Monografías del Seminario Matemático García de Galdeano **33**, 245–252 (2006)

LOGISTIC REGRESSION VERSUS NEURAL NETWORKS FOR MEDICAL DATA

Luis Mariano Esteban Escaño, Gerardo Sanz Saiz, Francisco Javier López Lorente, Ángel Borque Fernando and José María Vergara Ugarriza

Abstract. In this work we consider the usefulness of classical models as Logistic regression models versus newer techniques as neural networks when they are applied to medical data. We present the difficulties appearing in the building of both types of models and their validation. For the comparison of models we have used two types of medical data that allow us to validate our models and reinforce the conclusions. Although the neural network can fit the data a little better than the logistic model, the former models are less robust than the latter. This fact together with a greater simplicity and interpretation of the variables in the logistic models makes these models preferred from the point of view of clinical applications. These results agree with other published in literature [1].

Keywords: Logistic regression, neural networks, comparison of models.

AMS classification: 62P10, 62J02.

§1. Introduction

Nowadays statistical methods constitute a very powerful tool for supporting medical decisions. The volume of medical data that any analysis or test of patients provides, makes that doctors can be helped by statistical models to interpret correctly so many data and to support their decisions. Of course, although the models are a very powerful tool for doctors these can not substitute their viewpoint.

On the other hand, the characteristics of medical data and the huge number of variables to be considered have been a fundamental point for the development of new techniques as neural networks or the techniques designed for the analysis of microarray data.

In this context one of the first and more important decisions is to choose an adequate model. Many of the medical problems are related to questions of classification and prediction, many times with only two categories (disease or not disease, for instance). In those cases the use of classical techniques has to be restricted to some specific methods as logistic regression or similar. Moreover, the complexity of medical data and the lack of structure of themselves implies that some of the methods above mentioned have to be used with care. The neural networks do not require such restrictive hypothesis, that is why they are an interesting competitor of classical techniques.

In the last years a new culture has appeared in the scientist community vindicating the generalized used of these newer tools because their supposed flexibility an applicability. A

few years ago a lot of articles appeared in the literature trying to prove the advantage of these techniques over the classical ones, however some other articles have proved that the supposed superiority is not true.

In this work we have considered two different set of medical data from different fields (Urology and Children Neurophysiology) and we have found two main conclusions:

- The choice of one or other technique is not evident and none can be assumed better than other, the right election can not be guided by fashion or specific interest, depend on the particular data.
- Each type of data require a specific technique and model, it is impossible to generalize and to say that some family of models is better than other one.

Our datasets provides illustrations of this global conclusions.

§2. The datasets

2.1. Prostate cancer

This dataset was collected at the Miguel Servet Universitary Hospital. The purpose is to predict whether a tumor is confined or not. There are two categories in the outcome, 7 numerical attributes, 5 categorical attributes, and 621 observations. The numerical attributes are Age, PSA, PSAD, PSAD-AD, prostate's volume, adenoma's volume, rate of cylinder affectionate. The categorical attributes are clinical stage, Gleason of biopsy, First Gleason, Second Gleason, Perineural invasion.

For similar or identical data, most authors used logistic regression and neural networks mainly. For comparison of models most of them used classification tables and ROC curves, we used too. One characteristic of our work is that we have considered more variables as inputs than most works, including some potential interactions between them.

2.2. Obstructive Sleep Apnea Syndrome (OSAS)

This dataset was collected at the Miguel Servet Universitary Hospital. The purpose is to predict whether a saturation of O_2 is normal or abnormal for a children population. There are two categories in the outcome, 5 numerical attributes, 22 categorical attributes and 265 observations.

First attribute is *Age*, the categorical attributes are the answers to a questionnaire that doctors gives to the parents (you can see some examples in Table 1), there are two possibles answers (yes or no) that are recoded: yes$\rightarrow$1, and no$\rightarrow$0, because doctors think that a yes answer contributes to an abnormal case, in particular, we have 9 questions type A, 7 questions type B, 6 questions type C, and finally the sum of records type A, type B, type C and the total sum.

