Missing values imputation techniques for Neural Networks patterns

Thomás López-Molina* Anna Pérez-Méndez* Francklin Rivas-Echeverría**
Universidad de Los Andes
*Facultad de Ciencias Económicas y Sociales. Escuela de Estadística
**Laboratorio de Sistemas Inteligentes
Mérida, Venezuela 5101

Abstract:- This work presents the use of statistical techniques for data imputation for its use in artificial neural networks training. The Multiple imputation techniques used are: Metric Matching, Bayesian Bootstrap and Regression-based Minimal Square imputation. It is presented an application example for illustrating the appropriate use of these techniques.

*Key Words:-*Imputation techniques, Neural Networks, Statistical Analysis, Metric Matching, Bayesian Bootstrap.

1 Introduction

Artificial intelligence [2, 4, 5, 9, 29] is one of the scientific areas with greater diffusion and application in the last years. Every day is more common to find tools for industrial, commercial or academic use that involve the use of intelligent techniques in the resolution of critical and recurrent problems. Neural networks [5, 9, 10, 20] could be considered as one of the spread and more used techniques of artificial intelligence due to their simplicity, implantation facilities and design characteristics.

The wide use of neural networks in different human knowledge areas has created data processing requirements and additional necessities for the training systems. Users are interested, among other things, having tools that allow them to filter the input data, to select the best patterns and variables for the training and to fill missing values.

On the other hand, statistical data analysis techniques [3, 7, 12, 19, 21, 22] have been applied to an increasing number of knowledge areas in recent years. They are particularly appropriate for the study of great volumes of data in which it is impossible, due to its size, to observe structural characteristic easily.

One of the most frequent problems that neural networks users face is when the data have certain observations or patterns with missing values in some variables. Traditionally this problem has been solved by means of the following alternatives:

• Eliminating the patterns that present missing values (Case Deletion).
• Estimating these values (Simple Imputation).

These *ad hoc* methods in spite of being simple to implement bring serious problems which have been enough documented [14]. The first strategy presents two disadvantages: I. For data with many variables, the elimination can produce a high proportion of eliminated patterns, which is the case of a low missing values percentage but an incomplete patterns elevated percentage, II. If the patterns with missing observations are different from those completely observed, the network could present bad generalization results.

The Simple Imputation is the most common method for solving the missing values problem for two attractive reasons [24]: I. Once the values have been imputed, any Software can be used, because it would be already obtained a complete data set [13], II. In many cases, the imputations are created by the person who have collected the data and have a good knowledge about them; therefore the analyst can have better results trusting in such imputations that training considering the previously made eliminations.

The Simple Imputation presents, nevertheless, a big problem that can make it of small utility: Even considering that the missing values are not previously known, a neural networks training based on imputed data treat them as if they were the real ones, therefore, the obtained conclusions do not show the uncertainty produced by the absence of such values. Statistically, the variability or correlation estimations can be strongly biased.

The technique that will appear in this work, known as Multiple Imputation [24, 25, 27] maintains the two main virtues of the Simple Imputation and corrects its greater defects. The main idea of this technique is: for every missing value several values are imputed, presenting a series of possibilities that

take into account the variability produced by the absence of the missing value. This is not the only technique for estimating missing values; there exist numerical methods that sometimes give better results [18, 28]. If there is sufficient time and resources available, it is possible to think about techniques adapted for each problem in particular, but actually missing values is not a study object but it can be considered as a disadvantage and the proposed Multiple Imputation solutions is less complicated to implement.

The work is structured as follows: Section 2 presents the multiple imputation techniques for missing values estimation. Section 3 contains an example for evaluating the suggested techniques presented in this work and finally section 4 depicts the pertinent conclusions.

2. Missing Values Estimation Using Multiple Imputation

The Multiple Imputation is a technique that replaces each missing or deficient value by two or more acceptable values, representing a distribution of possibilities. This idea was originally propose by Rubin [23, 24]. Different investigations on missing data estimation can be found in Madow and Olkin [16], Madow, Nisselson and Olkin [17], Madow, Olkin and Rubin [18], Sande [26], Schafer [27] and Schafer and Olsen [28]. Figure 1 represents a nxp matrix which has some missing data. The Multiple Imputation replaces these missing values by a pointer to a row (record) of an auxiliary matrix that will have m>2 values or imputations.

