Proceedings of the 24th IASTED International Multi-Conference
BIOMEDICAL ENGINEERING
February 15-17, 2006, Innsbruck, Austria

MISSING DATA IMPUTATION IN BREAST CANCER PROGNOSIS

José M. Jerez
Escuela Técnica Superior de Ingeniería en Informática
Departamento de Lenguajes y Ciencias de la Computación
Universidad de Málaga, Campus de Teatinos S/N 29071, Málaga, Spain.
email: jja@lcc.uma.es
Ignacio Molina
Escuela Técnica Superior de Ingeniería en Telecomunicación
Departamento de Tecnología electrónica.
Universidad de Málaga, Campus de Teatinos S/N 29071, Málaga, Spain.
email: aimc@dte.uma.es
José L. Subirats
Escuela Técnica Superior de Ingeniería en Informática
Departamento de Lenguajes y Ciencias de la Computación
Universidad de Málaga, Campus de Teatinos S/N 29071, Málaga, Spain.
email: josuluco@hotmail.com
Leonardo Franco
Escuela Técnica Superior de Ingeniería en Informática
Departamento de Lenguajes y Ciencias de la Computación
Universidad de Málaga, Campus de Teatinos S/N 29071, Málaga, Spain.
email: lfranco@lcc.uma.es

ABSTRACT
Missing data are often a problem present in real datasets and different imputation techniques are normally used to alleviate this problem. In this paper we analyze the performance of two different data imputation methods in a task where the aim is to predict the probability of breast cancer relapse. Mean imputation and hot-deck methods were used to replace missing values found in a dataset containing 3679 records of patients. Artificial neural network models were trained with the standard dataset (containing no missing data but a restricted number of cases) and also with the data reconstructed by using the two imputation methods mentioned above. The results were analyzed in terms of the predictive accuracy and also in terms of the calibration of the results.

KEY WORDS
Missing data imputation, artificial neural networks, breast cancer, prognosis.

1 Introduction

This paper presents a study performed to analyze the effect of missing data imputation techniques in the generalization capabilities of an artificial neural network system constructed to model the prognosis of breast cancer in surgical operated patients.

Missing data is a very common problem in real datasets and thus different methods for handling this problem have been developed. A simple and common strategy is to ignore missing values, thus reducing the size of the useful dataset. Several authors ([31, 23]) have demonstrated the dangers of simply removing cases ('listwise deletion') from the original data set, as deletion can introduce substantial biases in the study, specially when missing data is distributed in a not random way. Missing data is an even greater concern when decisions must be made about the appropriateness of the care a patient should receive. Missing values either reduce the number of available cases for analysis or even worse might distort the analysis by introducing a bias into the estimation and/or prediction process. In this sense, the imputation of missing data is an area of statistics which has attracted much attention in the last decades ([23, 32, 1, 33, 25, 16]) and many different strategies have been developed to tackle this problem. In this work, we used two standard imputation techniques, Mean and hot-deck imputation to supply values for any datum variable containing a missing value.

Breast cancer prognosis is an important problem in medicine as knowing the probability of cancer relapse of a particular patient helps on the decision making process involved in the patient's treatment. Survival analysis have been mostly analyzed using regression methods from statistics (like the Cox regression model [7]) but different alternatives using techniques like artificial neural networks have been successfully applied [2, 6, 26, 18]. The success on the use of neural networks might be their inherently non-linearity and flexibility in comparison to the existing constraints in the regression models. The purpose of the research conducted in this work was to compare a neural

network model performance when different methods were used for the imputation of missing values on the problem of breast cancer prognosis . The results were analyzed in terms of their predictive accuracy and also in terms of their calibration.

2 Description of the dataset and Methods

2.1 The used dataset

Data were collected from the "El Alamo 1" Project, the largest database on breast cancer in Spain. The dataset analyzed in this study includes demographics, therapeutic and recurrence-survival information from 3679 women patients with operable invasive breast cancer diagnosed in 32 different hospitals belonging to the Spanish Breast Cancer Research Group (GEICAM) between the years 1990 and 1993. This study used the set of clinical and pathological variables selected in [18] as more significant prognostic factors in the prediction of patients outcome. Table 1 shows for the selected covariates(Age, Tumour size, Axillary lymph nodes, Tumour histological grade and Type of treatment) the range, the mean value and the number of missing cases present. The analysis was restricted to patients with follow-up time in the interval of one to 128 months (thirty-four percent of patients were relapsed in this period of study).

