

A new approach to training back-propagation artificial neural networks: empirical evaluation on ten data sets from clinical studies

Antonio Ciampi^{*,†} and Fulin Zhang

*Department of Epidemiology and Biostatistics, McGill University, 1020 Pine Avenue West,
Montreal, P.Q., H3A 1A2 Canada*

SUMMARY

We present a new approach to training back-propagation artificial neural nets (BP-ANN) based on regularization and cross-validation and on initialization by a logistic regression (LR) model. The new approach is expected to produce a BP-ANN predictor at least as good as the LR-based one. We have applied the approach to ten data sets of biomedical interest and systematically compared BP-ANN and LR. In all data sets, taking deviance as criterion, the BP-ANN predictor outperforms the LR predictor used in the initialization, and in six cases the improvement is statistically significant. The other evaluation criteria used (C-index, MSE and error rate) yield variable results, but, on the whole, confirm that, in practical situations of clinical interest, proper training may significantly improve the predictive performance of a BP-ANN. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: prediction; logistic regression; training; regularisation; cross-validation

1. INTRODUCTION

Artificial neural networks (ANN) are becoming an increasingly popular tool for data analysis in many areas of science and technology. In recent years they have been used to analyse data from a variety of human clinical studies, often in conjunction with more traditional data analytic approaches. By far the most used type of ANN is the back-propagation ANN (BP-ANN), and the work presented here is limited to this family of ANN. *Prediction*, as opposed to *explanation*, is the area where BP-ANN are considered to have an *a priori* advantage on traditional statistical modelling: see, for example, Tu [1]. Thus, not surprisingly, in most of these biomedical applications the central problem is to predict a binary outcome from a prespecified set of covariates. The prediction problem is solved by constructing from

*Correspondence to: Antonio Ciampi, Department of Epidemiology and Biostatistics, McGill University, 1020 Pine Avenue West, Montreal, P.Q., H3A 1A2 Canada

†E-mail: antonio.ciampi@mcgill.ca

Contract/grant sponsor: Canadian National Science and Engineering Research Council

data an artificial neural net, most frequently (but not always) of the back-propagation variety. This is often compared with a logistic regression model or, equivalently, discriminant analysis (Nielen *et al.* [2], Selker *et al.* [3], Marchevsky *et al.* [4], Tu *et al.* [5], Penny and Frost [6], Lapuerta *et al.* [7]). Occasionally, comparisons are made also with other statistical learning methods such as generalized additive models, classification trees, MARS etc. (for example, Ennis *et al.* [8], Carpenter and Markuzon [9], Curram and Mingers [10]). Results of the comparisons reported in the literature are variable, and authors' comments range from enthusiasm (for example, Jefferson *et al.* [11], Michael and Robert [12], Bostwick [13], Jain and Nag [14]) to scepticism (for example, Duh *et al.* [15], Manel *et al.* [16], Bertels *et al.* [17]) about the new tool.

Among the sceptics, Ennis *et al.* [8], perform a very thorough analysis of the 41 021 cardiac patients GUSTO-I database using a variety of non-linear statistical learning methods, among which is BP-ANN, only to conclude that none of these methods outperforms logistic regression. Commenting on this rather surprising result, they write: 'In human medical studies, the signal-to-noise ratio is often quite low (as its is here) and hence the modern methods may have less to offer... Thus our findings add evidence to support those who have suggested that adaptive non-linear algorithms might have limited applicability in clinical settings'.

Similarly, Duh *et al.* [15] analyse a database of 1674 patients with hepatic disorders and find that while BP-ANN clearly outperform logistic regression on the training set, they are statistically equivalent to logistic regression on the test set. This leads them to suggest that 'more generalizable modelling techniques for neural networks may be necessary before they are practical for medical research'. They cautiously conclude: 'Although this study shows that neural networks may not be as promising as they had seemed due to limited external validity, more comparative studies are needed to verify this finding.'

Thus there are (at least) two possible reasons to explain disappointing results with BP-ANN in clinical settings: (i) there is not much information in clinical data to warrant the effort of sophisticated BP-ANN modelling; (ii) a better effort is needed to improve generalizability of BP-ANN training – or, in a language more familiar to statisticians, to control the greater (with respect to logistic regression) overfitting bias inherent in highly adaptive non-linear modelling. The two explanations are, of course, *not* mutually exclusive. In fact more comparative studies are needed *and* more generalizable learning approaches are highly desirable.

