Volume 3, Issue 5, May 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Imputation Method for Missing Value Estimation of Mixed-Attribute Data Sets

Santosh Dane
M. Tech Student
Department of IT
SGGS IE&T, Nanded, India.

Dr. R. C. Thool
Professor
Department of IT
SGGS IE&T, Nanded, India.

Abstract*— Missing data imputation is an important issue in learning from incomplete data. Various techniques have been developed with great successes on dealing with missing values in data sets with homogeneous attributes (their independent attributes are all either continuous or discrete). We propose a new setting of missing data imputation that is by imputing missing data in data sets with heterogeneous attributes thus by contributing both continuous and discrete data. We propose two consistent estimators for discrete and continuous missing target values. Then mixture kernel based iterative estimator and spherical kernel based iterative estimator is advocated to impute mixed-attribute data sets.

Keywords*— Data Mining, Missing data, Classification, Kernel Function, Nonparametric.

I. INTRODUCTION

Missing data is a common problem for data quality. Rate of less than 1% missing data is generally considered trivial, 1-5% manageable. However, 5-15% requires sophisticated methods to handle, and more than 15% may severely impact any kind of interpretation [17]. Several methods have been proposed in the literature to treat missing data. Missing data imputation aims at providing estimations for missing values by reasoning from observed data. Because missing values can bias the impacts on the quality of learned patterns or/and the performance of classifications, missing data imputation has been a key issue in learning from incomplete data. Various techniques have been developed with great successes on dealing with missing values in data sets with homogeneous 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 gene databases, because these data sets are often with both continuous and discrete independent attributes [21]. This paper studies a new setting of missing data imputation, i.e., imputing missing data in mixed-attribute data sets.

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 to measure the relationship between instances in a mixed-attribute data set.
- How to construct hybrid estimators using the observed data in the data set.
- Impute the missing values.
- Check the accuracy of filled dataset.

To solve these problems, this research proposes a nonparametric iterative imputation method based on a mixture kernel for estimating missing values in mixed-attribute data sets. It first constructs a kernel estimator to infer the probability density for independent attributes in a mixed-attribute data set. And then, 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 utilize all the available observed information, Including observed information in incomplete instances (with missing values) [3].The proposed algorithm is experimentally evaluated in terms of root mean squared error (RMSE), classification accuracy and the convergence speed of the algorithm, compared with extant methods, such as the nonparametric imputation method with a single kernel, the nonparametric method for continuous attributes, and frequency estimator (FE). These experiments were conducted on UCI data sets and a real data set at different missing ratios.

II. PATTERN OF MISSING VALUE

The pattern of missing values as per Little and Rubin [11] classified into three categories.

1. Missing completely at random (MCAR): This is the highest level of randomness. It occurs when the probability of missing value for an attribute does not depend on either the observed values or the missing data.
2. Missing at random (MAR): When the probability of missing of an instance having a missing value for an attribute may depend on known values, but not on the value of the missing data itself.

3. Not missing at random (NMAR): When the probability of an instance having a missing value for an attribute could depend on the value of that attribute.

III. REASONS BEHIND MISSING VALUE

We extract data from different heterogeneous sources. These data sets are collected at central repository. When we collect these data from different sources data may be missing due to following reasons.

- Due to lack of response.
- Measurements not made by human.
- Machine error.
- Equipment malfunction.
- Inconsistent with recorded data.
- Data may not be considered important.
- History of data not registered.
- By corruption in transmission or storage.
- Data collection not done properly.
- Mistakes due to data entry operator.
- Unanswered questions.

IV. RELATED WORK

A. Treatment of Missing Data

1. Ignoring and Discarding Data: These methods determine the missing data on each instance and delete those instances. Other thing is determining each attribute/instances and remove that whole attribute/instances which having high level of missing data. This method is applicable only when the dataset is MCAR.
2. Parameter Estimation: This method is used to find the parameters for the complete data. This method uses the Expectation-maximization algorithm for handling the parameter estimation of the missing data.
3. Imputation Technique: This is one kind of procedure in which replaces the missing values based on estimated values.

B. Research into Missing Value Imputation

To impute missing value, commonly used techniques are as follows:

1. Parametric Regression Imputation Methods

The parametric methods, such as linear regression [10], [11], and [12], are superior while the data set are adequately modelled. However, in real applications, it is often impossible to know the distribution of the data set. Therefore, the parametric estimators can lead to highly bias, and the optimal control factor settings may be miscalculated.

