Imputing manufacturing material in data mining

Ruey-Ling Yeh · Ching Liu · Ben-Chang Shia · Yu-Ting Cheng · Ya-Fang Huwang

Received: March 2007 / Accepted: July 2007 / Published online: November 2007

Abstract Data plays a vital role as a source of information to organizations, especially in times of information and technology. One encounters a not-so-perfect database from which data is missing, and the results obtained from such a database may provide biased or misleading solutions. Therefore, imputing missing data to a database has been regarded as one of the major steps in data mining. The present research used different methods of data mining to construct imputative models in accordance with different types of missing data. When the missing data is continuous, regression models and Neural Networks are used to build imputative models. For the categorical missing data, the logistic regression model, neural network, C5.0 and CART are employed to construct imputative models. The results showed that the regression model was found to provide the best estimate of continuous missing data; but for categorical missing data, the C5.0 model proved the best method.

Keywords Data mining · C5.0 · Regression · BPNN · Missing data · Imputation

R.-L. Yeh (✉) · C. Liu
Division of Biometrics, Graduate Institute of Agronomy,
National Taiwan University, Taipei, Taiwan
e-mail: d90621202@ntu.edu.tw

B.-C. Shia
Department of Statistics and Information Science,
Fu Jen Catholic University, Taipei, Taiwan

Y.-T. Cheng · Y.-F. Huwang
Department of Statistics Science,
National Chengchi University,
Taipei, Taiwan

Introduction

In traditional statistical analysis, the application of databases is pretty straightforward. In order to obtain a good quality and representative database, one approach is to get the value-added database for data mining applications. The concept of the database with value added can be divided into three phases from the statistical viewpoint: (1) Sampling survey, (2) Functional, and (3) Application.

1. Any sampling survey can be subdivided into three parts as follows:
   (1) The imputation of missing data:
      Missing data indicates lost information. If imputation of the missing data is implemented, the database will display a better result.
   (2) Index and criteria:
      The structure of the sample and population is discussed using similarity and correlation. The prediction capability of the value-added data is measured against the index and criteria. The result is evaluated for improved prediction accuracy of the value-added data.
   (3) Sampling methods:
      Discuss the efficiency of the different sampling methods for different datasets.
2. Functional: Figure 1.
3. Application: After going through the sampling survey and functional phases, the data using data mining algorithms are analyzed and the output to determine are examined if what has been discovered is both useful and interesting. Decisions are made about whether to repeat previous steps using new parameters.

It is quite common to have missing data in a dataset. Missing data do present a problem for the outcome of the

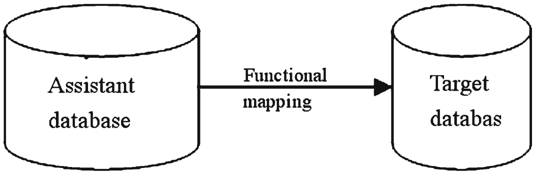


Fig. 1 Functional mapping

application (Kalton and Kasprzyk 1982). If more than one record has missing data, simply ignoring them may cause the remaining records unsuitable for data mining at all (Li et al. 2006).Therefore, with regard to the missing data in the database, in the light of the importance of missing value having an effect on the outcome, this research attempts to find applicable techniques to predict the values of the missing data; and it evaluates the accuracy of the estimation by comparing actual values with predicted values.

Related work

Imputation methods

Missing data are a part of most of the research, and missing data can seriously affect research results (Robert 1996). So, it has to be decided how to deal with it. If one ignores missing data or assumes that excluding missing data is acceptable, there is a risk of reaching invalid and non-representative conclusions. There are a number of alternative ways of dealing with missing data (Joop 1999). In general, one uses the weight adjustment method for unit non-response and the imputation method for item non-response. The present research examines item non-response and will introduce the handling of missing data involving item non-response. There are many methods of imputation (Litte and Rubin 1987) as follows:

1. *Mean imputation (MI)*: This method replaces the missing observations of a certain variable with the mean of the observed values in that variable. It is a simple method that generally performs well, especially when valid data are normally distributed.
2. *Regression imputation (RI)*: The missing values are estimated through the application of multiple regression where the variable with missing data is considered as the dependent one and all other variables as predictors.
3. *Expectation maximization (EM)*: The EM algorithm is an iterative two step procedure obtaining the maximum likelihood estimates of a model starting from an initial guess. Each iteration consists of two steps: the expectation (E) step that finds the distribution for the missing data based on the known values for the observed variables and the current estimate of the parameters and the maximization (M) step that replaces the missing data with the expected value.

In general, the overall mean imputation is not recommended. Kalton and Kasprzyk (1986) suggested that the sample be stratified into classes based on auxiliary variables after which one could then impute the class mean for nonrespondents within the class and call that ”within-class mean value imputation”. While this method may not be perfect, it certainly represents an improvement on the overall mean method.

