ISSN 2250-3153

Enhancing Iterative Non-Parametric Algorithm for Calculating Missing Values of Heterogeneous Datasets by Clustering

SUJATHA.R*

*Assistant Professor, CSE, Shivani Engineering College, Trichy-9, India

Abstract- Machine learning and data mining retort heavily on a large amount of data to build learning models and make predictions. There is a need for quality of data, thus the quality of data is ultimately important. Many of the industrial and research databases are plagued by the problem of missing values. A variety of methods have been developed with great success on dealing with missing values in data sets with uniform attributes. But in real life dataset contains heterogeneous attributes. In this paper, apart from the overview of imputation, then discussing about the proposed work i .e a new setting of handling missing data imputation (that is imputing missing data in data sets with mixed attributes and also in clustered data sets only with continuous attributes) in non-parametric mixture kernel based.

Index Terms— Data mining, Missing values, Mixed attributes, Imputation, Regression

I. INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. A common problem in data mining is that of automatically finding outliers or anomalies in a database. Outliers are an observation that is numerically distant from the rest of the data. Since outliers and anomalies are highly unlikely, they can be indicative of bad data or malicious behaviour. Bad data interns produce falls outcome. Examples of bad data include skewed data values resulting from measurement error, or erroneous values resulting from data entry mistakes, missing values, missing data. Missing data, or Missing values, occur when no data value is stored for the variable in the current observation. Common solution is either ignore the missing data is called as marginalization or fill in the missing values is called as imputation. Imputed values are treated as just as reliable as the truly observed data, but they are only as good as the assumptions used to create them.

Techniques of dealing with missing values can be classified into three categories [7], [12]. 1) Deletion, 2) Learning without handling of missing values, and 3) Missing value imputation The first technique is to simply omit those cases with missing values and only to use the remaining instances to finish the learning assignments [13].The deletion is classified in two categories they are, i) List wise or Case deletion ii) Pair wise deletion.

The second approach is to learn without handling of missing data, such as Bayesian Networks method, Artificial Neural Networks method 15, the methods in [10]. Missing data imputation is a procedure that replaces the missing values with some possible values, such as [11], [12]. A variety of methods have been developed with great success on dealing with missing values in data sets with uniform attributes. (their independent attributes are all either continuous or discrete).

However, these imputation algorithms cannot be applied to many real data sets, such as equipment maintenance databases, industrial data sets, and medical databases, because these data sets are often with continuous, discrete and categorical independent attributes. These heterogeneous data sets are referred to as mixed-attribute data sets and their independent attributes are called as mixed independent attributes. It advocates that a missing datum is imputed if and only if there are some complete instances in a small neighbourhood of the missing datum, otherwise, it should not be imputed. Further, a Non parametric iterative estimator is proposed to utilize all the available observed information, including observed information in incomplete instances with missing values.

In this paper, we present an imputation overview in that we discuss the problem of imputing the mixed attribute datasets and then we see how this problem can be solved by implementing the nonparametric iterative imputation method for estimating missing values in mixed-attribute data sets and also in clustered data sets (only clustering the continuous attributes).

II. IMPUTATION OVERVIEW

Missing data imputation is a procedure that replaces the missing values with some possible values. Imputed values are treated as just as reliable as the truly observed data, but they are only as good as the assumptions used to create them. The imputation consists of many types. In that some types of imputations are, (i) Single Imputation, (ii) Partial Imputation and (iii) Multiple Imputation, (iv) Iterative Imputation. According to our paper, previous work has been handling the missing values in heterogeneous data sets using semi parametric way of iterative imputation method [15].

Normally this method is inconsistent in some datasets. To avoid this problem, and also to improving the efficiency, the non parametric way is possible. So the proposed work based on handling the missing values in heterogeneous datasets and also in clustered data sets (only continuous attributes) using non parametric way of iterative imputation.

III. OBJECTIVE OF OUR WORK

The proposed work bring out the new setting of missing data imputation, i.e., imputing missing data in data sets with mixed attributes (their independent attributes are of different types i.e. the datasets consists of both discrete and continuous attributes), referred to as imputing mixed-attribute data sets in [13]. Although many real applications are in this setting, there is no estimator designed for imputing data sets with heterogeneous attributes. It first proposes two reliable estimators for discrete and continuous missing target values, respectively. Imputing mixed-attribute data sets can be taken as a new problem in missing data imputation because there is no estimator designed for imputing missing data in mixed attribute data sets.

