NOVEL CLUSTERING APPROACH TO FIND THE MISSING DATA IN LARGE DATASETS

[1]K.SUREKA, [2]MEGHANA.Y.M

[1]ASSISTANT PROFESSOR/CSE, [2]ME/CSE STUDENT
RVS COLLEGE OF ENGINEERING
DINDIGUL

***Abstract*: The data is growing tremendously and handling of data for various business and scientific applications is an on demand need. But, it is unavoidable to have missing values in any dataset. The reason for the missing values could be varying with different mode of applications and the concern here is that how those missing values are handled. The various imputation methods in practice are studied in detail and a new approach based on multiple linear regression model is proposed. The efficiency of the proposed imputation approach is tested with K-Means clustering by comparing the clustering of the original data and the imputed data. The proposed approach is found to perform than the other statistical imputation models.**

***Keywords*: missing data, imputation, clustering, Multiple Linear Regression**

I INTRODUCTION

Clustering a kind of machine learning model is in practice for the past decades, it cannot handle dataset which contains missing values in some of the features [1][2] . There are works that handle the missing values in both supervised [3][4] and unsupervised models[5].

The following figure 1 represents the various mechanisms. The figure is summarised from [6].

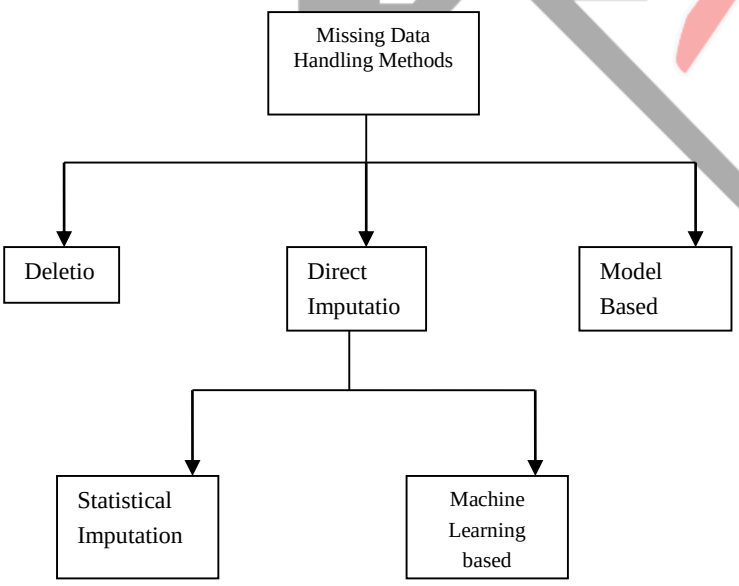


Figure 1 Various Imputation Methods

The objective of this work is to study the various imputation mechanisms that are in practice and experiment with few of the statistical models. A new approach is also proposed based on multiple linear regression for imputing missing data. The results of which are compared with the existing models.

The rest of the paper is organised as follows. The next section explains the various states of models and the third section gives the System model. The Fourth section explains the result obtained and their comparison. The last and the final section give the conclusion.

II STUDY OF THE VARIOUS MODELS

The studies include works that deals with the various imputation mechanisms and also works that experiments on how the clustering models handles missing data efficiently and hence improve the efficiency of the clustering.

[7] Provides a mechanism for finding a threshold value based on which a decision is made whether to include an element in to a cluster, not to include the element in the cluster or to delay the process of making a decision on it. Concept of game theory is applied and the result shows that the accuracy level obtained is similar to other models while the generality is improved.

[8] Employs convolutional Bidirectional LSTM model for filling up the missing data in spatio-temporal data. It is observed from the results that though this deep neural model is found to provide better performance than the other model, the error values are still high.

A two-step approach is provided in [9]. The authors work with the traffic data and K-Means Clustering is used for grouping of roads that has similar traffic pattern and further deep learning model is used for filling up the missing data with the observed relations. The model has been evaluated for its consistency for different missing rates. [10] deals with applications that has only proximity matrix as input and there are missing values in it. The authors have proposed an algorithm that imputes the values and completes the proximity matrix. The accuracy varies with different datasets and is not found to be remarkably high. In [11] authors have extended the K-means algorithm to design a new algorithm that can work even with the dataset that has missing values in it. The algorithm is designed in such a way that it acts as a conventional K-Means algorithm when the given dataset has no missing values.

[12] introduces a novel algorithm built on the K-Means algorithm that handles missing data scenario in a more robust way. Maximization Minimization approach is used for

optimization and the model shows better result in terms of clustering accuracy than the other models.