The purpose of this work is to propose a systematic approach to training which aims to control overfitting bias. At the same time a *very limited* comparison with logistic regression is presented. It is important to clarify, however, that a *proper and exhaustive* comparison of BP-ANN with LR is a formidable undertaking and is beyond the scope of this work. Indeed, it can be argued that most of the papers that claim to compare logistic regression (LR) with BP-ANN define 'comparison', implicitly or explicitly, in a rather narrow sense. Likewise, we limit ourselves to a comparison of one-hidden layer BP-ANN with a *specific logistic regression model, which is used to construct it*. Thus, the goal of our comparison effort is very limited: we intend to show that using statistically sound training approaches and *given a specific regression model*, a BP-ANN can be constructed that improves *on it* as far as predictive value is concerned. In fact, rather than considering a BP-ANN as an alternative to LR, we attempt to show that there is advantage in using everything we can learn from LR analysis as a basis for the construction of a BP-ANN. Put it another way, the main message of this work is *not* that BP-ANN are better predictors than LR models; it is, instead, that an LR-based predictor may *generate* a demonstrably better BP-ANN.

As our context is clinical biostatistics, we work with ten data sets, all of clinical interest and varying in size. To facilitate further comparisons, we have chosen data sets that are publicly available through the web. Also, we have used S-plus BP-ANN software, developed by Venables and Ripley [18]. While this may not be as sophisticated and user-friendly as some commercial packages, it has the advantage of being highly statistically thoughtful, especially for what concerns the problem of generalizability/overfitting, and easily accessible to statisticians.

2. THE DATA

For each of the ten databases used in this study, we give here the web site on which they can be downloaded and a brief description. Further details are available in the original web sites. Six of the ten data sets come from the UCI repository of machine learning databases.

2.1. Pima Indian diabetes database –

<ftp://ftp.ics.uci.edu/pub/machine-learning-databases/>

The task is to decide whether a patient is diabetic or not, based on eight clinical variables, all continuous: age; diabetes pedigree function; body mass index; 2-hour serum insulin level; triceps skin fold thickness; diastolic blood pressure; plasma glucose concentration; and number of pregnancies. Of the 768 patients, 268 are diabetic.

Published analyses of these data include Smith *et al.* (1988) [19], Prechelt (1994) [20], Carpenter and Markuzon (1998) [9] and Ripley (1996) [21].

This data set is notoriously difficult to analyse by BP-ANN. Both Ripley [21] and Carpenter and Markuzon [9] found that BP-ANN does not outperform logistic regression. Ripley [21] omitted serum insulin and used only 532 records, randomly split into a training set of size 200 and a test set of size 332. Ripley's best methods, logistic regression on the six explanatory variables, made 66 misclassification errors on the test set, an error rate of 19.8 per cent.

2.2. Wisconsin breast cancer database –

<ftp://ftp.ics.uci.edu/pub/machine-learning-databases/>

The task is to predict malignancy from nine continuous clinical variables: clump thickness; uniformity of cell size; uniformity of cell shape; marginal adhesion; single epithelial cell size; bare nuclei; bland chromatin; normal nucleoli, and mitoses. The database consists of 699 patients, of which 16 were eliminated due to missing values. Of the remaining cases, 239 were classified as having a malignancy.

Wolberg and Mangasarian (1990) [22] applied a multi-surface method of pattern separation, training on 246 of the 369 inputs available at that time, and obtained 96 per cent test set predictive accuracy. Carpenter and Markuzon (1998) [9] got nearly the same performance levels by application of specialized BP-ANN algorithms (96–97 per cent correct classifications, depending on the choices in the learning algorithm); they also reported the C-index for logistic regression and for the best BP-ANN of 0.993 and 0.987, respectively.

2.3. *Haberman's survival data* –

<ftp://ftp.ics.uci.edu/pub/machine-learning-databases/>

The data come from a study on breast cancer that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital. Of the 306 patients who had undergone surgery for breast cancer, 81 had died within 5 years. The outcome to predict is 5-year survival and the predictors are age(X_1), year of operation(X_2), and number of positive axillary nodes(X_3).

These data were analysed by Haberman (1976) [23], who suggested a model including X_1, X_2, X_3, X_3^2 and X_1X_2 , achieving a mean deviance of 1.046 (on 300 d.f.). Landwehr *et al.* (1984) [24] proposed graphical methods for assessing logistic regression, omitted one outlier, refitted a logistic model with $(X_1 - 52)$, $(X_1 - 52)^2$, $(X_1 - 52)^3$, $(X_2 - 63)$, $(X_1 - 52) \times (X_2 - 63)$, $\log(X_3 + 1)$, which gives mean deviance 0.984 (on 298 d.f.). Both authors use the whole set for the analysis and do not report misclassification errors, or other measures of predictive performance.

