ISSN:0975-9646

Sandhya.GN et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 1483-1485

An Enhanced Approach to Handle Missing Values in Heterogeneous Dataset

Saravanan.P [#], Sandhya.G.N[*]

[#]*Assistant Professor, Department of IT, Sathyabama University, Chennai, Tamil Nadu, India*
[*]*PG Scholar, Sathyabama University, Chennai, Tamil Nadu, India*

***Abstract*— Generally, data mining (sometimes called data or knowledge discovery, knowledge extraction, knowledge discovery) is the process of analyzing huge voluminous data from different perspectives and summarizing it into the useful information. Hence data quality is much important to get the high quality pattern as result. Quality decisions ought to be based on quality data. Data quality is affected by the presence of missing values called holes because of various reasons. In order to make the database as complete by filling the holes with plausible value, variety of imputation methods have been developed. But they are limited to handle missing values in homogenous attributes only. Few of the existing systems uses the mixture kernel function for imputing missing values in mixed attribute datasets. In the proposed work, new imputation framework has been developed to handle missing values in heterogeneous datasets. Firstly pre-imputation is performed using ENI (Encapsidated Neighbour Imputation) approach followed by the application of Gaussian Kernel function to both continuous and discrete attributes. The proposed framework is tested with its competitors for various standard missing rates over bench dataset UCI repository. The behaviour of the framework proposed in this paper is studied using the parameter RMSE and concluded that it is behaving good.**

Keywords— **Data Mining, Missing Value Imputation, Kernel Function, ENI.**

I. INTRODUCTION

Imputation refers to assigning a value to a missing value or holes introduced into the databases due to various reasons like optional attribute, instrument fault etc. Anyone who does statistical data analysis or data mining of any kind runs into the problems of missing data especially in real time databases. Characteristic dataset always land up in some missing values for attributes. For example in surveys people generally tend to leave the field of income blank or sometimes people have no information available and cannot answer the question. Also in the process of collecting data from multiple sources some data may be inadvertently lost. For all of this and various other reasons, missing data is a widespread problem in both social and health sciences. This is because every standard statistical method works on the fact that every problem has information on all the variables and it needs to be analyzed. The most regular and trouble-free solution to this problem is, in any case if an attribute has missing data to be analyzed it can be simply ignored. This will give a dataset which will not contain any missing value and standard methods can be used to process it further. But this method has a major drawback, which is deleting missing values sometimes might lead to ignoring a large section of the original sample.

A. Types of Missing Data

The choice of how to deal with missing data should be based on how the missing data are generated. Little and Rubin proposed a very useful classification. The three principal types of mechanism for the missingness are shown. They are

1) MCAR: If everyone in the population has an equal probability of being missing, Then the missing data mechanism is said to be missing completely at random or MCAR .in this case list wise deletion, the dropping of any subject with an incomplete set of measures, does not lead to bias and can be understood as being equivalent to simply having a lower sampling fraction and thus smaller sample size.

2) *MAR:* Just occasionally the loss of sample size may be too great to use list- wise deletion. Where the probability of missing is not constant but differences in probability depend solely on data that is not missing, then the missing data is said to be missing at random or MAR. Particularly simple case of MAR is covariate dependent missing data, where the probability of missingness depends only on the value of an explanatory covariate or independent variable.

3) *Non-ignorable MAR*: missing data are substantially more difficult to deal with. In the cases of both MAR and non-ignorable missing data, list-wise deletion will lead to bias.

B. Imputation Strategies

Many options for imputation exist. Some of the primary methods are

1. Mean based imputation
2. Medium based imputation
3. Stratified imputation
4. Regressed imputation-difficult with MCAR

1) *Mean based imputation*- This process is the most simple and straight forward. This involves replacing the missing values with the mean of the variable which is more suitable for discrete and continuous attributes.

2) *Median based imputation*- This process is also very simple. This involves replacing the missing values with the median value of the variable which is more suitable for categorical attributes.

3) *Stratified imputation*- This process is slightly more involved. This involves replacing the missing values with the mean or median of the variable but with consideration for similar of observations.

4) *Regressed imputation*- This process involves actually predicting the value of the missing values using regression. It works well if the variables are related to each other i.e., if there exists a dependency between the variables and if only have one or two variables with missing data.

The rest of the paper is organized as follows in section 2; the literature survey has been done which describes the related work done in this domain. In section 3, the proposed ENI based Gaussian kernel function for imputing missing values is explained with an algorithm, experimental description as been done and performance comparison is done followed by conclusion in section 4.

II. RELATED WORKS

Many methods were proposed earlier to impute the missing values which, makes the database as complete. Few of the notable work are described in this section. Shichao Zhang, Zhenxing Qin, Charles X. Ling, And Shengli Sheng [1] introduces a method which develop a decision tree contain minimal total cost of tests and misclassification of training

data. The advantage is that the total cost of test and misclassification is reduced by the presence of missing values. The disadvantage is that it is not probable to get complete data set because imputation is not executed. Deleting the attributes leads to losing the useful information in dataset. Yongsong Qin, Shichao Zhang, Xiaofeng Zhu, Jilian Zhang, Chengqi Zhang [2] suggest a fresh and well organized imputation method for semi parametric data, Which has two models, parametric and non parametric model. Kernel based stochastic semi parametric method is created, which switch with complex relation. It is enhanced than deterministic semi parametric regression imputation in competence. Imputation algorithm cannot be applied to many factual data set, for instance equipment maintenance database, industrial data set and medical database since these data set are frequently with continuous discrete and categorical independent attribute.