A method that contains the K-Means and RBF neural networked is used in[13] for predicting the missing values. Clustering is done with the K-Means algorithm and the neural network model is used for predicting the missing values. This model is found to produce better results when the missing rate is low and the accuracy degrades with the increase in the missing rate.[14] have employed autoencoders for finding the missing values in the dataset. The traditional architecture of the autoencoders is modified by the authors to get a better result.[15] combines the denoising encoders and generative adverisal networks for predicting the missing values and it proves to produce 30% higher accuracy than the other state of art models.

The missing data problem in the context of IoT is considered by [16] . a solution that combines the probabilistic matrix factorization method with K-Means clustering is is proposed which produces better result than the models with support vector machine and deep neural networks. [17] also develops a new k means algorithm that imputes the missing data when the clustering algorithm is in progress. The clustering accuracy iss found to be higher than the other state of art models.

There are also other models such as Bayesian framework[18] and Expectation Maximization [19]algorithm that deals with missing values and the later is used vastly. In addition to this there are various imputers that are prebuilt in python scikit-learn module[20].

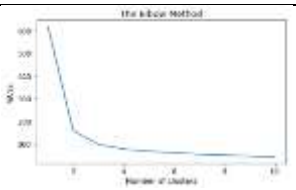
But from the study it has been observed that most of the models concentrate on the accuracy of the machine learning model rather than the accuracy of the imputed values. Regression model would help in addressing both.

III SYSTEM MODEL

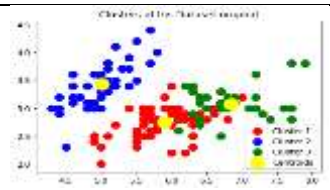
This section explains the various statistical methods that have been used for imputing the missing values. It has been observed from the literature [17] that statistical model outperforms the KNN filling model and the EM Model.

For experimenting, iris dataset is considered. To find the efficiency of the proposed imputation method, first K-Means algorithm is applied to the original data and later K-means is applied to data that contains missing values imputed with different methods.

For identifying the optimal number of clusters, Elbow method is followed. Elbow method is a proven method to find the number of clusters. In this method, a plot is made with different number of clusters and the point where the elbow occurs specifies the number of clusters that can be used. For the considered dataset, the following figure 2 represents the result obtained from the elbow method and the clusters formed from the optimal number of clusters 3.



Optimal Number of Clusters with Elbow Method



K-Means Clustering of the Iris Dataset without missing values

Figure 2 Elbow Graph and the clusters of the Iris Dataset

The Elbow method is based on the WCSS which stands for within cluster sum of squares.

Within cluster sum of squares = $\sum_1^k (x_i - y_i)$,

here,

k is the number of clusters,

x_i Represents the values in the particular cluster and

y_i Represents the centroid of that cluster.

Experimentation is done with two copies of the iris dataset with different percentage of the missing values. A dataset with 5% missing values and another dataset with 10% missing values is crested from the original dataset .
The missing values are filled with different methods and experimented. Commonly used statistical models of filling the missing values with mean and median is employed and a new approach of filling the missing values with regression model is also done. Approach of filling the missing values with regression is described below.

- From Dataset D, n Data sets are creates D_1, D_2,..D_n, where n represents the number of features
- For Each D_i, Multiple Linear Regression is applied and the missing feature is predicted. Multiple Linear regression can be defined as follows.

$yi=\beta_0+\beta_1 x_{i1}+\beta 2 x_{i2}+...+\beta p x_{ip}+\epsilon$
where
yi is the dependent variable
xi is the independent variables
β_0 is a constant
β_1, $\beta_{2...}$represents the slope coefficients
ϵ represents the residual
the dependent variables change for the individual datasets in the proposed case.

- All the Di are combined back together to form a new Imputed Dataset D^I
- K-Means clustering is applied to D and D^I , The performance of the clustering is measured with fowlkes_mallows_score. It is used to find the similarity between two clusterings. It can be defined as below.

