MEDINFO 98
B. Cesnik et al. (Eds)
Amsterdam: IOS Press

Diagnosing Breast Cancer from FNAs: Variable Relevance in Neural Network and Logistic Regression Models

Lucila Ohno-Machado and Donald Bialek

Decision Systems Group, Brigham and Women's Hospital and Harvard Medical School, Boston, USA

Abstract

We compared the selection of variables for building a classification model for the diagnosis of breast cancer using neural networks and logistic regression. A set of 460 cases was used to build neural network and logistic regression models that classify cell samples obtained by fine-needle aspiration (FNA) as malignant or benign, depending on nine pathology features. Variables selected by a step down logistic regression model were compared to those selected by a measure of relevance derived from neural network weights. Since both types of models resulted in similar predictive accuracy, we expected approximately the same variables to be selected. The variables with the highest relevance values for the neural network models corresponded to those of high significance in univariate logistic regression models, but were not the ones selected in the step down procedure of multivariate models. Variable relevance based on weights for neural network models does not seem to be a consistent index of the importance of that variable for multivariate models such as logistic regression.

Keywords

Breast Cancer; Logistic Regression; Neural Networks

Introduction

Although the use of neural networks in medical problems has increased considerably since the publication of the backpropagation algorithm in 1986 [1], this type of models still receive severe criticism for being "black boxes" that provide little help for researchers who want to be able to better understand the processes being modeled [2]. One of the main advantages of using neural network models is that they can model several types of non-linear phenomena, while classic statistic models are usually limited to linear or a limited number of non-linear models. One of the main drawbacks of neural network models is their inability to explain the relative importance of each variable in the solution of a problem. For example, in a classification problem that models the diagnosis of coronary heart disease from a set of findings, such as a certain type of chest pain, EKG features, etc., a neural network model may be an excellent predictor of disease, but may not be able to explain which findings were most relevant in reaching the diagnosis.

Attempts to devise a method for determining the importance of each input variable were tried in the medical domain [3], but they usually did not consider the possible interactions of a variable to other variables. For example, some findings may only be significant when they appear together. One method that tries to consider the relevance of a variable in the overall context of the model uses an index of relevance R

$$R = \Sigma (w_i)^2 / \Sigma (w)^2 \quad (1)$$

where w_i is the sum of squares of weights connecting input unit x_i to output unit y, and w is the sum of squares of all weights.

This index of relevance is implemented in the software Nevprop [4]. In this article, we describe an experiment that checks whether this index produces results that are equivalent to those of step down selection of variables in logistic regression, commonly implemented in statistical packages. We used the domain of breast cancer diagnosis from fine-needle aspirates for this experiment.

Carcinoma of the breast is one of the more common malignancies primarily affecting women. The ability to detect the tumor in its early stages in a minimally invasive way is important to the treatment of patients with this disease. Examination of the breast using techniques such as mammography can establish the presence of small masses within the breast tissue. To determine whether these masses are malignant, a fine needle is placed in the mass using radiographic guidance in order to aspirate a small tissue sample from the mass. Cytological examination of the cells obtained via fine needle aspiration is helpful in establishing whether the tumor mass is benign or malignant (carcinoma). If this examination shows the mass to be malignant, then open biopsy and invasive treatment is warranted; if benign, then periodic examination is indicated. It is important that cytological examination of the FNA be accurate in predicting whether the tumor is malignant. There is no one cytological characteristic of the FNA that is totally accurate in predicting whether the mass is a carcinoma; however with the use of several characteristics, it is possible that the diagnosis of malignancy can reliably be made.

Materials and Methods

We used a sample of cases from the University of Wisconsin

Hospitals, Madison, available at the UCI Repository of machine learning databases [5]. From the available 687 cases with no missing attributes, we randomly selected two thirds (460) to compose the training set and one third to compose the test set (227). The cases are classified as malignant (65.5%) and benign (34.5%). Nine pathology features characterize each case: (1) clump thickness, (2) uniformity of cell size, (3) uniformity of cell shape, (4) marginal adhesion, (5) single epithelial cell size, (6) bare nuclei, (7) bland chromatin, (8) normal nucleoli, and (9) mitoses. Each of these features was graded in a one to ten scale. Part of this data set was used by Wolberg to test mathematical approaches to classification based on linear programming [6]. Wolberg and colleagues showed that the data set contained information that allowed efficient classification of the cases [7,8,9], but has not elaborated on variable relevance.

