International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 15, 2017 ISSN: 2348-6600
Reference ID: IJCS-229 PAGE NO: 1422-1428
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

A Survey – Methods of Missing Data Imputation

Priya.S,Research scholar,
Research and Development Center,
Bharathiar University,Tamil Nadu,India

ABSTRACT- Missing values in attributes. Several schemes have been studied to overcome the drawbacks produced by missing values in data mining tasks; one of the most well known is based on preprocessing, formerly known as imputation This paper reviews methods for handling missing data in a research study.

1. Introduction

Missing data is a problem because nearly all standard statistical methods presume complete information for all the variables included in the analysis. A relatively few absent observations on some variables can dramatically shrink the sample size. As a result, the precision of confidence intervals is harmed, statistical power weakens and the parameter estimates may be biased. Appropriately dealing with missing can be challenging as it requires a careful examination of the data to identify the type and pattern of missingness, and also a clear understanding of how the different imputation methods work. Sooner or later all researchers carrying out empirical research will have to decide how to treat missing data. In a survey, respondents may be unwilling to reveal some private information, a question may be inapplicable or the study participant simply may have forgotten to answer it. Accordingly, the purpose of this report is to clearly present the essential concepts and methods necessary to successfully deal with missing data.

2. Missing data mechanisms

There are different assumptions about missing data mechanisms:

a) Missing completely at random (MCAR): Suppose variable Y has some missing values. We will say that these values are MCAR if the probability of missing data on Y is unrelated to the value of Y itself or to the values of any other variable in the data set. However, it does allow for the possibility that "missingness" on Y is related to the "missingness" on some other variable X. (Briggs et al., 2003) (Allison, 2001)

b) Missing at random (MAR)-a weaker assumption than MCAR-:

The probability of missing data on Y is unrelated to the value of Y after controlling for other variables in the analysis (say X). Formally: P(Y missing|Y,X) = P(Y missing|X) (Allison, 2001). c) Not missing at random (NMAR): Missing values do depend on unobserved values.

3. Literature Review

We can find more recent analysis and proposals of imputation methods, which considers an increasing number of techniques compared:

• Hruschka et al. [28] propose two imputation methods based on Bayesian networks. They

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 15, 2017 ISSN: 2348-6600
Reference ID: IJCS-229 PAGE NO: 1422-1428
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

compare them with 4 classical imputation methods: EM, Data Augmentation, C4.5, and the CMC method, using 4 nominal data sets from the UCI repository [3] with natural MVs (but inducing MVs in them as well). In their analysis, they employ 4 classifiers as follows: one-rule, Naïve-Bayes, C4.5, and PART. As performance measures, the authors measure the prediction value (i.e., the similarity of the imputed value to the original removed one) and the classification accuracy obtained with the four mentioned models. From the results, the authors state that better prediction results do not imply better classification results.

• Farhangfar et al. [18] take as the objective of their paper to develop a unified framework supporting a host of imputation methods. Their study inserts some imputation methods into their framework (Naïve-Bayes and Hot Deck) and compares this with other basic methods: mean, Linear Discriminant Analysis, Logreg, etc. All their experimentation is based on discrete data, so they use the "accuracy" of imputed values against randomly generated MVs. The relation of this imputation accuracy to classification accuracy is not studied. In Farhangfar et al. [19], the previous study is extended using discrete data, com- paring with more classical imputation methods. This study uses a representative method of several classifiers' types as follows: decision trees, instance-based learning, rule-based classifier, probabilistic methods, and SVMs by means of boosting [51]. The MVs are produced artificially in a wide-ranging amount for each of the data sets, and the results obtained from the classification of imputed data are compared with the ones with MVs. This study shows that the impact of the imputation varies among different classifiers and that imputation is beneficial for most amounts of MVs above 5% and that the amount of improvement does not depend on the amount of MVs. The performed experimental study also shows that there is no universally best imputation method.

• Song et al. [56] study the relationship between the use of the KNNI method and theC4.5 performance (counting with its proper MV technique) over 6 data sets of software projects. They emphasize the different MVs' mechanisms (MCAR, MAR, and NMAR) and the amount of MVs introduced. From their analysis, they found results that agree with Batista and Monard [6]: KNNI can improve the C4.5 accuracy. They ran a Mann– Whitney statistical test to obtain significant differences in this statement. They also show that the missingness mechanism and pattern affect the classifier and imputation method performance.

• Twala [58] empirically analyzes 7 different procedures to treat artificial MVs for decisiontrees over 21 real data sets. From the study, it can be concluded that listwise deletion is the worst choice, while the multiple imputation strategy performs better than the rest of the imputation methods (particularly those with high amounts of MVs), although there is no outstanding procedure.

