International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Missing Data Imputation in High Dimensional Data Set using Local Similarity

C.Nalini, J.Sudeeptha

Abstract: ***Data quality is an important aspect for any data mining and statistical tasks. Presence of missing values in the dataset affects the data quality. Missing values refers to the event did not happen or the value does not exist. Data mining algorithms are not robust towards incomplete data. Imputation of missing values is necessary to improve the data quality for performing data mining and statistical analysis. The existing methods such as Expectation Maximization Imputation (EMI), A Framework for Imputing Missing values Using co appearance, correlation and Similarity analysis (FIMUS) use the whole dataset to impute missing values. In such cases, due to the influence of irrelevant record the accuracy of imputation may be affected. This can be controlled by only considering locally similar records to impute missing values. Local similarity imputation can be done through clustering algorithms such as k-means algorithm. K-means clustering efficiency depends on the number of clusters is to be defined by users. To increase the clustering efficiency, first distinctive value is imputed in place of missing ones and this imputed dataset is given to stacked autoencoder for dimensionality reduction which also improves the efficiency of clustering. Initial number of clusters to k-means algorithm is determined using fast clustering. Due to initial imputation, some irrelevant records may be partitioned to a cluster. When these records are used for imputing missing values, accuracy of imputation decreases. In the proposed algorithm, local similarity imputation algorithm uses only top knearest neighbours within the cluster to impute missing values. The performance of the proposed algorithm is evaluated based on Root-Mean-Squared-Error (RMSE) and Index of Agreement (d2). University of California Irvine datasets has been used for analyzing the performance of the proposed algorithm.***

Keywords**: *Data quality, missing values, clustering, Root-Mean-Squared-Error, Index of Agreement***

I. INTRODUCTION

Data mining is the process of knowledge discovery where knowledge is gained by analysing the large amount of data which could be stored in database, data warehouses, or other information repositories. Data mining have various applications and these applications have enriched the various fields of human life including business, education, medical, scientific, etc. Missing data occur for several reasons during data collection . Data cleaning attempts to fill in missing values, smooth out noise while identifying outliers and correct inconsistencies in the data to improve the data quality.

Types of missing data

Missing data are often categorized into the following three types: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).

- **Missing Completely At Random (MCAR)**

Data are missing independently of both observed and unobserved data. For example, a participant flips a coin to decide whether to complete the depression survey.

- **Missing At Random (MAR)**

Data are missing independently of unobserved data. For example, male participants are more likely to refuse to fill out the depression survey, but it does not depend on the level of their depression.

- **Missing Not At Random (MNAR)**

Missing observations related to values of unobserved data. For example, participants with severe depression, or side-effects from the medication, were more likely to be missing at end. Participants with severe depression, or side-effects from the medication, were more likely to be missing at end.

Imputation is the process of replacing missing data with substituted values which would improve the data quality. Missing values are usually handled by ignoring the tuple or by imputing missing value manually or using a global constant or mean or median value of same or different class. The imputation techniques are widely classified into two types namely

- Single Imputation
- Multiple Imputation

In Single imputation the missing value is replaced by a single value. The imputed values are assumed to be the real value which makes the data would have been complete. Multiple imputation is a statistical technique in which all possible values for imputation is imputed in different copies of the dataset and appropriate value is found by analyzing each dataset for its accuracy and the value which provides greater accuracy has been imputed. This process is carried out in three steps namely imputation, analysis and pooling.

II. RELATED WORK

Azim et al (2014) proposed Hybrid model for data imputation:

Manuscript published on 30 September 2019
* Correspondence Author
C.Nalini*, Professor,Department of Information Technology,Kongu Engineering College,Erode,India. Email:nalinikec@gmail.com
J.Sudeeptha, Application development associate in Accenture Chennai India.Email: sudeepthajayaraman@gmail.com

Using fuzzy c means and multi layer perception Fuzzy-c-means imputation method performs FCM clustering to partition the dataset into k clusters. This, in turn, will result in membership values of each record with a particular cluster and the clusters centroids. Then, all the incomplete data objects are evaluated by using the membership degree of that object and the values of cluster centroids. The accuracy of imputation is significantly affected by clustering results. Besides, the selection of a suitable k may be challenging for data miners.

Deogun et al (2004) proposed an algorithm for missing data imputation. Unsupervised learning algorithm such as K-Means uses Euclidean distance for finding distance between data objects in which distance between missing object and others could not be found. One solution of this problem to delete all the records with missing values, but the clustering accuracy and reliability would be severely affected.