The percentage of missing data (counting every attribute for every patient cases a different data) represents a 7.13% of the overall data set. There are 1835 cases with at least one missing value. From these, there are 1520 patient cases where one value is missed, 150 with two data missed and only 5 cases had three data values missed. The last column in Table 1 under the heading of missingness, shows the distribution of missing data for each covariate. A first analysis on the distribution of missing data shows that: i) missingness is mainly present on the histological grade; ii) the amount of missingness for covariates age, tumour size and number of axillary lymph nodes, represents roughly a 16% of the amount of missing data for histological grade; iii) there is no data missing for the covariate related to the type of treatment.

For the purpose of testing the prediction accuracy of the system, the standard dataset (containing no missing data) was divided in training and validation sets, selected randomly in a $80\% - 20\%$ size relationship. Through the implementation of the imputation techniques the training sets used for the cases of the mean and hot-deck procedures are enlarged but the validation (or test) set remains the same in all the three cases analyzed.

Prognostic factors	Range	Mean	Missingness
Age, years	$25 - 90$	56.21	7
Tumour size	$0.2 - 13$	2.87	148
Axillary lymph nodes	$0 - 35$	2.48	104
Histological grade	$1, 2, 3$	2.04	1576
Type of treatment	$0, 1, 2, 3, 7,$ $8, 9, 10, 11$	5.37	0
Survival Status (S)	$0, 1$	0.34	0

Table 1. Some characteristics of the breast cancer dataset containing records from 3679 patients. For the five covariates considered as important for the prognosis (age, tumour size, number of axillary lymph nodes, histological grade and type of treatment) the range, mean and number of missing cases are shown in the columns. The survival status shown in the last row is the variable to be predicted by the system.

2.2 Imputation techniques

Among the different data imputation techniques, mean imputation is the simplest approach ([19, 1]). In general, mean imputation imputes the mean values of each variable on the respective missing variables, as an estimate of the missing value. Mean imputation is a simple method commonly used in the social sciences, easy to implement and a fast alternative to listcase deletion. On the other hand, hot-deck imputation is an intuitively simple method for accommodating incomplete data. For this reason, it has been very successful in imputing missing values in large data sets ([21, 23]). When values are missing at random (MAR), it produces unbiased estimates of population means ([20]). In fact, the hot-deck method is asymptotically equivalent to the mean-score method for the estimation of a regression model parameter and thus the hot-deck procedure can be understood in the context of likelihood methods. In general, the hot-deck imputation replaces a missing value of a receptor respondent for that taken from a similar donor respondent who has complete data, but other alternatives exists. For random hot-deck imputation, the missing value is replaced by a responding value from a donor randomly selected from a set of candidate donors. hot-deck sequential uses the observed value of the respective variable from the immediately preceding complete record as an estimate of the missing value. Finally, hot-deck distance uses a predefined distance metric (usually a Euclidean distance function) to find the donor record from their potential donors.

Both of these methods accommodate data where missingness may depend on the observed variables but not on the unobserved value of the incomplete covariate (MAR) ([23]). However, hot-deck imputation is commonly used because it presents some advantages: i) unlike mean imputation, it preserves the distribution of item values, ii) permits the use of the same sample weight for all items, and iii) results obtained from different analysis are consistent with one another.

2.2.1 Mean imputation

Mean imputation is accomplished simply by averaging the corresponding variables belonging to complete patient cases. Thus, the missing values are systematically substituted by the mean value. The only categorical variable to

be imputed is the histological grade, and for this variable, the mode is used instead of the mean.

