COMPARATIVE ANALYSIS OF ACCURACY ON MISSING DATA USING MLP AND RBF METHOD

V.B. Kamble[1], S.N. Deshmukh[2]
[1]P.E.S. College of Engineering, Aurangabad. (M.S.) India.
[2]Dr. Babasaheb Ambedkar Marathwada University, Aurangabad. (M.S.), India

ABSTRACT
Missing data creates various problems in analyzing and processing data in databases. In this research paper we introduce a new method aimed at approximating missing data in a database using neural networks. Multi-Layer Perceptron (MLP) and Radial Basis Function (RBF) networks were employed to train the neural networks. Our focus also lies on the investigation of using the Multi-Layer Perceptron (MLP) and Radial Basis Function (RBF) in accurately predicting missing data as the number of missing cases within a single record increases. It is observed that there is no significant reduction in accuracy of results as the number of missing cases in a single record increases. This paper helps for to found out accuracy by using MLP and RBF method by using threshold value like 0.1, 0.15, 0.2, and 0.25. It is also found that results obtained using RBF method is superior to MLP method.

General Terms: Missing Data, Accuracy, Imputation Method, Missing Values, Dataset, Multi-Layer Perceptron (MLP), Radial Basis Function (RBF) etc.
Keywords: Dataset, Incomplete data, Missing Values, MLP, RBF etc.

1. INTRODUCTION

The serious quality problems that reduce the Performance of data mining algorithms are due to incomplete, redundant, inconsistent and noisy data. Missing data is a common issue in almost every real dataset. If the rate of missing is less than 1%, missing data it does not make any trouble for the discovery of Knowledge, 1-5% manageable, 5-15% requires sophisticated methods to handle and more than 15% may severely impact any kind of interpretation. Due to this reason, missing data has been an area of research in various disciplines for a quite long time.

The missing data treatment methods have to be carefully chosen based on missing mechanism. According to Little and Rubin [1], missing data can be divide the following three mechanisms:

i. Missing at Random

Data are said to be Missing At Random (MAR) when the probability that Y is missing depends on the set of other observed responses X, but is unrelated to the specific missing values.
Pr(Y is missing | X, Y) = Pr (Y|X) for all missing Y

ii. Missing completely at Random

Data are said to be Missing Completely At Random(MCAR) if the probability that Y is missing is unrelated to Y or other variables X (where X is a vector of observed variables).
Pr(Y is missing |X, Y) = Pr(Y is missing)

iii. Not missing at Random

Data are said to be Not Missing At Random (NMAR) when the probability that responses are missing depends on the specific missing values itself. NMAR is often referred as non-ignorable, non-random missingness. When missingness is non-ignorable, it means that future unobserved responses, conditional on past observed responses cannot be predicted.
The organization of the paper is Section1: Introduction Section 2: Literature Reviews, Section 3: Related Work Section 4: Missing Data Imputation Technique, Section 5: Methodology Section 6: Experimental result and Analysis, Section 7: Conclusions

2. Literature Reviews

Missing value imputation is challenging issue in data mining. Missing value may generate bias result and affect the quality of the data mining task. But in practice a suitable method for dealing missing values is necessary. Missing values may appear either in conditional attributes and in class attribute. There are many approaches to deal with missing values described.

First approach usually lost more useful information, whereas the second one is time-consuming and expensive, so it is not suitable in many applications. Third approach assumes that all missing values are same value, probably leading to considerable distortions in data distribution. The method of imputation is a popular strategy in comparison with other methods; it uses as more information as possible from the observed data to predict missing values [14, 17]. Missing value imputation techniques can be classified into parametric imputation and non -parametric imputation. The parametric regression imputation is applicable if a dataset can be adequately modeled parametrically and users can correctly specify the parametric forms for the dataset [16, 17].