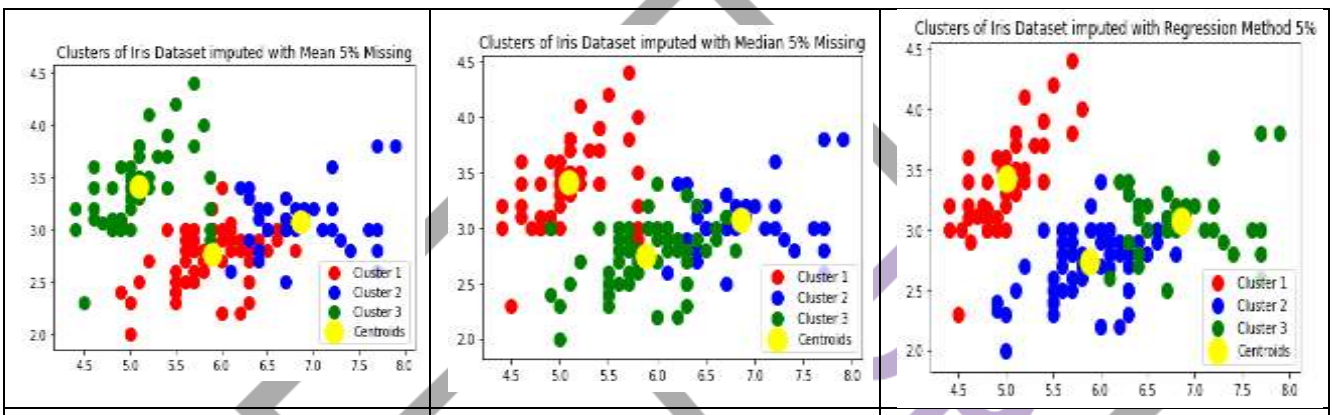
fowlkes_mallows_score = TP/ sqrt((TP + FP) * (TP + FN))

where,
TP stands for True positive and it represents the number of pairs of elements that belongs to the same cluster in both the actual cluster and predicted cluster, here actual cluster refers to the cluster of original data and predicted cluster refers to cluster dof data that contains imputed values
FP stands for False positive and it represents the number of pairs of elements that belongs to the same cluster in the actual cluster but not in the predicted cluster
FN stands for False Negative and it represents the number of pairs of elements that belongs to the same cluster in the predicted cluster but not in the actual cluster

The following figure 3 represents the result of K-Means clustering applied on dataset the missing values of which are imputed with different approaches. The percentage of the missing data is 5%.

IV RESULTS

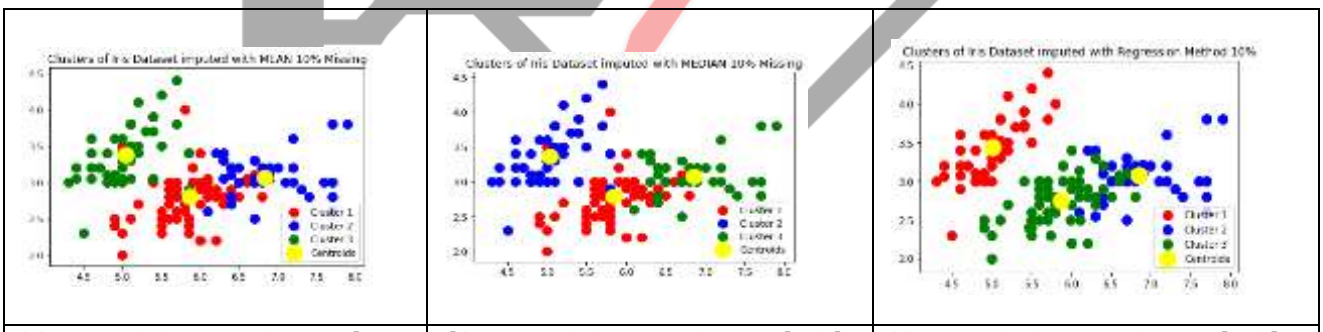
The accuracy of the clustering is measured with fowlkes_mallows_score and the result of different method of imputation are compared for different percentage missing data. The following figure 5 represents the fowlkes_mallows_score obtained with comparison of the original data with the data that has 5% missing values imputed with different approaches.



(a)5% Missing Data imputed with Mean	(b) 5% Missing Data imputed with Median	(c)5% Missing Data imputed with Multiple Regression

Figure 3 K-Means clustering of Data imputed with different approaches (5% Missing Data)

The following figure 4 represents the clustering obtained with 10% missing data.



(a)10% Missing Data imputed with Mean	(b) 10% Missing Data imputed with Median	(c)10% Missing Data imputed with Multiple Regression

Figure 4 K-Means clustering of Data imputed with different approaches (10 % Missing Data)

The results obtained and the performance comparison of the different imputation models are given in the next section.

