IEEE- Fourth International Conference on Advanced Computing, ICoAC 2012
MIT, Anna University, Chennai. December 13-15, 2012

Missing Value Imputation Techniques Depth Survey And an Imputation Algorithm To Improve The Efficiency Of Imputation

1.S.Thirukumaran
Research scholar
thirukumaran75@gmail.com

2.Dr. A.Sumathi
sumathi_2005@rediffmail.com

Abstract :

Missing data in Medical database is an issue which makes lose of data integrity, solution for missing value is imputing the relevant value for every missing value(here data and value takes same meaning) it is the scope of imputation and it gives the data integrity. According to the title so many imputation Techniques available. This paper aims to describe the depth survey of types of imputation techniques and which is categorized in the form of table with the attributes like Technique, Description, when to be used, Advantages, disadvantages, Almost different imputation Techniques ideas were exposed in this paper after detailed study. After feasible study here we exposed the concept to improve the imputation technique more worthy than other techniques that Clustering imputation Algorithm proposed which reduce the error rate of imputed value for missing data into Medical database and makes the imputation perfect to the maximum level. And the results elaborates the reduced error rate for dataset of 786 samples with 8 features.

I. Introduction :

Data Integrity is the foremost aim of database Missing data helps to degrade the integrity and to avoid the degradation or to improve or to maintain the data integrity missing data to be imputed properly. Handling of imputation causes the three major issues 1. Loss of information, as a consequence, a loss of efficiency. 2. Data handling is an issue, computation and analysis due to irregularities in the data structures. 3. Systematic difference among the data. The scope of the paper Chapter wise ie. chapter I. introduction, chapter II. Producing the survey experience about what is imputation?. How many types of imputation technique available?, distinguished the types of imputation techniques. Chapter III. proposed the Clustering imputation Algorithm is not compared with any of the available method but we trust and proved that our method reduces the imputation error rate so it improves the imputation technique better. Chapter VI Experimentation of clustered algorithm Result. Chapter V Conclusion.

II. Background Work :

In this section we would like to share the survey experience about imputation Techniques like Mean-Mode imputation, Hotdeck imputation, K-nearest neighbour's imputation, multiple imputation, Multivariate imputation by chained equations (MICE).

Mean-Mode imputation (MMimpute) : MMimpute filling the missing data by the mean or mode(qualitative) from all Know data set [H. Kim, G.H. Golub, H. Park, 2005][P.D. Allison,2001].

Hotdeck (HD) imputation: Given an incomplete pattern, HD replaces the missing data with values form input data vector that is closest in terms of the attributes that are Know in both patterns[H. Kim, G.H. Golub, H. Park, 2005]. HD attempts to preserve the distribution by substituting different observed values for each missing data [J.L. Schafer,1997]. The similar method of HD is Cold deck imputation method which takes other data source than current dataset [H. Kim, G.H. Golub, H. Park, 2005].

K-nearest neighbor`s imputation (KNNimpute) : training dataset incomplete pattern where missing data K will be selected with the help of known data such that they minimize some distance measure. Once the K nearest neighbours have been found, are placement value to substitute for the missing attribute value must be estimated. How the replacement value is calculated depends on the type of data; the mode can be used for qualitative data and the mean for

978-1-4673-5584-1/12/$31.00©2012 IEEE

continuous data. The main benefits of KNNimpute are: (1) KNNimpute can easily handle and predict both quantitative features and qualitative features. (2) KNNimpute does not create explicit predictive models, because the training dataset is used as a'lazy' model. Also, this method can easily treat cases with multiple missing values. The major drawback of this approach is that when ever the KNNimpute looks for the most similar instances, the algorithm searches through all the dataset. Nevertheless, even thought his limitation can be very critical for large databases, it has been shown that KNNimpute can provide a robust procedure for missing data estimation [G.E. Batista, M.C. Monard,2003][G.E. Batista, M.C. Monard, 2002][O. Troyanskaya, M. Cantor, 2001]. To apply the KNN approach to impute missing data, one of the most important issues is to select.

Multiple imputation (MULimpute) : MULimpute three step procedure an appropriate imputation model generates the N completed dataset which contains N sets of possible values for imputation of missing data. Second Procedure each completed dataset can be analyzed using standard software. Finally result of the an analyses are combined and get the exact imputation for the missing data[Schafer, 1997][Schafer, 1999][Rubin,1987,1996]. The main feature of multiple imputation is that a very small value of N will usually suffice (2_*m*_5).

Multivariate imputation by chained equations (MICE) : MICE is an approach that imputes data variable by variable this means the imputation model is specified separately for each variable with missing values while other variables are predictors[Ulrike, Grittner,2011].

The Features of the different imputation methods mentioned in the Table 2.1.

