Decision Trees for the Imputation of Categorical Data

Tobias Rockel, Dieter William Joenssen and Udo Bankhofer

Abstract Resolving the problem of missing data via imputation can theoretically be done by any prediction model. In the field of machine learning, a well known type of prediction model is a decision tree. However, the literature on how suitable a decision tree is for imputation is still scant to date. Therefore, the aim of this paper is to analyze the imputation quality of decision trees. Furthermore, we present a way to conduct a stochastic imputation using decision trees. We ran a simulation study to compare the deterministic and stochastic imputation approach using decision trees among each other and with other imputation methods. For this study, real datasets and various missing data settings are used. In addition, three different quality criteria are considered. The results of the study indicate that the choice of imputation method should be based on the intended analysis.

1 Introduction

Missing data, an occurrence in many areas of empirical research, is problematic because common analysis methods require complete datasets (Graham, 2009).

Tobias Rockel, Dieter William Joenssen, Udo Bankhofer
TU Ilmenau, Helmholtzplatz 3, 98693 Ilmenau, Germany
✉ {tobias.rockel,dieter.joenssen,udo.bankhofer}@tu-ilmenau.de

ARCHIVES OF DATA SCIENCE (ONLINE FIRST) DOI 10.5445/KSP/1000058749/14
KIT SCIENTIFIC PUBLISHING ISSN 2363-9881
Vol. 2, No. 1, 2017

One solution for this problem is to replace missing values with estimations – a process called imputation.

In principle, any prediction model can be used to estimate the missing values. Simple approaches for imputation substitute an appropriate location parameter (e.g. the mode or mean) or a randomly drawn value from the observed data (also known as a random hot deck imputation). However, more advanced prediction models, which use more available information, usually lead to better imputation results (Little and Rubin, 2002). One of these more advanced methods is the decision tree, first proposed for imputation by Kalton and Kasprzyk (1982). The usage of decision trees for prediction is already well researched in machine learning, but still has not been thoroughly examined for imputation. Unfortunately, it is not possible to infer decision tree performance for imputation from machine learning contexts. For machine learning prediction, accuracy is usually of interest, whereas for imputation subsequent parameter estimation is key. To this end, imputation is not the same as prediction (van Buuren, 2012), and the distribution of the resulting data plays a role. Further, it depends on the missing data mechanism which imputation method suits best for resolving missing data.

The concept of the missing data mechanism was first developed by Rubin (1976). This concept is based on treating the missing data indicator matrix $M = (m_{ij})_{n \times q}$, a matrix that shows whether a datum x_{ij} from the data matrix $X = (x_{ij})_{n \times q}$ is observed ($m_{ij} = 0$) or not observed ($m_{ij} = 1$), as a random variable. The concept defines three classes of dependencies between the data matrix X and the missing data indicator matrix M: missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR). If the data is MCAR, the distribution of M is independent of all data

$$f(M \mid X, \phi) = f(M \mid \phi) \text{ for all } X, \phi, \tag{1}$$

where ϕ is the parameter vector of the missing data mechanism. In this case, the observed data is a simple random sample of all data. If the distribution of M is related to the observed values X_{obs}, but not to the missing X_{mis}

$$f(M \mid X, \phi) = f(M \mid X_{obs}, \phi) \text{ for all } X_{mis}, \phi, \tag{2}$$

a MAR mechanism is present and the observed values are no longer a simple random sample. If the distribution of M also depends on values that are not observed, the mechanism is called NMAR (Little and Rubin, 2002). Thus, the prediction of missing data based solely on observed data is only meaningful in the first two cases.

Due to the caveats surrounding the application of decision trees in the context of missing data, this paper seeks to analyze the performance of classical decision trees in the context of missing data. To this end, the remainder is organized as follows. Section 2 illustrates the induction of decision trees and how decision trees can be used for imputation. In Sect. 3, the design of the simulation study used to determine the performance of decision trees for imputation is presented. The results of this study are given in Sect. 4 and implications from these results are discussed in Sect. 5.

2 Decision trees for imputation

The idea to use decision trees for missing data dates back to Kalton and Kasprzyk (1982). They propose to use a decision tree for imputation, but fail to provide further guidance on how. Later Kalton (1983) used decision trees to identify covariates subsequently used to construct imputation classes. Creel and Krotki (2006) went one step further. These authors used the nodes of a decision tree to define imputation classes and then applied different imputation methods within the classes.