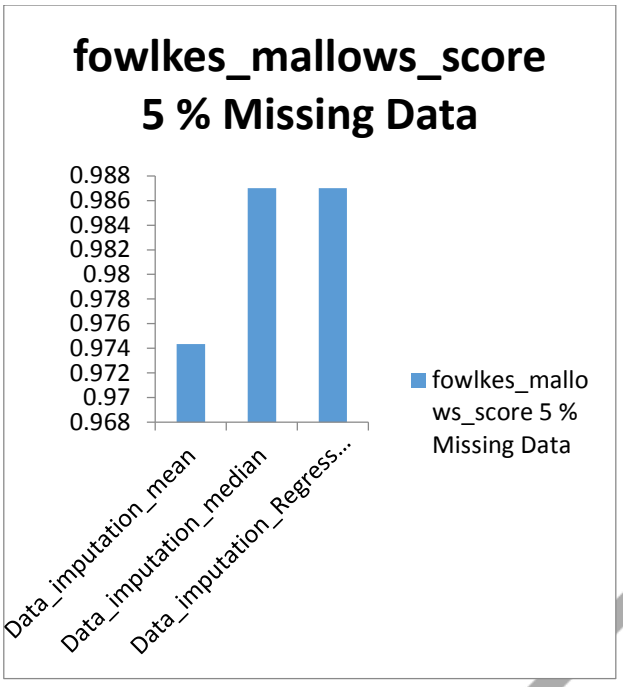


Figure 5 comparison of fowlkes_mallows_score (5 % Missing Data)

The following figure 6 represents the results obtained with 10% missing data.

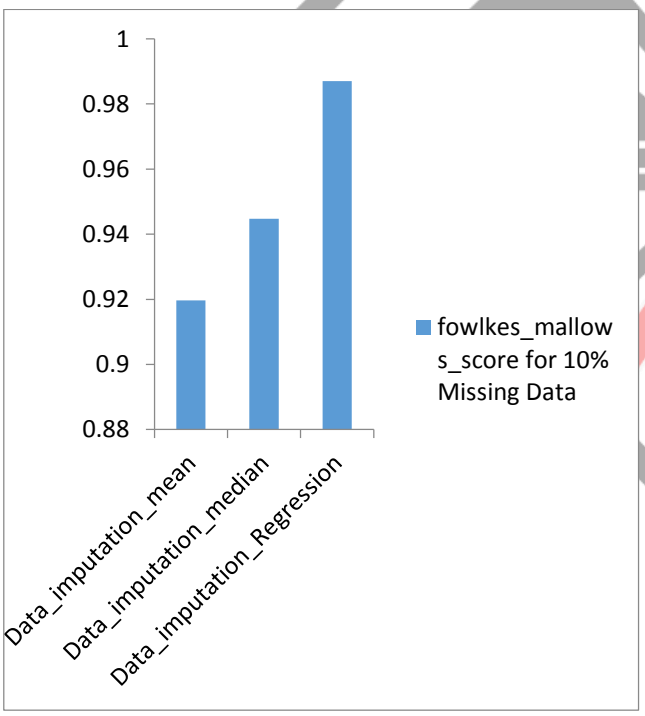


Figure 6 comparison of fowlkes_mallows_score (10 % Missing Data)

The following figure 7 represents the comparison of the result obtained with different imputation methods for different percentage of missing data. It can be observed from the result that in case of 5% missing data imputation with regression performs equally well as the imputation made with median and comparatively higher than the one done with mean. When there is an increase in the percentage of the missing data, there is considerable performance degradation in case of imputation with mean and median which is not true in case of imputation with regression.

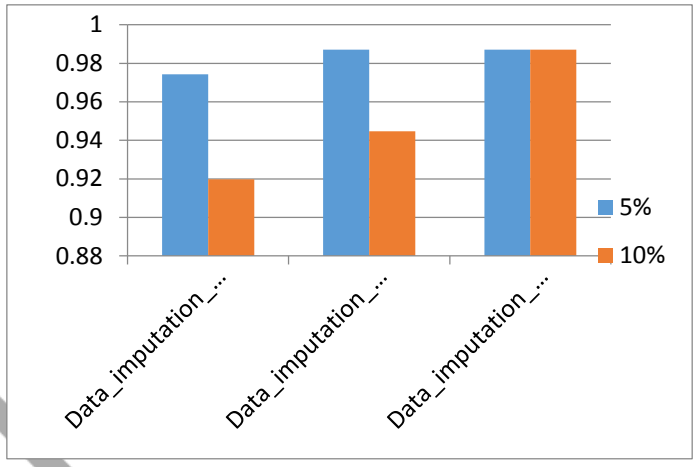


Figure 7 Comparison of fowlkes_mallows_score obtained for different percentage of missing data

V CONCLUSION

Statistical models are generally followed for imputing missing data in the dataset before applying any kind of machine learning algorithms on it. A new approach of filling the missing values with multiple linear regression model is proposed and found to perform well than the existing statistical models. It is also observed that there is no degradation in the performance of the model when the percentage of the missing value increases.

