Journal of Computer Science, 9 (3): 327-334, 2013
ISSN 1549-3636

doi:10.3844/jcssp.2013.327.334 Published Online 9 (3) 2013 (http://www.thescipub.com/jcs.toc)

Radial Basis Function Network Dependent Exclusive Mutual Interpolation for Missing Value Imputation

[1]Somasundaram, R.S. and [2]R. Nedunchezhian

[1]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

Received 2012-09-04, Revised 2013-04-10; Accepted 2013-04-24

ABSTRACT

The success of data mining relies on the purity of the data set. Before performing the data mining, the data has to be cleaned. An unprocessed data set may contain noisy or missing values which is a critical research issue in the pre-processing stage. Imputation methods are being used to solve the missing value problems. In this proposed work, a machine learning based imputation method is proposed by using the mutual information by exclusively interpolating two different section of the same dataset. For designing the proposed model, a radial basis function based neural network has been used. The performance of the proposed algorithm has been measured with respect to different rate or percentage of missing values in the data set and the results has been compared with existing simple and efficient imputation methods also. To evaluate the performance, the standard WDBC data set has been used. The proposed algorithm performs well and was able to impute the missing values even in the worst cases with more than 50% of missing values. Instead of using simple quality measure such as Mean Square Error (MSE) to evaluate the imputed data quality, in this study, the quality is measured in terms of classification performance. The results arrived were more significant and comparable.

Keywords: Datamining, Preprocessing, Imputation Methods

1. INTRODUCTION

Data cleaning processes apply routines that can handle incomplete, noisy and inconsistent data. A missing data is defined as an attribute or feature in a dataset which has no associated data value. Correct treatment of these data is crucial, as they have a negative impact on the interpretation and result of data mining processes. Incomplete data is an unavoidable problem in dealing with most of the real world data sources. The topic has been discussed and analyzed by researchers Zhang *et al.* (2004) and Kotsiantis *et al.* (2006) in the field of machine learning. Generally, there are some important factors to be taken into account when processing unknown feature values. The most important one of them is the source of 'unknowingness'.

Missing values occur when no data value is stored for an attribute or feature in the dataset. Missing values are a common occurrence and it can severely disturb the conclusions drawn from the data if handled inappropriately in empirical research. Missing values may occur because of non-availability of information for several items or whole unit. Non-availability of data is sensitive in various applications like database storing information about private subjects' items such as income. Dropout is a type of missing value that occurs mostly when studying development over time. In this type of study the measurement is repeated after a certain period of time. Missing occurs when participants drop out before the test ends and one or more measurements are missing.

Corresponding Author: Somasundaram, R.S., Research and Development Centre, Bharathiar University, Coimbatore, Tamilnadu, Zip-641046, India

Sometimes missing values are caused by the researchers themselves. For example, as Ader and Mellenberg (2008) stated when data collection is not done properly or when mistakes were made in data entry and as Messner (1992) discussed, a great deal of missing data arise in cross-national research in economics, sociology and political science because governments choose not to, or fail to, report critical statistics for one or more years.

Generally, missing values can occur in datasets in different forms. They can be classified into three categories and a clear knowledge on which category the missing values lies is a clear step towards a positive solution:

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

As methods used for each of these categories differ, therefore selection of correct algorithm is significant. Normally, missing rates less than one per cent are considered trivial, 1-5% are considered to be manageable. But databases with 5-15% missing data values rate needs sophisticated methods to handle them correctly and more than 15% requires careful handling as they affect interpretation. It is in the last category most of the solutions have been proposed and it is understood that several alternative ways of dealing with missing data exists.

Efficient treatment of missing values requires a complete understanding behind it. The following topic outlines some fundamental aspects of incomplete or missing values.

1.1. Types of Missing Values

Little and Rubin (2002) define a list of missing mechanisms, which are widely accepted by the community.

Missing Completely At Random (MCAR)-MCAR is the probability that an observation (X_i) is missing, is unrelated to the value of X_i or to the value of any other variables and the reason for missing is completely random. Typical examples of MCAR are when a tube containing a blood sample of a study subject is broken by accident (such that the blood parameters cannot be measured) or when a questionnaire of a study subject is accidentally lost (Donder *et al*., 2006). This situation is rare in real world and is usually discussed in statistical theory.