In this paper, we will go beyond the existing approaches and use the decision tree directly for imputation. Decision trees will be inducted using complete observations (Quinlan, 1986b; Lobo and Numao, 1999; Twala, 2009). The resulting tree is then used to replace the missing values. Further, we limit this paper to the prediction of categorical data. Imputation of categorical data is equivalent to a classification in machine learning. However, the concepts developed may also be useful for imputing quantitative data.

2.1 Induction of decision trees

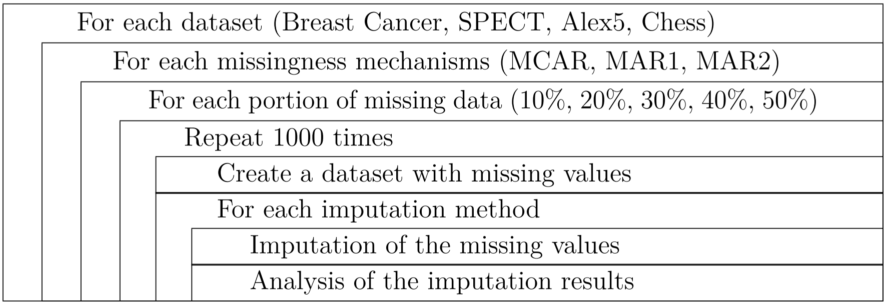
Popular algorithms for the induction of decision trees are C4.5 (Quinlan, 1993), CART (Breiman et al, 1993), or CHAID (Kass, 1980). These algorithms vary in details, but they are all based on the divide and conquer strategy. This means the algorithms divide a dataset recursively into subsets that are more and more homogeneous with respect to a criterion.

The mentioned algorithms use a measure of impurity as decision criterion to determine the optimal division of the data. For example, the C4.5, CART, and CHAID algorithms use the Information Gain, Gini-Index, and chi-squared-value as impurity measure. All these impurity measures increase with the number of unequal objects in the subset, which in this context means that they belong to different classes (Rokach and Maimon, 2008). The decision tree algorithm calculates the criterion for every possible split and chooses the one with the highest gain of purity. The algorithms stop splitting when a subset consists only of objects belonging to the same class or a different rule is satisfied. In the tree, a node represents a subset, a branch represents a split, and the final nodes are called leafs (Han et al, 2011).

2.2 Choosing the value for imputation

After the decision tree is induced, it can be used to impute missing data. For this, two basic possibilities exist: The majority rule and the probability rule. When applying the majority rule, the predicted class of an object corresponds to the most common class of the final subset. This leads to deterministic imputation and can be seen as a mode imputation by subset. In contrast to this, when applying the probability rule the predicted class of an object is set to a random class, with probabilities equal to the empirical class frequencies in the final subset. This leads to stochastic imputation and corresponds to a random hot deck imputation by subset. So far, only the majority rule has been considered to substitute missing values (e.g. Quinlan, 1986b; Lobo and Numao, 1999; Twala, 2009). Thus, we propose to use the probability rule for the prediction of missing values.

It is not clear, whether the majority or the probability rule will yield superior results. In the context of missing data it is argued that draws, not means, should be used as imputation values (Little and Rubin, 2002). Thus, a stochastic imputation method, such as a decision tree using the probability rule, should produce better imputation values. However, with respect to accuracy, Quinlan (1986a) stated that the majority rule is always better than the probability rule. But in turn, Quinlan’s argumentation assumes that a simple random sample of the data is used for decision tree induction, which in the context of missing data is equivalent to the presence of an MCAR mechanism. This assumption is very strong and the consequences for the effectiveness of an imputation method are unforeseeable, if this assumption is not met. Furthermore, as already stated



```
For each dataset (Breast Cancer, SPECT, Alex5, Chess)
  For each missingness mechanisms (MCAR, MAR1, MAR2)
    For each portion of missing data (10%, 20%, 30%, 40%, 50%)
      Repeat 1000 times
        Create a dataset with missing values
        For each imputation method
          Imputation of the missing values
          Analysis of the imputation results
```

Fig. 1 The design of the simulation study

Table 1 Datasets

	Few Attributes	Many Attributes
Few Objects	Breast Cancer (10 attributes, 277 objects)	SPECT (23 attributes, 267 objects)
Many Objects	Alex5 (11 attributes, 5000 objects)	Chess (37 attributes, 3196 objects)

in Sect. 1, higher accuracy does not imply better imputation. Thus, empirical evidence is needed to decide on the preferable strategy.

