International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 4 (2015) pp. 9969-9982
© Research India Publications
http://www.ripublication.com

An Alternate Imputation Technique of A Mean Method For Missing Values and Comparative Study With Neighbor Methods

Thirukumaran S[1], and Sumathi A [2]

[1] Research scholar, Anna University, 635025, Chennai, INDIA
[e-mail: thirukumaran.mca@adhiyamaan.in]
[2] Electronics and Communication Engineering, Adhiyamaan college of Engineering,
635109, Hosur,
Tamilnadu, India
[e-mail: sumathi_2005@rediffmail.com]
**Corresponding author: Thirukumaran.S*

Abstract

Inadequate data negligible for knowledge discovery Missing value imputation is real confront to the world evolved more than five to six decades, to provide the solution for inadequate data the technique available with so many types and research team focuses either on imputation and improving imputation accuracy level. Idea behind this work is imputing the value(s) to missing data (attribute) of different types of Medical Datasets taken by proposed Algorithm called Mean method by Step digression (MMSD) and the other three Algorithms projected here is usual and existing algorithm which is utilized for comparative study with proposed algorithm. The results we got for proposed method is compared and outlined with the existing imputation algorithms.

Keywords: Missing value imputation, KK-Nearest missing value imputation, Hot deck imputation, Regression imputation, Multiple imputation.

Introduction

In the decades of 1990`s and around 2000 the technique missing value imputation has wide scope in the area of industry, research datasets, Medical database, Sensor data, survey data editing [1] etc...The complete medical dataset documentation is the knowledge helps to take decision or to build the analysis for the treatment of patient which is followed in the developed country survey reveals. The Fig. 1 gives the feel to realize how would be the missing value data set available, the missing value in the dataset, incomplete dataset will not support for the analysis leads to bias and it raises the issues no knowledge producing to resolve the issues the imputation Technique experimented by the researchers. The single data or group of data not available from the database treated as a missing value(s) can be filled .Generally the imputation methods are established based on statistical algorithms disseminated into two areas 1.Model based method, 2. Data driven method [2] [3]. Model based methods are outcomes of random attributes (variables) estimated by unknown parameters and data driven method are not outcomes of random attributes estimated by known parameters. The missing data, the techniques exist to directly calculating the MEAN value [4] to get the impute value so called base method from this point the imputation techniques grown as simple approach (data driven) and multidimensional approach. **Fig. 1** Shows data set with missing value and data availability Simple approach are Hot deck(HD), cold deck(CD), Multidimensional techniques called regression[5], decision trees[6], maximum likelihood, and least squares approximation techniques are example for model based method. Techniques ahead diversification into multiple imputation, nested multiple imputation, multiple regression method etc.. The Multidimensional and diversification methods use all the data for the imputation process and parallel imputation will be performed simultaneously. The point to be known is what are the challenges arises when the imputation is performed are What mechanism need to measure the complexity of the missing values?, how to handle the inappropriate missing values?.

The scope of this paper an alternate algorithm for mean method proposed and named as Algorithm1 is phenom called: MMSD estimates the Ximpute by Digression method for grouped data allows to impute the same to the missing place to the ten medical dataset. The dataset must be classified as a grouped data with missing data before imputation starts. Next Direct mean Method (DM) an Algorithm2 exist in the survey used for comparison with proposed Algorithm1, Hence it is the baseline technique and the other technique is a Short-Cut Method mean(SCM) an Algorithm3 designed which also used for comparison with an algorithm1. Finally a proposed algorithm impute results are compared with other existing techniques results to reduce the complexity and proved to increase the efficiency of imputation values to the medical database.

Attribute 1	**Attribute 2**	**Attribute 3**	**Attribute 4**	**Attribute 5**	**Attribute 6**	**Attribute …**	**Attribute N**
1 Data	Data	Data	Data	?	Data	Data	Data
2 Data	Data	Data	?	Data	Data	Data	?
3 Data	Data	Data	Data	?	Data	Data	?
4 Data	?	?	Data	Data	Data	Data	Data
N Data	?	Data	Data	?	Data	Data	?

Figure 1: Understanding of Missing values of database Structures

Related Work Data Imputation Method

A Small Review of Missing imputation Technique

Missing completely at random (MCAR): The possibility of the missing value will not depend on other attributes of the dataset. Mathematical notation said by Rubin.

$$\Pr\left({}^{\mathrm{Xm}}\!\!\diagup_{\mathrm{Xmiss}}, \mathrm{Xobs} \right) = \Pr\left(\mathrm{Xm} \right) \tag{1}$$

Where Xm denotes the missing value, and Xmis and Xobs denote the unobserved data and observed data of database.

