View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by CiteSeerX

International Journal of Computer Trends and Technology- volume3Issue2- 2012

Imputation Framework for Missing Values

K. Raja [#1], G. Tholkappia Arasu [#2], Chitra. S. Nair *[3]

[1] Professor, Dept. CSE, Adhiyamaan College of Engineering, Hosur, TamilNadu, India.
[2] Principal, Jayam College of Engineering and Technology, Dharmapuri, TamilNadu,India.
[3] PG Scholar, Dept. CSE, Adhiyamaan College of Engineering, Hosur, TamilNadu, India.

***Abstract*—Missing values may occur for several reasons and affects the quality of data, such as malfunctioning of measurement equipment, changes in experimental design during data collection, collation of several similar but not identical datasets and also when respondents in a survey may refuse to answer certain questions such as age or income. Missing values in datasets can be taken as a common problem in statistical analysis. This paper first proposes the analysis of broadly used methods to treat missing values which are either continuous or discrete. And then, an estimator is advocated to impute both continuous and discrete missing target values. The proposed method is evaluated to demonstrate that the approach is better than existing methods in terms of classification accuracy.**

***Keywords*— Classification, data mining, methodologies**

1. INTRODUCTION

Data quality is a major concern in Machine Learning and other related areas such as Data Mining and Knowledge Discovery from Databases. Various techniques have been developed with great successes with dealing with missing values in datasets with homogeneous attributes. Imputation is a term that denotes a procedure that replaces the missing values in a dataset by some estimated values. However, the traditional imputation algorithms cannot be applied to many real datasets. The real datasets are those whose probability distribution is not known in advance, such as equipment maintenance databases, industrial data sets, and gene databases. These data sets are often with both continuous and discrete independent attributes [1]. These heterogeneous data sets are referred to as mixed-attribute datasets in this paper. This work includes the surveys of most widely used methods for missing data treatment.

The challenging issues include, such as how to measure the relationship between instances in a mixed-attribute datasets. To address the issue, this work proposes a nonparametric iteration imputation method based on a mixture kernel functions. A grid search method is used to obtain optimal bandwidth for the proposed mixture kernel estimators, instead of data-driven method in [2]. A mixture kernel formed by combining both local kernel and global kernel.

The missing independent attributes are imputed iteratively. The resulting dataset is an imputed dataset with higher accuracy than several other imputation methods.

The paper is organized as follows: Section 2 begins with related work as proposed in [3] to classify the degree of randomness of missing data and most widely used methods for missing data treatment. Section 3 describes the nonparametric imputation methods. Section 4 describes the nonparametric regression framework. Section 5 describes Kernel regression imputation framework. And Section 6 presents the conclusion of this paper.

2. Related work

2.1 Missing data mechanisms

Missing data randomness can be divided into three classes as proposed by [4]:

2.1.1 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.

*2.1.2 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.

*2.1.3 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.

2.2 Missing data techniques

In general way, missing data treatment methods can be divided into the following three categories [4]:

2.2.1 Ignoring and discarding data. There are two main ways to discard with missing values. The first way is known as *complete case analysis*. This method consists of discarding all cases with missing data. The second method is known as *discarding instances and/or attributes*. This method consist of determining the extent of missing data on each instance and attribute, and delete the instances and/ or attributes with high levels of missing data. Both methods should be applied only if missing data are MCAR, because missing data that are not MCAR have non-random that can bias the results.

2.2.2 Parameter estimation. Maximum likelihood procedures are used to estimate the parameters of a model defined for the complete data. Maximum likelihood procedures that use variants of the Expectation-Maximization algorithm [5] that can handle parameter estimation in the presence of missing data.

2.2.3 Imputation. Imputation is a class of procedures that aims to fill in the missing values with estimated ones. The objective is to employ known relationships that can be identified in the valid values of the data set to assist in estimating the missing values. This paper focus on imputation of mixed attribute missing data.

2.3 Missing data imputation methods

Missing data imputation is a procedure that replaces the missing values with some plausible values. A description of the widely used imputation methods following the idea from [6], [7], [8]:

2.3.1 Case substitution. This method is typically used in sample surveys. Missing values are replaced by a global constant(NULL) but this method causes inconsistencies for data mining algorithms if frequent occurrence of such constants in training sets and also leads to discrepancies on the analysis. Used only if missing category is Missing Completely at Random.

