A Preprocessing Method to Deal with Missing Values by Integrating Clustering and Regression Techniques

Shin-Mu Tseng[†], Kuo-Ho Wang[†] and Chien-I Lee [‡]
[†]Department of Computer Science and Information Engineering
National Cheng Kung University, Taiwan, R.O.C.
E-mail: tsengsm@mail.ncku.edu.tw
[‡]Institute of Computer Science and Information Education
National Tainan Teachers College, Tainan, Taiwan, R.O.C.
E-mail: leeci@ipx.ntntc.edu.tw

Abstract

Data preprocessing is a critical task in the knowledge discovery process in order to ensure the quality of the data to be analyzed. One widely studied problem in data preprocessing is the handling of missing values, with the aim to recover its original value. Based on numerous studies on missing values, it is shown that different methods are needed for different types of data missingness. In this work, we proposed a new method to deal with missing values in datasets where cluster properties exist among the data records. By integrating the clustering and regression techniques, the proposed method can predict the missing values with higher accuracy. To our best knowledge, this is the first work combining regression and clustering analysis to deal with the missing values problem. Through empirical evaluation, the proposed method was shown to perform better than other methods under different types of datasets.

Keywords: data preprocessing, missing values, regression, clustering, knowledge discovery

1 Introduction

In recent years, the rapid development of data mining techniques has enabled the successful knowledge discovery applications in various industries [Han and Kamber 2000]. Data preprocessing is a critical task in the knowledge discovery process for ensuring good data quality. One important problem in data preprocessing is the handling of missing values in a dataset, with the aim to recover the missing values to be as close as possible to the original values. Although a number of studies have been made on dealing with missing values [Liu et al. 1997, Kalton and Kasprzyk 1982, Little and Rubin 1987, McQueen 1967, Pyle 1999, Ragel and Cremilleux 1998, 1999, Lee et al. 1976], it was observed that the accuracy in recovering the missing values is determined by the suitable matching between the type of dataset and the good analysis methods. It seems that no single method can handle well all kinds of missingness of datasets. Hence, an important issue in designing missing-value handling methods is to take into account the inherent property of a dataset.

In this research, we proposed a new method for handling missing values in datasets where cluster properties exist for the data records. This kind of phenomenon exists in many reallife applications. For example, consider a customer database in a bank, where the main attributes of customers include age, gender, occupation, income, etc. Based on these attributes, the customers can be classified into several clusters, where the ones in a same cluster have similar properties. By using the knowledge of the cluster properties, the recovery of the missing values in the data records can be done better with higher accuracy.

The main idea of our approach is to integrate the clustering and regression techniques for estimating the missing values. For a given dataset D, firstly we use the regression method to estimate the values of the missing ones and fill them with the estimated values to form a temporary dataset D'. Then, the clustering and validation analysis is conducted on D' to discover the best clustering of all data records. For each cluster, the data values missing in D are calculated again by applying the regression method only over the data records within the same cluster since these records have high similarities. In this way, the missing values in the dataset can be estimated more accurately due to the considerations of the cluster properties. To our best knowledge, this is the first work integrating the regression and clustering techniques to deal with the missing values problem.

To evaluate the performance of the proposed method, some experiments were conducted under datasets with different types of data distributions and missingness. The empirical results show that the proposed method deliver higher accuracy in recovering the missing values than other methods under different circumstances.

The remainder of this paper is organized as follows: some previous studies on this problem are introduced in Section 2. The proposed method is described in details in Section 3, and the experimental setup and results are given in Section 4. Finally, the conclusions and future work are given in Section 5.

2 Previous Work

The problem of missing value handling has been studied for many years, with numerous methods have been proposed. The existing methods can be categorized into two types: imputation-based and data-mining based methods. The former types of methods are primarily for handling missing values of numerical data, while the latter for category data. The principle of imputation methods is to estimate the missing values by using the existing values as an auxiliary base. The underlying assumption is that there exist certain correlations between different data tuples over all attributes. The existing methods include mean imputation, hot-deck imputation, cold-deck imputation, regression imputation, expectation maximization (EM), composite imputation, etc. For the data-mining based methods, techniques like associations [Ragel and Cremilleux 1998], clustering[Lee 1976] and classifications[Liu 1997] are used to discover the similar patterns between data tuples so as to predict the missing values.

