Математички Билтен
41(LXVII) No. 1
2017(23-31)
Скопје, Македонија

ISSN 0351-336X (print)
ISSN 1857-9914 (online)
UDC: 519.237:519.257-77

IMPUTATION OF MISSING CATEGORICAL VALUES IN SURVEY RESEARCH DATA

VESKA NONCHEVA AND VENELIN VALKOV

Abstract. The imputation of missing data is often a crucial step in data analysis and prediction. This study discusses an approach for imputation of missing data in surveys with a large number of categorical variables. The proposed method is based on models of Correspondence Analysis and Data Envelopment Analysis.

1. Introduction

In a survey each respondent is asked several questions. A ubiquitous problem in the real-world datasets is missing data. The missing data could be unintentional and intentional.

When the amount of missing data is small, often a good data analysis can be performed assuming that the missing data are ignorable. As the fraction of missing information increases, the ignorability assumption becomes more critical.

In some cases, half or more of the information is missing, and then a recovery of missing data is essential.

The aim of this work is to find an approach for reducing the amount of the missing values.

1.1. Background.

1.1.1. *Correspondence Analysis (CA).* CA is a method of data analysis for representing tabular data graphically [8]. Particular emphasis is placed on how the distance between points is interpreted. In some cases, two variables are similar entities with comparable scales and interpoint distances that can be interpreted as a measure of difference, or dissimilarity, between the plotted points in a spatial map of the data. The points are weighted differently and the centroid tends to lie in a position closer to the points with higher weight. Distance in correspondence analysis is measured using the χ^2 distance and this distance is the key to our data imputation approach.

Date: October 8, 2017.
2000 *Mathematics Subject Classification.* 62H25, 62P25.
Key words and phrases. intentional and unintentional missing data.

Correspondence analysis gives a way for expressing the data in a pictorial form for ease of interpretation called CA table. This option will help us to present our data imputation approach.

The total inertia of a cross-tabulation is a measure of how much variation there is in the table. The inertia can be further broken down into row and column components along individual principal axes. The overall quality of a two-dimensional correspondence map can be measured by the amount of inertia accounted for by the first two principal axes. Correspondence analysis is performed with the objective of accounting for a maximum amount of inertia among the first axis. The second axis accounts for a maximum of remaining inertia. Any additional row of a data matrix can be positioned on the existing correspondence map. They are called supplementary points. The supplementary points have a position but no mass at all. Their contribution to the inertia is zero and they have no influence on the principal axes.

For ordinal or nominal data, we can use correspondence analysis to find groups of close levels. The close levels of the columns of X define a rule for imputation. Namely the levels of the passive points predict the level of the response. We can impute missing data once all groups of close levels have been found.

Once all groups of close levels have been found, the missing values can be imputed using the predicted values from the groups.

1.1.2. *Data Envelopment Analysis (DEA).* DEA is a popular non-parametric method used to measure efficiency. It uses linear programming to identify the most efficient units [5].

DEA makes it possible to identify efficient and inefficient units in a framework where results are considered in their particular context. The units to be assessed were originally called Decision-Making Units (DMUs). DEA is an extreme point method and compares each DMU with only the best DMUs.

DEA handles multiple input and multiple output models. Inputs and outputs can have very different units. The estimated efficiency for any DMUs depends on the number of inputs and outputs included in the model.

DMUs are directly compared against a peer or combination of peers. DEA can tell how well peers are doing compared to others peers but not compared to a theoretical maximum.

2. Related work

In the last couple of years, there have been major advancements in the domain of missing data imputation. Various techniques include amongst others: Bayesian methods [19, 13, 1], Nearest Neighbours, Mean and Mode, Random Forests [18], Latent Class Models [21], Multiple Correspondence Analysis [2] Hybrid approaches [17, 3, 9] and Neural Networks [10] more

recently. Different methods have variable performance based on the missing data mechanism and the structure of the data.

Most of the methods work with continuous data only. However, some techniques support both categorical [6] and mixed-type [18] data imputation.

Empirical evaluations of different approaches appear in [7, 12, 4, 16]. Despite the significant number of techniques, there is no clear winner when imputation is needed for surveys with categorical data only.

3. Problem description

We are given a data table with n rows and m columns containing only categorical data. Some cells of the table might contain missing data. Each cell can contain only a single value.

Our goal is to impute the missing values.