The ‘hot deck’ imputation method is the technique where the data file is stratified into classes and cases and respondents within classes in the current survey file are used to impute blank values in incomplete records (Ford 1983). In other words, the hot deck imputation method replaces missing values by values from a similar unit and then chooses observed values randomly from donors that are in the same imputation class to impute the missing value. In order to create imputation classes, auxiliary variables that are related to the missing mechanism are needed. The ‘cold deck’ imputation method is of historical interest, but little used in practice (Lessler and Kalsbeek 1992). It may easily give rise to multiple uses of donors, thus leading to a lack of precision in the survey estimates (Kalton and Kasprzyk 1986).

The goal of the imputation method is to reduce the bias of survey estimates. Imputation of missing data minimizes bias and allows for analysis using a rectangular dataset, so that standard analysis can then proceed. But, imputed data is not real data and variance estimates need to reflect the instability of the data. Since a single imputation nearly always reduces variance estimates, these cannot reflect the instability of the missing data (Judi 2002).The single imputation method always imputes the same value, thereby ignoring the variance associated with the imputation process. The multiple imputations method imputes several imputed values and the effect of the chosen imputed values on the variance can be taken into account. Multiple imputing is the only way of estimating the total variance, which includes the variance within imputations (Rubin 1987).

Artificial neural network

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems interacts, such as the brain processes information. The Neural Network receives and propagates the messages from and by the Artificial Neuron. By using the mathematical function in the artificial intelligence neuron to do the transformation, the predicted results are exported.

Backpropagation Neural Network (BPNN) is one of the most popular neural network learning algorithms. Werbos (1974) proposed the learning algorithm of the hidden layers

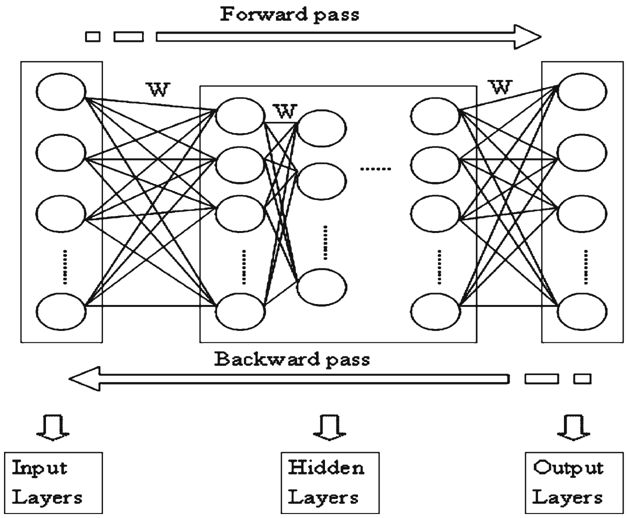


Fig. 2 The structure of BPNN

and applied to the prediction in the economy. With a more sophisticated learning rule, BPNNs overcome the limitations that single-layer networks have. Input vectors and the corresponding target vectors are used to train a network until it can approximate a function, associate input vectors with specific output vectors, or classify input vectors in an appropriate way as defined by the user (Craven 1997).

BPNNs start as a network of nodes arranged in three layers—the input, hidden, and output layers. The input and output layers serve as nodes to buffer input and output for the model respectively, and the hidden layer serves to provide a means for input relations to be represented in the output. Before any data has been run through the network, the weights for the nodes are random. The usual implementation of the backpropagation algorithm consists of two phases:(i) Forward pass; (ii) Backward pass. The forward pass in which an input pattern is presented to the network and the actual outputs are calculated; and a backward pass in which the errors are calculated and the weights are adjusted. A backward pass is always carried out after each forward pass. The structure of BPNN is as follows: Figure 2.

Decision tree

Classification is an important technique in data mining. And the decision tree is the most efficient approach to classification problems (Friedman 1997). The input to a classifier is a training set of records, each of which is a tuple of attribute values tagged with a class label. A set of attribute values defines each record. A decision tree has the root and each internal node labeled with a question. The arcs emanating from each node represent each possible answer to the associated question. Each leaf node represents a prediction of a solution to the problem under consideration.