In this work, we confine the study to handling missing values of numerical type. For the existing methods, it was observed that specialized methods are needed for different types of data missingness, and no single method can handle well all kinds of missingness of datasets.

3 Proposed Method

The main ideas of our approach are two folded: 1) Utilize the cluster property of the given dataset, 2) Integrate the prediction methods with the clustering methods to produce more accurate results. In the following, we first introduced the main concepts of the proposed method. Then, the proposed method is stated in details.

3.1 Main Concepts

As described previously, in recovering the missing values, the clustering-based methods have the advantages of capturing the cluster properties among data records, while the regression methods can effectively predict the values by finding the correlations between data records. However, the regression methods used in previous researches incur the problem that the sample base used for imputing the missing values is too large when the whole dataset is considered. Consequently, the precision of the imputation may be low. On the other side, the main problem encountered in clustering-based imputations is how to determine the correct and best clustering for the data records with missing values.

The main concept of our method is to combine the regression and clustering methods for handling the above problems. For a given dataset, firstly we use the regression method to obtain roughly estimated values for the missing ones. Then, the clustering analysis is conducted on the new dataset to find out the best clustering. For each discovered cluster, regression analysis is applied again to predict the originally missed values with only the data records in the same cluster as the base. In this way, the missing values can be recovered more accurately due to the considerations of the cluster properties while conducting regression analysis.

As an example, consider the simple 2-dimensional (X, Y) dataset as shown in Figure 1. It is obvious that a good linear regression model can be generated with high R^2 (coefficient of determination) since the base of datum exhibits good locality. However, for the dispersed data as shown in Figure 2, it is difficult to establish a good regression model with high R^2 . By using our approach, the whole dataset can be partitioned into three clusters first as shown in Figure 3, then good regression models can be built for clusters C_1, C_2 and C_3, with R^2 as 0.926, 0.813 and 0.862, respectively. Therefore, more accurate values can be predicted for the missing values by using the built regression models.

3.2 Proposed Method: $RegressionandClustering(RC)$

We propose a new method, namely $RegressionandClustering(RC)$, for handling the missing values. Given a dataset D , the RC method consists of three steps as follows:

1. The whole dataset D is divided into two parts, namely D_c and D_m , where D_m are composed of all data records with missing values and D_c = D - D_m . Firstly, the missing values in D_m are recovered using regression method with D_c as the base. Denote $D\prime$ as the temporary dataset in which all missing values are replaced by the imputation values.
2. Apply clustering analysis on $D\prime$ to discover the best clustering result. Suppose k clusters are produced, namely C_1, C_2, $\cdots$, C_k, , where $\sum \mid C_i \mid = \mid D \mid$.

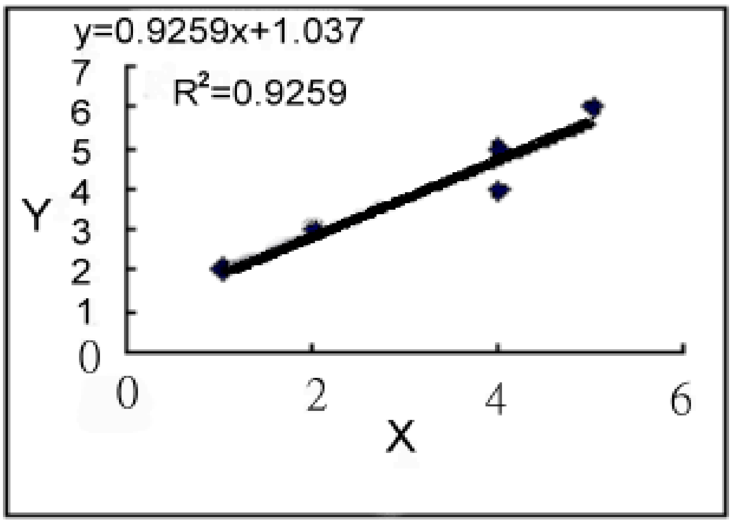

Figure 1: Dataset with good regression model

3. For each cluster C_i, the regression analysis is applied on all records R_j for predicting the missing values, where $R_j \in D_m \bigcap C_i$, and the base used for regression is the set $\{R_c, \mid R_c \in D_c \bigcap C_i$ $\}$.

   All the missing values are replaced by the finally predicted values.