2.2.2 *Hot-deck imputation*

In the hot-deck, the whole set of patient cases (which includes missing values) is splitted into 22 different groups. This number of groups is obtained by multiplying the cardinalities of the categorical variables type of treatment and survival status. Since there is no missingness in these two variables, a patient case can always be classified and assigned to its corresponding group. As described, the variable grade is also categorical, but unlike treatment and survival status, it is ordinal. Therefore, differences between the values of grade within two different patients can be computed and is explained below.

Basically, the hot-deck method is implemented as follows: when a missing variable is to be imputed, the right set of candidate donors consisting of patient cases with complete data belonging to the same group is selected. From this set of candidates, the patient case closest to the receptor patient is selected as the donor from which the missing values are taken. For simplicity, the squared Euclidean norm is used as the similarity measure between pairs of patient cases.

In more detail, let a generic patient case *i* be defined by $\mathbf{p}_i = \{p_0, p_1, p_2, p_3, p_4\}$, where p_0 represents the covariate age, p_1 is the tumour size, p_2 is number of axillary lymph nodes, p_3 is the survival status , and p_4 represents the histological grade. As stated above, the first step in our implementation of the hot-deck imputation method, begins by determining the donor group for any receptor $\mathbf{p_i}$. Each group is formed by the set of complete cases with the same pair of values for the variables: type of treatment and survival status. For each donor patient case, the squared Euclidean norm between the receptor patient case to be imputed and the donor is computed. The variables involved in this computation are only those which are not missed in the receptor patient. In order to ensure that the weight of all the variables used in the computation of the Euclidean distance is the same, they are normalized according to:

$$p_i^n = \frac{p_i - E(p_i)}{\sigma_i}, \text{with } i <= 3 \qquad (1)$$

where $E(p_i)$ is the mean of p_i and σ_i is the standard deviation. As explained, p_4 is an ordinal categorical variable, and consequently it can not be normalized according to Eq. 1. Therefore, it has to be incorporated in the similarity measure in a different way. The resulting similarity function between the receptor patient $\mathbf{r_j}$ and a potential donor $\mathbf{d_k}$ is given by:

$$D(\mathbf{r_j}, \mathbf{d_k}) = \mathbf{K} \times \sum\nolimits_{\mathbf{p,q} \leq \mathbf{3}} \left(\mathbf{r_p^n} - \mathbf{d_q^n}\right)^{\mathbf{2}} + \left(\mathbf{r_3} - \mathbf{d_3}\right)^{\mathbf{2}}. \qquad (2)$$

In Eq. 2, the term $(r_3 - d_3)^2$ accounts for the squared difference between the ordinal categorical variable grade, which ranges from 0 to 4. Thus, constant K serves to scale up the summation to this range. As stated, the number of variables involved in this summation is that of complete variables in the receptor patient case.

2.3 Artificial neural networks

Several neural network approaches have been proposed to model survival data, where the aim is to predict the probability of survival (or the instantaneous hazard) at different intervals of time. In some cases (see for example [13, 29, 17, 5, 24]) the prognostic covariates have been used as inputs to the neural system while the time to relapse is the output of the neural network. A more efficient representation of time is to include it as a covariate, and in this case the output of the system becomes an indicator of relapse or not at a given time. This kind of approach (also refereed as time-coded models) has been implemented by several authors [9, 8, 22, 28, 2, 4] and can be interpreted as the discrete time implementation of the proportional hazards model [26]. Moreover, this kind of neural network model for survival prediction has proved to be very stable in monthly studies over follow-up periods of several years [3]. Time-coded models generate a prognostic index that can be interpreted as conditional probabilities or cumulative probabilities depending on the preprocessing performed on the input data [10].

The neural approach adopted in this work lies within those known as time-coded models, in which the time of follow-up is included as an additional covariate. In the modelling process, the input vectors are replicated from the first time interval until the interval previous to the maximum follow-up, setting the survival status to 0 and the time of follow-up to the mean value of the corresponding interval (5 months, 15 months, etc). Besides, for a patient who has died, data vectors are included with survival status 1 for all time intervals after the occurrence of the event. The selective replication of cases for all the patients depending on the censoring status at maximum follow-up makes the output of the network to represent directly the cumulative relapse probability for a given patient, and it has also the advantage that a single neural network can be used to obtain predictions for every time interval.

