Missing Value Imputation using Refined Mean Substitution

R.S. Somasundaram[1] and R. Nedunchezhian[2]
[1]Research Scholar, Research and Development Centre, Bharathiar University,
Coimbatore, Tamilnadu,Zip–641046, India

[2]Department of Information Technology, Sri Ramakrishna Engineering College,
Coimbatore, TamilNadu, Zip-641022, India

Abstract

In a previous work, it was clearly shown that the performance of the very simple imputation method based on "Most Common Attribute Value" called MC gave performance better than that of several complex imputation algorithms. And in that work [1] it was shown that the performance of MC was almost equal to that of best performing imputation method called "Event Covering" (EC). So in this work, It is tried to improve the performance of the simple imputation method MC and proposed a new algorithm.

The performance of the proposed algorithm has been compared with the other simple and efficient imputation methods. The performance has been measured with respect to different rate or different percentage of missing values in the data set. To evaluate the performance, the standard WDBC data set has been used. The proposed algorithm performed very well and the arrived results were more significant and comparable.

Keywords:*Datamining, Preprocessing, Missing Value.*

Introduction

One of the most significant processes in data preprocessing phase is that finding missing attribute values and it is a very important issue in data mining. Missing attribute value is more common in several real-world data sets. They possibly will come from the data collecting process or repeated diagnoses tests, any transformation in the experimental set up, indefinite data and so on. Removal of all data containing the missing attribute values cannot completely maintain the characteristics of the real data. Understanding and handling of original circumstance and background knowledge to allocate the missing values seem to be a most favorable approach for handling missing attribute values. But in actual fact, it is extremely complicated to know the unique meaning for the missing data or attributes.

Several approaches have been in practice to handle the missing information in an uncomplicated manner, for instance, substituting missing values with the global or class-conditional mean/mode. On the other hand, several real world data include missing attribute values, making it hard to produce constructive knowledge from training data and to provide precise result. As a result, many strategies to deal with incomplete data have been developed.

In general, there are three categories of missing data, They are Missing Completely At Random (MCAR), Missing At Random (MAR) and Not Missing At Randam(NMAR).

Missing value imputation can happen in data sets in several forms. In practice, missing values in datasets are classified into three classes.

- Missing values occur in several attributes (columns),
- Missing values occur in a number of instances (rows),
- Missing values occur randomly in attributes and instances.

Also technique of handling missing values is classified into two groups of methods.

- Pre-replacing methods: This method replace the missing values before the data mining process, and
- Embedded methods: This method deal with missing values during the data mining process.

The occurrence cases of missing values can have an effect on the result of missing value methods, as a result the selection of appropriate missing value methods in each case is more important.

The Normal and Proposed Imputation Methods

In this section, a standard mean based imputation technique as well as out proposed imputation techniques are addressed.

Let us assume D as a dataset of m records in which, each record contains n attributes. So, there will be m x n values in that dataset D. If the dataset D contains some missing attribute values, then, in side that dataset, it may be represented by a non numeric string. (in matlab the missing values as NaN can be represented – not a number)

2.1. Replacing Missing Values with a Constant Numeric Value

Numeric computation on a dataset is not possible if it is containing non numeric attribute values like "unknown", "N/A" or minus infinity along with other numeric data. So

before taking the data in to calculations or computation process, all the instances of such non numeric missing value attributes can be replaced with a constant numeric value such a 0 or 1 or any vale depending upon the magnitudes of the individual attributes.

After this process, the data set can be used for any numeric calculation or data mining process.

Pseudo code of Method 1

For r=1 to N
For c = 1 to M
If D(r,c) is not a Number (is a missing value), then
Substitute zero to D(r,c)

2.2. Filling Missing Values with Random Attribute Values

Pseudo code of Method 2

For c = 1 to M
Find mean value "Am" of all the attributes of the column 'c'
Min(c) = (min of all the values of column c)
Max(c) = (max of all the values of column c)

For r=1 to N
For c = 1 to M
If D(N,M) is not a Number (missing value), then Substitute a random value between Min(c) and Max(c) to D(N,M)

2.3. Replacing Missing Values with Attribute Mean

The following pseudo code explains the very commonly used mean substitution method which is also commonly known as "Most Common Attribute Value" substitution method(MC)

Let
$D = \{ A_1, A_2, A_3, \ldots\ldots A_n \}$

Where
D is the set of data with missing values
A_i – is the i^{th} attribute column of values of D with missing values in some or all columns
n - is the number of attributes.