The challenging issues include, such as how measuring the relationship between instances (transactions) in a mixed-attribute data set, and how to construct hybrid estimators using the observed data in the data set. To address the issue, this research proposes a nonparametric iterative imputation method based on a mixture kernel for estimating missing values in mixed-attribute data sets. A mixture of kernel functions (a linear combination of two single kernel functions, called mixture kernel) is designed for the estimator in which the mixture kernel is used to replace the single kernel function in traditional kernel estimators. These estimators are referred to as mixture kernel estimators.

Based on this, two consistent kernel estimators are constructed for discrete and continuous missing target values, respectively, for mixed-attribute data sets. Further, a mixture-kernel-based iterative estimator is proposed to utilizes all available observed information, including observed information in incomplete instances (with missing values), to impute missing values, whereas existing imputation methods use only the observed information in complete instances (without missing values).To improve the accuracy cluster based non-parametric iterative imputation is proposed. Fig 1 shows that proposed system architecture. It initially considers the database with missing values, and then identifies the attribute type by using appropriate techniques to find attributes of either continuous or discrete attribute. If it is a continuous attribute Mean Pre-Imputation is applied otherwise Mode Pre-Imputation is applied. This is the basic step of imputation techniques. Then by using the pre imputed data sets kernel function is applied separately to both the attributes.

This imputation is said to be single imputation. Mixture kernel function is obtained by integrating both the discrete and continuous kernel function. Estimated value is calculated by the standard formulas. Finally Iterative kernel estimator is applied separately for continuous as well as discrete attributes to get final value for imputation. This data will be imputed in the missing data set to make it as a complete dataset. Further to improve the accuracy clustering algorithm is applied. This clustered data set considered as a first step of the framework.

There are five steps in our proposed system. They are (i) Data Preparation (ii) Single Imputation Using Kernel Function (iii) Constructing the Estimator and Iterative Imputation (iv) Pre-Processing dataset Using Clustering Algorithm (v) Performance Analysis

