Applied Artificial Intelligence, 17:535–544, 2003

0883-9514/03 $12.00 + .00
DOI: 10.1080/08839510390219318

Taylor & Francis
Taylor & Francis Group

A PRE-PROCESSING METHOD TO DEAL WITH MISSING VALUES BY INTEGRATING CLUSTERING AND REGRESSION TECHNIQUES

SHIN-MU TSENG and KUO-HO WANG
Department of Computer Science and Information Engineering, National Cheng Kung University, Taiwan, China

CHIEN-I. LEE
Institute of Computer Science and Information Education, National Tainan Teachers College, Tainan, Taiwan, China

Data pre-processing 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 pre-processing 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 missing data. In this work, we propose a new method to deal with missing values in data sets 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 data sets.

INTRODUCTION

In recent years, the rapid development of data-mining techniques has enabled successful knowledge discovery applications in various industries (Han and Kamber 2000). Data pre-processing is a critical task in the knowledge discovery process for ensuring good data quality. One important problem in data pre-processing is the handling of missing values in a data set

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

Address correspondence to Shin-Mu Tseng, Department of Computer Science and Information Engineering, National Cheng Kung University, Taiwan, China. E-mail: tsengsm@mail.ncku.edu.tw

with the aim to be to recover the missing values 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 data set and good analysis methods. It seems that no single method can handle well all kinds of missing data sets. Hence, an important issue in designing missing-value handling methods is to take into account the inherent property of a data set.

In this research, we propose a new method for handling missing values in data sets where cluster properties exist for the data records. This kind of phenomenon exists in many real-life 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 the 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 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 data set 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 data set 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 data set 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 data sets with different types of data distributions and missingness. The expirical results show that the proposed method delivers higher accuracy in recovering the missing values than other methods under different circumstances.

PREVIOUS WORK

The problem of missing value handling has been studied for many years with numerous methods 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 exists 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 such as associations (Ragel and Cremilleux 1998), clustering (Lee et al. 1976), and classifications (Liu et al. 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 the numerical type. For the existing methods, it was observed that specialized methods are needed for different types of missing data, and no single method can handle well all kinds of missing data sets.

PROPOSED METHOD

The main ideas of our approach are two-fold: 1) utilize the cluster property of the given dataset and 2) integrate the prediction methods with the clustering methods to produce more accurate results. In the following, we first introduce the main concepts of the proposed method. Then, the proposed method is stated in detail.

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 research incur the problem that the sample base used for imputing the missing values is too large when the whole data set is considered. Consequently, the precision of the imputation may be low. On the other hand, 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 data set, first, we use the regression method to obtain roughly estimated values for the missing ones. Then, the clustering analysis is conducted on the new data set 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) data set 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 data set can be partitioned into three clusters as shown in Figure 3. Then good regression models can be built for clusters C_1, C_2, and

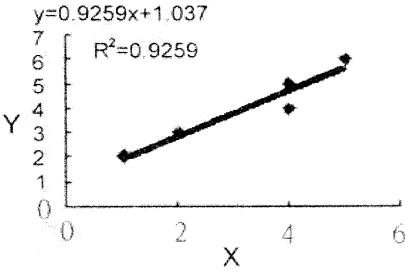

FIGURE 1. Data set with good regression mode.

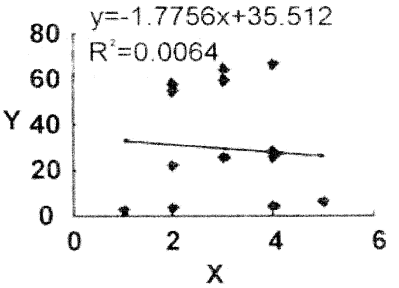

FIGURE 2. Data set with poor regression model.

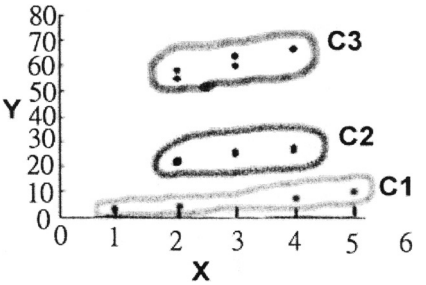

FIGURE 3. Clustering of the example data set.

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.

Proposed Method: *RegressionandClustering(RC)*

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

1. The whole data set D is divided into two parts, namely D_c and D_m, where D_m is composed of all data records with missing values and $D_c = D - D_m$. First, the missing values in D_m are recovered using regression method with D_c as the base. Denote D' as the temporary data set in which all missing values are replaced by the imputation values.
2. Apply clustering analysis on D' to discover the best clustering result. Suppose k clusters are produced, namely $C_1, C_2, \ldots, C_k$, where $\Sigma|C_i| = |D|$.
3. For each cluster C_i, the regression analysis is applied on all records R_j for predicting the missing values, where $R_j \in D_m \cap C_i$, and the base used for regression is the set $\{R_c | R_c \in D_c \cap C_i\}$.

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.