3 Study design

To compare the imputation quality of decision trees with other imputation methods, a simulation study was conducted using R (R Core Team, 2014). Figure 1 gives a schematic overview of the study.

Four real datasets with different characteristics were selected to determine the possible influences that the number of objects and attributes have on the imputation results. Properties of these datasets are summarized in Table 1. Breast Cancer, SPECT, and Chess are freely available from the UCI Machine Learning Repository (Lichman, 2013). The dataset Alex5 consists of the first 5000 objects from the training dataset used in a machine learning competition (Causality Workbench, 2014). Before applying the imputation mechanism, all datasets are reduced to complete cases.

We used one MCAR mechanism and one MAR mechanism at two different parameter levels to create missing values in one binary attribute. For the

Breast Cancer, SPECT, Alex5, and Chess this attribute was Class, Overall Diagnosis, attribute 11, and Classes, respectively. The MCAR mechanism was achieved by deleting values depending on independent Bernoulli trials with success probability equal to the proportion of missing values p_{mis}

$$f(M \mid p_{mis}) = \prod_{i=1}^{n} \left((p_{mis})^{m_{ik}} (1 - p_{mis})^{1-m_{ik}} \prod_{j \neq k} 1^{1-m_{ij}} 0^{m_{ij}} \right) \tag{3}$$

where k is the index of the attribute with missing data, n is the number of objects and m_{ij} is the missing data indicator as defined in Sect. 1.

The MAR mechanism is achieved by deleting different proportions of values, dependent on the value of a second, binary attribute. Let l be the index of this second attribute, $x_{il} \in \{0,1\}$ and 1 be the mode value of this second attribute. Then the MAR mechanism is defined as

$$\begin{aligned} f(M \mid X_l, p_{MAR}, c) = \prod_{i=1}^{n} \Big[& \left((p_{MAR})^{m_{ik}} (1 - p_{MAR})^{1-m_{ik}} \right)^{1-x_{il}} \times \\ & \left((c\, p_{MAR})^{m_{ik}} (1 - c\, p_{MAR})^{1-m_{ik}} \right)^{x_{il}} \times \\ & \prod_{j \neq k} 1^{1-m_{ij}} 0^{m_{ij}} \Big] \end{aligned} \tag{4}$$

where p_{MAR} was chosen so that the expected number of missing values equals np_{mis}. We used this MAR mechanism at two levels of the parameter c. For MAR1 we set $c = 2$, so that the proportion of missing values was two times higher, when the object had the mode value of the second attribute. To increase the effect we set $c = 4$ for MAR2. The second attribute used in generating the missing values is attribute 5 (Breast Cancer), 22 (SPECT), 1 (Alex5) and 33 (Chess). All attributes are moderately related (corrected contingency coefficient about 0.4) to the attribute in which the missing values are generated. The contingency coefficient $C = \left(\chi^2\right)^{\frac{1}{2}} \left(\chi^2 + n\right)^{-\frac{1}{2}}$ is based upon Pearson's χ^2 between two nominal variables. The corrected contingency coefficient is then defined as $C_{corr} = CM^{\frac{1}{2}} \left(M - 1\right)^{-\frac{1}{2}}$, where M is the minimum of the number of rows and columns of the contingency table on which χ^2 is based. Further, the proportion of missing values was varied from 10% to 50% in steps of 10 percentage points.

As imputation methods, we applied the decision tree algorithms CART (using rpart version 4.1-8, Therneau et al 2014) and C4.5 (using RWeka version 0.4-23, Witten et al 2011; Hornik et al 2009), both with the majority and probability rule. C4.5 was run, using the recommended setting from Quinlan (1993).

rpart was run with no pre-pruning except for minsplit=2 and 0-SE-Pruning. For further comparison, we also applied a random hot deck, a mode imputation and a nearest neighbor hot deck (using HotDeckImputation (Joenssen, 2013)).

To single out the superior imputation methods, we used three quality criteria. The first criterion measures the distortion of the micro structure – the single objects – of the dataset. This imputation accuracy is defined as the number of times the imputed value is equal to the deleted value, divided by the total number of imputed values.