4. Model

Let $y = (y_1, y_2, ..., y_n)$, where $y_i = (y_{i1}, y_{i2}, ..., y_{ij})$, $i = 1, ..., n$, be the data matrix. Let some values y_{ij} be missing. Our aim is to recover all missing values.

Let $x = (x_1, x_2, ..., x_k)$ be a vector of variables that are fully observed for all units in this survey.

The general idea is the following: We are interested in the distribution of y and x provides information about it. If x is fully observed then a model for $p(y \mid x)$ can lead to more precise inference about missing values of y than would be obtained by modeling y alone.

The model of missing data is random variables with their distribution conditional on observed data.

5. Algorithm

Suppose the survey dataset is presented by data table $T*$ containing only categorical data. Some of the cells in one column of $T*$ have missing values.

Let $y*$ be the name of a response column for which missing data must be imputed. Let $X*$ be a set of fully observed variables (columns of T* without missing values). Then $T* = (y*, X*)$. Let y be the vector of observed values of the response y* and X be the corresponding sub-table of X*. Then $T = (y, X)$ is data table with n rows and m columns containing only observed categorical data.

Let x be a column of X. Now we use the cross-tabulation of the following two variables from the survey x and y. The result is a table W of counts for each combination of values of the nominal variables x and y. The cells of the two-way table called profiles indicate the frequency with which each combination occurred in the sample. We are interested in the relationship

between the levels of these variables. The χ^2 distance can measure the dissimilarity between profiles. We say that two levels are close if the distance between the corresponding profiles is a small number. This small number is denoted by λ.

5.1. **Input.**

- T - data table with n rows and m columns containing only categorical data. Some of the cells might contain missing data. Each possible level for a column must be present at least once.
- y - name of the response column for which missing data must be imputed
- λ - threshold for considering two levels of two category variables "close" to each other.

5.2. **Output.** A data table with n rows and m columns with all (or some) values imputed.

Our algorithm uses three different methods to obtain it's final predictions:

5.3. **Correspondence Analysis (CA).** Given a data table T and column y for which we must impute its values, a two-way table W is generated with y's levels as columns and all possible levels from columns contained in X (all columns which will be used as predictors) are represented as rows.

Correspondence analysis is then applied on W by specifying all columns in X (except the first one) as supplementary rows. The χ^2 distance d between each level of y and each level in X is computed using the resulting CA table. The obtained distances are stored in table D.

5.4. **Clustering.** The distances in D are divided into clusters using Partitioning Around Medoids (PAM) algorithm. The number of clusters k for PAM is chosen by maximizing a Silhouette score.

Mean distance for each cluster is calculated. Distance indicator is created for each distance $d \in D$ with value equal to 1 if the threshold $\lambda > d$ and 0 otherwise. New table CT is created using the columns of D, the cluster to which each row belongs, the indicator for each d and the sum of all indicators.

5.5. **Data Envelopment Analysis (DEA).** This method computes effectiveness when a set of input data X (distances from CA) and output data Y (sum of indicators) are provided. Furthermore, we require the Returns to Scale (RTS) to be variable and the model is oriented towards its input data.

We apply DEA to CT and create new table PT which contains all columns of CT and add a new column for effectiveness. This final table will be used for predicting values.

5.6. **Prediction.** For each row with missing value we:

- Create a set of prediction rules of all possible levels for y with fixed levels for X.
- If no rules have effectiveness equal to 1 the algorithm does not predict a value for this row.
- Otherwise we select the subset of rows with minimum mean cluster distance and we pick the rule with highest effectiveness. The value for y in this rule is the predicted value for this row.
- The process repeats until all rows are evaluated.

5.7. **Accuracy.** Once we have accomplished the data imputation step we should test the accuracy of the imputed data. A vector y' with true labels corresponding to y is required for the evaluation.

Our algorithm uses two simple metrics to evaluate its performance - a percentage of predicted values and percentage of correctly predicted values.

6. Example

The proposed algorithm has been implemented using the R programming language [15] along with the following example.

We are going to use an artificial survey dataset generated in order to demonstrate the algorithm.

6.1. **Survey dataset.** A data table T that contains 4 columns corresponding to 4 questions in a fictional survey is generated. Each column contains categorical variables with 5 levels.

