Imputing Missing Values for Data Samples In Educational Empirical Research Using Machine Learning Techniques

Pooja Manghirmalani Mishra and **Rabiya Saboowala
**Machine Intelligence Research Labs, India.*
***Independent Researcher, India*
**Corresponding email: pmanghirmalani@ieee.org*

Abstract

The major concern that arises in every research that involves physical data collection is the loss of data which happens because certain units of sample are rejected due to missing responses or values of the participants. The problem of missing values in the data is more profoundly seen where the amount of time spent by the participant in going through the test is in hours. For this study, data was collected from in-service teachers in order to understand their acceptance to adopting blended learning pedagogy as the new normal post Covid era. To solve the proposed problem, two methods are investigated for imputing the missing values; network prediction using multi-layered feedforward neural network model which predicts the missing value and multi-layered feed-backward neural network model which imputes missing values using estimated values from predictor networks, and their performance is compared. It has been found that the multiple-layered feedback neural network model was superior in terms of accuracy. When we combine the technology of machine learning with educational research problems, a lot of issues can be catered to, out of which, one of the common ones is of the missing values.

*Key words**:** Backpropagation Algorithm, Imputing, Missing Values, Feedback, Feedforward, Neural Networks*

Introduction

The advancement and trend in technology has changed the way knowledge is acquired by learners, making education more approachable to all. Lately, Education has undergone a massive revolution of its own and the era of COVID-19 pandemic considered as a “black swan” event is a stark example of this. As per statistics put out on its website, UNESCO states that more than over 190 countries faced interference in formal teaching and learning impacting more than 1.5 billion learners at one point. As per the same report, it was estimated that about 320 million learners have been impacted in India alone (UNESCO, 2021)

Blended learning (BL), also known as hybrid learning (Graham 2013), has been widely adopted across all sectors of education including learning for teachers with some scholars referring it to as the “new traditional model” (Ross and Gage 2006, p. 167) or the “new normal” (Norberg et al. 2011, p. 207). BL has been around globally for a long time but has come to the light and gained importance due to the pandemic. In simple words, BL approach is a cocktail of e-learning (also referred as online learning) and traditional learning (face-to-face) which is increasingly being adopted by all educational institutes as the most reliable and sustainable approach. BL brings together the best of both worlds where the delimitations of one are satisfied by the pros of the other, thus enabling learners to stay engaged and motivated (Roy, n.d.). This approach helps to cater keeping in mind the individual learning styles and needs. It is a convenient pedagogy, hassle-free and cost-effective model which can be used to ensure engaging, holistic and personalised learning. BL leverages emerging technologies such as artificial intelligence and machine learning to tailor learning experiences as per the pace and performance of learners meeting their learning needs.

It has been observed that even in a well-designed and controlled research study, missing data occurs and this problem is relatively common. Missing data present various problems to state a few, the absence of data or incomplete data reduces the statistical power, which refers to the probability that the test will falsely reject the null hypothesis, the results obtained from such data can cause biasness in the estimation of parameters, it can reduce the representativeness of the samples and it can complicate the analysis of the study. Missing data may arise due to various reasons like during data recording when responses are related to sensitive questions (drug use, violence, abuse) or when the research survey conducted is for a longer time or is lengthy and while data collection individuals skip few items due to boredom or lack of interest (Jamshidian, 2004).

The problem of missing data can have a drastic effect on the conclusions and generalisations that can be drawn from the data. Also, it may threaten the validity of the trials and can lead to invalid conclusions. Some researchers have focused on the issue arising due to missing data and the methodology that can be adopted in order to avoid or minimise this. By far the most common and widely accepted approach to the missing data that has been adopted in the field of education is to omit those cases with the missing data and analyse the remaining data, which is at the cost of sample size deduction (Kang, 2013). Some researchers insist that doing so may introduce bias in the estimation of the parameters. The present research study also has tried to focus on dealing with the problem of missing data by applying knowledge of machine learning techniques to input this missing values. The further discussed techniques train from the existing database where one or more parameters are missing. With the heavy integration of technologies and evolving areas of machine learning, there can be improvements in educational research where the missing values of data serve as limitation leading to discarding of huge data entries.