2.3.2 Global imputation based on missing attribute. If we look at the other values taken by a variable with a missing item, and we find some of them are most frequent than the others, we may decide to use this fact to assign the most frequent values to a missing one. In particular *mean or mode imputation* used o fill the holes. Used only if the statistical property of missing data should be known in advance or probability density function of a datasets are of normal distribution.

2.3.3 Global imputation based on non-missing attributes. If there are correlations between missing and non-missing variables, we may learn these relationships and use them to predict missing values. Thus this method also called as prediction model, k-nearest neighbor approach. The imputation processing time depends only on the construction of prediction model, once a prediction model is constructed, and then the missing values are imputed in a constant time. On the other hand, the main drawbacks of this approach are that, if there is no relationship exists among one or more attributes in the dataset and the attributes with missing data, then the prediction model will not be suitable to estimate the missing value. If there is no correlation among the attributes then the method is not good.

2.3.4 Local imputation. Hot Deck imputation is a procedure where the imputation comes from other records in the same data, and these are chosen on the basis of the incomplete record [9].

A Hot Deck imputation acts in two steps:

1. Records are subdivided into classes.
2. For every incomplete record, the imputing values are chosen on the basis of the records in the same class. This can be done by calculating the mean or mode of the attribute within a class or cluster. Good only for MCAR missing.

2.4 Parametric Imputation

Commonly used methods to impute missing values include parametric and nonparametric regression imputation methods

The Parametric method, such as *imputation by regression [5]*, missing values are treated as dependent variables, and a regression is performed to impute missing values. Linear regression and logistics regression are typical choices. In linear regression, if the relationship is not linear may lead to wrong results

Attribute1	Attribute2	Attribute3	Attribute3
data	?	?	data
data	data	?	?
data	?	data	?

Fig.1. Database containing missing values

3. Nonparametric Imputation

The nonparametric methods are designed for either continuous attributes or discrete independent attributes. Continuous attributes are sometimes called numeric attributes [2], which are either real or integer valued. Discrete attributes are called nominal attributes [2], have values that are distinct symbols. The values themselves serve just as labels or names.

3.1 Discrete Attributes Imputation Methods

The well established imputation methods such as [11], [12], and [13], developed for only discrete attributes.

3.1.1 C4.5 algorithm: C4.5 consists of a collection of training cases, each having a tuple of values for a fixed set of attributes or independent variable and a class attribute or dependent variable. The goal is to learn from the training cases that maps from the attribute values to a predicted class. Decision tree structure is used which consists of a leaf node and a test node. Missing values are common occurrences in data and affect the way a decision tree is constructed, and its use to classify a new case. Decision tree has been determined from classifying case using class probability distribution resulting from the class with highest probability is chosen as predicted class.

3.1.2 Association rule based method: Many association rule mining algorithms, such as the Apriori algorithm, is that only database entries which exactly match the candidate patterns may contribute to the support of the candidate pattern. This creates a problem for databases containing missing values [13].

3.1.3 Rough set based method [9]: The rough set theory was found to assist in the increased speed of convergence and in avoiding local minima problem. For example, a hot-deck data imputation method, based on rough set computations

In these algorithms, continuous attributes are always discretized before imputing. This possibly leads to a loss of useful characteristics of continuous attributes.

3.2 Continuous Attributes Imputation Methods

The well established imputation methods in [4], [14] are developed for only continuous attributes.

A simple imputation method is just to use the average value for the attribute, for continuous attributes, we use the mean or average of all the specified values for that attribute in the training data set, but there are more robust techniques, including a k-nearest neighbors (KNN) approach [9] and the use. If a training example contains one or more missing values, we measure the *distance* between the example and all other examples that contain no missing class attribute (or dependent variable).

The difficulty with this method is that the KNN does not actually map the points to the higher-dimensional space, it evaluates a *kernel function* that is chosen such that its result is proportional to the inner product in the higher dimensional space.

4. Nonparametric Regression Framework

The natural extension of this paper is to model the settings of discrete and continuous independent attributes in [3] to a fully nonparametric regression frame work.

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. In [15], iterative approaches impute missing values several times and can be usefully developed for missing data imputation. But 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 [5], and nonparametric iterative methods but based on a k-nearest neighborhood framework.