Non-parametric imputation can provide superior fit by capturing structure in the dataset, offers a nice alternative if users have no idea about the actual distribution of a dataset [7]. Statistical methods consist of linear regression, replacement under same standard deviation, and mean/mode method. But these methods are not completely satisfactory ways to handle missing value problems [5].

3. Related Work

3.1 Treatment of Missing Data

A) Ignoring Tuple: This method determine the missing data on each instance and delete instances and remove the entire attribute which having high percentage of missing data in the dataset. This method is applicable only when the dataset is MCAR pattern.

B) Parameter Estimation: This method is used to find out parameters for the complete data. This method uses the Expectation and maximization algorithm for handling the parameter estimation of the missing data.

C) Imputation Technique: This technique is very popular and frequently used in which replaces the missing values based on estimated values. Imputation techniques are classified like 1. Mean Imputation 2. Mode Imputation, 3.Median Imputation 4. Standard Deviation Imputation etc. [6].

3.2 Missing Value Imputation

To impute missing value following techniques are used:

A*) Parametric Regression Imputation Methods:* The parametric methods [18], [19],[20], are superior while the data set are adequately modeled but in real applications it is impossible to know the distribution of the data set. Therefore, the parametric estimators can lead to highly bias, and the optimal control factor settings may be miscalculated [18, 19, 20].

B) *Non Parametric Regression Imputation Methods:* By using The Nonparametric imputation method. we know the structure of the data set. These imputation methods are designed for either continuous or discrete independent attributes. And these estimators cannot handle discrete attributes efficiently. Some methods, such as C4.5 algorithm [13], association rule based method [16], and rough set based method are designed to deal with only discrete attributes. In these algorithms, continuous attributes are always makes discrete before imputing. This possibly leads to a loss of useful information of the continuous attributes [17].

4. Multi-Layer Perceptron

Multi-Layer Perceptron (MLP) is a popular architecture used in ANN. The MLP can be trained by a back propagation algorithm. Typically, the MLP is organized as a set of interconnected layers of artificial neurons, input, hidden and output layers. When a neural group is provided with data through the input layer, the neurons in this first layer propagate the weighted data and randomly selected bias through the hidden layers. Once the net sum at a hidden node is determined, an output response is provided at the node using a transfer function. Two important characteristics of the MLP are its non- linear processing elements which have a non-linear activation function that must be smooth (the logistic function and the hyperbolic tangent are the most widely used) and its

massive interconnectivity (*i.e.* any element of a given layer feeds all the elements of the next layer)[9]. The structure of a Basic MLP (Multilayer Perceptron) Structure in **Figure1.**

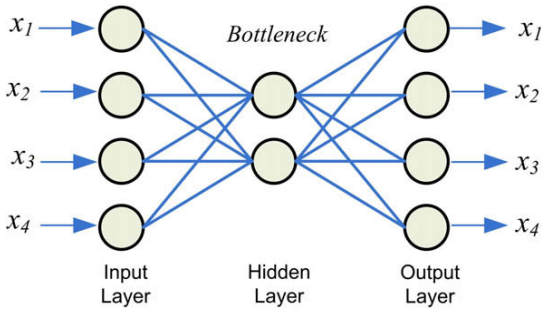

Figure1. Basic MLP(Multilayer Perceptron) architecture

4.1 Radial Basis Function

The Radial Basis Function (RBF) is another popular architecture used in ANN. The RBF, which is multilayer and feed-forward, is often used for strict interpolation in multi-dimensional space. The term "feed-forward" means that the neurons are organized as layers in a layered neural network. The basic architecture of a three-layered (RBF) neural network is shown in **Figure 2**.

The RBF network comprises three layers, *i.e.* input, hidden and output. The input layer is composed of input data. The hidden layer transforms the data from the input space to the hidden space using a non-linear function. The output layer, which is linear, yields the response of the network. The argument of the activation function of each hidden unit in an RBF network computes the Euclidean distance between the input vector and the center of that unit [7].

