Generalized Linear Models, Generalized Additive Models and Neural Networks: Comparative Study in Medical Applications

Ana Luisa Papoila, Cristina Rocha, Carlos Geraldes, and Patricia Xufre

Abstract

During the last two decades, evaluating severity of illness and predicting mortality of critical patients became a major concern of all professionals that work in intensive care units all over the world. Due to the binary nature of the response variable, logistic regression models were a natural choice for modelling this kind of data. The objective of this study is to compare the performance of generalized linear models (GLMs) with binary response (McCullagh and Nelder, Generalized Linear Models. Chapman and Hall, London, 1989), with the performance of generalized additive models (GAMs) with binary response (Hastie and Tibshirani, Generalized Additive Models. Chapman and Hall, New York, 1990) and also with the performance of artificial neural networks (ANNs) (Bishop, Neural Networks for Pattern Recognition. Clarendon Press, Oxford, 1995), in what concerns their predictive and discriminative power. A dataset of 996 patients was collected and the entire sample was used for the development

A.L. Papoila (✉)
CEAUL and Faculdade de Ciências Médicas da, Universidade Nova de Lisboa, Campo Mártires da Pátria 130 1169-056 Lisboa, Portugal
e-mail: ana.papoila@fcm.unl.pt

C. Rocha
CEAUL and Faculdade de Ciências, Universidade de Lisboa, Campo Grande 1149-016 Lisboa, Portugal
e-mail: cmrocha@fc.ul.pt

C. Geraldes
Universidade Nova de Lisboa, Campo Mártires da Pátria 130 1169-056 Lisboa, Portugal
e-mail: carlos.geraldes@fcm.unl.pt

P. Xufre
CIOUL and Instituto Superior de Estatística e Gestão de Informação da, Universidade Nova de Lisboa, Campus de Campolide 1070-312 Lisboa, Portugal
e-mail: pxufre@fe.unl.pt

J. Lita da Silva et al. (eds.), *Advances in Regression, Survival Analysis, Extreme Values, Markov Processes and Other Statistical Applications*, Studies in Theoretical and Applied Statistics, DOI 10.1007/978-3-642-34904-1_33,

of the models and also for the validation process, due to the nonexistence of an external, independent dataset. The performance of the proposed methodologies was assessed, not only by the evaluation of the agreement between observed mortality and predicted probabilities of death through the use of calibration plots, but also by their discriminating ability, measured by the area under the receiver operating characteristic (ROC) curve.

1 Introduction

Since 1981 numerical scoring systems and multivariable statistical models have been used to assess the severity of illness of critically ill patients. The former assign, subjectively, weights to variables reflecting the degree of physiologic derangement. In fact, the acute physiology and chronic health evaluation score, referred to as APACHE [4], the simplified acute physiology score, referred to as SAPS [7] and the APACHE II score [5] were built using a panel of experts to select variables and weights. More recently, and because the subjectivity of these procedures may lead to undesired discrepancies, multivariable statistical models were considered. Mortality probability models, referred to as MPM [9–11], the APACHE III score [6] and the SAPS II [8], were then developed, making use of more objective methods such as multiple logistic regression. However, the fact that a non-linear dependence between the binary response variable and the continuous covariates may exist led us to adopt generalized additive models (GAMs) to accomplish the fitting process. In fact, the more recently developed severity of illness scores, SAPS 3 [14, 15] and APACHE IV [17] also make use of more flexible strategies, such as splines and regression trees, to model the data. So, in this chapter, we propose the use of GAMs to estimate the probabilities of death and/or to obtain new adjusted cut-off points with the purpose of categorizing the continuous independent variables, if the main interest is the obtainment of a severity of illness score. SAPS II variables were used because this was the severity of illness score adopted by the clinicians of the Portuguese intensive care unit (ICU) where the dataset analysed in the present study was collected. Since artificial neural networks (ANNs) are an alternative to some statistical methodologies, namely, regression models [16], this study also aims to evaluate the performance of ANNs to predict the outcome under study. Finally, a comparison of the several approaches was carried out.

