Global Journal of Engineering Science and Research Management

A NOVEL HYBRID APPROACH FOR PREDICTION OF MISSING VALUES IN NUMERIC DATASET

V.B.Kamble*[1], S.N.Deshmukh[2]
*[1]Department of Computer Science and Engineering, P.E.S. College of Engineering, Aurangabad. India.

KEYWORDS: Classifier, Incomplete data ,DM(Missing Dataset), DI(Imputed Dataset), Missing Values, Imputation Technique, Absolute Error, Relative Error, Co-Relation coefficient and Root Mean Squared Error, weka Tool.

ABSTRACT
Issue of incomplete data exists across the entire field of data mining. In this paper, a proposed method with Mean Imputation, Mode Imputation and Median Imputation is used to handle the missing value in the dataset. In this work student class test dataset were used of engineering college. By comparing the Absolute Error, Relative Error, Co-Relation coefficient and Root Mean Squared Error of the Proposed Method with imputation techniques and Classifier applied on the imputed dataset by using Weka Tool. Proposed Method with Mean Imputation was performing efficient result with compares with classifier applied on the dataset.

INTRODUCTION
Missing data is the absence of data items; they hide some information that may be important. In practice, missing data affecting data quality. The presence of missing data is a general and challenging problem in the data analysis field. Fortunately, missing data imputation techniques can be used to improve the data quality. Missing data imputation techniques refer to any strategy that fills missing values of a dataset so that standard data analysis methods can be applied to analyze complete dataset [1]. Information quality is important to organization. People use information attribute as a tool for accessing information quality. Information quality is measured based on users as well as experts opinion on the information attributes. The commonly known information attributes for information quality including accuracy, objectivity, reputation, access, security, relevancy, value added, timeliness, completeness, amount of data, and ease of understanding and consistent representation etc. The boundary between data and information can never be unambiguous, these attribute can also be applicable to data quality. Commonly, one can rarely find a data set that contains complete entries [2].

Related Work: Methods for dealing with missing value can be classified into three categories 1) Case Deletion, 2) Learning Without Handling Missing value, and 3) Missing Value Imputation. The Case Deletion is to simply omit those cases with missing value and only to use the remaining instances to finish the learning assignments. Second approach is to learn without handling missing data. Missing data imputation methods advocates filling missing values before a learning application. Missing data imputation is a procedure that replaces the missing value with some possible values [3].

Missing Data Handling: Several methods have been applied in data mining to handle missing value in database. Data with missing value could be ignored, or a global constant could be used to fill missing value, such as attribute mean, attribute mean of the same class, or an algorithm could be applied to find the missing values. Missing data imputation techniques means a strategy to fill missing value of a dataset in order to apply the standard methods which require complete data set for analysis. These techniques retain data in incomplete cases, as well as impute values of correlated variables.

Missing data imputations techniques are classified as ignorable missing data imputations methods, which include single imputation methods and multiple imputation methods, and non-ignorable missing data imputations methods which include likelihood based methods and the non-likelihood based methods. Single imputation methods could fill one value for each missing values and it is more commonly used at present than multiple imputations which replace each missing value with several possible values and better reflects sampling variability about actual value [4].

Types of Classifier: There is a number of different classifier used for data analysis. The classifier used in the data analysis is 1) Linear Regression 2) K Star 3) Multi-Layer Perceptron 4) IBK (K-Nearest Neighbor) 5) Radial Basis Function. This classifier applied on the student class test dataset with help of Weka tool.

A) K star

The K* algorithm can be defined as a method of cluster analysis which mainly focus on partition of n observation into k clusters in which each observation belongs to the cluster with the nearest mean. K* algorithm as an instance based which uses entropy as a distance measure. The benefits are it provides a consistent approach to handling of real valued attributes, symbolic attributes and missing values. K* is a simple, instance based classifier, similar to K Nearest Neighbor (K-NN). New data instances, *x*, are assigned to the class that occurs most frequently amongst the k-nearest data points, y, where j = 1, 2…k. Entropic distance is then used to retrieve the most similar instances from the data set. By means of entropic distance as a metric has a number of benefits including handling of real valued attributes and missing values.

The K* function can be calculated as:

$$K^*(y_i, x) = -\ln P^*(y_i, x) \quad (1)$$

Where *P** is the probability of all transformational paths from instance *x* to *y*. It can be useful to understand this as the probability that *x* will arrive at *y*.

B) IBK (K - Nearest Neighbor)

IBK is a k-nearest neighbor classifier that uses distance metric. The number of nearest neighbors can be specified explicitly in the object focus to an upper limit given by the specified value. IBK is a nearest neighbor classifier. Different types of search algorithms can be used to speed up the task of finding the nearest neighbors. A linear search is the default but further options like KD-trees, ball trees, and so-called cover trees etc.

The distance function used is a parameter of the search method for example 1. Euclidean distance 2.Chebyshev Manhattan 3. Murkowski distance. Predictions from more than one neighbor can be weighted according to the distance from the test instance and two different formula are implemented for converting the distance into a weight [5].