The purpose of step 1 is to replace the missing values with estimated ones such that they can also be taken into account for clustering in the further steps. This is also to resolve the problem encountered in clustering-based imputation that too many samples are lost due to the existence of missing values. After the best clustering is produced in step 2, the regression analysis performed in step 3 will produce more accurate results for the missing values since the base scope is confined to the correlated cluster.

3.3 Finding the Best Clustering Result

One problem arisen in clustering-based imputation is how to produce the best clustering result for a given dataset. To discover the best clustering result for a given dataset, we adopt the approach similar to [Tseng and Kao 2002], which used an iterative validation approach. The main ideas of the proposed clustering method are as follows. First, the CAST [Ben-Dor et al. 1999] algorithm was used as the basic clustering method, which efficiently generates a clustering result based on the value of an input parameter named $affinity$ *threshold* t, where $0 < t < 1$. The average similarity between the items in a generated cluster is guaranteed to be above t.

Secondly, a quality validation method is applied to find the best clustering result. To validate the quality of the clustering result, the *Huberts* Γ *statistic* is used to measure it. Let X=[$X(i,j)$] and Y=[$Y(i,j)$] be two $n \bigodot n$ matrix, $X(i,j)$ indicates the similarity of data record i and j, $Y(i,j)$

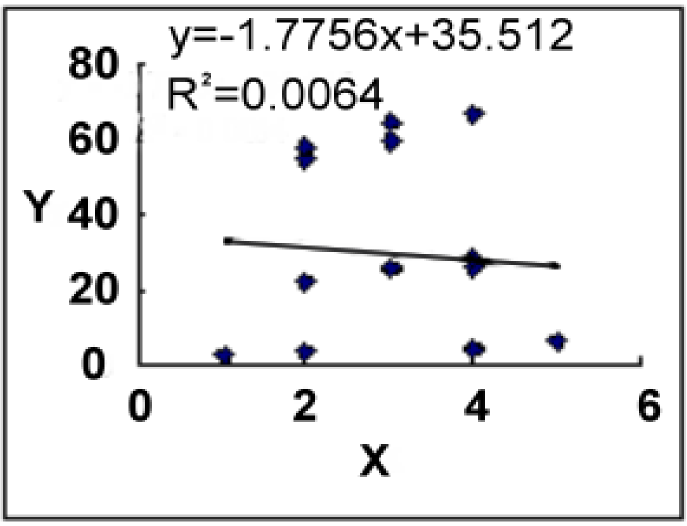

Figure 2: Dataset with poor regression model

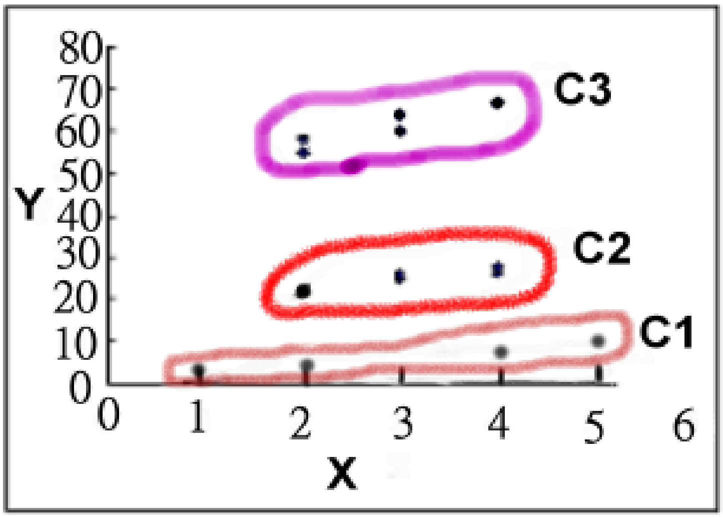

Figure 3: Clustering of the example dataset

is defined as follows:

$$Y(i,j) = \begin{cases} 1, & \text{if items i and j in same clusetr,} \\ 0, & \text{otherwise} \end{cases}$$

Hubert's Γ statistic represents the point serial correlation between the matrix X and Y, and is defined as follows:

$$\Gamma = \frac{1}{M} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\frac{X(i,j) - \bar{X}}{\sigma_x} \right) \left(\frac{Y(i,j) - \bar{Y}}{\sigma_y} \right)$$

where $M = n(n-1)\ /\ 2$ and Γ is between [-1, 1]. A higher value of Γ represents the better clustering quality.