Missing at random (MAR): the missing value does not depend on the other attributes data rather it depends on corresponding attribute. Mathematical notation said by Rubin.

$$\Pr\left({}^{\mathrm{Xm}}\!\!\diagup_{\mathrm{Xmiss}}, \mathrm{Xobs} \right) = \Pr\left({}^{\mathrm{Xm}}\!\!\diagup_{\mathrm{Xobs}} \right) \tag{2}$$

Missing not at random (MNAR): When neither MCAR nor MAR hold, the missing value of attribute will depend on the same attribute itself. MCAR and MAR dataset are imputable whereas MNAR is not.

Missing data imputation techniques can be categorize to the ignorable missing data imputation technique are the single imputation and the multiple imputation, and the non-ignorable missing data imputation technique are the likelihood based methods and the non-likelihood based methods [7].

Hot Deck Method

Imputation method root classification into Hot deck and cold deck the focus is on Hot deck method is base method next level classification into Weighted Sequential Hot Deck, Weighted Random Deck.

Weighted Sequential Hot Deck: The weighted sequential hot deck imputation technique [8-9] enforced by two issues weighted and unweighted : unweighted sequential hot deck method, the imputed values are certainly equals to the weight it drives to biased imputation. Sorting methodology of the unweighted sequential hot deck method preserved by weighted sequential hot deck it is not necessary to widely implemented the weighted sequential hot deck.

Weighted Random Hot Deck: Sample random sampling are used to get the impute value and sampling weight ignorance tends to estimators bias [10]. The approach can remove the bias when the imputation value is so constant as every time in the hot deck. Properties of Hot Deck Estimates the complete sample dataset taken for the hot deck imputation produces the consistent estimates if Data are missing completely at random (MCAR) [11].

K-NN (K-Nearest Neighbors) Imputation

The hot deck single imputation in other form is K-NN (K-nearest neighbors) imputation technique works to fill the imputation value by looking values of other observations from the same dataset. By the mean or mode estimation the value imputes to the missing place [12]. The K-NN ensures the list of advantages which are no explicit missing value imputation performed on assumption whereas approximation imputation possible notably, The K-NN can placed to perform on continues and dataset. Imputation done for multiple missing values, no creation of predicative model for each attribute with missing data.

Multiple Imputation

X_i the attribute to be imputed by predicative distribution produces a pool of multiple values M>2 from that a suitable value will be consider for imputation called Multiple imputation first proposed by Rubin[13]. All the survey says the multivariate of incomplete data belongs to MAR and assumption test on MAR dataset requires relevant information [14]. (Schafer & Olsen 1998)[15] indicated Four different classes of multivariate complete data models for multiple imputation which are 1) Normal model performs multiple imputation by a multivariate normal distribution; 2) Loglinear model, traditional model helps to describe associations among variables in cross-classified data; 3) General location model, which combines a loglinear model for the categorical variables with a multivariate normal regression model for the continuous variables; 4) Two-level linear regression model is commonly implemented to multi-level data. The advantage of the multiple imputation helps easy to estimate variances for sample data while calculating totals and means.

Regression Imputation

Multivariate linear regression or logistic regression imputation technique[16-17] uses the equation Y(Mi)=Xir of the conditional mean E(Yi |Xi) = Xi imputes the regress value from the regression equation to corresponding missing place of the dataset exactly and enforces the relationship for the missing attribute and regress value alone which is not applicable for other imputing value of the complete dataset. Random linear regression imputation technique imputes value for missing Yi, is a simple random sample of size m drawn with replacement from the residuals.

Artificial Neural Networks (Anns)

In survey we found C. G. Wilmot, S. Shivananjappa is base for AANs imputation, Artificial neural networks ANNs[18] is an area uses to impute the missing value into dataset a tool constructed and trained with available data it find the missing place and gives the impute value. An advantage of the technique is to generates the probabilities for each of the possible values to be imputed for the attribute and imputation can be repeated this advantage enables to estimate the variance [19].

Design and Experimentation

Algorithm1: Mean Method by Step Digression (MMSD)

Step1: Get the grouped Attribute value and apply the frequency distribution on the values.

Step2: h is the interval to be calculated form frequency distribution.

Step3: Calculate the $\overline{X}_{impute}$ by implementing the following steps.

Step4: Get the two inputs: group input of AttrXi, FREQi i ranges from 1 to n…..

Step5: Generate Midi from AttrXi by taking average where i ranges from 1 to n…..

Step6: Perform $\sum_{i=1}^{n} FREQ$.

Step7: Calculate the A as Midvalue from the Midi values,

Step8: Calculate the Mi from AttrXi,the difference between vaue1 and value n.