Missing at Random (MAR)-MAR is the probability of the observed missing pattern, given the observed and unobserved data, does not depend on the values of the unobserved data. An example of this is accidentally or deliberately skipping an answer on a questionnaire by the participant. This mechanism is common in practice and is generally considered as the default type of missing data.

Not Missing At Random (NMAR)-If the probability that an observation is missing depends on information that is not observed, this type of missing data is called NMAR. For example, high incomers may be more reluctant to provide their income information (Donder *et al*., 2006). This situation is relatively complicated and there is no universal solution.

This study is organized as follows: Section-1 discuss the introduction about missing value, section-2 details about various imputation methods taken for discussion and comparison, section-3 narrates the performance of the proposed imputation method and section-4 concludes.

1.2. Missing Value Imputation Methods under Evaluation

A standard mean based imputation technique is addressed as well as out proposed imputation techniques.

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, inside that dataset, it may be represented by a non numeric string. (In matlab, the missing values can be represented as NaN-not a number)

1.3. 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 = {A1, A2, A3, ….. An}

Where:

D = The set of data with missing values
Ai = The ith attribute column of values of D with missing values in some or all columns
n = The number of attributes.

```
Function MC
Begin
    For i=1: n
        ai ← Ai ∩ mi
Where
ai is the column of attributes without missing values
mi is the set of missing values in Ai (missing values denoted by a symbol)
```

```
            Let $\mu i$ be the mean of ai
            Replace all the missing elements of Ai
            with $\mu i$
        end
        Finally imputed data set is generated.
End
```

1.4. 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
        $ai \leftarrow Ai \cap mi$
Where
ai is the column of attributes without missing values
mi is the set of missing values in Ai (missing values denoted by a symbol)
        Let $\mu i$ be the mean of ai
        Replace all the missing elements of Ai with $\mu i$
    end
    Let
Dnew = {R1, R2, R3,…., Rm}
Where
Dnew be the approximately imputed data set of D
R1, R2, R3,…., Rm are the m rows of the data set.
For
For j=1:m
    $d \leftarrow$ dist ( Dnew , Rj )
    $I \leftarrow$ 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 Dnew(m,n) is originally a missing element
begin
 Let $\mu j$ be the mean of elements Dnew(I, n)
        $Rj(k) \leftarrow \mu j$
        end
end
end
Finally the imputed data set is generated.
end
```

1.5. The Proposed RBFN Dependent EMI Imputation Method

The following algorithm explains the proposed method. In this algorithm. The data set is processed column-wise. The knowledge of one column or one set of columns of data and its relationship with the other column or another set of columns of data will be used to mutually impute one from the other.

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 a novel interpolation technique.

```
Let:
 D = {A1, A2, A3, ….. An}
Where
        D is the set of data with missing values
Ai-is the ith attribute column of values of D with missing values in some or all columns
        n - is the number of attributes.
Function EMI_RBF
Begin
        For i=1:n
            $ai \leftarrow Ai \cap mi$
where
ai is the column of attributes without missing values
 mi is the set of missing values in Ai (missing values denoted by a symbol)
            Let $\mu i$ be the mean of ai
        Replace all the missing elements of Ai with $\mu i$ $\rightarrow$ Dtemp
        end
        Let
Dtemp = {C1, C2, C3,…., Cn}
Where
Dtemp be the approximately imputed data set of D
C1, C2, C3,…., Cn are the n columns of the new data set.
        Separate the data column-wise to for two mutually related data sets Dx and Dy.
Dx = { C1, C2, C3,…., Cn/2}
Dy= { C n/2+1, C n/2+2, C n/2+3,…., Cn}
```

Construct an RBF neural network N1 and train it with Dx for interpolating Dy. Test the network with same Dx to predict the estimated value of Dy namely Dy_new:

$$\text{N1}(\text{Dx}) \rightarrow \text{Dy\_new}$$

Construct another RBF neural network N2 and train it with Dy for interpolating Dx. Test the network with same Dy to predict the estimated value of Dx namely Dx_new:

$$\mathrm{N2}(\mathrm{Dx}) \rightarrow \mathrm{Dy\_new}$$