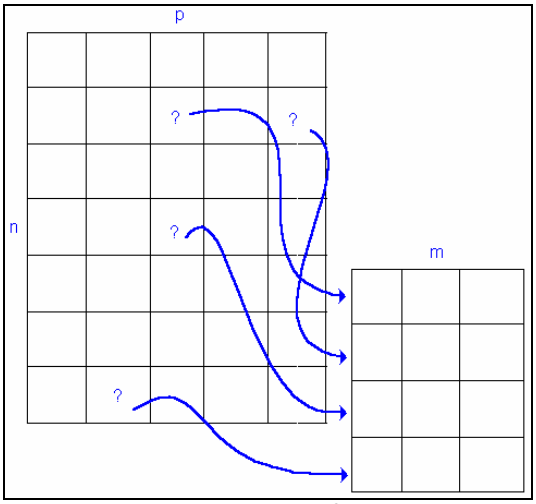

Fig 1. Data with m imputed values for each missing value.

The m values are ordered such that the missing values replaced by the first components of the records form a data set, replaced by the second components of the records form a second data set and so on. The imputed values are kept in an auxiliary matrix with a row for each missing value and m columns. Rubin [24] indicates that the data sets obtained by means of Multiple Imputation are more useful when the ratio of missing values is not excessive and m is between 2 and 10.
Multiple Imputation is attractive by a great amount of reasons: I. It is compatible with the methods and software for complete data. II. A set of m imputations can be used for a great variety of analysis and there is no necessity to impute again when a new analysis is going to be made. III. Inferences, standard errors and correlations obtained from Multiple Imputation are generally valid because they incorporate the uncertainty due to the missing values. Additionally it is highly efficient even when m is low. In most applications, only 3 to 5 imputations are necessary for obtaining excellent results.

The Multiple Imputation was proposed more than 20 years ago [46, 48] but the method has remained unused. The main reason of it has been the absence of computational tools for generating imputations. Recently it has appeared software for multivariate incomplete data. These programs are easy to understand and for using, in addition they are implemented in graphical environment for Windows (95/NT). They can be obtained freely from the Web site: http://stat.psu.edu/~jls/misoftwa.html. Additionally, there exists software for sale that uses the Multiple Imputation techniques propose by Rubin [47]. It can be obtained from Statistical Solutions through his Web site: http://www.statsolusa.com.

2.1 Multiple Imputation Techniques

- *Metric Matching*

This method defines a distance measurement (d) between the patterns with missing values and the complete ones, this is $d(x^{|}, x^{||})$, where $x^{|} \in X_{obs}$ and $x^{||} \in X_{mis}$ and it will select as donors for the pattern with the missing value those complete patterns that are nearer according to the selected distance. One of the most popular distances measures are:

$d_I(x',x'')^2 = (x'-x'')(x'-x'')^t$
(Euclidean Distance)

$d_S(x',x'')^2 = (x'-x'')S^{-1}(x'-x'')^t$
(Statistical Distance)

The defined distances are well known and the last one appears frequently in statistical Literature; in it "S" is the calculated variance and covariance matrix for the p variables of X_{obs}.

- *Bayesian Bootstrap*

Let's consider the vector $d = (d_1,...,d_k)$ of all the possible values of $Y_i \in Y$ and $\theta = (\theta_1,...,\theta_k)$ a vector of associated probabilities, let's suppose that Y_i $(i = 1,...,n)$ given θ are independent identically distributed. The data probability distribution is given by:

$$P(Y_i = d_j / \theta) = \theta_j$$

$$\sum_{j=1}^{k} \theta_j = 1$$

Let's suppose additionally the following a priori distribution of θ:

$$P(\theta) = \prod_{j=1}^{k} \theta_k^{-1} \qquad \text{If} \quad \sum_{j=1}^{k} \theta_j = 1$$

$$P(\theta) = 0 \qquad \text{Otherwise}$$

Ericson and Rubin [6] have called to this distributions specification the Bayesian Bootstrap. Under these assumptions it is possible to develop an easy Multiple Imputation method that provides valid inferences for great values of "n": From the observed patterns set, there are found randomly the donors for the pattern with missing observations.