Parametric and nonparametric regression imputation methods are commonly used methods to impute missing values. The parametric method, such as linear regression R.Little and D.Rubin [3], is superior while the dataset are adequately modelled. The demerits is that in real applications, the parametric estimators can lead to highly bias and the optimal control factor settings may be miscalculated as it is impossible to know the distribution of the dataset. There are a number of usual imputation methods which were discussed in [7], [8], [9], designed for discrete type of attributes using a "frequency estimator" which means a data set is separated into several subsets or "cells."On the other hand, when the number of cells is outsized, observations in each cell possibly will not be enough to non-parametrically estimate the relationship along with the continuous attributes in the cell.

For this case, nonparametric imputation method Q.H. Wang and R. Rao [4], can provide superior fits by capturing the structure of the dataset. Drawbacks are that these imputation methods are designed for either continuous or discrete independent attributes. For example the well established imputation methods in [4] are developed for only continuous attributes. And these estimators cannot handle discrete attributes well.

The method namely association-rule-based method W.Zhang [5] and rough-set-based method C.Peng and J.Zhu [6], are designed to deal with only discrete attributes. In these algorithms, continuous attributes are always discretized before imputation. This possibility leads to a loss of useful characteristics of the continuous attributes.

A. Existing System

Existing work brings out imputation framework to handle missing values in heterogeneous datasets. Non-Parametric iterative imputation is done. It first considers the database with missing values, then it identifies the type of attributes either continuous or discrete attribute. Mean pre-imputation is applied if it continuous otherwise Mode pre-imputation is applied. This is the basic step of imputation. Then by using pre-imputed datasets kernel function is 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. Here spherical kernel function is used. Estimated value is calculated. Final step is 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 dataset to make it as a complete dataset.

III. PROPOSED SYSTEM

The proposed work brings out the new imputation framework. i.e., the non-parametric iterative algorithm is extended from mixture kernel to high order kernel such as Gaussian kernel based ENI(Encapsidated- neighbor imputation) for missing value imputation is designed .In the existing system basic imputation i.e. pre-imputation is done using MMI(mean mode imputation)method. In the proposed ENI is used to perform pre-imputation. Then Gaussian kernel function is applied to both continuous and discrete attribute. Iterative imputation is carried out based on the imputed results until the filled-in values converge or satisfy the demands of the users. All imputed values are stored into DB called imputed DB.

ENI (Encapsidated-neighbor imputation), is alike the kNNI method. The ENI approach considers the left and right nearest neighbors of a missing data, while the kNNI method selects k nearest neighbors. The kNNI method uses a fixed k, but in the ENI approach the number of selected nearest neighbors is variable determined by data during imputation process.

A. Algorithm for Encapsidated Neighbor Imputation

In ENI method, the imputation process is carried as follows,

1. For each incomplete data search all the left or right nearest neighbor.
2. Using the eqn (2) weight w_i is calculated.

   The weight of a left or right nearest neighbor can be obtained as follows,

   For a left or right nearest neighbor $T_i = (X_{i1}, X_{i2}, \ldots . X_{in}, Y_i, l, 1)$ of a missing data $T=(X_{l1}, X_{l2}, \ldots . X_{ln}, Y_l, 0)$, we get

   $$d_i=\sqrt{(X_i - X_{l1})^2 + \cdots + (X_{in} - X_{ln})^2} \quad (1)$$

   From this we can get w_i ,

   $$w_i = 1 - \frac{d_i}{d_1 + d_2 + \cdots + d_m} \quad (2)$$

   m-is the number of selected left or right nearest neighbor of the missing data.
3. Using the eqn (3) Y_l is estimated.

   $$Y_l = \sum_{i=1}^{n} (w_{i-} Y_{i-} + w_{i+} Y_{i+}) \quad (3)$$
4. Repeat step 1-3 until no incomplete data in the dataset.

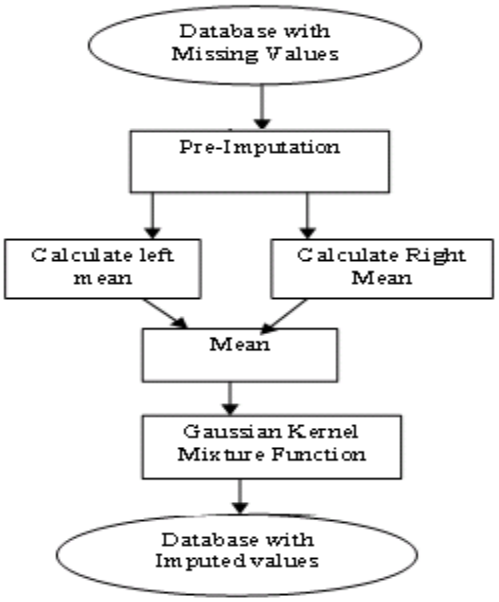

Fig.1.Flow Diagram

B. *Advantages:*

ENI imputation algorithm is model free well-organized compared with other methods. Thus, the ENI can be easily adapted to work with any attribute as class, by just modifying which attributes will be considered in the distant metric. Also, this approach can easily treat examples with multiple missing values.