```
Combine the two interpolated Data sets and form Dnew
Dnew = {Dx_new , Dy_new}
Let Dfinal= D
For j=1:m
        For k=1:n
                If Dfinal(m,n) is originally a missing element
Begin
Dfinal(m,n) → Dnew(m,n)
                end
end
end
Finally the imputed data set Dfinal is generated.
end
```

1.6. 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. As per literature survey, FC means provides better performance. So, FC-means is used to evaluate the imputation 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 Bezdek (1981) as an improvement on earlier clustering methods Equation 1:

$$J(\mathrm{wqk}, z(k)) = \Sigma(k=1,K)\Sigma(k=1,K)(\mathrm{wqk}) \left\| x(q) - z(k) \right\| 2 \qquad (1)$$

$$\Sigma(k=1,K)(\mathrm{wqk}) = 1 \text{ for each q } \mathrm{wqk} = (1/(\mathrm{Dqk})2) \; 1/(p-1)/\Sigma(k=1,K) \; (1/(\mathrm{Dqk})2)1/(p-1),\ p>1 \qquad (2)$$

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.

1.7. 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. It is selected for the following reasons:

- it is having no missing values so that missing values may be simulated and have the control over the evaluation process
- All the records are having corresponding clean class label
- It is having sufficiently large number of attributes and records
- 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 **Table 1-5**, the results arrived on a Windows XP laptop equipped with Intel core 2 duo CPU at 2GHz and 2GB RAM is presented. The Matlab implementations of the algorithms were used for evaluation.

Table 1. Performance in terms of rand index

	Clustering accuracy in terms of (average of five runs)			
Percent of missing values	MC/Mean value substn	Proposed RMS method	Proposed IRMS method	Proposed EMI-RBF method
10	0.848177	0.842323	0.845244	0.854081
20	0.851123	0.854081	0.848177	0.857051
30	0.845244	0.866036	0.860034	0.866036
40	0.839414	0.854081	0.845244	0.863029
50	0.827904	0.825058	0.839414	0.854081
Avg.	0.842372	0.848316	0.847623	0.858856

Table 2. Performance in terms of accuracy

	Clustering accuracy in terms of (average of five runs)			
Percent of missing values	MC/Mean value substn	Proposed RMS method	Proposed IRMS method	Proposed EMI-RBF method
10	91.740	91.390	91.560	92.090
20	91.920	92.090	91.740	92.270
30	91.560	92.790	92.440	92.790
40	91.210	92.090	91.560	92.620
50	90.510	90.330	91.210	92.090
Avg.	91.388	91.738	91.702	92.372

Table 3. Performance in terms of specificity

	Clustering accuracy in terms of (average of five runs)			
Percent of missing values	MC/Mean value substn	Proposed RMS method	Proposed IRMS method	Proposed EMI-RBF method
10	83.490	82.550	82.550	83.960
20	82.550	82.550	81.600	84.430
30	81.600	84.430	83.960	85.850
40	81.130	87.260	82.550	85.850
50	79.720	91.980	84.910	86.790
Avg	81.698	85.754	83.114	85.376

