IJCAT International Journal of Computing and Technology, Volume 1, Issue 3, April 2014
ISSN : 2348 - 6090
www.IJCAT.org

Missing Value Estimation and Impact on Heterogeneous Data set

[1] Prutha Deshmukh, [2] Swati G. Kale

[1] Department of Information Technology, Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India.

[2] Department of Information Technology, Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India

Abstract - Missing Data are question without answer or variables without observation. Missing Data can be treacherous because it is difficult to identify the problem. Numerous industrial and research database include missing values. It is not uncommon to encounter database that have up to half of the entries missing making it very difficult to mine them using data analysis method that can work only with complete data. Missing Value or data result in bias that impacts on the quality of learned patterns or/and performance. Missing Data Imputation is a key issue in learning from incomplete data. Mixture kernel based iterative estimator is advocated to impute mixed-attribute data sets.

Keywords **- Classification, data mining, methodologies, machine learning**

1. Introduction

Missing data imputation aims at providing estimations for missing values by reasoning from observed data. Because missing values can result in bias that impacts on the quality of learned patterns or/and the performance of classifications and missing data imputation has been a major factor in learning from incomplete data. Different techniques have been developed with for dealing with missing values in data base with their independent attributes is all either continuous or discrete. These imputation algorithms cannot be applied to many real data sets, because these data sets are often with both continuous and discrete independent attributes. These heterogeneous data sets are referred to as mixed-attribute data sets and their independent attributes are called as mixed independent attributes. To fulfill the above practical requirement, this paper studies a new setting of missing data imputation, i.e., imputing missing data in mixed-attribute data sets. In almost any research performed there is a potential for missing or incomplete data. The issues with the missingness is that nearly all classic and modern techniques assume or require complete data most common statistical packages default to least desirable option for dealing with missing data: deletion of the case from the analysis. Methods for dealing with missing values can be classified into three categories by following the idea from: 1) Case deletion, 2) learning without handling of missing values, and 3) missing value imputation. The case deletion is to simply omit those cases with missing values and only to use the remaining instances to finish the learning assignments. The second approach is to learn without handling of missing data, such as Bayesian Networks method or Artificial Neural Networks method. Different from the former two, missing data imputation method advocates filling in missing values before a learning application. Missing data imputation is a procedure that replaces the missing values with some plausible values. While the imputation method is regarded as a more popular strategy.

Missing Data are question without answer or variables without observation. Missing Data can be treacherous because it is difficult to identify the problem. Numerous industrial and research database include missing values. Missing information can diminish the confidence on the concepts learned from data. It is not uncommon to encounter database that have up to half of the entries missing making it very difficult to mine them using data analysis method that can work only with complete data. Missing Value or data result in bias that impacts on the quality of learned patterns or/and performance. When correct and mismeasured data are obtained for a sample of the observations, it is obviously possible to estimate the parameters relating the results to the covariate of interest. Realistic models for the mis measurement process are difficult to construct

We propose an easily implemented method that is nonparametric with respect to the mis-measurement process and that is applicable when mis-measurement is due to the problem of incomplete missing data, errors in

variables, or use of not so perfect surrogate covariates. Missing Data Imputation is a key issue in learning from incomplete data. Missing Data imputation is an important step in the process of machine learning and data mining when certain values are missed Mixture kernel based iterative estimator is advocated to impute mixed-attribute data sets. Even a small amount or percentage of missing data can cause a serious problem with your analysis leading you to draw wrong conclusion.

Imputation method allows to use standard complete-data methods, Can incorporate data collector's knowledge to reflect the uncertainty about imputed values (sampling variability and uncertainty about the reasons for no response) Increases efficiency of estimation Provides valid inferences (for variance estimators) under an assumed model for no response .Allows one to study sensitivity to various models. In real life application missing values imputation is an actual and challenging problem confronted by machine learning and data mining.

