World Applied Sciences Journal 9 (5): 561-566, 2010
ISSN 1818-4952

An Application of Expectation and Maximization, Multiple Imputation and Neural Network Methods for Missing Value

[1]Yılmaz Kaya, [2]Abdullah Yeşilova and [3]M. Nuri Almali

[1]Department of Computer Programming, Vocational High School, Yuzuncu Yil University, Van, Turkey
[2]Department of Animal Science, Faculty of Agricultural, Biometry and Genetic Unit,
Yuzuncu Yil University, Van, Turkey
[3]Department of Electrics and Electronics Engineering Unit, Faculty of Engineering and Architecture,
Yuzuncu Yil University, Van, Turkey

Abstract: In order to ensure reliable results in research, it is crucial that the data obtained is complete and accurate. In this respect, it is expected that there are not any missing values in the data set. However, due to various reasons, a certain part of the study might not be observed and certain data can be missing. The missing values can be associated with one variable or more, as well. Expectation maximization, multiple imputation and neural networks are different models for estimation missing values in data. This study aimed to examine the success of expectation and maximization, multiple imputation and neural network methods in estimation of missing values in cases where data are missing at random. The analyses have proven that neural network method was more successful in the estimation of missing values compared to multiple imputation and expectation and maximization methods. The standard errors and correlation coefficients of missing values by neural network method produced much closer results to the standard errors of real values compared to other methods.

Key words: Missing Value · Expectation and Maximization · Multiple Imputation · Neural Network

INTRODUCTION

Missing values is a frequently encountered problem in statistical data analysis. Completeness of data is vital for achievement of accurate results in research. Missing values can be associated with one variable or more, as well. Additionally, stemming from various reasons, a certain part of the study might not be observed. In such case and certain data can be missing. Missing values can generate problems of inaccuracy and biased estimations based on statistical analyses. In some studies, missing values are excluded from the analyses. Nevertheless, the exclusion of missing values from analysis might lead to biased parameter estimations [1,2].

Different methods have been developed for estimation of missing values. A couple of these are expectation and maximization (EM), multiple imputation (MI) and neural network (NN) methods. The appropriateness of the method depends on the type of nonresponse mechanism In literature, there are three types of nonresponse mechanism. These can be listed as missing completely at random (MCAR), missing at random (MAR) and non ignorable (NI) [1,3,4].

In "missing completely at random", the missing value for X is unrelated to any other variables or X, itself [5]. It is assumed that the missing values in the data set occurred completely randomly. In "missing at random", the missing value for X is dependent on other variables in the data set other than the variable, X. In "nonignorable", the missing value for X is dependent on all the variables in the data set including X. In other words, occurrence of missing value is not random [6,7, 8].

EM algorithm is a method used for finding maximum likelihood estimates of missing values. EM is a two-step process in which E step provides maximum likelihood estimates of the missing value and M step provides estimates on mean, standard deviation or correlation in case of imputation of missing values [9]. These two-step iterations continue until where the differences between the estimated values are insignificant [9, 10]. In other

Corresponding Author: Yuzuncu Yil University, Faculty of Agricultural, Department of Animal Science, Biometry and Genetic Unit, Van, Turkey. E-Mail: yesilova@yyu.edu.tr

words, iteration continues until the convergence criterion is satisfied.

Multiple imputation is an important method for estimation of missing values when dealing with data set with missing values. Instead of filling in a single value for each missing value, MI procedure replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute [11]. The MI procedure consists of three stages for imputing missing values. In the first phase, the missing data are filled in m times to generate m complete data sets. In the second phase, the m complete data sets are analyzed by using standard procudures. In the third phase, the results from the m complete data sets are combined for the inference [11].

Neural network method is based on the findings of biological nervous system. In this artificial nervous system, there are nerve cells which are joined together in a variety of ways to form networks. NN consists of three layers, namely input layer, inter layer (hidden layer) and output layer [12, 13]. The input layer receives data from external world. The output layer presents the data to the user. The hidden layer in-between these two layers is where the data are processed. The number of the nerve cells in the hidden layer is significant for the performance as well as the length of the network.

MLP (multilayer perceptron) neural networks consist of multiple layers of computational units, usually interconnected in a feed-forward way Each neuron in one layer is directly connected to the neurons of the subsequent layer. MLP networks apply different learning techniques, the most popular being back-propagation [14, 15].

The goal of this study is to assess the efficiency of expection-maximization, multiple imputation and neural network methods in estimation of missing values in case of missing completely at random.