The second criterion evaluates the effect on the marginal distribution of the attribute with missing values $X_k = (x_{1k}, \dots, x_{nk})^T$. For this purpose, we calculated the absolute bias between the estimate $\hat{p}(X_k^{orig})$ based on the originally observed values X_k^{orig} and the estimate $\hat{p}(X_k^{imp})$ based on the imputed variable X_k^{imp}

$$\left| \hat{p}(X_k^{imp}) - \hat{p}(X_k^{orig}) \right|, \tag{5}$$

where p is the mode value probability (before deletion) of the attribute with missing values X_k.

The last criterion is the root mean square error (RMSE) of the corrected contingency coefficients. It measures the distortion of the relationship between the attribute with missing values X_k and the other attributes X_j, $j \neq k$, in the dataset. It is defined as

$$\Delta C_{corr} = \sqrt{\frac{1}{q-1} \sum_{j \neq k} \left(C_{corr}\left(X_j, X_k^{orig}\right) - C_{corr}\left(X_j, X_k^{imp}\right) \right)^2} \tag{6}$$

where q is the number of attributes in the dataset and $C_{corr}(X_k, X_k^{orig})$ and $C_{corr}(X_j, X_k^{imp})$ are the corrected contingency coefficients between the attributes X_j, $j \neq k$, and X_k^{orig}, and X_j and X_k^{imp}, respectively. We averaged the results of each criterion over 1000 simulation runs.

4 Results

This section presents the results of the simulation study, separated by the three quality criteria. To investigate whether the induction of a decision tree is worthwhile, we stress the difference between the random hot deck and the probability rule, as well as the difference between the mode imputation and the majority rule. Furthermore, we emphasize differences between the stochastic and the deterministic imputation methods. Finally, since the differences between the C4.5 and CART algorithms, for both majority and probability rule, are negligible, only the results of the C4.5 algorithm are discussed.

4.1 Accuracy

Figure 2 shows the results for the accuracy criterion. Here, higher accuracy values indicate better imputation results. The graphs show that a decision tree using the majority rule always leads to one of the most accurate imputations, whereas the random hot deck is normally the least accurate method.

Furthermore, Fig. 2 shows that mode imputation achieves relatively high accuracy, except for the Chess dataset. The nearest neighbor hot deck is usually midway between the most accurate and least accurate method. It is observable that the decision tree methods improve with more objects as well as more attributes. An increase in both leads to more accurate imputations. This is especially obvious for the probability rule, which leads to poor results in the small Breast Cancer dataset, but is very competitive in the big Chess dataset.

In addition, the influence of the missing data mechanism on the results is negligible. The sole exception is mode imputation in the Chess dataset. Here, the combination of a high percentage of missing values and a MAR mechanism leads to a decline in accuracy.

As expected from the literature, deterministic imputation methods generally lead to a higher accuracy than the corresponding stochastic imputation methods in case of MCAR. Also, the deterministic methods are often more accurate when the missing data mechanism is MAR. But, as the results for the Chess dataset show, this need not be the case. Further, when comparing random hot deck to the probability rule or mode imputation to the majority rule, the results indicate that the decision tree methods are usually at least as good as the corresponding simple imputation method.

4.2 Estimation of p

Figure 3 gives the results for the absolute bias of the estimator $\hat{p}$. For this quality criterion, lower values indicate better results. The results are unambiguous as to