Table 4. Performance in terms of sensitivity

	Clustering accuracy in terms of (average of five runs)			
Percent of missing values	MC/Mean value substn	Proposed RMS method	Proposed IRMS method	Proposed EMI-RBF method
10	96.640	96.640	96.920	96.920
20	97.480	97.760	97.760	96.920
30	97.480	97.760	97.480	96.920
40	97.200	94.960	96.920	96.640
50	96.920	89.360	94.960	95.240
Avg.	97.144	95.296	96.808	96.528

Table 5. Performance in terms of MSE

	Clustering accuracy in terms of (average of five runs)			
Percent of missing values	MC/Mean value substn	Proposed RMS method	Proposed IRMS method	Proposed EMI-RBF method
10	0.151952	0.1555890	0.1552460	0.152350
20	0.149573	0.1555880	0.1548850	0.150090
30	0.148509	0.1563690	0.1550530	0.148211
40	0.147007	0.1559430	0.1552450	0.146585
50	0.143774	0.1558650	0.1526840	0.143409
Avg.	0.148163	0.1558708	0.1546226	0.148129

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

Figure 1-5 show 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 **Table 1**, the performance if imputation with reconstructed WDBC data is indirectly measured using the classification performance measure rand index. The better classification performance (high Rand Index) signifies the better imputation of missing values.

Figure 1 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 the state of the art EMI-RBF provided excellent performance.

In the **Table 2**, the performance in terms of accuracy measure. The better classification performance (high accuracy) signifies the better imputation of missing values.

Figure 2 shows the average performance in terms of accuracy. It is obvious that all the three proposed algorithms performed better than the standard MC/mean value substitution algorithm and the state of the art EMI-RBF provided excellent performance.

In the **Table 3**, the performance in terms of Specificity. The better classification performance (high Specificity) signifies the better imputation of missing values.

Bar chart in **Fig. 2** shows the average performance in terms of Accuracy. It is obvious that all the three proposed algorithms performed better than the standard MC/mean value substitution algorithm and the proposed algorithms RMS and EMI-RBF provided excellent performance.

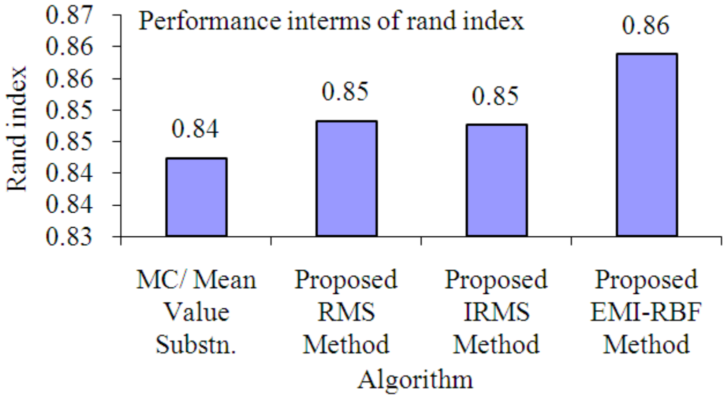

Fig. 1. Average Performance in terms of rand index

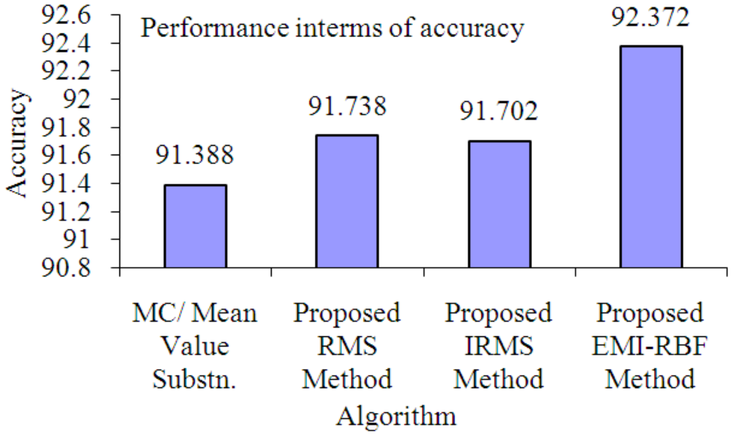

Fig. 2. Average performance in terms of accuracy

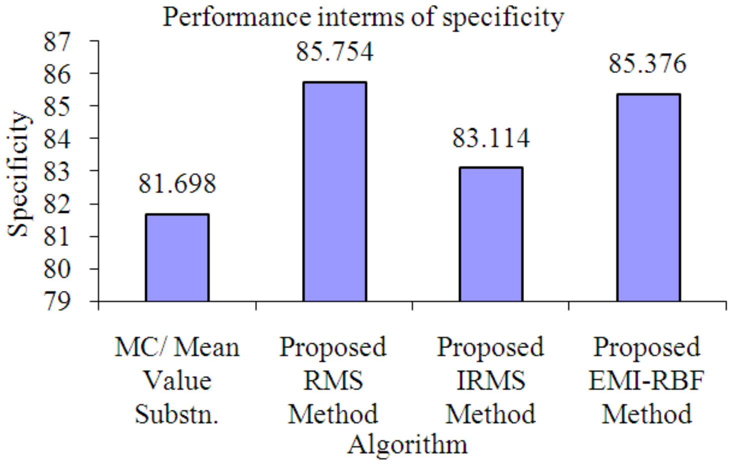

Fig. 3. Average performance in terms of specificity

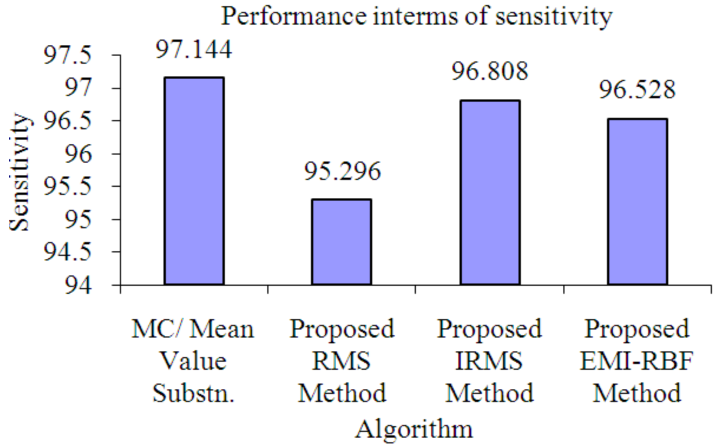

Fig. 4. Average performance in terms of sensitivity

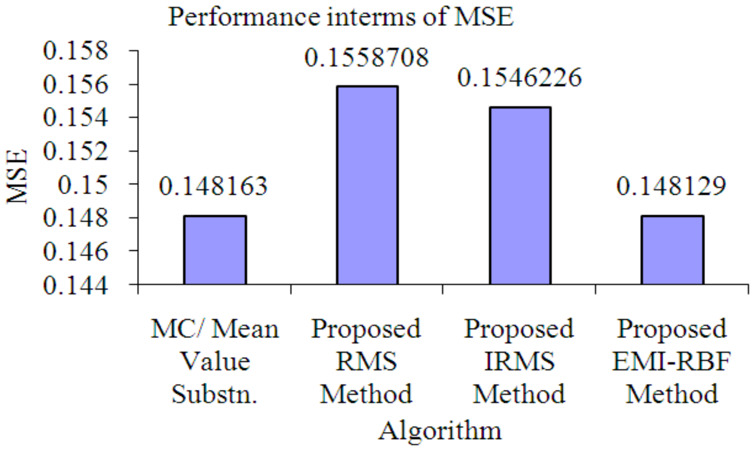

Fig. 5. Average performance in terms of MSE

In **Table 4**, the performance in terms of Sensitivity is given. 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.