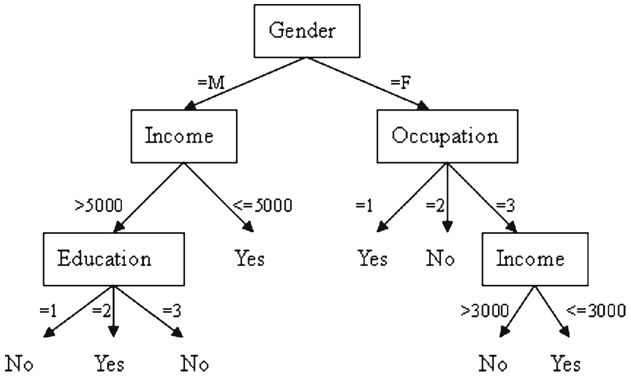


Fig. 3 Basis structure of C5.0

C5.0 is a commercial version of C4.5 now widely used in many data mining packages. A C5.0 model works by splitting the sample based on the field that provides the maximum information gain. The Gain Ratio is defined as $GainRatio\left(D,S\right) = \frac{Gain(D,S)}{H\left(\frac{|D_1|}{D},\ldots,\frac{|D_S|}{D}\right)}$, where D is a database state, $H(\cdot)$ finds the amount of order in that state, when the state is split into s new states $S = \{D_1, D_2, \ldots, D_S\}$. C5.0 used the larger than average information gain. This is to compensate for the fact that the Gain Ratio value is skewed toward splits where the size of one subset is close to that of the starting one. The algorithm C5.0 is to split the sub-sample which is defined by the first split, then divide it again by another different field. One repeats this step until the sub-sample cannot be split. It would re-examine the lower level split in the end, then remove any sub-sample that does not contribute significantly to the value of the model.

The method of C5.0 is very robust for handling missing data and in a large number of input fields. It usually does not require a long training time to make estimates. In addition, C5.0 is far easier to use than any other similar package. It also provides the powerful boosting method to increase the accuracy of the classification. The tree structure of C5.0 is as follows: Figure 3.

The split rule in the classification and regression trees (CART) and C5.0 is different. Classification and regression trees (CART) are a technique that generates a binary decision tree. Entropy is used as a measure to choose the best splitting attribute and criterion. At each step, an exhaustive search is used to determine the best split, where best is defined by $\phi\left(s/t\right) = 2P_L P_R \sum_{j=1}^{m} \left|P\left(C_j|t_L\right) - P\left(C_j|t_R\right)\right|$, where t is the current node, s is each possible splitting attribute and criterion, and L and R are used to indicate the left and right sub-trees of the current node in the tree. P_L and P_R are the probability that a tuple in the training set will be on the left or right side of the tree. $P\left(C_j|t_L\right)$ and $P\left(C_j|t_R\right)$ is the probability that a tuple is in this class, C_j, and in the left or right sub-tree.

The advantages of CART are the same as for C5.0. The only difference is that the output field of CART can provide the numerical and character type, but C5.0 only provides the character type.

Research methodology

Data mining is frequently defined as finding hidden information in a database. It has great potential to help companies focus on the important information in their data warehouses (Margaret 2002). This research adopts the process of CRISP-DM (Cross-Industry Standard Process for Data Mining) to build the imputative model. The research steps are as follows:

Data understanding

Before applying the technique of data mining, data understanding is an indispensable step. The datasets are taken from the databases of the industry, commerce and services (ICS) census from Census Bureau, Directorate-General of Budget, and from Accounting & Statistics for this research purpose. First, the background of the ICS should be understood, i.e. the investigative purpose, manner, subject, contents and so forth.

Data preparation and data cleaning

When the amount of data in the database is large, the task at hand may not require all of the data points; or there may be specific data requirements under which certain data preprocessing is necessary. Under such circumstances, data preparation and cleaning is an inevitable step. Before constructing the model, all efforts need to be taken to make the data useful and qualitative, because proper data preparation and cleaning can increase the accuracy rate of prediction.

Sampling

The study took the sample from the ICS database as our research datasets. The research goals were to impute missing data. The sampling method used in this research was simple random sampling. Ten percent of the sample data were taken and treated as missing data; the other 90% were used for model building. The 10% that was treated as missing values would be compared against the predicted value to find the model's accuracy.

Modeling

After the first three steps, different imputative models for different research goals were used in accordance with different types of missing data. When the type of missing data is continuous, the regression model and neural network were used to build imputative models. And when the type of missing data is categorical, the logistic regression model, neural network, C5.0 and CART were used to build imputative models.

Evaluation

For the continuous missing data, this research uses the root mean square error (RMSE) to evaluate the two imputative models, which are regression and neural network. For the categorical missing data, this research uses the accuracy and classification table to evaluate the four imputative models, which are logistic regression, neural network, and decision tree (C5.0, CART) and select the one that possesses the quality of accuracy prediction. Different models that best suited different types of data were applied. If during the stage of model evaluation, the results do not make any satisfactory, the imputative model is modified and reconstructed until favorable results are achieved.

