Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 31 No. 7s (2024)

Handling Missing Data through Artificial Neural Network

Geeta Chhabra

Data Informatics & Innovation Division, Ministry of Statistics & PI, Govt. of India, India,
geeta_chhabra@rediffmail.com

Article History:

Received: *07-06-2024*

Revised: *07-07-2024*

Accepted: *31-07-2024*

Abstract:

Missing data is a difficult problem to solve when working with real-world datasets. It is vital to improve data quality by imputing missing values in order to obtain effective data learning. Deep learning has recently risen to prominence as the most effective sort of machine learning technique for uncovering hidden knowledge in massive datasets and making accurate predictions. We have used "back-propagation artificial neural network" to impute categorical missing data. The main goal is to see how well "neural network" compares to statistics and machine learning for resolving categorical missing data. The results of "back-propagation" are compared with multiple imputation and random forest. It consistently outperforms alternative methods both in "training" and "test" data sets, showing that "neural network" is a suitable for reconstructing "missing values" in "multivariate analysis".

Keywords: Artificial Neural Network, Multiple Imputation, Random Forest, backpropagation.

1. Introduction

In the modern era, a great amount of data is gradually generated in various fields, and the quick increase in data size has demonstrated the importance of big data analysis. Because low quality data cannot be used to construct credible models, it is vital to ensure that the data acquired is trustworthy and valuable. Missing values in the acquired data set are a regular and inevitable issue in practise, and they can lead to uncertainty in data analysis. Missing data can be found in a variety of areas, including gene expression, traffic control, industrial informatics, image processing, and software development. Without addressing the aforementioned issue, data analysis can yield deceptive conclusions. As a result, it is vital to improve the situation.

Missing data could be due to a variety of factors, including questionnaire non-response, failure to follow up on research subjects, data entry errors, equipment failure, or incomplete or deleted records. Simply excluding instances of "missing data" could lead to "biased inference", lower "statistical power" and outcomes, that are less generalizable. The issue of "missing data" can be categorised in three groups based on missingness assumptions: "missing completely at random ("MCAR"), "missing at random" ("MAR"), and "missing not at random" ("MNAR"). The majority of "missing data" research is presumed to be MAR in general.

MICE ("multiple imputation by chained equations") is most widely used "statistical" approach for dealing with "missing data", especially MAR data. Many statistical software packages, such as SPSS, STATA, and R, have MICE. Because MICE defaults to "predictive mean matching" and "logistic

regression" are limited to handle non-linear connections and interactions between variables and may produce biased results.

“Random forest”, an “ensemble machine-learning” technique for “multi-classification” or “decision tree regression”, is one way to overcome “non-linearity”. “Random forest” beat “MICE” in both real-world and simulated datasets, according to subsequent research (Shah et al., 2014).

Back-propagation “artificial neural network” (ANN) is a prominent “deep learning” method that has been widely used for prediction, classification, and decision-making. In three open-source datasets, it outperformed previous approaches in terms of imputation accuracy while replacing MAR data. The primary goal of this research was to see how accurate ANN was in imputing missing values in categorical variables. This research also sought to compare its results to those of MICE and random forest.

2. Related work

This section examines few representative works that use deep learning to handle missing data. Many researches have been undertaken to address the issue of “missing data”, which is a widespread problem in many data-driven systems.

In their article “Analysis of backpropagation neural neural network algorithm on student ability based cognitive aspects”, Izhari et al., (2020), used a dataset for predictions using "artificial neural networks" with "back propagation" for the maximum accuracy in “binary” and “bipolar sigmoid” to estimate children's cognitive abilities. With considerable success, the “back propagation artificial neural network” technique is utilised to predict students’ ability on cognitive assessments.

Cihan, P. (2020) in his article “Deep Learning-Based Approach For Missing Data Imputation” used Denoising Autoencoder on DNA, Wine & Shuttle data and compared the results with kNN and MICE. He concluded that Deep Learning-Based Approach outperforms statistical methods.

Phung et (2019) in their manuscript “A deep learning technique for imputing missing healthcare data” used denoising autoencoder on “Linked Birth/Infant Death Cohort” data records and compared the outcome with matrix factorization, KNN, SVD mean, median and concluded that deep learning method has better accuracy. Experiments show that their method produces lower imputation mean squared error than other imputation methods.