Finding the Best Clustering Result

One problem that has arisen in clustering-based imputation is how to produce the best clustering result for a given data set. To discover the best clustering result for a given data set, 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 and Yakhini 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.

Second, 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 n matrix; $X(i, j)$ indicates the similarity of data record i and j; $Y(i, j)$ is defined as follows:

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

Huberts 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* Γ *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 (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, such as k-means, in both of accuracy and efficiency.

In calculating the similarity between two data records, we adopt Pearsons 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 high correlation exists between A and C. Hence, it is very likely that they will be grouped into the same cluster if correlation coefficient is used as the similarity measure. However, this will not hold if Euclidean distance is used as the similarity

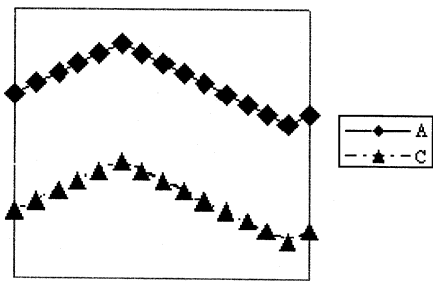

FIGURE 4. Two data records with high correlation.

measure. Take 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.

EMPIRICAL EVALUATION

We evaluate the performance of RC and compare it with other methods under different kinds of data sets. Two main types of data sets are used: randomness-based data set and cluster-based data set. 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.
- *ClusterNum*: The number of main clusters to be generated.
- *TupleNum*: The number of tuples in the data set.
- *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 paramter 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:

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

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

Experiment 1: Cluster-Based Data Set

In this experiment, a data set 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 noted as missing for a data record selected to be in proportion with missing values.

TABLE 1 Base Parameter Settings

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

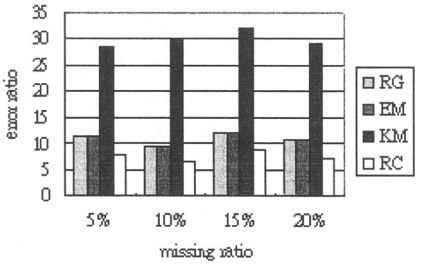

FIGURE 5. RAD under cluster-based data.

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

- Overall, RC outperforms other methods under varied missing ratios; 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 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 a high degree of scattering (70%) in the data set, 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 exists outliers in the data set.

To examine the effect of cluster property of a data set, Figure 6 shows the experimental result with the parameter *ClusterNum* being changed to 8. It is obvious that *RC* performs much better than other methods as compared to

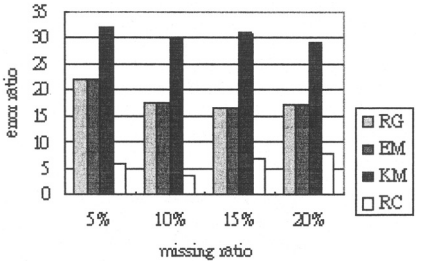

FIGURE 6. RAD results with ClusterNum = 8.

the results in Figure 5. This indicates that RC is a good method for dealing with data set-bearing cluster structures.

Experiment 2: Random Data Set

The purpose of this experiment is to examine the performance of tested methods under a random data set, which means that the data values are in random distributions without obvious cluster structures. The parameters are set the same as 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 data sets. In particular, RG, EM, and AVE perform similarly and are outperformed by KM and RC. This indicates that it is difficult to predict the missing values by using regression analysis directly for a random data set. However, the accuracy can be improved by incorporating the clustering methods.

CONCLUSIONS

A new method, namely *RegressionandClustering*(*RC*), is proposed for handling the missing values in a large data set for ensuring data quality. The main feature of RC is it integrates the clustering method with regression

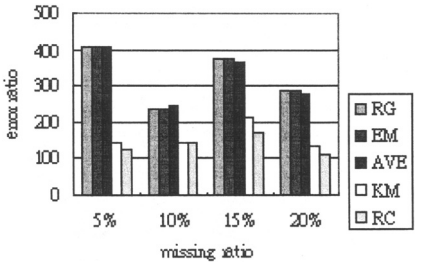

FIGURE 7. RAD under random data.

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 data sets, especially for the cluster-based ones.

For future work, we will apply the RC method on real data sets in different domains, such as 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 how to reduce the computation time in conducting the tasks of value predictions.

REFERENCES

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

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

Han, J., and M. Kamber. 2000. *Data Mining: Concepts and Techniques*. San Mateo, CA: Morgan Kaufmann.

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

Lee R. C. T., J. 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*, pages 539–544, San Francisco, CA, IEEE Press.

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

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 Analysis*, pages 527–536.

McQueen, J. B. 1967. Some methods of classification and analysis of multivariate observations. In *Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, Berkeley, CA.

Pyle, D. 1999. *Data Preparation for Data Mining*. San Mateo, CA: Morgan Kaufmann.

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*, pages 258–270.

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

Tseng, S.-M., and C.-P. Kao. 2002. Efficient clustering methods for gene expression mining: A performance evaluation. In *Proc. Sixth Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 432–437.