MATERIALS AND METHODS

Data Set: Data set used in this study was composed of physical features of 155 students attended in special ability exam of the Department of Physical Education and Sports, Faculty of Education, Yüzüncü Yil University in 2008. According to measurements before the exam data on the subjects' age, height, weight, foot size, sports branch, their history in sports, how they use their hands, smoking habits, the number of sit-ups before the exam, time of preparing for the exam, how many times they attended the exam and the number of sit-ups in the exam were gained. Data gained and their codes are shown in the following.

Table 1: Data set

Code	Variables
S1	Age
S2	Height
S3	Weight
S4	foot size
S5	their history in sports
S6	The number of sit-ups before the exam
S7	time of preparing for the exam
S8	how many times they attended the exam
S9	The number of sit-ups in the exam

Methods

Expectation and Maximization (EM): Expectation and maximization algorithm is estimated estimation of mean, covariance matrix and correlation of quantitative variables that include the missing values by the use of an iterative process [16]. In EM, estimations of missing values, mean vector and covariance matrix are calculated respectively based on the equation 1 [9].

$$x0 = \left[\bar{x}_j^0\right] = Diag(\bar{x}^P) = \left[\bar{x}_{JJ}^P\right]$$
$$C0 = \left[c_{jk}^0\right] = C^P = \left[c_{jk}^P\right] \quad (1)$$

In Equation 1, j and k indises display quantitative variables. In this equation m=1 to M or until the convergence provided. if x_{ij} is not completed, then

$$x_{ij}^m = x_{ij}$$

If x_{ij} is missing, the estimation value provided in m. iteration outcome is calculated as following,

$$x_{ij}^m = \beta_{0,ij}^{m-1} + \sum_1 \beta_{1,ij}^{m-1} * x_{i1};$$
$$1 \in J_2 = J(1 : x_{i1} \text{ if not missing and } 1 \neq j) \quad (2)$$

where $\left[\beta_{0,ij}^{m-1}, \beta_{1,ij}^{m-1}\right]$ and C_{m-1} is calculated using $\bar{x}_{m-1}$ and 1 indis displays expected variables. Besides $\bar{x}_m$ and C_m are calculated as following [9].

$$\bar{x}_m = \left[\bar{x}_j^m\right] = \left[\sum_i w_i * x_{ij}^m / \sum_i w_i; \qquad i \in I\right] \quad (3)$$

$$C_m = \left[c_{jk}^m\right] = \left[\frac{\sum_i w_i * x_{ij}^m \left(x_{ij}^m - \bar{x}_j^m\right) * \left(x_{ik}^m - \bar{x}_k^m\right) + \sum_i \sum_s c_{J,s|J2}^{m-1}}{(n-1) * \sum_i w_i / n}\right] \quad (4)$$

$$; i \in J_2, s \notin J_2 \; ve \; s \neq j$$

where $c_{J,s|J2}^{m-1}$ is member of J. lines member of of J_2 converted (pivoted) the C_{m-1} value. The w_i value is,

$$w_i = \begin{cases} 1 & \text{for multivariate normal} \\ 1-\alpha+\alpha*\lambda^{1+p/2}*\exp((1-\lambda)*D^2/2) & \text{for ruined normal} \\ 1-\alpha+\alpha*\lambda^{p/2}*\exp((1-\lambda)*D^2/2) & \\ (sd+p)/(sd+D^2) & \text{for t(sd) distrubution} \end{cases} \quad (5)$$

Where,

α is ratio of ruined, λ is ratio of standart deviation, p is number of independent variables, D^2 is....D^2 is the unit square of Mahalanobis distance from mean. D^2 is,

$$D2 = \sum_{jk}\left(x_{ij}^m - \overline{x}_j^m\right)*\left(c_{jk}^m\right)^{-1}* \left(x_{ik}^m - \overline{x}_k^m\right) \quad (6)$$

where, $\left(c_{jk}^m\right)^{-1}$ expression is the jk. of C_m^{-1} [17].

In order to end the EM algorithm, it is necessary to check it with definity convergence criterion. Convergence criterion used in providing suitable parameter value for imcomplate observation can be given as,

$$\left|c_{ij}^m - c_{ij}^{m-1}\right| / c_{ij}^m \leq \text{Converged Value}$$

it is looked fort he validity of convergence criterion. The iteration is stopped when the difference is much smaller then convergence criterion or equal to and the estimation parameter provided at the same moment is taken as suitable values [18, 9].