Function MC
Begin
For i=1:n
$a_i \leftarrow A_i \cap m_i$
where
a_i is the column of attributes without missing values
m_i is the set of missing values in A_i (missing values denoted by a symbol)
Let μ_i be the mean of a_i
Replace all the missing elements of A_i with μ_i
end
Finally the imputed data set will be generated.
End

2.4. The Proposed Refined Mean Substitution Method (RMS Method)

This algorithm also starts with mean value substitution (or constant/random value substitution). But, by assuming that the initially imputed values are not accurate, the algorithm, again re-estimates the new values based on the Euclidean distance of the missing value records and the remaining records. For mean value calculations, the records with minimum Euclidean distance with the missing value record were not taken in to account.

Function RMS
Begin
For i=1:n
$a_i \leftarrow A_i \cap m_i$
where
a_i is the column of attributes without missing values
m_i is the set of missing values in A_i (missing values denoted by a symbol)
Let μ_i be the mean of a_i
Replace all the missing elements of A_i with μ_i
end
Let
$D_{new} = \{ R_1, R_2, R_3, \ldots, R_m \}$
Where
D_{new} be the approximately imputed data set of D
$R_1, R_2, R_3, \ldots, R_m$ are the m rows of the data set.
For j=1:m
$d \leftarrow dist (D_{new} , R_j)$
$I \leftarrow find(D > mean (d))$
Where
d is the distance matrix
I is the index of elements which are having distance higher than mean(d).
For k=1:n
If $D_{new}(m,n)$ is originally a missing element
begin
Let μ_j be the mean of elements $D_{new}(I, n)$
$R_j(k) \leftarrow \mu_j$
end
end
end
Finally the imputed data set will be generated.
end

FC Mean Clustering

To evaluate the quality of imputation, the imputed data is clustered with fuzzy C means clustering algorithm and the classification the performance of classification is measured with different quality metrics. FC-means was selected to evaluate the imputation performance because,

in our previous work[2] it was observed that FC-means provided better performance.

Fuzzy c-means (FCM) is a data clustering technique wherein each data point belongs to a cluster to some degree that is specified by a membership grade. This technique was originally introduced by Jim Bezdek in 1981 as an improvement on earlier clustering methods.

It provides a method that shows how to group data points that populate some multidimensional space into a specific number of different clusters. The Fuzzy c-means algorithm starts with an initial guess for the cluster centers, which are intended to mark the mean location of each cluster. The initial guess for these cluster centers is most likely incorrect. Additionally, Fuzzy c-means algorithm assigns every data point a membership grade for each cluster. By iteratively updating the cluster centers and the membership grades for each data point, Fuzzy c-means algorithm iteratively moves the cluster centers to the right location within a data set. This iteration is based on minimizing an objective function that represents the distance from any given data point to a cluster center weighted by that data point's membership grade.

The fuzzy c-means (FCM) algorithm was introduced by J. C. Bezdek []. The idea of FCM is using the weights that minimize the total weighted mean-square error:

$$J(w_{qk}, \mathbf{z}^{(k)}) = \Sigma_{(k=1,K)} \Sigma_{(k=1,K)} (w_{qk}) \| \mathbf{x}^{(q)} - \mathbf{z}^{(k)} \|^2$$

……(1)

$$\Sigma_{(k=1,K)} (w_{qk}) = 1 \text{ for each } q$$

$$w_{qk} = (1/(D_{qk})^2)^{1/(p-1)} / \Sigma_{(k=1,K)} (1/(D_{qk})^2)^{1/(p-1)} , \; p > 1$$

……(2)

The FCM allows each feature vector to belong to every cluster with a fuzzy truth value (between 0 and 1), which is computed using Equation (2). The algorithm assigns a feature vector to a cluster according to the maximum weight of the feature vector over all clusters.

Implementation and Evaluation

To evaluate the algorithms, a suitable and standard data set is needed. It is decided to use Wisconsin Diagnostic Breast Cancer (WDBC) dataset for our experiments. The original dataset was provided by Dr. William H. Wolberg, W. Nick Street and Olvi L. Mangasarian of University of Wisconsin. this data set was selected because of the following reasons,

1. it is having no missing values so that missing values can be simulated and have the control over the evaluation process.
2. All the records are having corresponding clean class label.
3. It is having sufficiently large number of attributes and records.
4. Since the attributes (except the ID and class attribute) are real values features, it is well suited for this evaluation process.