2.4. *BUPA liver disorders* – <ftp://ftp.ics.uci.edu/pub/machine-learning-databases/>

The database contains records of 345 patients, 145 of which have liver disorders. The outcome to predict is liver disorder. There are six predictors, of which the first five are results of blood tests thought to be sensitive to liver disorders that might arise from excessive alcohol consumption: mean corpuscular volume; alkaline phosphatase; alamine aminotransferase; aspartate aminotransferase; gamma-glutamyl transpeptidase; number of half-pint equivalents of alcoholic beverages per day.

The web site contains no information on past use of these data, and we did not find any literature on their statistical analysis.

2.5. *Trauma data* – <http://www.math.unm.edu/~fletcher>

The trauma database is described in *Log-Linear Models and Logistic Regression*, second edition; by Ronald Christensen (1997) [25]. It consists of a randomly selected subset of 300 patients admitted to the University of New Mexico Trauma Center between the years 1991 and 1994. Of these, 22 died.

The task is to predict survival from four attributes: ISS; TI; RTS, and AGE. ISS is an overall index of injury based on the (approximately) 1300 injuries catalogue of the Abbreviated Injury Scale: it can take on values from 0 for a patient with no injuries to 75 for a patient having injuries in three or more body areas. The TI is a binary variable denoting the type of injury: TI=0 for bland injury, for example, the result of a car crash, and TI=1 for penetrating injuries, for example, gunshot wounds. The RTS is an index of physiologic injury and is constructed as a weighted average of an incoming patient's systolic blood pressure, respiratory rate and Glasgow Coma Scale. The RTS takes on values from 0, for a patient with no vital signs, to 7.84, for a patient with normal vital signs.

The analysis proposed by Christensen (1997) [25] is based on Bayesian binomial regression and the results are not directly comparable with ours.

2.6. *Heart disease* – <ftp://ftp.ics.uci.edu/pub/machine-learning-databases/>

The outcome is diagnosis of heart disease. We worked with the 13 predictors suggested in the web site: age; sex; chest pain type; blood pressure; cholesterol level; fasting blood

sugar; resting electrocardiograph results; maximum heart rate; angina; ST depression (induced by exercise relative to rest); slope of peak exercise ST segment; number of major vessels coloured by fluoroscopy, and thalassaemia.

There are 920 observations in the whole database, including many observations with missing values. After deleting the records containing missing attributes, we used 299 observations in this study. Of the 299 patients, 139 (46.49 per cent) were diagnosed as having heart disease.

Several authors studied this data, such as Gennari *et al.* (1989) [26] and Aha *et al.* (1991) [27]. Results of early analyses are summarized in Carpenter and Markuzon [9] who used 303 observations and obtained 80 per cent correct prediction rate using the ARTMAP-IC neural network.

2.7. *Contraceptive method choice* – <http://www.stat.wisc.edu/~limt/compare.ps> or [compare.pdf](#)

This database is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey.

Cases are married women who were either not pregnant or did not know if they were pregnant at the time of interview. The problem is to predict use of a contraceptive method. As done by other authors, we only considered as outcome ‘use of contraception’ versus ‘no use’. The following nine predictors are available in the database and were used in our work: age; education; husband’s education; number of children ever born; religion; working status; husband’s occupation; standard-of-living index, and media exposure. The number of cases is 1473, and of these, 844 used long-term or short-term contraception.

Lim *et al.* (1999) [28], who analysed these data using two neural networks, reports test set error rates of 0.491 and 0.458, respectively. When using linear discriminant analysis and logistic discriminant analysis, they also found test set error rates of 0.492 and 0.489, respectively.

2.8. *Thyroid disease* – <ftp://ftp.ics.uci.edu/pub/machine-learning-databases/>

This is a database on thyroid disease containing 3772 records. The task is to predict thyroid disease from the 20 variables that remain after removing from the original set the variables with identical or almost identical values. Of these, 15 are binary: age; sex; use of thyroxine; use of antithyroid medication; presence of other illness; pregnancy; thyroid surgery; I131 treatment; hypothyroidism; hyperthyroidism; thium; goitre; tumour; psych. The remaining five are continuous: TSH; T3; TT4; T4U; FTI. One six-level categorical variable contained in the original database, referral source, was not used in this analysis. Other variables in the database were not used for having either an excess of missing values, or an (almost) null variance.

After deleting all records with missing attributes, we were left with 2643 cases, of which 212 (8.02 per cent) had thyroid disease.