Therefore, it is clear that the best clustering result can be obtained by running CAST algorithm with different values for $affinity\ threshold\ t$, and choosing the clustering result with the highest value of $Huberts\ \Gamma\ statistic$. To determine the best value for the $affinity\ threshold\ t$, the easiest way is to fix the increment of the value of $affinity\ threshold\ t$, e.g., 0.05 to 0.95 in steps of 0.05. A heuristic was also proposed in [Tseng and Kao 2002] to reduce the iterations of computations effectively. Through experimental evaluations, this approach was shown to perform much better than other clustering methods like *k*-means in both of accuracy and efficiency.

In calculating the similarity between two data records, we adopt Pearson's correlation coefficient for better integration with regression analysis. To illustrate why correlation coefficient is a better choice, consider Figure 4 which shows two data records, namely A and C, with the value of each attribute plotted over the X-axis. It is clear that there exists high correlation between A and C. Hence, it is very likely that they will be grouped into the same cluster if correlation coefficient is used as similarity measure. However, this will not hold if Euclidean distance is used as the similarity measure. Given another data record B, whose attribute values are about the means of those of A and C originally but some attribute values are missing. Consequently, the missing values will be recovered correctly by our approach with correlation coefficient used as the similarity measure since A, B and C will be grouped into the same cluster while applying the regression imputation.

4 Empirical Evaluation

We evaluate the performance of RC and compare it with other methods under different kinds of datasets. Two main types of datasets are used: randomness-based dataset and cluster-based one. The former was produced by a random number generator, while the latter by a volumetric-clouds type clusters simulator [Chen 2002] with the following parameters:

- *FieldNum*: the number of attributes for each data tuple.

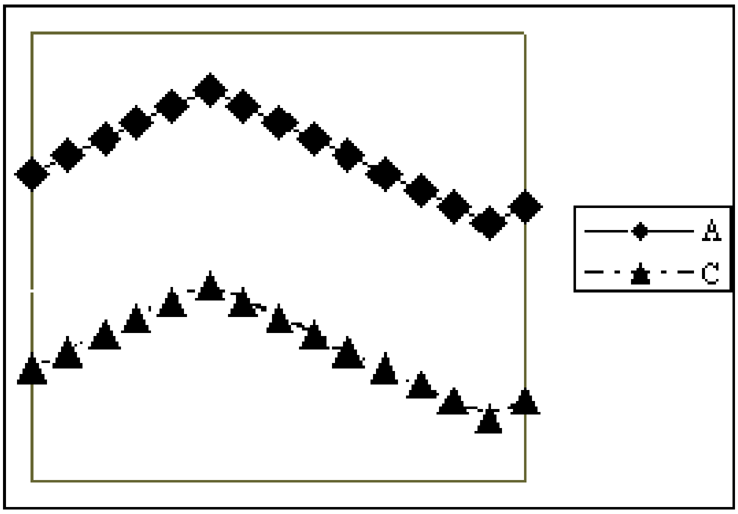

Figure 4: Two data records with high correlation.

- $ClusterNum$: the number of main clusters to be generated.
- $TupleNum$: the number of tuples in the dataset.
- $ScatterNum$: the degree of scatter of the clusters.
- $MissingRatio$: the ratio of the number of tuples with missing values.

The methods compared to RC include EM, regression (denoted as RG), average (denoted as Ave) and k-means, where the value of k is set between 3 and 48 in steps of 3 and the input parameter t for RC is set between 0 and 1 in steps of 0.1.

For the performance metric, the *Relative Absolute Deviation*(RAD) was used to measure the correctness of the prediction on the missing values. The RAD is defined as follows:

$$RAD = \frac{1}{P_k} \sum_{i=1}^{P_k} \frac{\mid Z_i - Z_i^* \mid}{Z_i}$$

where Z_i is the original value, Z_i^* is the predicted value and P_k is the total number of missing values.

4.1 Experiment 1: Cluster-based Dataset

In this experiment, a dataset with cluster structure is produced by setting the parameters as in Table 1. The *Missing Ratio* was varied from 5% to 20% in steps of 5% as the percentage of the total number of data records. To simplify the evaluation, only one attribute was made as missing for a data record selected to be in the proportion with missing values.

Table 1: Base parameter settings.