A case study

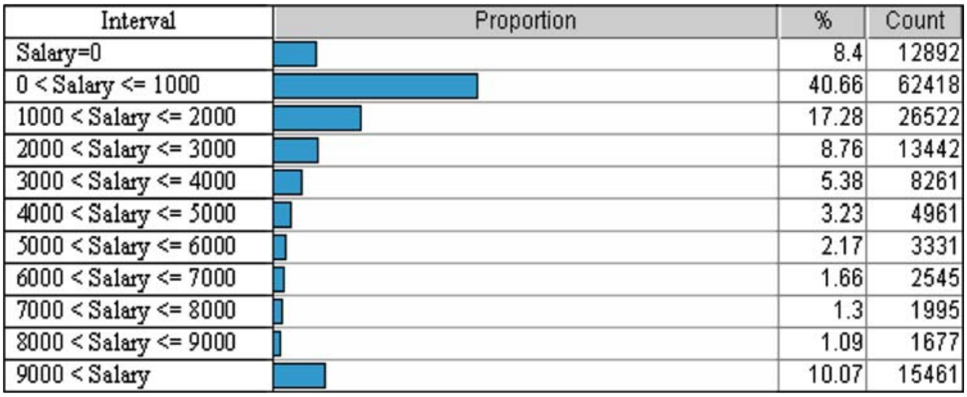
Data understanding

This research will take the database of the industry, commerce and services (ICS) census to provide its research datasets. The ICS census is a survey concerning national competitiveness conducted by the government on a regular basis in compliance with the Statistics Laws. The results of the census would be used for reference by the government in its further evaluating the performance of policy implementation and enacting the appropriate industrial and commercial policies.

The survey is carried out by means of field personnel interview. From 1st April to 15th July, 2002, identification of the census objects and examination of their basic operating data were drawn from more than 980,000 industrial and commercial firms throughout the country; and then statistical methods were used to select representative enterprises of a certain scale to conduct interview and form filling, so as to collect detailed data.

Data preparation

Due to the huge database of ICS, the task of data preprocessing is complicated. As such, this research will use only the manufacturing data of ICS database with 153,505 records for the research analysis. The principles for screening out the non-relevant variables from ICS database are as follows:



Interval	Proportion	%	Count
Salary=0		8.4	12892
0 < Salary <= 1000		40.66	62418
1000 < Salary <= 2000		17.28	26522
2000 < Salary <= 3000		8.76	13442
3000 < Salary <= 4000		5.38	8261
4000 < Salary <= 5000		3.23	4961
5000 < Salary <= 6000		2.17	3331
6000 < Salary <= 7000		1.66	2545
7000 < Salary <= 8000		1.3	1995
8000 < Salary <= 9000		1.09	1677
9000 < Salary		10.07	15461

Fig. 4 Distribution of the continuous dependent variable

1. Delete variables that are not used in the present paper.
2. Delete variables that are zero-value in all fields, i.e. value rate and profit rate.
3. Delete the ID variables.
4. Delete variables that have linear correlation. In other words, if a variable has a linear correlation with other variables, then delete this variable.

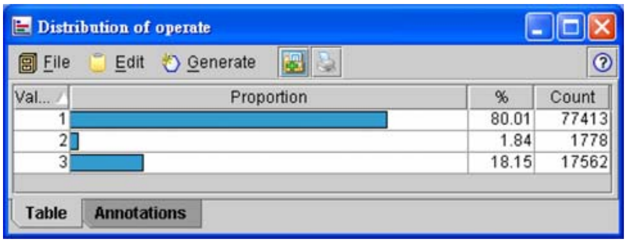
Using the above-mentioned principles, 49 variables out of 92 from the ICS database have been selected.

Modeling

The study used the imputative method to analyze the data from manufacturing for missing data. The model is built based upon the data type. Before building the model, the data structures are displayed by descriptive statistics that will help in constructing the model. The field named ‘salary’ has missing data of the continuous data type. The ‘salary’ in the database means that the salary and welfare in the manufacturing. We use a bar chart to present the structure of ‘salary’ (Fig. 4). It is found that 40.66% of salary + welfare is less than 1,000 dollars and only 10.07% of salary + welfare is more than 9,000 dollars. Worse yet, 8.4% of salary and welfare is zero dollar.

The field named ‘operate’ has the discrete missing data type. The ‘Operate’ in the database means the main manner of the operation in the manufacturing and has three elements, its distribution is very disproportionate (Fig. 5). The accuracy of the model is seemingly very high on the surface, but this result seemingly has some problems after observing the classification table. In order to avoid or reduce the influences of disproportionate data on the imputative model, a weight is assigned to the dependent variable to reduce the disproportionate gap among them.

After understanding the basic structure of the imputative field, five data mining techniques will be used to build the imputative model. Before building the model, 90% of the data were taken from the dataset of manufacturing at random as the sample model. Furthermore, the sample model is divided into 70% training data and 30% test data. And the other 10% of the data in manufacturing is regarded as missing data and is used as a benchmark to test for the accuracy of the predicted model.