- *Regression-based Minimal Square Imputation*

An intuitive method for generating imputations is using Minimal Square Regression [1]. It is found Y with X using the complete patterns for obtaining the parameters of the model, and this model is used for obtaining the missing values. The assumptions on which the technique rests are the following ones:

$Y_i \sim N(X_i\beta, \sigma^2)$ (i=1,2,...,n)

$P(\beta, \sigma^2) \propto const$

With the data distribution and the assumption of "noninformative" distribution for the model parameters, Rubin [24] presents the theoretical bases for a method that assures valid inferences for the variable Y.

3 Multiple imputation techniques application example for missing values estimation in neural networks training

The data used for this example is well known in Multivariate Analysis Literature, presented by Johnson and Wichern [11] of a study made by Gerrild and Lantz [8], concerning 56 analyzed crude samples originated on the Elk Hills Oil Field in California (USA). For each sample, five variables were measured:

X1: Vanadium (%)
X2: Iron (%)
X3: Beryllium (%)
X4: Saturated Hydrocarbons (%)
X5: Aromatic Hydrocarbons (%)

Each one of the samples could belong to one of the three following zones:

C1: Whilhelm
C2: Sub-Mulinia
C3: Upper (Mulinia, second sub-scales, first sub-scales)

Once the data is standardized with general mean and variance (r = 15%; α= 10%), it was obtained the results displayed in table 1.

N	56
N1	6
N2	12
N3	38
ne	32
n1	3
n2	7
n3	22

Table 1. Partition Results.

From the 56 (N) data set, classified in the three groups of size 6 (N1), 12 (N2) and 38 (N3), the technique suggests to use 32 (ne) data for training (and therefore 24 for validation), selecting 3(n1) of the first group, 7(n2) of the second and 22 (n3) of third.

Once the data is standardized, 15% of the matrix cells were randomly selected and they were replaced by missing values. The Multiple Imputation techniques of Metric Matching and Bayesian Bootstrap were applied. The general characteristics of the matrix with missing data appear in Table 2.

Number of Patterns (n)	56
Number of Variables (p)	5
Number of Patterns with Missing Values (n0)	31
Number of cells in the matrix (nxp)	280
Number of cells with Missing Values	41
Percentage of cells with Missing Values	15%
Number of Imputations Made (m)	3

Table 2. Data Matrix General Characteristics

An intuitive form for evaluating the quality of the procedures is through the Imputation Errors analysis. For each missing value two error measures are defined:

$$RECM = \sqrt{\frac{\sum_{j=1}^{m}\left(Y_j^* - Y\right)^2}{m}}$$

$$EAM = \frac{\sum_{j=1}^{m}\left|Y_j^* - Y\right|}{m}$$

Where Y_j^* are the obtained values by imputation and Y is the real value of the cell that artificially became missing value. The calculated errors for each of the variables were averaged and they appear in table 3. It can be observed that the Metric Matching have presented smaller error values than the Bayesian Bootstrap in almost all the variables, being significantly smaller in that with smaller number of missing values.

	Missing		Metric Matching	Bayesian Bootstrap
Variable 1	3	RECM	0.7189	1.4027
		EAM	0.5362	1.0908
Variable 2	15	RECM	1.0299	1.4563
		EAM	0.9869	1.2243
Variable 3	10	RECM	1.0376	1.0783
		EAM	1.0066	0.9503
Variable 4	2	RECM	0.4578	2.3063
		EAM	0.4578	2.2287
Variable 5	11	RECM	1.3968	1.0839
		EAM	1.3854	0.9333

Table 3. Obtained errors using diverse methods of Multiple Imputation

Next, an artificial data set with missing values was obtained, for comparing the performance of the Bayesian Bootstrap technique with the Case Deletion, Imputation with Averages and original data. These data will be use for training an Artificial Neural Networks for classification, therefore a form to compare different networks is from the Percentage of correct classification (PCC) and Percentage of Incorrect classification (PIC). It is obtained using the previously trained network with a testing data set and counting the number of mistaken classification. When the testing data set is independent to the training data set the estimation is unbiased, but its variance could be elevated if there is not many data available. When there is not testing data set available, a partition is made as it was previously suggested and there are considered PIC values for training and validation.

Once it is generated 10% of artificial missing values in the original matrix, it was separately applied the Multiple Imputation Bayesian Bootstrap technique for each of the three data categories, as well as the Case Deletion and the Imputation by Average. Later it was selected the different matrices partitions by means of Stratified Random Sampling (r=0.15; α=10%). This way it was found the data sets shown in table 4 and the sizes of partition of table 5.