A three-layer neural network model was constructed with an ad-hoc software developed in C++ and R languages ([27, 12]). The Back-propagation learning algorithm together with the cross-entropy error function were used for training the network. Transfer functions for all neurons in the network were sigmoidal and overfitting problems were avoided using the weight decay regularization technique [14]. The architecture used contained 20 neurons in the hidden layer and was selected using only the training data between the different architectures considered with a number of hidden neurons between 5 and 50. The selected architecture was the best one in terms of the validation error achieved and was selected using the standard dataset (containing no missing data). The input data fed into the neural

architecture was pre-processed by Gaussian normalization making the data to be normally distributed around 0.5 with standard deviation equals to 1. Outliers with covariate values 3 standard deviations larger than the mean were also eliminated from the dataset.

2.4 Model evaluation

The internal validation of the predictive models was evaluated by measures of calibration and discrimination on the test (validation) set. Calibration or goodness-of-fit ([11]) refers to the ability of the model to assign the correct probabilities of outcome to groups of patients. This ability was assessed using the Hosmer-Lemeshow C statistic in which a high p value would indicate a good model fit. Model discrimination refers to the ability of the model to assign higher probabilities of death (outcome) to patients who actually die than to those patients who live. This was evaluated by the area under the receiver operator characteristic (ROC) curve([15]).

3 Results

The accuracy of the neural networks trained on the different datasets generated by using the imputation methods was tested on the validation set by means of the area of the ROC curve. In figure 1 the ROC curve computed from the results obtained by using the standard dataset (no missing data present), and the ROC curves obtained by applying both used imputation methods (mean and hot-deck) are plotted.

The area under the ROC curve (AUC) was quite similar for the three analyzed cases and not statistically significantly different ($p > 0.01$) between any pair of them: 0.816 (standard dataset), 0.801 (mean imputation), 0.825 (hot-deck imputation).

Calibration plots for the three different datasets used were also computed. In figure 2 the calibration chart for the observed and computed outcomes is shown. The results obtained from using the standard set (containing no missing data) are indicated by Normal. The two other cases shown correspond to the two imputation methods used, Mean and hot-deck methods. Bars represent the observed and predicted outcome in deciles of ascending risk. The Hosmer-Lemeshow C statistic was used to analyze the accuracy of the predictions respect to the actual observed values. The results indicate that the calibration of the models using the data that include the imputation values were more accurate than the one obtained with the standard dataset $(t = 28.1, p < 0.001)$. For the case of using the method of imputation of the mean, the values were $(t = 11.5, p = 0.17)$ while for the hot-deck method were $(t = 12.14, p = 0.14)$ indicating in both cases that the calibration of the predictions obtained with these two imputation methods were not significantly different from the ones corresponding to the observed data.

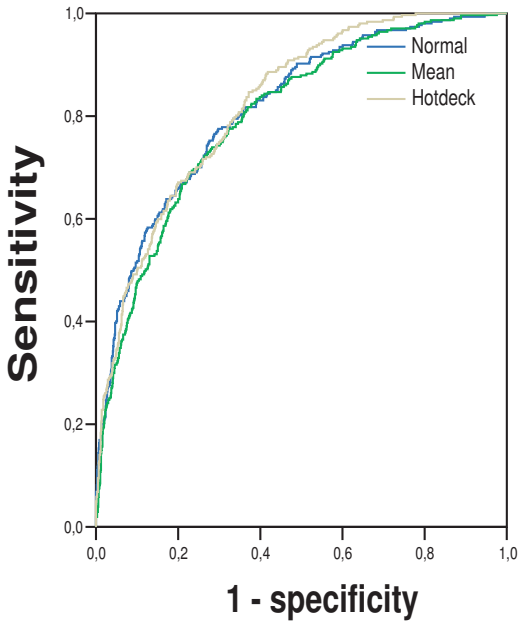

Figure 1. ROC curves obtained for the prediction of the neural network models trained on the standard datasets (no missing data present, 'labeled normal') and on augmented dataset in which the mean and hot-deck imputation methods were used to replace the missing values present.