Technique	Description	Advantages	Disadvantages	Studies
Deletion-based Listwise deletion	Eliminates from further analysis all cases with any missing data	Easy to use (default in most statistical packages) ''Conservative'': hard to find statistical significance Preserves more data and is more accurate than listwise deletion	Sacrifices a large amount of data and has a negative impact on statistical power Correlations or covariances may be biased	Kim and Curry (1977), Raymond (1986), Malhotra (1987), Little and Rubin (1987) Gleason and Staelin (1975), Kim and Curry (1977), Raymond (1986), Roth (1994)
Replacement-based Mean substitution	Missing value is replaced by the mean (see, text below for variants	Preserves the data and is easy to use	Negative impact on variance estimates and degrees of freedom	Ford (1976), Raymond (1986), Little and Rubin (1987), Kaufman (1988), Hawkins and Merriam (1991), Quinten and Raaijmakers (1999)
Total mean substitution	Missing value is replaced by the mean on the item for all respondents answering the question	Easy to use (built-in in most statistical packages), sample retention	Downward biased variance/covariance estimates	Little and Rubin (1987), Quinten and Raaijmakers (1999)
Subgroup mean substitution	Missing value is replaced by the mean on the subgroup of which the respondent is a member	Gives better estimates, when compared to the total mean substitution procedure	Downward biased variance, arbitrary nature of defining subgroups in some situations	Ford (1976)
Regression imputation	Estimates relationships among variables, and then uses coefficients to estimate the missing value	Estimated data preserve deviations from the mean and the shape of the distribution	Distorts the number of degrees of freedom and could artificially increase the relationships	Frane (1976), Cohen and Cohen (1983), Raymond and Roberts (1987), Little and Rubin (1987), Little (1988)

Hot-deck imputation	Replaces a missing value with the actual score from a similar case in the dataset	Missing data are replaced by realistic values and not means that distort distributions	Little theoretical or empirical work to determine its accuracy, problematic if no other case is closely related in all aspects of the data set	Ford (1983), Roth et al. (1999)
Model-based Maximum likelihood	Parameters are estimated by available data and missing scores are estimated based on the parameters	Increased accuracy if model is correct	The distributional assumptions required by the technique are relatively strict	Donner and Rosner (1982), DeSarbo et al. (1986), Lee and Chiu (1990)
Expected maximization	An iterative process that continues until there is convergence in the parameter estimates	Increased accuracy if model is correct	The algorithm takes time to converge and is too complex	Laird (1988), Little and Rubin (1987), Malhotra (1987), Azen et al. (1989), Ruud (1991), Graham and Donaldson (1993)

Table 2.1 The classification of imputation methods.

III. Missing Value Block Generation :

Concept of Missing value block generation that the entire database is scanned to group the records into various blocks in such a way that a block contains the records which are missing with same set of attributes. This phase helps the system to extract the incomplete records from the database and also to process the same to make it complete for effective data mining process. The maximum number of blocks will be (2^n - 2) Where n = total number of attributes in the dataset. After formulating the blocks, the records in every block will be considered as the testing dataset and the same will be passed as the seed-points (reference points) for the clustering process. In our testing, we have assumed that at most one or more attribute is missing.

Clustering and Missing Value Imputation

In this phase, clusters will be generated by passing every record in every block as the seed-point-record. The records in the training set will be formulated into number of clusters in which the missing-attribute-records are the seed points. In most of the clustering models, the concept of similarity is based on distance. Here we explore a new approach for measuring the similarity, wherein a record is said to be similar to the seed-point- record only if it has the maximum weight. The training dataset is scanned for measuring the similarity. The standard deviation of the attributes is taken into account for measuring the similarity. The weights are assigned to the records based on the number of similar attributes. The cluster is then formed by grouping the records which have higher degree of weights. Finally, the missing-attribute-value(s) in each cluster resulted by computing the mean value of the respective attribute(s) in the cluster and complete dataset is generated without any missing any data. The Figure 3.1 show the flow of clustering.

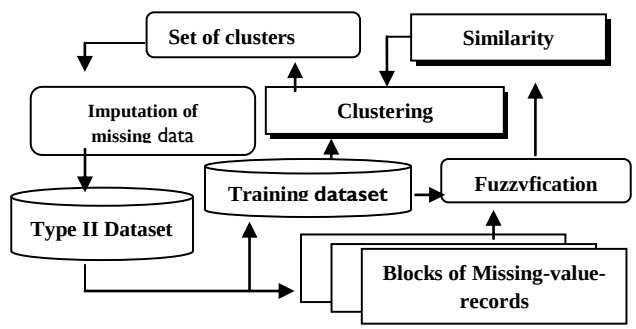

Figure 3.1 Flow of Clustering algorithm to impute the value.