Step9: Generate

$$T_i = \frac{(Mid_i - A)}{M_i}$$

where i ranges from 1 to n…..

Step10: Generate FREQi*Ti $\sum_{i=1}^{n} FREQ * Ti$ where i ranges from 1 to n…..

Step11: Execute

$$\overline{X}_{impute} = \frac{\sum_{i=1}^{n} FREQ * T_i}{\sum_{i=1}^{n} FREQ} \times h$$

Algorithm2: Direct Mean method (DM)

Step1: Get the grouped Attribute value and apply the frequency distribution on the values.

Step2: Calculate the $\overline{X}_{impute}$ by implementing the following steps.

Step3: Get the two inputs: group input of AttrXi, FREQi i ranges from 1 to n….

Step4: Generate Midi from AttrXi by taking average where i ranges from 1 to n….

Step5: Perform $\sum_{i=1}^{n} FREQ$

Step6: Compute FREQi * Midi where i ranges from 1 to n.

Step7: Perform the sum FREQi* Midi.

Step8: Execute

$$\overline{X}_{impute} = \frac{\sum_{i=1}^{n} FREQ * Mid_i}{\sum_{i=1}^{n} FREQ}$$

Algorithm3: Short-Cut Method Mean (SCM)

Step1: Get the grouped Attribute value and apply the frequency distribution on the values.

Step2: Calculate the $\overline{X}_{impute}$ by implementing the following steps.

Step3: Get the two inputs: group input of AttrXi, FREQi i ranges from 1 to n….

Step4: Generate Midi from AttrXi by taking average where i ranges from 1 to n….

Step5: Perform $\sum_{i=1}^{n} FREQ$

Step6: Calculate the A as Midvalue from the Midi values,

Step7: Generate Di = Midi - A

Step8: Generate FREQi * Di and sum it.

Step9: Execute

$$\overline{X}_{impute} = A + \frac{\sum_{i=1}^{n} FREQ * D_i}{\sum_{i=1}^{n} FREQ}$$

The Architecture of imputation process in **Fig. 2** explores how the imputation process flows stage by stage evaluation. Assumes 10 Dataset taken into account which contains full-fledged Data replicated into Training dataset with different percentage of Missing values. Every Data set attribute is grouped based on the similarities of attributes attribute1{X1, X2, X3…, Xi,}, attribute2{X1, X2, X3…, Xi,}, attribute3{X1, X2, X3…, Xi,}, attribute4{X1, X2, X3…, Xi,}. . . attributeN{Xi, Xi, Xi…, Xi,} before imputation algorithm begins the frequency distribution applied on the attribute1,attribute2… attributeN.

Now, the classified Training Dataset attributes 1 to N are the inputs to the imputation algorithms we taken the four imputation algorithm into account, the algorithm estimates $\overline{X}_{impute}$ to fill the missing values to the corresponding attribute. Seeing the literature the simple method mentioned is statistical mean method used for the imputation, taking this as a base we proposed the algorithm called MMSD imputation algorithm in section A is the alternate approach of mean method.

$$\overline{x}_{impute}=\frac{\sum_{i=1}^{n} FREQ*T\,i}{\sum_{i=1}^{n} FREQ}\times h \tag{3}$$

Where A is the Middle value of the attribute can be calculated from the attribute classification, h is the interval calculated for the frequency distribution.

$$\overline{x}_{impute}=\frac{\sum_{i=1}^{n} FREQ*Mid\,i}{\sum_{i=1}^{n} FREQ} \tag{4}$$

Equation (4) mentioned in the Algorithm2 in section B implements Direct Mean method for the attribute.

$$\overline{x}_{impute}=A+\frac{\sum_{i=1}^{n} FREQ*\boldsymbol{D}\,i}{\sum_{i=1}^{n} FREQ} \tag{5}$$

Equation (5) described in Algorithm3 in section C implements SCM method where A is the Middle value of the attribute can be calculated from the attribute classification.

Discussion

The main objective of the implementation is to empirically evaluate the effect of missing data imputation on the proposed method. We disseminate the datasets used in experiment follow up with experimental results and process of analysis.

Performance Comparative

All the four algorithm base is to calculate the mean for imputation but algorithms have different strategy produces the mean with accuracy variation. MMSD, DM, SCM algorithms imputation initial work before database classification frequency distribution applied on the dataset to progress the attribute classification where as in fourth technique simple mean calculation with weight to be assumed for imputation. The weighted method adds the weight to mean to manage approximate accuracy and this leads a advantage of the weighted method and variance is not defrosted the accuracy. The proposed method(MMSD) imputation accuracy dependence on the factor of Defining variable A and h. so the A and h is adjusting factor of the MMSD which takes the mean accuracy to better level. The mean value of all the four methods compared in the result.