	C	D	E	F
Training	Original Data Set ne = 32	Imputed data Set using Bayesian Bootstrap ne = 45	Imputation with average Data Set ne = 32	Case Deletion Data Set ne = 19
Validation	nv = 24	nv = 42	nv = 24	nv = 5

Table 4. Data Partition for training and Validation

	C	D	E	F
C1	3	6	3	1
C2	7	10	7	3
C3	22	29	22	15

Table 5. Number of Training patterns (n_e) using Stratified Random Sampling

The training was given using two layers perceptron networks [15]: Hidden layer (10 neurons and logistic activation function) and three neurons output (logistic activation function). The initial results of the three training appear in table 6.

	Correct Classification Rate	Training Error	Validation Error
C	0.833333	0.01153	0.3259853
D	0.833333	0.278334	0.3074587
E	0.913043	0.129556	0.2371633
F	1.000000	0.05689	0.1518498

Table 6. Correct Classification Rate. Training and Validation Error

The Correct Classification rate is a performance measurement in the training of the network and its proximity to one (1) indicates a better performance. The networks trained with sets E and F have presented the best performance, but the trained with the imputed data by means of Bayesian Bootstrap (D) have presented an exactly behaviour as the trained with the original data.

Neural Networks trained with sets E and F have an appreciable difference in the Training and Validation errors, indicating the possibility of over training. In ideal situations more data would have to be acquired in order to improve the results and for generating more reliable networks. Table 7 displays Percentage of Correct Classification (PCC) and Percentage of Incorrect Classification (PIC) of the Training phase and table 8 presents the corresponding ones of validation for the different data sets.

	C	D	E	F
Total Training Patterns	32	45	32	19
PCC	100	91.11	100	100
PIC	0.00	8.89	0.00	0.00
Unknown classification	0.00	0.00	0.00	0.00

Table 7. Neural Networks behaviour with diverse data groups. Training

	C	D	E	F
Total Validation Patterns	24	42	24	5
PCC	83.33	83.33	87.5	100
PIC	16.67	16.67	12.5	0.00
Unknown classification	0.00	0.00	0.00	0.00

Table 8. Neural Networks behaviour with diverse data groups. Validation

It can be appreciated that the network trained with the Multiple Imputation technique turned out to have a greater PIC of Training, nevertheless in the validation the PIC was equal to the value obtained with the real data, whereas the networks trained with sets E and F have satisfactory but far statistics of the values obtained with C, originating from the data without missing values.

The Simple Imputation and the Case Deletion have favorable results, but far from the real values. They are the most popular techniques for handlings missing values in the Neural Networks training systems. They diminish the sample size and can produce over training as it was shown in the previous example. The Multiple Imputation, however, presented results that at first sight seem inferior but they are very similar the obtained with the original data, incorporating the originating uncertainty of the missing values.

4 Conclusions

The data analysis with missing values is a statistical area where have been found great advances. The modern technologies for their handling surpass the old ad hoc ones and finally they are available to the analysts. Between these techniques the Multiple Imputation is especially powerful by its generality characteristics.

Multiple Imputation gives better results that Simple Imputation and Case Deletion procedures, because it does not diminish the sample size available and also it takes into account the uncertainty due to the presence of missing values, incorporating such variability in the estimations. In the presented example although the obtained imputations are not absolutely precise, they are statistically more reasonable than to impute with averages or to eliminate patterns (which would reduce its number in 55,4%). Training results with the estimations of missing values by Multiple Imputation are more realistic than those of the commonly used techniques, providing to the network greater generality characteristics.

The Multiple Imputation is not the only modern technology for the handling of missing values available to the investigators. Some houses of software are beginning to incorporate characteristic related to missing values to certain routines of models adjustment. These procedures are similar to the Multiple Imputation because they are based on a predictive distribution, but the used methods are analytical or numerical.

The Multiple Imputation can be applied to a variety of problems and will possibly be common to data analysts. For neural networks users, it can be seen as a coherent available procedure, supported by the Statistical Theory, which can be used for solving the problem of missing values presence in training patterns.