Rahman et al (2014) proposed a Framework for Imputing Missing values Using co-appearance, correlation and Similarity analysis. FIMUS technique fills in missing values combined with correlation, co-appearance and similarity analysis. In this method, each imputed value is related to all other records and attributes, making the imputation complexity high. When dealing with large data, the time cost is high.

Rahman et al (2011) proposed an algorithm Decision tree-based Missing value Imputation technique for data preprocessing. DMI uses an existing decision tree algorithm to find a similar data subset, which belongs to a leaf that represents a given attribute. Then it estimates a missing value of the attribute according to the subset of similar records. The method also identifies that the similarities of the attributes within the records belonging to a leaf are usually higher than that of attributes within the records of the whole dataset. The supervised decision tree algorithm would produce the some heterogeneous leaves which affects accuracy of imputation. Schneider et al (2001) proposed an algorithm Estimation of mean values and covariance matrices and imputation of missing values. Expectation Maximization Imputation (EMI) technique utilizes correlations between attributes and their mean values are used to impute the missing ones. This method uses the whole dataset to impute a missing value, in which some irrelevant records could have unfavorable influences on the accuracy or complexity of imputation.

Fuzzy c-means and k-means algorithm provide easiest way for clustering but choosing suitable k value is difficult and k-means algorithm is not robust to incomplete or missing data. EMI algorithm uses the whole dataset to impute a single value where influence of irrelevant records would degrade the accuracy of imputation. FIMUS algorithm has same shortcomings of EMI along with it takes huge time to impute a value when the dataset is large. DMI algorithm uses decision algorithm for classification of dataset in which imputation accuracy affects on efficiency of classification. In existing algorithms such as k-means and FCM choosing suitable number of cluster values is difficult. Fast clustering algorithm imposes good results but back tracking is not possible. The proposed algorithm apply fast clustering algorithm to determine the number of clusters possible to partition the dataset

III. PROPOSED METHOD

Presence of missing values affect the data quality and makes it unfit for data mining and statistical tasks. Imputing appropriate value in place of missing one is necessary to improve the data quality. Missing data are imputed in the dataset by considering locally similar data in the dataset. Locally similar data could be identified by clustering the dataset. As clustering efficiency decreases due to the presence of missing data, distinctive value is imputed in place of missing ones. Then, the dataset is given to a stacked auto-encoder where higher dimensional dataset is transformed to a lower dimension. This process also reduces the effects of distinctive imputation. The features extracted from the stacked auto-encoder is given to a hierarchal clustering algorithm named as fast clustering algorithm to find the possible number of clusters could be formed from the extracted features. The number of clusters found after fast clustering and the extracted features is given as input to k-means algorithm for clustering. Top k-nearest neighbours for each missing values within the cluster is found. Probability of influence of each top k-nearest neighbour is found which is used for calculating the value for imputation for that missing record. Product of the probability and attribute value of the corresponding top k-nearest neighbour is calculated and this value is imputed in place of missing ones. The imputed value is compared with the previous iteration value and error rate is calculated and the process is terminated when error rate is found to be below the threshold value. The Frame work for Missing Data Imputation using Local Similarity based on Fast Clustering is illustrated in the figure 1.

Fast Clustering Algorithm (SAE_FC)

Input: Extracted Features H

Output: Number of clusters k_c.

1: Set first t samples of H as the cluster Centers;
2: for j = t + 1 to n do
3: minDis ← GetMinDis (H_j , Centers);
4: [minDisC,maxDisC] ← GetDisC(Centers);
5: [Centers] ← UpdateC(minDis, minDisC,maxDisC);
6: end for
7: k_c← NumberofClusters
8: Return number of cluster k_c.

In SAE_FC algorithm, first t records are set as cluster centers in Step 1. Minimum distance between the other record and the cluster centers have been calculated and Centers have been updated accordingly (Steps 2 to 6). Number of clusters are then calculated and value is returned to the calling function.

Retrieval Number: C6435098319/2019©BEIESP
DOI:10.35940/ijrte.C6435.098319
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

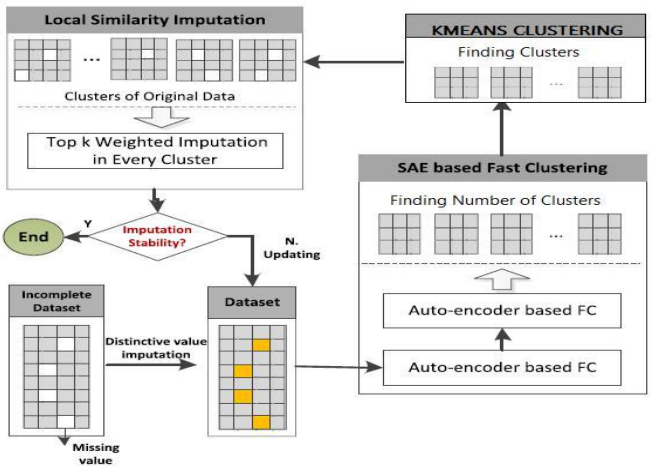

Figure 1. Frame work

K-Means Clustering Algorithm (KMeans)

Input: Extracted Features H, Number of clusters kc .