Datasets

The implementation were performed using 10 Medical datasets taken from the UCIML repository and the KDD repository[20] Each dataset is described by a specifications such as attributes, Records, Training sets dataset characteristics and attribute characteristics Referred in the **Table 1**. the selected datasets include only discrete data (i.e., discrete numerical and categorical data) and cover a full spectrum of values for each of the characteristics. Missing data were introduced randomly, using the MCAR mechanism, into each of the datasets. The missing values were introduced into all attributes in all datasets in the following five units: 29%, 31%, 32%, 44%, and 54%. Each original dataset randomly divided into equal size training and test subsets and five amount of missing values were introduced for the test subsets and imputation algorithm are executed.

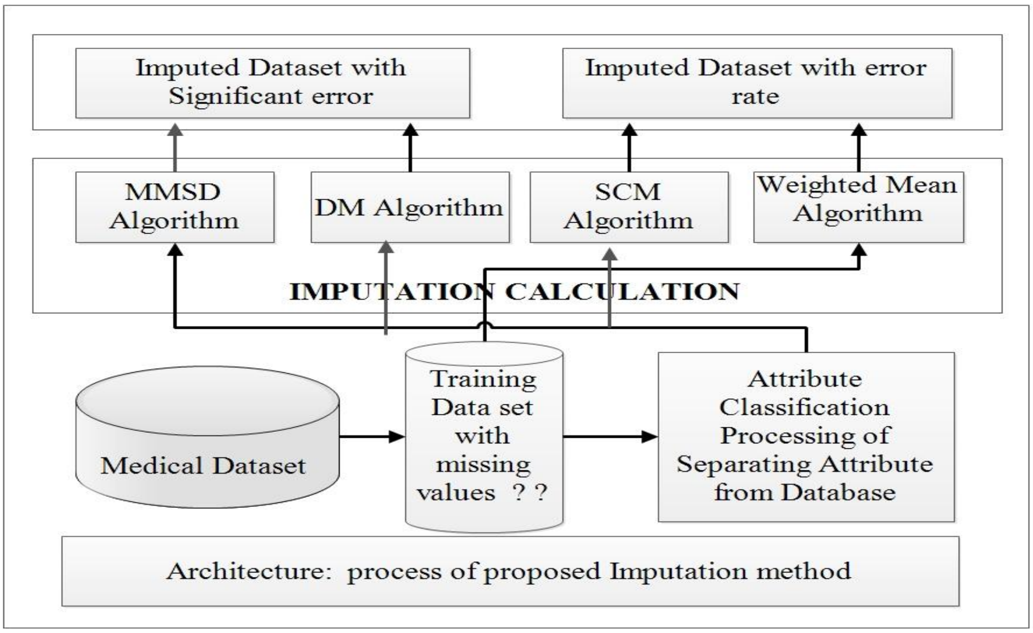
Imputed Dataset with Significant error
Imputed Dataset with error rate
MMSD Algorithm
DM Algorithm
SCM Algorithm
Weighted Mean Algorithm
IMPUTATION CALCULATION
Medical Dataset
Training Data set with missing values ? ?
Attribute Classification Processing of Separating Attribute from Database
Architecture: process of proposed Imputation method

Figure 2: Architecture of Imputation process

Table 1: Ccharacteristic of Data Set Utilized In Imputation Process

Data set	Attribute	Records	Training Set	Data Set Characteristic	Attribute Characteristic	Missing ratio overall
Pima	8	786	16	Multivariate	Integer Real	4%
Heart	13	303	65	Multivariate	Integer Real	68%
Thyroid	5	215	15	Multivariate, Domain	Real	34%

				Theory		
Liver	6	345	12	Multivariate	Categorical Integer Real	51%
Intubation	17	302	34	Multivariate	Categorical Integer Real	8%
Diabeties	20	345	12	Multivariate, Time-Series	Categorical Integer	33%
Wbc	30	569	60	Multivariate	Categorical Integer Real	66%
Lung Cancer	56	32	56	Multivariate, Time-Series	Categorical	18%
Breast Cancer	9	286	14	Multivariate	Categorical	45%
Spect	22	267	45	Multivariate	Categorical	49%

Results and Analysis

Table 2, Table 3, Table 4 and **Table 5** explicit the mean rate at different level of missing rate. **Table 6** refers the Standard Deviation of Imputation methods for the different missing values like 29%, 31%, 32%, 44% and 54%. Among four Techniques the mean values almost equal and slight variation it shows that MMSD technique imputation is nearly better compared with W-Mean method. The 29% missing value of MMSD 0.500723 produces viable accuracy comparatively than W-Mean for 29 % Missing values 0.420723 which is shown in Bold. In the same way in the column 31% of MMSD with W-Mean Second time the accuracy level differs slightly and it continues for 54% of MMSD with W-Mean. Almost considering of proposed method with W-Mean, 00.3% only accuracy level improvement we got it and proposed method with DM, SCM the imputation results are equally same. Taking the trail and error for missing percentage of 57%, 60 % the imputation can be done whereas the accuracy level cannot shown the difference.