Val...	Proportion	%	Count
1		80.01	77413
2		1.84	1778
3		18.15	17562

Fig. 5 Distribution of the categorical dependent variable

1. Training for the continuous fields

If the imputative data type is continuous, the regression model and neural network are adopted to estimate the imputed value. The continuous dependent variable is ‘Salary’ for the missing data. The independent variable is selected if the Pearson correlation between the independent variable and dependent variable is greater than 0.7 (Table 1).

Further, the independent variables that have high collinearity are dropped (John et al. 1998). In other words, the independent variables are dropped where the VIF is greater than 10 (Table 1). The dependent variable and independent variables of the continuous imputative model for missing data are shown below.

$$Y_i \equiv \text{Salary, where } i = 1, 2, \ldots, 96753$$

$$X_{i1} \equiv \text{ground; } X_{i2} \equiv \text{empno\_tatle; } X_{i3} \equiv \text{tax;}$$

$$X_{i4} \equiv \text{goods; } X_{i5} \equiv \text{ cash; } X_{i6} \equiv \text{house,}$$
$$\text{where } i = 1, 2, \ldots, 96753$$

After trying out the model, the accuracy and credibility of the best training model will be tested using the 30% test dataset.

2. Training for the categorical fields

If the data type of the imputative field is categorical, the logistic regression model, neural network, C5.0 and CART will be adopted to estimate the value. The dependent variable

Table 1 Pearson correlation and results of the regression

Variable	Pearson correlation	Coefficient	Sig.	VIF	Interpret variables
constant		−351.009	0.001		
ground	0.782	0.01342	0.000	1.336	The area of lands in use
empno_tatle	0.794	184.944	0.000	2.024	The total of worker force
tax	0.732	0.2196	0.000	2.758	The tax
goods	0.787	0.2961	0.000	4.775	Products and materials in stock
cash	0.735	0.03227	0.000	2.580	Cash and other Current assets
house	0.774	0.2867	0.000	2.731	House and building

of the imputative categorical field is 'Operate' for missing data. As a result of the disproportionate distribution of the dependent variable 'Operate', the weighted proportion will be assigned to the variables before modeling.

The independent variables are selected for modeling according to the importance of the dependent variables towards them using decision tree such as C5.0 and CART. When C5.0 and CART build the decision tree, they select the important variables 'first3_income', 'first4_income' and 'material cost'.

The dependent variable and independent variables of the categorical imputative model for missing data are as follows. This research will use these variables to train the model.

$Y_i \equiv$ Operate, where $i = 1, 2, \ldots, 96753$

$X_{i1} \equiv$ first3_income (the income of making repairs and supplying replacements);

$X_{i2} \equiv$ first4_income (the income of the expenses for processing);

$X_{i3} \equiv$ material cost (total value of raw materials consumed), where $i = 1, 2, \ldots, 96753$

After training the model, the study will test the accuracy and the credibility of the training model by using the other 30% test dataset.

Evaluation

1. Results of model training for continuous field

Seventy percent of the training dataset out of the ninety percent was used complete dataset to develop several models for the missing data. Determination of how well the model performs is made using root mean square error for continuous models and the accuracy rate for categorical models.

(1.1) Regression

This research analyzed the imputative model of the continuous missing data in the manufacturing by regression model. From the analytic results of the regression model (Table 1), the regression equation is

$$\begin{aligned} \text{salary} = {} & \text{ground} \times 0.01342 + \text{empno\_tatle} \times 184.944 \\ & + \text{tax} \times 0.2196 + \text{goods} \times 0.2961 + \text{cash} \\ & \times 0.03227 + \text{house} \times 0.2867 - 351.009 \end{aligned}$$

The tests of regression coefficients are all significant and there is no collinearity among the independent variables. From the coefficient of the regression model, it is clear that 'empno_tatle' has the highest coefficients signifying a good fit for the model. The coefficients of "ground" and "cash" are relatively low. Therefore, it was decided to remove these two variables from the equation. The impact of such action would be a reduction of the coefficient of determination by 10% points from the original of 93.5% to 83.8%. Because this research respects the importance of the predicted value in explaining the variation in the actual value, it is better to sacrifice two degrees of freedom for the regression model, and adopt six independent variables to construct the model.

The next step is to determine how accurate the regression equation is at estimating the predicted values. RMSE was used to evaluate the difference between actual and predicted values in the case, the adjusted R^2 of the regression model is 0.935 with a linear correlation of 0.967, and RMSE is 31323.206 (Table 2). This indicates that the linear correlation of the dependent variable and the independent variables is quite high and that the regression equation is quite a good predictor of the estimated value. And the independent variables can reflect the 93.5% variation of the dependent variable.

(1.2) Neural network

The study uses the manufacturing data to construct an imputative model with one hidden layer. After trying all methods of neural network, the BBNP method of neural network produces the best result. Therefore, this research would settle for applying this method. The linear correlation is only 0.427. This means that the linear correlation between the independent variable and dependent variables is not strong. RMSE was used to evaluate the validity of the model. There are quite a few differences between predicted values and actual ones; RMSE is 111586.933 (Table 2).