For this data set we try other models like CART or Classification tables appeared in some articles [2, 3] with poor results. Finally Logistic Regression and Neural Networks were used, which have revealed as very similar models. This fact, confirms our conclusion that some of the ideas appeared in some papers about the general fitness of neural networks can not be supported.

While sleeping, does your child...	Does your child...	Often, your child...	Does your child...
snore more than half of the time?	wake up feeling tired?	does not seem to listen when you talk to him	usually breathe with his mouth open during the day?
snore always?	have a headache in the morning?	is badly organized to do its tasks	usually wake up with a dry mouth?
snore noisily?	Is it difficult to wake him up?	is distracted when doing something	ever wet his bed?
breathe noisily?	Is he sleepy during the day?	can't be still while sitting	
breathe with difficulty?	Does his teacher say he's sleepy at school?	can't stop moving	
Have you noticed if he has ever stopped breathing by night?	Has he ever stopped growing up normally?	interrupts others (their conversation or games)	

Table 1: Obstructive Sleep Apnea Syndrome. Examples of questions

§3. The models

Here we are going to describe the models we used to fit our data base, but first of all we must explain that in supervised learning algorithms there are a problem about over-fitting, data must be structured in subsets: training, validation and test data. Training and validation data are used to build a good model, but perhaps you are fitting noise, if this happens, prediction of new data are very bad; that's why we used another set, test data. For this data, you can compare the predictions of the model with their outcomes, that's how you can know if you model fit well new data. Our partition was 70% -30%.

Logistic Regression [4], Multilayer Perceptron [5] and Radial Basis Function Networks [5] were used in both datasets. We briefly introduce these neural networks.

A multilayer perceptron (MLP) network consists of a set of source nodes forming the input layer, one or more hidden layers of computation nodes, and an output layer of nodes. An MLP is a network of simple neurons called perceptrons. The perceptron computes a single output from multiple real-valued inputs by forming a linear combination according to its input weights and then possibly putting the output through some nonlinear activation function. This can be written as $y = \varphi(\sum_{i=1}^{n} w_i x_i + b)$ where $(w_1, \ldots, w_n)$ denotes the vector of weights, $(x_1, \ldots, x_n)$ is the vector of inputs, b is the bias and φ is the activation function. Radial Basis Function Networks have three layers: input layer, one hidden layer and an output layer. They are characterized by an hybrid learning. Although they could be similar to an MLP, they have clear differences. The nodes in the hidden layer do not calculate a weighted sum of inputs and then apply an activation function. These nodes calculate the distance between the synaptic weight vector called centroid and the input, and then apply a radial basis function to this distance.

3.1. Prostate Cancer

Logistic regression. Our best model include the inputs: Age, PSA, rate of cylinder affectionate, clinical stage, Gleason. The equation that gives us the predicted probability for a non confined tumor is $p = (1+e^{-z})^{-1}$, where

$$\begin{aligned} z = {} & 3.805 + 0.058 \cdot \textit{Age} + 0.086 \cdot \textit{PSA} + 0.013 \cdot \textit{Rate} - 2.527 \cdot \textit{Gleason(2-6)} \\ & - 1.861 \cdot \textit{Gleason(7)} - 7.121 \cdot \textit{ClinicalStage(T2a)} \\ & - 6.322 \cdot \textit{ClinicalStage(T2b)} - 8.172 \cdot \textit{ClinicalStage(T1c)}. \end{aligned}$$

Multilayer Perceptron. This is the Multilayer Perceptron neural network implemented in the program Neural Connection. We worked in two different ways, first model was building using all inputs and second using the same variables that appeared like predictive in Logistic Regression.

1. All variables: The network architecture is 24-5-2, with a nodal output activation function tanh for the hidden layer and linear for the output layer. The learning algorithm was Conjugate Gradient and 137 weights were calculated.

2. Variables LR: The network architecture is 10-3-2, with a nodal output activation function tanh for the hidden layer and linear for the output layer. The learning algorithm was Conjugate Gradient and 41 weights were calculated.