T is generated by creating a 4×5 (4 questions each with 5 levels) matrix of which each row is then replicated 100 times. Thus, a total of 500 rows are obtained. The resulting table has the following unique rows:

Table 1. Unique rows in the artificial survey dataset

question 1	question 2	question 3	question 4
Never	One	Less than 5000	Child
Often	Two	5000 - 10000	Teen
Rarely	Three	10001 - 15000	Young
Very often	Four	15001 - 20000	Adult
Very rarely	Five	20001 - 25000	Elderly

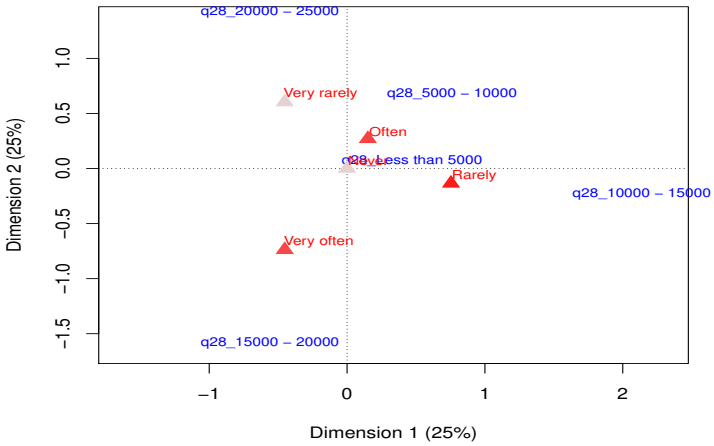

FIGURE 1. CA applied to the artificial survey dataset

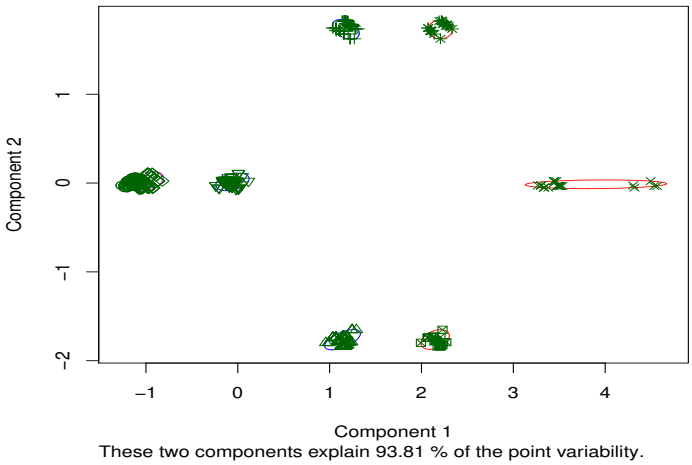

FIGURE 2. Clustering on the artificial survey dataset with k=8

6.2. **Results.** We provide data table T as an input to our algorithm along with the name of the first column as y (which values we want to predict) and the threshold $\lambda = 1.0$. We then set each value of y in T to *none*.

A plot of the CA result is shown on Figure 1. The result of the PAM algorithm is displayed on Figure 2.

Initially the levels of the response variable are uniquely predicted by the levels of the predictors. In this setting, the algorithm predicts 100% of the values with an accuracy of 100%.

We then add a new row to T with the following levels: Never, Two, Less than 5000, child. The results do not change - 100% of the values are predicted with an accuracy of 100%.

Then we introduce noise in data adding the following row 10 times: Often, Three, 10000 - 15000, adult. When we run the algorithm 100% of the values are predicted but our accuracy has dropped to about 97%. All newly added rows are incorrectly predicted.

Continuing with the previous example we set the threshold $\lambda = 0.49$ and run the algorithm again. This time, the results are switched - about 97% of the values are predicted and the accuracy is 100%. This demonstrates that we can manipulate the trade-off between a percentage of predicted values and the overall accuracy.

7. Conclusion

We presented a technique for data imputation which is useful to all who collect categorical data, for example data collected in social surveys. The method is based on Correspondence Analysis with supplementary points and Data Envelopment Analysis.

Correspondence analysis models are helpful in analyzing cross-tabular data in the form of numerical frequencies and result in elegant but simple graphical displays which permit more rapid interpretation and understanding of the data. Namely, the ability of correspondence analysis methods to communicate complex tables of numerical data through the medium of graphics is used by the authors to introduce their ideas for data imputation.

We used the freely available R statistical software, which has become the standard in statistical computing. We would like to thank the authors of the following R packages which greatly helped our work: *ca* [14], *rDEA* [20] and *cluster* [11].

