International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 11 | Nov 2018 www.irjet.net p-ISSN: 2395-0072

Evidence Chain for Missing Data Imputation: Survey

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 is one of the major problems in datasets which reduce the integrity and deviate data mining. Imputation technique is used to fill the missing data, which give the complete knowledge of the dataset. The missing data imputation technique is applied in the phase of data pre-processing. This will help to reduce the data missing due to human caused errors. Existing imputation methods have less accuracy and low stability. To improve accuracy a new method Missing value Imputation Algorithm based on an Evidence Chain (MIAEC) is used. MIAEC provide accuracy for increasing rates of missing data. Map-reduce programming model is applied with MIAEC for large-scale data processing.*

Key Words: **Evidence chain, Map-Reduced programming model, Missing value imputation, Expectation Maximization, KNN.**

1. INTRODUCTION

The missing data means for some unit one or more data are not observed. Incorrect manual data entry and measurements, equipment error are some reasons of missing data. One of the methods used to deal with missing data is to delete the record that contains missing value. But this method is not good if the dataset contains large amount of missing values. This method misses important information and also reduces the size of datasets. Data pre-processing is an important step before data mining. It is difficult to perform data mining in incomplete data sets. Traditional imputation methods are mean and median also known as universal methods. Hot-deck, Cold-deck methods are the extended versions of mean method. Regression and likelihood based methods are also known as multivariate methods. These methods are based on the model based method. In regression using real data variable predict the missing value. Maximum likelihood is used to predict the missing value in likelihood method. The incomplete data reduce the quality of data and generate the wrong data mining model. Three major problems caused by missing data: 1) Datasets with missing data cannot be processed with most of the data processing algorithms; commonly used algorithms are not suitable. 2) Missing records have often been overlooked this will lead to poor statistical results.3) Mining datasets with missing fields.

According to the degree of randomness of deletion there are three types of data loss: 1) Missing Completely At Random (MCAR), 2) Missing At Random (MAR), and 3) Not Missing At Random (NMAR). The choice of imputation algorithm and the final effect of imputation are based on missing type of a data set. For MCAR commonly used imputation method is a mean-imputation method. The prior knowledge of the dataset itself is used for the imputation of NMAR data.

In proposed system a missing data imputation approach based on a chain of evidence (MIAEC) is extended using Map-Reduce programming model. By mining MIAEC obtains all relevant evidence of missing data in each tuple of missing data. This is combined to form a chain of evidence to estimate the missing attribute value. The chain of evidence is used to estimate the value of missing data. The existing algorithms are designed to process small data sets, but the new method help to process large data sets. MIAEC algorithm is extended using Map-Reduce programming model for large-scale datasets processing.

2. LITERATURE SURVEY

Various methods are used for missing data imputation. Some of them are discussed below.

Missing Value Estimation for Mixed-Attribute Data Sets

Xiaofeng Zhu et al. [1] 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. Homogeneous attributes means their independent attributes are either continuous or discrete. For discrete and continuous missing target values two consistent estimators are used. These estimators are constructed for discrete and continuous missing target values, for mixed-attribute data sets. Then, 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.

2.1 Semi-parametric optimization for missing data imputation

Yongsong Qin et al. [2] suggest a semi-parametric optimization for missing data imputation. To handle complex relation here proposes a stochastic semi-parametric method. This method overcomes some limitations in linear models and non-parametric models by making an optimal

inference. Consider an example Consider the sale of ice cream in summer. There is a linear relation between the sale of ice cream and weather. It is difficult to find the relation between the sale of ice cream and other factors. The relation between the sale of ice cream and all the factors is difficult to find. To handle this complex relation a stochastic semi-parametric method is used. A semi-parametric model has two parts: In this example, the first part is the relation between the sale of ice cream and weather. And the second part is the relation between the sale of ice cream and the other factors. By making an optimal inference on RMSE (Root Mean Square Error), distribution function semi-parametric regression imputation aims to overcome some shortcomings in non-parametric models and linear models.