Radial Basis Function Network. We worked in the same way that MLP using the program Neural Connection again.

1. All variables: The best result was performed with 15 centers, the non-linear radial basis function used was Multi-Quadratic($\beta = 0.6$) and the error distance measure was Euclidean.

2. Variables LR: The best result was performed with 5 centers, the non-linear radial basis function used was Multi-Quadratic ($\beta = 0.1$) and the error distance measure was Euclidean.

3.2. Obstructive Sleep Apnea Syndrome

Logistic regression. We discussed the models detailed in Table 2.

Multilayer Perceptron. The network architecture is 27-5-2, with a nodal output activation function sigmoid for the hidden layer and linear for the output layer. The learning algoritm used was Steepest Descendent and 152 weights were calculated.

Radial Basis Function Network. The best result was performed with 8 centers, the non-linear radial basis function used was Thin Plate Spline and the error distance measure was City Block (Manhattan).

Included inputs	$z =$
Age, *A*1, *A*3, *A*9, *A*, *B*1, *B*6, *C*1, *C*	$-3.319 - 0.180\,Age + 1.229\,A1 - 0.924\,A3 + 0.750\,A9 + 0.524\,A$ $- 0.733\,B1 - 1.274\,B6 - 1.113\,C1 + 0.202\,C$
Age, *A*3, *A*, *B*1, *B*6	$-3.305 - 0.184\,Age - 1.031\,A3 + 0.733\,A - 0.753\,B1 - 1.128\,B6$
Age, *A*3, *A*, *B*6	$-3.094 - 0.169\,Age - 1.049\,A3 + 0.673\,A - 1.351\,B6$
Age, *A*3, *A*	$-2.945 - 0.169\,Age - 1.022\,A3 + 0.629\,A$

Table 2: OSAS. Models considered in Logistic Regression

§4. Comparison of models

As we introduced before, classification is very usual for medical data. There are two ideas that are very important for this point: sensitivity and specificity. When you consider the results of a particular test in two populations, one population positive, the other population negative, you will rarely observe a perfect separation between the two groups. Indeed, the distribution of the test results will overlap.

For every possible cut-off point or criterion value you select to discriminate between the two populations, there will be some positive cases classified as positive (TP = True Positive fraction), but some positive cases will be classified negative (FN = False Negative fraction). On the other hand, some negative cases will be correctly classified as negative (TN = True Negative fraction), but some negative cases will be classified as positive (FP = False Positive fraction).

Sensitivity is the probability that a test result will be positive when the category is positive (true positive rate, expressed as a percentage). Specificity is the probability that a test result will be negative when the category is negative (true negative rate, expressed as a percentage). Our purpose is 100% sensitivity and specificity, but this point is not present in real databases, we must search for the best combination of them that can be compatible.

In a Receiver Operating Characteristic Curve (ROC) Sensitivity is plotted as function of 1-Specificity for different cut-off probability points. Each point on the ROC plot represents a sensitivity/specificity pair corresponding to a particular decision threshold. The area under the ROC curve is a measure of how well a parameter can distinguish between two diagnostic groups (positive/negative).

When the variable under study can not distinguish between the two groups, i.e. where there is no difference between the two distributions, the area will be equal to 0.5 (the ROC curve will coincide with the diagonal). When there is a perfect separation of the values of the two groups, i.e. there no overlapping of the distributions, the area under the ROC curve equals 1 (the ROC curve will reach the upper left corner of the plot).

For comparison of models, we used the test developed by Hanley & McNeil, 1983, [6] that study significance of the difference between the areas under ROC Curves from identical samples. The null hypothesis is that the areas are equal, and the alternative hypothesis is that the areas are different.

Prostate cancer. The pictures in Figure 1 show the curves for our data base. Likewise, in Tables 3 and 4 we can see the area under ROC Curves in the first row and the significance of

the Hanley & McNeil test in the middle of them. As we can see there, some model of Neural Network can be considered better with a significance of 0.05 in training+validation data, but none of them still better in test data, this confirm our proposal.

