Missing Data Imputation by Evidence Chain

Anaswara R[1], Sruthy S[2]

[1]M. Tech. Student, Computer Science and Engineering, Sree Buddha College of Engineering, Kerala, India.
[2]Assistant Professor, Computer Science and Engineering, Sree Buddha College of Engineering, Kerala, India.

---***---

Abstract – *Missing data in a dataset decrease the accuracy of data analysis. Missing value occurs when no data value is stored for a variable in an observation. A data mining problem that adversely affects data analysis and decision making processes. Missing data handling is an important step in data pre-processing. Imputation technique is used to fill missing data. Imputation is a technique of replacing missing data with substituted values. This will increase the accuracy and efficiency of the dataset. Here use a new method an evidence chain for missing data imputation.*

***Key Words*: Imputation, MapReduce, Attribute combination, Possible value, Marked dataset, Data pre-processing.**

1. INTRODUCTION

Missing data handling is an important step in data pre-processing. Data pre-processing is an indispensable step before data mining. Data preprocessing includes the missing data imputation, entity identification, and outlier detection. Missing data padding is an important problem to be solved in data preprocessing. Because of lack of information, the omission of information and man-made operation, some data are missing in the dataset. These incomplete data will affect the quality of data mining and even lead to the establishment of the wrong data mining model, making the data mining results deviate from the actual data. The preprocessing phase helps users to understand the data and to make appropriate decisions to create the mining models and to make appropriate decisions. This phase will identify the data inconsistencies like missing data and fix the missing or wrong data using appropriate techniques. This will increase the accuracy and integrity of data mining.

Missing data is a common problem that and is the current focus of this research. This is a rapidly growing area. There have been many techniques for finding missing data and solving missing data problems [1]. If a dataset that contain a small number of missing data, then the simplest method to handle missing data is “Deletion technique”. In deletion techniques eliminate the attributes or cases. This is a default method for handling missing data. This method is not good if the data set contains a large number of missing data. In this case use imputation techniques for handling missing data. Imputation technique is a process of replacing missing data with substituted values. If an important data is missing for a particular instance, then it can be estimated from the data that are present by using these imputations.
There are many studies on imputation methods, including the use of neural network methods, using regression, and Bayesian network. A new imputation technique is used here. This method identifies the missing value and imputes the value to the missing place.

The new method used here use evidence chain for missing data imputation and this will combine with the MapReduce programming model to handle large amount of data [7].

Here the new imputation method is implemented by creating an application. That generates numerical data automatically and in between that generates missing data fields. This is used as the dataset for missing data imputation.

2. RELATED WORKS

There is various imputation techniques are used in data mining. The selection of imputation technique may be depending on datasets or may be related to the mechanisms. Some of the methods used in different areas are given below: Xiaofeng Zhu et al. [2] suggest a method to handle missing value estimation in mixed-attribute datasets. Various techniques have been used to dealing missing value in a dataset with homogeneous attributes. A mixture-kernel based iterative estimator is advocated for missing value imputation. Mixture-kernel-based iterative estimator utilizes all the available observed information, including observed information in incomplete instances. Author propose a new algorithm which has better classification accuracy and the convergence speed of the algorithm than extant methods, such as the nonparametric imputation method with a single kernel, frequency estimator (FE), and the nonparametric method for continuous attributes. The imputed values are used to impute subsequent missing values in new algorithm.

Bankat M. Patil et al. [3] provide a novel approach for estimation of missing data method using cluster based k-mean weighted distance algorithm (CMIWD). Here proposed an efficient missing value imputation method which on clustering with weighted distance. Based on the user specified value K, divide the data set into clusters. Then find a complete valued neighbour that is nearest to the missing valued instance. Then compute missing value by taking the average of the centroid value, and the centroidal distance of the neighbour and this is used as impute value.

Ingunn Myrtveit et al. [4] provide a Full Information Maximum Likelihood (FIML) method. The FIML is a model-based method. FIML method can analyze incomplete data sets directly. This is an incomplete case analysis method.

Thirukumaran. S et al. [5] suggest a method to improving accuracy rate of imputation of missing data using classifier methods. Mean method by step Digression (MMSD) impute missing data containing varying amount of missing values. MMSD is alternate to mean method. The MMSD overcomes the limitation of mean method.

Dan Li et al. [6] provide a Fuzzy K-means Clustering method for missing data imputation. Author present a missing data imputation method based on one of the most popular techniques in Knowledge Discovery in Databases (KDD), i.e., clustering technique.

Yongsong Qin et al. [8] suggest a semi-parametric optimization for missing data imputation. This method overcomes some limitations in linear models and non-parametric models by making an optimal inference.

The Prereplacing method replaces the missing values before the data mining process [9]. It works in the pre-process phase. The Embedded method is able to handle missing values during the data mining process. The missing value is imputed in the same time while creating the model. There are many imputations with different features to impute the missing data for these methods. The pre-replacing method contains Mean-and Mode, Linear Regression, K-Nearest Neighbor (KNN), Expectation Maximization (EM), Hot-Deck (HD), and Autoassociative Neural Network techniques. The embedded method contains the Casewise deletion, Lazy decision tree, Dynamic path generation, C5.0, and Surrogate split techniques.