Definition 1 Expectation-Maximization (EM) algorithm: A framework based on maximum likelihood density estimation for learning. It initially guess a parameter vector (k-means partitioning).Iteratively refine parameters based on E-step and M-step. Hence it is an example for parametric imputation method in missing value data sets.

EM algorithm is applied only to discrete population with MAR assumption. This method is not flexible with all models. As it is an iterative algorithm it takes long processing time, and also continuous variables are discretized.

Definition 2 K-nearest neighborhood framework: In this case the missing data treatment is independent of learning algorithm used. The method can predict attributes. The difficulty is in the analysis of large database. For this implementation, if a training example contains one or more missing values, we measure the *distance* between the example and all other examples that contain no missing values. For discrete attributes, this distance is 0 if the values are the same, and 1 otherwise. In order to combine distances for discrete and continuous attributes, perform a similar distance measurement for continuous attributes. If the absolute difference between the two values is less than half a standard deviation, the distance is 0, otherwise, it is 1.

5. Kernel Regression Imputation Framework

5.1 Mixture Kernel Function into Missing Values

As pointed in [16] that a global kernel can present better extrapolation at lower order degrees, but need more higher order degree for receiving a good interpolation. But the degree of polynomial must be between 0 and 1. A local kernel has better interpolation, but fails to provide stronger extrapolation. It also demonstrated that a mixture of kernels can lead to much better extrapolation and interpolation than using either local or global kernels. In this paper the proposed imputation is based on mixture kernel function and is as shown in figure below.

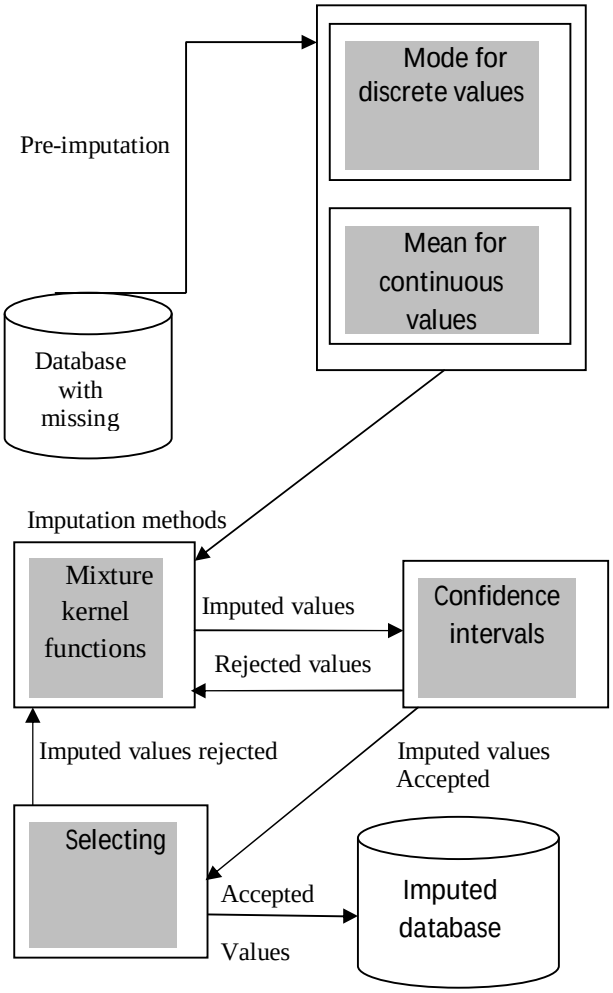

Fig.2. Proposed framework

The overall architecture of the proposed framework is visualized in Fig.2. It consists of three main functional modules:

1) Pre-imputation, 2) Mixture kernel functions 3) Setting Confidence intervals and 4) Selecting. All of those are visualized as shadowed boxes.