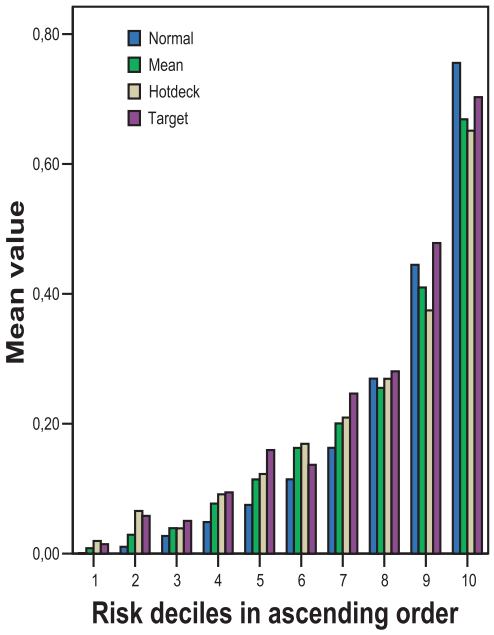

Figure 2. Calibration chart displaying the results obtained from the ANN models trained with three different data sets in comparison to the observed outcomes. The results obtained from using the standard set containing no missing data but a less number of cases are indicated by Normal. The two other cases shown correspond to the two imputation methods used, the Mean and hot-deck methods. Bars represent the predicted and observed outcome in deciles of ascending risk.

4 Discussion and Conclusion

We have applied two different methods (mean and hot-deck) to treat the problem of missing data present in data records from breast cancer operated patients. In terms of predictive accuracy, even if the size of the datasets used were twice as large for the cases in which the imputation methods were applied no significant improvement was observed as measured by the area under the ROC curve. However, in terms of model calibration, measuring the accuracy of the predictions when the data is grouped according to the patients risk of cancer relapse, the results for both datasets containing the imputation values were significantly better than those obtained when the missing values were simply discarded. This preliminary results show that the use of imputation methods can be beneficial for this problem, but given that the predictive accuracy when the enlarged datasets were used has not improved, it might be necessary to implement more sophisticated methods (as for example, multiple imputation [30]). We are currently in the process of pursuing this task, also taking a previous step of validating the appropriateness of the imputation methods by artificially creating missing data and analyzing the obtained results.

Acknowledgements

This work has been supported by the University of Málaga through a pre-competitive Grant, by CICYT (Spain) project TIN2005-02984 and by FEDER funds. LF acknowledges support from the Spanish Ministry of Education and Science through a Ramón y Cajal fellowship.

References

[1] P.D. Allison. *Missing data*. Sage Pub., Thousand Oaks, CA, 2002.

[2] E. Biganzoli, P. Boracchi, D. Coradini, M. Daidone, and E. Marubini. Prognosis in node-negative primary breast cancer: a neural network analysis of risk profiles using routinely assessed factors. *Ann Oncol*, 14:1484–1493, 2003.

[3] E. Biganzoli, P. Boracchi, L. Mariani, and E. Marubini. Feed forward neural networks for the analysis of censored survival data: A partial logistic regression approach. *Stat Med*, 17:1169–1186, 1998.

[4] P. Boracchi, E. Biganzoli, and E. Marubini. Modelling cause-specific hazards with radial basis function artificial neural networks: application to 2233 breast cancer patients. *Stat Med*, 20:3677–3694, 2001.

[5] S. Brown, A. Branford, and W. Moran. On the use of artificial neural networks for the analysis of survival data. *IEEE Trans Neu Net*, 8:1071 –1077, 1997.

[6] H. Burke, P. Goodman, D. Rosen, D. Henson, J. Weinstein, F. Harrel, J. Marks, D. Winchester, and D. Bostwick. Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer*, 79, 1997.

[7] D. Cox. Regression models and life tables. *J R Stat Soc*, 34:187–202, 1972.