Obstructive Sleep Apnea Syndrome. The ROC Curves for this data base are given in Figure 2. Again the area under ROC Curves is in the first row and the significance of the Hanley & McNeil test is in the middle of them (cf. Tables 5 and 6). The comparison of area under ROC Curves gives us models with no significative difference in training-validation data and test data.

§5. Conclusion

Several points must be remarked:

- For the data we used to build the model, some of the Neural Networks can be considered better with a significance of 0.05 in the area under ROC curve.
- For the data we used to validate the model none of them are better.
- The Neural Networks models that can be considered better in building data include a greater number of inputs.
- This can be a clear difficulty to diagnose, the complexity of these models makes that usually doctors use logistic regression which includes less variables and have an easy meaning in the model.
- Prediction can be done with a simple equation in Logistic Regression and it needs a computer tool in Neural Network, this can delay the diagnose for the doctors.
- Based on these, we can say that none of the models have reveal better results, it must be choose by its application, for medical data we recommend Logistic Regression.

References

[1] BORQUE, A., SANZ, G., ALLEPUZ, C., PLAZA, L., GIL, P., AND RIOJA, L. A. The use of neural networks and logistic regression analysis for predicting pathological stage in men undergoing radical prostatectomy: a population based study. *J. Urol. 166* (2001), 1672–8.

[2] CHERVIN, R. D., HEDGER, K., DILLON, J. E., AND PITUCH, K. J. Pediatric Sleep Questionnaire (PSQ): validity and reliability of scales for sleep-disordered breathing, snoring, sleepiness, and behavioral problems. *Sleep Medicine I*, (2000), 21–32.

[3] CHAU, K. W., NG, D. K. K., KWOK, C. K. L., CHOW, P.Y. Clinical risk factors for obstructive sleep apnea in children. *Ho. Singapore Med J. 44*, 11 (2003), 570–573.

[4] DAVID W. HOSMER, D. W., AND LEMESHOW, S. *Applied Logistic Regression*, 2nd edition. Wiley Series in Probability and Statistics, 2000.

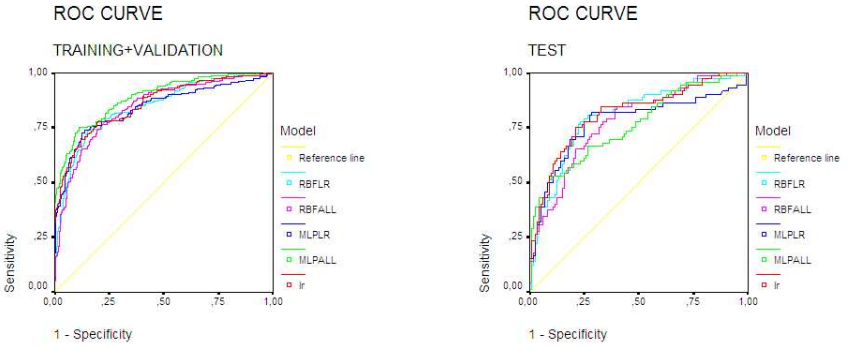

Figure 1: Prostate cancer. ROC curves

Model	LR	MLP all	MLP lr	RBF all	RBF lr
Area	0.858	0.889	0.841	0.843	0.845
95%C.I.	0.821-0.889	0.858-0.917	0.803-0.879	0.805-0.876	0.807-0.878
LR	1	*0.035*	0.208	0.283	0.257
MLP all	*0.035*	1	*0.004*	*0.007*	*0.008*
MLP lr	0.208	*0.004*	1	0.924	0.820
RBF all	0.283	*0.007*	0.924	1	0.895
RBF lr	0.257	*0.008*	0.820	0.895	1

Table 3: Prostate cancer. Comparison of ROC curves for 70% Training+Validation data