8. Future work

It might be possible to further control the trade-off between the percentage of made predictions versus their accuracy by introducing effectiveness threshold parameter.

We are planning to provide our implementation of the algorithm as free and open-source R package.

Acknowledgements. This study was supported by the Plovdiv University "Paisii Hilendarski" - NPD, grant number NI15 FMI-004.

References

[1] A. Agresti and D. B. Hitchcock. Bayesian inference for categorical data analysis. *Statistical Methods and Applications*, 14(3):297–330, 2005.

[2] V. Audigier, F. Husson, and J. Josse. Mimca: Multiple imputation for categorical variables with multiple correspondence analysis. *Statistics and Computing*, pages 1–18, 2015.

[3] S. Azim and S. Aggarwal. Hybrid model for data imputation: using fuzzy c means and multi layer perceptron. In *Advance Computing Conference (IACC), 2014 IEEE International*, pages 1281–1285. IEEE, 2014.

[4] M. Celton, A. Malpertuy, G. Lelandais, and A. G. de Brevern. Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments. *BMC genomics*, 11:15, 2010.

[5] A. Charnes, W. W. Cooper, and E. Rhodes. Measuring the efficiency of decision making units. *European journal of operational research*, 2(6):429–444, 1978.

[6] S. J. Cranmer and J. Gill. We Have to Be Discrete About This: A Non-Parametric Imputation Technique for Missing Categorical Data. *British Journal of Political Science*, 43(May):1–25, 2012.

[7] P. J. García-Laencina, P. H. Abreu, M. H. Abreu, and N. Afonoso. Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. *Computers in Biology and Medicine*, 59:125–133, 2015.

[8] M. Greenacre. *Correspondence analysis in practice*. CRC press, 2007.

[9] X.-Y. Jing, F. Qi, F. Wu, and B. Xu. Missing data imputation based on low-rank recovery and semi-supervised regression for software effort estimation. In *Proceedings of the 38th International Conference on Software Engineering - ICSE '16*, pages 607–618. ACM, 2016.

[10] C. Leke, T. Marwala, and S. Paul. Proposition of a Theoretical Model for Missing Data Imputation using Deep Learning and Evolutionary Algorithms. *arXiv:1512.01362 [cs]*, pages 1–14, 2015.

[11] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2016.

[12] S. P. Mandel J. A Comparison of Six Methods for Missing Data Imputation. *Journal of Biometrics & Biostatistics*, 06(01):1–6, 2015.

[13] D. Manrique-Vallier and J. P. Reiter. Bayesian multiple imputation for large-scale categorical data with structural zeros. *Survey Methodology*, 40(1):125–134, 2014.

[14] O. Nenadic and M. Greenacre. Correspondence analysis in r, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software*, 20(3):1–13, 2007.

[15] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[16] M. M. Rahman and D. N. Davis. Machine learning-based missing value imputation method for clinical datasets. In *IAENG Transactions on Engineering Technologies*, pages 245–257. Springer, 2013.

[17] P. Rey-del castillo. Fuzzy min max neural networks for categorical data : application to missing data imputation. *Neural Computing and Applications*, 21(6):1349–1362, 2012.

[18] A. D. Shah, J. W. Bartlett, J. Carpenter, O. Nicholas, and H. Hemingway. Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. *American Journal of Epidemiology*, 179(6):764–774, 2014.

[19] Y. Si and J. P. Reiter. Nonparametric Bayesian Multiple Imputation for Incomplete Categorical Variables in Large-Scale Assessment Surveys. *Journal of Educational and Behavioral Statistics*, 38(5):499–521, 2013.
[20] J. Simm and G. Besstremyannaya. *rDEA: Robust Data Envelopment Analysis (DEA) for R*, 2016.
[21] D. Vidotto, J. K. Vermunt, and M. C. Kaptein. Multiple Imputation of Missing Categorical Data using Latent Class Models : State of the Art. *Psychological Test and Assessment Modeling*, 57(4):542–576, 2015.

Faculty of Mathematics and Informatics, Plovdiv University "Paisii Hilendarski" - Bulgaria.
E-mail address: wesnon@uni-plovdiv.bg

Faculty of Mathematics and Informatics, Plovdiv University "Paisii Hilendarski" - Bulgaria.
E-mail address: venelin@curiousily.com