We built logistic regression models using STATA and we used a backward selection procedure based on Wald's statistic [10]. We constructed a feedforward neural network with nine input nodes (which corresponded to the nine independent variables of the logistic regression model), five hidden nodes, and one output node (which corresponded to the dependent variable, "malignancy," of the logistic regression model). The network was trained by backpropagation. Variable relevance was determined by comparing the weights related a certain input to the overall weights.

Results

Logistic Regression Model

Univariate Analysis

Analysis of the univariate models indicated that each variable was significant at the 0.05 level, as shown in Table 1. The likelihood ratio test compares each univariate model to the intercept only model.

Multivariate Analysis

Since each of the explanatory variables was a significant predictor of malignancy in the univariate analysis, we began by building a multivariate model that included all of these variables entered as continuous variables scaled 1-10. When all variables were taken together in this multivariate model, the p-values from the Wald test lost significant for some of the variables (uniformity of size, uniformity of shape, single epithelial cell size, and normal nucleoli; alpha=0.05). Using a step down procedure, we removed the variable from the model with the largest Wald test p-value. The large p-value indicated that this variable was the least significant predictor of malignancy in the model including all variables. Single epithelial cell size had the largest p-value (0.646) and was deleted. Using the likelihood ratio test, we compared the model excluding single epithelial cell size to the original model containing all variables. The likelihood ratio test statistic was not significant on a chi squared distribution with one degree of freedom, leading us to conclude that single epithelial cell size did not significantly contribute to the model. We, therefore, chose the more parsimonious model that excluded single epithelial cell size.

†Covariates: Clump thickness (clump), Uniformity of cell size (usize), Uniformity of cell shape (ushape), Marginal adhesion (adhes), Single epithelial cell size (csize), Bare nuclei (bare), Bland chromatin (bland), Normal nucleoli (nucleoli), Mitoses (mitoses).

Table 1 - Univariate logistic regression.

Variable†	b*	SE(b)*	Log-likelihood	G*
constant	-0.706	0.098	-299.502	
clump	0.870	0.083	-161.582	275.84
usize	1.542	0.152	-91.371	416.26
ushape	1.344	0.127	-100.317	398.37
adhes	1.023	0.106	-165.527	267.95
csize	1.194	0.120	-173.128	252.75
bare	0.828	0.084	-119.297	347.92
bland	1.384	0.141	-132.539	333.93
nucleoli	0.915	0.096	-165.967	267.07
mitoses	1.223	0.199	-247.297	104.41

* b, estimated beta coefficient; SE, standard error; CI, confidence interval; G, likelihood ratio test statistic.

Based on our new model, we again excluded the variable with the largest Wald test p-value (*uniformity of cell size*, p=0.667). The likelihood ratio test was not significant in this case also, suggesting that uniformity of cell size need not be included in our model.

Table 2 - Multivariate logistic regression.

Explanatory variable	Adjusted OR* estimate	95% CI*
Clump thickness	1.58	1.14-2.18
Uniformity of cell size	-	-
Uniformity of cell shape	1.66	1.13-2.43
Marginal adhesion	1.63	1.20-2.21
Single epithelial cell size	-	-
Bare nuclei	1.40	1.12-1.76
Bland chromatin	1.97	1.29-3.01
Normal nucleoli	-	-
Mitoses	2.07	1.10-3.89

* OR, odds ratio; CI, confidence interval

We had similar results for the next variable we removed, *nucleoli* (G=1.17, p=0.279). Of all the variables left, *mitoses* had the largest Wald test p-value. Comparing the model without *mitoses* to the model containing *mitoses* produced a significant likelihood ratio test statistic (G=5.844, p=0.016). Therefore, *mitoses* was an important variable that significantly added to the model. Deciding to retain *mitoses* in our model, at this point we ended our step down procedure. The Hosmer-Lemeshow goodness-of-fit test produced a chi squared value of 3.25 with a p-value of 0.9180 with 8 degrees of freedom. This model had

acceptable fit.

Neural Network Model

The neural network model had a good fit and good predictive performance, as indicated by areas under the ROC curves of .9901 and .9795, respectively.

Over fitting is a serious problems in neural networks [11]. In order to reduce it, one half of the training set was used as a "hold-out" subset. We stopped training when error in the hold-out set started to increase. Thirty different randomly selected holdout sets were used to determine the ideal minimum error. The ideal minimum error consisted of the average minimum error of the thirty sets. The whole training set (with no holdout set) was used to build the final neural network model, which was trained until the average minimum error was achieved.