Model	LR	MLP all	MLP lr	RBF all	RBF lr
Area	0.813	0.759	0.774	0.777	0.803
95%C.I.	0.744-0.866	0.691-0.819	0.707-0.832	0.710-0.835	0.738-0.857
LR	1	*0.034*	0.063	0.139	0.581
MLP all	*0.034*	1	0.637	0.550	0.111
MLP lr	0.063	0.637	1	0.920	0.233
RBF all	0.139	0.550	0.920	1	0.288
RBF lr	0.581	0.111	0.233	0.288	1

Table 4: Prostate cancer. Comparison of ROC curves for 30% Test data

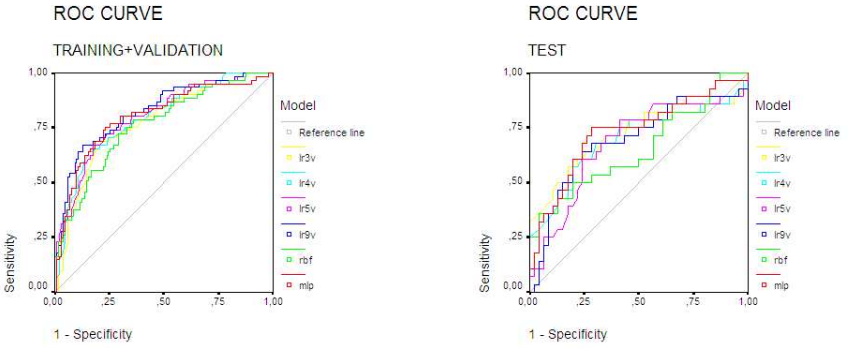

Figure 2: OSAS. ROC curves

Model	MLP	RBF	LR9v	LR5v	LR4v	LR3v
Area	0.803	0.762	0.825	0.801	0.792	0.774
95%C.I.	0.739-0.857	0.695-0.821	0.763-0.876	0.737-0.855	0.728-0.847	0.708-0.831
MLP	1	0.220	0.340	0.928	0.645	0.208
RBF	0.220	1	0.095	0.268	0.416	0.744
LR9v	0.340	0.095	1	0.213	0.195	0.068
LR5v	0.928	0.268	0.213	1	0.541	0.204
LR4v	0.645	0.416	0.154	0.541	1	0.238
LR3v	0.208	0.744	0.068	0.204	0.238	1

Table 5: OSAS. Comparison of ROC curves for 70% Training+Validation data

Model	MLP	RBF	LR9v	LR5v	LR4v	LR3v
Area	0.729	0.655	0.691	0.678	0.698	0.715
95%C.I.	0.613-0.826	0.535-0.761	0.573-0.793	0.559-0.782	0.580-0.799	0.599-0.814
MLP	1	0.183	0.342	0.141	0.445	0.690
RBF	0.183	1	0.547	0.654	0.475	0.289
LR9v	0.342	0.547	1	0.701	0.878	0.598
LR5v	0.141	0.654	0.701	1	0.463	0.349
LR4v	0.445	0.475	0.878	0.463	1	0.572
LR3v	0.690	0.289	0.598	0.349	0.572	1

Table 6: OSAS. Comparison of ROC curves for 30% Test data

[5] HAYKIN, S. *Neural Networks: A Comprehensive Foundation*, 2nd edition. Prentice Hall, 1999.

[6] HANLEY, J. A., AND MCNEIL, B. J. A method of comparing the areas under receiver operating characteristics curves derived from the same cases. *Radiology 148* (1983), 839–843.

Luis Mariano Esteban Escaño
Applied Mathematics Department. School of Engineering La Almunia.
C/ Mayor s/n. 50100 La Almunia de Doña Godina (Zaragoza).
luis.esteban@eupla.unizar.es

Gerardo Sanz Saiz and Francisco Javier Lopez Lorente.
Statistics Department. University of Zaragoza.
C/ Pedro Cerbuna 12. Building Mathematics. 50009 Zaragoza.
gerardo@unizar.es and javierl@unizar.es

Ángel Borque Fernando and José María Vergara Ugarriza.
Miguel Servet Universitary Hospital. Isabel la Católica 1-3. 50009 Zaragoza.
aborque@comz.org and vergeur@comz.org