Data assigment given as,

$$X_i^E = \left[x_{ij}^E\right] = \left[x_{ij}^{m'}\right]$$

where m is the value provided from the last iteration of m. The mean and covariance are calculated as follows [17, 9, 19, 10].

$$\overline{x}^E = \left[\overline{x}_j^E\right] = \overline{x}_{m'} = \left[\overline{x}_j^{m'}\right] \quad (7)$$

$$C^E = \left[c_{jk}^E\right] = C_{m'} = \left[c_{jk}^{m'}\right]$$

$$R^E = \left[r_{jk}^E\right] = \left[c_{jk}^E / (c_{jj}^E * c_{kk}^E)^{1/2}\right] \quad (8)$$

Multiple Imputation: With m imputations, m different sets of the point and variance estimates for a parameter can be computed. $\hat{Q}$ and U are point and variance estimates [1]. The MI point estimate for Q is the avarage of the m-complate data. Q etimates can be written as follow [4],

$$\overline{Q} = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i \quad (9)$$

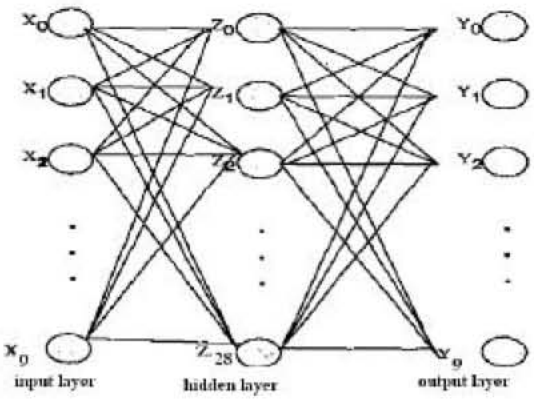

Fig. 1: Structure of 9-28-9 MLP

Let U be the within-imputation variance, which is the average of the m complete-data estimates. Therefore, U is,

$$\overline{U} = \frac{1}{m}\sum_{i=1}^{m}U_i \quad (10)$$

Between imputation variance (B) is,

$$B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{Q}_i - \overline{Q})^2 \quad (11)$$

Then the variance estimate associated with Q is the total variance,

$$T = \overline{U} + (1+\frac{1}{m})B \quad (12)$$

The statistic (Q - Q)T-1/2 is approximately distributed as a t-distribution with df degrees of freedom [5], Where,

$$df = (m-1)\left(1+\frac{m\overline{U}}{(m+1)B}\right)^2$$

95% confidence interval is,

$$\overline{Q} \pm t_{df}\sqrt{T} \quad (13)$$

where, t_{df} denotes a quantile of Student's t-distirbution with degrees of freedom.

Neural Network: When the relationship between variables is non-linear it is difficult to model and solve such kind of problems. It is needed to provide some assumptions for the solution. This causes difference between modeled system and real system. However NN provides better and more realistic solutions than common methods among non-linear relationship [20]. Compared to classical statistical approaches, NN can be used in the model and solution of more complicated problems.

In addition to this NN does not require any assumption compared to classical statistical models [14, 21, 22].

Neural Network is composed of hierarchically connected and parallel processing artificial nerve cells. Each cell is called as neuron, nerve or node. There are weight parameters representing the magnitude and the impact of the networks connecting the nerves. NN is composed by the connection of the nerves [12, 6, 13, 23].

Neural Network model is [14],

$$\vec{Y} = f(\vec{X},\vec{W}) \tag{14}$$

Equation 14, $\vec{Y}$ is output vector, $\vec{X}$ is input vector, $\vec{W}$ is weight vector. In NN, Between the expected value and the value it generated may be error. In applications, minimization of error is required. This error functions given as following,

$$e = (Y - f(\vec{X},\vec{W}))^2 \tag{15}$$

X data set having missing value are divided two gruops because of all the observations can not be used. So, $\vec{X}_k$ is known vector and $\vec{X}_u$ is unknown vector. Here, error term can be written as [1, 14],

$$e = \left(\begin{Bmatrix} \vec{Y}_k \\ \vec{Y}_u \end{Bmatrix} - f \begin{Bmatrix} \vec{X}_k \\ \vec{X}_u \end{Bmatrix}, \vec{W} \right)^2 \tag{16}$$

As a result of experiments carried out for NN model, the most suitable outputs are obtained by multilayer MLP (9:28:9) type NN. It means 9 nerve cells are used in input layer, 28 in hidden layer and 9 in output layer. By 10 times working the model which is found suitable due to aleatory appointment at the beginning of weight, the average of results is taken. Hyperbolic tangent function is used an activation function in hidden and output layers. In order to minimize the fault, back-propagation algorithm is used. This education method, namely standard back-propagation, aims to diminish error sum of squares. The learning value is chosen as 0.01and 500 iterations are carried out for the training of net.

