Document details

International Journal of Advanced Computer Science and Applications
Volume 9, Issue 6, 2018, Pages 442-447

Performance analysis of machine learning algorithms for missing value imputation (Article)

Abidin, N.Z.[a], Ismail, A.R.[a], Emran, N.A.[b]

[a]Department of Computer Science, Kulliyyah of Information and Communication Technology, International Islamic University Malaysia, P.O Box 10, Kuala Lumpur, Malaysia
[b]Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM), Hang Tuah Jaya, Durian Tunggal, Melaka, Malaysia

Abstract

View references (19)

Data mining requires a pre-processing task in which the data are prepared, cleaned, integrated, transformed, reduced and discretized for ensuring the quality. Missing values is a universal problem in many research domains that is commonly encountered in the data cleaning process. Missing values usually occur when a value of stored data absent for a variable of an observation. Missing values problem imposes undesirable effect on analysis results, especially when it leads to biased parameter estimates. Data imputation is a common way to deal with missing values where the missing value's substitutes are discovered through statistical or machine learning techniques. Nevertheless, examining the strengths (and limitations) of these techniques is important to aid understanding its characteristics. In this paper, the performance of three machine learning classifiers (K-Nearest Neighbors (KNN), Decision Tree, and Bayesian Networks) are compared in terms of data imputation accuracy. The results shows that among the three classifiers, Bayesian has the most promising performance.

Author keywords

Bayesian networks | Data mining | Decision tree | Imputation | KNearest neighbors | Machine learning

ISSN: 2158107X
Source Type: Journal
Original language: English

DOI: 10.14569/IJACSA.2018.090660
Document Type: Article
Publisher: Science and Information Organization

References (19)

1 Vivek, A.
Research on Data Preprocessing and Categorization Technique for Smartphone Review Analysis
(2015) *International Journal of Computer Applications*, 131 (4), pp. 30-36.

2 Emran, Nurul, A.
Data completeness measures, Pattern Analysis
(2015) *Intelligent Security and the Internet of Things*, pp. 117-130.

3 Aljuaid, T., Sasi, S.
Proper imputation techniques for missing values in data sets
(2016) *Proceedings of the 2016 International Conference on Data Science and Engineering, ICDSE 2016*, art. no. 7823957. Cited 6 times.
ISBN: 978-150901280-0
doi: 10.1109/ICDSE.2016.7823957

View at Publisher

4 Somasundaram, R.S., Nedunchezhian, R.
Evaluation of Three Simple Imputation Methods for Enhancing Preprocessing of Data with Missing Values
(2011) *International Journal of Computer Applications*, 21 (10), pp. 14-19. Cited 14 times.

5 Kang, H.
The prevention and handling of the missing data
(2013) *Korean Journal of Anesthesiology*, 64 (5), pp. 402-406. Cited 25 times.
http://ekja.org/Synapse/Data/PDFData/0011KJAE/kjae-64-402.pdf
doi: 10.4097/kjae.2013.64.5.402

View at Publisher

6 Schmitt, P., Mandel, J., Guedj, M.
A Comparison of Six Methods for Missing Data
(2015) *Imputation Journal of Biometrics & Biostatistics*, 6 (1).

7 Marsh, H.W.
Pairwise deletion for missing data in structural equation models: Nonpositive definite matrices, parameter estimates, goodness of fit, and adjusted sample sizes
(1998) *Structural Equation Modeling*, 5 (1), pp. 22-36. Cited 35 times.
doi: 10.1080/10705519809540087

View at Publisher

8 Soley-Bori, M.
(2013) *Dealing with missing data: Key assumptions and methods for applied analysis*. Cited 17 times.
Boston University

9 Susianto, Y., Notodiputro, K.A., Kurnia, A., Wijayanto, H.
A Comparative Study of Imputation Methods for Estimation of Missing Values of per Capita Expenditure in Central Java (Open Access)
(2017) *IOP Conference Series: Earth and Environmental Science*, 58 (1), art. no. 012017.
http://www.iop.org/EJ/volume/1755-1315
doi: 10.1088/1755-1315/58/1/012017

View at Publisher