Computational prediction requires the data to be in a particular format. As this work deals with teachers from across India, their time, importance of the data and the length of the questionnaire need to be accepted and hence quite a lot of data was bound to have missing values. Collecting this type of data (from different states, educational institutes across India) is a tedious task which requires financial and time inputs. Any loss of data proves to be very expensive and further reduces small sized database.

This paper explores in detail different computational methods and models applied for predicting the missing values in a BL database. Having elaborately explored different approaches, researchers have found that there are still possible ways of approaching the given problem. Section II gives an exclusive review of the existing literature whereas Section III discusses the BL dataset. Section IV elaborates on the Machine learning techniques used to impute missing values. In section V comparative results are shown and Section VI elaborates on the outcomes of the results whereas Section VII discusses future objectives respectively.

Taxonomy

Various methodologies of imputing have been used so far for the problem of missing values in the data. Few of the closet works to the present study are stated below: Twala and Cartwright (2005) analysed the ensembles of imputation methods in order to improve effort prediction accuracy and classifier learning efficiency. The issue addressed in the study is the impact of missing value to the prediction accuracy. Twala et al. (2006)

investigate the randomization of decision tree building algorithms to improve prediction accuracy. The main objective here was to investigate the impact of missing values to prediction accuracy and how the ensemble missing data techniques could be implemented to improve effort prediction accuracy. However, the results did not support when the data set is small with many attributes and gives different performance as the proportion of the missing data is increased.

Molloken and Jorgensen (2003) mentioned that most of the predictions done in a software project are based on expert judgments because there is no evidence that a formal prediction models will lead to better prediction accuracy. Menzies and Shepperd (2007) addressed the repeatable issues specifically in software engineering prediction. The conclusion drawn from the study is the algorithm instability which means the conclusion stands true to one project but does not hold in other projects. Mirkin et al (2006) proposed the nine strategies to reduce conclusion instability which again pose the algorithm instability issues.

Highlighting some hybrid approaches to deal with the similar problem of missing data, some of the most competitive results were drawn by Wasito and Mirkin (2006) who combined HDI, Likelihood Imputation with machine learning based imputation which is a hybrid model and gives a very competitive output. Song and Shepperd combine k-NN imputation with CMI which results in a MINI algorithm, Sentas and Angelis (2006) used Kernel based imputation which is a machine learning method with RI. There are some fairly new techniques being evaluated especially after 2008 afterward such as Genetic Algorithm based imputation, Kernel based, Multi-Layer Perceptron and Neural Network with the concern of empirical evaluation of estimation accuracy such as parameter, feature subset, and outlier presence (Menzies and Shepperd, 2007).

Manghirmalani et al (2015) in their study have already explored the implementation of the Back Propagation model to impute missing values. The results obtained in this study are a benchmark for this study and are tested and compared with a multi-Layer feed forward Perceptron model.