RESULTS

The missing values were consist of randomly using the real data set. For randomly generating the missing values (Table 1), a program produced by Microsoft Visual Basic was used. Table 1 presents the missing values and their associated percentages for all variables. Missing values were not imputed only for S1 variable. Moreover, the highest number of missing value was produced for S9 variable.

Correlation between estimated values gained with EM, NN and MI methods and real values were given in Table 3.

In Table 2, a strong correlation between gained by using EM, NN and MI methods and values at the real data set was found out. Mean of variables and standard deviations in the real data set and in the data set missing values of which were completed by estimating with the use of EM, NN and MI were shown in Table 3.

As seen in Table 3, the missing values were estimated by using EM, NN, MI methods. While smaller mean values were obtained for the variables S2, S3, S4, S5, S9, higher mean values were imputed for the variables S6, S7, S8 with respect to those in the real data set. Similarly, the standard deviations of the data set with completed missing values for S2, S6 and S9 variables were smaller than the standard deviation of the real data set,

Table 1: Number of missing value and percentages for each variables

Variables	S1	S2	S3	S4	S5	S6	S7	S8	S9
The number of missing value	0	9	12	9	13	22	15	16	9
percentages %	0	5,8	7,7	5,8	8,3	14,1	9,6	10,3	5,8

Table 2: Correlation between estimated values gained with EM, NN and MI methods and real data set

Variables	EM	NN	MI
S1	1.000	0.982	1.000
S2	0.979	0.964	0.954
S3	0.932	0.958	0.934
S4	0.962	0.971	0.953
S5	0.961	0.975	0.923
S6	0.902	0.892	0.894
S7	0.923	0.958	0.923
S8	0.946	0.976	0.946
S9	0.962	0.964	0.974

Table 3: Mean and std. deviation between estimated values gained with EM, NN and MI methods and real data set

Variable	Mean and std. Deviation	EM	MI	NN	Real data set
S2	Mean	174.93	174.80	174.74	175.02
	Std. deviation	5.50	5.52	5.30	5.63
S3	Mean	64.55	64.98	64.94	65.0
	Std. deviation	7.95	7.96	6.82	7.60
S4	Mean	41.79	41.79	41.78	41.8
	Std. deviation	1.04	1.06	1.01	1.05
S5	Mean	5.12	5.08	5.23	5.71
	Std. deviation	3.16	3.31	3.02	3.13
S6	Mean	131.63	132.26	131.62	131.48
	Std. deviation	15.12	15.24	13.79	15.74
S7	Mean	2.59	2.61	3.23	2.50
	Std. deviation	1.35	1.37	1.01	1.34
S8	Mean	1.67	1.67	1.69	1.65
	Std. deviation	1.00	1.02	0.95	0.99
S9	Mean	122.27	121.77	121.89	121.97
	Std. deviation	18.96	18.93	18.28	18.43

Table 4: Standart error between estimated values gained with EM, NN and MI methods and real data set

Model	S2	S3	S4	S5	S6	S7	S8	S9
EM	1,32	8.57	0.08	0.77	46.72	0.28	0.105	26.97
NN	2,35	4.9	0.062	0.687	50.51	0.13	0.046	17.39
MI	2.90	8.08	0.104	1.62	51.14	0.29	0.109	18.38

whereas the standard deviations for the other variables were higher than that of the real data set. Stadart error between methods and real data were given Table 6.

CONCLUSION

R statistical software was used in this study. Within the R software, neural, amore, norm, mi, mitools modules were utilized. In the data set which EM, NN and MI methods were applied to, different quantities of missing values were generated for each variable.

In this study, EM, NN and MI methods were used for estimation of missing values. These methods were compared in terms of completed data set and real data set. It was determined that NN method produced much closer results to the real values compared to the EM and MI methods.

A correlation coefficient close to the value of "1" indicates strong relationships between variables. As seen in Table 3, strong correlations were determined between the variables in the real data set and those for which missing values were estimated by EM, NN and MI methods. In other words, EM, MI, NN methods proved to be very successful in estimation of missing values.

NN method was more efficient than the other methods regarding estimation of missing values for the variables including different quantities of missing values. NN method was followed by EM and MI methods respectively, in terms of efficiency. In general, the means and standard deviations of the data sets with completed missing values by NN were much closer to the mean and standard deviation of the real data set.

As indicated by Table 5, the model with the smallest standard error was the most successful in the estimation of missing values. In general, NN method was more successful in estimating much closer values of standard errors associated with missing values to those associated with the real values.