Published analyses seem to be using different versions of the database also available in the web site. For example, Prechelt (1994) [20] used 7200 observations, 21 inputs and 3 outputs, partitioned the data to training, validation, test subset, and permuted it in three ways; the classification errors are 6.56, 6.56 and 7.23 per cent, respectively. Lim (1999) [28] trained 3772 records, estimated the error rate on the other 3428 records; classification error rates for two neural networks are 7.12 per cent and 3.18 per cent, respectively, and for linear discriminant analysis (LDA) and logistic discriminant analysis (LOG) are 0.0619, and 0.0405, respectively. We have preferred to use the present version since it was the best documented one at the time we initiated our work.

2.9. *Cardiac arrhythmia* – <ftp://ftp.ics.uci.edu/pub/machine-learning-databases/>

The outcome to predict is presence of arrhythmia. The number of cases is 452 and the total number of predictors is 279. Of these, we only considered 18, eliminating those that were not clearly defined, those for which class distribution was not reported, and those that had almost constant value over the sample. After deleting records with missing values, 420 cases remained in our study. The 18 predictors are: age; sex; height; weight; average of QRS duration; average duration between onset of P and Q waves; average duration between onset of Q and offset of T waves; average duration of T wave; average duration of P wave; QRS; T; P; QRST; heart rate; Q wave; R wave; S wave; number of intrinsic deflections.

2.10. *Prostate cancer* – <http://hesweb1.med.virginia.edu/biostat/s/data/index.html>

This database consists of data from a randomized clinical trial on 506 patients with stage 3 and 4 prostate cancer, of which 502 were followed for at least 2 years. The outcome is survival status at 2 years. There are 14 predictors of which six are discrete (binary or ordinal) and eight continuous: stage; treatment; age; weight; performance rating; history of cardiovascular disease; systolic blood pressure; diastolic blood pressure; electrocardiogram code; serum haemoglobin; size of tumour; combined index of tumour stage and histologic grade; serum prostatic acid phosphatase; bone metastasis. After dropping observations with at least one missing value, we are left with a sample of 475 patients.

3. METHODS

The same systematic approach was used for the analysis of each data set.

First, each data set was split into a learning and a test set by randomly selecting the elements of the learning set with probability $2/3$. On the training set, two logistic models were built and a BP-ANN was trained until stable connection weights were obtained. With regression coefficients and connection weights fixed to the values determined on the training set, the resulting three predictors were evaluated on the test set. Evaluation was performed by computing for each predictor the following quantities: (i) deviance (twice the negative log-likelihood), (ii) C-index (area under the ROC curve, in the case of binary response); (iii) error rate (ER), based on a 50 per cent cut-off; (iv) mean squared error (MSE). These are widely used measures of performance, see Harrell (1998) [29], more specifically of *predictive accuracy*. As also remarked in reference [29], another important aspect of model evaluation is *calibration*, which can be studied by graphical approaches. However, since our main focus is on prediction, we did not systematically investigate the *calibration* of our models.

3.1. *Logistic regression models*

On the training set, two logistic regressions were performed using the `glm` function of S-plus: one, henceforth referred to as *full model*, included all variables, the other, referred to as *stepwise*, performed an automatic variable selection using the default option of the generic function `step`, see Venables and Ripley (reference [18], pp. 218–222).

No attempt was made to model interactions, which is of course a very arguable choice. Indeed, as one referee suggested, one could deploy a general strategy for detecting pairwise

interactions and, in addition, fit *additive* models to detect non-linear effects of individual variables, or, as another referee suggested, one could use partial least squares, principal component and all-subset selection LR to find better models. Conceivably one could find an extremely sophisticated model building strategy to reflect the practice of conscientious, skillful analysts – but there is no end to possible suggestions on how to build a ‘good’ LR model.

We acknowledge that had we used a more sophisticated LR modelling strategy, we would have come closer to a convincing *general* comparison of LR and BP-ANN. These considerations notwithstanding, we chose the simplest approach, in part for pragmatic considerations – lack of time and resources. To support our ‘minimalist’ choice, however, we wish to point out that the ten data sets used here have been previously used in comparative studies, but usually without recourse to sophisticated strategies; in particular we are not aware of published studies in which interaction detection strategies or generalized additive models were applied to these data sets. In particular, the Pima Indian data set, one of the two we studied most extensively, has already been used for comparison with BP-ANN by Ripley [21], he did *not* report interaction tests, and we have found ourselves that neither interactions nor the addition of non-linear terms in individual variables seem to improve the fit of the model.

Another argument in favour of our choice is that a more sophisticated effort would obscure rather than clarify the issue. In fact the main goal of this work is to show that given a *specific* LR model, one can do better by constructing a BP-ANN which uses the given LR model as a starting point. Clearly the closest the LR model is to the ‘truth’, the more modest one would expect the improvement to be; demonstration of this intuitive reasoning, however, would require in itself a major careful study. In conclusion, we chose the ‘quick and dirty’ approach.