Figure 4 shows the average performance in terms of sensitivity. Only with this metric, the performance of the standard MC/mean value substitution is high. The proposed IRMS algorithm provided little bit lower performance while the percentage of missing value is high.

In **Table 5**, the performance in terms of MSE is presented. Generally, the lower MSE signifies the better imputation of missing values. But in our experiments, it is observed that, even for higher MSE, the proposed methods RMS and IRMS provided higher performance in terms of other metrics.

Bar chart in **Fig. 5** sh**o**ws the average performance in terms of MSE. The state of the art EMI-RBF provided excellent performance. Even though the Average MSE in the case of other two previously proposed algorithms seems to be poor, those algorithms provided better results in terms of all other metrics.

2. CONCLUSION

The proposed imputation method has been successfully 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 performance of reconstruction was compared with the original WDBC data set.

In various previous works, 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 the proposed algorithm performs better than the most popular and standard methods.

The performance of the algorithms was evaluated with five different metrics. In almost all the metrics, proposed algorithms performed better than mean value substitution method. It is proved that the proposed EMI-RBF imputation method provided excellent performance than all other methods.

3. ACKNOWLEDGEMENT

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

4. REFERENCES

Ader, H.J. and G.J. Mellenberg, 2008. Chapter 13: Missing data, Advising on Research Methods: A Consultant's Companion. 1st Edn., Johannes Van Kessel Publishing, Huizen, ISBN-10: 9079418013, pp: 572.

Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. 1st Edn., Plenum Press, New York, pp: 256.

Donder, A.R.T., G.J.M.G. Van Der Heijdenc, T. Stijnend and K.G.M. Moons, 2006. Review: A gentle introduction to imputation of missing values. J. Clin. Epidemiol., 59: 1087-1091. DOI: 10.1016/j.jclinepi.2006.01.014

Kotsiantis, S.B., D. Kanellopoulos and P. E. Pintelas, 2006. Data preprocessing for supervised leaning. Int. J. Comput. Sci., 1: 111-117.

Little, R.J.A. and D.B. Rubin, 2002. Statistical Analysis With Missing Data. 2nd Edn., Wiley, New York, ISBN-10: 0471183865, pp: 381.

Messner, S.F., 1992. Exploring the consequences of erratic data reporting for cross-national research on homicide. J. Quantitative Criminol., 8: 155-173. DOI: 10.1007/BF01066742

Zhang, S., C. Zhang and Q. Yang, 2004. Guest editors' introduction-Information enhancement for data mining. IEEE Intell. Syst., 19: 12-13. DOI: 10.1109/MIS.2004.1274905