Let us briefly discuss the functionality of each of these modules. The missing values are first pre-imputed (module1), i.e., temporarily filled with a value to identify the missing values uniquely, and is used to perform imputation, using a fast linear mean imputation method for continuous values and mode used to perform imputation for discrete values. Next, each missing pre-imputed value is imputed using a base imputation method, called Mixture kernel function (module 2) where a polynomial kernel is used along with a pinch of local kernel to obtain better interpolation and extrapolation in real datasets. Thus imputed value is filtered by using confidence intervals (module 3) using standard deviation and mean of values. Confidence intervals are used to select the most probable imputed values while rejecting possible outlier imputations. Once all the values are imputed and filtered, each of them is assigned with a value that quantifies its quality, that is, it might be expressed as a probability or a distance. Based on these values, the selecting module (module 4) accepts the best high-quality imputed values falls near or above threshold or mean, whereas the remaining imputed values are rejected, and the process repeats with the new partially imputed database. After ten iterations, all the remaining imputed values are accepted, and the imputed database is outputted. Note that finally, all the observed information including observed information in incomplete instances with missing values uses a method, i.e., grid search method is used to obtain the optimal bandwidth for the proposed mixture kernel estimators, instead of the data driven method in [15].

5.1.1 Pre-imputation Module: The mean pre-imputation module was developed based on the premise that the base imputation method would benefit, i.e., improve its accuracy, by having a complete database to develop a model and impute the missing data. Completion of the database enhances its information contents, which, if done correctly, ultimately results in the ability to generate a better imputation model. A simple and efficient way of completing the database is to initially impute the missing values and subsequently use the pre-imputed values to perform the actual imputation.

5.1.2 Confidence Intervals: The filter is based on the premise that imputed values, which are close to the mean (for numerical attributes) or mode (for nominal attributes) of an attribute, have the highest probability of being correct. The filter is designed by computing confidence intervals.

Imputed values for a given attribute that fall within the interval are kept, whereas the values outside of the interval are discarded.

5.1.3 Selecting: Selecting is a procedure where the most appropriate value is iteratively selected from imputed data based on threshold associated with records.

The ultimate goal of the proposed work is to improve accuracy of the imputation by accepting only high-quality imputed values and using them, i.e., additional and reliable information, to impute the remaining values. In general, the module works iteratively and is appended at the end of the imputation process, when all imputed values have been already filtered out. At each iteration, high-quality imputed values are selected and accepted, whereas the remaining values are rejected. In this way, a partially imputed database is created and fed back to the base imputation algorithm. Next, the imputation is repeated, but this time, the concentration is on imputing the remaining values. All the remaining imputed values are accepted at the last iteration.

6. CONCLUSION

This paper discussed different methods to impute the missing values. Missing values are replaced by probability distributions over possible values for the missing feature, which allows the corresponding transaction to support all datasets that could possibly match the data. Transactions which do not exactly match the candidate item set may also contribute a partial amount of support this behavior is beneficial for databases with many missing values or containing numeric data. Handling missing values using the most probable information for all the samples belonging to the same class gives better result as compare to other techniques because presented technique is a hybrid approach of class technique and probability technique. The cases where dataset have both continuous and discrete independent attributes are imputed using mixture kernel based iterative nonparametric estimators are also discussed. Missing values filled with better accuracy leads to better results, this phenomenon is also observed.

ACKNOWLEDGMENT

I gratefully acknowledge Computer Science Department at Adhiyamaan College of Engineering for the constant support and guidance.

REFERENCES

[1] 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.

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

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

[4] 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.

[5] 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.

[6] S.C. Zhang et al., "Missing Is Useful: Missing Values in Cost- Sensitive Decision Trees," *IEEE Trans. Knowledge and Data Eng.*,vol. 17, no. 9, pp. 1689-1693, Dec. 2005

[7] 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. 91-99, 1994

[8] A. Dempster and D. Rubin, *Incomplete Data in Sample Surveys*:Theory and Bibliography, W.G. Madow, I. Olkin, and D. Rubin, eds., vol. 2, pp. 3-**10**, Academic Press, 1983

[9] D. Rubin, Multiple Imputation for Nonresponse in Surveys. Wiley, 1987.

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

[11] 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.

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

[13] Q.H. Wang and R. Rao, "Empirical Likelihood-Based Inference under Imputation for Missing Response Data," *Annals of Statistics*,vol. 18, pp. 896-912, 2002.

[14] V.C. Raykar and R. DuraiswamiFast, "Fast Optimal Bandwidth Selection for Kernel Density Estimation",*Proc. SIAM Int'l Conf. Data Mining (SDM '06)*, pp. 512-511, 2006.

[15] 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.

[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.