2.2 Learning Bayesian Networks from Incomplete Databases

Marco Ramoni et al. [3] suggest a method Learning Bayesian Networks from Incomplete Databases. A Bayesian network is used to represent how information about one event can change the probability of another. Bayesian Belief network (BBN) is a special type of diagram ad which associated with a set of probability table. Here variables are represented as nodes, can be continuous or discrete. Arc represents the relationship between variables. To learn the graphical structure of a BBN from a possibly incomplete database proposed a deterministic method. This is an iterative method. Deterministic method estimates the conditional probabilities. Which defining the dependencies in a BBN, but does not rely on the Missing Information Principle and this method is called Bound and Collapse (BC). Started with bound and then collapse is performed. Bounding is the set of possible estimates consistent with the available observations in the database. Depending on the assumed pattern of missing data a convex combination of the extreme estimates with weights is used for collapses the resulting interval to a point. May be an external source of information or may be estimated from the available information provide the pattern of missing data under the assumption that data are missing at random. BC can be used to both assess the conditional probabilities of a BBN and induce the graphical structure.

2.3 Learning Bayesian networks from incomplete databases using a novel evolutionary algorithm

Man Leung Wong et al. [4] proposed learning Bayesian networks from incomplete databases using a novel evolutionary algorithm. In the presence of missing value here propose a novel method for learning Bayesian Network. This is a combination of a novel evolutionary algorithm the traditional Expectation Maximization (EM) algorithm. For learning and evaluating the candidate networks a data completing procedure is presented. New method overcomes the problem of getting stuck in sub-optimal solutions. Hybrid Evolutionary Algorithm (HEA) does not deal with incomplete databases. The novel evolutionary algorithm also known as EBN (Evolutionary Bayesian Network learning method). This utilizes the global search ability of the HEA and applies EM to handle missing values.

2.4 Iterative KNN Imputation Based on GRA (Grey Relational Analysis) for Missing Values

Ming Zhu et al. [5] suggest an iterative KNN Imputation based on GRA for Missing Values. To analyzing the relationship between discrete data sets GRA (Grey Relational Analysis) is used. Iterative KNN methods perform poorly in the estimation of missing values in the trash pickup logistics management system (TPLMS). The missing values in the TPLMS data set are significant and result in unserviceable decision-making. Here introduces an iterative KNN imputation method which contains weighted k nearest neighbor (KNN) imputation and the grey relational analysis (GRA). This is an instance based imputation method. Instead of Euclidean distance or other similarity measures using a grey relational grade, this method takes advantage of the correlation of attributes to search k-nearest neighbor instances. From these nearest neighbour instances the plausible values for the missing values are estimated iteratively. The iterative imputation allows all available values, including the imputed values from previous iteration and the attribute values in the instances with missing values to be utilized for estimating the missing values. Specifically, regardless of the missing rate the imputation method can fill in all the missing values with reliable data.

2.5 Cluster-based KNN (CKNN) Missing Value Imputation for DNA Microarray Data

Phimmarin Keerin et al. [6] proposed a Cluster-based KNN (CKNN) Missing Value Imputation for DNA Microarray Data. Author proposes a new to impute missing values in microarray data. The proposed algorithm is CKNN impute. This algorithm is an extension of *k* nearest neighbor imputation with local data clustering being incorporated for improved quality and efficiency. CKNN kicks off by finding a complete dataset this is done by the removal of rows with missing value(s). Then by applying a clustering technique on the complete dataset *k* clusters and their corresponding centroids are obtained. A set of similar genes of the target gene (with missing values) are those belonging to the cluster. Their centroid is the closer the target. The target gene is imputed by applying *k* nearest neighbor method with similar genes previously determined. The proposed technique performs better than the classical *k* nearest neighbor method. This CKNN contain three major steps, fist, finding a number of clusters, in a second step generating a clustering model that will be used as a reference, and apply the KNN impute in conjunction with the discovered clusters.