Description of the Dataset:
Number of instances: 569
Number of attributes: 32
(ID, diagnosis and 30 real-valued input features)
Missing attribute values: none
Class distribution: 357 benign, 212 malignant

The ID is a number to denote the patient/record and the Diagnosis may be M (malignant) or B (benign). All the other features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

According to the original descriptions, the ten real-valued features are computed for each cell nucleus. They are :

(1) radius (mean of distances from center to points on the perimeter) (2) texture (standard deviation of gray-scale values) (3) perimeter, (4) area, (5) smoothness (local variation in radius lengths) (6) compactness (perimeter^2 / area - 1.0), (7) concavity (severity of concave portions of the contour), (8) concave points (number of concave portions of the contour), (9), symmetry and (10) fractal dimension ("coastline approximation" - 1)

The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features in total. For example, field 3 is Mean Radius, field 13 is Radius SE, and field 23 is Worst Radius.

In the following table the results arrived on a Windows XP laptop equipped with Intel core 2 duo CPU at 2GHz and 2GB RAM is presented. The Matlab implementation of the algorithms was used for evaluation.

This dataset is selected for evaluating the three missing data imputation methods because; it has original classification labels along with the records. So our results with original classification can be compared. Further, this data set is not having any missing values. So missing values can be simulated and then do missing values imputation and then compare the accuracy of clustering with recreated missing data.

Missing attribute values in the original data set is none. But missing values were synthetically introduced in arbitrary locations. The percentage of Missing Value Attributes each case clustering was made three times and the average value is calculated.

The following figure shows the performance of the imputation algorithms with respect to different metrics. To measure this performance, the original class labels of WDBC data set is compared with the calculated class labels of the imputed data using different performance measures.

In the following tables, the performance if imputation with reconstructed WDBC data is indirectly measured using the classification performance measures. The better classification performance (high Rand Index) signifies the better imputation of missing values.

TABLE 1 PERFORMANCE IN TERMS OF RAND INDEX

% of Missing Values	Clustering Accuracy in Terms of (Average of five runs)			
	Constant Value Substn.	Random Value Substn.	MC/ Mean Value Substn.	Proposed RMS Method
10	0.836518	0.839414	0.848177	0.842323
20	0.783844	0.799995	0.851123	0.854081
30	0.781195	0.851123	0.845244	0.866036
40	0.511820	0.658531	0.839414	0.854081
50	0.500248	0.631934	0.827904	0.825058
Avg	0.682725	0.756199	0.842372	0.848316

The following chart shows the performance of the algorithms (in terms of Rand Index) with respect to different percentage of missing values. The proposed RMS imputation algorithm performed little bit better than the mean value substitution method and better than all other methods.

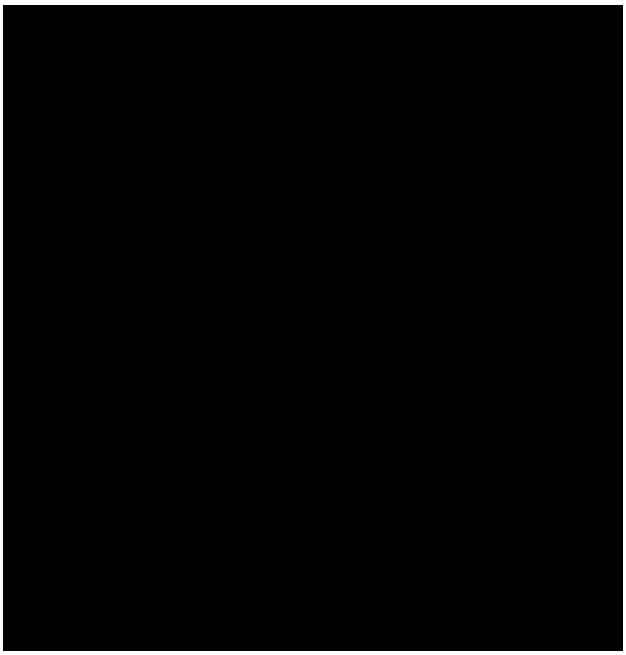

Figure 1 : Percentage of Missing Values vs. Rand Index

The following bar chart shows the average performance in terms of Rand Index. It is obvious that all the three proposed algorithms performed better than the standard MC/mean value substitution algorithm and better than all other methods.

