ESTIMATING THE MISSING VALUES IN ANALYSIS OF VARIANCE TABLES BY A FLEXIBLE ADAPTIVE ARTIFICIAL NEURAL NETWORK AND FUZZY REGRESSION MODELS

Ali Azadeh - Zahra Saberi –Hamidreza Behrouznia-Farzad Radmehr
Peiman Pazhoheshfar

Abstract

Missing data are a part of almost all research, and we all have to decide how to deal with it from time to time. There are a number of alternative ways of dealing with missing data. The problem of handling missing data has been treated adequately in various real world data sets. Several statistical methods have been developed since the early 1970s, when the manipulation of complicated numerical calculations became feasible with the advance of computers. The purpose of this research is to estimate missing values by using artificial neural network (ANN), fuzzy regression models, and approach in a complete randomized block design table (analysis of variance) and to compare the computational results with two other methods, namely the approximate analysis and exact regression method. It is concluded that ANN provides much better estimation than the conventional approaches. The superiority of ANN is shown through lower error estimations.

Key words: Missing Values; artificial neural network; fuzzy Regression; ANOVA.

JEL Code: C13,C45,C63.

Introduction

Missing data are a part of almost all research, and we all have to decide how to deal with it from time to time. Data values may be absent from a dataset for numerous reasons, for example, the inability to measure certain attributes (Van Hulse & Khoshgoftaar, 2008). In such cases, the most popular and simple method of handling missing data is to ignore either the projects or the attributes with missing observations. This technique causes the loss of valuable information and therefore may lead to inaccurate cost estimation models. Gad and Ahmed (2006), claimed that ignoring the missing values in this case leads to biased inferences. Moreover, when an attribute contains a missing value in a test case, it may or may not be worthwhile to take the extra effort in order to obtain a value for that attribute(s) (Yang et al., 2006). There are a number of alternative ways of dealing with missing data.

The structure of the paper is as follows: In section 1, the related work and the literature in this area is outlined. In Section 2, the different mechanisms used to create missing data and the most common techniques for handling them are described. Section 3 explains the general outline of applied algorithms. In section 4, is designated to the computational results, while final section is a conclusion to this research.

1 Literature Review

The problem of handling missing data has been treated adequately in various real world data sets. Several statistical methods have been developed since the early 1970s, when the manipulation of complicated numerical calculations became feasible with the advant of computers. Some of the most important review papers on the subject are Afifi and Elashoff (1966), , Little and Rubin (1983).

In the field of software engineering there are rather few published works concerning missing data. In Song and Shepperd’s study (2003), two imputation methods, class mean imputation (CMI) and k-nearest neighbors (k-NN), were considered with respect to two mechanisms of creating missing data: missing completely at random (MCAR) and missing at random (MAR). Bashir et al. (2006) introduced a novel partial matching concept in association rules mining to improve the accuracy of missing values imputation. Their imputation technique combined the partial matching concept in association rules with k-nearest neighbor approach.

In Reis and Saraiva’s study (2006), the problem of extending the multi-scale decomposition framework based upon the wavelet, transformed to situations where datasets contain any type of missing data patterns (e.g., random, multirate). Their proposed approaches integrate data uncertainty information into their algorithms to explore all knowledge available about data during the decomposition stage. These frameworks, called generalized multi-resolution decomposition frameworks (GMRD), also lead to new developments in data-analysis tools based upon the information they provide.

Sehgal et al. (2008) presented an Ameliorative Missing Value Imputation (AMVI) technique which has ability to exploit global/local and positive/negative correlations in a given dataset by automatic selection of the optimal number of predictor genes k using a wrapper nonparametric method based on Monte Carlo simulations.

2 Missing Data Mechanism and Missing Data Techniques

2.1 Missing Data Mechanisms

The methods of handling missing data are directly related to the mechanisms that caused the incompleteness. Generally, these mechanisms fall into three classes (Little & Rubin, 2002):

Missing completely at random (MCAR): The missing values in a variable are unrelated to the values of any other variables, whether missing or valid.

Non-ignorable missingness (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).

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 variables. Formally, we denote the data matrix by $X = (x_{ij}), i=1,. . . ,n , j=1,. . . ,k$, where x_{ij} is the value of variable x_j for case i. When there are missing data, we define the missing-data indicator matrix $M =(m_{ij})$ where $m_{ij}=1$ if x_{ij} is missing and $m_{ij} =0$ if x_{ij} is present. The matrix M defines the pattern of the missing data. Moreover, the of unknown parameters and missingness do not depend on X, i.e., $f(M/X,\theta)=f(M/\theta)$; then the mechanism is MCAR.