In general, NN method was more successful in estimating much closer values of standard errors associated with missing values to those associated with the real values.

In terms of the standard errors given in Table 5, it was concluded that the best model for S3, S4, S5, S7, S8, S9 variables was NN, the best model for S2, S6 variable was EM.

REFERENCES

1. John, F., K. Mike and L. Marvin, 2006. Effects of the neural network s Sigmoid function on KDD in the presence of imprecise data. Computers and Operations Research, 33: 3136-3149.
2. Bal, C. And K. Özdamar, 2004. Eksik Gözlem Sorununun TüretilmiŞ Veri Setleri Yardımıyla Çözümlenmesi. Osmangazi Üniversitesi Tıp Fakültesi Dergisi., 26(2): 67-76.

3. Alzola, C. And F. Harrell, 2003. An Introduction to S and The Hmisc and Design Libraries, University of Virginia School of Medicine, Charlottesville Va, USA, pp: 68-99.
4. Colleen, M., F.C. Ennett and W. Robin, 2001. Influence of Missing Values on Artificial Neural Network Performance. Medinfo., 10(1): 449-53.
5. Adele, A., S. Mahmoud, A. Hossein and B. Farid, 2010. Comparison of Regression Pedotransfer Functions and Artificial Neural Networks for Soil Aggregate Stability Simulation. World Appl. Sci. J., 8(9): 1065-1072.
6. Aitkin, M. and D.M. Titterington, 2000. Statistics and Neural Network. The Statistican, 49: 627-628. A
7. Mohamed, S. And T. Marwala, 2005. Neural Network Based Techniques for Estimating Missing Data in Databases, 16th Annual Symposium of the Patten Recognition Association of South Africa, Langebaan, pp: 27-32.
8. Ibrahiem, M., M. El Emary and S. Ramakrishnan, 2008. On the Application of Various Probabilistic Neural Networks in Solving Different Pattern Classification Problems. World Appl. Sci. J., 4(6): 772-780.
9. Pelckmans, K., J.D. Brabanter, J.A.K. Suykens and B.D. Moor, 2005. Handling missing values in support vector machine classi?ers. Neural Networks, 18: 684-692.
10. Barnard, J. And D.B. Rubin, 1999. Small-Sample Degrees of Freedom with Multiple Imputation. Biometrika, 86: 948-955.
11. Rubin, D.B., 1976. Inference and Missing Data, Biometrika, 63: 581-590.
12. Draper, N. and H. Smith, 1998. Applied regression analysis. J. Wiley, New York, third edition.
13. Jamal, M., N.M. Ibrahim and A.N. Salam, 2008. Multilayer Perceptron Neural Network (MLPs) For Analyzing the Properties of Jordan Oil Shale. World Appl. Sci. J., 5(5): 546-552.
14. Little, R.J.A. and D.R. Rubin, 1987. Statistical Analysis with Missing Data. Wiley, New York, 18: 41-77.
15. Dempster, A.P., N.M. Laird and D.B. Rubin, 1977. Maximum Likelihood FromIncompleteData via the EMAlgorithm. Journal of the Royal Statistical Society, Series B., 39: 1-22.
16. Scheffer, J., 2002. Dealing with Missing Data. Res. Lett. Inf. Math. Sci., 3: 153-160.
17. Little, R.J.A. and D.R. Rubin, 2002. Statistical Analysis with Missing Data, Second Edition, Wiley, New York, 84: 123.
18. Rubin, D.B., 1978. Multiple imputations in sample surveys – a phenomenological bayesian approach to nonresponse. In The Pro- ceedings of the Survey Research Methods Section of the American Statistical Association, pages, 20: 34.
19. SAS., 2007. SAS/Stat Software Hangen and Enhanced, SAS Institute Incorporation, USA.
20. Fulufhelo, V., N. Muhamed and T. Marwala, 2007. Missing data: A comparison of neural network and expectation maximization techniques. Current Science, 93: 11.
21. Jerez, J.M., I. Molina, J.L. Subirats and L. Franco, 2006. Missing Data Imputation In Breast Cancer Prognosis. Processing of the 24th IASTED International Multi-Conference Boimedical Enginerring. February 15-17. Innsbruck, Austria.
22. Hill, M. SPSS., 1987. Missing Value Analysis 7.5, SPSS Inc., Chicago.
23. Mohsen, H. and M. Zahra, 2007. Temperature Forecasting Based on Neural Network Approach. World Appl. Sci. J., 2(6): 613-620.