• García-Laencina et al. [23] evaluate the influence of imputing MVs into the classificationaccuracy obtained by an artificial neural network (multilayer perceptron). Four imputa- tion techniques are considered as follows: KNNI, SOM imputation, MLP imputation, and EM over one synthetic and two real data sets, varying the amount of MVs introduced. They conclude that in real-life scenarios a detailed study is required in order to

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 15, 2017 ISSN: 2348-6600
Reference ID: IJCS-229 PAGE NO: 1422-1428
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

evaluate which MVs estimation can help to enhance the classification accuracy.

• Luengo et al. [35] study several imputation methods for RBFNs classifiers, both for nat-ural and artificial (MCAR) MVs. From their results can be seen that the EC method has a good synergy with respect to the RBFN methods, as it provides better improvements in classification accuracy.

• Ding and Simonoff [14] investigate eight different missingness patterns, depending onhe relationship between the missingness and three types of variables, the observed pre- dictors, the unobserved predictors (the missing values), and the response variable. They focus on the case of classification trees for binary data (C4.5 and CART) using a model- ing bankruptcy database, showing that the relationship between the missingness and the dependent variable, as well as the existence or non-existence of MVs in the testing data, is the most helpful criterion to distinguish different MVs methods.

• Gheyas and Smith [25] propose a single imputation method and a multiple imputation method, both of them based on a generalized regression neural network (GRNN). Their proposal is compared with 25 imputation methods of different natures, from machine learning methods to several variants of GRNNs. Ninty-eight data sets are used in order to introduce MVs with MCAR, MAR, and NMAR mechanisms. Then, the results of the imputation methods are compared by means of 3 different criteria, using the following three classifiers: MLP, logistic regression, and a GRNN-based classifier, showing the advantages of the proposal.

Recently, the treatment of MVs has been considered in conjunction with other hot topics in classification, like imbalanced data sets, semi-supervised learning, temporal databases, scalability, and the presence of noisy data. Nogueira et al. [41] presented a comparison of techniques used to recover values in a real imbalanced database, with a massive occurrence of MVs. This makes the process of obtaining a set of representative records, used for the recovering techniques, difficult. They used C4.5, Naïve-Bayes, K-NN, and multilayer per- ceptron as classifiers. To treat the MVs, they applied several techniques as follows: default value substitution or related attribute recovery. The latter tries to obtain the missing value from the information of another attribute. In addition to this, cleaning of instances/attributes with too many MVs was also carried out.

Saar-Tsechansky and Provost [52] compare several different methods (predictive value imputation, the distribution-based imputation used by C4.5 and using reduced models) for applying classification trees to instances with missing values. They distinguish between MVs in "training" (usual MVs) and MVs in "prediction" time (i.e., test partition) and adapt the novel-reduced models to this scenario. The results show that for the predictive value impu- tation and C4.5 distribution based, both can be preferable under different conditions. Their novel technique (reduced models) consistently outperforms the other two methods based on their experimentation.

Matsubara et al. [36] present an adaptation of a semi-supervised learning algorithm for imputation. They impute the MV using the C4.5 and Naïve-

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 15, 2017 ISSN: 2348-6600
Reference ID: IJCS-229 PAGE NO: 1422-1428
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

Bayes classifiers by means of a ranking aggregation to select the best examples. They compare the method with three qualitative UCI [3] data sets applying artificial MVs and perform a similar study to the one presented by Batista and Monard [6], comparing with the KNNI and MC methods. Using a non-parametric statistical test, they demonstrate the better performance of the new method over the other two in some cases.

Merlin et al. [38] propose a a new method for the determination of MVs in temporal databases based on self-organizing maps. Using two classifiers for the spatial and tempo- ral dependencies, improvements in respect of the EM method in a hedge fund problem

Conclusion

Our main objective was comparison of the methods to deal with missing data imputation. These methods are promising for the best performances. However it is risky for us to conclude that they are the best methods among all other methods . Our future work is to conclude a best method to find missing data.