To deal with attributes that had more than 1% "missing data," Desiani et al. (2021) employed "artificial neural network" in their manuscript “Handling Missing Data Using Combination of Deletion Technique, Mean, Mode, and Artificial Neural Network Imputation for Heart Disease Dataset”. The outcomes of approaches and procedures used to deal with “missing data” were evaluated using the performance of classification method. “Artificial neural networks”, “nave bayes”, “support vector machines” and “K-nearest neighbor” are some of the classification techniques employed in this research. The accuracy, sensitivity, specificity, and recall of classification without missing data were compared to the accuracy, sensitivity, specificity, and recall of classification after “imputation”. Furthermore, the mean squared error findings were compared to assess how similar was the predicted label in categorization with the original label. Artificial Neural Network had the lowest Mean Squared Error, indicating that when compared to other approaches, the Artificial Neural Network performed

exceptionally well on datasets with missing data. The accuracy, specificity, and sensitivity results for “classification” revealed that “imputation” of “missing data” could improve classification performance, particularly with the artificial neural network method.

Yoon et al (2018) in their manuscript “Gain: Missing data imputation using generative adversarial nets” used Imputation Nets (GAIN) on UCI Machine Learning Repository dataset namely Breast, Spam, Letter, Credit, and News. The results were compared with MICE, MissForest, Auto-encoder and Expectation-maximization. GAIN outperforms other “state-of-the-art” imputation methods.

Dong et al (2021) in their manuscript “Generative adversarial networks for imputing missing data for big data clinical research” used “Generative adversarial imputation (GAIN)” on mixed type of data and compared the results with “MICE” and “missForest”. In large "clinical datasets", GAIN outperformed "MICE" & "missForest" as an "imputation" strategy, and was more resilient to high "missingness" rates (50 %).

3. Methodology

The traditional methods for dealing with “missing data” can be categorized in two groups. The first is deletion, which is intended to erase all instances with missing values for specified features. Imputation is the second strategy, which seeks to replace missing values with some suitable ones (Cardoso Pereira et al., 2020).

Furthermore, “deep learning” has been widely applied in a variety of domains in recent years, including missing data imputation, resulting in a significant improvement in imputation performance by utilising a vast amount of training data (Khare et al., 2021, Yang et al., 2019; Fei et al., 2017).

As a result, of their amazing success, “artificial neural networks” (ANN) have received a lot of attention in recent years. (Wei & Yang, 2021; Özden et al., 2017; Wang et al., 2016)

3.1 Artificial Neural Network

It is made up of artificial neurons that look like human brain neurons. It is a collection of artificial neurons (known as processing units) that are linked to one another. A weight is assigned to each link, which symbolises the influence of one neuron on the other. The signals to conduct the activities are sent by neurons in the brain. Artificial neurons link in a neural network to complete tasks in a similar way. (Silva-Ramrez and colleagues, 2015)

The term "network" refers to the connections between "neurons" in different layers. Every system is made up of three layers: an “input layer”, a “hidden layer” and an “output layer” (Saputra et al., 2017). The “input layer” contains input “neurons” that send data to the “output layer” via “synapses”.

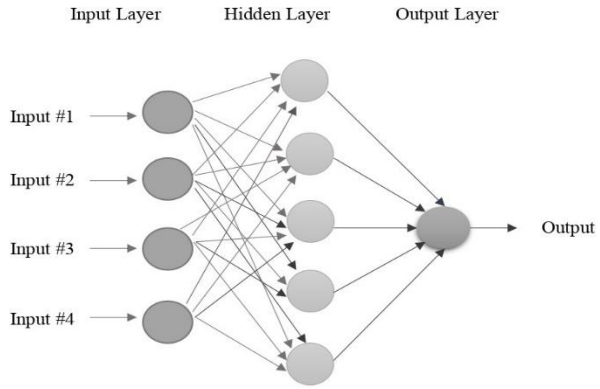

Figure 1: Backpropagation Neural Network

The basic concept underlying training is to select a set of weights at random, apply inputs to "neural net" and compare output to the weights assigned. The calculated value is compared with real one. Using generalised delta rule, difference is used to update the weights of each layer. Back propagation is the name of the training algorithm. When the error between actual output and computed output is smaller than previously determined number, after several training epochs, the neural net is considered trained. After being trained, the "neural net" can be used to analyse fresh data and classify it according to the information it requires. The "neuralnet" library in R makes this simple.