2.6 Missing Value Imputation Method Based on Clustering and Nearest Neighbors

Satish Gajawada and Durga Toshniwal et al. [7] provide a missing value imputation method based on Clustering and Nearest Neighbors. The available complete objects are less

when objects with missing values are high. Good results are not getting when imputing missing values with limited amount of complete objects. Complete objects only used for imputation. The imputed objects are treated as complete objects and the number of completed objects is increased. For further imputation the imputed objects are used along with the available complete objects. In this system author present a missing value imputation method based on K-Means and nearest neighbours which uses the imputed objects for further imputations. K-means clustering based imputation is performed in 2 steps. First apply K-means clustering to get clusters. In second step cluster information is used to impute the missing values.

A. Description of Proposed Method

a) K-means Clustering

The dataset is divided into two sets. One set contain complete instance that do not contain any missing values. The second set contains incomplete instances with missing values. Applying K-means clustering on complete instances set to obtain clusters of complete instances.

B. Proposed Method

a) Imputing Incomplete Instance

In the order of their missing values incomplete instances in the incomplete set are arranged. That means the instance with less number of missing values comes first in the list and the instance with large number of missing values comes later in the list. The incomplete instance coming first in the list, which is taken and applying clustering on it. By this found the instance nearer. K nearest neighbours of the incomplete instance in the cluster is found. Here K is cluster size, if K is greater than cluster size. Using weighted KNN the missing values in the incomplete instance are imputed.

b) Moving to Complete Instance Set

The imputed incomplete instance move to the complete instance set. Then assigned to the cluster used for imputing that instance. After assigning imputed instance to the cluster cluster centre is updated to the centroid of all points in the cluster. Like any other complete instances in the complete instance set the imputed instance will be used for all further imputations of other incomplete instances.

c) Imputing All Incomplete Instances

Above described steps are repeated until all incomplete instances in the incomplete set are imputed.

2.7 Clustering Method Imputation with Weighted Distance

Bankat M. Patil et al. [8] 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 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. The proposed method use K-means technique with weighted distance. This method is better than K-means.

2.8 FIML (Full Information Maximum Likelihood)

Ingunn Myrtveit et al. [9] provide a Full Information Maximum Likelihood method. The FIML is a model-based method. FIML method can analyse incomplete data sets directly. This is an incomplete case analysis method. FILM is based on maximizing the log-likelihood principle. The Maximum Likelihood or ML-estimator is efficient and is implemented in most statistical software for handling multivariate analysis with complete data sets. ML estimation of incomplete data sets has computational effort. FIML assumes that the data comes from the likelihood of the theoretical model and a multivariate given the observed data. FIML estimates are efficient and consistent even if the MAR condition is not strictly met.

2.9 Mean method by step Digression (MMSD) method

Thirukumaran. S et al. [10] suggest a method to improving accuracy rate of imputation of missing data using classifier methods. MMSD impute missing data containing varying amount of missing values. MMSD is alternate to mean method. The MMSD overcomes the limitation of mean method. Limitation of mean method overcome in a way that the frequency distribution applied to the dataset is separated into blocks and subjects. The first block of set1 for imputation, using MMSD technique. The formula for MMSD imputation technique is

$$T = \left(\frac{\mu - \rho}{h}\right)$$

The vector calculates $\mu = \frac{Xi + Xn}{n}$

$$\overline{X} = \left[\frac{\sum_1^n T * V}{\sum_1^n V} * h\right]$$

$\overline{X}$ is mean value of MMSD imputation technique.

Y is total number of missing entry and **h** is the interval between numbers of missing value.

2.10 Handling missing data in trees: surrogate splits or statistical imputation

Ad Feelders et al. [11] discuss about handling missing data in trees: Surrogate splits or Statistical imputation. Surrogate split is mimic or substitute for the primary split of a node. One advantage of statistical imputation is the imputation phase is separated from analysis phase. Allow

different data mining algorithms to be applied to the completed data sets. The completed data sets analysed with any appropriate data mining algorithm. The imputation model does not have to be the "true" model but good enough for generating the imputations. An advantage is that imputation is able to handle missing data in the covariates as well as in the dependent variable. Multiple imputation shows a consistently superior performance. The variance reduction is achieved by averaging the resulting trees. For high variance models such as trees and neural networks multiple imputation get a substantial performance improvement.