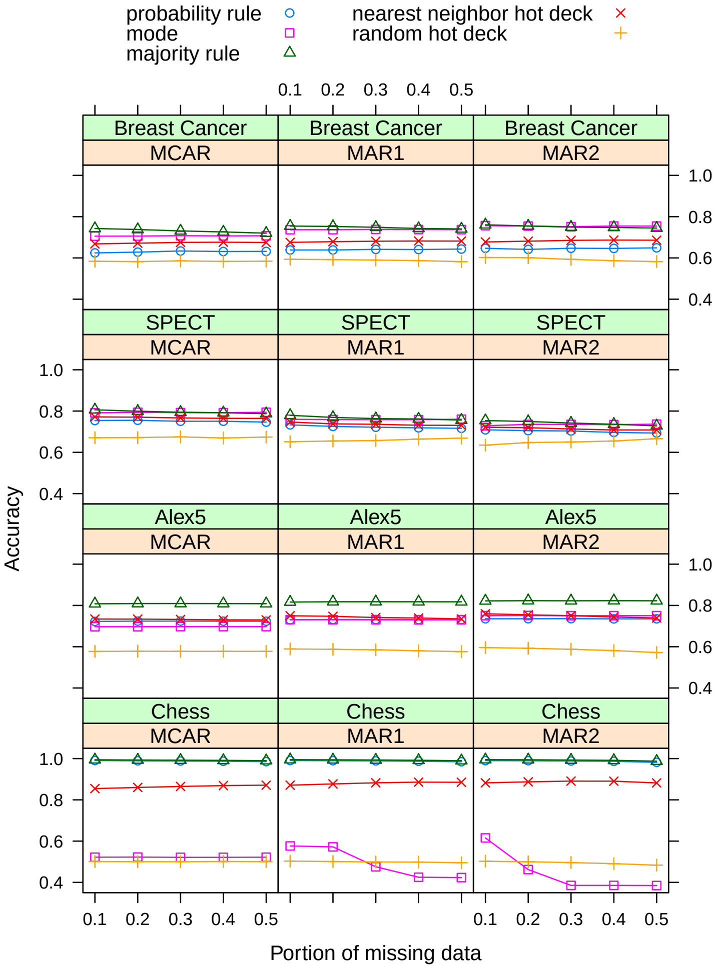


Fig. 2 Accuracy of the imputation methods

which method performs worst. For all simulated cases, mode imputation leads to the highest absolute bias. In contrast, the probability rule and the random hot deck perform best when the missing data mechanism is MCAR. When the mechanism is MAR, results for the random hot deck are dependent on the portion of missing data, whereas the probability rule is nearly unaffected and remains the method of choice.

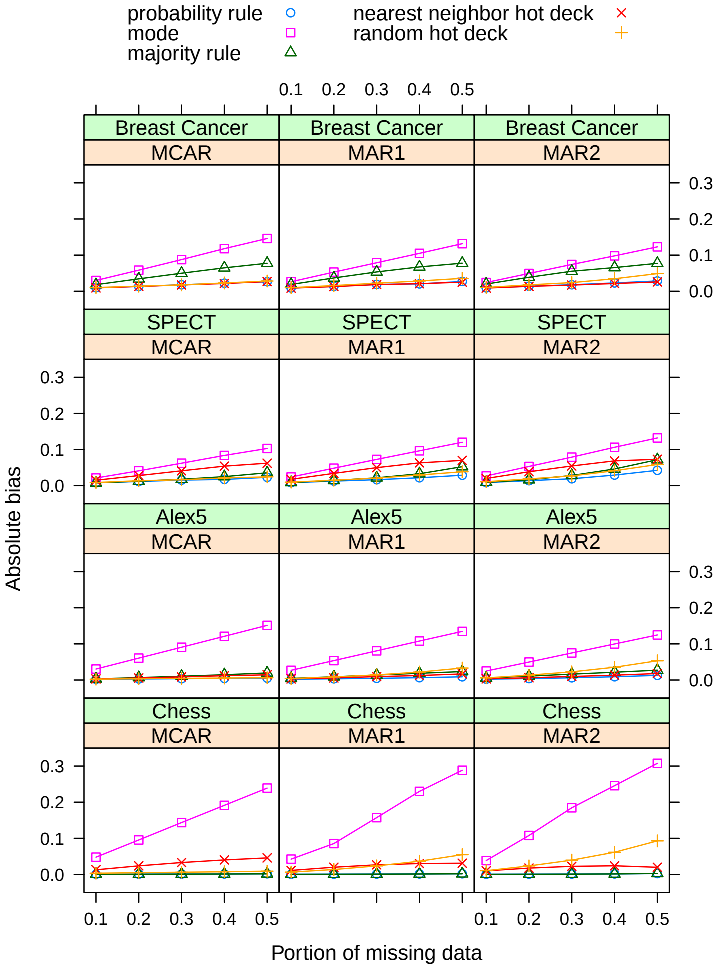


Fig. 3 Absolute bias of *p* for the different imputation methods

In accordance with previous results the nearest neighbor hot deck ranks midfield. Further, confirming previous results, the decision tree with majority rule performs poorly in the small Breast Cancer dataset, but improves with both more objects and more attributes. In contrast, we observe that stochastic methods generally outperform deterministic methods for this criterion. However, the general order between simple and more complex imputation methods still holds. The majority rule is better than mode imputation, and the probability rule is still at least as good as the random hot deck.