Table 2 Evaluated results of the regression and the neural network

Model	Regression		Neural network	
Parameter	RMSE	Linear correlation	RMSE	Linear correlation
Training	31323.206	0.967	111586.933	0.427
Testing	35040.592	0.957	107278.152	0.47

2. Results of model testing for continuous field

The other 30% of test dataset out of the 90% of the complete dataset were used to test for the stability of the selected training model. The test results for both the training and test models by regression and neural network do not differ much from each other, signifying the good fit of the model (Table 2).

3. Results of model training for categorical field

The categorical imputative model is constructed by applying the logistic regression, neural network, C5.0 and CART on the categorical variables of missing data from manufacturing.

(3.1) Logistic regression

By the results of the model, the logistic regression equation for operate = 1 is

$$\pi(\mathrm{X}) = \Pr(\mathrm{Y} = 1|\mathrm{X}) = \frac{\exp(0.787 + 0.002 \times \text{first3\_income} - 0.005 \times \text{first4\_income} + 0.008 \times \text{material\_cos t})}{1 + \exp(0.787 + 0.002 \times \text{first3\_income} - 0.005 \times \text{first4\_income} + 0.008 \times \text{material\_cos t})},$$

since the regression coefficient of ‘material cost’ equals zero which means that the corresponding explanatory variable ‘material cost’ is not associated with the occurrence of the dependent variable (Alan 1996), one would not consider the explanatory variable ‘material cost’ for the final logistic regression model. Thus, the logistic regression equation for operate = 2 is

$$\pi(\mathrm{X}) = \Pr(\mathrm{Y} = 2|\mathrm{X}) = \frac{\exp(0.399 + 0.007 \times \text{first3\_income} - 0.025 \times \text{first4\_income})}{1 + \exp(0.399 + 0.007 \times \text{first3\_income} - 0.025 \times \text{first4\_income})}$$

It excludes the situation in which the coefficient of independent variables ‘material cost’ is equal to zero; when the dependent variable ‘operate’ is equal to 2, the regression coefficients of the other independent variables are all significant except ‘material cost’ (Table 3).

The accuracy rate of prediction for the logistic regression model as a whole is 97.47% (Table 4) with an error rate of 2.53%; the error rates of the three categories are 0.48%, 4.33% and 2.78% respectively (Table 5). It shows that whether in the whole or individual classification, the overall result is properly accepted.

(3.2) Neural network

Manufacturing data are used to construct the imputative model with one hidden layer. The relative importance between the neuron of the input layer and output layer is as follows (Table 6).

The accuracy rate of prediction for neural network model as a whole is 96.05% (Table 4) with an error rate of 3.95%; the error rates of the three categories are 3.44%, 4.27% and 4.13% respectively (Table 5). It shows that whether in the whole or individual classification, the overall result is properly accepted.

(3.3) C5.0

Manufacturing data are used to construct the imputative model by the C5.0 decision tree method. The accuracy rate of prediction for C5.0 model as a whole is 97.64% (Table 4) with an error rate of 2.36%; the error rates of three categories are 0.41%, 4.11% and 2.56% respectively (Table 5). It shows that whether in the whole or individual classification, the overall result is properly accepted.

(3.4) CART

The accuracy rate of the CART model is 96.38% (Table 4) with an error rate of 3.62%; the error rates of three categories are 3.81%, 4.16% and 2.86% respectively (Table 5). It shows that whether in the whole or individual classification, the overall result is properly accepted.

4. Results of model testing for categorical field

The test results for both of the training and test models by logistic regression, neural network, C5.0 and CART do not differ much between them, signifying a good fit for the model. The results of the training and test model are quite consistent with one another as shown in Tables 4 and 5, showing the validity of the training model.

5. Results of the simulation for missing data

This research adopts a model that takes into consideration for the suitability of the data type. The simulation is carried out for every model for 20 iterations. RMSE was used as an accuracy measure for both of the regression and neural network model. The accuracy rate was used as accuracy measure for the four models. The results and the multiple plot are given below (Fig. 6, Table 7).