It is therefore necessary, in order to identify the mechanism behind the missing data, to perform statistical analysis of the available data and make inferences for the distributions of the individual variables of X and search for dependencies of the matrix M on these distributions of X. Unfortunately, this is a very difficult task in a real data set where usually there is no prior information on any distribution. A reasonable approach and a common practice in order to test various methods for incomplete data sets is to generate artificial missing values from complete databases. In this way we can check the robustness of the proposed methods under different missingness situations.

2.2 Missing Data Techniques (MIDTs)

The techniques used in this paper for handling missing data and for comparisons with MLR are (Little and Rubin, 2002):

Listwise deletion (LD): It is a typical method that belongs to a broader class, namely the deletion methods. According to LD, cases with missing values for any of the variables are omitted from the analysis. The procedure is quite common in practice because of its simplicity, but when the percentage of missing values is high, it results in a small complete subset of the initial data sets and therefore in difficulties in constructing a valid cost model.

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.

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.

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.

Multinomial logistic regression (MLR): This method is a generalization of the logistic regression (LR), which is used to model the relationship between a dichotomous (binary) dependent variable and a set of k predictor variables *{*x_1*,*x_2*, . . . ,*x_k*}*, which are either categorical (factors) or numerical (covariates). As the binary dependent variable can be always interpreted as the occurrence or not of an event *E*, the logistic regression model is an expression of the form:

$$\log\left(\frac{prob(E)}{1-prob(E)}\right)=b_0+\sum_{i=1}^{k}b_i x_i \qquad (1)$$

Where the b_i denotes the unknown logistic regression coefficients (b_0 is the intercept) while *prob(E)* denotes the probability that event *E* will occur.

2.2.1 Maximum Likehood

The principle of maximum likelihood is fairly simple, but the actual solution is computationally complex. There are a number of ways to obtain maximum likelihood estimators, and one of the most common is called the Expectation-Maximization algorithm, abbreviated as the EM algorithm. The basic idea is simple enough, but the calculation is complicated.

In order to solve the EM algorithm by hand, we would first take estimates of the variances, covariances and means, perhaps from listwise deletion. We would then use those estimates to solve for the regression coefficients, and then estimate missing data based on those regression coefficients. (For example, we would use whatever data we have to estimate $Y = bX + a$, and then use X to estimate Y wherever it is missing.) This is the "estimation step of the algorithm. Having filled in missing data with these estimates, we would then use the complete data (including estimated values) to recalculate the regression coefficients. (The new estimates would be adjusted to model sampling error, but that is a technical issue.) This is the "maximization" step. Having new regression coefficients, we would re-estimate the missing data, calculate new coefficients, etc. We would continue this process until the estimates no longer change noticeably. At that point we have maximum likelihood estimates of the parameters, and we can use those to make the maximum likelihood estimates of the regression coefficients.

The solution from the EM algorithm is better than we can do with coding for missing data, but it will still underestimate the standard errors of the coefficients. There are alternative maximum likelihood estimators that will be better than the ones obtained by the EM algorithm, but they assume that we have an underlying model (usually the multivariate normal distribution) for the distribution of variables with missing data.

2.2.2 Multiple Imputation

An alternative the maximum likelihood is called Multiple Imputation. Each of the solutions discussed involves estimating what the missing values would be, and using those "imputed" values in the solution. With dummy variable coding, a constant is substituted (often the variable mean) for the missing data. For the EM algorithm, a predicted value on the basis of the variables that were available is substituted. In multiple imputation, we will substitute random data.

In multiple imputation we generate imputed values on the basis of existing data, just as we did with the EM algorithm. But suppose that we are estimating Y on the basis of X. For every situation with X=5, for example, we will impute the same value of Y. This leads to an underestimate of the standard error of our regression coefficients, because we have less variability in our imputed data than we would have had if those values had not been missing. One solution was the one used in the EM algorithm, where we altered the calculational formulae by adding error in the calculation. With multiple imputation we are going to take our predicted values of Y and then add, or subtract, an error component drawn randomly from the residual distribution of $Y - \hat{Y}$. This is known as a "random imputation".