All statistical analyses were performed using S-PLUS (version 8.0, 2007; Insightful Corporation, Seattle, WA) and, to implement the ANNs, a new software was developed using a standard commercially available mathematics package format (MATLAB R2006b, The Math-Works Inc., 3 Apple Hill Drive, Natick, MA 01760).

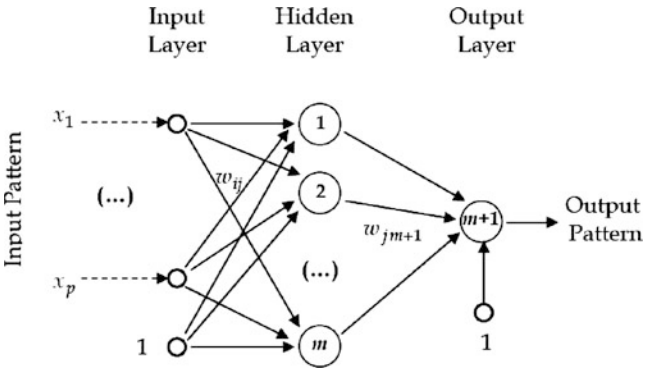

Fig. 1 Multi-layer perceptron architecture

2 Generalized Linear Models and Generalized Additive Models

Let Y be a response variable and $(X_1, \ldots, X_p)$ a vector of p associated covariates that characterize each of n individuals. A GAM is defined by the expression $E(Y|X_1, \ldots, X_p) = h(\beta_0 + \sum_{j=1}^{p} f_j(X_j))$, where Y has a probability mass or density function that belongs to the exponential family, $h(\cdot)$ is the link function and $f_j(X_j)$, $j = 1, \ldots, p$, known as the partial functions, are arbitrary univariable functions that must be estimated from the data and represent the effect of the covariates on the response [3]. A generalized linear model (GLM) is a particular case of a GAM when $f_j(X_j) = \beta_j X_j$ [12].

3 Artificial Neural Networks

An ANN is, fundamentally, a mathematical model composed by a set of units (nodes), where information is processed [2]. These units are connected through unidirectional communication links, which carry numerical data. One of the most studied and used ANN architecture is the multi-layer perceptron (MLP). Fundamentally, one MLP consists of an input–output network, which has the neurons distributed by several layers, fully connected between adjacent layers, and where the flow of information is done in a feed-forward way. The following figure shows an MLP with three layers: an input layer, without neurons, a hidden layer and a layer with one output neuron.

If we have an MLP such as the one represented in Fig. 1 and with the same activation function, f, in all its neurons, then it can be described mathematically as

$$y(x) = f(\omega_0^T f(\omega_H^T x)),$$

where x is the input pattern and ω_0 and ω_H are the matrices of the parameters related with the links of the output and hidden layers, respectively. As it can be seen from the equation above, this is a relatively complex model since it is non-linear in the parameters. Therefore, it is difficult to identify and estimate it correctly. The method traditionally used to perform the training of such networks is the error backpropagation algorithm [2], which consists of a variant of the instantaneous gradient descent procedure. The network is trained, using the steepest descent algorithm, in order to minimize an error such as the mean squared error (MSE) given by

$$MSE \equiv E_N = \frac{1}{2N}\sum_x (e(x))^2 = \frac{1}{2N}\sum_x (y(x) - d(x))^2,$$

where $d(x)$ corresponds to the desired output for the input pattern x and N is the number of individuals of the training dataset. It can be viewed as a sort of non-linear and non-parametric regression. The updating of the synaptic weights is

$$\omega = \omega - \alpha \frac{\partial E_N}{\partial \omega_{ij}},$$

where α is the learning rate. However, this kind of searching methods does not guarantee convergence of the objective function to a global minimum, and the convergence rate is typically very slow during most of the training process. To help in both respects, it is common to consider the inclusion of a momentum term in the weights updates:

$$\Delta\omega_{ij}^{(k)} = -\alpha \frac{\partial E_N}{\partial \omega_{ij}} + \beta \Delta\omega_{ij}^{(k-1)}.$$