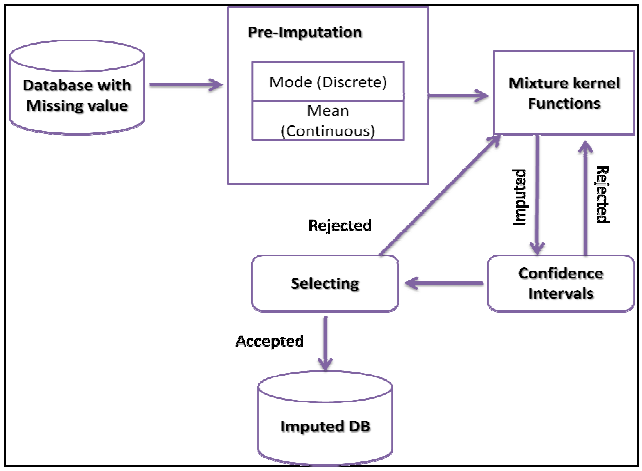

Figure1 : Flow Diagram

The figures describes the flow of how it will work step by step, firstly we have to collect the data, as we know numerous industrial and research database include missing values. It is not uncommon to encounter database that have up to half of the entries missing making it very difficult to mine them using data analysis method that can work only with complete data. After loading the data sets that have missing values we have to perform the pre-imputation process, in pre-imputation process values are being imputed in the place of missing values, so that further process can be carried out. And that is what kernel function is being applied to it, and after performing kernel function some filtering work is done and then what we get is the final imputed data base.

2. Nonparametric Iterative Imputation Method

Before presenting the new imputation algorithm first recall the previous work which reported on kernel functions for discrete attributes. Then, a mixture kernel function is proposed by combining a discrete kernel function with a continuous kernel function. Furthermore, a new estimator is constructed based on the mixture kernel, develops novelty kernel estimators for discrete and continuous target values, respectively. In further section the nonparametric iterative imputation algorithm is extended from a single kernel to a mixture of kernels, and the nonparametric iterative imputation algorithm is designed and simply analyzed.

Single Imputation Using Kernel Function

This module shows about the kernel function. After getting the basic imputation, then apply the kernel function separately for both the discrete and continuous attributes. Then integrate both the discrete and kernel function to get the mixture kernel function

1) Discrete Kernel Function

$L(X^d_{t\,i}, x^d_t) = [1 \text{ if } X^d_{t,I=} x^d_t \text{ and } \lambda \text{ if } X^d_{t,i} = x^d_t]$

Where,
X^d_i -- Discrete Variable or attributes
λ -- Smoothing Parameter normally discrete attributes are contains a binary format values example is either it will be 0 or 1.so for this step, the output will shows about the similar values as the imputation for the missing values by taking one attribute as a relation**.**

2) Continuous Kernel Function

K(x - Xi /h)
K(.) is a mercer kernel, i.e., positive definite kernel.

3) Mixture Kernel Function

$K\ h, \lambda, ix = K(x - X_i/h)\ L(X^d_i, x_i^d, \lambda)$
Where, h->0 and λ->0 (λ, h is the smoothing parameter for the discrete and continuous kernel function, respectively),
Kh,λ,ix -- symmetric probability density function.
$K(x-X_i/h)$ -- Continuous Kernel Function
$L(X_i^d, x_i^d, \lambda)$ -- Discrete Kernel Function

Constructing the Estimator and Iterative Imputation

Construct the estimator, separately for continuous and discrete. Estimator is nothing but, it attempts to approximate the unknown parameter using the

measurements. Then by the idea of the estimator calculate the iterative value for each attributes by using the formula. The iterative method explains that all the imputed values are used to impute subsequent missing values, i.e., the $(t+1)^{th}$ ($t\geq1$) iteration imputation is carried out based on the imputed results of the t^{th} imputation, until the filled-in values converge or begin to cycle or satisfy the demands of the users. Normally first imputation is single imputation. It cannot provide valid. standard confidence intervals. Therefore running extra (imputation) iterative imputation based on the first imputation is reasonable and necessary for better dealing with the missing values. Since the second iteration imputation is carried out based on the former imputed results. Here, a stopping criterion is designed for nonparametric iterations. With t imputation times, there will be (t-1) chains of iterations. Note that the first imputation won't b considered when talking about the convergence because the final results will be decided mainly by imputation from the second imputation. Of course, the result in the first imputation always generates, to some extent, effects for the final results.

Iterative Kernel Estimator for Continuous Target Variable.