[8] M. De Laurentis, S. De Placido, A. Bianco, G. Clark, and P. Ravdin. A prognostic model that makes quantitative estimates of probability of relapse for breast cancer patients. *Clin Cancer Res*, 5:4133–4139, 1999.

[9] M. De Laurentis and P. Ravdin. A technique for using neural network analysis to perform survival analysis of censored data. *Cancer Lett*, 77:127–138, 1994.

[10] S. Dreiseitl and L. Ohno-Machado. Logistic regression and artificial neural network classification models: a methodology review. *J Biomed Inf*, 35:352–359, 2002.

[11] Hosmer D.W. and Lemeshow S. *Applied Logistic Regression*. Wiley, 2000.

[12] B.S. Everitt. *An R and S-Plus Companion to Multivariate Analysis*. Springer, New York, 2005.

[13] D. Faraggi, R. Simon, E. Yaskil, and A. Kramar. Bayesian neural network models for censored data. *Biometrica J.*, 5:519–532, 1997.

[14] R. Golden. *Mathematical Methods for Neural Network Analysis and Design*. MIT Press, 1996.

[15] J. Hanley and B. McNeil. The meaning and use of the area under the receiver operating characteristic (roc) curve. *Radiology*, 143:29–36, 1982.

[16] J.G. Ibrahim, M.H. Chen, and S.R. Herring. Missing-data methods for generalised linear models: A comparative review. *J Am Stat Assoc*, 100(469):332–326, 2005.

[17] J. Jerez, J. Gómez, G. Ramos, J. Muñoz, and E. Alba. A combined neural network and decision trees model for prognosis of breast cancer relapse. *Artif Intell Med*, 27:45–63, 2003.

[18] J.M. Jerez, L. Franco, E. Alba, A. Llombart-Cussac, A. Lluch, N. Ribelles, B. Munrriz, and M. Martn. Improvement of breast cancer relapse prediction in high risk intervals using artificial neural networks. *Br Can Res Tr*, in press, 2005.

[19] J.H. Jihn and J. Sedransk. Effect on secondary data analysis of common imputation methods. *Soc Method*, 19:213–241, 1989.

[20] J. Kaiser. The effectiveness of hot-deck procedures in small samples. In *Proc. Ann. Meeting of the Am. Statistical Assoc.*, 1983.

[21] G. Kalton and D. Kasprzyk. Imputing for missing survey responses. In *proceedings of the section on survey research methods*, pages 22–31. Am Stat Assoc, 1982.

[22] K. Liestol and P. Andersen. Updating of covariates and choice of time origin in survival analysis: problems with vaguely defined disease states. *Stat Med*, 21:3701–3714, 2002.

[23] R.J. Little and D.B. Rubin. *Statistical analysis with missing data*. John Wiley & Sons, Inc., New York, NY, USA, 2002.

[24] M. Lundin, J. Lundin, H. Burke, S. Toikkanen, Pylkken, and H. Joensuu. Artificial neural networks applied to survival prediction in breast cancer. *Oncology*, 57:281–286, 1999.

[25] C. Manski. Partial identification with missing data: concepts and findings. *Int J Aprox Rea*, 39(2-3):151–165, 2005.

[26] L. Ohno-Machado. A comparison of cox proportional hazards and artificial neural network models for medical prognosis. *Comput Biol Med*, 27:55–65, 1997.

[27] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005. ISBN 3-900051-07-0.

[28] P. Ravdin and G. Clark. A practical application of neural network analysis for predicting outcome of individual breast cancer patients. *Breast Cancer Res Treat*, 22:285–293, 1992.

[29] R. Ripley, A. Harris, and L. Tarassenko. Neural network models for breast cancer prognosis. *Neural Comput. Appl.*, 7:367–375, 1998.

[30] D. Rubin. Multiple imputation after +18 years. *J Am Stat Assoc*, 91:473–489, 1996.

[31] D. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley & Sons Inc., 2004.

[32] J.L. Schafer. *Missing data*. Chapman and Hall, New York, 1997.

[33] J.L. Schafer and J.W. Graham. Missing data: Our view of the state of the art. *Psychological Methods*, 7:147–177, 2002.