References

1. Acuna E, Rodriguez C (2004) Classification, clustering and data mining applications. Springer, Berlin, pp 639–648
2. Alcalá-fdez J, Sánchez L, García S, Jesus MJD, Ventura S, Garrell JM, Otero J, Bacardit J, Rivas VM, Fernández JC, Herrera F (2009) Keel: a software tool to assess evolutionary algorithms for data mining problems. Soft Comput 13(3):307–318
3. Asuncion A, Newman D (2007) UCI machine learning repository. http://archive.ics.uci.edu/ml/
4. Atkeson CG, Moore AW, Schaal S (1997) Locally weighted learning. Artif Intell Rev 11:11–73
5. Barnard J, Meng X (1999) Applications of multiple imputation in medical studies: From aids to nhanes.
Stat Methods Med Res 8(1):17–36
6. Batista G, Monard M (2003) An analysis of four missing data treatment methods for supervised learning.
Appl Artif Intell 17(5):519–533
7. Bezdek J, Kuncheva L (2001) Nearest prototype classifier designs: an experimental study. Int J Intell Syst
16(12):1445–1473
8. Broomhead D, Lowe D (1988) Multivariable functional interpolation and adaptive networks. Complex
Syst 11:321–355
9. Clark P, Niblett T (1989) The cn2 induction algorithm. Mach Learn J 3(4):261–283
10. Cohen W (1995) Fast effective rule induction. In: Machine learning: proceedings of the twelfth interna- tional conference, pp 1–10
11. Cohen W, Singer Y (1999) A simple and fast and and effective rule learner. In: Proceedings of the sixteenth national conference on artificial intelligence, pp 335–342
12. Cover TM, Thomas JA (1991) Elements of information theory, 2nd edn. Wiley, NY
13. Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7:1–30

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 15, 2017 ISSN: 2348-6600
Reference ID: IJCS-229 PAGE NO: 1422-1428
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

14. Therese D. Pigott Loyola University Chicago, Wilmette, IL, USA-A Review of methods of Missing Data
16. Ennett CM, Frize M, Walker CR (2001) Influence of missing values on artificial neural network perfor- mance. Stud Health Technol Inform 84:449–453
17. Fan R-E, Chen P-H, Lin C-J (2005) Working set selection using second order information for training support vector machines. J Mach Learn Res 6:1889–1918
18. Farhangfar A, Kurgan LA, Pedrycz W (2007) A novel framework for imputation of missing values in databases. IEEE Trans Syst Man Cybern Part A 37(5):692–709
19. Farhangfar A, Kurgan L, Dy J (2008) Impact of imputation of missing values on classification error for discrete data. Pattern Recognit 41(12):3692–3705
20. Fayyad U, Irani K (1993) Multi-interval discretization of continuous-valued attributes for classification learning. In: Proceedings of 13th international joint conference on uncertainly in artificial intelligence (IJCAI93), pp. 1022–1029
21. Feng H, Guoshun C, Cheng Y, Yang B, Chen Y (2005) A svm regression based approach to filling in missing values. In: Khosla R, Howlett RJ, Jain LC (eds) 'KES (3)', vol 3683 of lecture notes in computer science. Springer, Berlin, pp 581–587
22. Frank E, Witten I (1998) Generating accurate rule sets without global optimization. In: Proceedings of the fifteenth international conference on machine learning, pp 144–151
23. García-Laencina P, Sancho-Gómez J, Figueiras-Vidal A (2009) Pattern classification with missing data:
a review. Neural Comput Appl. 9(1):1–12
24. García S, Herrera F (2008) An extension on "statistical comparisons of classifiers over multiple data sets"
for all pairwise comparisons. J Mach Learn Res 9:2677–2694
25. Gheyas IA, Smith LS (2010) A neural network-based framework for the reconstruction of incomplete data sets. Neurocomputing In Press, Corrected Proof
26. Grzymala-Busse J, Goodwin L, Grzymala-Busse W, Zheng X (2005) Handling missing attribute values in preterm birth data sets. In: Proceedings of 10th international conference of rough sets and fuzzy sets and data mining and granular computing(RSFDGrC), pp 342–351
27. Grzymala-Busse JW, Hu M (2000) A comparison of several approaches to missing attribute values in data mining. In: Ziarko W, Yao YY (eds) Rough sets and current trends in computing, vol 2005 of lecture notes in computer science, Springer, pp 378–385
28. Hruschka ER Jr., Hruschka ER, Ebecken NF (2007) Bayesian networks for imputation in classification problems. J Intell Inf Syst 29(3):231–252
29. Kim H, Golub GH, Park H (2005) Missing value estimation for dna microarray gene expression data:
local least squares imputation. Bioinformatics 21(2):187–198
30. Kwak N, Choi C-H (2002) Input feature selection by mutual information based on parzen window. IEEE Trans Pattern Anal Mach Intell 24(12):1667–1671

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 15, 2017 ISSN: 2348-6600
Reference ID: IJCS-229 PAGE NO: 1422-1428
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