3.2. BP-ANN architecture and training

We have used throughout one hidden layer BP-ANN. While other authors have used two hidden layers and other more complex architectures, it is not clear that for a comparative study such as ours, the advantages would be substantial; on the other hand, considering such architectures would require a much greater computational effort.

To train all BP-ANN, we have used a standardized approach, described below, based on *regularization*, as in the work of Ennis *et al.* [8]. Regularization has been preferred to *early stopping*, another popular approach in recent work (for example, Duh *et al.* [15]). The choice was based on the theoretical soundness of the method (Ripley) as well as on some empirical evidence (see Results section). We have performed regularization training by means of the S-plus function `nnet` contributed by Venables and Ripley [18]. Regularization consists in learning optimal weights by minimizing a penalized objective function, with a penalty proportional to the sum of the squares of the weights. The proportionality factor, λ , referred to as the *regularization parameter*, was constrained to be the same for all weights, a choice implicit in `nnet`. Therefore, once the objective function is selected, the regularization parameter λ and the number of hidden units n_h are the only additional choices to make in the learning algorithm; the weights are then determined from the minimization of the penalized objective function. As (non-penalized) objective function we have chosen throughout the deviance, since this is also the criterion used in logistic regression. A clear description of the function `nnet` can be found in reference [18] (pp. 337–341).

It should be remarked that many BP-ANN applications use least squares as the objective function. In contrast, we decided to use the deviance as the objective function as well as our primary evaluation criterion, since minimizing the deviance is equivalent to maximizing the likelihood. This choice is very compelling from a statistical point of view as the likelihood is the base of statistical inference. In particular, the comparison of our BP-ANN with a particular logistic regression model becomes very direct; indeed, logistic regression models are fitted by minimizing the deviance (maximizing the likelihood). Our deviance-based approach is most easily implemented in the S-plus function `nnet`. This is easily achieved with `nnet` ('entropy = T') and has the great advantage of associating a likelihood to a BP-ANN.

The first distinctive feature of our approach consists in choosing λ and n_h by 10-fold cross-validation (for example, Prechelt [20] and Tourassi and Floyd [30]), which has the desirable property of using the entire training sample. More specifically, λ and n_h were chosen so as to minimize the cross-validated deviance (average over the ten validation sets of the deviance of the model with parameters estimated on the corresponding training set). This choice is slightly different from that of other authors (Prechelt [20], Duh *et al.* [15], Jain and Nag [14]) who maximize the cross-validated C, or ER, or MSE; it was motivated by some empirical evidence (see Results), and also by logical consistency, since what we minimize on the training set is the penalized deviance. We systematically calculated the cross-validated deviance for values of n_h ranging from 1 to 13 and for $\lambda = 0, 0.0001, 0.001, 0.01, 0.1, 0.2, 0.3, 0.4$. When calculations suggested that the minimum of the deviance with respect to λ was located between two of the above values, we continued the exploration at a finer level, for example if the minimum seemed to be between 0.01 and 0.1, we also looked at the values 0.03, 0.05, 0.08 etc. It should be noticed that we did not use a Bayesian argument in determining the regularization parameter, but proceeded in a purely empirical fashion. Similarly, we chose the number of hidden units automatically, proceeding from one hidden unit upward (see Results section). It should also be noticed that full model logistic regression can be performed by minimizing a regularized deviance (that is, a deviance penalized by a term proportional to the sum of the squares of the regression coefficients); however, our calculations showed a general tendency to obtain regularization parameters tending towards zero, thus for simplicity we used the standard (non-regularized) logistic regression.

We preferred 10-fold cross-validation since work in the context of tree growing suggests that it may be more effective than other forms of cross-validation and the bootstrap in controlling overfitting. Also, results reported by Tourassi and Floyd [30], who compare split-sample cross-validation, leave-one-out cross-validation (also known as round robin) and bootstrap for BP-ANN training, show that each of these methods has its drawbacks; at the same time, their results indirectly suggest that 10-fold cross-validation might represent a reasonable compromise between perturbing the sample too much (split-sample) and too little (leave-one-out and bootstrap).

The other distinctive feature of our approach to training is the initialization procedure, which is based on the observation that a BP-ANN always *contains* a logistic regression model as a particular case.* As a consequence, the resulting predictor should not be worse

*An LR model may be defined from a BP-ANN as follows: all units are taken to be logistic; the weights of the connections between the outer and the hidden layers are selected so that the hidden units operate as though they were linear (identity); finally the weights of the connections between hidden and output units are taken to be identical to the coefficients of the LR model. Other approaches are discussed below.