3.2 Multiple Imputation

It is a three-stage statistical strategy for coping with missing values, i) data imputation, which generates m imputed data sets from a distribution that yields m full data sets, ii) complete data analysis, which is done on each of the m imputed data sets, and iii) data pooling, which uses simple principles to pool the data acquired through data analysis.

Multiple Imputation is a method for replacing missing values with potential solutions. Using imputation methods, the incomplete dataset is turned into a complete dataset that may then be analysed using any standard analysis approach. As a result, multiple imputations have become popular for dealing with missing data.

The process is done numerous times for all variables with missing values in the multiple imputation technique, as the name implies, and then analysed to aggregate m number of imputed data sets into one imputed data set. For this, "R" supplies the MICE package, which is simple to use.

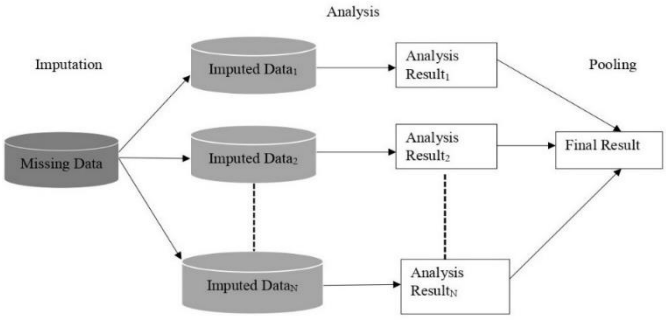

Figure 2: Multiple Imputation Mechanism

In the case of quantitative variables, “predictive mean matching” is an appealing technique for “missing value” substitution. It is comparable to “regression”, in which “observed value” is equated with “missing value” to get it close to “predicted mean”. (Vink et al, 2014). The values are estimated using a combination of “linear regression” and closest “neighbour”.

3.3 Random Forest

Random forest is a frequently used “supervised machine-learning” approach for “classification” & “regression”. It uses majority vote for classification & average for regression to create decision trees from samples.

It can handle data sets with both continuous and categorical variables, as in “regression” and “classification”. When it comes to categorization difficulties, it excels the competitors. The problem of decision trees overfitting their training set is addressed by random decision forests. The “randomForest” library from R makes this simple.

Steps involved in random forest algorithm:

- From data set of k records, it selects n records at random.
- For each sample, a decision tree is built.
- Result is generated by each decision tree.
- For “classification” and “regression”, the final output is based on majority vote or averaging, as applicable.

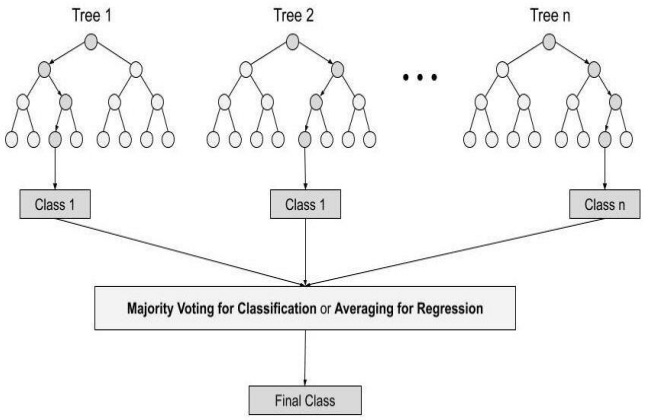

Figure 3: Random Forest Mechanism

4. Experimental framework

This manuscript utilizes data sets from the UCI machine learning repository 1) iris data set includes three classes, each with 50 instances, each class is related to a different type of iris plant. This is multiclass classification to predict class of iris plant, 2)seed dataset that comprises wheat kernels from three distinct wheat kinds: Kama, Rosa and Canadian, 70 elements each, selected randomly for the experiment is again a multiclass classification 3) infert dataset that is about infertility after spontaneous and induced abortion with 248 instances of matched case-control study with 83 elements of case and 164 control, a binary classification problem having imbalanced dataset.(Dua & Graff, 2019; Charytanowicz et al, 2012; Trichopoulos et al,1976)

The data was divided equally i.e. 50% into training and testing. On these data set back-propagation artificial neural network with three “hidden layers”, “ensemble learning” method “random forest” and “statistical” technique “multiple imputation” with “predictive mean matching’ have been applied for imputation of categorical variables which is binary and multiclass The “artificial neural net” performance is significant better than others in terms of accuracy. Machine Learning packages in R/Revolution environment like caret, neuralnet, MICE and randomForest have been used.