Fig. 3 graph represents Missing rate at 29% implicates four imputation method mean values, In that way the SCM imputation method meets the proposition level of significance for T-distribution. At this 29% the SCM imputation method leads imputation smoothly comparing to other methods including proposed method. Enduringly **Fig. 4** Scenario represents missing rate at 31% explores clearly DM imputation, SCM imputation and W-mean imputation meets the level of significance for T-test and proves that MMSD method is comparatively less productivity than other three methods at 31% missing values.

Similarly **Fig. 5** and **Fig. 6** replicate the same point but with alternate methods. For example In Fig. 5 graph sample missing rate at 44% projects strongly DM and MMSD imputation method meets the level of significance for T-test and exhibits two methods increases the imputation performance relatively and DM and W-mean method not attained to meet the level of the significance.

Table 2: Mean value for the missing value at 29%,31%,32%,44%,54%

Method 1: DM Imputation algorithm				
Missing values				
29%	**31%**	**32%**	**44%**	**54%**
0.48	0.78	0.871875	0.609091	0.762963
0.43	0.90	0.984375	0.609091	0.4
0.48	0.78	0.981259	0.525	0.462932
0.62	0.78	0.871875	0.609091	0.881481
0.36	0.54	0.673875	0.934091	1.081481
0.48	1.153704	0.622274	0.834091	1.153704
0.50	0.483821	0.572864	0.65	0.522396
0.51	0.244537	0.521346	0.565909	1.031481
0.36	0.783871	0.871875	0.609091	0.962963
0.93	0.526633	0.871875	0.559091	0.957407

Table 3: Mean Value for the Missing Value At 29%,31%,32%,44%,54%

Method 2: W-Mean Imputation algorithm				
Missing values				
29%	**31%**	**32%**	**44%**	**54%**
0.386206	**0.783871**	0.871875	0.609091	0.762963
0.231034	**0.906452**	0.984375	0.609091	0.4
0.286211	0.783872	0.981259	0.525	0.462932
0.227581	0.783842	0.871875	**0.609091**	0.881481
0.263962	0.641233	0.673875	**0.934091**	1.081481
0.486206	1.153704	0.622274	**0.834091**	1.153704
0.506896	**0.483821**	0.572864	**0.65**	0.522396
0.517241	0.244537	0.521346	**0.565909**	1.031481
0.363962	**0.783871**	0.871875	**0.609091**	0.962963
0.937931	0.526633	0.871875	**0.559091**	0.957407

Table 4: Mean Value for the Missing Value At 29%,31%,32%,44%,54%

Method 3: SCM Imputation algorithm				
Missing values				
29%	**31%**	**32%**	**44%**	**54%**
0.486206	0.78	0.871875	0.609091	0.762963
0.431034	0.906452	0.984375	0.609091	0.4
0.486211	0.783872	0.981259	0.525	0.462932
0.627581	0.783842	0.871875	0.609091	0.881481
0.363962	0.541935	0.673875	0.934091	1.081481
0.486206	1.153704	0.622274	0.834091	1.153704
0.506896	0.483821	0.572864	0.65	0.522396

0.517241	0.244537	0.521346	0.565909	1.031481
0.363962	0.783871	0.871875	0.609091	0.962963
0.937931	0.526633	0.871875	0.559091	0.957407

Table 5: Mean Value for the Missing Value At 29%,31%,32%,44%,54%

Method 4: MMSD Imputation algorithm				
Missing values				
29%	31%	32%	44%	54%
0.386206	0.883871	0.871875	0.609091	0.762963
0.331034	1.206452	0.984375	0.929191	0.4
0.486211	0.687862	0.981259	0.525	0.462932
0.627581	0.783842	0.871875	0.609091	0.881481
0.363962	0.541935	0.673875	0.934091	1.081481
0.486206	1.153704	0.622274	0.834091	1.153704
0.506896	0.522313	0.572864	0.65	0.522396
0.517241	0.244537	0.521346	0.565909	1.031481
0.363962	0.883871	0.871875	0.609091	0.962963
0.937931	0.526633	0.871875	0.559091	0.957407