2.11 Parametric and Non-parametric Methods

Faraj Bashir et al. [12] introduced parametric and non-parametric methods to enhance prediction performance in the presence of missing data. Most missing data analysis techniques use model parameter estimation which depends on modern statistical data analysis methods such as multiple imputation and maximum likelihood. Most of the modern approaches based on linear parametric regression. They do not provide good results.

Kernel methods can be used for selecting the proper fit of data. These methods cannot be used directly in the case of missing data, because they depend on the observed data or original values. Kernel approach needs the original value to determine the weight function. A linear missing data technique imputes missing values and takes them as original values. A combination of these algorithms is able to solve the problems of model selection and nonparametric estimation with incomplete information.

2.12 Incorporating an EM-Approach for Handling Missing Attribute-Values in Decision Tree Induction

Amitava Karmaker et al. [13] present an EM-approach for handling missing attribute-values in decision tree induction. Data with missing attribute-values are common for many classification problems. Here author provide an Expectation-Maximization (EM) inspired approach for filling up missing values to decision tree learning and the objective is to improving classification accuracy. Each missing attribute value is filled using a predictor constructed from the known values and predicted missing attribute-values from the previous iteration.

EM consist of two steps.

Estimation (E) step: Estimate the parameters in the model for the data source by using the known attribute-values and estimates of the missing attribute values obtained in the previous iteration of the M-step.
Maximization (M) step: Find the attribute which has maximum likelihood.

The proposed method can handle both continuous and nominal attributes. If not use a parametric model it is not able to prove that the algorithm converges for continuous value attributes. The non-convergence is expected since the attributes are nominal and datasets contain noise. This occurs infrequently and has no impact on the overall prediction accuracy.

2.13 Missing data imputation for fuzzy rule-based classification systems

Julia´n Luengo et al. [14] proposed Missing data imputation for Fuzzy Rule-Based Classification Systems (FRBCSs). The proposed Fuzzy rule-based classification systems (FRBCSs) are known due to their ability to treat with low quality data and obtain good results in these scenarios. Analyzing the influence of imputation methods with respect to the two measures. These two measures are the average mutual information difference and the Wilson's noise ratio. Wilson's noise ratio quantifies the noise induced by the imputation method in the instances which contain MVs. With respect to the class label the average mutual information difference examines the increment or decrement in the relationship of the isolated input attributes. The CMC (Concept Most Common) and EC (Event covering) methods are providing less noise and maintain the mutual information better. This corresponds to the best imputation methods observed for each FRBCS types.

Wilson's noise ratio: Wilson's noise ration used to observe the noise in the dataset. For each instance of interest, the method looks for the K nearest neighbours using the Euclidean distance, and uses the class labels of such neighbours in order to classify the considered instance. The variable noise is increased by one unit if the instance is not correctly classified.

Mutual information: Mutual information (MI) is a good indicator of significance between two random variables.

The EC method contains the following processes:

(1) Detect and synthesize from data inherent patterns which indicate statistical interdependency.

(2) Based on this detected interdependency group the given data into inherent clusters

(3) For each cluster identified interpret the underlying patterns.

2.14 iDMI: A Novel Technique for Missing Value Imputation

Md. Geaur Rahman et al. [15] proposed iDMI, a novel technique for missing value imputation using a decision tree and Expectation-Maximization algorithm. The propose a novel technique called iDMI that combines a Decision Tree (DT) algorithm (such as C4.5) and an Expectation Maximization (EMI) algorithm for imputing missing values of a data set. First divide a dataset into horizontal segments. This is done by applying a DT algorithm such as C4.5. In

order to impute the missing values belongs to the segment then apply an EMI algorithm on each segment.