2. Non Parametric Regression Imputation Methods

For this case, nonparametric imputation method [13], [14], [15] can provide superior fits by capturing the structure of the data set. However, these imputation methods are designed for either continuous or discrete independent attributes. For example, the well-established imputation methods in [9], [15] are developed for only continuous attributes. And these estimators cannot handle discrete attributes well. Some methods, such as C4.5 algorithm [13], association-rule-based method [16], and rough-set-based method [17], are designed to deal with only discrete attributes. In these algorithms, continuous attributes are always makes discrete before imputing. This possibly leads to a loss of useful characteristics of the continuous attributes. There are some conventional imputation approaches, such as [6], [1], and [11], designed for discrete attributes using a "frequency estimator" in which a data set is separated into several subsets or "cells." However, when the number of cells is large, observations in each cell may not be enough to non parametrically estimate the relationship among the continuous attributes in the cell. When facing with mixed independent attributes, some imputation methods take the discrete attributes as continuous ones, or other methods are used.

However, all the above methods were designed to impute missing values with only the observed values in complete instances, and did not take into account observed information in incomplete instances. On the other hand, all the above methods are designed to impute missing values one time. John et al. [4] thought that iterative approaches impute missing values several times and can be usefully developed for missing data imputation. Zhang et al. [5] thought it is necessary to iteratively impute missing values while suffering from large missing ratio. Hence, many iterative imputation methods have been developed, such as the Expectation-Maximization (EM) algorithm which is a classical parametric method. Zhang et al. [5] and Caruana [7] proposed nonparametric iterative methods but based on a k-nearest neighbourhood framework. In this paper, the proposed iterative imputation method is a nonparametric model specially designed for those data sets with both continuous and discrete attributes, which is based on a kernel regression imputation framework.

C. Kernel Function Selection

Kernel function is popularly used in building imputation models, such as [15], [8] and [3], denoted by kernel imputation. When kernel imputation method is employed to impute missing values, it usually consists of two parts: kernel function selection and bandwidth adjustment. During the process for selecting kernel functions, what we need to consider is not only the ability to learn from the data (i.e., "interpolation"), but also the ability to predict unseen data (i.e., "extrapolation"). For example, a global kernel (such as the polynomial kernel) has better extrapolation abilities at lower order degrees, but requires higher order degrees for good interpolation. A local kernel (such as the RBF kernel or

Gaussian kernel) has good interpolation abilities, but fails to provide longer range extrapolation. Jordan [18] demonstrated that a mixed kernel, a linear combination between poly kernel and Gaussian kernel, gives the extrapolation and interpolation much better than either a local kernel or a global kernel. In this paper, a mixture of kernels is employed to replace the single kernel in continuous kernel estimator.

V. SYSTEM DESIGN

A. *Non-Parametric Iterative Imputation Method*

A Non-Parametric Iterative Imputation Method is proposed for imputing missing values in mixed attribute data sets. Definition 1 and 2 gives discrete and continuous kernel function respectively [3]. Definition 3 gives mixture kernel function [8]. Definition 4 and 5 construct two consistent kernel estimators are designed for both continuous and discrete missing target values in mixed attribute data sets. Definition 6 and 7 construct Iterative kernel estimator is designed for both continuous and discrete missing target values in mixed attribute data sets. Then Definition 8 gives iterative mixture kernel estimator is designed to utilize all the available information's in the incomplete instance.

Definition 1. Discrete kernel function

The discrete kernel function is as follows:

$$L\left(X_{i,}^{d} x_i^{d}, \lambda\right) = \lambda^{d_{x_i} x} + \delta_{x_i^x}$$

Definition 2. Continuous kernel function

For a $K \times 1$ vector with continuous values, such as, $x \in R^{p}$, then its kernel function is $K\left(x - \frac{X_i}{h}\right)$, and the K(.) is a mercer kernel, i.e., positive definite kernel.

Definition 3. Mixture kernel function

With integrating the discrete and continuous kernel functions, a mixture kernel function is constructed as follows:

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

Where $h \to 0$ and $\lambda \to 0$ (λ , h is the smoothing parameter for the discrete and continuous kernel functions, respectively), and $K_{h,\lambda,ix}$ is a symmetric probability density function.

Definition 4. Estimator for Missing Continuous Attribute

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

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

Definition 5. Estimator for Missing Discrete Attribute

$$m(x) = \frac{n^{-1} \sum_{i=1}^{n} Y_i K_{h,\lambda.ix}}{n^{-1} \sum_{i=1}^{n} K_{h,\lambda.ix} + n^{-2}} + \lambda \frac{n^{-1} \sum_{i=1}^{n} \sum_{y \in D_y y \neq y_i} k_{n\lambda}}{n^{-1} \sum_{i=1}^{n} K_{h,\lambda}}$$