Data Set

The questionnaire for the present study was adapted from Birbal et al. (2018) study on learners' readiness for blended learning. The instrument consisted of 34 items that measured learners' attitudes towards six different aspects of blended learning: learning flexibility (4 items); online learning (8 items); study management (6 items); technology (4 items); classroom learning (5 items) and online interaction (7 items). The present study adopted a descriptive survey method for collecting data. The sample was collected by circulating 500 forms out of which the sample size selected was 313 teachers. Out of the total in-service teachers 221 were females and 92 were males. 187 forms were discarded due to data loss. As humans are involved in empirical research, data loss is normal. Minor loss can be accepted but loosing on major data can affect the validity of the study, hypothesis testing and also leads to fault in data accuracy. Unusual values are often obtained in dataset, and they can distort statistical analyses and violate assumptions. Some reasons noticed for data loss during the present research were that During data entry, weird values were produced. For example, the present questionnaire was scored on a 5-point rating scale, but few values against item were written as 41 or 51 which not only stand out but are also not apt. The probable reason for writing value 51 or 61 could be that the value for 2 proceedings items were written together and the other were left blank. Due to lack of time or under certain circumstances there are chances that the subject may skip certain items unconsciously or may be due to boredom or lack of time or unknowingly. This was also one of the reasons noticed while collecting BL data. Not all such forms can be discarded as they can contribute in capturing valuable information that is part of the interest study area, here in the present study, understanding various dimensions contributing the most towards BL was also given importance. Thus, if data loss occurs in single dimension, then the overall attitude is also affected. Thus, by proper application of machine learning techniques data loss can be avoided and results obtained can be more valid and reliable. Circulating forms again in order to collect the desired number of samples just because of few incomplete or missed items, it becomes time consuming and expensive on the part of the researcher, thus the following research question was taken into account for the present study.

Machine Learning Techniques For Imputing Missing Values

The aim of this study is to impute the missing values in the data and avoid any data loss. Having neural networks constitute a class of predictive modelling system that works by iterative parameter adjustment (Manghirmalani Mishra. P., 2017). This study compares two machine learning techniques; Multi-Layer Perceptron (MLP) Algorithm and Backpropagation (BP) algorithm for imputing the missing values. Both algorithms belong to the class of supervised learning that uses training data to classify the new or unseen instance. Both of these algorithms have their own advantages. MLP is considered good due to its simplicity, less time consumption while maintaining a good level of accuracy. On the other hand, BP achieves promising results due to its weighted error feed backward approach.

A. General Steps for Imputing Missing Values

Gupta & Lam (1998) introduced the following procedure for reconstruction of missing values using neural networks.

i. Collect all training cases without any missing value and call them the complete set.

ii. Collect all training and test cases with at least one missing value and call them the incomplete set.

iii. For each pattern of missing values, construct a multi-layered network with the number of input nodes in the input layer equal to the number of non-missing attributes, and the number of output nodes in the output layer equal to the number of missing attributes. Each input node is used to accept one non-missing attribute, and each output node to represent one missing attribute.

iv. Use the complete set and train each network constructed in Step iii. Since the complete set does not have missing values, different patterns of input-output pairs can be obtained from the complete set to satisfy the input-output requirements for different networks from Step iii. As the output of a network is between 0 and 1, data has to be converted to values between 0 and 1 for this reconstruction procedure. The trained networks from Step iv calculate the missing values in the incomplete set.

v. To construct such a Neural Network, a strategy is applied where one must start with a simple network and add extra nodes to the network until such addition does not improve the network performance. The system is implemented using Java.

B. Multi-Layer Feedforward Model

MLP model adjusts weights of hidden layers to reduce the error. This layer, based on weights (random numbers between 0-1) predicts the next value. The multilayer part is used to identify the error (which is verified using the existing complete dataset) and makes adjustments to give the most suitable output (Manghirmalani Mishra. P., 2015). The weight aspect of this network is adjusted (manually) multiple times until the network gives the highest possible accuracy. The number of weights depends upon the number of columns present in the data. Number of hidden layers depends upon the number of columns with missing values.

C. Multi-Layer Backpropagation Model

BP is a technique that applies to MLP to adjust the weights of hidden layers to reduce the error (Manghirmalani Mishra. P., 2017). It is named “back propagation” because it propagates the weighted error backward to hidden layers to update weight. For this it uses a generalised delta rule. The convergence of BP to local minima for error depends on the value of α (learning rate coefficient), too large α makes it difficult for networks to find the gradient (narrow peak) while too small α usually increases the chance of getting trapped into local maxima. The problem is solved using another term called momentum (β), which reduces the chance of getting stuck in local peaks as well as accelerates learning over smooth surfaces. By changing the values of α and β the network can be tuned to perform better.