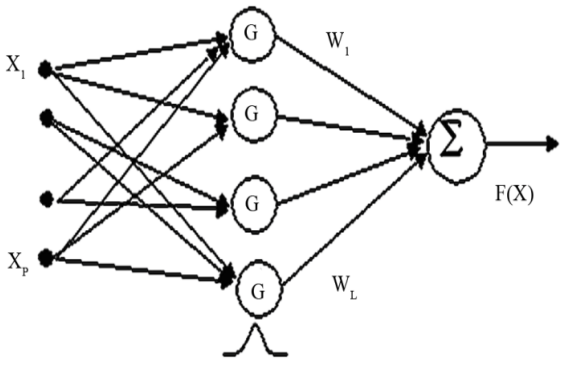

Figure 2. Basic RBF (Radial Basis Function) architecture

5. Methodology

5.1 Multi-Layer Perceptrons (MLP)

MLP neural networks consist of multiple layers of computational units, usually interconnected in a feed-forward way [7, 8]. A fully connected two layered MLP architecture was used in the experiment. Each neuron in one layer is directly connected to the neurons of the subsequent layer.

A two-layered MLP architecture was used because it had better results which state that a two layered architecture is adequate for MLP [9].

Figure 1 shown the architecture of the MLP used in this paper. The parameters (number of neurons, training cycles, activation function) used to train the neural network were chosen after training the neural network with different parameters. Combination of parameters that gave better results was selected for training the actual neural network. The MLP network contains inputs layer, a hidden layer and the output layer.. A linear activation function was used, as it gave better results.

5.2 Radial-Basis Function (RBF)

RBF networks are feed-forward networks trained using a supervised training algorithm [7]. They are typically configured with a single hidden layer of units whose activation function is selected from a class of functions called basis functions. A fully connected two layered RBF architecture was used in the experiment. Each neuron in one layer is directly connected to the neurons of the subsequent layer. The network has inputs, Hidden layer and output units. Gaussian Function was used as hidden unit activation function. The RBF network used in this research is depicted in Figure 1. Figures2 represent the non-linear activation functions. [10]

By implementing Multilayer Perceptron (MLP) and Radial Basis Function on the dataset and to find out predicted missing values at the respective feature from the Five Thousand Dataset, Ten Thousand Dataset, Fifteen Thousand Dataset and Twenty Thousand Dataset respectively. And also find out the Accuracy by combining the MLP and RBF Method.

6. Experimental Result and Analysis

6.1Dataset Used

In this work dataset having characteristics is given below.

Number of Instances: 5000, 10,000, 15,000 and 20,000

Number of Attributes: 05

(Record No., M1, ECE, EM, EE)

Dataset contains marks of four different subjects of engineering college. In dataset randomly distributed the missing values in each attribute to become the incomplete dataset. Record. No. in the Dataset is used are imaginary and generated for the data analysis purpose in data mining process. In dataset M1, ECE, EM, EE are the subject in any engineering college and class test marks for each subject is out of twenty marks for each subject repectively.The structure of Dataset as shown in the Table No.1

Table No.1 Dataset

Record. No.	Subject 1	Subject 2	Subject 3	Subject 4
01	X1	X2	X3	X4
..	..	..	..	..
N	XN	XN	XN	XN

6.2 Accuracy by using Multi-Layer Perceptrons (MLP) Method.

In dataset missing data randomly distributed for each attribute/column. Implement Multi-Layer Perceptrons (MLP) Method on dataset and found out corresponding missing value for each attribute/column. After finding the missing value for each attribute find out the accuracy by using threshold value like 0.1, 0.15, 0.2 and 0.25. The result shown in following Table 2.