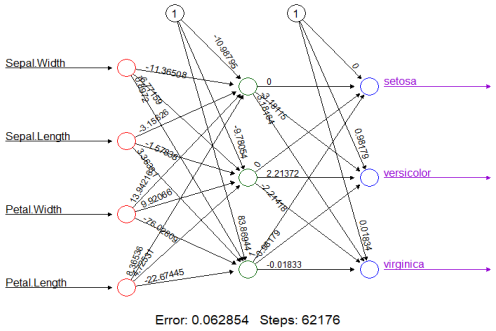

Figure 4.1: ANN with Iris dataset

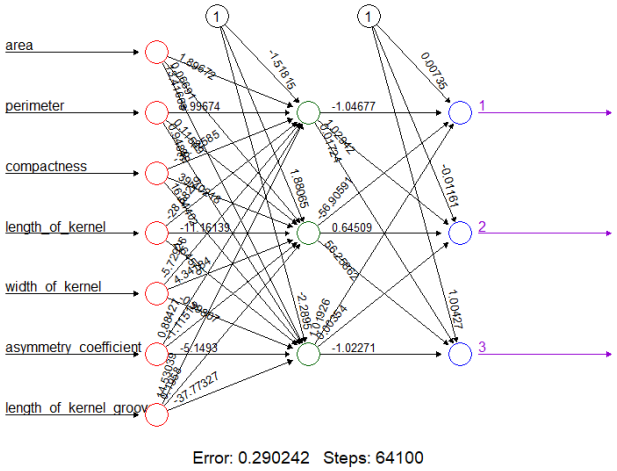

Figure 4.2. ANN with Seed dataset

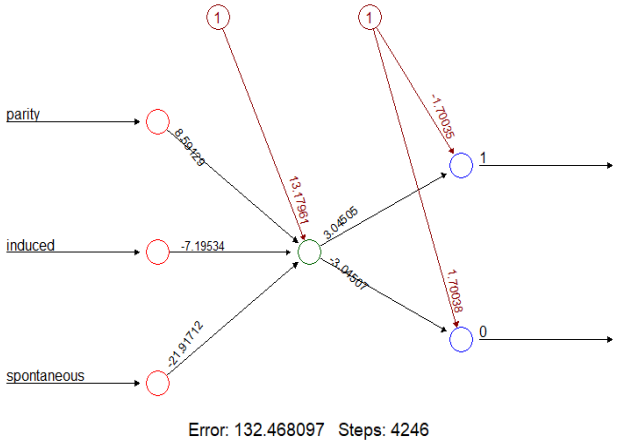

Figure 4.3: ANN with Infert dataset

Table 1: Comparison of accuracy on different dataset with different classifier

Dataset	Multiple Imputation	Random Forest	Artificial Neural Net
Iris	90.67	94.67	96
Infert	73.39	75	76.61
Seed	73.3	86.67	94.29

5. Conclusion & Future Work

We compared “artificial neural net-based” missing data imputation approaches in this manuscript, with the goal of transforming partial data into complete data. Overall, “artificial neural networks” outperformed “MICE” and ‘random forests” in imputation of missing data for categorical variables, even when the data was imbalanced and the missingness rate was high i.e. about 50 %.

There is no single ideal algorithm for dealing with missing data. Indeed, the approach used will be determined by “missingness” assumption as well as “auxiliary factors” that could explain why data is missing. In some cases, for example, “complete case analysis” may be preferable than multiple imputation. Based on our findings, we recommend that while selecting an imputation method, consider the “missingness rate”, “variable distribution”, and “expected computation” time. Furthermore, the use of “multiple imputation” methods and “sensitivity” analysis may improve the results' reliability.

References

[1] Shah, A., Bartlett, J., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). 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. https://doi.org/10.1093/aje/kwt312

[2] Dong, W., Fong, D., Yoon, J., Wan, E., Bedford, L., Tang, E., & Lam, C. (2021). Generative adversarial networks for imputing missing data for big data clinical research. *BMC Medical Research Methodology*, *21*(1). https://doi.org/10.1186/s12874-021-01272-3

[3] Cardoso Pereira, R., Seoane Santos, M., Pereira Rodrigues, P., & Henriques Abreu, P. (2020). Reviewing Autoencoders for Missing Data Imputation: Technical Trends, Applications and Outcomes. *Journal Of Artificial Intelligence Research*, *69*, 1255-1285. https://doi.org/10.1613/jair.1.12312