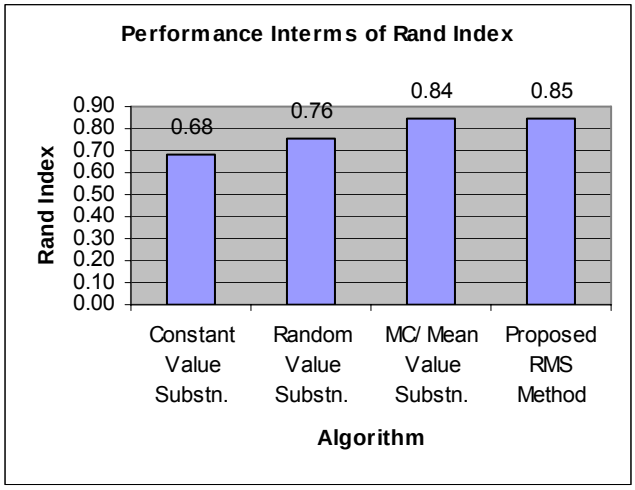

Figure 2 : Average Performance in terms of Rand Index

In the following tables, the performance in terms of accuracy measure. The better classification performance (high accuracy) signifies the better imputation of missing values.

TABLE 2 PERFORMANCE IN TERMS OF ACCURACY

% of Missing Values	Clustering Accuracy in Terms of (Average of five runs)			
	Constant Value Substn.	Random Value Substn.	MC/ Mean Value Substn.	Proposed RMS Method
10	91.04	91.21	91.74	91.39
20	87.70	88.75	91.92	92.09
30	87.52	91.92	91.56	92.79
40	57.96	78.21	91.21	92.09
50	52.37	75.75	90.51	90.33
Avg	75.318	85.168	91.388	91.738

The following chart shows the performance of the algorithms (in terms of Accuracy) with respect to different percentage of missing values. The proposed RMS imputation algorithm performed little bit better than the mean value substitution method and significantly better than all other methods. In fact, Rand Index and Accuracy are the same kind of metrics. So that, It is having almost similar shape of graphs in both cases (but the x-scale values are entirely different)

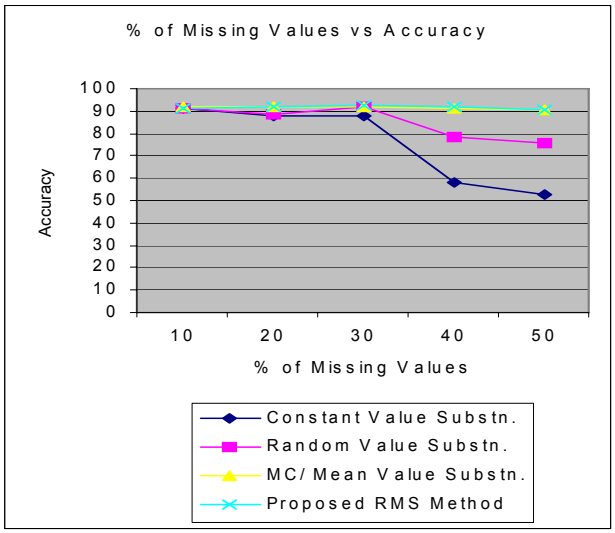

Figure 3 : Percentage of Missing Values vs. Accuracy

The following bar chart shows the average performance in terms of Accuracy. The proposed RMS imputation algorithm performed little bit better than the mean value substitution method and significantly better than all other methods.

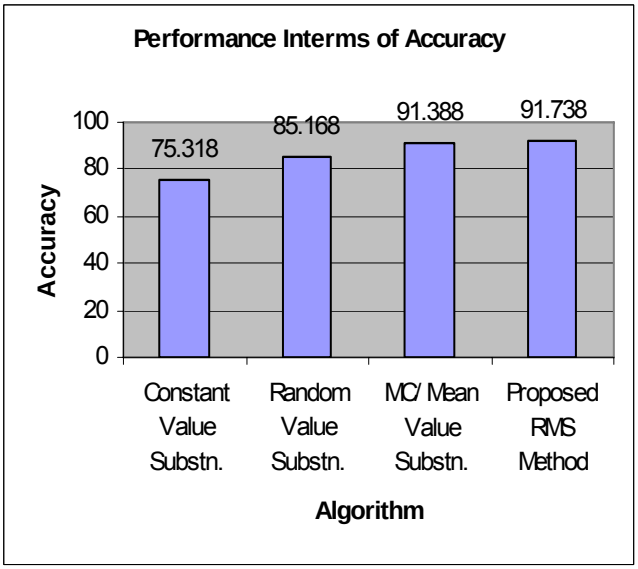

Figure 4 : Average Performance in terms of Accuracy

In the following tables, the performance in terms of accuracy Specificity. The better classification performance (high Specificity) signifies the better imputation of missing values.