Output: Clusters

1: while 1 do
2: Set first kc samples of H as the cluster Centers;
3: for j = 1 to k do
4: for i = 1 to n do
5: $minDis_i \leftarrow GetMinDis(H_i, Centers)$;
6: end for
7:end for
8: [Centers] ← Mean(minDis);
9: if $Centers_{pre}$ a = $Centers_{cur}$ then break; end if
10: end while
11: Return Clusters

In KMeans algorithm, first k samples are kept as cluster centers (Step 2) and distance between each record to the cluster centers is calculated (Steps 3 to 7). Mean value of the distance of particular center is kept as new center (Step 8) and process is iterated until there is no change in the centers (Steps 9 to 12).

Each input samples are mapped to hidden layer feature by using the Encoding function which is expressed in Eq.1

$$h_i = f_\theta x_i = f\left(w^{(1)} x_i + b^{(1)}\right) \quad (1)$$

Where $w^{(1)}$ is initial weight, $b^{(1)}$ is initial bias and x_i denotes input records of the dataset.

Reconstruction of hidden layer to output layer has been done using decoding function which is expressed in Eq.2

$$z_i = f_\theta'(h_i) = f\left(w^{(2)} h_i + b^{(2)}\right) \quad (2)$$

Sigmoid Function is used as activation function which is expressed in Eq.3

$$f(x) = \frac{1}{(1+e^{-x})} \quad (3)$$

Where, x refers to the input of the activation function.Weight updation formula is expressed in Eq. 4 and Eq. 5

$$w_{ijnew} = w_{ijold} + \Delta w_{ij} \quad (4)$$

$$\Delta w_{ij} = \alpha \delta_j x_i \quad (5)$$

Where, Δw_{ij} is the basic delta rule, α is the learning rate, δ_j is difference between the actual and observed output and is the input value.

Distance between cluster centers can be calculated using Euclidean distance formula which is expressed in Eq. 6

$$Dis(c_p, c_q) = \sqrt{\sum_{k=1}^{m} (c_{pk`} - c_{qk`})^2} \quad (6)$$

Where c_p and c_q are cluster centers, m is the number of fields in the dataset.

Local Similarity Imputation

Missing values are imputed based on local similarity of data which is found using k-nearest neighbour algorithm. Within each cluster, for each record with missing values top k-nearest neighbour is found using Euclidean distance formula which is expressed in Eq. 6. For this k-nearest record, their probability of influencing the missing record is estimated using the formula expressed in Eq. 7.

$$p_i = \frac{1/dist_i}{\sum_{i=1}^{k} \frac{1}{dist_i}} \quad (7)$$

Where, is the distance between the missing record and the top k-nearest neighbour record The imputation value is calculated using the formula expressed in Eq. 8

$$I = \sum_{i=1}^{k} p_i x_{iu} \quad (8)$$

Where, is the probability of influence of a record to the missing record and missing attribute value of top k-nearest neighbours. When all missing values are imputed, the error between current imputation Icur and previous imputation is estimated based on Root Mean Squared Error which is expressed inEq. 9

$$err = \sqrt{\frac{1}{t} \sum_{i=1}^{t} (I_{pre.i} - I_{cur.i})^2} \quad (9)$$

Local Similarity Imputation Algorithm

Input: Incomplete dataset ID (n×m). Parameter γ, k.

Output: Imputed dataset D.