Weight relevance

Table 3 lists the variables, R* (the average relevance of each variable in the thirty runs), and R† (the relevance of each variable in the final model). The variables with highest R* and R† were uniformity of cell shape, single epithelial cell size, and bare nuclei. Variables with low R* and R† were mitoses, and marginal adhesion. The relative ranking of other variables varied for R* and R†.

Table 3 - Variable relevance in neural network models

Explanatory variable	*R**	R†
Clump thickness	0.0618	0.0583
Uniformity of cell size	0.2189	0.3441
Uniformity of cell shape	0.1012	0.0567
Marginal adhesion	0.0630	0.0390
Single epithelial cell size	0.1256	0.1236
Bare nuclei	0.2143	0.1741
Bland chromatin	0.0860	0.1076
Normal nucleoli	0.0773	0.0836
Mitoses	0.0519	0.0130

*Average relevance of 30 models in which 30 different holdout sets were used.

†Relevance of model trained until error reached average minimum in the holdout sets of 30 models

Discussion

Variables selected by the step down procedure based on the odds ratio in the logistic regression model did not have high relevance in the neural network model, and the reverse was also true.

A variable considered particularly important for the determination of malignancy, according to an independent pathologist, is the number of mitoses. This variable was the least relevant in the neural network model. *Uniformity of cell size*, which was the first variable to be removed from the multivariate logistic regression model, was the most relevant in the neural network model and in the univariate logistic regression. We concluded that the measure of variable relevance in the neural network model is not a good indicator of the importance of the variable in this classification problem. Providing measures of relevance does not help to explain the importance of variables in this problem, nor helps the researcher to get any insight on the model. The step down procedure used in the multivariate regression model is easier to understand and provides more insight in this problem.

The study of variable importance in neural networks is a subject of active research. As mentioned previously, attempts to select variables based on univariate analyses have been made. The main advantage of neural networks is, however, its ability to integrate multivariate information with little need to transform variables or predetermine interactions. The explanatory capability of neural networks needs to be derived from indices that take into account these factors. Univariate analyses of variable relevance are not sufficient.

Acknowledgments

We thank Dr. Dan Schwartz, Dr. Ahsan Arozullah, Ms. Melissa Johnson, and Mr. Yoninah Segal for their help with the logistic regression models.

References

[1] Rumelhart DE; Hinton GE; Williams RJ. Learning internal representation by error propagation. In Rumelhart, D.E., and McClelland, J.L. (eds) *Parallel Distributed Processing*. MIT Press, Cambridge, 1986.

[2] Cheng B; Titterington DM. Neural networks: A review from a statistical perspective. *Statistical Science 1994: 9(1):2–54.*

[3] Baxt WG. Analysis of the clinical variables driving decision in an artificial neural network trained to identify the presence of myocardial infarction. *Annals of Emergency Medicine* 1992: Dec, 21(12):1439–44.

[4] Goodman P. *NevProp 3.0. Reno: University of Nevada,* 1997.

[5] Merz CJ, Murphy PM. (1996*). UCI Repository of machine learning databases.* Irvine: University of California, Department of Information and Computer Science, 1996.

[6] Wolberg WH, Mangasarian OL. Multisurface Method of Pattern Separation for Medical Diagnosis Applied to Breast Cytology. *Proc Nat Acad Sci* 1990: 87: 9193-9196.

[7] Wolberg WH; Street WN; Heisey DM; Mangasarian OL.Computerized breast cancer diagnosis and prognosis from fine-needle aspirates. *Archives of Surgery* 1995: May, 130(5):511-6.

[8] Wolberg WH, Street WN, Mangasarian OL.Image analysis and machine learning applied to breast cancer diagnosis and prognosis. *Analytical and Quantitative Cytology and Histology* 1995: Apr, 17(2):77-87.

[9] Wolberg WH; Street WN; Mangasarian OL. Machine learning techniques to diagnose breast cancer from image-

processed nuclear features of fine needle aspirates. *Cancer Letters* 1994: Mar 15, 77(2-3):163-71.

[10] SAS Institute. SAS/STAT *User's Guide, Version 6,* Fourth Edition. SAS Institute Inc., Cary, 1990.

[11] Bishop CM. Neural *Networks for Pattern Recognition.* Oxford: Oxford University Press, 1996.

Address for correspondence

Lucila Ohno-Machado
Decision Systems Group, Birgham and Women's Hospital
75 Francis Street
Boston, MA 02174
Email: machado@dsg.harvard.edu