This solution will still underestimate the standard errors. This problem can be solved by repeating the imputation problem several times, generating multiple sets of new data whose coefficients varying from set to set. This variability can then be captured and added back into the estimates.

3 Applied algorithms for estimating missing values in ANOVA tables

3.1 Artificial neural network (ANN)

In general, ANNs are simply mathematical techniques designed to accomplish a variety of tasks. ANNs consists of an inter-connection of a number of neurons. There are many varieties of connections under study, however here we will discuss only one type of network which is called the Multi Layer Perceptron (MLP). In this network, the data flow forward to the output continuously without any feedback. The model can be written in (2):

$$y_t = \alpha_0 + \sum_{j=1}^{n} \alpha_j f\left(\sum_{i=1}^{m} \beta_{ij} y_{t-i} + \beta_{0j}\right) + \varepsilon_t \qquad (2)$$

where m is the number of input nodes, n is the number of hidden nodes, f is a sigmoid transfer function such as the logistic: $f(x) = \frac{1}{1+\exp(-x)} \{\alpha_j, j = 0,1,...,n\}$ is a vector of weights from the hidden to output nodes and $\{\beta_{ij}, i = 1,2,...,m; j = 0,1,...,...,n\}$ are weights from the input to hidden nodes. α_0 and β_{0j} are weights of arcs leading from the bias terms which have values always equal to 1. Note that Eq. (2) which indicates a linear transfer function is employed in the output node as desired for forecasting problems.. At the beginning of the learning, stage all weights in the network are initialized to small random values. The algorithm uses a learning set, which consists of input – desired output pattern pairs. Each input – output pair is obtained by the offline processing of historical data. These pairs are used to adjust the weights in the network to minimize the Sum Squared Error (SSE), which measures the difference between the real and the desired values overall output neurons and all learning patterns. After computing SSE, the back propagation step computes the corrections to be applied to the weights.

We present the network with training examples, which consist of a pattern of activities for the input units together with the desired pattern of activities for the output units. For this reason, each ANN uses a set of training rules that define training method. Generalization or testing evaluates network ability in order to extract a feasible solution when the inputs are unknown to network and are not trained to network. We determine how closely the actual output of the network matches the desired output in new situations. In the learning process, the values of interconnection weights are adjusted so that the network produces a better approximation of the desired output.

3.2 Fuzzy Regression

Conventional regression analysis is one of the most used statistical tools to explain the variation of a dependent variable Y in terms of the variation of explanatory variables X as: $Y = f(X)$ where $f(X)$ is a linear function. It refers to a set of methods by which estimates are made for the model parameters from the knowledge of the values of a given input–output data set. The goal of the conventional regression analysis is:

(a) to find an appropriate mathematical model and

(b) to determine the best fitting coefficients of the model from the given data.

The use of statistical conventional regression is bounded by some strict assumptions about the given data. This model can be applied only if the given data are distributed according to a statistical model and the relation between X and Y is crisp. Overcoming such limitations, fuzzy regression is introduced which is an extension of the conventional regression and is used in estimating the relationships among variables where the available data are very limited and imprecise and variables are interacting in an uncertain, qualitative and fuzzy way.

4 Computational results

In this section, we propose the ANOVA table .The missing values have been estimated by artificial neural network (ANN), fuzzy regression models, approximate and exact methods and for comparison with each others; their MAPEs (mean absolute percentage error) have been reported. The number of hidden neurons has been considered 10. The computational results of methods and their errors are shown in Table 1.

Tab. 1: The normal ANOVA table and compression of approaches

7	7	15	11	9
12	17	12	**18**	18
14	18	**18**	19	**19**

19	25	22	19	23
7	10	11	15	11

Tab. 1: The normal ANOVA tables and compression of approaches (Continue)

Actual	7	MAPE	18	MAPE	18	MAPE	19	MAPE
approximate	6.3	0.1	18.25	0.01	16.1	0.11	18.3	0.04
exact	6.3	0.1	18.3	0.02	16.1	0.11	18.3	0.04
ANN	7.0143	0.002	18.1031	0.0057	18.0008	4.3507×10^{-5}	19.013	6.8183×10^{-4}
hojati	5.47	0.22	16.41	0.088	13.84	0.23	19.31	0.016
ozelkan	5.33	0.24	16	0.11	13.94	0.23	19.28	0.015
peters	6.125	0.125	18.375	0.02	18	0	24.125	0.27
Tanaka 1989	7.75	0.107	23.25	0.29	27.17	0.509	34.92	0.84
Tanaka 1982	7.75	0.107	23.25	0.29	20.42	0.13	28.17	0.48
Chang	5.35	0.236	16.05	0.108	16.2	0.1	21.55	0.134
Lee	7.75	0.107	23.25	0.292	23.27	0.293	31.02	0.633
Peters 2	4.11	0.414	12.31	0.316	11.996	0.334	16.099	0.153
Wang	5.09	0.272	15.28	0.15	15.396	0.14	20.49	0.078
Wangtsaur	7.75	0.107	23.25	0.292	24.97	0.387	32.72	0.72