Table 6: Mean Value for the Missing Value At 29%,31%,32%,44%,54%

Imputation Method	**29%**	**31%**	**32%**	**44%**	**54%**
	S.D	**S.D**	**S.D**	**S.D**	**S.D**
DM	0.16	0.25	0.17	0.13	0.27
W-Mean	0.21	0.25	0.17	0.13	0.27
SCM	0.16	0.25	0.17	0.13	0.27
MMSD (Proposed)	0.17	0.30	0.17	0.13	0.28

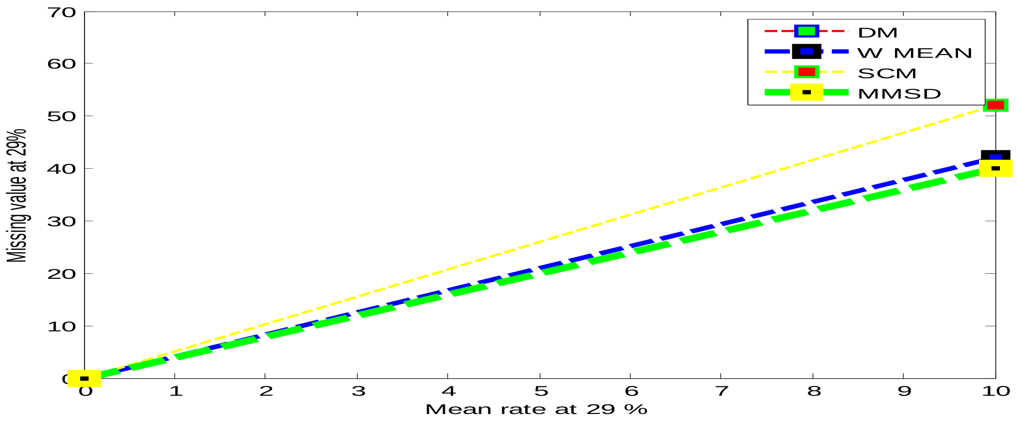

Figure 3: Mean rate at 29% of proposed method for 9 attribute of Example data set

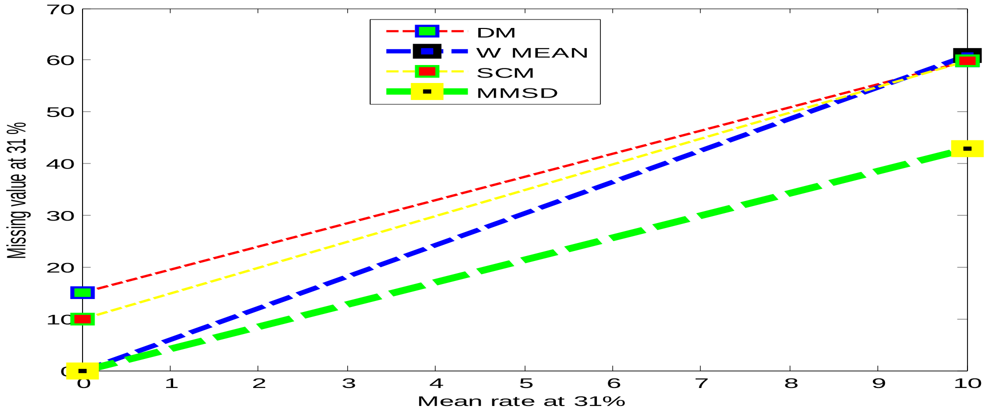

Figure 4: Mean rate at 31% of proposed method for 9 attribute of Example data set