Where $l(Y_{i,y,\lambda}) = 1$ if $y = Y_i$ and λ if $y \neq Y_i$

Definition 6. Iterative Kernel Estimator for Continuous Target Attribute

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

Definition 7. Iterative Kernel Estimator for Discrete Target Attribute

$$m_t(x) = \frac{\sum_{i=1}^{n} \sum_{y \in D_y y \neq y_i} l\left(Y_i^{d}, y_t, \lambda\right) yt K_{n\lambda}}{\sum_{i=1}^{n} K_{h,\lambda}}$$

Definition 8. Iterative Kernel Estimator for Discrete Target Attribute

Let $K_{poly} = \left(<x, x_i>+1\right)^q, K_{rbf} = \exp\left(-\left(x-x_i\right)^2 / \sigma^2\right)$, a linear mixture kernel function is defined as follows:

$$K_{mix} = \rho K_{poly} + (1-\rho) K_{rbf}$$

B .Single Kernel Imputation

1. Polynomial Kernel Imputation

The Polynomial kernel is a non-stationary kernel. Polynomial kernels are well suited for problems where all the training data is normalized. As a global kernel, the polynomial kernel is good at capturing general trends and extrapolation behavior. A global kernel has better extrapolation abilities and it requires higher order degrees for good interpolation. The extrapolation behavior of the model becomes erratic and shows sudden increases or decreases in the response surface when the value of q is too high. So, a lower degree for the polynomial kernel may be chosen. Target Variable.

2. RBF Kernel Imputation

A local kernel such as the RBF kernel or Gaussian kernel has good interpolation abilities, but fails to provide longer range extrapolation. The Gaussian kernel is an example of radial basis function kernel. Since the RBF-kernel is a very powerful kernel for modeling local behavior, it will not need much of its effects in order to see a huge improvement in the model.

C. Mixture Kernel Imputation

A global kernel (such as the polynomial kernel) can present better extrapolation at lower order degrees, but need more higher order degrees for receiving a good interpolation. And a local kernel has better interpolation, but fails to provide stronger extrapolation. They also demonstrated that a mixture of kernels can lead to much better extrapolation and interpolation than using either the local or global kernels [18]. A single kernel is usually used for continuous target variables in traditional imputation techniques. Whereas Mixture Kernel replaces the single kernel in continuous missing attribute estimator.

VI. ALGORITHM DESIGN

NIIA (Nonparametric Iterative Imputation Algorithm) method is designed for imputing iteratively missing target values [16]. The NIIA method imputes missing values several times until the algorithm converges. First iteration, all complete instances are used to estimate missing values. The information within incomplete instance is utilized since the second iteration. In the first iteration of imputation in the NIIA algorithm, all the missing values are imputed using the mean for continuous attributes and the 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. Imputing with the mean (or mode) will be valid if and only if the data set is chosen from a population with a normal distribution. This is usually impossible for real applications because the real distribution of a data set is not known in advance. On the other that a single imputation cannot provide valid standard errors and confidence intervals, since it ignores the uncertainty implicit in the fact that the imputed values are not the actual values. Therefore, running extra iteration imputations based on the first imputation is reasonable and necessary for better dealing with the missing values.

//the first imputation

FOR each MVi in Y

MV_i^t=mode$\left(S^r inY\right)$; //if Y is a discrete variable

MV_i^t=mean$\left(S^r inY\right)$; //if Y is a continuous variable

END FOR

//t-th iteration of imputation (t>1)

t=1;

REPEAT

t++

FOR each MVi in Y

$MV_i^t = MV_i^{t-1}$, pε Sm, p=1… m, p≠i

MV_i^t is got based on definition 6 //if discrete variable

MV_i^t is got based on definition 7 //if continuous variable

END FOR

UNTIL

| CAt-CAt-1 |≥ ε // if discrete variable

Convergence or cycling //if continuous variable

//finishing the iterative imputation

OUTPUT

//t is iterative times

//Completed dataset;

VII. PROPOSED METHOD

The main purpose of the proposed system is to impute the missing values in a mixed attribute data set. Initially a nonparametric iterative imputation method is presented. In this method kernel functions for the discrete attributes and continuous attribute are studied and then a mixture kernel function is proposed by combining a discrete kernel function with a continuous one. Further an estimator is constructed based on the mixture kernel.

The higher order kernel here we propose for missing value estimation is spherical kernel and polynomial kernel. These kernels provide maximum functionalities on higher dimensional space. Here we are proposing new mixture kernel function i.e. spherical kernel function is in linear combination with polynomial kernel. In another method mixture kernel function i.e. spherical kernel function is in linear combination with RBF kernel. These mixture kernel functions provide better extrapolation and interpolation.

VIII. EXPERIMENTAL STUDY AND RESULT

We considered several data sets from real applications and data sets taken from the UCI data set in [7] (see Table 1) in this section. Some of the data sets taken from UCI repository are completed data sets. For experimental purpose we generate some missing value in the percent of 10,20,30,40 for each data set.