4 The New Simplified Acute Physiology Score (SAPS II)

The SAPS II is a severity of illness score, used in ICUs, that has received a lot of attention in Europe for its simplicity and applicability. It includes 17 variables: 12 physiology variables (heart rate, systolic blood pressure, body temperature, the ratio $\frac{PaO_2}{FiO_2}$ for ventilated patients, urinary output, serum urea level, white blood cells count, serum potassium, serum sodium level, serum bicarbonate level, bilirubin level and Glasgow coma score), age, type of admission (scheduled surgical, unscheduled surgical or medical) and three underlying disease variables (acquired immunodeficiency syndrome, metastatic cancer and hematologic malignancy). To develop and validate this score, a large international sample of surgical and medical patients, collected by an European/North American multicentre study, was used [8]. The development phase used 65 % of the available patients, randomly selected, while the remaining 35 % became the validation set. The cut-off points for each of the continuous covariates that revealed to be statistically significant in the univariable

analysis were found by using the LOWESS (locally weighted scatterplot smoothing) technique. After the categories were defined, a multiple logistic regression was used and the total severity score was obtained by adding the estimated coefficients of the corresponding design variables multiplied by 10 and rounded off to the nearest integer. Finally, for converting the SAPS II into a probability of hospital mortality, a multiple logistic regression model was fitted with SAPS II and ln(SAPS II + 1) as independent variables. However, when applied to different populations, this model is often unable to adequately predict the outcome, and so, a customization may be done by fitting that model to the new datasets.

Model calibration was evaluated by analysing the agreement between the estimated probabilities of death and the actual observed mortality using the Hosmer–Lemeshow goodness-of-fit test, having obtained a p-value = 0.104 for the validation sample. To evaluate the ability of the model to distinguish between patients who live from patients who die, usually referred to as discrimination, receiver operating characteristic (ROC) curves were used and an area under the curve of 0.86 was achieved for the validation sample. Indeed, both results are highly satisfactory; however, when SAPS II was applied to some external databases, the results obtained were far worse (e.g. [1, 13]).

5 Results

Data from 996 patients, consecutively admitted to a Portuguese mixed (medical and surgical) ICU, were analysed. All SAPS data were collected during the first 24 hours after ICU admission. The mean age of the patients was 60.3 (95 % C.I. : 59.3,61.4) years with a median SAPS score of 41 (interquartile range 20–60) and a hospital mortality of 36 %. The original SAPS II scoring system did not produce very good results, namely, in what concerns calibration (p-value < 0.001) (Fig. 2, left), although an area under the ROC curve of 0.82 (95 % C.I. : 0.79, 0.84) was achieved, showing a satisfactory discrimination ability. After customization, by using a logistic regression model with SAPS II and ln(SAPS II + 1) as independent variables, a new equation for the hospital mortality prediction was derived and a better performance was obtained (Fig. 2, right), with a p-value = 0.517 attained by the Hosmer–Lemeshow goodness-of-fit test and with the same area under the ROC curve.

The same dataset was used to implement a 3-layered perceptron with 17 input nodes, 5 hidden units, a single output node and a sigmoidal activation function. Firstly, this network was trained using the steepest descent algorithm so to minimize the MSE (Fig. 3, left). The obtained area under the ROC curve was 0.82 (95 % C.I. : 0.79, 0.84). Secondly, the Kullback–Leibler (KL) distance was used instead of the MSE criterium (Fig. 3, right) and the obtained area under the ROC curve was 0.81 (95 % C.I. : 0.78, 0.84).

At last, GAMs were used to analyse the data. Based on the partial functions estimates, we found new cut-off values for each continuous covariate adjusted by the remaining covariates and we fitted a logistic regression model with these new categorical independent variables (Fig. 4, left).