31. Kwak N, Choi C-H (2002) Input feature selection for classification problems. IEEE Trans Neural Netw
13(1):143–159
32. Cessie S le, van Houwelingen J (1992) Ridge estimators in logistic regression. Appl Stat 41(1):191–201
33. Li D, Deogun J, Spaulding W, Shuart B (2004) Towards missing data imputation: a study of fuzzy k-means clustering method. In: Proceedings of 4th international conference of rough sets and current trends in computing (RSCTC), pp 573–579
34. Little RJA, Rubin DB (1987) Statistical analysis with missing data, wiley series in probability and statis- tics, 1st edn. Wiley, New York
35. Luengo J, García S, Herrera F (2010) A study on the use of imputation methods for experimentation with Radial Basis Function Network classifiers handling missing attribute values: the good synergy between RBFNs and EventCovering method. Neural Netw 23(3):406–418
36. Matsubara ET, Prati RC, Batista GEAPA, Monard MC (2008) Missing value imputation using a semi- supervised rank aggregation approach. In: Zaverucha G, da Costa ACPL (eds) 'SBIA', vol 5249 of lecture notes in computer science. Springer, Berlin, pp 217–226
37. McLachlan G (2004) Discriminant analysis and statistical pattern recognition. Wiley, NY
38. Merlin P, Sorjamaa A, Maillet B, Lendasse A (2010) X-SOM and L-SOM: a double classification approach for missing value imputation. Neurocomputing 73(7–9):1103–1108
39. Michalksi R, Mozetic I, Lavrac N (1986) The multipurpose incremental learning system aq15 and its testing application to three medical domains. In: Proceedings of 5th international conference on artificial intelligence (AAAI), pp 1041–1045
40. Moller F (1990) A scaled conjugate gradient algorithm for fast supervised learning. Neural Netw 6:525–533
41. Nogueira BM, Santos TRA, Zárate LE (2007) Comparison of classifiers efficiency on missing values recovering: application in a marketing database with massive missing data. In: 'CIDM', IEEE, pp 66–72
42. Oba S, aki Sato M, Takemasa I, Monden M, ichi Matsubara K, Ishii S (2003) A bayesian missing value estimation method for gene expression profile data. Bioinformatics 19(16):2088–2096
43. Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max- dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238
44. Pham DT, Afify AA (2005) Rules-6: a simple rule induction algorithm for supporting decision making.
In: Industrial electronics society, 2005. IECON 2005. 31st annual conference of IEEE, pp 2184–2189
45. Pham DT, Afify AA (2006) Sri: A scalable rule induction algorithm. Proc Inst Mech Eng Part C J Mech
Eng Sci 220:537–552
46. Plat J (1991) A resource allocating network for function interpolation. Neural Comput 3(2):213–225

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 15, 2017 ISSN: 2348-6600
Reference ID: IJCS-229 PAGE NO: 1422-1428
Sri Vasavi College, Erode Self-Finance Wing 3rd February 2017
National Conference on Computer and Communication NCCC'17
http://www.srivasavi.ac.in/ nccc2017@gmail.com

47. Platt JC (1999) Fast training of support vector machines using sequential minimal optimization. In: Advances in kernel methods: support vector learning. MIT Press, Cambridge, pp 185–208
48. Pyle D (1999) Data preparation for data mining. Morgan Kaufmann, Los Altos
49. Qin B, Xia Y, Prabhakar S (2010) Rule induction for uncertain data. Knowl Inf Syst, doi:10.1007/
s10115-010-0335-7, pp 1–28 (in press)
50. Quinlan J (1993) C4.5:programs for machine learning. Morgan Kauffman, Los Altos
51. Reddy C, Park J-H (2010) Multi-resolution boosting for classification and regression problems. Knowl
Inf Syst, doi:10.1007/s10115-010-0358-0, pp 1–22, (in press)
52. Saar-Tsechansky M, Provost F (2007) Handling missing values when applying classification models.
J Learn Res 8:1623–1657
53. Safarinejadian B, Menhaj M, Karrari M (2010) A distributed EM algorithm to estimate the parameters of a finite mixture of components. Knowl Inf Syst 23(3):267–292
54. Schafer JL (1997) Analysis of incomplete multivariate data. Chapman & Hall, London
55. Schneider T (2001) Analysis of incomplete climate data: estimation of mean values and covariance matri- ces and imputation of missing values. J Clim 14:853–871
56. Song Q, Shepperd M, Chen X, Liu J (2008) Can k-NN imputation improve the performance of C4.5 with small software project data sets? A comparative evaluation. J Syst Softw 81(12):2361–2370
57. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB (2001) Missing value estimation methods for dna microarrays. Bioinformatics 17(6):520–525
58. Twala B (2009) An empirical comparison of techniques for handling incomplete data using decision trees.