

Reporte Avance 2

Proyecto de sistemas inteligentes

Profesor:

Dr. Raul Ramírez Velarde

Bernardo García

A00570682

6 de octubre de 2023

1. Applying AI methodologies to improve data quality.

Data cleaning is a crucial step in predictive modeling as it directly influences the accuracy and reliability of the results. Ensuring that the data used for training machine learning models is accurate, complete, and consistent is essential. Dirty or inconsistent data can lead to misleading insights and inaccurate predictions. By cleaning the data, removing errors, filling in missing values, and addressing inconsistencies, analysts create a solid foundation for the predictive models. This meticulous process not only enhances the model's performance but also increases the trustworthiness of the conclusions drawn from the analysis, making the predictions more reliable and valuable for decision-making processes (Brownlee J, 2020).

Currently, the project relies on mean imputation to handle missing data, a method that, while straightforward, has limitations. Even though statistical and probabilistic approaches have long been employed in data cleansing, their usage has mainly centered around identifying outliers in numerical data. Now, with the rise of large-scale machine learning capabilities and accessible resources, integrating ML techniques into data cleaning processes is gaining traction as a popular and promising avenue (Ilyas I.F, Chu Xu, 2019). By comprehensively analyzing relationships between variables, these models offer a more accurate and nuanced approach to filling in missing or erroneous values.

Transitioning to supervised learning techniques for data cleaning would mark a significant improvement. Unlike mean imputation, which simply inserts average values, supervised learning models delve into the complexities of the data. Considering multiple factors, they capture intricate correlations often missed by mean imputation (Batista G, Monard M C. 2003). This shift to supervised learning ensures a more precise estimation of missing values, elevating data quality and, subsequently, enhancing the reliability of the predictive models.

As described by M.K Hasan (et al. 2021), the mean/median/mode-based missing value imputation is one of the most employed methods (2nd) for missing data handling. Nevertheless, this is not a suitable imputation method for categorical values since they maintain the average value of an attribute but does not consider the relationships between attributes. This limitation can be problematic in applications where the interconnections between attributes are crucial. Additionally, mean value imputation might produce fractional values that need to be truncated, leading to information loss.

On the other hand, Hasan also includes that linear/logistic regression-based imputation for missing data are also on the list for most used methods, and these analyze the connections between attributes and use regression coefficients to fill in the missing values. These methods are reported to be applied for predicting both numerical and categorical attribute values.

Here goes a transition paragraph to explore used techniques

**Deep Learning Methods**

For instance, Smieja M (et al. 2019) suggested an adapted version of MLP to address diverse domain problem datasets, ranging from 0.25% to 23.8% missing rates. Moreover, Chen (et al. 2015) evaluates how well DBN performs on three small-scale datasets with simulated missing rates ranging from 5% to 40%. Also, in Jie Lin’s (et al. 2020) research, a DBN is employed to fill in missing data in three incomplete datasets gathered from a cluster monitoring system, where the missing rates vary from 1% to 15%.

Nevertheless, in these last-mentioned studies, deep neural networks are solely pitted against baseline imputation methods using various statistical and machine learning techniques to draw a final conclusion. In simpler terms, a direct performance comparison between different deep neural networks has not been conducted.

Additionally, Lin (et al. 2022) examines the performance of two deep learning models for missing value imputation: a multilayer perceptron (MLP) and deep belief networks (DBN), which outperformed the accuracy of the mean value imputation and even other ML and statistical techniques (CART, KNN, SVN methods) in datasets with 20%-50% missing data rate, being DBN better to MLP in the different dataset tests.

**Machine Learning Methods**

Also, according to his literature research, Lin (et al. 2022) references the work of Dougherty (et al. 1995), Garcia (et al. 2013), and Liu (et al. 2002) about data discretization using machine learning algorithms (decision trees, Apriori, and Naïve Bayes) to create more effective models rather than the deep learning alternatives. However, data discretization is an important data preparing process that can be done before or after filling missing values in the dataset. Still, this could mean machine learning techniques can be used when the data that must be predicted is already discrete, finite or numerical.