TABLE 3 PERFORMANCE IN TERMS OF SPECIFICITY

% of Missing Values	Clustering Accuracy in Terms of Specificity(Average of five runs)			
	Constant Value Substn.	Random Value Substn.	MC/ Mean Value Substn.	Proposed RMS Method
10	80.66	81.60	83.49	82.55
20	79.72	79.72	82.55	82.55
30	86.79	86.32	81.60	84.43
40	56.60	70.28	81.13	87.26
50	49.53	75.47	79.72	91.98
Avg	70.66	78.678	81.698	85.754

The following chart shows the performance of the algorithms in terms of Specificity with respect to different percentage of missing values. It is obvious that the RMS imputation algorithm outperforms all other algorithms. Even all the three proposed algorithms performed better than the standard MC/mean value substitution algorithm.

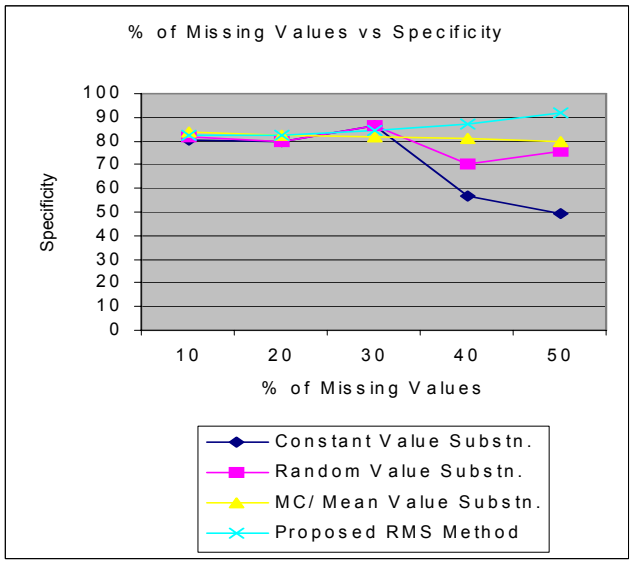

Figure 5 : Percentage of Missing Values vs. Specificity

The following bar chart shows the average performance in terms of Accuracy. The proposed RMS imputation algorithm performed little bit better than the mean value substitution method and significantly better than all other methods.

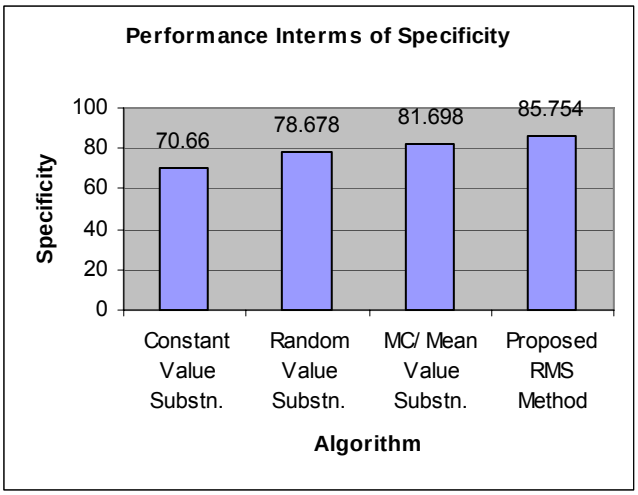

Figure 6 : Average Performance in terms of Specificity

In the following tables, the performance in terms of accuracy Sensitivity. The better classification performance (high Sensitivity) signifies the better imputation of missing values. In terms of sensitivity the algorithms MC, EMI-RBF and IRMS were almost provided equal performance.

