Comparison of Imputation Techniques after Classifying the Dataset Using Knn Classifier for the Imputation of Missing Data

Ms.R.Malarvizhi[1], Dr. Antony Selvadoss Thanamani[2]

[1]Research Scholar ,Research Department of Computer Science, NGM College, 90 Palghat Road,Pollachi, ,Bharathiyar.
[2]Professor and HOD, Research Department of Computer Science, NGM College 90 Palghat Road, Pollachi Bharathiyar University, Coimbatore.

Abstract:

Missing data has to be imputed by using the techniques available. In this paper four imputation techniques are compared in the datasets grouped by using k-nn classifier. The results are compared in terms of percentage of accuracy. The imputation techniques Mean Substitution and Standard Deviation show better results than Linear Regression and Median Substitution.

Keywords: Mean Substitution, Median Substitution, Linear Regression, Standard Deviation, k-nn Classifier.

1. Introduction

Missing Data Imputation involves imputation of missing data from the available data. Improper imputation produces bias result. Therefore proper attention is needed to impute the missing values. Imputation techniques help to impute the missing value. The accuracy can be measured in terms of percentage. Pre-processing has to be done before imputing the values using imputation techniques. kNN classifier helps to classify the datasets into several groups by using the given training dataset. The imputation techniques are separately imputed in each dataset and checked for accuracy. The results are then compared.

2. Missing Data Mechanisms

In statistical analysis, data-values in a data set are missing completely at random (MCAR) if the events that lead to any particular data-item being missing are independent both of observable variables and of unobservable parameters of interest. Missing at random (MAR) is the alternative, suggesting that what caused the data to be missing does not depend upon the missing data itself. Not missing at random (NMAR) is data that is missing for a specific reason.

3. Imputation Methods

Imputation is the process of replacing imputed values from the available data. There are some supervised and unsupervised imputation techniques. The imputation techniques like Listwise Deletion and Pairwise Deletion deletes the entire row. The motivation of this paper is to impute missing values rather than deletion. In this paper, unsupervised imputation techniques are used and the results are compared in terms of percentage of accuracy.

3.1. Mean Substitution

Mean Substitution substitutes the mean value of data available. It can be calculated from the available data. Mean can be calculated by using the formula

$$\bar{X} = \frac{\Sigma X}{N}$$

Where:

$\bar{X}$ is the symbol for the mean.
Σ is the symbol for summation.
X is the symbol for the scores.
N is the symbol for the number of scores.

3.2. Median Substitution

Median Substitution is calculated by grouping up of data and finding average for the data. Median can be calculated by using the formula

Median =L+h/f(n/2-c)

where
L is the lower class boundary of median class
h is the size of median class i.e. difference between upper and lower class boundaries of median class
f is the frequency of median class

c is previous cumulative frequency of the median class
n/2 is total no. of observations divided by 2

3.3. Standard Deviation

The standard deviation measures the spread of the data about the mean value. It is useful in comparing sets of data which may have the same mean but a different range. The Standard Deviation is given by the formula

$$s_N = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}.$$

where $\{x_1, x_2, \ldots, x_N\}$ are the observed values of the sample items and $\overline{x}$ is the mean value of these observations, while the denominator N stands for the size of the sample.

3.4. Linear Regression

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. Linear regression can be calculated by using the formula

$$\boldsymbol{Y = a + bX},$$

where $\boldsymbol{X}$ is the explanatory variable and $\boldsymbol{Y}$ is the dependent variable. The slope of the line is $\boldsymbol{b}$, and $\boldsymbol{a}$ is the intercept.

4. Database and Data Pre-Processing

The dataset consists of 10000 records which is taken from a UK census report conducted in the year 2001. It has 6 variables. The data has to be pre-processed before it is used for experiment. The original dataset is modified by making some data as missed. The missing percentage may vary as 2, 5, 10, 15, and 20 percentages. Next data is classified into several groups. For grouping of data, knn classifier is used. Each group is then separately taken for the experiment. The imputation techniques are implemented one by one and the performance is measured by comparing with original database in terms of accuracy.

5. K-Nearest Neighbor Classifiers

The idea in k-Nearest Neighbor methods is to identify k samples in the training set whose independent variables x are similar to u, and to use these k samples to classify this new sample into a class, v. f is a smooth function, a reasonable idea is to look for samples in our training data that are near it (in terms of the independent variables) and then to compute v from the values of y for these samples. The distance or dissimilarity measure can be computed between samples by measuring distance using Euclidean distance.

The simplest case is k = 1 where we find the sample in the training set that is closest (the nearest neighbor) to u and set v = y where y is the class of the nearest neighbouring sample. In k-NN the nearest k neighbors of u is calculated and then use a majority decision rule to classify the new sample. The advantage is that higher values of k provide smoothing that reduces the risk of over-fitting due to noise in the training data. In typical applications k is in units or tens rather than in hundreds or thousands.

6. Experimental Analysis

In our research the database is grouped into 3, 6 and 9 groups for experiment. The above said imputation techniques are implemented in each group. The missing percentages were 2, 5, 10, 15 and 20. The imputation techniques are Mean Substitution, Median Substitution, Standard Deviation and Linear Regression.

Table 1 describes the percentage of accuracy for various groups. For accuracy each imputed data set is compared with the original dataset. Linear Regression and Mean Substitution shows same result as well as poor result. Median Substitution and Standard Deviation shows same result as well as better result. When the group size is large there is some improvement in the result.

Table 2 describes the overall average of accuracy.