4.3 RMSE of contingency coefficients

Figure 4 shows the results for the last criterion, the RMSE of contingency coefficients. As for the last criterion, a smaller RMSE indicates better imputation results. Hence, we can observe that the probability rule is typically the method of choice, whereas the random hot deck usually leads to the worst results.

Contrary to the previous results, the nearest neighbor hot deck is one of the best methods in the datasets with a low number of covariates, namely Breast Cancer and Alex5. For the other two datasets however, it ranks in the middle again and the decision tree imputation methods offer superior performance. This shows once more, that decision trees benefit more substantially from an increase in available covariates than the nearest neighbor hot deck. However, the nearest neighbor hot deck seems to be a strong contender when the preservation of the relationships in small datasets is of interest.

In contrast to the previous criteria, there is no consistent order in the stochastic and deterministic imputation methods. On the one hand, the probability rule as a stochastic method is typically better than the deterministic majority rule. On the other hand, the random hot deck is worse than mode imputation. Nonetheless, the induction of a decision tree is worthwhile. Hence, the probability rule is always better than the random hot deck, and the majority rule usually leads to better results than mode imputation.

5 Conclusion

In this paper, we investigated decision trees for imputation and introduced the probability rule to achieve a stochastic imputation based on decision trees. The results of the simulation study indicate that decision tree imputation is generally at least as good as the corresponding simple imputation method, independent of the chosen quality criterion. Thus, the results of using a decision tree with the majority rule for imputation are at least as good as the results when using mode imputation. The same applies to decision trees using the probability rule and random hot deck. Furthermore, the effort required to induce a decision tree

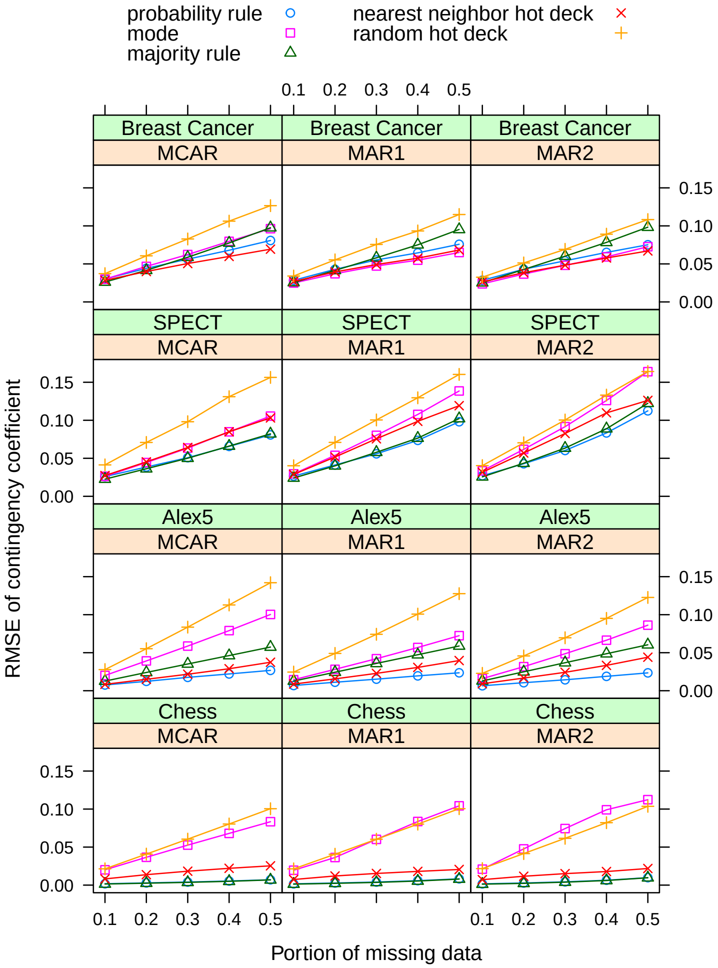


Fig. 4 RMSE of contingency coefficients for the different imputation methods

is especially worthwhile when many attributes or objects are available in the dataset.