C) Linear Regression

Linear Regression is most commonly used technique for determining relation between a scalar dependent variable y and one or more explanatory variables denoted X. The case of explanatory variable is called simple linear regression. For more than one explanatory variable is called as Multiple Linear Regression [2].

Linear Regression Model has many Practical uses.

1. To describe the linear dependence of one variable on another.
2. To predict values of one variable from values of another, for which more data are available.
3. To correct for the linear dependence of one variable on another, in order to clarify other features of its variability.

D) 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). The structure of a four input, four output auto encoder shown in **Figure1.**

ISSN 2349-4506
Impact Factor: 2.265

Global Journal of Engineering Science and Research Management

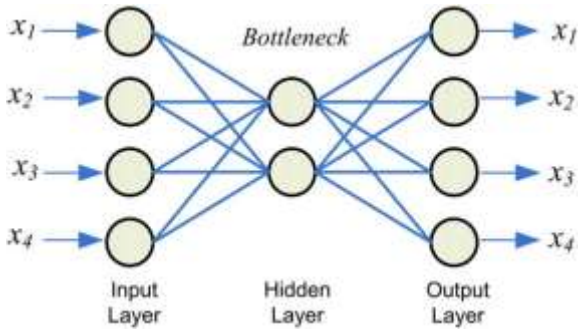

Figure1. The structure of a four-input, four-output auto-encoder

E) 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 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].

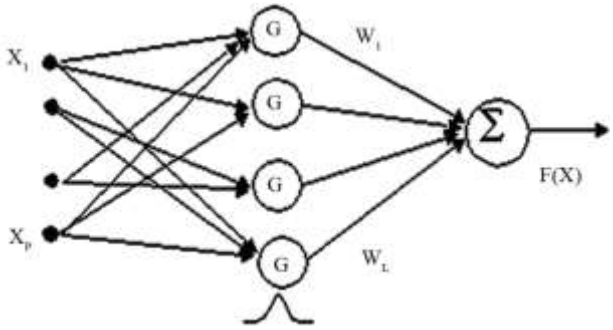

Figure 2. Basic RBF architecture

MATERIALS AND METHODS

In this work student class test dataset is used of engineering college. In the dataset some missing values are present. We developed a Proposed Method with combination of Mean Imputation, Mode Imputation and Median Imputation for handling the missing value in the dataset. We calculate the Absolute Error, Relative Error, Co-Relation coefficient and Root Mean Squared Error of the Proposed Method.

Proposed method consists of following step.

1. Generation of Missing data in Dataset, DM
2. Application of Imputation Method on Missing Dataset, DM
3. Imputation of Missing Values at Respective Feature
4. To find Mean, Mode, Median and SD from Imputed Dataset, DI
5. To Calculate Average of all Mean, Mode, Median and SD from DI Dataset Horizontally
6. .Multiply Average of all Mean, Mode, Median and SD from Imputed Dataset, DI with Mean, Mode, Median and SD from Missing Dataset, DM
7. It Divided by Original Values of Respective Feature to Find the Predicted Values

By using different types of classifiers using Weka Tool apply on student class test dataset with Mean Imputation dataset, Mode Imputation dataset and Median Imputation dataset getting values of the Absolute Error, Relative Error, Co-Relation coefficient and Root Mean Squared Error of the dataset.

RESULTS AND DISCUSSION

4.1 By using imputation method like Mean Imputation, Mode Imputation and Median Imputation find out different types errors with proposed method. As per experiment result show that Mean Imputation Method with Proposed Method performance is more as compare to other methods.

Table 1. Comparison of Errors

Error	Mean Imputation	Mode Imputation	Median Imputation
	Proposed Method	Proposed Method	Proposed Method
Absolute Error	0.003721	0.008	0.0058
Relative Absolute Error	0.0004	0.000826	0.0006
Co-Relation Coefficient	0.31118332	0.82160087	0.631659835
RMSE	1.680183661	1.95065218	1.651157889

4.2 By using the Mean Imputation dataset of student class test record apply different classifiers using Weka tool getting the result as per shown in the table and compare it with the errors of proposed method with imputation method. It shown that proposed method with imputation method well perform

Table 2. Errors of Mean Imputation

Error/Classifier	K-Star	Linear Regression	MLP	RBF	IBK
Absolute Error	3.1336	3.4679	4.8738	4.3353	4.4266
Relative Absolute Error	88.1794	97.5878	111.2733	99.881	79.5683
Co-Relation Coefficient	0.4077	0.1398	0.1366	0.0235	0.471
RMSE	3.9605	4.2938	4.8738	4.3353	4.4266

4.3 Similar way By using the Median Imputation dataset of student class test record apply different classifiers using Weka tool getting the result as per shown in the table and compare it with the errors of proposed method with imputation method. It shown that proposed method with imputation method well perform

