International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 4, Issue 7 (November 2012), PP. 14-16

Framework for Missing Value Imputation

Ms.R.Malarvizhi[1], Dr.Antony Selvadoss Thanamani[2],
[1]*Scholar, Department of CS, NGM College of Arts and Science ,Pollachi ,Bharathiyar University,Coimbatore*.
[2]*HOD, Department of CS, NGM College of Arts and Science, Pollachi, Bharathiyar University, Coimbatore.*

Abstract— *Missing values may occur due to several reasons. In this paper, data is imputed by comparing the two most popular techniques. Mean Substitution the traditional method replaces mean value in K-means Clustering and in groups of kNN classifier. When compared in terms of accuracy of imputing missing data, the proposed kNN classifier is evaluated to demonstrate that the approach is better than the existing K-means clustering.*
Keywords— *K-Means Clustering, kNN Classifier, Missing Value, Mean Substitution, Imputation*

I. INTRODUCTION

Researchers are often faced with the issues of missing in Survey. Errors may occur due to human or machine when processed a sample and stored data values into their respective records. Traditional methods are there for handling missing data. List-wise deletion and pair-wise deletion exclude the data from analyses. These methods are unaccepted by the researchers as they produce biased result. K-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. kNN classifier operate on the premises that classification of unknown instances can be done by relating the unknown to the known according to some distance function. Mean substitution replaces missing data with the average of the valid data. This paper proposed a framework which compares two techniques.

II. TYPES OF VARIABLES IN MISSING DATA

Data that are Missing Completely at Random (MCAR) can be considered as a simple random sample of observed data. Data are Missing at Random (MAR) when the probability of the value being missing is dependent on some measurement characteristics of the individual but not on the missing value itself.(1). Data are missing not at Random (MNAR), variable value being missing is directly related value of the variable itself.

III. CONSTRAINTS FOR MISSING VALUE REPLACEMENT

Any method which is used for replacing the missing value should follow certain constraints. They are (i) they should produce unbiased result (ii) They should not change any values of other variables (iii) Estimation should minimize the cost.

IV IMPUTATION METHODS

A. List-Wise Deletion

The List-wise Deletion leads to the loss of large amount of data, as the whole record gets deleted even if one single variable was missing. So finally the database gets reduced.

B. Pair-Wise Deletion

Pair-wise deletion is an appropriate method when missing data is completely at random. It is easy to implement but the resultant data are difficult to interpret

C. Regression Imputation

Regression imputation is also known as conditional mean imputation in which linear regression equation is used to replace the missing values.

D. Hot-Deck Imputation

Hot-Deck imputation is a type of imputation in which missing data value is replaced with actual data by estimating similar data set which is currently in use.

E. Cold-Deck Imputation

Cold-Deck imputation follows the same procedure of Hot-Deck Imputation but it compares with the similar data set which is not currently in use.

F. Expectation Maximization (EM)

The EM strategy is based on a recursive process: The missing data have information that is useful in estimating various parameters, and the estimated parameter has information that is useful in finding the most likely value of the missing

data (Bennett, 2001). Thus, the EM method is an iterative procedure with two steps in each iteration. The disadvantage of EM is that the standard errors and confidence intervals are not provided, so obtaining those statistics requires an additional step. For inferential analyses for which those are essential, EM may not suffice.

G. Mean Imputation

In Mean Imputation, missing values are imputed with the mean value of that variable on the basis of the non missing values for that variable. This method assumes that data are MCAR and results in biased means when this assumption is false. Furthermore, imputing the mean value into cases tends to reduce the variance, which also attenuates covariance that the variable has with other variables. This method produces biased means with data that are MAR or NMAR and underestimates variance and covariance's in all cases. Experts strongly advise against this method (Allison 2001; Bennett 2001; Graham et al., 2003; Pallant, 2007).

IV. K-MEANS CLUSTERING

K-Means (Mac Queen, 1967) is one the simplest unsupervised learning algorithms that solve the well known clustering problem. The intra-cluster dissimilarity is measured in K-Means clustering by the summation of distances between the objects and the centroid of the clusters they are assigned. In K-Means clustering method, the data set X is divide into K-Clusters. Each cluster is represented by the centroid of the set of objects in the cluster.

The objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the n data points from their respective cluster center.