In addition, the results indicate that there is no best imputation method. A decision tree with the majority rule typically leads to the best accuracy, whereas a decision tree with the class probability rule is a better choice with respect to the absolute bias in the estimation of p and regarding the RMSE of the contingency coefficients. In short, the majority rule preserves the micro structure (the single objects) of the dataset better, whereas the probability rule is more appropriate for the macro structure (for the single attribute and the dependencies between attributes) of the dataset. The nearest-neighbor hot deck appears to offer a compromise between preserving the micro and macro structure of the dataset. In summary, the results indicate that, on the one hand, a deterministic imputation is preferable when the following analysis relies strongly on the values of single objects. On the other hand, a stochastic imputation seems to be more appropriate when the subsequent analysis involves statistical parameter estimations. Therefore, the choice between deterministic and stochastic imputation methods should not be made without the intended analysis in mind.

References

Breiman L, Friedman JH, Olshen RA, Stone CJ (1993) Classification and Regression Trees. Chapman & Hall, New York

Causality Workbench (2014) Active learning challenge. URL http://www.causality.inf.ethz.ch/activelearning.php?page=datasets

Creel DV, Krotki K (2006) Creating imputation classes using classification tree methodology. In: American Statistical Association (ed) Proceedings of the Section on Survey Research Methods, pp 2884–2887

Graham JW (2009) Missing data analysis: Making it work in the real world. Annual Review of Psychology 60:549–576, DOI 10.1146/annurev.psych.58.110405.085530

Han J, Kamber M, Pei J (2011) Data Mining: Concepts and Techniques, 3rd edn. The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann, Waltham

Hornik K, Buchta C, Zeileis A (2009) Open-source machine learning: R meets weka. Computational Statistics 24(2):225–232, DOI 10.1007/s00180-008-0119-7

Joenssen DW (2013) Hotdeckimputation: Hot deck imputation methods for missing data: R package version 0.1.0. URL http://CRAN.R-project.org/package=HotDeckImputation

Kalton G (1983) Compensating for missing survey data. Tech. rep., The University of Michigan, Ann Arbor, Michigan

Kalton G, Kasprzyk D (1982) Imputing for missing survey responses. In: American Statistical Association (ed) Proceedings of the Survey Research Methods Section, pp 22–31

Kass GV (1980) An exploratory technique for investigating large quantities of categorical data. Journal of the Royal Statistical Society Series C (Applied Statistics) 29(2):119–127, DOI 10.2307/2986296

Lichman M (2013) Uci machine learning repository. URL `http://archive.ics.uci.edu/ml`

Little RJA, Rubin DB (2002) Statistical Analysis with Missing Data, 2nd edn. Wiley series in probability and statistics, Wiley, Hoboken

Lobo OO, Numao M (1999) Ordered estimation of missing values. In: Zhong N, Zhou L (eds) Methodologies for Knowledge Discovery and Data Mining, Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, vol 1574, pp 499–503, DOI 10.1007/3-540-48912-6_67

Quinlan JR (1986a) The effect of noise on concept learning. In: Michalski RS, Carbonell JG, Mitchell TM (eds) Machine Learning: Volume II, Morgan Kaufmann, Los Altos, pp 149–166

Quinlan JR (1986b) Induction of decision trees. Machine Learning 1(1):81–106, DOI 10.1007/BF00116251

Quinlan JR (1993) C4.5: Programs for Machine Learning. The Morgan Kaufmann series in machine learning, Morgan Kaufmann, San Mateo

R Core Team (2014) R: A language and environment for statistical computing. URL `http://www.R-project.org/`

Rokach L, Maimon O (2008) Data Mining with Decision Trees: Theory and Applications. Series in machine perception and artificial intelligence, World Scientific, Singapore

Rubin DB (1976) Inference and missing data. Biometrika 63(3):581–592, DOI 10.1093/biomet/63.3.581

Therneau T, Atkinson B, Ripley BD (2014) rpart: Recursive partitioning and regression trees: R package version 4.1-8. URL `http://CRAN.R-project.org/package=rpart`

Twala B (2009) An empirical comparison of techniques for handling incomplete data using decision trees. Applied Artificial Intelligence 23(5):373–405, DOI 10.1080/08839510902872223

van Buuren S (2012) Flexible Imputation of Missing Data. Interdisciplinary Statistics Series, Chapman and Hall/CRC, Boca Raton

Witten IH, Frank E, Hall MA (2011) Data Mining: Practical Machine Learning Tools and Techniques, 3rd edn. Morgan Kaufmann series in data management systems, Morgan Kaufmann, Burlington