TABLE 4 PERFORMANCE IN TERMS OF SENSITIVITY

% of Missing Values	Clustering Accuracy in Terms of Sensitivity (Average of five runs)			
	Constant Value Substn.	Random Value Substn.	MC/ Mean Value Substn.	Proposed RMS Method
10	97.20	96.92	96.64	96.64
20	92.44	94.12	97.48	97.76
30	87.96	95.24	97.48	97.76
40	58.77	82.91	97.20	94.96
50	54.06	75.91	96.92	89.36
Avg	78.086	89.02	97.144	95.296

The following chart shows the performance of the algorithms in terms of Specificity with respect to different percentage of missing values. In terms of sensitivity the performance of the proposed algorithm is little bit lower than the mean value substitution method.

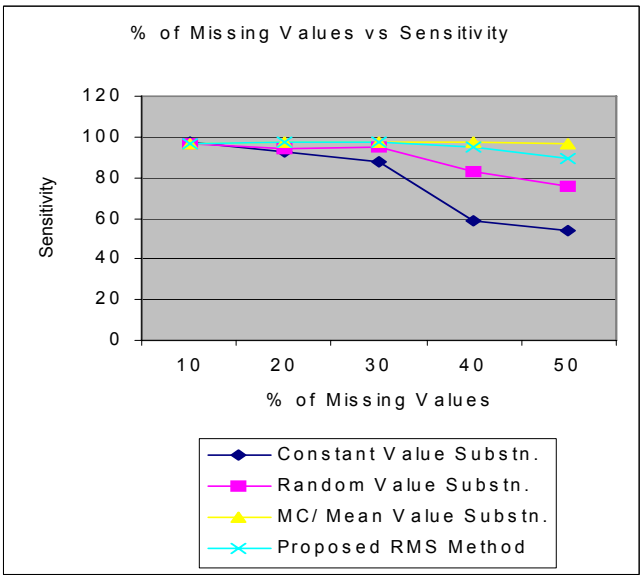

Figure 7 : Percentage of Missing Values vs. Sensitivity

The following bar chart shows the average performance in terms of sensitivity. with this metric, the performance of the standard MC/mean value substitution is little bit better than the proposed method.

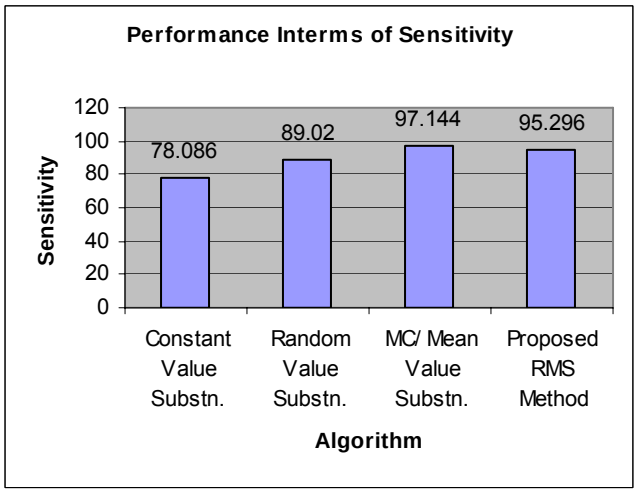

Figure 8 : Average Performance in terms of Sensitivity

In the following tables, the performance in terms of accuracy MSE. Generally, the lower MSE signifies the better imputation of missing values.

TABLE 5 PERFORMANCE IN TERMS OF MSE

% of Missing Values	Clustering Accuracy in Terms of Mean Square Error (Average of five runs)			
	Constant Value Substn.	Random Value Substn.	MC/ Mean Value Substn.	Proposed RMS Method
10	0.138478	0.147872	0.151952	0.155589
20	0.121282	0.137835	0.149573	0.155588
30	0.106783	0.130256	0.148509	0.156369
40	0.091499	0.124055	0.147007	0.155943
50	0.076607	0.117663	0.143774	0.155865
Avg	0.1069298	0.1315362	0.148163	0.1558708

The following chart shows the performance of the algorithms in terms of Mean Square Error (MSE) with respect to different percentage of missing values. It is obvious that the proposed RMS imputation algorithm outperforms all other algorithms. The following line char and bar chats shows the difference in performance.