[4] Desiani, A., Dewi, N., Fauza, A., Rachmatullah, N., Arhami, M., & Nawawi, M. (2021). Handling Missing Data Using Combination of Deletion Technique, Mean, Mode and Artificial Neural Network Imputation for Heart Disease Dataset. *Science And Technology Indonesia*, *6*(4). https://doi.org/10.26554/sti.2021.6.4.303-312

[5] CİHAN, P. (2020). EKSİK VERİLERİ TAMAMLAMADA DERİN ÖĞRENME TEMELLİ YAKLAŞIM. *Eskişehir Teknik Üniversitesi Bilim Ve Teknoloji Dergisi B - Teorik Bilimler*, *8*(2), 337-343. https://doi.org/10.20290/estubtdb.747821

[6] Phung, S., Kumar, A., & Kim, J. (2019). A deep learning technique for imputing missing healthcare data. *2019 41St Annual International Conference Of The IEEE Engineering In Medicine And Biology Society (EMBC)*. https://doi.org/10.1109/embc.2019.8856760

[7] Khare, P., Wadhvani, R., & Shukla, S. (2021). Missing Data Imputation for Solar Radiation Using Generative Adversarial Networks. *Proceedings Of International Conference On Computational Intelligence*, 1-14. https://doi.org/10.1007/978-981-16-3802-2_1

[8] Yang, Y., Wu, Z., Tresp, V., & Fasching, P. (2019). Categorical EHR Imputation with Generative Adversarial Nets. *2019 IEEE International Conference On Healthcare Informatics (ICHI)*. https://doi.org/10.1109/ichi.2019.8904717

[9] Izhari, F., Zarlis, M., & Sutarman. (2020). Analysis of backpropagation neural neural network algorithm on student ability based cognitive aspects. *IOP Conference Series: Materials Science And Engineering*, *725*(1), 012103. https://doi.org/10.1088/1757-899x/725/1/012103

[10] Silva-Ramírez, E., Pino-Mejías, R., & López-Coello, M. (2015). Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns. *Applied Soft Computing*, *29*, 65-74. https://doi.org/10.1016/j.asoc.2014.09.052

[11] Saputra, W., Tulus, Zarlis, M., Sembiring, R., & Hartama, D. (2017). Analysis Resilient Algorithm on Artificial Neural Network Backpropagation. *Journal Of Physics: Conference Series*, *930*, 012035. https://doi.org/10.1088/1742-6596/930/1/012035

[12] Wei, W., & Yang, X. (2021). Comparison of Diagnosis Accuracy between a Backpropagation Artificial Neural Network Model and Linear Regression in Digestive Disease Patients: an Empirical Research. *Computational And Mathematical Methods In Medicine*, *2021*, 1-10. https://doi.org/10.1155/2021/6662779

[13] Özden, S., Saylam, B., & Tez, M. (2017). Is artificial neural network an ideal modelling technique?. *Journal Of Critical Care*, *40*, 292. https://doi.org/10.1016/j.jcrc.2017.06.011

[14] Wang, J., Wang, F., Liu, Y., Xu, J., Lin, H., & Jia, B. et al. (2016). Multiple Linear Regression and Artificial Neural Network to Predict Blood Glucose in Overweight Patients. *Experimental And Clinical Endocrinology & Diabetes*, *124*(01), 34-38. https://doi.org/10.1055/s-0035-1565175

[15] Fei, Y., Hu, J., Li, W., Wang, W., & Zong, G. (2017). Artificial neural networks predict the incidence of portosplenomesenteric venous thrombosis in patients with acute pancreatitis. *Journal Of Thrombosis And Haemostasis*, *15*(3), 439-445. https://doi.org/10.1111/jth.13588

[16] Yoon, J., Jordon, J., and Van Der Schaar, M. (2018). Gain: Missing data imputation using generative adversarial nets. Proceedings of the 35th International Conference on Machine Learning. https://proceedings.mlr.press/v80/yoon18a.html.

[17] Vink G, Frank LE, Pannekoek J, Buuren SV.(2014).Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica*, 68(1), 61–90.

[18] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[19] Charytanowicz, Magorzata, Niewczas, Jerzy, Kulczycki, Piotr, Kowalski, Piotr & Lukasik, Szymon. (2012). Seeds. UCI Machine Learning Repository.

[20] Trichopoulos et al (1976) Br. J. of Obst. and Gynaec. 83, 645–650