5 References

[1] Afifi, A. y Elashoff, R. (1969). *Missing observations in multivariate statistics III: Large sample analysis of simple linear regression*. Journal of the American Statistical Association. 64, 337-358

[2] Amador-Hidalgo, L. (1997) *Inteligencia Artificial y Sistemas Expertos.* Servicio de Publicaciones de la Universidad de Córdoba

[3] Anderson, T. (1958). *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons Inc., New York

[4] Cohen, P. (1989) *The Handbook of Artificial Intelligence*, New York. Addison Wesley

[5] Colina, E., Rivas, F. (1998) *Introducción a la Inteligencia Artificial*. Cuadernos de Control. Postgrado en Ingeniería de Control. Universidad de Los Andes

[6] Ericson, W. (1969). *Subjective Bayesian models in sampling finite populations*. Journal of the Royal Statistical Society, B31, 195-233

[7] Freixa, M., Salafranca, L., Ferrer R., Guardia, J. y Turbany, J. (1992) *Análisis Exploratorio de Datos*. Promociones y Publicaciones Universitarias. Primera Edición. Barcelona España

[8] Gerril y Lantz (1969). *Chemical Analysis of 75 Crude Oil Samples from Pliocene Sand Units, Elk Hills Oil Field, California*. US Geological Survey File Report. In. Applied Multivariate Statistical Analysis. Johnson, R. y Wichern, D. (aut). Prentice Hall, New Jersey

[9] Hagan, M., Demuth, H., Beale, M. (1996) *Neural Networks Design*. PWS Publishing Company. U.S.A.

[10] Honik, K., Stinchcombe M., y White, H. (1989). *Multilayer feedforward networks are universal aproximators*. Neural Networks, vol 2, 5, pp 359-366

[11] Johnson, R. A. and Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis*. Prentice Hall, New Jersey

[12] Lebart, L. Morineau, A. y Warwick, K. (1984). Multivariate Descrptive Statistical Analysis. John Wiley & Sons Inc., New York

[13] Littel, R. et al. (1996). *SAS System for Mixed Models*. SAS Institute, Cary, NC.

[14] Little, R. y Rubin, D. (1987). *Statistical Analysis with Missing Data*. John Wiley & Sons Inc., New York

[15] López, T. e Izarra, E. (2000). *Aplicación de Redes Neurales Perceptrónicas para Clasificación de Observaciones y Comparación con el Análisis Discriminante Múltiple*. II jornadas de Alpicaciones de la Inteligencia Artificial, Postgrado en Ingeniería de Control y Automatización, Universidad de los Andes, Mérida, Venezuela

[16] Madow, W. y Olkin, I. (1983). *Incomplete Data in Sample Surveys*, Proceedings of the Sympsosium. Academic Press, New York

[17] Madow, W., Nisselson, H. y Olkin, I. (1983). *Incomplete Data in Sample Surveys*, Volume I, Report and case Studies. Academic Press, New York

[18] Madow, W., Olkin, I. y Rubin, D. (1983). *Incomplete Data in Sample Surveys*, Volume 2, Theory and Bibliographies. Academic Press, New York

[19] Mardia, K., Kent, J., y Bibby, J. (1979). Multivariate Analysis. Academic Press, London

[20] MathWorks (1994). Neural Networks Toolbox for MatLab. MathWorks

[21] Morrison, D. (1990) *Multivariate Statistical Methods*. Tercera Edición. McGraw Hill Publishing Company

[22] Neural Works. Neuralware Inc. Website: http://www.teleport.com/~cognizer/eet/Almanac/COMPANY/42088521.HTM

[23] Rubin, D. (1978). *Multiple imputations in sample surveys – a phenomenological Bayesian approach to nonresponse*. Proceedings of the Survey Research Methods Section of the American Statistical Association, 20-34

[24] Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. J. Wiley & Sons, New York

[25] Rubin, D. (1996*). Multiple Imputation after 18+ Years*. Journal of the American Statistical Association, 91, 434. Application and Case Studies

[26] Sande, I. (1983). *Hot-deck imputation procedures. In Incomplete Data in Sample Surveys*, Volume 3, Proceedings of the Symposium. Madow, W., y Olkin, I. (eds). Academic Press, New York

[27] Schafer, J. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London

[28] Schafer, J. y Olsen, M. (1998). *Multiple imputation for multivariate missing-data problems: a data analyst's persepective*. The Pennsylvania State University, en http://www.stat.psu.edu

[29] Zadeh, F. (1965) *Fuzzy Sets*. Information and Control pp.338-353