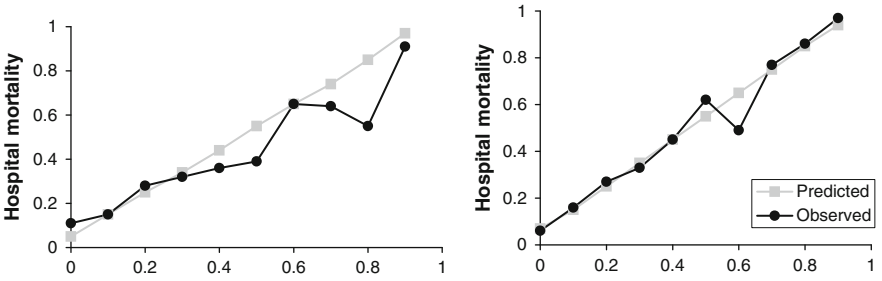

Fig. 2 Predicted versus observed probability of death. Original SAPS II (*left*) and customized SAPS II (*right*)

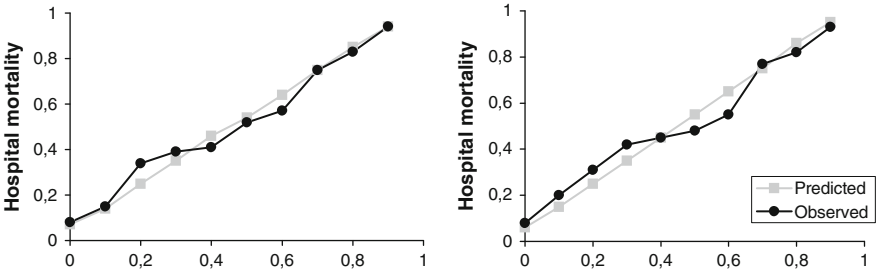

Fig. 3 Predicted versus observed probability of death. Artificial neural network using MSE (*left*) and using the Kullback–Leibler distance (*right*)

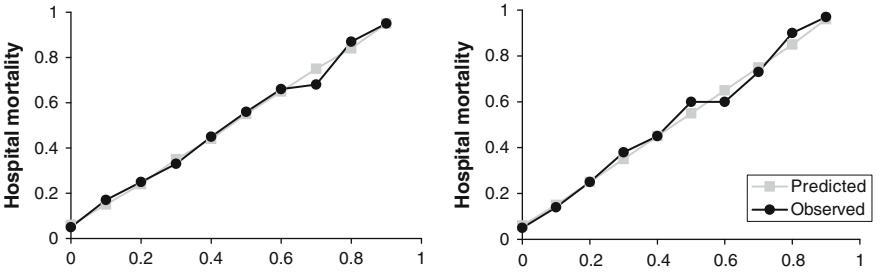

Fig. 4 Predicted versus observed probability of death. Logistic regression with the new categorical covariates (*left*) and a GAM without categorizing the continuous covariates (*right*)

The entire sample was used for model estimation and validation was accomplished by randomly splitting the whole sample into five mutually exclusive groups. Five regression models were then fitted, with each model excluding one group and used to calculate predictions for the excluded group (fivefold cross validation). An area below the ROC curve equal to 0.85 (95 % C.I. : 0.82, 0.87) and a calibration p-value $=$ 0.74 were obtained. The substantial improvements in both calibration and discrimination, even without introducing new prognostic variables, were

interesting findings. However, since some information is lost in the categorization process, we also used GAMs to estimate the probabilities of death without categorizing the continuous covariates. After fitting a GAM to our cross-validated sample, good calibration curves (Fig. 4, right) and an area under the ROC curve of 0.87 (95 % C.I. : 0.85, 0.89) were obtained. As it can be seen from Fig. 4, GAMs obtained better results than those presented by the other approaches.

6 Conclusions and Future Work

The performance of GAMs is clearly superior to the GLMs and neural networks used in this study. When comparing these last two approaches, in what concerns their discriminative power, results are according to the ones referred elsewhere (no substantial differences between the areas under the ROC curve). The same did not happen for the predictive power since neural network calibration plots showed a weaker performance, independently of the used criterium (MSE or KL distance). This means that, in our study, there was no relevant advantage in using ANN-MLPs. As future work, other ANN structures, such as Generalized Additive Neural Networks, will be implemented with the purpose of obtaining better results.