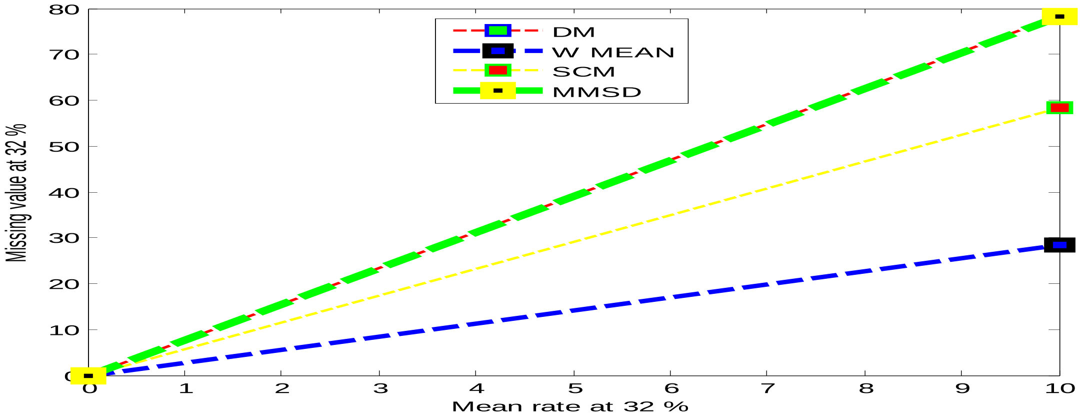

Figure 5: Mean rate at 32% of proposed method for 8 attribute of Example dataset

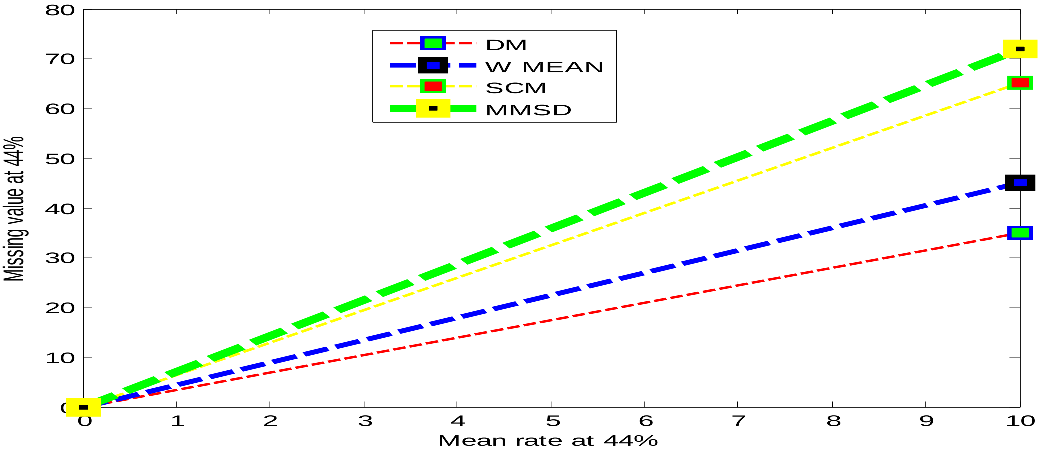

Figure 6: Mean rate at 44% of proposed method for 8 attribute of Example data set

Considering the Fig. 6 at the missing rate of 54% empirical evidence MMSD imputation alone meets the level of significances for T-test to execute the hypothesis condition.

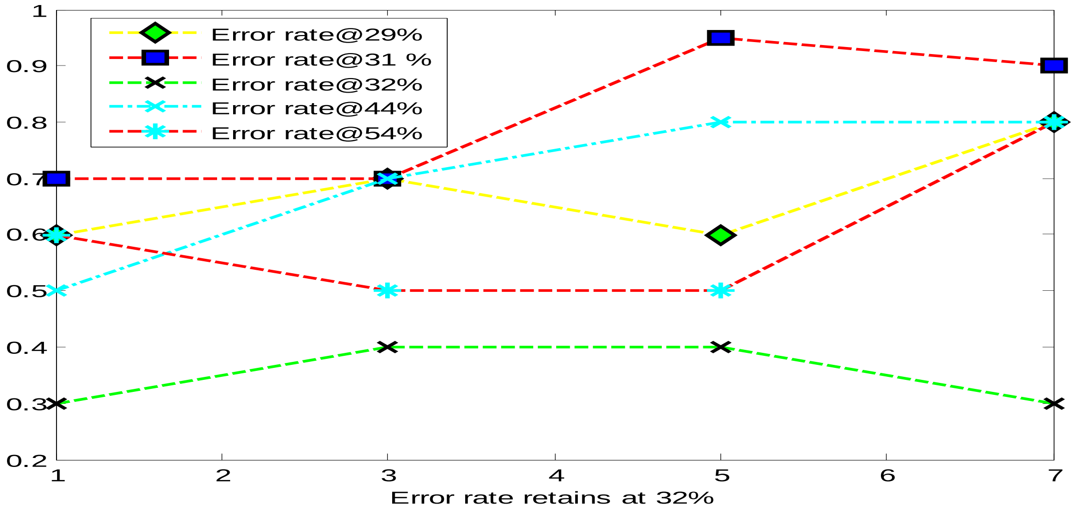

Figure 7: Error rate estimation chart for 32%

Error rate estimation implemented on the Diabetes dataset with 20 attributes of 345 records includes 12 training dataset. The **Fig. 7** graph chart produced the result for MMSD technique at 29%, 31%, 32%, 44% and 54%. The empirical evidence at all level of missing values the imputation error rate estimation lies in between 1 to 10 percentage in that ratio at 32 % of missingness our proposed technique sustains the error rate at 3% viable 4% and again drops to 3% to maintain the accuracy level of imputation.

Apparently the error estimation rate for 29% missing attains at 6% viable 5% and surges to 7% increase the error rate and not guarantee to sustain the error rate. Same as like the error rate estimation of 31%, 44% and 54% takes increase the error rate and it is must be tuned to improve the error accuracy.