C. *Algorithm Design*

From the input heterogeneous dataset the records with the missing values will be identified and categorized based on attribute type of missing values.

For each missing values (MVi) in dataset Y

Calculate Left mean $y_{i-} = \sum_{i=1}^{n} Y_{i-}$

Calculate Right mean $y_{i+} = \sum_{i=1}^{n} Y_{i+}$

Calculate Mean $\frac{y_{i-} + y_{i+}}{2}$

Update the values in dataset

End For

For each missing value (MVi) in dataset Y

Repeat

Apply the Gaussian kernel function for both discrete and continuous attributes

MVi' is got based on (5)//if continuous

MVi' is got based on (6)//if discrete

Until MVi' is between ML and MH (low & high limit)

$$K_{w,Y,X} = \left[\frac{||Y_i - Y_{i-}||^2}{2 * w_{i-}} + \frac{||Y_i - Y_{i+}||^2}{2 * w_{i+}}\right] \quad (4)$$

$$mt(x) = \frac{n^{-1} \sum_{i=1}^{n} (w_{i-} Y_{i-} + w_{i+} Y_{i+}) \; K_{w,Y,X}}{n^{-1} K_{w,Y,X} + n^{-2}} \quad (5)$$

$$mt(x) = \frac{\sum_{i=1}^{n} \sum_{y \epsilon Dy \neq Yi} l(Y_i^y, yt, w) \; yt \; K_{w,Y,X}}{\sum_{i=1}^{n} K_{W,Y,X}} \quad (6)$$

IV. Experimental Description

Two UCI datasets, Abalone dataset for discrete and Autompg dataset for continuous are taken, to test the validity of the proposed work. Abalone contains 300 instances and 9 attributes. Autompg includes 397 instances and 9 attributes. To basically examine the effectiveness and validity and ensure the systematic nature of the research, we artificially generated a lack of data at standard missing ratios 5%, 10%, 15% under three different modalities, in the complete datasets.

Performance Analysis

Performance of the proposed system is evaluated using the RMSE method. The existing system is compared with the proposed system and the graph is plotted as follows.

$$\text{RMSE} = \sqrt{\frac{(X_i - Y_i)^2}{m}}$$

X_i- Actual value, Y_i-Estimated value, m- no. of predictions

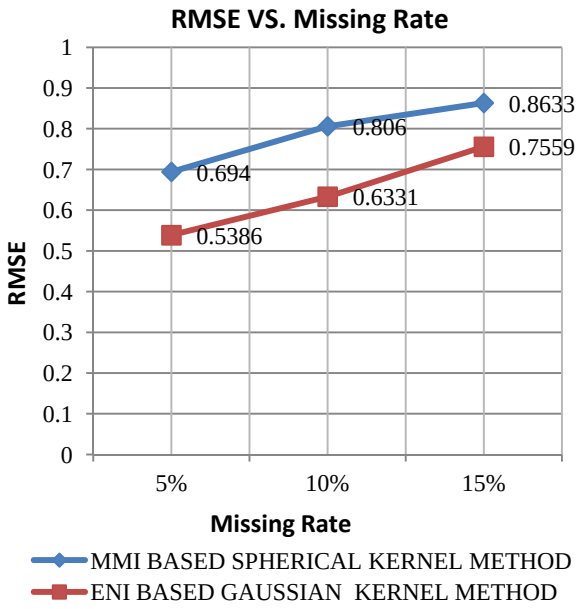

Fig.2 Performance Graph

V. Conclusion

Missing values are usually regarded as a tough problem and should be imputed before learning is applied. In this paper, ENI based Gaussian kernel function is used for imputing missing values in heterogeneous datasets. The experimental results have demonstrated that the proposed algorithm outperforms the existing one in imputing both discrete and continuous missing values in different missing ratios. In future, we plan to implement in many real datasets.

References

[1] Santosh et al., "Imputation Method for Missing Value Estimation of Mixed-Attribute Data Sets" International Journal of Advanced Research in Computer Science and Software Engineering 3(5), May - 2013, pp. 729-734

[2] Shichao Zhang, Zhenxing Qin, Charles X. Ling, and Shengli Sheng "Missing Is Useful": Missing Values in Cost-Sensitive Decision Trees, IEEE Transactions On Knowledge And Data Engineering, Vol. 17, No. 12, December 2005

[3] Yongsong Qin, Shichao Zhang, Xiaofeng Zhu, Jilian Zhang, Chengqi Zhang "Semi-parametric optimization for missing data imputation" Published online: 18 January 2007, Appl Intell (2007) 27:79–88

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

[5] 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.

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

[7] 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.

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

[9] 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.

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