Parameters	Values
FieldNum	10
ClusterNum	4
TupleNum	5000
ScatterNum	70%
MissingRatio	5%-20%

Figure 5 shows the experimental results, on which the following observations were made:

- In overall, RC outperforms other methods under varied missing ratio; RG and EM have very similar performance, which is ranked as the next best; KM has the worst. Performance.
- For the effect of varying *Missing Ratio*, KM performs worse with MissingRatio increased; RG delivers stable performance under different MissingRatio; RC performs slightly worse under higher MissingRatio. This is because that the accuracy of regression-based analysis is dominated by the data distribution instead of the percentage of missing values.
- The poor performance of KM is resulted from the property of high degree of scatters (70%) in the dataset, meaning that the values of data tuples in the same cluster might have scaled correlations and obvious differences in the absolute values. Under such conditions, KM will produce high *RAD* since it directly adopts the averaged value of all tuples as the predicted value. In addition, it is difficult to find the best clustering by using KM, especially when there exist outliers in the dataset.

To examine the effect of cluster property of a dataset, Figure 6 shows the experimental result with the parameter $ClusterNum$ being changed as 8. It is obvious that *RC* performs much better than other methods, as compared to the results in Figure 5. This indicates that *RC* is a good method for dealing with dataset bearing cluster structures.

4.2 Experiment 2: Random Dataset

The purpose of this experiment is to examine the performance of tested methods under a random dataset, which means that the data values are in random distributions without obvious cluster structures. The parameters are set as same in Experiment 1, while the values for each data attribute are generated randomly between (0, 999). Figure 7 shows that all methods perform much worse than under the cluster-based datasets. I n particular, RG, EM and AVE perform similarly and are outperformed by KM and RC. This indicates that it is hard to predict the missing values by using regression analysis directly for a random dataset. However, the accuracy can be improved by incorporating the clustering methods.

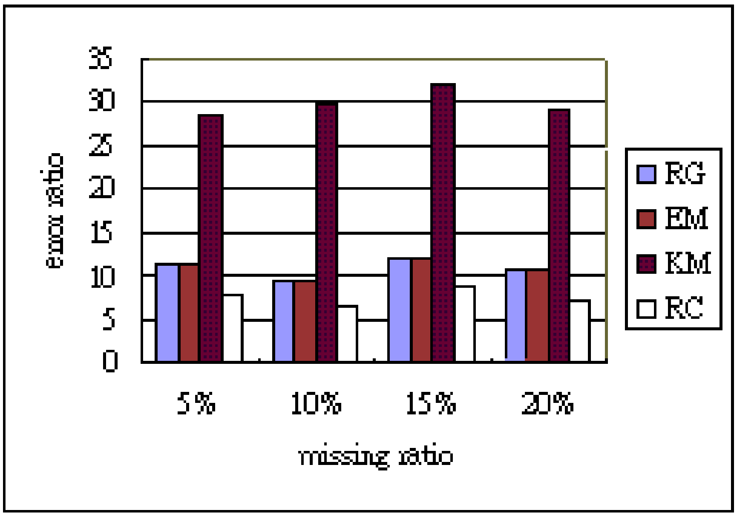

Figure 5: RAD under cluster-based data.

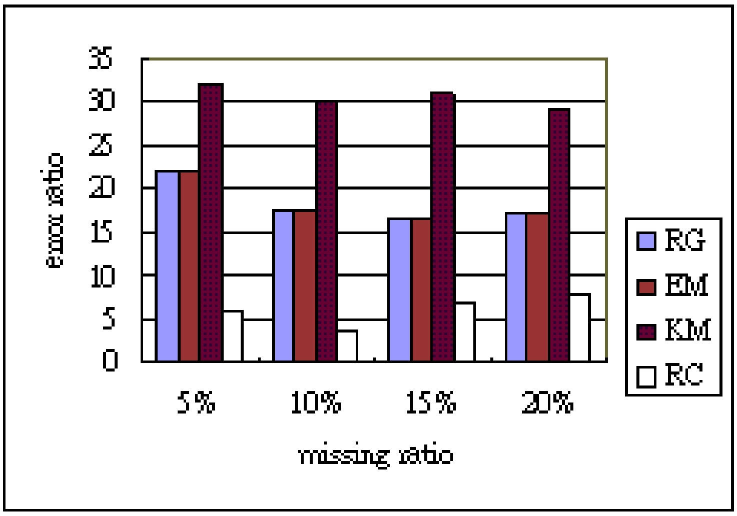

Figure 6: RAD results with ClusterNum =8.