The below given Equation to calculate the average imputation error (E) [Mariso, Giardina, 2005] which gives the accuracy and consistency to the dataset which helps to implement the Clustering algorithm.

$$E = \left(\sum_{k=1}^{m} \left(\left(\sum_{i=1}^{n} \left(| Oij - Iij | / \left(\max j - \min j \right) \right) \right) / n \right) \right) / m$$

Where n is the number of imputed values, m is the number of random simulations for each missing value, O_{ij} is the original value of the attribute j, I_{ij} is the imputed value, Max_j is the maximum value of the of the attribute j, Min_j is the is the minimum value of the of the attribute j, j is the corresponding attribute to which O_i and I_i belong.

	Attr1	Attr 2	Attr 3	Attr 4	Attr 5	Attr6	Attr7	Attr8
5%	0.76838	5.3796	3.0884	3.2895	27.325	1.5552	0.037879	2.8886
10 %	0.77199	5.1483	3.1775	3.3446	26.806	1.559	0.039566	2.8777
15 %	0.77314	5.1351	3.103	3.364	25.659	1.5749	0.042627	2.9149
20 %	0.77024	5.1167	3.174	3.4214	25.299	1.5905	0.040265	2.8274
25 %	0.7831	5.2077	3.1112	3.4126	25.917	1.5633	0.037698	2.8599
30 %	0.7734	5.1942	3.134	3.4264	25.733	1.5521	0.040259	2.8335
35 %	0.76873	5.1878	3.1666	3.3207	26.24	1.5977	0.039485	2.8683

Table 3.1: Average Imputation Error.

IV. Experiments and Results :

The proposed frameworks was tested with a comprehensive set of Pima Indian diabetes(PIMA) dataset has 786 samples with 8 features, and the other dataset used is kala-azar disease(KZAR) dataset consists of 68 samples, each with 7 input features. There are 2 classes, 0 for normal and 1 for presence of disease. The goal is to provide motivation for the proposed design of the frameworks. The effect of each of the frameworks on the accuracy of imputation improvement is experimentally demonstrated and for the first approach the Table 3.1 of Chapter III and Figure 4.1 shows the experimental evaluations performed.

This section summarizes that how the error rate reduced for eight attributes of with 786 samples of PIMA dataset, according to the error rate source Table1 through the experiment we got it clear hint that for different % of missing value after applying the clustering imputation process or algorithm the error rate maintained to low level and attribute5 shows the error rate is 26.24 for 35% of missing value, for 30% the error rate is 25.73 it means the error rate is high and imputation which could not be better for this attribute5 alone.

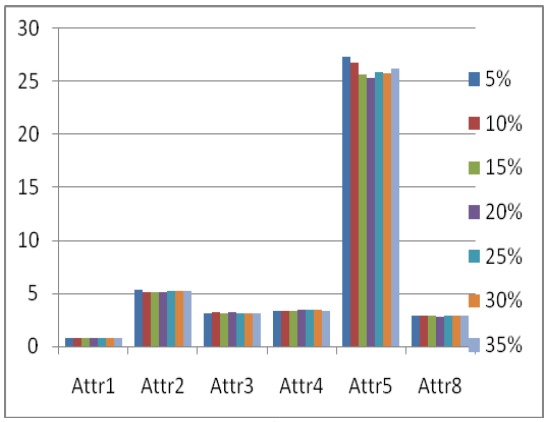

Figure 4.1 X-Axis Level Of Missing Value In(%).

V. Conclusion :

We trust that this paper gives the complete views about the Imputation Techniques and categories of imputation methods additionally the concept clustering algorithm proposed and proved to reduce the error rate of the imputation for the data sample taken .

References :

[1] Azen, S.P., Van Guilder, M., Hill, M.A., 1989. Estimation of parameters and missing values under a regression model with non-normally distributed and non-randomly incomplete data. Statistics in Medicine 8, 217–228.

[2] Cohen, J., Cohen, E., 1983. Applied Multiple Regression/Correlational Analysis for the Behavioral Sciences. Hillsdale, Erlbaum, NJ.

[3] DeSarbo, W.S., Green, P.E., Carroll, J.D., 1986. Missing data in product-concept testing. Decision Sciences 17, 163–185.

[4] Donner, A., Rosner, B., 1982. Missing value problems in multiple linear regression with two independent variables. Communications in Statistics 11, 127–140.

[5] Ford, B.L., 1976. Missing data procedures: a comparative study. In: Statistical Reporting

Serviceunknown:book, U.S. Department of Agriculture, Washington, DC.