Results

The performance of the various classifiers may be presented in terms of accuracy, correctness and coverage:

$$\text{accuracy} = \frac{\text{No. of cases correctly classified}}{\text{Number of Cases}} \times 100\%$$

$$\text{correctness} = \frac{\text{No. of cases correctly classified}}{\text{Number of cases classified}} \times 100\%$$

$$\text{coverage} = \frac{\text{No. of cases classified}}{\text{Number of cases}} \times 100\%$$

Accuracy may be measured on the training set or on a test set. A high figure for training set accuracy does not mean that the performance of the classifier will be good in practice. In this paper, accuracy results on the test set since these give a better indication of how well the classifier is able to impute the missing value on a new data entry.

A. Implementation

The system is implemented using Java. The experiments were conducted on a workstation with an Intel Core i7 CPU, 8 Ghz, 16 GB of RAM, running on Microsoft Windows 10 Home Edition.

B. Accuracy of Multi-Layer Feed Forward Perceptron Model

Table 2 represents the accuracy of MLP and its corresponding graph, Graph 1 shows the steady increment in the level of accuracy as the size of the training data base increases.

Detection Measure	Percentage
Accuracy	93%
Correctness	89%
Coverage	88%

Table 1: Accuracy of MLP

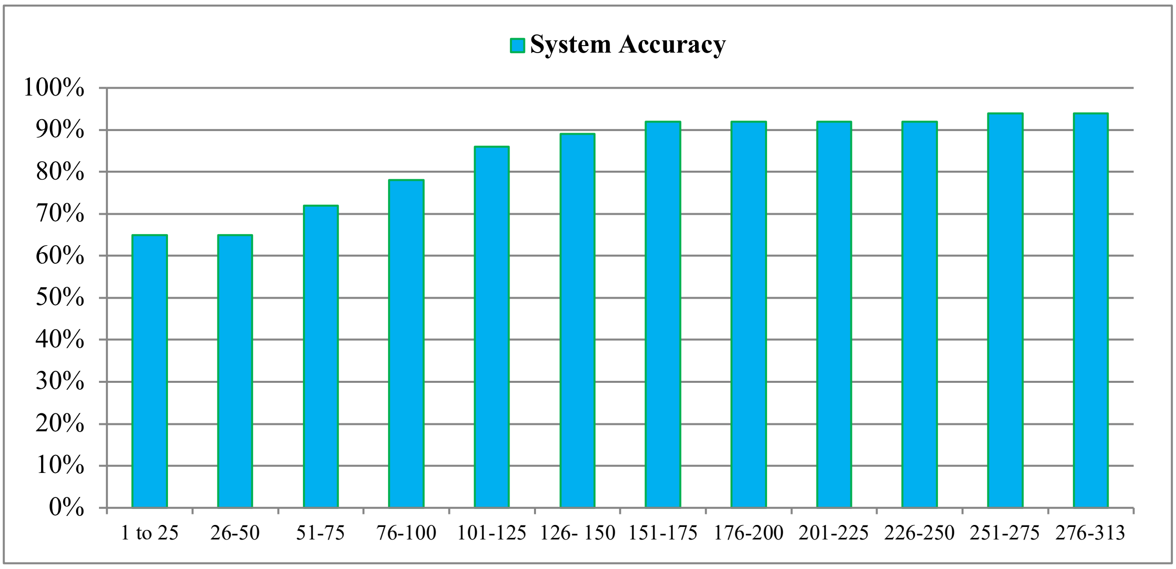

Figure 1: System Accuracy vs the Testing Samples for MLP where x-axis denotes the number of testing samples and y-axis denotes the accuracy in %

C. Accuracy of Multi-Layer Feed Backward Back Propagation Model

Table 3 represents the accuracy of BP and its corresponding graph, Graph 2 shows the steady increment in the level of accuracy as the size of the training data base increases.