Then impute them by the mean values of the attributes of the records belong to a segment where the record falls in, if all numerical attribute values of a record are missing. Thereby reduce the computational time complexity of iDMI compare to an existing technique DMI. DMI calculate the mean value of an attribute by using all records of a dataset.

2.15 Four missing data treatment methods for supervised learning an analysis

Maria Carolina et al. [16] proposed an analysis of four missing data treatment methods for supervised learning. One major problem in data quality is missing data. Rather than naive way many machine learning algorithms are used to handle missing data in a rather naive way. Missing data treatment are carefully treated, otherwise bias might be introduced into the knowledge induced. In this author analyze the use of the k-nearest neighbour as an imputation method. Imputation is a procedure which replaces the missing values in a dataset using some plausible values. An advantage of this approach is that the missing data treatment is independent of the learning algorithm used.

This help to choose the most suitable imputation method for each situation. Missing data imputation based on the k-nearest neighbour algorithm can perform better than the internal methods used by C4.5 and CN2 to treat missing data and also perform better than the mean or mode imputation method. The mean or mode imputation methods are broadly used to treat missing values. Mean method replacing every missing value of an attribute by the mean, if the attribute is quantitative. In mode is used to replace missing data if the attribute is qualitative of its known values.

Here author proposes the use of the k-nearest neighbour algorithm to estimate and substitute missing data. The advantages are:

(i) K-nearest neighbour predict both quantitative attributes (the mean among the k-nearest neighbours) and qualitative attributes (the most frequent value among the k-nearest neighbours).
(ii) For each attribute with missing data there is no necessity for creating a predictive model.

The main drawback of the k-nearest neighbor is, the algorithm searches through all the datasets whenever the k-nearest neighbour looks for the most similar instances. This is very critical for KDD. One method used to solve this limitation is the creation of a reduced training set for the k-nearest neighbour composed only by proto-typical examples. This uses an access method called M-tree it is implemented in the k-nearest neighbour algorithm. Based on a generic metric space M-trees can organize and search data sets. M-trees reduce the number of distance computations in similarity queries drastically.

This work analyzes four missing data treatment methods: the 10-NNI method using a k-nearest neighbour algorithm, the internal algorithms used by C4.5 and CN2 to treat missing data and the mean or mode imputation. These methods analyzed by inserting different percentages of missing data into different attributes of four data sets showing promising results. Even for training sets having a large amount of missing data the 10-NNI give very good results.

2.16 Fuzzy K-means Clustering Method

Dan Li et al. [17] 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. Here combine the clustering method with soft computing. Which is more tolerant of imprecision and uncertainty, and apply a fuzzy clustering algorithm to deal with incomplete data. Analysis shows that the fuzzy imputation algorithm has better performance than the basic clustering algorithm.

To extend the K-means clustering method use a fuzzy version to impute missing data. Fuzzy approach is applied because fuzzy clustering provides a better description tool when the clusters are not well separated. The original K-means clustering trapped in a local minimum status if the initial points are not selected properly. To get stuck in local minimum situation the continuous membership values in fuzzy clustering make the resulting algorithms less susceptible. If the percentage of missing values is high then the overall performance of the fuzzy K-means method is better than the basic K-means method.

2.17 A Review of Hot Deck Imputation

Hot deck imputation [18] replaces missing values with values from a "similar" unit. Usually use data from surveys. Replacing missing values of one or more variables of a non-respondent (called the recipient) with observed values from the donor (or a respondent) .Similar to the non-respondent. In hot deck method a missing attribute value is filled with a value from an estimated distribution for the missing value from the current data. There are two stages for hot deck implementation. In first stage the data are partitioned into clusters and in the second stage, one cluster each instance with missing data is associated. To fill in the missing values the complete cases in a cluster are used. This can be performed by calculating the mean or mode of the attribute within a cluster.

Hot Dock Imputation is two types:

- Random hot deck methods: donor is selected randomly from a set of potential donors.

- Deterministic hot deck methods: single donor is identified. Values are imputed "nearest" in some sense.