1. ID ← PreIm(dv, ID);
2. v_i=m; V =ID;
3. while 1 do
4. for i = 1 to l do
5. AE ← SetupAE (v_i, h_i);
6. [AE,H] ← GetHiddenRepresentation (AE, V);
7. k_c ← SAE_FC (H);
8. [AE,Clusters, k_c] ← TrainAE (AE,V, H,Centers,opts);
9. KM ← KMeans(H, k_c)
10. v_i +1=size(H, 2);V =H;
11. end for
12. for i = 1 to k_c do
13. Per ← NumberofRecords(Clusters.i);
14. if Per < k then
15. Partition all the items to other clusters;
16. end if
17. end for
18. for i = 1 to k_c do
19. [InData, p] ← GetInData (Clusters.i);
20. for j = 1 to p do
21. Using DisTK to impute InData[j] via (6),(7) and (8).
22. end for
23. Getting the imputation set I_{cur} of Clusters.i;
24. I_{cur} ← AddSet (I_{cur});

25. end for
26. Calculate the err between current and last imputation by (9);
27. if err<γ or loop> 100 then
28. D ← OutputDataset (ID, I_{cur});
29. break;
30. else
31. ID ← UpdateDataset (I_{cur});
32. I_{pre} ← I_{cur} ;
33. end if
34. end while
35. Return the complete data set D;

In Local similarity imputation algorithm, the missing value is imputed with distinctive value (Step 1). Then the imputed dataset is given to autoencoder from where hidden representation of dataset has been extracted (Steps 4 to 11). If number of samples in a cluster is less than the k value then partition the clusters into others (Steps 12 to 17). Top k-nearest neighbour for a missing record within the cluster is found and imputation is performed (Steps 18 to 34).

IV. RESULT ANALYSIS

The proposed algorithm is implemented using R programming language .The performance of the algorithm has measured by using Root- Mean-Squared Error(RMSE) and Index of Agreement(d2). Root- Mean-Squared Error is one of common evaluation criteria for missing data.The value of RMSE can range from 0 to ∞. A lower value means a better imputing.RMSE value is calculated using the formula expressed in Eq. (10)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{N}(P_i - O_i)^2} \qquad (10)$$

Index of Agreement (d2) is also one of the most commonly used evaluation criteria for missing values. The value of d2 can vary from 0 to 1. A higher value indicates a better accuracy. This can be calculated by the formula expressed in Eq. (11)

$$d = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(|P_i - \bar{O}| + |O_i - \bar{O}|)^2} \quad , \quad 0 \le d \le 1 \qquad (11)$$

Figure 5.1 illustrates the Root Mean Squared Error of the dataset glass and wine. When the missing ratio is 5%, the error rate is 0.05 and when the missing ratio increases to 10%, the error rate also increases to 0.13. As the missing ratio of the dataset increases, the error rate also increases.

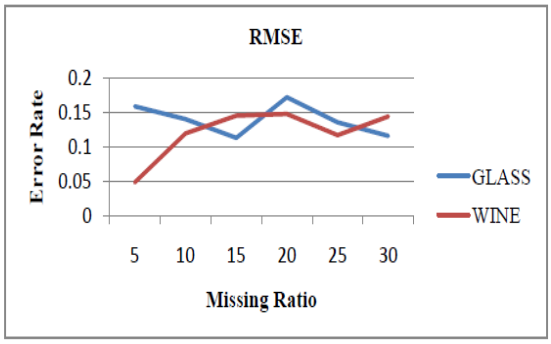

Figure 2 RMSE for the two datasets Glass and Wine

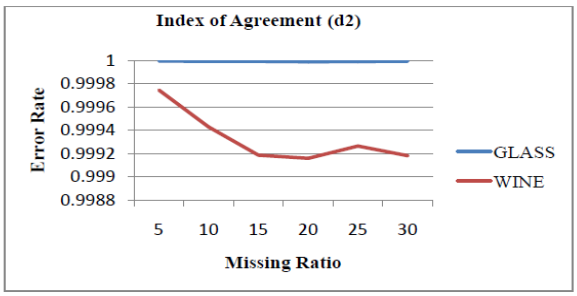

Figure 5.2 Index of Agreement (d2) for the two datasets Glass and Wine

Figure 5.2 illustrates the Index of Agreement (d2) of the dataset glass and wine. When the missing ratio is 5% the error rate is 0.99975 and when the missing ratio is 10%, the error rate is 0.9994 which means the accuracy decreases in wine dataset. In glass dataset, the error rate is almost constant as the values in attribute Iron (Fe) and Barium (Ba) is almost zero and their standard deviation as 0.4972 and 0.0974 which shows they are not uniformly distributed. This may be a reason for the constant error rate.