Table No. 2 Accuracy by using MLP Method

Accuracy/Threshold Value	0.1	0.15	0.2	0.25
Five Thousand Dataset	51.72	55.17	58.62	68.96
Ten Thousand Dataset	52.94	56.86	62.74	78.43
Fifteen Thousand Dataset	54.70	57.74	70.42	80.64
Twenty Thousand Dataset	61.29	65.59	73.11	81.69

As per result shown in table No. 2 It was observed that when threshold value is increase then accuracy also increase gradually.

6.3 Graphical representation of Accuracy by using MLP Method.

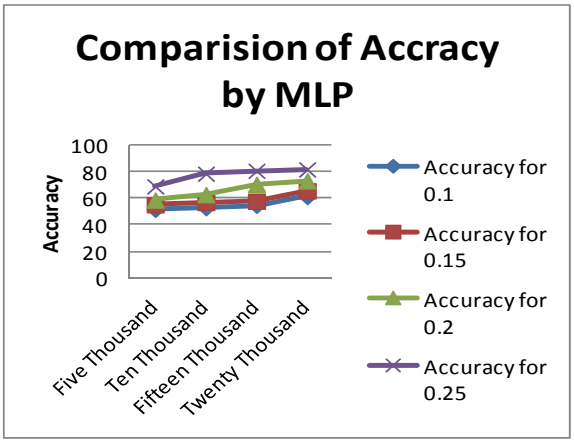

Figure 2. Graphical Representation of Accuracy by Using MLP.

6.4 Accuracy by using Radial-Basis Function (RBF) Method.

In dataset missing data randomly distributed for each attribute/column. Implement Radial-Basis Function (RBF) Method on dataset and found out corresponding missing value for each attribute/column. After finding the missing value for each attribute find out the accuracy by using threshold value like 0.1, 0.15, 0.2 and 0.25. The result shown in following Table 3.

Table No. 3 Accuracy by using RBF Method

Accuracy/Threshold Value	0.1	0.15	0.2	0.25
Five Thousand Dataset	58.62	62.06	65.17	75.86
Ten Thousand Dataset	59.90	63.78	66.66	82.35

Fifteen Thousand Dataset	61.52	65.56	73.23	82.79
Twenty Thousand Dataset	64.51	67.74	75.26	83.09

As per result shown in table No. 3 It was observed that when threshold value is increase stepwise up to 0.25 then accuracy also increase gradually.

6.5 Graphical representation of Accuracy by using RBF Method.

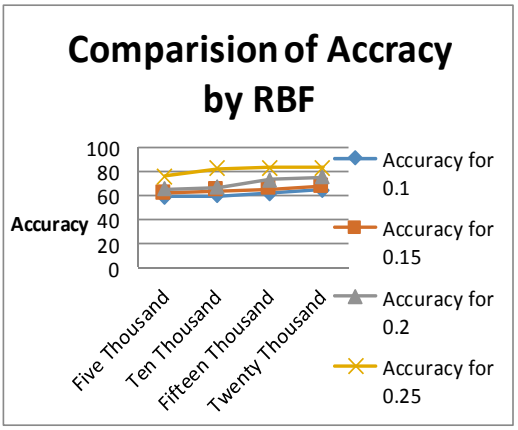

Figure 3. Graphical Representation of Accuracy by Using RBF.

7. CONCLUSIONS

To handle the missing value in the dataset accuracy for RBF Method is more suitable than accuracy for MLP Method. To implement RBF and MLP Method use of threshold value like 0.1, 0.15, 0.2 and 0.25 etc. As per observation from experimental result shown that Accuracy for RBF method is more superior to the Accuracy for MLP method.

Hence for handling missing values MLP and RBF method is more suitable to handle the incomplete dataset. It is observed that accuracy obtained from RBF method is more as compares to the MLP method.

This research paper helps how to handle missing data from the numerical dataset. In future work it is demanded to work on to handle the missing data from categorical dataset and also on image dataset for doing data analysis work to retrieve the useful and needed information from the missing dataset so as to take policy decision.

