression, 0.283 (or 1.99%) for Average, 0.06 (or 0.09%) for AvgNeg, and 0.031 (or 0.02%) for SimNeg. The rate of positive or negative agents who act strategically is the same regardless of a decision maker's method. Therefore, if a decision maker is more negative risk averse, her best strategy should be the D method.

TABLE 6.
THE COMPARISON RESULTS OF TUKEY'S RANGE TEST

TotMisc			PosMisc			NegMisc			NotClassified		
Group	**Mean**	**Methods**	**Group**	**Mean**	**Methods**	**Group**	**Mean**	**Methods**	**Group**	**Mean**	**Methods**
A	2.205	7	A	2.192	0	A	0.402	3	A	1.902	2
B	2.086	3	B	1.893	4	B	0.324	6	A	1.902	3
B	2.065	6	C	1.629	5	B	0.322	2	A	1.902	6
B	2.062	2	D	1.283	1	C	0.283	1	A	1.902	7
C	1.128	0	E	0.783	7	D	0.06	5	B	0	0
D	0.992	4	F	0.185	3	E	0.031	7	B	0	1
E	0.895	5	F	0.167	6	FE	0.013	4	B	0	4
F	0.822	1	F	0.158	2	F	0	0	B	0	5

V. CONCLUSION

We compare eight different imputation methods in the case where data are missing at random and in the case where data are missing strategically. We find that as the percent of missing data increases, the performance of all the eight imputation methods decreases. When data are missing strategically, the D method or DNeg method gives the lowest misclassification rate.

REFERENCES

[1] Allison, P.D. Missing data. Woburn, MA, USA: Sage Publications Inc., 2001.

[2] Chan, L.S., and Dunn, O.J. The treatment of missing values in discriminant analysis-1. The sampling experiment. Journal of the American Statistical Association, 67, 338 (1972), 473-477.

[3] Chan, L.S., Gilman, J.A., and Dunn, O.J. Alternative approaches to missing values in discriminant analysis. Journal of the American Statistical Association, 71, 356 (1976), 842-844.

[4] Donders, A.R.T., van der Heijden, G., Stijnen, T., and Moons, K.G.M. Review: A gentle introduction to imputation of missing values. Journal of Clinical Epidemiology, 59, 10 (2006), 1087-1091.

[5] Healy, P.M., and Palepu, K.G. The challenges of investor communication the case of cuc international, inc. Journal of financial economics, 38, 2 (1995), 111-140.

[6] Hirshleifer, D., and Teoh, S.H. Limited attention, information disclosure, and financial reporting. Journal of accounting and economics, 36, 1-3 (2003), 337-386.

[7] Jackson, E.C. Missing values in linear multiple discriminant analysis. Biometrics, 24, 4 (1968), 835-844.

[8] Roth, P.L. Missing data: A conceptual review for applied psychologists. Personnel Psychology, 47, 3 (1994), 537-560.

[9] Schafer, J.L., and Graham, J.W. Missing data: Our view of the state of the art. Psychological Methods, 7, 2 (2002), 147-177.

[10] Stam, A., and Joachimsthaler, E.A. Solving the classification problem in discriminant analysis via linear and nonlinear programming methods. Decision Sciences, 20, 2 (1989), 285-293.

[11] Wilks, S.S. Moments and distributions of estimates of population parameters from fragmentary samples. The Annals of Mathematical Statistics, 3, 3 (1932), 163-195.

[12] Zhang, J. Linear discrimination with strategic missing values. Information Systems and Operations Management, Gainesville, FL, USA: University of Florida, 2011.

[13] Zhang, J., Aytug, H., and Koehler, G.J. Discriminant analysis with strategically manipulated data. Information Systems Research, 25, 3 (2014), 654-662.

[14] Zhang, J., Liu, X., and Li, X. Support vector regression for handling strategically hidden data. Lowell, U.o.M., 2015