Detection Measure	Percentage
Accuracy	97%
Correctness	96%
Coverage	96%

Table 2: Accuracy of BP

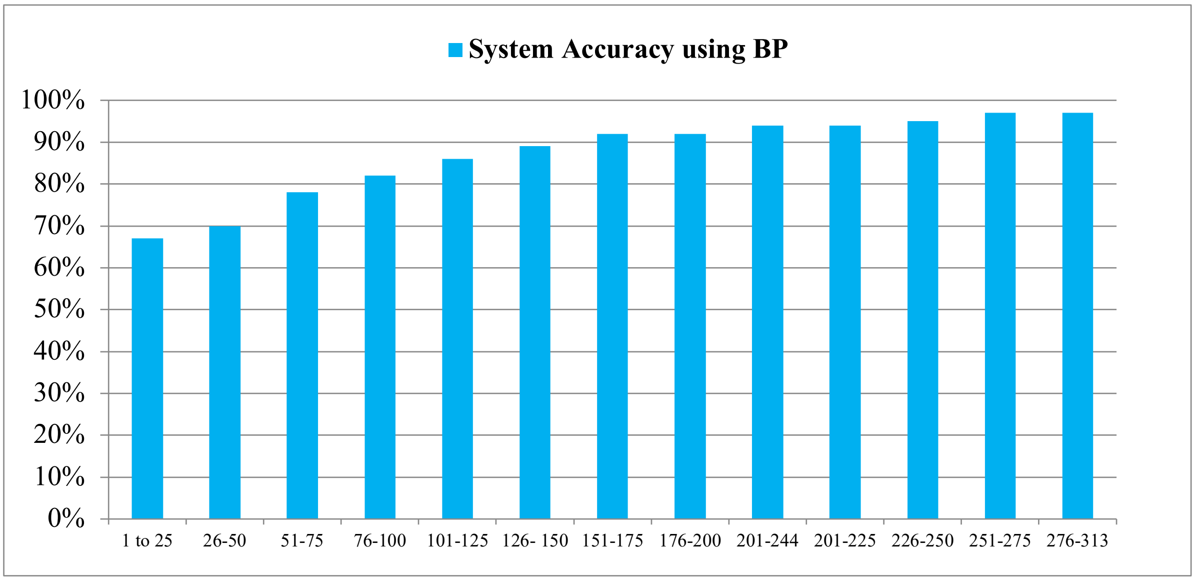

Figure 2: System Accuracy vs the Testing Samples for BP where x-axis denotes the number of testing samples and y-axis denotes the accuracy in %

Conclusion and Discussions

It could be concluded that for this study the BP network works more impressively. The predicted missing values derived here are of higher precision when compared with a feed-forward network. Though both the systems give accuracy of over 90%, it could be concluded that BP gives good coverage of the data with high accuracy and correctness. As we know that the Missing values in input data are a particular problem in the medical field and this phenomenon is, therefore, of particular consequence in medical diagnostic systems which are neural network based.

Though, in theoretical research we state that incomplete data should be discarded but this leads to waste of information and the efforts one takes to collect it in the first place. In medical data sets that are collected with the purpose of creating a diagnostic system which relates to local conditions, and where the data set is likely to be small, the loss may be critical. Hence, the proposition of such a blend of technology in educational research can be a critical turnover towards a multidisciplinary approach and can encourage more such ventures.

The proposed methodology can work for almost all types of databases and the prediction approach of the missing value can hold true for all forms of studies. The main focus of this work was, therefore, to produce systems which when presented with BL data containing missing values either (a) a feed forward network to predict and impute the missing values, or (b) feed-back network to predict, error check, feed backward and then impute the missing values. The performance of these systems was then compared with each other to obtain the highest precision algorithm.