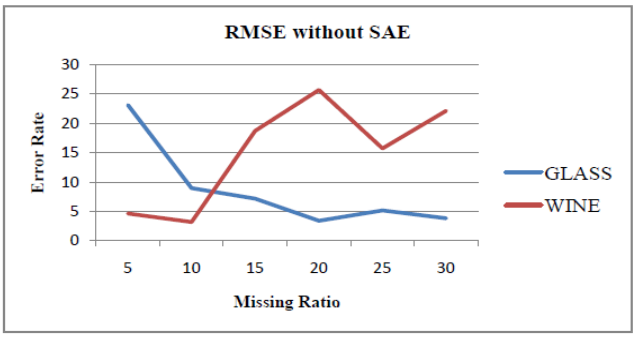

Figure 5.3 RMSE without SAE for the two datasets Glass and Wine

Figure 5.3 illustrates the Root Mean Squared Error for the dataset glass and wine without implementation of SAE algorithm. In wine dataset, the error rate increases with increase in the missing ratio. In glass dataset, the error rate decreases with increase in missing ratio which may due to the values in the attribute Iron(Fe) and Barium(Ba) is almost zero and their standard deviation is also very low. Without implementation of SAE algorithm the error rate increases when compared to implementation with SAE.

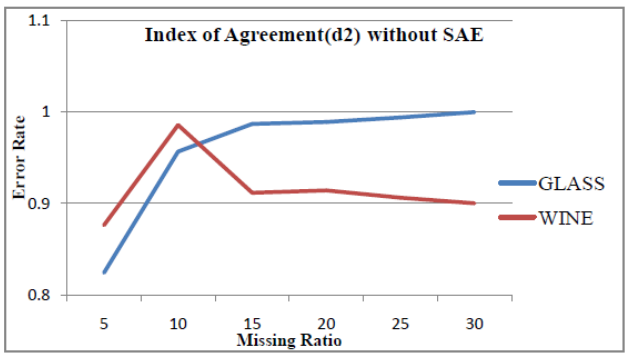

Figure 5.4 Index of Agreement (d2) without SAE for the two datasets Glass and Wine

Figure 5.4 illustrates Index of Agreement (d2) of the dataset glass and wine without SAE implementation. When the missing ratio is 10%, the error rate is 0.9859 and when missing ratio increases to 15% the error rate increases to 0.9117 in wine dataset. As the missing ratio increases the error rate also increases in wine dataset. In glass dataset, the error rate decreases with increase in missing ratio and this might be due to the values in attribute Iron (Fe) and Barium (Ba) which is mostly zero and their standard deviation is also very low. The result shows that implementation of SAE plays a major role to increase the accuracy of imputation.

V. CONCLUSION

Imputation of missing values is made easier by using principle features of the dataset which is extracted using stacked autoencoder. As existing clustering algorithms are not robust towards missing data, the extracted features upgrade the clustering efficiency. Fast clustering algorithm determines the number of clusters could be formed with the extracted features and k-means clustering algorithm clusters the dataset efficiently. K-nearest neighbor imputation algorithm is used for missing value imputation and evaluation techniques such as RMSE and Index of Agreement (d2) is used to measure the accuracy of imputation. The result shows that implementation of the proposed algorithm produce lower RMSE value and higher d2 value. The model has developed for numeric dataset. In future the algorithm can be extended for mixed dataset and has to be extended to remove the missing values in data streaming.

REFERENCES

1. Arslan A and Aydilek I B (2013), 'A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm', Information Sciences,vol. 233, no. 6, pp. 25–35.
2. Islam Z Rahman G and (2011), 'A decision tree-based missing value imputation technique for data preprocessing', in Proceedings of the 9th Australasian Data Mining Conference ,Computer Society, vol. 121 pp. 41–50.
3. Islam Z Rahman G and (2014), 'FIMUS: A framework for imputing missing values using co appearance, correlation and similarity analysis', Knowledge.-Based System, vol. 56, no. 1, pp. 311–327.
4. Liang Zhao, Zhikui Chen, Zhennan Yang, Yueming Hu, Mohammad S. Obaidat (2016),"Local Similarity Imputation Based on Fast Clustering for Incomplete Data in Cyber-Physical Systems", IEEE Systems Journal, Online, 2016. DOI: 10.1109/JSYST.2016.2576026.
5. Schneider T(2001), 'Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values', Journal of Climate, vol. 14, no. 5, pp. 853–871.

AUTHORS PROFILE

Dr.C.Nalini is working as a professor in Kongu Engineering College, Erode, has 27 years teaching experience. Her research interests include Data mining, Machine learning. and Optimization techniques. She has published more than 25 papers in various international journals and national and international conferences. She is a life member of Computer Society of India.

J.Sudeeptha is working as a Application development associate in Accenture Chennai. She has completed her Under graduate programme in Computer Science Engineering in 2019. Her Research interest includes Data mining, Machine Learning and Deep Learning. She has published 5 papers in various national and international conferences.