3. METHODOLOGY

A new method is used for missing data imputation. The new imputation method is implemented by creating an application. That generates numerical data automatically and in between that generates missing data fields. This is used as a dataset for missing data imputation. This first identifying the presence of missing data and replaces it with possible value. Here missing value imputation is performed with the help of a combination of combination attribute and possible value. In this method first estimate the missing value of the imputation value, the algorithm first scans the each data tuple in the entire dataset. Then marking tuple with missing value as "-1" as incomplete data tuples, and combine different associated attribute values of missing data in the incomplete data tuple as the evidence of the estimated missing value. So that many of combination of the relevant attributes for missing data in the incomplete tuple constitute the estimated missing data value of the chain of evidence. The repeated evidence with maximum possible value is taken as the value for missing value [7].

The imputation process is done through the following steps:
Step 1: First scan each data tuple in the entire dataset to identify the position of missing value. Then mark the missing value as "-1". This dataset is called marked dataset. This dataset is used for further operations.

Step 2: Scan each tuple in the dataset and combine the entire associated attribute with the missing value. This will serve as an evidence for estimating the missing data values. The result will be used as the chain of evidence to estimate the missing data. In this stage use a map function calculate the associated attribute value combinations of the missing data in the incomplete data tuple combi_attri.

The Reduce function gets a set of associated attribute value combinations.

Step 3: Find all the possible values in the dataset. The possible values are taken from the corresponding columns of missing data. Then calculate the probability value of possible values.

Step 4: Check the missing data's attribute combination with the combi_attri table if corresponding combination is present in the database then take its corresponding possible value. If there more than one possible value for same combi_attri then takes the corresponding possible value with maximum probability. If there is no repeated attribute combination, then use the possible value with the highest possible value for missing value imputation.

Stage 5: The value of the missing value estimate of step 4 is filled into the original missing dataset.
Modules present in this project are:

1) Data capturing: This module includes a control panel that can activate/deactivate the sensors periodically. The sequence of sensor operations causes missing data. The captured data undergo a MapReduce algorithm for handling efficient storage and quick retrieval.
2) Dataset management: The captured data is used as the dataset for missing data imputation.
3) Missing data imputation: In this module apply above method to impute missing data.

4. CONCLUSION

The processing of missing data has become a very important step in the process of data preprocessing. The limitations of data acquisition or improper operation of the data lead to data errors, incomplete results, and inconsistencies. This will reduce accuracy of data mining. To overcome these limitations missing value imputation methods are used. Evidence chain is used to predict the value of missing data.

Estimation of missing value by using the set of combinations of missing attributes values. The Map-Reduce framework is used to process large data.

REFERENCES

[1] T. Aljuaid and S. Sasi, "Proper imputation techniques for missing values in data sets,"*2016 International Conference on Data Science and Engineering (ICDSE)*, Cochin, 2016, pp. 1-5. doi: 10.1109/ICDSE.2016.7823957.

[2] X. Zhu, S. Zhang, Z. Jin, Z. Zhang, and Z. Xu, "Missing value estimation for mixed- attribute datasets," Knowledge and Data Engineering, IEEE Transactions on, vol. 23, pp. 110-121, 2011.

[3] B. M. Patil, R. C. Joshi, and D. Toshniwal, "Missing Value Imputation Based on K-Mean Clustering with Weighted Distance", Communications in Computer and Information Science, 1, Contemporary Computing, Part 11. 94, pp. 600-609.

[4] Ingunn Myrtveit, Erik Stensrud, and Ulf H. Olsson, "Analyzing Data Sets with Missing Data: An Empirical Evaluation of Imputation Methods and Likelihood-Based Methods", IEEE Transactions On Software Engineering, Vol. 27, No. 11, November 2001.

[5] S. Thirukumaran and A. Sumathi, "Improving accuracy rate of imputation of missing data using classifier methods", 10th International Conference on Intelligent Systems and Control (ISCO), Coimbatore, India, pp. 1-7, 2016.

[6] Li, D., Deogun, J., Spaulding, W., Shuart, B.: Towards "Missing Data Imputation: A Study of Fuzzy K-means Clustering Method". In: Tsumoto, S., Słowiński, R., Komorowski, J., Grzymała-Busse, J.W. (eds.) RSCTC 2004. LNCS (LNAI), vol. 3066, pp. 573–579.

[7] X. Xu, W. Chong, S. Li, A. Arabo and J. Xiao, "MIAEC: Missing Data Imputation Based on the Evidence Chain," in *IEEE Access*, vol. 6, pp. 12983-12992, 2018. doi: 10.1109/ACCESS.2018.2803755.

[8] Y. Qin, S. Zhang, X. Zhu, J. Zhang, and C. Zhang, "Semi-parametric optimization for missing data imputation," Applied Intelligence, vol. 27, pp. 79-88, 2007.

[9] T. Aljuaid and S. Sasi, "Proper imputation techniques for missing values in data sets," *2016 International Conference on Data Science and Engineering (ICDSE)*, Cochin, 2016, pp. 1-5. doi: 10.1109/ICDSE.2016.7823957.

BIOGRAPHIES

Anaswara R, she is currently pursing Master's Degree in Computer Science and Engineering from Sree Buddha college of Engineering, Elavumthitta, India. Her research area of interest includes the field Data mining.

Sruthy. S, she is an Assistant Professor in the Department of Computer Science and Engineering, Sree Buddha College of Engineering. Her main area of interest is Computer Vision and Image Processing and Data mining.