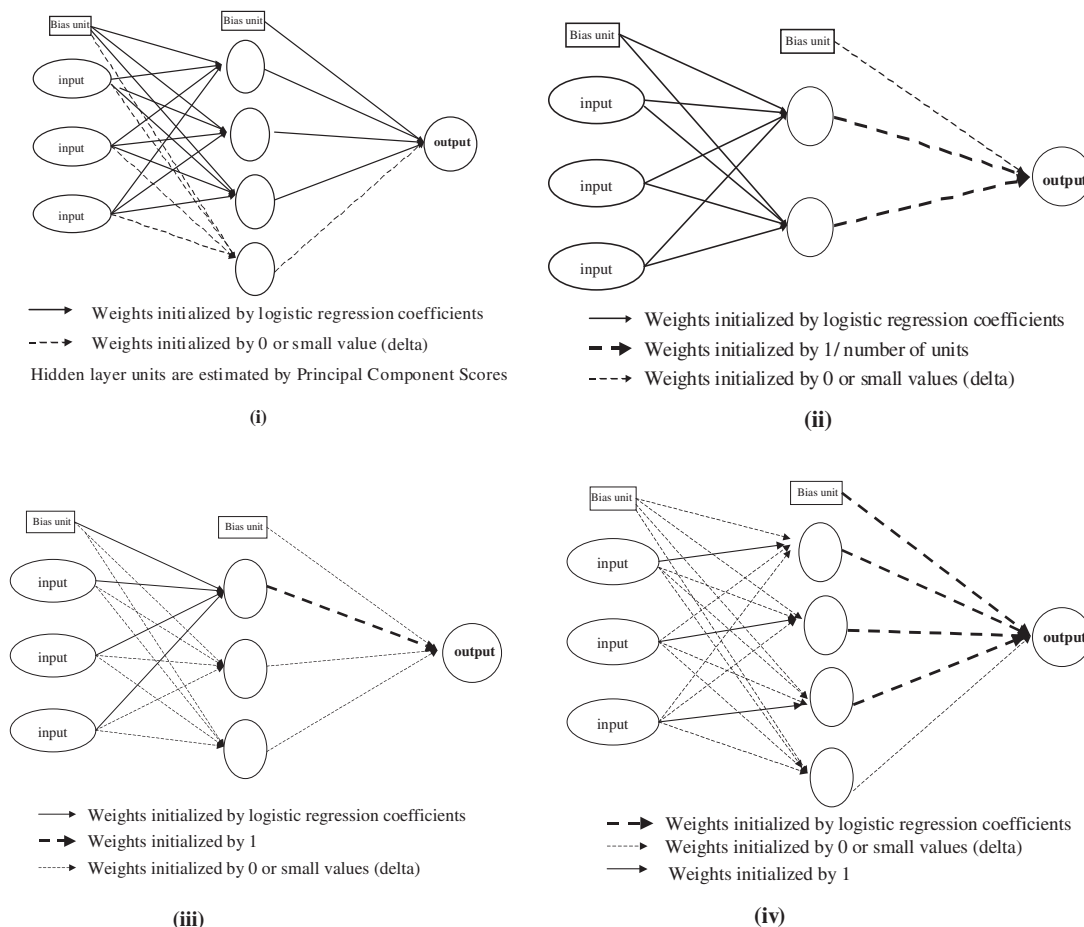


Figure 1. Initialization procedures for the connection weights of one hidden layer artificial neural network (ANN).

than a logistic regression model. Since the minimization procedure for BP-ANN generally leads to local minima, even with the benefit of regularization, we should aim at reaching a local minimum, which is an improvement on logistic regression at least on the training set. Although there is no guarantee that this improvement is generalizable, we can hope, owing to the regularization device, that the advantage is maintained ‘on the average’ also on future data. More specifically, we considered initializations corresponding to four different realizations of logistic regression in a one hidden layer BP-ANN, and chose the one with minimum deviance. These are summarized in Figure 1 and are as follows:

- (i) For the initial weights of the connections between the input and the hidden layer, we took the standardized scores of the first n_h principal components (PC) extracted from the predictors, except for the bias unit connection, which was initialized by zero or values close to zero (by this we mean, here and in what follows, 0, 0.0001 and 0.1);

when n_h exceeded the number of PC, the connections between inputs and the extra hidden units were initialized as the bias unit. For the initial weights of the connections between hidden and outer layers, we took the coefficients of the logistic regression to predict the outcome from the PC.