Conclusion

The problem of handling missing data has been treated adequately in various real world data sets. Several statistical methods have been developed since the early 1970s, when the manipulation of complicated numerical calculations became feasible with the advance of computers. In this research, the missing values are being estimated using a flexible artificial neural network (ANN), fuzzy regression models in randomized block design tables, and the computational results are compared with two other methods namely the regression method. Computational results indicated that the best answer obtained by ANN is frequently the same as the missing value, with the mean value being close to the missing observation too. We presented the MAPE (mean absolute percentage error), in each approach, therefore we studied approaches in each ANOVA table. The MAPE can indicate the accuracy of our proposed flexible ANN and other presented methods. With comparing the results, flexible artificial neural network could produce the best results.

References

Afifi A. A. ,Elashoff R. M., Missing observations in multivariate statistics: review of the literature, *Journal of the American Statistical Association* 61(315), 1966, 595-604.

Bashir Sh. , Razzaq S. , Maqbool U. , Tahir S., Baig A. R. , Introducing partial matching approach in association rules for better treatment of missing values, *WSEAS Transactions on Computers* 5(10), 2006, 2388-2393.

Gad A. M., Ahmed A. S. , Analysis of longitudinal data with intermittent missing values using the stochastic EM algorithm, *Computational Statistics and Data Analysis* 50(10), 2006, 2702-2714.

Hulse J. Van , Khoshgoftaar T.M., , A comprehensive empirical evaluation of missing value imputation in noisy software measurement data, *The Journal of Systems & Software* 81(5), 2008, 691-708.

Little R. J. A. , Rubin D. B., Incomplete data, *Encyclopedia of Statistical Sciences* 4, 1983, 46-53.

Little R. J. A. , Rubin D. B., Statistical analysis with missing data, seconded, Wiley, New York, 2002.

Reis M. S. ,Saraiva P. M., Generalized multiresolution decomposition frameworks for the analysis of industrial data with uncertainty and missing values, *Industrial and Engineering Chemistry Research* 45(18), 2006, 6330-6338.

Sehgal M. Sh. B., Dooley L. S., and Coppel R., Ameliorative missing value imputation for robust biological knowledge inference, *Journal of Biomedical Informatics* 41(4), 2008, 499-514.

Song Q. , Shepperd M., A short note on safest default missingness mechanism assumptions, *ESERG Technical Report TR02-07,* Bournemouth University, 2003.

Yang Q., Ling C., Chai X., and Pan R., Test-cost sensitive classification on data with missing values, *IEEE Transactions on Knowledge and Data Engineering* 18(5), 2006**,** 626-638.

Contact

Ali Azadeh

Department of Industrial Engineering, Center of Excellence for Intelligent-Based
Experimental Mechanics, College of Engineering, University of Tehran , P.O. Box 11365-
4563, Tehran, Iran
Aazadeh@ut.ac.ir

Zahra Saberi

Department of Industrial Engineering, Amirkabir University of Technology, Tehran, Iran

Zahra.saberi.aut@gmail.com

Hamidreza Behrouznia

Young Researchers Club, Tafresh Branch, Islamic Azad University, Tafresh, Iran

Moallem Square,Moien Abad st. Islamic Azad University, P.O. Box 39515-164, Tafresh, Iran
hamidrezabehroznia@gmail.com

Farzad Radmehr

Department of Industrial Engineering, Center of Excellence for Intelligent-Based
Experimental Mechanics, College of Engineering, University of Tehran , P.O. Box 11365-
4563, Tehran, Iran
Probit_mor@yahoo.com

Peiman Pazhoheshfar

Young Researchers Club, Tafresh Branch, Islamic Azad University, Tafresh, Iran

Moallem Square,Moien Abad st. Islamic Azad University, P.O. Box 39515-164, Tafresh, Iran
p.pazhohesh@gmail.com