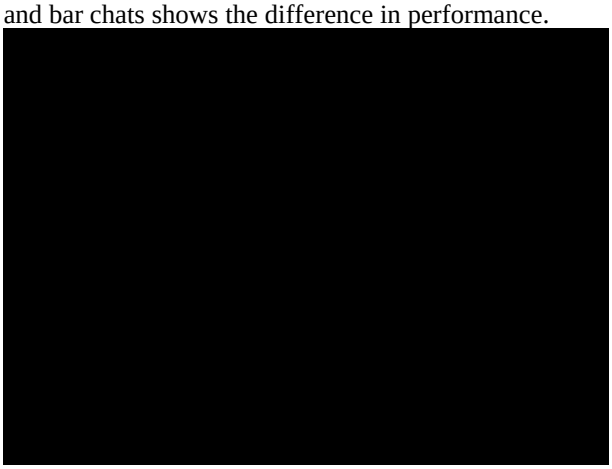

Figure 9 : Percentage of Missing Values vs. MSE

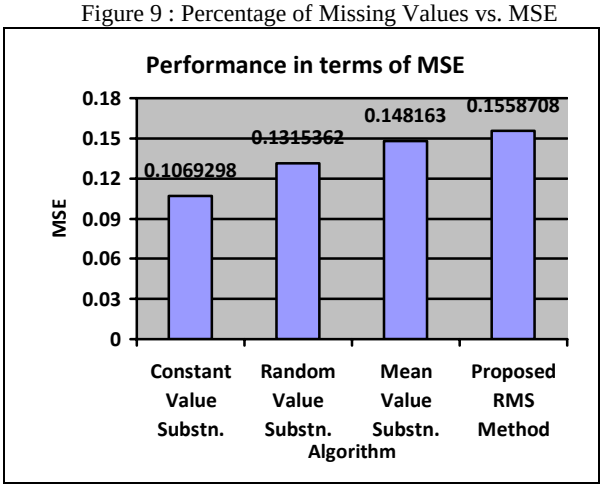

Figure 10 : Average Performance in terms of MSE

Conclusion and Scope for Further Enhancements

In this paper, the proposed RMS imputation methods has been implemented and evaluated. The performance of the missing value imputation algorithms were measured with respect to different percentage of missing values in the data set. The perforce of reconstruction was compared with the original WDBC data set.

In various previous works including [2], it was shown that the performance of "Most Common Attribute Value"(MC) or Mean Value Substitution based method performed better than most of the complex algorithms. But in our case, the proposed algorithms provided better performance than the most popular and standard method.

The performance of the algorithms was evaluated with five different metrics. In almost all the cases or metrics, our proposed algorithms performed better than MC/mean value substitution method.

Acknowledgement

The authors wish to thank the Management of Sri Ramakrishna Engineering College, Coimbatore – 641 022, India for providing resources and support for pursuing research.

References

Julián Luengo, Salvador García, Francisco Herrera, "A study on the use of imputation methods for experimentation with Radial Basis Function Network classifiers handling missing attribute values: The good synergy between RBFNs and Event Covering method "Neural Networks 23 (2010) 406-418, Elsevier.

R.S. Somasundaram, R. Nedunchezhian, "Evaluation on Three Simple Imputation Methods for Enhancing Preprocessing of Data with Missing Values", International Journal of Computer Applications, Vol21-No. 10, May 2011, pp14-19.

Qi Duan, Elsa Angelini, Susan L. Herz, Christopher M. Ingrassia, Olivier Gerard, Kevin D. Costa, Jeffrey W. Holmes, Shunichi Homma, Andrew Laine, Dynamic Cardiac Information From Optical Flow Using Four Dimensional Ultrasound.,

YUE Rentian, SHI Qingyan & LUO Yun, "Forecasting of Aviation Accidents Based on Radial Basis Function Neural Network"

Thomas Lumley, "Missing data", A Lecture Note, BIOST 570, 2005-11-9

Zhang, S.C., et al., (2004). Information Enhancement for Data Mining. IEEE Intelligent Systems, 2004, Vol. 19(2): 12-13.

Qin, Y.S., et al. (2007). Semi-parametric Optimization for Missing Data Imputation. Applied Intelligence, 2007, 27(1): 79-88.

Zhang, C.Q., et al., (2007). An Imputation Method for Missing Values. PAKDD, LNAI, 4426, 2007: 1080-1087.

Quinlan, J.R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, USA, 1993.

Han, J., and Kamber, M., (2006). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2006, 2nd edition.