Conclusion

In this scenario mining is widely explored were we proposed a statistical model for the value to impute. We proposed a method MMSD algorithm for imputation implemented to impute the missing value into Medical databases is alternate method of Mean and median methods. The Proposed method obtained results compared with other three methods mentioned in this paper, the results outcome determines the imputation accuracy level of proposed method is 0.03% better than other methods and we have implemented the root mean square error rate to check the accuracy level and the error rate of MMSD method show in the relevant figure is produces the improved results than comparison methods ten datasets.

In Future work the same proposed algorithm will be compared with least square algorithm to see the performance of the imputation, next the classifiers will be posed on the proposed algorithm to improve the accuracy level and the error rate will be tested with significant level.

References

[1] P. Davies, P. Smith, “Model Quality Reports in Business Statistics”, ONS, UK, pp. 619-622, 1999.

[2] K. Lakshminarayn, S. A. Harp, T. Samad, R. P. Goldman, “Imputation of missing data in industrial databases, Appl”, Intell. 11, pp. 259-275, 1999.

[3] H. L. Oh and F. J. Scheuren, “Weighting adjustments for unit nonresponse, in Incomplete Data in Sample Surveys”, Theory and Bibliographies, edited by W.G. Madow, I. Olkin, and D. B. Rubin, Volume 2, Academic Press: New York, pp. 143–183, 1983.

[4] R. J. A. Little, D. B. Rubin, “Statistical Analysis with Missing Data”, John Wiley and Sons, 1987.

[5] R. J. A. Little, “Regression with missing x’s: a review”, Journal of The American Statistical Association 87 (420), pp.1227–1237, 1992.

[6] K. Lakshminarayn, S. A. Harp, T. Samad, R. P. Goldman, “Imputation of missing data using machine learning techniques”, in: E.Simoudis, J. Han,U.Fayyad(Eds.)., Second Internationals conference on Knowledge Discovery and Data Mining, Oregon, PP.140-145, 1996.

[7] Qinbao Song, Martin Shepperd, Xiangru Chen a, Jun Liu, “Can k-NN imputation improve the performance of C4.5 with small software project data sets? A comparative evaluation”, The Journal of Systems and Software, PP. 2361–2370, 2008.

[8] B. G. Cox “The weighted sequential hot deck imputation procedure. ASA Proc Section on Survey Res Methods”, pp. 721–726, 1980.

[9] B. G. Cox, R. E. Folsom. “An evaluation of weighted hot deck imputation for unreported health care visits. ASA Proc Section on Survey Res Methods”, pp.412–417, 1981.

[10] R. Platek, G. B. Gray, ”Imputation methodology: Total survey error in Incomplete Data in Sample Surveys.”, Theory and Bibliography(eds.W. G. Madow, I. Olkin, D. B. Rubin), San Diego: New York: Academic Press, Vol. 2. pp. 249– 333, 1983.

[11] R. J. A. Little, D. B. Rubin, “Statistical Analysis with Missing Data”, second ed., Wiley, NJ, USA, 2002.

[12] J. M. Robins, “Non-response models for the analysis of non-monotone nonignorable missing data”, Statistics in Medicine 16, 21–38. 1997.

[13] D. B. Rubin, “Multiple imputation after 18_ years”, Journal American Statistical Association vol .91, pp.473–89, 1996.

[14] J. L Schafer, M. K. Olsen, “Multiple imputation for multivariate missing-data problems:a data analyst’s perspective”, Multivariate Behavioral Research, vol. 33, pp. 545–571, 1998.

[15] I. H. Witten and E. Frank Data Mining: “Practical Machine Learning Tools and Techniques with JAVA Implementations”, Morgan Kaufmann, San Francisco, CA, 2000.

[16] S. Laaksonen “Regression-based nearest neighbor hot decking”. Computational Statistics, 2000.

[17] R.J.A. Little, “A test of missing completely at random for multivariate data with missing values”, Journal of the American Statistical Association, Vol. 83, pp.1198-1202, 1988.

[18] C. G. Wilmot, S. Shivananjappa, “Comparison of hot-deck and neural-network imputation”. Invited paper, International Conference on Transport Survey Quality and Innovation, Kruger National Park, South Africa, August 2001.

[19] B. S. Larsen, B. Madesn, “Error identification and imputations with neural networks. Contributed working Paper” No. 26,UN/ ECE Work Session on Statistical Data Editing, 1999.

[20] S. Hettich, S. D. Bay, The UCI KDD Archive, Department of Information and Computer Science, University of California, Irvine, CA, 1999. (http://kdd.ics.uci.edu).