Table 1. Comparison of Imputation Techniques in Groups Classified Using k-NN Classifier

Percentage of Missing	3 GROUPS				6 GROUPS				9 GROUPS			
	Mean Sub	Med Sub	Std Dev	Linear Regress	Mean Sub	Med Sub	Std Dev	Linear Regress	Mean Sub	Med Sub	Std Dev	Linear Regress
2	67	70	70	70	69	71	71	66	71	73	72	67
5	69	72	72	69	75	76	76	75	74	77	77	75
10	66	71	71	66	72	78	78	72	71	80	80	71
15	64	72	72	64	66	77	77	66	66	79	79	66
20	60	72	72	60	55	78	77	56	50	80	80	52

Table 2: Performance of Above Imputation Methods in Terms of Percentage of Accuracy

Percentage of Missing	Mean Substitution	Median Substitution	Standard Deviation	Linear Regression
2	69	71	71	68
5	73	75	75	73
10	70	76	76	70
15	65	76	76	65
20	55	77	76	56
% of Accuracy	66.4	75	74.8	66.4

Figure 1: Comparison of Imputation Methods

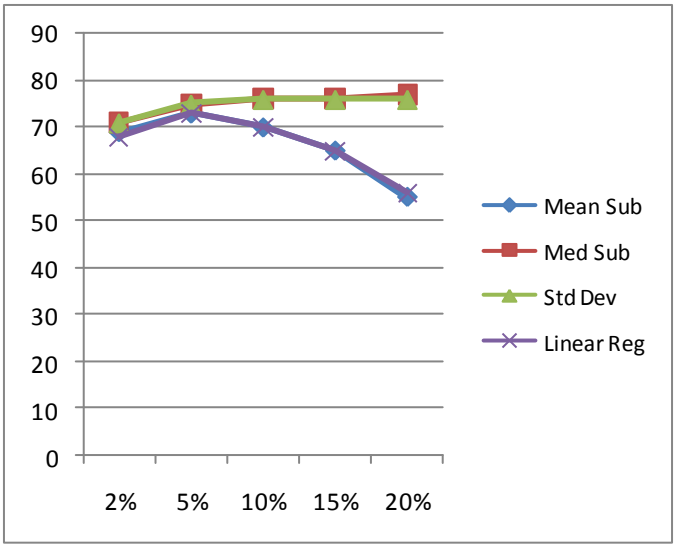

Figure 1 shows the difference graphically.

7. Conclusions and Future Enhancement

In this paper, as the imputation techniques like Median Substitution and Standard deviation shows better result, it can be used for further research. Our research also finds out that when the data gets groups into several sizes, there is some drastic improvement in the accuracy of percentage. In future data algorithms can be generated for grouping of data. After grouping, the two imputation methods can be applied.

References

[1]. Graham, J.W, "Missing Data Analysis: Making it work in the real world. Annual Review of Psychology", 60, 549 – 576 , 2009.

[2]. Jeffrey C.Wayman , "Multiple Imputation For Missing Data : What Is It And How Can I Use It?" , Paper presented at the 2003 Annual Meeting of the American Educational Research Association, Chicago, IL ,pp . 2 -16, 2003.

[3]. A.Rogier T.Donders, Geert J.M.G Vander Heljden, Theo Stijnen, Kernel G.M Moons, "Review: A gentle introduction to imputation of missing values" , Journal of Clinical Epidemiology 59 , pp.1087 – 1091, 2006.

[4]. Kin Wagstaff ,"Clustering with Missing Values : No Imputation Required" -NSF grant IIS-0325329,pp.1-10.

[5]. S.Hichao Zhang , Jilian Zhang, Xiaofeng Zhu, Yongsong Qin,chengqi Zhang , "Missing Value Imputation Based on Data Clustering", Springer-Verlag Berlin, Heidelberg ,2008.

[6]. Richard J.Hathuway , James C.Bezex, Jacalyn M.Huband , "Scalable Visual Assessment of Cluster Tendency for Large Data Sets", Pattern Recognition ,Volume 39, Issue 7,pp,1315-1324- Feb 2006.

[7]. Gabriel L.Scholmer, Sheri Bauman and Noel A.card "Best practices for Missing Data Management in Counseling Psychology", Journal of Counseling Psychology, Vol. 57, No. 1,pp. 1–10,2010.

[8]. R.Kavitha Kumar, Dr.R.M Chandrasekar,"Missing Data Imputation in Cardiac Data Set" ,International Journal on Computer Science and Engineering , Vol.02 , No.05,pp-1836 – 1840 , 2010.

[9]. Jinhai Ma, Noori Aichar –Danesh , Lisa Dolovich, Lahana Thabane , "Imputation Strategies for Missing Binary Outcomes in Cluster Randomized Trials"- BMC Med Res Methodol. 2011; pp- 11: 18. – 2011.

[10]. 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 ,pp.14-19 ,May 2011.

[11]. K.Raja , G.Tholkappia Arasu , Chitra. S.Nair , "Imputation Framework for Missing Value" , International Journal of Computer Trends and Technology – volume3Issue2 – 2012.

[12]. Paul D. Allison, Statistical Horizons, Haverford, PA, USA," Handling Missing Data by Maximum Likelihood", SAS Global Forum 2012- Statistics and Data Analysis

[13]. BOB L.Wall , Jeff K.Elser – "Imputation of Missing Data for Input to Support Vector Machines" ,

[14]. Ms.R.Malarvizhi, Dr.Antony Selvadoss Thanamani – "K-Nearest Neighbor in Missing Data Imputation", International Journal of Engineering Research and Development, volume 5Issue 1-November - 2012