- (ii) We initialized the weights of the connections between the inputs and each hidden unit by the coefficients of the logistic regression to predict the outcome from the original predictors. We initialized the weights of the connections between hidden units and output by $1/n_h$, and between bias unit and output by 0 or very small values.
- (iii) The weights of the connections between inputs and first hidden units were initialized by the coefficients of the logistic regression, those of the connections between the first hidden units and the output were initialized to 1, and all other weights to 0 or values close to zero.
- (iv) This initialization was used only when the number of hidden units is equal to or larger than the number of inputs. The hidden units were copied from the inputs, that is, each input is connected with initial weight 1 to only one hidden unit, with the remaining connection weights being initialized to zero or values close to zero. The connection weights between the hidden layer and the output are initialized to the coefficients of the logistic regression; when n_h exceeded the number of inputs, again the initial values of the corresponding weights was initialized to zero or values close to zero.

Moreover, for each BP-ANN, we also repeated ten times the 'standard' initialization corresponding to randomly selected connection weights (each time with a different seed).

The motivation for this multiple initialization is, of course, the existence of local minima which implies that even a very little change in the initialization may change the results of the back-propagation. Since 'zero or values close to zero' corresponds to three different values of initial weights, the above mentioned four initialization procedures give, in fact, 12 or 9 different starting points for the training algorithms, according to whether n_h is equal to or larger than the number of inputs, or whether n_h is smaller than such a number. Thus, for each value of n_h and λ , the optimal BP-ANN, the one with smallest deviance, was chosen among 22 or 19 candidates. It should be emphasized that 'optimal' means 'with the smallest *cross-validated* deviance'; this is a crucial point, as also previously stated. Failing to cross-validate the selection criterion is, unfortunately, a frequently encountered oversight in the neural net clinical application literature and it may explain why naively trained nets perform so poorly when properly compared with other approaches on real data. This point is most clearly put forward in a recent paper by Schwarzer *et al.* [31].

3.3. Evaluating the average performance

Choosing one test set only for evaluation purposes may be criticized on the grounds that a result, whether favourable or unfavourable, may be due to chance; this is especially true with relatively small data sets. It is preferable to make evaluation on an 'average' test set. Therefore, for two of the databases (Pima Indian diabetes and BUPA liver disorder) we have carried out the process further. In order to assess the 'average' performance of the proposed approach for a given prediction problem, the random splitting was performed ten times. Evaluations of the two logistic regression models and the BP-ANN were made by calculating the means and standard errors of DEV, C, ER and MSE over the ten test sets. Similarly, each logistic regression model was compared with the BP-ANN by calculating the mean and the standard

error over the ten test sets of the differences of the corresponding measures of performance. Limited resources did not allow us to complete the process for the ten data sets, but we chose the two that appeared more challenging.

4. RESULTS

In Table I we report summary characteristics of the ten databases and, for each of them, of the three models considered in this work: full model logistic regression; stepwise logistic regression, and BP-ANN (trained by the approach described in Section 3.2). We also give the size of training and test sets used in the single test set performance comparison described below. The number of hidden units range from 1 to 13 and it does not appear to be increasing with the sample size. It should also be noted that eight of the ten empirical choices of the regularization parameter fall within the range suggested by Ripley according to a Bayesian argument (0.0001–0.1), although in two cases we obtain $\lambda=0.1$. The heart disease and the prostate cancer data are two notable exceptions with $\lambda=0.23$ and 0.2, respectively.

4.1. Single test set performance

First, we report the comparison of the three models based on a single test set for each database. The results are given in Table II.

If we consider the deviance as comparison criterion, we can see that for all ten data sets the BP-ANN constitutes an improvement as compared to both logistic regressions (LRs). The

Table I. A summary description of the data sets and their trained logistic regression (LR) and back-propagation artificial neural networks (BP-ANN) models.*

Data sets	Sample size			Events	Number of continuous (categorical) predictors	Stepwise LR	BP-ANN	
	Total	Train	Test				Units	λ
Pima Indian diabetes	768	523	245	268	8 (0)	1, 2, 6, 7	3	0.08
Wisconsin breast cancer	683	457	226	239	9 (0)	2, 4, 7–10	3	0.08
Haberman's survival data	306	194	112	81	3 (0)	1, 3	3	0.01
BUPA liver disorder	345	221	124	145	6 (0)	1–5	3	0.03
Trauma data	300	192	108	22	3 (1)	1, 3, 4	9	0.05
Heart disease	299	197	102	139	5 (8)	2, 3, 5, 8, 10, 12, 13	13	0.23
Contraceptive method choice	1473	936	537	844	2 (7)	1, 2, 4–6, 8	3	0.1
Thyroid disease	2643	1739	904	212	6 (14)	6, 10, 11, 15–18, 20	5	0.05
Cardiac arrhythmia	420	286	134	183	17 (1)	2, 3, 5, 6, 8, 9, 12, 15–17	7	0.1
Prostate cancer	475	316	159	180	8 (6)	3–6, 10–14	1	0.2

* Full logistic regression takes all predictors into the equation.