TABLE I
DATA SET USED FOR EXPERIMENTS

Name	No of Attribute	No of instances
Auto-mpg	26	205
Housing	14	506
Abalone	8	4177
annealing	38	798
Vowel	10	528

The proposed method is evaluated with some traditional algorithm like Frequency estimator (FE), polynomial kernel, and RBF kernel. The result demonstrates that the proposed approach is better than these existing imputation methods in terms of classification accuracy and root mean square error (RMSE) at different missing ratios.

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 for better extrapolation and interpolation. This kernel function used in this method utilizes all available observed information, including observed information in incomplete instances to impute missing values, whereas existing imputation methods use only the observed information in complete instances. The experimental results have demonstrated that the proposed algorithms outperform the existing ones for imputing both discrete and continuous missing values. Missing values filled with better accuracy leads to better results.

REFERENCES

[1] I.A. Ahamad and P.B.Cerrito, "Nonparametric Estimation of Joint Discrete-Continuous Probability Densities with Applications," J. Statistical Planning and Inference, vol. 41, pp. 349-364, 1994.

[2] M.A. Delgado and J.Mora, "Nonparametric and Semi-Parametric Estimation with Discrete Regressors," Econometrica, vol. 63, pp. 1477-1484, 1995.

[3] 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.

[4] G. John et al., "Ir-Relevant Features and the Subset Selection Problem," Proc. 11th Int'l Conf. Machine Learning, W. Cohen and H. Hirsch, eds., pp. 121-129, 1994.

[5] C. Zhang, X. Zhu, J. Zhang, Y. Qin, and S. Zhang, "GBKII: An Imputation Method for Missing Values," Proc. 11th Pacific-Asia Knowledge Discovery and Data Mining Conf. (PAKDD '07),pp. 1080-1087, 2007

[6] H. Bierens,"Uniform Consistency of Kernel Estimators of a Regression Function under Generalized Conditions," J. Am. Statistical Assoc., vol. 78, pp. 699-707, 1983.

[7] R. Caruana, "A Non-Parametric EM-Style Algorithm for Imputing Missing Value," Artificial Intelligence and Statistics, Jan. 2001. [10] K. Cios and L. Kurgan, "Knowledge Discovery in Advanced Information Systems," Trends in Data Mining and Knowledge Discovery, N. Pal, L. Jain, and N. Teoderesku, eds., Springer, 2002.

[8] 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.

[9] J. Barnard and D. Rubin, "Small-Sample Degrees of Freedom with Multiple Imputation," Biometrika, vol. 86, pp. 948-955, 1999.

[10] A. Dempster, N.M. Laird, and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," J. Royal Statistical Soc., vol. 39, pp. 1-38, 1977.

[11] R. Little and D. Rubin, Statistical Analysis with Missing Data, second ed. John Wiley and Sons, 2002.

[12] D. Rubin, Multiple Imputations for Nonresponse in Surveys. Wiley, 1987.

[13] J.R. Quinlan, C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.

[14] 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.

[15] S.C. Zhang, "Parimputation: From Imputation and Null-Imputation to Partially Imputation," IEEE Intelligent Informatics Bull., vol. 9, no. 1, pp. 32-38, Nov. 2008.

[16] W. Zhang, "Association Based Multiple Imputation in Multivariate Data Sets: A Summary,"Proc. Int'l Conf. Data Eng. (ICDE),p. 310, 2000.

[17] C. Peng and J. Zhu, "Comparison of Two Approaches for Handling Missing Covariates in Logistic Regression," Educational and Psychological Measurement, vol. 68, no. 1, pp. 58-77, 2008.

[18] E.M. Jordan, "Development of Robust Inferential Sensors: Industrial Application of Support Vector Machines for Regression," PhD thesis, Technical University Eindhoven, 2002.

[19] C.Blake and C. Merz UCI Repository of Machine Learning Database, http://www.ics.uci.edu/~mlearn/MLResoesitory. Html, 1998.

[16] Shichao Zhang, Zhi Jin and Zhuoing Xu, "Missing Value Estimation for Mixed-attribute Data Sets", IEEE Trans.Knowledge and Data Eng., vol.23, no.1, Jan 2011.

[17] Acuna E. and Rodriguez C., The treatment of missing values and its effect in the classifier accuracy. In D. Banks, L. House, F.R. McMorris, P. Arabie, W.Gaul (Eds).Classification, Clustering and Data Mining Applications. Springer-Verlag Berlin-Heidelberg, 639-648. 2004.

[18] K. Lakshminarayan et al., "Imputation of Missing Data in Industrial Databases," Applied Intelligence, vol. 11, pp. 259-275, 1999.