Ahmed Sobhy Sherif , Hany Harb and Sherif Zaky, "A New Data Imputing Algorithm", International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 2, May 2011, pg No. 133-139.

Chen, J., and Shao, J., (2001). Jackknife variance estimation for nearest-neighbor imputation. J. Amer. Statist. Assoc. 2001, Vol.96: 260-269.

Lall, U., and Sharma, A., (1996). A nearest-neighbor bootstrap for resampling hydrologic time series. Water Resource. Res. 2001, Vol.32: 679-693.

Chen, S.M., and Chen, H.H., (2000). Estimating null values in the distributed relational databases environments. Cybernetics and Systems: An International Journal. 2000, Vol.31: 851-871.

Chen, S.M., and Huang, C.M., (2003). Generating weighted fuzzy rules from relational database systems for estimating null values using genetic algorithms. IEEE Transactions on Fuzzy Systems. 2003, Vol.11: 495-506.

Magnani, M., (2004). Techniques for dealing with missing data in knowledge discovery tasks. Available from http://magnanim.web.cs.unibo.it/data/pdf/missingdata.pdf, Version of June 2004.

Kahl, F., et al., (2001). Minimal Projective Reconstruction Including Missing Data. IEEE Trans. Pattern Anal. Mach. Intell., 2001, Vol. 23(4): 418-424.

Gessert, G., (1991). Handling Missing Data by Using Stored Truth Values. SIGMOD Record, 2001, Vol. 20(3): 30-42.

Pesonen, E., et al., (1998). Treatment of missing data values in a neural network based decision support system for acute abdominal pain. Artificial Intelligence in Medicine, 1998, Vol. 13(3): 139-146.

Ramoni, M. and Sebastiani, P. (2001). Robust Learning with Missing Data. Machine Learning, 2001, Vol. 45(2): 147-170.

Pawlak, M., (1993). Kernel classification rules from missing data. IEEE Transactions on Information Theory, 39(3): 979-988.

Forgy, E., (1965). Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications. Biometrics , 1965, Vol. 21: 768

Blake, C.L and Merz, C.J (1998). UCI Repository of machine learning databases.

Hamerly, H., and Elkan, C., (2003). Learning the k in k-means. Proc. of the 17th intl. Conf. of Neural Information Processing System.

Zhang, S.C., et al., (2006). Optimized Parameters for Missing Data Imputation. PRICAI06, 2006: 1010-1016.

Wang, Q., and Rao, J., (2002a). Empirical likelihood-based inference in linear models with missing data. Scand. J. Statist., 2002, Vol. 29: 563-576.

Wang, Q. and Rao, J. N. K. (2002b). Empirical likelihood-based inference under imputation for missing response data. Ann. Statist., 30: 896-924.

Silverman, B., (1986). Density Estimation for Statistics and Data Analysis. Chapman and Hall, New York.

Friedman, J., et al., (1996). Lazy Decision Trees. Proceedings of the 13th National Conference on Artificial Intelligence, 1996: 717-724.

John, S., and Cristianini, N., (2004). Kernel Methods for Pattern Analysis. Cambridge.

Lakshminarayan, K., et al., (1996).Imputation of Missing Data Using Machine Learning Techniques.KDD-1996:140-1

R.S. SomaSundaram holds B.Sc(Mathematics), M.C.A., I.C.W.A(I), M.Phil. At present he is pursuing his Ph.D under the able guidance of Dr. R. Nedunchezhian. He is serving as an Associate Professor in the department of Computer Applications, Sri Ramakrishna Engineering College, Coimbatore, India. His area of interest includes Data Mining, Data Compression and Finger Print recognition. He is a member of IEEE, IAENG and ISTE. He is having 15 y ears of teaching experience. He has published 7 research publications in National/International Conference/Journal to his credit.

Dr. Nedunchezhian. R. is working as the Professor and Head in the department of Information Technology , Sri Ramakrishna Engineering College, Coimbatore. He has more than 19 y ears of experience in research and teaching. Currently, he is guiding many Ph.D scholars of the Anna University , Coimbatore, and the Bharathiar University. His re search interests are know ledge discovery and data mining, distributed computing, and database security. He has published many research papers in national/international conferences and journals. He has edited a book entitled "Handbook of Research on Soft Computing Applications for Database Tec hnologies: Techniques and Issues" which was published by IGI publications, USA in April, 2010. He is a Life member of Advanced Computing and Communication Society and ISTE.