The kernel estimator, ^ mðxÞ, for continuous missing target values m(x) for data sets with mixed independent attributes is defined as follows:

$$mt(x) = \frac{n^{-1}\sum_{i=1}^{n} Y_i^t\, K_{h,\lambda,ix}}{n^{-1}\sum_{i=1}^{n} K_{h,\lambda,ix} + n^{-2}}$$

Iterative Kernel Estimator for Discrete Target Variable
Let Dy=(0,1------cy-1) denote the range of m(x), one could estimate m(x) by

$$mt(x) = \frac{\sum_{i=1}^{n}\sum_{y\in D_{y,y\neq Y_i}} l(Y_i^t, y_t, \lambda)\, y_t\, K_{n,\lambda}}{\sum_{i=1}^{n} K_{h,\lambda}}$$

ALGORITHM DESIGN

In this approach, the i^{th} missing value is denoted by MV_i and the imputed value of MV_i in t^{th} iteration imputation is regarded as MV_i^t . From the algorithm, all the imputed values are used to impute subsequent missing values, i.e.$(t+1)^{th}$ ($t\geq1$) iteration imputation is carried out based on the imputed results of the t^{th} imputation, until the filled-in values converge or begin to cycle or satisfy the demands of the users.

```
//the first imputation
FOR each MV_i in Y
        MV_i^1=mode ( S^r in Y);//if Y is discrete variable
        MV_i^1=mean ( S^r in Y);//if Y is continuous variable
END FOR
// t-th iteration of imputation (t>1)
t=1;
REPEAT
t++;
FOR each MV_i in Y
        MV_i= MV_i^(t-1) , pε S_m ,p!=i
        MV_i^t from Eq. (1) // if discrete variable
        MV_i^t from Eq. (2) // if continuous variable
END FOR
UNTIL
|CA_t – CA_(t-1) | >= ε // if discrete variable
Convergence or Cycling // if continuous variable
```

In the first iteration of imputation in the above algorithm, all the missing values are imputed using the mean for continuous attributes and mode for discrete ones. Using the mean or mode of an attribute to replace missing values is a popular imputation method in machine learning and statistics. However, imputing the mean or mode will be valid only if the data set is chosen from a population with a normal distribution. This is usually not possible for real applications because the real distribution of a data set is not known in advance. On the other hand, single imputation cannot provide valid errors and confidence intervals, since it ignores the uncertainty implicit in the fact that the imputed values are not the actual values. So, running extra iteration-imputations based on the first imputation is reasonable and necessary for better dealing with the missing values. Since the second iteration of imputation, each of iteration imputation is carried out based on earlier imputed results with the nonparametric kernel estimator. During the imputation process, when the missing value is imputed, all other missing values are regarded as observed values The iteration imputation for missing continuous attributes will be end up when the filled-in values converge or begin to cycle,and for discrete missing values, the imputation algorithm will be terminated if $|CA_t - CA_{t-1}| \geq \varepsilon$ this condition satisfy,where ε is a nonnegative constant specified by users. The classification accuracy for the t^{th} imputation is denoted by CA_t. Then the period of iteration of the algorithm is t for imputing a discrete missing attribute because the first imputation has been finished.

3. Impact Factor

The main and foremost data mining process deals with projection, approximation, classification, pattern recognition. Therefore, the significance of the analysis depends heavily on the accuracy of the data-set and on the chosen sample data to be used for training and testing. The problem of missing data must be addressed since ignoring this problem can introduce bias into the models being evaluated and lead to inaccurate data mining conclusions. The objective of this research is to address the impact of

missing data on the data mining process. This is how the impact has been displayed in this project.

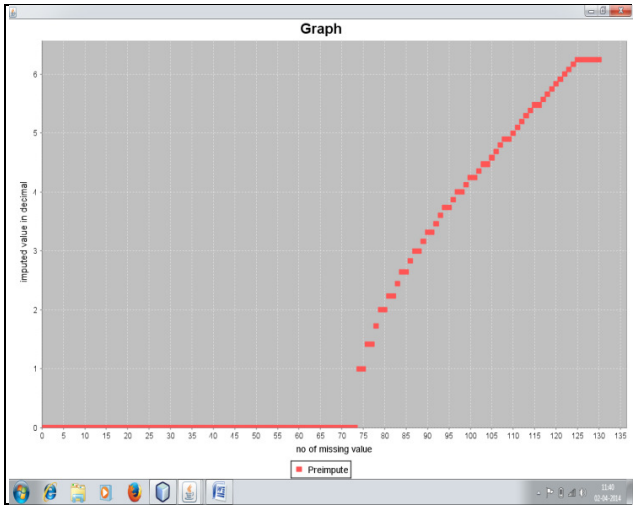

Figure 2: Impact of imputation

4. Future Scope and Conclusion:

In this paper, a consistent kernel regression has been proposed for imputing missing values in a mixed-attribute data set. The mixture kernel- based iterative nonparametric estimators are proposed against the case that data sets have both continuous and discrete isolateted attributes. It utilizes all available revealed information, including reveal information in incomplete instances with missing values, to impute missing values, whereas existing imputation methods use only the observed information in complete instances ie without missing values.