8. REFERENCES

[1].S.Kanchana, Dr. Antony Selvadoss Thanaman,“ Classification of Efficient Imputation Method for Analyzing Missing Values”, International Journal of Computer Trends and Technology (IJCTT) – volume 12 number 4 – Jun 2014

[2].Minakshi , Dr. Rajan Vohra, Gimpy, “ Missing Value Imputation in Multi Attribute Data Set”,(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5315-5321.

[3].Alireza Farhangfara , Lukasz Kurganb , Witold Pedrycz “Experimental analysis of methods for imputation of missing values in databases”

[4].Dinesh J. Prajapati ,Jagruti H. Prajapati, “Handling Missing Values: Application to University Data Set” , International journal of emerging trends in Engineering and Development Issue1, Vol 1August2011

[5].R.S. Somasundaram, R. Nedunchezhian, “Evaluation of three Simple Imputation Methods for Enhancing Preprocessing of Data with Missing Values”, International Journal of Computer Applications (0975 – 8887) Volume 21– No.10, May 2011

[6]. Santosh Dane, Dr. R. C. Thool, “Imputation Method for Missing Value Estimation of Mixed-Attribute Data Sets”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 5, May 2013, ISSN: 2277 128X

[7].Ms.R.Malarvizhi, Dr. Antony Selvadoss Thanamani, “ Comparison of Imputation Techniques after Classifying the Dataset Using KNN Classifier for the Imputation of Missing Data”, International Journal Of Computational Engineering Research (ijceronline.com) Vol. 3 Issue. 1, ISSN No. 2250-3005(online),January 2013

[8].Arnaud Ragel, Bruno Cremilleux & J. L. Bosson, “An Interactive and Understandable Method to Treat Missing Values: Application to a Medical Data Set”, In ACM Comput. Surv. 1985.

[9].Shalu Bhati, Manoj Kumar Gupta, "Missing Data Imputation for Medical Database: Review", international Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 6, Issue 4, April 2016.

[10] Mussa Abdella, Tshilidzi Marwala, "The Use Of Genetic Algorithms And Neural Networks To Approximate Missing Data In Database" Computing and Informatics, Vol. 24, 2005, 577–589

[11]. Luai Al Shalabi, "A comparative study of techniques to deal with missing data in data sets", In Proceedings of the 4th International Multiconference on Computer Science and Information Technology /CSIT 2006.

[12]. A. Pujari, Data Mining Techniques, Universities Press, India, 2001.

[13]. Ragel, A. and Cremilleux, B., "MVC - a preprocessing method to deal with missing values", In Proceedings of Knowl.-Based Syst 1999, 285-291.

[14].Chih-Hung Wu, Chian-Huei Wun, Hung-Ju Chou, "Using Association Rules for Completing Missing Data," Fourth International Conference on Hybrid Intelligent Systems (HIS'04), 2004 pp.236-241.

[15]. Lakshminarayan, K., Harp, S., Goldman, R., and Samad, T. 1996, "Imputation of missing data using machine learning techniques", In Proceedings of the Second International Conference on Knowledge Discovery in Databases and Data Mining
.

[16]. Zhang, S.C., et al., "Information Enhancement for Data Mining", IEEE Intelligent Systems, 2004, Vol. 19(2): 12-13, (2004).

[17].Qin, Y.S., "Semi-parametric Optimization for Missing Data Imputation". Applied Intelligence, 2007, 27(1): 79-88.

[18] J.R. Quinlan, "C4.5: Programs for Machine Learning. Morgan Kaufmann", 1993.

[19].Q.H. Wang and R. Rao, "Empirical Likelihood-Based Inference under Imputation for Missing Response Data," Annals of Statistics, vol. 30, pp. 896-924, 2002.

[20].S.C. Zhang, "Par imputation: From Imputation and Null-Imputation to Partially Imputation", IEEE Intelligent Informatics Bull., vol. 9, no. 1, pp. 32-38, Nov. 2008.