Future Work

Proposed work for the future is to further increase the accuracy of the imputing missing value by trying the hybrid approach of machine learning. Once this is achieved, further enhancement in the data pre-processing could be achieved by reducing the number of attributes of sample study in order to reduce the data collection time and choose only the most important features out of the given attributes. This could be achieved by using pruning techniques of machine learning like decision trees. Combination of the above discussed could help the overall community of researchers irrespective of their field of research.

References

1. B. Twala, M. Cartwright., and M. Shepperd (2006). Ensemble of Missing Data Techniques to Improve Software Prediction Accuracy. ICSE.
2. B.Mirkin and I. Wasito (2006). "Nearest neighbours in least-squares data imputation algorithms with different missing patterns." Computational Statistics & Data Analysis 50: 926 – 949.
3. Birbal, R., Ramdas, M., and Harripaul, C. (2018, June). Student Teachers' Attitudes towards Blended Learning. Journal of Education and Human Development, 7(2), 9-26. doi:10.15640/jehd.v7n2a2

4. Debarati Roy. (n.d.). Blended learning — the new normal. Retrieved from Collins: https://collins.in/events/blog/blended-learning-the-new-normal/

5. Graham, C. R. (2013). Emerging practice and research in blended learning. In M. G. Moore (Ed.), Handbook of distance education, (3rd ed., pp. 333–350). New York: Routledge.

6. Gupta, A. & Lam, M., The weight decay Backpropagation for generalizations with missing values, Annals of Operations Research 78, pp. 165-187, 1998

7. J. Scheffer(2002). "Dealing with Missing Data." R.L.I.M.S 3.

8. Jamshidian, M. (2004). Strategies for analysis of incomplete data. In Handbook of data analysis (pp. 112-130). SAGE Publications, Ltd, https://dx.doi.org/10.4135/9781848608184

9. K. Molloken, and M. Jorgensen (2003). A Review of Surveys on Software Effort Estimation. Proceedings of the 2003 International Symposium on Empirical Software Engineering, IEEE Computer Society: 223.

10. Kang H. (2013). The prevention and handling of the missing data. Korean journal of anesthesiology, 64(5), 402–406. https://doi.org/10.4097/kjae.2013.64.5.402

11. M. Cartwright, B. Twala. and. Menzies. (2005). Ensemble Imputation Methods for Missing Software Engineering Data. METRICS

12. M. Jørgensen and M. Shepperd (2007). "A Systematic Review of Software Development Cost Estimation Studies." IEEE Transactions on Software Engineering 33.

13. Manghirmalani Mishra P., Kulkarni S.T, 2017, Attribute Reduction to Enhance Classifier's Performance- a LD Case Study, Journal of Applied Research DOI:10.15373/2249555X

14. Manghirmalani-Mishra P., Kulkarni S.T, Magre S., A Computational Based study for Diagnosing LD amongst Primary Students, National Conference on Revisiting Teacher Education; ISBN: 97-81-922534, 2015.

15. Norberg, A., Dziuban, C. D., & Moskal, P. D. (2011). A time-based blended learning model. On the Horizon, 19(3), 207–216. https://doi.org/10.1108/10748121111163913.

16. P. Sentas, and L. Angelis. (2006). "Categorical missing data imputation for software cost estimation by multinomial logistic regression." Journal of Systems and Software 79(3): 404-414.

17. Ross, B., & Gage, K. (2006). Global perspectives on blended learning: Insight from WebCT and our customers in higher education. In C. J. Bonk, & C. R. Graham (Eds.), Handbook of blended learning: Global perspectives, local designs, (pp. 155–168). San Francisco: Pfeiffer.

18. T. Menzies and M. Shepperd (2012). "Special issue on repeatable results in software engineering prediction." Empirical Software. Engineering. 17(1-2): 1-17.

19. UNESCO. (2021). Education: From disruption to recovery. Retrieved from https://en.unesco.org/covid19/educationresponse