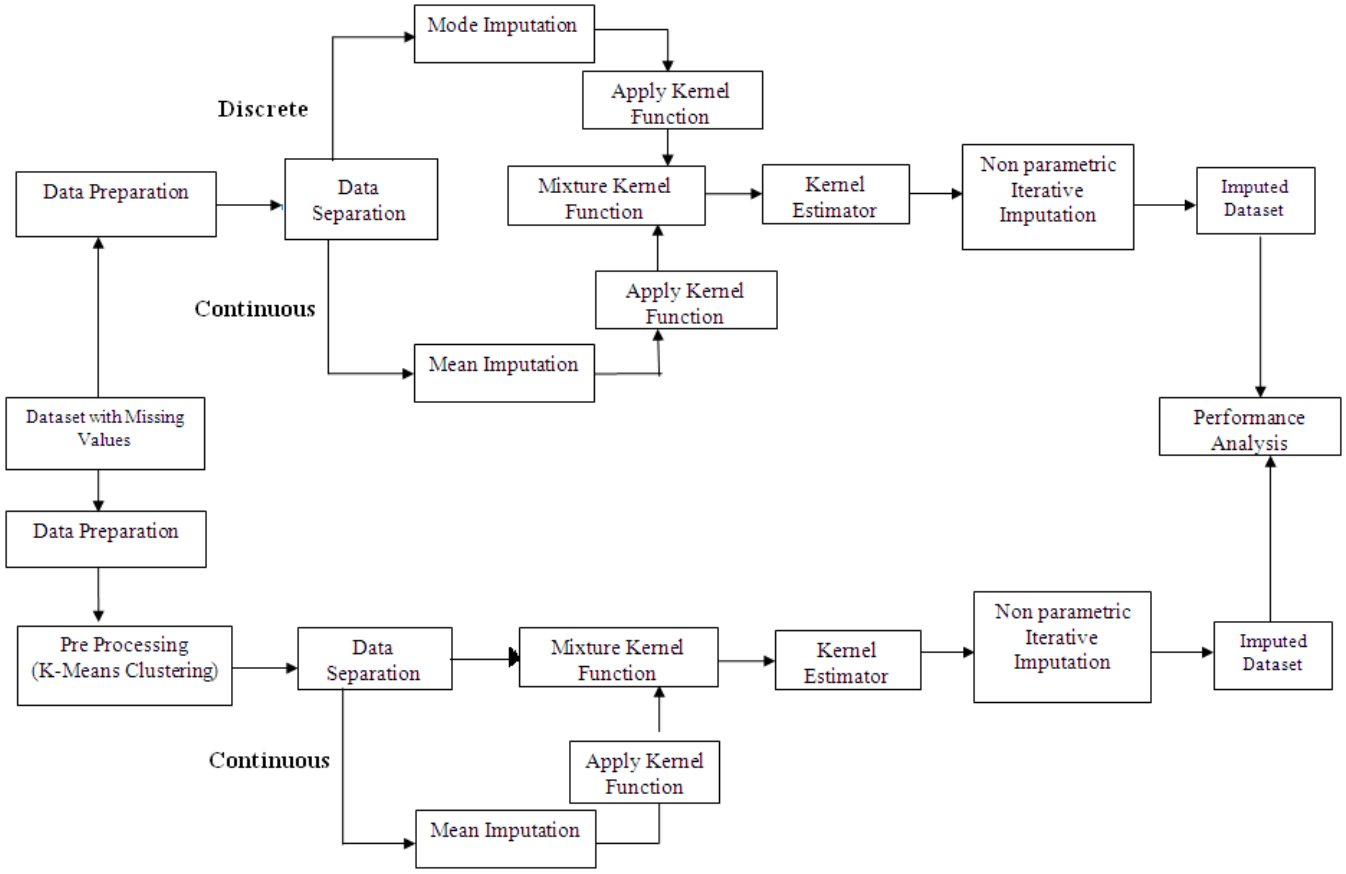

Fig. 1.System Architecture for Proposed System

A. Data Preparation

In this module, from the input heterogeneous data set the records with missing values will be identified and categorized based on attribute type of missing values, attributes are grouped. Mean and mode value for continuous and discrete category is calculated separately. Basic imputation has been done with this calculated value.

B. 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

$$L\left(X_{t,i}^{d}, x_{t}^{d}\right)=\begin{cases}1 \text{if } X_{t,i}^{d}=x_{t}^{d}\\ \lambda \text{if } X_{t,i}^{d}\neq x_{t}^{d}\end{cases} \quad \text{---------- (B.1)}$$

Where,

X_i^d -- 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

$K(x - X_i/h)$ ---------- (B.2)

K(.) is a mercer kernel, i.e., positive definite kernel.

3) Mixture Kernel Function

$K_{h,\lambda,ix} = K(x-X_i/h)\, L(X_i^d, x_i^d, \lambda)$ ----------(B.3)

Where,

h->0 and λ->0 (λ, h is the smoothing parameter for the discrete and continuous kernel function , respectively),

$K_{h,\lambda,ix}$ -- symmetric probability density function.

$K(x-X_i/h)$ -- Continuous Kernel Function

$L(X_i^d, x_i^d, \lambda)$ -- Discrete Kernel Function

C. Constructing the Estimator and Iterative Imputation

Construct the estimator, separately for both attributes. 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≥1) 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

1) Kernel estimator for Continuous Missing attributes

$$\hat{m}(x)=\frac{n^{-1}\sum_{i=1}^{n} Y_i K_{h,\lambda,ix}}{n^{-2}\sum_{i=1}^{n} K_{h,\lambda,ix}+n} \quad \text{---------- (C.1)}$$

where ,

item n^{-2} m(x) -- only used for avoiding the denominator to be 0.

Y_i -- Denoting the ith Missing Value.

2) Kernel estimator for Discrete Missing attributes

When the missing value m(X) is in a discrete attribute ,the estimator is, let $D_{m(x)} = (0,1,.....,c_u-1)$ denote the range of m(x).One could estimate m(x) by,

$$\hat{m}(x)=\frac{n^{-1}\sum_{i=1}^{n} Y_i K_{h,\lambda,ix}}{n^{-2}\sum_{i=1}^{n} K_{h,\lambda,ix}+n^{-2}} + \frac{\lambda n^{-1}\sum_{i=1}^{n}\sum_{Y\in Dy, y\neq Y_i} K_{h,\lambda,ix}}{n^{-2}\sum_{i=1}^{n} K_{h,\lambda,ix}+n} \quad \text{---------- (C.2)}$$

Where $l(Y_i, y, \lambda)=1$ if $y= Y_i$ and λ if $y\neq Y_i$.

3) Iterative Kernel Estimator for continuous Missing attributes

$$\hat{m}(x)=\frac{n^{-1}\sum_{i=1}^{n} Y_i^{t} K_{h,\lambda,ix}}{n^{-2}\sum_{i=1}^{n} K_{h,\lambda,ix}+n} \quad \text{---------- (C.3)}$$

Where,

Y_i^t -- tth imputation of the ith Missing Value

4) Iterative Kernel Estimator for discrete Missing attributes

$$\hat{m}(x)=\frac{\sum_{i=1}^{n}\sum_{y\in Dy, y\neq Y_i} I(Y_i, y, \lambda) y_t K_{h,\lambda}}{\sum_{i=1}^{n} K_{h,\lambda}} \quad \text{-------- (C.4)}$$

Where,

$y_i^t = \{ Y_i$ if $\delta i= 0$ or i=1,.....,r,

Y^t_i if δ_i =1 or i=r+1,.....,n

In particular, Y^t_i is the best common class in the discrete target variable, and

Y^i_t =0, i = r+1,...., n.