The algorithm is composed of the following steps:

Procedure 1:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

VI. KNN CLASSIFIER

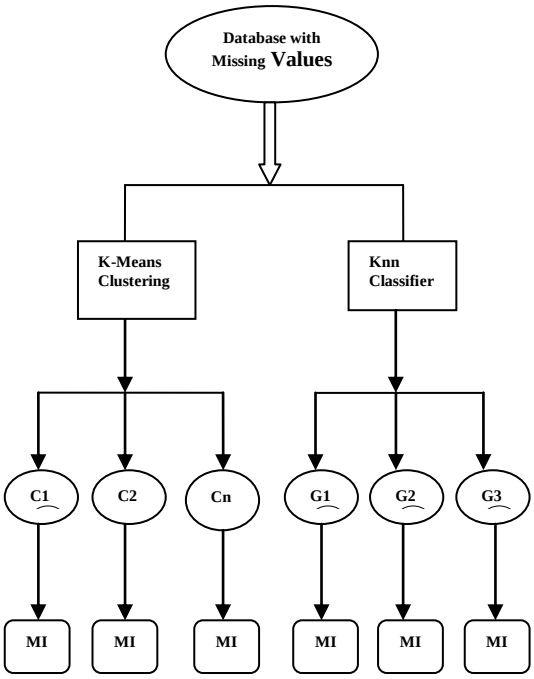
The *k*-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its *k* nearest neighbors (*k* is a positive integer, typically small). If *k* = 1, then the object is simply assigned to the class of its nearest neighbor. The neighbors are taken from a set of objects for which the correct classification is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. The *k*-nearest neighbor algorithm is sensitive to the local structure of the data

Procedure 2:

1. Compute Euclidean or Mahalanobis distance from target plot to those that were sampled.
2. Order samples taking for account calculated distances.
3. Choose heuristically optimal *k* nearest neighbor based on RMSE done by cross validation technique.
4. Calculate an inverse distance weighted average with the *k*-nearest multivariate neighbors.

VII. PROPOSED FRAMEWORK

The above two techniques are implemented separately in a dataset which has missing value. When K-Means Clustering is implemented, the missing values in each cluster are replaced by finding mean of the remaining values. Similarly mean value is imputed for each group from kNN classifier. When both the datasets are compared for accuracy, kNN imputation seems to perform better than the K-Means Clustering.



VIII. CONCLUSION AND FUTURE ENHANCEMENT

This paper discussed different methods to impute the missing values. It also explains about the framework to be implemented with dataset with missing data. The presented two techniques deal with grouping of dataset so that each group can be dealt with mean substitution or any other traditional method. The results are compared for accuracy.

REFERENCES

[1] Allison, P.D-"Missing Data", Thousand Oaks, CA: Sage -2001.
[2] Bennett, D.A. "How can I deal with missing data in my study? Australian and New Zealand Journal of Public Health", 25, pp.464 – 469, 2001.
[3] Graham, J.W. "Adding missing-data-relevant variables to FIML- based structural equation models. Structural Equation Modeling", 10, pp.80 – 100, 2003.
[4] Graham, J.W, "Missing Data Analysis: Making it work in the real world. Annual Review of Psychology", 60, 549 – 576 , 2009.
[5] Gabriel L.Schlomer, Sheri Bauman, and Noel A. Card : " Best Practices for Missing Data Management in Counseling Psychology" , Journal of Counseling Psychology 2010, Vol.57.No 1,1 – 10.
[6] 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.
[7] 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.
[8] Kin Wagstaff ,"Clustering with Missing Values : No Imputation Required" -NSF grant IIS-0325329,pp.1-10.
[9] S.Hichao Zhang , Jilian Zhang, Xiaofeng Zhu, Yongsong Qin,chengqi Zhang , "Missing Value Imputation Based on Data Clustering", Springer-Verlag Berlin, Heidelberg ,2008.
[10] 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.
[11] Qinbao Song, Martin Shepperd ,"A New Imputation Method for Small Software Project Data set", The Journal of Systems and Software 80 ,pp,51–62, 2007.
[12] 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.
[13] 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.
[14] 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.
[15] 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.
[16] K.Raja , G.Tholkappia Arasu , Chitra. S.Nair , "Imputation Framework for Missing Value" , International Journal of Computer Trends and Technology – volume3Issue2 – 2012.
[17] BOB L.Wall , Jeff K.Elser – "Imputation of Missing Data for Input to Support Vector Machines" ,