Table 3 Results of the logistic regression

Operate		Coefficient(B)	SE	Wald	Sig.	Exp(B)
1	Intercept	0.787	0.017	2081.165	0.000	
	First3_income	0.002	0.000	42.571	0.000	1.002
	First4_income	−0.005	0.000	15747.374	0.000	0.995
	Material_cost	0.008	0.000	12654.847	0.000	1.005
2	Intercept	0.399	0.020	397.257	0.000	
	First3_income	0.007	0.000	970.236	0.000	1.007
	First4_income	−0.025	0.008	10.079	0.001	0.975
	Material_cost	0.000	0.000	0.401	0.527	1.000

Table 4 Accuracy rate of the categorical training and testing model (%)

Model	Logistic regression	Neural network	C5.0	CART
Training	97.47	96.05	97.64	96.38
Testing	98.99	96.43	99.05	96.36

Table 5 Error rate of four models (%)

Model	Error rate	Logistic regression	Neural network	C5.0	CART
Training model	Categorical 1	0.48	3.44	0.41	3.81
	Categorical 2	4.33	4.27	4.11	4.16
	Categorical 3	2.78	4.13	2.56	2.86
Testing model	Categorical 1	0.51	3.41	0.47	3.8
	Categorical 2	3.99	3.99	3.84	3.87
	Categorical 3	2.90	4.20	2.77	2.95

Table 6 Relative importance in neural network

Input layer	Relative importance
First3_income	0.618225
First4_income	0.625505
Material_cost	0.488321

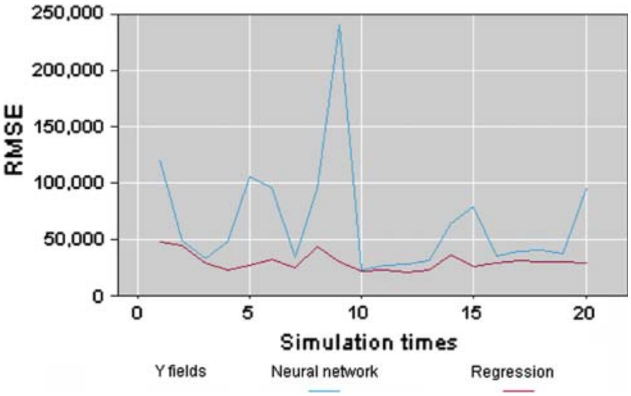


Fig. 6 Multiple plot of the continuous missing data

Conclusions and suggestions

Conclusions

The treatment of missing data is important in the application of statistical analysis. In this paper, several methods are applied to try imputing missing data. In accordance with the type of data, RMSE and the accuracy rate are used to evaluate for the performance of the model.

In the case of the application of regression and neural network model to missing data where the data are of continuous, the result from the simulation clearly indicates better model performance from regression. The RMSE values of regression seem to be quite stable compared to those of the neural network. Based on the above observation, it can reasonably be concluded that the regression model does a better job of imputing missing data than the neural network for the continuous data type.

For the categorical missing data, the accuracy rate of prediction from four methods (logistic regression, CART, C5.0 and neural network) is above 96%. The rate is even higher, up to 98%, for C5.0 and logistic regression. So all of the four models' predictive qualities are not bad, and the C5.0 model is the best method.

Table 7 Simulation results of the categorical missing data (accuracy rate, %)

Times	C5.0	Neural	CART	Logistic	Times	C5.0	Neural	CART	Logistic
1	99.08	96.30	96.24	98.40	11	99.17	96.66	96.52	98.63
2	98.99	96.67	96.62	98.52	12	99.04	96.47	96.37	98.44
3	98.97	96.09	96.07	98.30	13	98.97	96.34	96.2	98.41
4	99.02	96.56	96.44	98.46	14	99.11	96.60	96.54	98.56
5	99.13	96.69	96.47	98.54	15	99.05	96.48	96.37	98.5
6	99.14	96.52	96.46	98.56	16	99.14	96.37	96.26	98.65
7	99.17	96.68	96.58	98.70	17	99.00	96.58	96.49	98.48
8	99.11	96.48	96.32	98.57	18	99.12	96.52	96.40	98.54
9	98.97	96.49	96.38	98.45	19	99.20	96.72	96.59	98.63
10	99.03	96.45	96.25	98.49	20	98.98	96.31	96.20	98.46

Suggestions

Variable selection

There are many methods of selecting the variables. In general, the two most widely used are: all-possible-subsets regression and stepwise regression. Many researchers tend to choose stepwise regression method for its time-saving advantages. However there are some drawbacks worth discussions which are noted below.

1. Use of incorrect degrees of freedom
2. Cannot recognize the independent variables which make the best combination to the dependent variable.
3. The stepwise regression evaluates the importance of the independent variables by the order of the independent variable entering the model.

Although stepwise regression can be improved by the other method, it is still time-consuming and laborious. Huberty (1989) even proposes a substitute method for the stepwise regression with certain assumptions, as described below.

1. The variables in this research must be chosen circumspectly.
2. The numbers of variables in this research is not less than 30.
3. There are adequate reasons or the support of theory for the deletion of some variables.
4. The researcher is dealing with the data by the software of the statistic.

In view of this, the present research suggests the use of correlation for selecting the independent variables. If the correlation is greater than 0.7, then this represents that the coefficient of determination of the models will be above 50%. And if the correlation of the models is greater than 0.95, the coefficient of determination of the models will be above 90%.