Since the last couple of decades, machine learning methods have been tested as missing data solutions. In the research of Lakshminarayan (et al. 1996) conducted their research with Autoclass, a Bayesian unsupervised learning method, and C4.5, a decision-tree based supervised learning method, for missing data imputation in industrial data bases. Finding that Autoclass performs accurate results when used to predict multiple target variables, or multiple choices for an attribute. On the other hand, C4.5 has a high accuracy when used to predict single values for missing data.

A more recent research by Batista and Monard (2003) uses K-nearest neighbors algorithm (KNN) to predict missing data, and results in a superior performance when compared to CN2, C4.2 and mean imputation methodologies in several datasets with high amount of missing data. Nevertheless, the tests when the database has attributes with high correlation, showed that KNN should be avoided in occasions like this one. KNN can give a precise value, or suddenly can change the treated attribute by another one with also high correlation. This is an interesting situation since it lets us recall the limitations of ML methods and the importance of human interpretation of the data. As Batista and Monard recalls in their research conclusions: “Missing data imputation can be harmful because even the most advanced imputation method is only able to approximate the actual (missing) value.”

Furthermore, when getting into machine learning techniques for missing data imputation, several random forest algorithms are often used for the process. Random forests consist of multiple tree predictors, where the structure of each tree is determined by values sampled randomly and independently from a common distribution across all trees in the forest (Breiman L, 2001).

For instance, Anoop D Shah (et al. 2013) contrasts a traditional Multivariate imputation by chained equations (MICE) approach with imputation through missForest and introduces a proposed version of MICE that employs random forest for imputing each variable. These methods are evaluated using a complex survival analysis involving patients from the CALIBER database. Such study resulted in: for categorical variables, the outcomes obtained from random forest MICE with 10 trees and random forest MICE with 100 trees were nearly identical, resulting in a coverage of 95% confidence intervals. In the case of *missForest* resulted in imputed values nearer to real values, but with a lower confidence value.

Stekhoven and Buhlmann’s *missForest* (2011) is a random forest algorithm that enables missing value imputation for any type of data. It can manage multivariate data that includes both continuous and categorical variables at the same time. It operates without the necessity for fine-tuning parameters and avoids making assumptions about the data's distributional aspects. Its superiority is demonstrated over established imputation methods such as KNN imputation and multivariate imputation using chained equations on various real datasets from diverse biological and medical fields.

In addition to this, Guo Y-C (et al. 2021) experiments with *missForest* in various missing data scenarios and compares them against KNN and other random forest variations evaluated by the root mean square error (RMSE), the proportion of falsely classified entries (PFC), and type-I errors from the Cox proportional hazard model. The results showed that all four algorithms performed well and similar int RMSE and PFC standards. Although, when coming to type-I errors, *missForest* is the only approach that does not exhibit exaggerated flaws across all missing mechanisms.

A very similar research made by Alsaber Ahmad R (et al. 2021) proposed missForest against other machine learning techniques such as random forest (RF), k-nearest neighbor (KNN), Bayesian principal component analysis (BPCA), multiple imputation using expectation maximization with bootstrapping (EM with Bootstrapping), and predictive mean matching (PMM), for 5%-40% missing data imputation related to air quality monitoring. The performance results of the study, based in RMSE and mean squared error (MAE), demonstrated that missForest was the sole imputation method consistently displaying a comparatively lower error rate of 0.82. The approach resulted in a root mean square error of 1.04 and showed the smallest variance in the predicted values of pollutants. Nevertheless, a potential link between pollutant values and the missing variables exists. Consequently, these findings might not be applicable in situations where missing data are due to non-random causes. Additionally, proficiency in R programming is necessary, making it more demanding than KNN or PMM methods.

*missForest* can be an interesting alternative for mixed-type data series for its several advantages such as: computational efficiency, easy access through programming packages and its ability to manage large amount of data (Arriagada P, et al. 2021).