D. Pre-Processing Data set using cluster Algorithm

Before sending data to the data preparation module, clustering take place to group similar data object. By applying the formula mentioned below, the data sets are grouped in two sets with respect to every attribute.

E. Performance Analysis

Imputed values without using clustering and using k-means clustering are compared. The performance analysis takes place by using both the method

IV. CONCLUSION AND FUTURE WORK

Imputation is the best solution for handling the Missing values. Missing data imputation is a procedure that replaces the missing values with some possible values. But this is not appropriate solution for discrete and categorical missing values. A consistent kernel regression has been proposed for imputing missing values in a mixed-attribute data set and uses the techniques of data driven method for bandwidth selection. The data-driven (i.e., automatic) bandwidth selection procedures are not guaranteed always to produce good results due to perhaps the presence of outliers or the rounding/discretization of continuous data, among others. The nonparametric estimators are proposed against the case that data sets have both continuous and discrete independent attributes and also in clustered data sets. It utilizes all available observed information, including observed information in incomplete instances (with missing values), to impute missing values, whereas existing imputation methods use only the observed information in complete instances (without missing values).That is the work includes exploring a framework for non parametric iterative imputation based on mixture kernel estimation in both mixture data sets and also in clustered data sets (only continuous attributes).In future work furthermore, this paper could be extended to handle this imputation process in more than one missing value in a single attribute.

REFERENCES

[1] A. Dempster, N.M. Laird, and D. Rubin, "MaximumLikelihood from Incomplete Data via the EM Algorithm," J. Royal Statistical Soc., vol. 39, pp. 1-38, 1977.
[2] D.Rubin, Multiple Imputation for Nonresponse in Surveys. Wiley, 1987.
[3] H. Bierens, "Uniform Consistency of Kernel Estimators of a Regression Function under Generalized Conditions," J. Am. Statistical Assoc., vol. 78, pp. 699-707, 1983.
[4] J. Han and M. Kamber, Data Mining Concepts and Techniques, second ed. Morgan Kaufmann Publishers, 2006
[5] J. Racine and Q. Li, "Nonparametric Estimation of Regression Functions with Both Categorical and Continuous Data," J. Econometrics, vol. 119, no. 1, pp. 99-130, 2004.
[6] J.R. Quinlan, "Unknown Attribute values in Induction," Proc. Sixth Int'l Workshop Machine Learning, pp. 164-168, 1989.
[7] Y.S. Qin et al., "Semi-Parametric Optimization for Missing DataImputation," Applied Intelligence, vol. 21, no. 1, pp. 79-88, 2007.
[8] M. Alirera, "A Novel Framework for Imputations of Missing Values in Databases," IEEE Transactions Vol 37,No. 5, 2007.
[9] Q.H. Wang and R. Rao, "Empirical Likelihood-Based Inference under Imputation for Missing Response Data," Annals of Statistics, vol. 30, pp. 896-924, 2002.
[10] R. Caruana, "A Non-Parametric EM-Style Algorithm for Imputing Missing Value," Artificial Intelligence and Statistics, Jan. 2001.
[11] R. Little and D. Rubin, Statistical Analysis with Missing Data, second ed. John Wiley and Sons, 2002.
[12] S.C. Zhang, "Par imputation: From Imputation and Null-Imputation to Partially Imputation," IEEE Intelligent Informatics Bull., vol. 9, no. 1, pp. 32-38, Nov. 2008.
[13] Xiaofeng Zhu, Shichao Zhang Senior Member, IEEE,(2011)"Missing Value Estimation for Mixed Attribute Data Sets" , IEEE Transactions on Knowledge and Data Engineering, vol.23,No.1.2011
[14] Y.S. Qin et al., "POP Algorithm: Kernel-Based Imputation to Treat Missing Values in Knowledge Discovery from Databases," Expert Systems with Applications, vol. 36, pp. 2794-2804, 2009.
[15] Zhang.S.C, "Estimating Semi-Parametric Missing Values with Iterative Imputation," International Journal of Data Warehousing and Mining, pp. 1-10.2010

AUTHORS

First Author – Sujatha.R, M.E, Assistant Professor, Shivani Engineering College, Trichy-9, sujamole@gmail.com.