Table 8 All kinds of the correlation

Type of data	Correlation coefficient
Continuous versus continuous	Pearson's r
Continuous versus categorical (ordinal)	Jaspen's multiserial coefficient (M)
Continuous versus categorical (nominal)	Eta
Categorical (ordinal) versus categorical (ordinal)	Spearman's r (Rho) or Kendall's t (Tau)
Categorical (nominal) versus categorical (ordinal)	Somers'd
Categorical (nominal) versus categorical (nominal)	Lamda

Using the correlation to select independent variables, it can achieve time efficiency and it can guarantee the coefficient of determination to be in or above a certain range. After selecting the independent variables by the correlation, one theoretically has adequate support in deleting redundant variables from the equation so as to maximize the accuracy of the model prediction. Because of there being different types of the missing data, the calculation of the correlation has to differ accordingly, as shown above (Table 8).

The model process

General speaking, the handling of missing data is on a case-by-case basis. There is no one definite approach as to how to handle missing data. This research suggests the application of CRISP-DM for building the model imputing missing data in the database gives the best value. CRISP-DM has been widely accepted for its practical application among many researchers on data mining topics. In the process of model building, the same path is followed as in the real world application of data mining techniques, in which the dataset is divided into a 70:30 proportion. The 70% of the data is assigned as training data; the other 30% is to provide test data. Training data are used to build the model. Once the model is obtained, it is tested for the accuracy of the model using 30% test set. If the results from both the training and testing model are consistent with one another, then the model is applied to impute the missing data.

Future work

The methods of handling missing data are directly related to the mechanisms that caused the incompleteness. Generally, these mechanisms fall into three classes (Little and Rubin 2002): (i) Missing completely at random (MCAR): The missing values in a variable are unrelated to the values of any other variables, whether missing or valid; (ii) Non-ignorable missing data (NIM): NIM can be considered as the opposite of MCAR in the sense that the probability of having missing values in a variable depends on the variable itself (for example a question regarding skills may be not answered when the skills are in fact low); (iii) Missing at random (MAR): MAR can be considered as an intermediate situation between MCAR and NIM. The probability of having missing values, does not depend on the variable itself but on the values of some other variable. Future work focuses on three main areas as follows: (i) to identify the mechanism behind the missing data; (ii) to evaluate the methods of handling missing data in accordance with different mechanisms (MCAR,NIM,MAR); and (iii) to apply the multiple imputations method to estimate the variance of imputing missing data.

Acknowledgements We wish to thank the anonymous referees for their valuable comments which helped us to improve the paper.

References

Alan, A. (1996). *An introduction to categorical data analysis*. Wiley Interscience.

Craven, M. P. (1997). A faster learning neural network classifier using selective backpropagation. In *Proceedings of the fourth IEEE international conference on electronics, circuits and systems*. Cairo, Egypt, 1, 254–258.

Ford, B. L. (1983). *An overview of hot-deck procedures*, In W. G. Madow, I. Olkin, & D. B. Rubin (Eds.) *Incomplete data in sample surveys. Volume 2. Theory and Bibliographies* (pp. 185–207). New York, NY: Academic Press.

Friedman, J. H. (1997). A recursive partitioning decision rule for nonparametric classifiers. *IEEE Transactions on Computers, 26*, 404–408.

Huberty, C. J. (1989). Problems with stepwise methods—better alternatives. In B. Thompson (Ed.), *Advances in social science methodology*, (Vol. 1, pp. 43–70). Greenwich: JAI Press Inc.

John, O. R., Sastry G. P., & David, A. D. (1998). *Applied regression analysis—a research tool*, 2nd ed. Springer.

Joop, J. H. (1999). A review of current software for handing missing data. *Kwantitatieve Methoden, 62*, 123–138.

Judi, S. (2002). Dealing with missing data. *Research Letters in the Information and Mathematical Sciences, 3*, 153–160.

Kalton, G., & Kasprzyk, D. (1982). Imputing for missing survey responses. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 22–23.

Kalton, G., & Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology, 12*(1), 1–16.

Lessler, J. T., & Kalsbeek, W. D. (1992). *Nonsampling error in surveys*. New York: John Wiley & Sons, Inc.

Li, J. R., Khoo, L. P., & Tor, S. B. (2006). RMINE: A rough set based data mining prototype for the reasoning of incomplete data in condition-based fault diagnosis. *Journal of Intelligent Manufacturing, 17*, 163–176.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: John Wiley & Sons.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*, 2nd ed. New York: John Wiley & Sons.

Margaret, H. D. (2002). *Data mining—introductory and advanced topics*. Prentice Hall.

Robert, E. F. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association, 91*(434), 490–498.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley & Sons.

Werbos, P. (1974). *Beyond regression: New tools for prediction and analysis in the behavioral sciences*. Harvard University.