REFERENCES

[1] A. Trivedi, P. Rai, H. Daumé, III, and S. L. DuVall, ``Multiview clustering with incomplete views," in *Proc. NIPS Workshop*, 2010.

[2] P. Ezatpoor, J. Zhan, J. M.-T.Wu, and C. Chiu, ``Finding top-*k* dominance on incomplete big data using MapReduce framework," *IEEE Access*, vol. 6, pp. 7872_7887, 2018.

[3] J. Shen, E. Zheng, Z. Cheng, and D. Cheng, ``Assisting attraction classication by harvesting Web data," *IEEE Access*, vol. 5, pp. 1600_1608,2017.

[4] H. Timm, C. Döring, and R. Kruse, ``Different approaches to fuzzy clustering of incomplete datasets," *Int. J. Approx. Reasoning*, vol. 35, no. 3,pp. 239-249, 2004.

[5] D. Lam, M.Wei, and D.Wunsch, ``Clustering data of mixed categorical and numerical type with unsupervised feature learning," *IEEE Access*, vol. 3, pp. 1605-1613, 2015.

[6] García Laencina P.J et al. *Pattern Classification with Missing Data: A Review*. Neural Comput Applied. 2009. 9(1): 1–12.

[7] Mohammad Khan Afridi, Nouman Azam, JingTao Yao, Eisa Alanazi, 2018,
A three-way clustering approach for handling missing data using GTRS,
International Journal of Approximate Reasoning,Volume 98,pp11-24.

[10] Samira Karimzadeh, Sigurdur Olafsson,2019,Data clustering using proximity matrices with missing values,Expert Systems with Applications,Volum e 126,pp. 265-276,

[11] Lithio, A, Maitra, R. 2018, An efficient *k*-means-type algorithm for clustering datasets with incomplete records. *Stat Anal Data Min: The ASA Data Sci Journal*. 11: 296– 311.

[12] Jocelyn T. Chi, Eric C. Chi & Richard G. Baraniuk (2016) k-POD: A Method for k-Means Clustering of Missing Data, The American Statistician, 70:1, 91-99, DOI: 10.1080/00031305.2015.1086685

[13] Shi Z., Li X., Su Z. (2018) Power Missing Data Filling Based on Improved k-Means Algorithm and RBF Neural Network. In: Sun X., Pan Z., Bertino E. (eds) Cloud Computing and Security. ICCCS 2018. Lecture Notes in Computer Science, vol 11067. Springer, Cham

[14] Xiaochen Lai, Xia Wu, Liyong Zhang, Wei Lu, Chongquan Zhong,2019
Imputations of missing values using a tracking-removed autoencoder trained with incomplete data,Neurocomputing,Volume 366,pp 54-65.

[8] Asadi, R., & Regan, A. (2019). A convolution recurrent autoencoder for spatio-temporal missing data imputation. *ArXiv, abs/1904.12413*.

[9] W. C. Ku, G. R. Jagadeesh, A. Prakash and T. Srikanthan, "A clustering-based approach for data-driven imputation of missing traffic data," *2016 IEEE Forum on Integrated and Sustainable Transportation Systems (FISTS)*, Beijing, 2016, pp. 1-6.

[15] Huan Wang, Zhaolin Yuan, Yibin Chen, Bingyang Shen, Aixiang Wu,2019
An industrial missing values processing method based on generating model,
Computer Networks,Volume 158,Pages 61-68,

[16] B. Fekade, T. Maksymyuk, M. Kyryk and M. Jo, "Probabilistic Recovery of Incomplete Sensed Data in IoT," in *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 2282-2292, Aug. 2018.

[17] S. Wang *et al*., "K-Means Clustering With Incomplete Data," in *IEEE Access*, vol. 7, pp. 69162-69171, 2019.

[18] R. J. Little and D. B. Rubin, *Statistical Analysis With Missing Data*,vol. 333. Hoboken, NJ, USA: Wiley, 2014.

[19] A. P. Dempster, N. M. Laird, and D. B. Rubin, ``Maximum likelihood from incomplete data via the *EM* algorithm," *J. Roy. Stat. Soc., B (Methodol.)*, 1977, pp. 1_22.

[20] Scikit-learn: Machine Learning in Python, Pedregosa *et al*., JMLR 12, pp. 2825-2830, 2011.