In future, we plan to further explore global or local kernel functions, instead of the existing ones, in order to achieve better extrapolation and interpolation abilities in learning algorithms. And also Impact factor can be shown with the help of report generation.

References

[1] Pedro J. Garcı´a-Laencina Æ Jose´-Luis Sancho-Go´mez Æ Anı´bal R. Figueiras-Vidal "Pattern classification with missing data: a review", Neural Comput & Applic (2010) 19:263–282

[2] Shichao Zhang, Senior Member, IEEE "Parimputation: From Imputation and Null-Imputation to Partially Imputation" IEEE Inteligent Informatics bulletin november 2008 vol 9 no 1

[3] Judi Scheffer "Dealing with Missing Data " judi Res. Lett. Inf. Math. Sci. (2002) 3, 153-160 R.L.I.M.S. Vol. 3, April, 2002

[4] Alireza Farhangfara, Lukasz Kurganb, Jennifer Dy" Impact of imputation of missing values on classification error for discrete data" A. Farhangfar et al. / Pattern Recognition 41 (2008) 3692 – 3705

[5] Yulei He Missing Data Analysis Using Multiple Imputation: Getting to the Heart of the Matter Circ Cardiovasc Qual Outcomes. 2010;3:98-105

[6] Youting Sun, Ulisses Braga-Neto, Edward R. Dougherty " Impact ofMissing Value Imputation on Classification for DNAMicroarray Gene Expression Data—AModel-Based Study" Hindawi Publishing Corporation EURASIP Journal on Bioinformatics and Systems Biology Volume 2009

[7] Schafer, J., & Graham, J. "Missing data: Our view of the state of the art." Psychological Methods, 7(2), 147–177. 2002

[8] Stuart, E. A., Azur, M., Frangakis, C., & Leaf, P." Multiple imputation with large data sets: A case study of the children's mental health initiative." American Journal of Epidemiology, 169(9), 1133–1139. 2009

[9] SPSS White Paper "Missing Data:The Hidden Problem"

[10] Olga Troyanskaya, Michael Cantor , Gavin Sherlock , Pat Brown, Trevor Hastie , Robert Tibshirani , David Botstein and Russ B. Altman "Missing value estimation methods for DNA microarrays" BIOINFORMATICS Oxford University Press 2001Vol. 17 no. 6 2001

[11] Mehmet G¨onen Ethem Alpaydın "Multiple Kernel Learning Algorithms" Journal of Machine Learning Research 12 (2011) 2211-2268

[12] G. Batista and M. Monard, "An Analysis of Four Missing Data Treatment Methods for Supervised Learning," Applied Artificial Intelligence, vol. 17, pp. 519-533, 2003.

[13] JonathanA C Steme, Ian R White, Johan B Carlin Michael Spratt, Patrick Royston, Michael G Kenward, Angela M Wood, James R Carpenter "Multiple Imputations for missing data in epidemiological and clinical research: potential and pitfalls" BMJ : British Medical Journal 2009 June 29.

[14] Muhammad Shoaib B. Sehgal, Iqbal Gondal, Laurence S. Dooley "Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data" BIOINFORMATICS Vol. 21 no. 10 2005, pages 2417–2423 February 24, 2005

[15] Little, Roderick J., Donald Rubin" Statistical Analysis with Missing Data." John Wiley & Sons, Inc: Hoboken.2002

[16] LSimpute "Accurate estimation of missing values in microarray data with least squares method". Nucleic Acids Res., 32, e34. Bø,T.H. et al. (2004)

[17] K. Lakshminarayan, S.A. Harp, T. Samad, "Imputation of missing data in industrial databases", Appl. Intell. 11 (1999) 259–275.

Prutha a Deshmukhreceived the B.E. degree in Information Technology from DBACER , Nagpur, INDIA in 2012 and perusing MTech, degree in Information Technology from YCCE, Nagpur in 2013.

Swati G. Kale She now currently working as a Asst. Prof. in Information Technology Department at YCCE Nagpur.