Table 3. Errors of Median Imputation

Error/Classifier	K-Star	Linear Regression	MLP	RBF	IBK
Absolute Error	3.1368	3.4695	3.8339	3.5495	2.8432
Relative Absolute Error	88.2644	97.6266	107.8796	99.8754	80.0025
Co-Relation Coefficient	0.4073	0.1388	0.1524	0.0231	0.4684
RMSE	3.962	4.2951	4.7574	4.3361	4.4415

4.4 Again by taking the Mode Imputation dataset of student class test record apply different classifiers using Weka tool getting the result as per shown in the table and compare it with the errors of proposed method with imputation method. It shown that proposed method with imputation method well perform

Table 4. Errors of Mode Imputation

Error/Classifier	K-Star	Linear Regression	MLP	RBF	IBK
Absolute Error	3.1367	3.4678	3.8318	3.548	2.861
Relative Absolute Error	88.2897	97.6093	107.8535	99.8669	80.5294
Co-Relation Coefficient	0.407	0.1394	0.1533	0.0244	0.4639
RMSE	3.9601	4.2922	4.7588	4.3333	4.5553

CONCLUSION

Missing values are regarded as serious problems in most of the information systems due to unavailability of data and must be impute before the dataset is used. To handle these missing values we developed proposed method with imputation methods and by comparing with the different classifiers are applied using Weka Tool on the imputed dataset like Mean Imputation dataset, Median Imputation dataset and Mode Imputation dataset. Conclude that Proposed method with Imputation perform efficient result as compare with the classifiers applied on Imputed dataset using Weka Tool.

The proposed work handles missing values only for the numerical attributes. Further it can be extended to handle a categorical attribute. Different classification algorithm can be used for comparative analysis of missing data techniques. Missing data techniques can also be implemented in mat lab.

ACKNOWLEDGEMENTS

The authors wish to thank the Management and Principal of P.E.S.College of Engineering, Nagsenvana, and Aurangabad-431001 for providing resources and support for pursuing research. Also the author wishes to thank Data Analytics Laboratory of Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad.

REFERENCES

1. Dinesh J. Prajapati, Jagruti H. Prajapati," Handling Missing Values: Application to University Data set". Issue 1, Vol. 1(August-2011), ISSN 2249-6149
2. Shamsher Singh, Prof. Jagdish Prasad, "Estimation of Missing Values in the Data Mining and comparison of Imputation Methods". Mathematical Journal of Interdisciplinary Sciences Vol. 1, Issue 1,March 2013,pp. 75–90
3. Xiao Feng Zhu, Shichao Zhang, Senior Member, IEEE, Zhi Jin, Senior Member, IEEE, Zili Zhang, and Zhuoming Xu, "Missing Value Estimation for Mixed-Attribute Data Sets". IEEE Transactions On Knowledge And Data Engineering, Vol. 23, No. 1, January 2011
4. T.R.Sivapriya, V. Thavavel, A.R.Nadira Banu Kamal, "Imputation And classification of Missing Data Using Least Square Support Vector Machines- A New Approach in Dementia Diagnosis". International Journal of Advanced Research in Artificial Intelligence, Vol.1, No.4, 2012
5. Ms S. Vijayarani, Ms M. Muthulakshmi "Comparative Analysis of Bayes and Lazy Classification Algorithms" International Journal of Advanced Research in Computer and Communication engineering,Vol. 2, Issue 8, August 2013 ISSN (Print) : 2319-5940, ISSN (Online) : 2278-1021
6. Mr.M.B.Shelke, Mr.K.B.Badade"Processing of Incomplete Data Sets: Prediction of Missing Values by using Multiple Regression" international Journal of Computer and Electronics Research. Volume 2, Issue 5, October 2013
7. Hadi Memarian, Siva Kumar Balasundram" Comparison between Multi-Layer Perceptron and Radial Basis Function Networks for Sediment Load Estimation in a Tropical Watershed" Journal of Water Resource and Protection, 2012, 4, 870-876
8. Yann-Yann Shieh, "Imputation Methods on General Linear Mixed Models Of Longitudinal Studies", American Institutes For Research
9. Edgar Acu˜Na1 and Caroline Rodriguez, "The Treatment Of Missing Values And Its Effect In The Classifier Accuracy Studies In Classification", Data Analysis, And Knowledge Organization, 2004, Springer.Com
10. Anjana Sharma, Naina Mehta, Iti Sharma, "Reasoning With Missing Values in Multi Attribute Datasets". International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue5, May 2013 ISSN: 2277 128X.
11. V.B.Kamble, S.N.Deshmukh," Comparison of Percentage Error by using Imputation Method On Mid Term Examination Data", International Journal of Innovations in Engineering Research and Technology (IJIERT),Impact Factor 2.766, Volume 2, Issue 12,2015
12. V.B.Kamble, S.N.Deshmukh,"Comparative Analysis Of Standard Error Using Imputation Method", International Conference on Innovations and Technological Developments in Computer, Electronics and Mechanical Engineering, 28-29, December 2015, VACOE Ahmednagar.