[6] Ford, B.L., 1983. An overview of hot-deck procedures. In: Madow, W.G., Olkin, I., Rubin, D.B. (Eds.), Incomplete Data in Sample Surveys. Theory and Bibliographies, vol. II. Academic Press, New York, pp. 185–207

[7] Frane, J.W., 1976. Some simple procedures for handling missing data in multivariate analysis. Psychometrika 41, 409–415.

[8] Gleason, T.C., Staelin, R., 1975. A proposal for handling missing data. Psychometrika 40, 229–252.

[9] G.E. Batista, M.C. Monard, A study of k-nearest neighbour as an imputation method, in: Second International Conference on Hybrid Intelligent Systems, vol. 87, Santiago, Chile, 2002, pp. 251–260.

[10] G.E. Batista, M.C. Monard, An analysis of four missing data treatment methods for supervised learning, Applied Artificial Intelligence 17 (5–6) (2003) 519–533.

[11] Graham, J.W., Donaldson, S.W., 1993. Evaluating interventions with differential attrition: the importance of nonresponse mechanisms and use of follow-up data. Journal of Applied Psychology 78, 119–128.

[12] Hawkins, M.R., Merriam, V.H., 1991. An overmodeled world. Direct Marketing 21–24.

[13] H. Kim, G.H. Golub, H. Park, Missing value estimation for DNA microarray gene expression data: local least squares imputation, Bioinformatics 21 (2) (2005) 187–198.

[14] J.L. Schafer, Analysis of Incomplete Multivariate Data, Chapman & Hall, Florida, USA, 1997.

[15] Kim, J.O., Curry, J., 1977. The treatment of missing data in multivariate analysis. Sociological Methods and Research 6, 215–217

[16] Kaufman, C.J., 1988. The application of logical imputation to household measurement. Journal of the Market Research Society 30, 453–466.

[17] Laird, N.M., 1988. Missing data in longitudinal studies. Statistics in Medicine 7, 305–315.

[18] Lee, S.Y., Chiu, Y.M., 1990. Analysis of multivariate polychoric correlation models with incomplete data. British Journal of Mathematical and Statistical Psychology 43, 145–154.

[19] Malhotra, N.K., 1987. Analyzing marketing research data with incomplete information on the dependent variable. Journal of Marketing Research 24, 74–84.

[20] Mariso Giardina, Yongyang Huo, Francisco Azuaje, Paul McCullagh, Roy Harper, "A Missing Data Estimation Analysis in Type II Diabetes Databases", Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05) 1063-7125/05 2005 IEEE.

[21] O. Troyanskaya, M. Cantor, O. Alter, G. Sherlock, P. Brown, D. Botstein, R. Tibshirani, T. [22] Hastie, R. Altman, Missing value estimation methods for DNA microarrays, Bioinformatics 17 (6) (2001) 520–525.

[23] P.D. Allison, Missing data, Sage University Papers Series on Quantitative Applications in the Social Sciences, Thousand Oaks, California, USA, 2001.

[24] Quinten, A., Raaijmakers, W., 1999. Effectiveness of different missing data treatments in surveys with Likert-type data: introducing the relative mean substitution approach. Educational and Psychological Measurement 59 (5), 725–748.

[25] Raymond, M.R., 1986. Missing data in evaluation research. Evaluation and the Health Profession 9, 395–420.

[26] Raymond, M.R., Roberts, D.M., 1987. A comparison of methods for treating incomplete data in selection research. Educational and Psychological Measurement 47, 13–26.

[27] Roth, P.L., 1994. Mising data: a conceptual review for applied psychologists. Personnel Psychology 47 (3), 537–560.

[28] Roth, P.L., Switzer, F.S., Switzer, D.M., 1999. Missing data in multiple item scales: a Monte Carlo analysis of missing data techniques. Organizational Research Methods 2 (3), 211–232.

[29] Rubin, D.B., 1987. Multiple Imputation of Nonresponse in Surveys.Wiley, NewYork.

[30 Rubin, D.B., 1996. Multiple imputation after18+ years. J. Amer. Statist. Assoc. 91, 473–489.

Ruud, P.A., 1991. Extensions of estimation methods using the EM algorithm. Journal of Econometrics 49, 305–341.

[31] Schafer, J.L., 1997. Analysis of Incomplete Multivariate Data. Chapman & Hall, London.

[32] Schafer, J.L., 1999. Multiple imputation: a primer. Statist. Methods Med. Res. 8, 3–15.

[33] Ulrike Grittner,[1] Gerhard gmel,[2,3,4,5] samuli ripatti,[6] Kim bloomfield[1,7] & Matthias wicki[2] "Missing value imputation in longitudinal measures of alcohol consumption." International Journal of Methods in Psychiatric Research

[34] Int. J. Methods Psychiatr. Res. 20(1): 50–61 (2011).