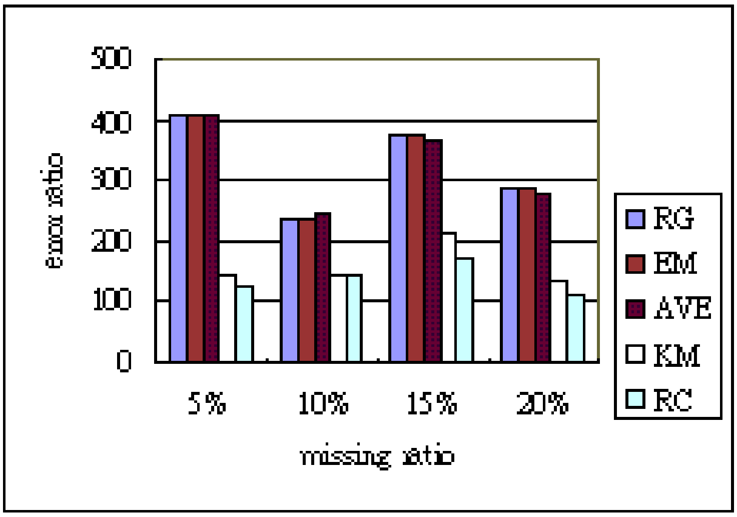

Figure 7: RAD under random data.

5 Conclusions

A new method, namely $RegressionandClustering(RC)$, is proposed for handling the missing values in a large dataset for ensuring the data quality. The main feature of RC is it integrates the clustering method with regression analysis such that the regression can be applied on several clusters with narrower scope. In this way, the accuracy of the predicted values for the missing ones can be improved substantially, as verified by the empirical evaluation under different datasets, especially for the cluster-based ones.

For future work, we will apply the RC method on real datasets in different domains, like biological and financial data. More experiments will also be conducted for obtaining more detailed evaluation of the RC method under various conditions. In addition, we will also investigate about how to reduce the computation time in conducting the tasks of value predictions.

Acknowledgement

This work was partially supported by National Science Council, Taiwan, R.O.C., under grant no. NSC91-2213-E-006-074.

References

Ben-Dor, and Z. Yakhini. 1999. Clustering gene expression patterns. In *Proceedings of the 3rd Annual International Conference on Computational Molecular Biology.*

Liu, W.Z., A.P White, S.G Thompson and M.A. Bramer. 1997. Techniques for Dealing with Missing

Values in Classification. In *2nd Int. Symp. Intelligent Data Analaysis*, 527-536.

Tseng, Shin-Mu and Ching-Pin Kao. 2002. Efficient Clustering Methods for Gene Expression Mining: a performance Evaluation. In *Proc. Sixth Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 432-437.

Han, Jiawei and M. Kamber. 2000. Data Mining: Concepts and Techniques. *Morgan Kaufmann Publishers.*

Kalton, G. and D. Kasprzyk. 1982. Imputing for missing survey response. In Proc. Sect. Survey Res. Meth., Amer. Statist. Assoc., 22-23.

Chen, Lien-Chin. 2002. A Correlation-Based Approach for Validating Gene Expression Clustering. Master Thesis, Department of Computer Science and Information Engineering, National Cheng Kung University, Taiwan, R.O.C..

Little, R.J.A. and D.B. Rubin. 1987. Statistical Analysis with Missing Data. *John Wiley and Sons Publishers.*

McQueen, J.B. 1967. Some Methods of Classification and Analysis of Multivariate Observations. In *Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability*, 281-297.

D. Pyle, 1999. Data Preparation for Data Mining. Morgan Kaufmann Publishers.

Ragel, A. and B. Cremilleux. 1998. Treatment of Missing Values for Association Rules. In *Proc. 2nd Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, 258-270.

Ragel, A. and B. Cremilleux. 1999. MVC: a preprocessing method to deal with missing values. *Knowledge-Base System*, 12(5):205-332.

Lee Richard C.T., James R. Slagle, and C.T. Mong. 1976. Application of Clustering to Estimate Missing Data and Improve Data Integrity. In *Proc. Int'l Conf. Software Engineering*, 539-544.