2.18 GBKII: An Imputation Method for Missing Values

Chengqi Zhang et al. [19] proposed Grey-Based KNN Iteration Imputation (GBKII) method for missing value. Missing data imputation is challenging issue in data mining and machine learning. To handle this issue present a Grey-Based K-NN Iteration Imputation method, called GBKII. GBKII is an instance-based imputation method. GBKII is referred to a non-parametric regression method in statistics and also efficient for handling with categorical attributes. Evaluate GBKII approach and demonstrate that which is more efficient than the k-NN and mean-substitution methods. This approach uses a GRG (Grey Relational Grade) to substitute for Minkowski distance or other alternative similarity measures during the process of searching for the nearest neighbor.

GBKII is an EM-like iteration imputation method and it is different from the EM and MI algorithms. GBKII algorithm is a non-parametric method and which is different from EM algorithm in which both the E and M steps depend on parametric models.

2.19 Feature Weighted Grey KNN (FWGKNN)

Ruilin Pan et al. [20] present a missing data imputation by K-nearest neighbours based on grey relational structure and mutual information. The *K* nearest neighbours (KNN) based classic imputation strategies are used to solve plague problem. To impute missing data a novel method is proposed it is known as Feature Weighted Grey KNN (FWGKNN) imputation algorithm. The proposed method imputes missing data iteratively in each class. The FWGKNN method imputes instances with missing data according to the amount of missing data in ascending order to improve imputation performance. Figure. 1 shows how FWGKNN algorithm work.

The mean or mode imputation approach is used to obtain a preliminary dataset in the first imputation. The instances imputed in the first imputation iteration are used as complete instances for the second imputation iteration. From the second imputation iteration, one by one all the missing data are imputed again, to improve the imputation performance. First suppose that the imputed value from the former iteration is missing when imputing the missing data, and the rest of them are viewed as known values (the latest imputed values). The similarity of instances is measured by the GRA; relative weight based on MI is taken into account. To determine the nearest neighbours of an instance with missing values the MI-weighted GRA metric is employed. At last, the missing value is imputed as the weighted value of the nearest neighbours.

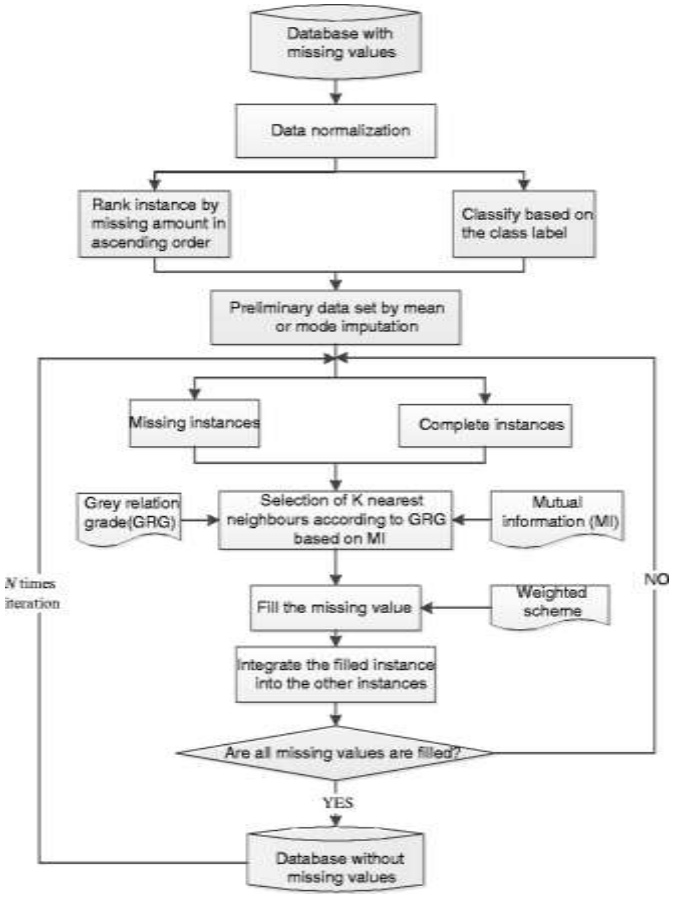

Fig -1: Flowchart of the proposed model for imputing missing data.