Acknowledgements Ana Luisa Papoila's and Cristina Rocha's research is partially sponsored by national funds through the Fundação Nacional para a Ciência e Tecnologia, Portugal—FCT—under the project PEst-OE/MAT/UI0006/2011. The authors would like to thank the medical team of UCIP from the Centro Hospitalar de Lisboa Central (Lisbon), in particular to Dr. Eduardo Silva and Dr. Miguel Robalo, for their collaboration in data collection.

References

1. Apolone, G., Bertolini, G., D'Amico, R., et al.: The performance of SAPS II in a cohort of patients admitted to 99 ICUs: Results from GiViTI. Int. Care Med. **22**, 1368–1378 (1996)
2. Bishop, C.M.: Neural Networks for Pattern Recognition. Clarendon Press, Oxford (1995)
3. Hastie, T., Tibshirani, R.: Generalized Additive Models. Chapman and Hall, New York (1990)
4. Knaus, W.A., Zimmerman, J.E., Wagner, D.P., et al.: APACHE-Acute physiology and chronic health evaluation: A physiologically based classification system. Crit. Care Med. **9**, 591–597 (1981)
5. Knaus, W.A., Draper, E.A., Wagner, D.P., Zimmerman, J.E.: APACHE II: A severity of disease classification system. Crit. Care Med. **13**, 818–829 (1985)
6. Knaus, W.A., Wagner, D.P., Draper, E.A., et al.: The APACHE III prognostic system: Risk prediction of hospital mortality for critically ill hospitalized adults. Chest **100**, 1619–1636 (1991)
7. Le Gall, J.-R., Loirat, P., Alperovitch, A., et al.: A simplified acute physiology score for ICU patients. Crit. Care Med. **12**, 975–977 (1984)
8. Le Gall, J.-R., Lemeshow, S., Saulmier, F.: A New Simplified Acute Physiology Score (SAPS II) based on an European/North American multicenter study. JAMA vol. 270 **24**, 2957–2963 (1993)
9. Lemeshow, S., Teres, D., Pastides, H., et al.: A method for predicting survival and mortality of ICU patients using objectively derived weights. Crit. Care Med. **13**, 519–525 (1985)

10. Lemeshow, S., Teres, D., Avrunin, J.S., Gage, R.W.: Refining intensive care unit outcome prediction by using changing probabilities of mortality. Crit. Care Med. **1**, 470–477 (1988)
11. Lemeshow, S., Teres, D., Klar, J., et al.: Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. JAMA vol. 270, **20**, 2478–2486 (1993)
12. McCullagh, P., Nelder, J.: Generalized Linear Models, 2nd edn. Chapman and Hall, London (1989)
13. Metnitz, P.G.H., Valentin, A., Vesely, H., et al.: Prognostic performance and customization of the SAPS II: Results of a multicenter Austrian study. Int. Care Med. **25**, 192–197 (1999)
14. Metnitz, P.G.H., Moreno, R.P., Almeida, E., et al.: SAPS 3: From evaluation of the patient to evaluation of the intensive care unit, Part 1: Objectives, methods and cohort description. Int. Care Med. **31**, 1336–1344 (2005)
15. Moreno, R.P., Metnitz, P.G.H., Almeida, E., et al.: SAPS 3: From evaluation of the patient to evaluation of the intensive care unit, Part 2: Development of a prognostic model for hospital mortality at ICU admission. Int. Care Med. **31**, 1345–1355 (2005)
16. Schumacher, M., Roβner, R., Vach W.: Neural networks and logistic regression: Part I. Elsevier, Comput. Stat. Data An. **21**, 661–682 (1996)
17. Zimmerman, J.E., Kramer, A.A., McNair, D.S., Malila, F.M.: Acute physiology and chronic health evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. Crit. Care Med. **34**, 1297–1310 (2006)