Continue with: <https://scholar.google.com.mx/scholar?hl=es&as_sdt=0%2C5&q=random+forest+for+missing+data&btnG=>

1. Applying AI techniques to optimize the selection of model parameters.

References:

1. Brownlee Jason (2020). Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python. (p.12-14)
2. Batista G and Monard M C. (2003). An analysis of four missing data treatment methods for supervised learning. (p. 522-532).
3. Ilyas Ihab F, Chu Xu (2019). Data Cleaning. (Ch 7. p. 195-196)
4. Hasan K, Alam A, Roy S, Dutta A, Jawad T, Das S. (2021). Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021) (p. 1,13-14)
5. Smieja M, Struski L, Tabor J, Zielinski B, Spurek P. (2018). Processing of missing data by neural networks. (p.7-9).
6. Chen Z, Liu S, Jiang K, Xu H, Cheng X. (2015). A data imputation method based on deep belief network.
7. Lin J, Li N H, Alam Md A, Ma Y. (2020) Data-driven missing data imputation in cluster monitoring system based on deep neural network. (p. 860–877).
8. Lin W-C, Tsai C-F, Zhong J R. (2022). Deep learning for missing value imputation of continuous data an effect of data discretization. (p. 2-8).
9. Dougherty J, Kohavi R, Sahami M. (1995). Supervised and unsupervised discretization of continuous features. (p. 194-202).
10. Garcia S, Luengo J, Saes J A, Lopez V, Herrera F. (2013). A survey of discretization techniques: taxonomy and empirical analysis in supervised learning. (p. 734-750).
11. Liu H, Hussain F, Tan C L, Dash M. (2002). Discretization: an enabling technique. (p. 393-423).
12. Lakshminarayan K, Harp S A, Goldman R, Samad T. (1996). Imputation of missing data using machine learning techniques.
13. Breiman Leo. (2001). Random Forests. (p 5-32).
14. Anoop D S, Bartlett J W, Carpenter J, Nicholas O, Hemingway H. (2013). Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study. (p. 764-772).
15. Stekhoven Daniel J and Buhlmann Peter. (2011). MissForest -non-parametric missing value imputation for mixed-type data. (p. 112-113, 117).
16. Guo C-Y, Yang Y-C and Chen Y-H. (2021). The Optimal Machine Learning-Based Missing Data Imputation for the Cox Proportional Hazard Model. (p 2-8)
17. Alsaber A R, Pan J, Al-Hurban A. (2021). Handling Complex Missing Data Using Random Forest Approachforan Air Quality Monitoring Data set :A Case Study of Kuwait Environmental Data (2012 to 2018). (p. 2-17).
18. Arriagada P, Karelovic B, Link O. (2021). Automatic gap-filling of daily streamflow time series in data-scarce regions using a machine learning algorithm.

--------------------------------------------------------------------------------------------------------------------------

Links

1. <https://books.google.com.mx/books?hl=es&lr=&id=uAPuDwAAQBAJ&oi=fnd&pg=PP1&dq=the+importance+of+data+cleaning+in+predictive+models&ots=Cl9Kvi9RwT&sig=pm54rEJbwm5NDsu4-rE614mVo6A#v=onepage&q=the%20importance%20of%20data%20cleaning%20in%20predictive%20models&f=false>
2. https://books.google.com.mx/books?hl=es&lr=&id=RxieDwAAQBAJ&oi=fnd&pg=PP2&dq=importance+of+data+cleaning&ots=bTCkYC3pXd&sig=enZDkdm69dCKgS45mxubxFRld\_c#v=onepage&q=importance%20of%20data%20cleaning&f=false
3. <https://www.sciencedirect.com/science/article/pii/S2352914821002653?ref=pdf_download&fr=RR-2&rr=8112a02d4f63219e>
4. <https://www.sciencedirect.com/science/article/pii/S0950705121011527>
5. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://arxiv.org/pdf/1805.07405.pdf
6. ND
7. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/file:///C:/Users/berna/Desktop/Adopted%20Files/Tec%20de%20Monterrey/11vo%20Semestre/Proyectos%20de%20Sistemas%20Inteligentes/Referencias/ML%20for%20Data%20cleaning/s10489-019-01560-y.pdf
8. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://cdn.aaai.org/KDD/1996/KDD96-023.pdf
9. <https://www.tandfonline.com/doi/epdf/10.1080/713827181?needAccess=true>
10. <https://www.readcube.com/articles/10.3389/fpubh.2021.680054>
11. <https://academic.oup.com/bioinformatics/article/28/1/112/219101?login=false>
12. <https://www.mdpi.com/1660-4601/18/3/1333>

Very useful (pay)

<https://link.springer.com/article/10.1007/s42044-020-00065-z>

<https://link.springer.com/article/10.1007/s10115-022-01661-0>

<https://link.springer.com/article/10.1007/s00232-015-9856-z>