The third imputation iteration performs same as the procedure in the second. Finally, the imputed data converges or starts looping when the process terminates.

To overcome the difficulties of existing methods develop a novel framework for missing data imputation. Missing data not only damage the integrity of the data but also lead to the deviation of data analysis and data mining, The existing imputation methods has low accuracy and poor stability. Due to this limitations propose a system Missing value Imputation Algorithm based on Evidence chain (MIAEC). This algorithm is extended using Map-Reduce programming model. MIAEC algorithm help to process large amount of data.

3. CONCLUSION

The literature survey could fetch a number of existing missing data imputation systems. 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. Here propose a Missing value Imputation Algorithm based on Evidence Chain is used.

REFERENCES

[1] 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.

[2] 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.

[3] M. Ramoni and P. Sebastiani, " Learning Bayesian networks from incomplete databases", in Proceedings of the Thirteenth Conference on Uncertainty in artificial intelligence pp. 401-408, 1997.

[4] Man Leung Wong, Yuan Yuan Guo, "Learning Bayesian networks from incomplete databases using a novel evolutionary algorithm", Decision Support Systems 45 (2008), 368–383.

[5] M. Zhu and X.B. Cheng, "Iterative KNN imputation based on GRA for missing values in TPLMS", 2015 4th International Conference on Computer Science and Network Technology (ICCSNT), Harbin, China, pp. 94-99.

[6] P. Keerin, W. Kurutach and T. Boongoen, "Cluster-based KNN missing value imputation for DNA microarray data", IEEE International Conference on Systems, Man, and Cybernetics (SMC), Seoul, pp. 445-450, 2012.

[7] S. Wu, X.D. Feng and Z.G. Shan, "Missing Data Imputation Approach Based on Incomplete Data Clustering", Chinese Journal of Computers, vol. 35, no. 28, pp. 1726-1738, Aug. 2012.

[8] 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.

[9] 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.

[10] 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.

[11] Feelders, A.J. (1999), "Handling missing data in trees: surrogate splits or statistical imputation", in: Proceedings of the 3rd European Conference on Principles and Practice of Knowledge Discovery in Data Bases (PKDD99), Springer, Berlin, pp. 329– 334.

[12] F. Bashir and Hua-Liang Wei, "Parametric and non-parametric methods to enhance prediction performance in the presence of missing data", System Theory, Control and Computing (ICSTCC),19th International Conference on, Cheile Gradistei, pp. 337-342, 2015.

[13] A. Karmaker and S. Kwek, "Incorporating an EM-approach for handling missing attribute-values in decision tree induction", Fifth International Conference on Hybrid Intelligent Systems (HIS'05), pp. 6, 2005.

[14] Julián Luengo, José A. Sáez, Francisco. Herrera, Missing data imputation for Fuzzy Rule Based Classification Systems. Soft Computing, 16(5):863-881, May 2012.

[15] M. G. Rahman and M. Z. Islam, "iDMI: A novel technique for missing value imputation using a decision tree and expectation-maximization algorithm", Computer and Information Technology (ICCIT), 2013 16th International Conference on, Khulna, 2014, pp. 496-501.

[16] Batista, G.E.A.P.A., Monard, M.C.: An analysis of Four Missing Data Treatment Methods for Supervised Learning. J. Applied Artificial Intelligence 17, 519–533 (2003).

[17]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.

[18] R.J.A. Little and D.B. Rubin, Statistical "Analysis with Missing Data", 2nd ed, United States of America: Wiley-Interscience, 2002, pp. 200-220.

[19] 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.

[20] Ruilin Pan, Tingsheng Yang, Jianhua Cao, Ke Lu, Zhanchao Zhang, "Missing data imputation by K nearest neighbours based on grey relational structure and mutual information", Springer Science,Business Media New York 2015.

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.