Table II. Performance comparison between logistic regression (LR) and back-propagation artificial neural networks (BP-ANN) on test data.*

Data sets and criterion	Estimated values			Standard deviation			Full LR and BP-ANN			Stepwise LR and BP-ANN		
	Full LR (1)	Stepwise LR (2)	BP-ANN (3)	Full LR (4)	Stepwise LR (5)	BP-ANN (6)	Difference (7)	P-value (8)	Difference (9)	P-value (10)		
<i>Pima Indian diabetes</i>												
DEV	259.075	258.974	246.888	1.3340	1.3165	1.2671	12.1870	0.0224	12.0860	0.0743		
MSE	0.1758	0.1756	0.1676	0.2465	0.2428	0.2377	0.0082	0.0342	0.0080	0.1063		
C	0.8070	0.8076	0.8203	0.0197	0.0191	0.0189	-0.0133	0.0226	-0.0127	0.0844		
ER	27.35	26.94	24.90	0.4457	0.4436	0.4324	2.4500	0.6071	2.0400	0.6801		
<i>Wisconsin breast cancer</i>												
DEV	74.466	78.859	63.902	1.3433	1.4433	1.0886	10.5640	0.0169	14.9570	0.0410		
MSE	0.0485	0.0487	0.0448	0.1916	0.1920	0.1765	0.0037	0.0174	0.0039	0.0581		
C	0.9951	0.9942	0.9955	0.0011	0.0017	0.0010	-0.0004	0.0219	-0.0013	0.2023		
ER	6.19	6.19	6.19	0.2411	0.2411	0.2411	0.0000	0.8453	0.0000	0.8453		
<i>Haberman's survival data</i>												
DEV	124.762	122.483	119.302	1.2593	1.2160	1.0614	5.4600	0.4546	3.1810	0.6548		
MSE	0.1761	0.1724	0.1742	0.2274	0.2219	0.2155	0.0019	0.8131	-0.0018	0.8141		
C	0.6198	0.6416	0.6302	0.0683	0.0682	0.0692	-0.0104	0.3038	0.0114	0.2141		
ER	21.43	22.32	26.79	0.4103	0.4164	0.4428	-5.3600	0.4348	-4.4700	0.5346		
<i>BUPA liver disorder</i>												
DEV	159.665	160.394	156.787	0.8376	0.8338	0.9842	2.8780	0.6655	3.6070	0.5882		
MSE	0.2276	0.2286	0.2239	0.1786	0.1783	0.2028	0.0037	0.7507	0.0047	0.6816		
C	0.6532	0.6478	0.6705	0.0303	0.0308	0.0344	-0.0173	0.4171	-0.0227	0.3124		
ER	36.29	37.10	38.71	0.4808	0.4831	0.4871	-2.4200	0.7931	-1.6100	0.8959		
<i>Trauma data</i>												
DEV	49.680	49.310	46.044	1.4626	1.4503	1.2805	3.6360	0.2528	3.2660	0.3048		
MSE	0.0650	0.0647	0.0612	0.1986	0.1979	0.1926	0.0038	0.1832	0.0035	0.2128		
C	0.8889	0.8900	0.8844	0.0298	0.0297	0.0293	0.0045	0.3592	0.0056	0.2776		
ER	9.26	9.26	9.26	0.2899	0.2899	0.2899	0.0000	0.8144	0.0000	0.8144		

<i>Heart disease</i>										
DEV	73.195	82.746	71.586	1.1141	1.3023	0.8629	1.6090	0.6664	11.1600	0.0674
MSE	0.1091	0.1234	0.1045	0.1935	0.2136	0.1667	0.0046	0.4633	0.0189	0.0388
C	0.9250	0.9083	0.9389	0.0169	0.0197	0.0153	-0.0139	0.0017	-0.0306	0.0038
ER	11.76	15.69	13.73	0.3222	0.3637	0.3441	-1.9700	0.8337	1.9600	0.8433
<i>Contraceptive method choice</i>										
DEV	665.700	665.983	610.223	0.7985	0.7999	0.8888	55.4770	0.0001	55.7600	0.0001
MSE	0.2159	0.2159	0.1927	0.1769	0.1766	0.1870	0.0232	0.0001	0.0232	0.0001
C	0.6911	0.6922	0.7589	0.0138	0.0138	0.0118	-0.0678	0.0000	-0.0667	0.0000
ER	34.26	34.26	28.68	0.4746	0.4746	0.4523	5.5800	0.0567	5.5800	0.0567
<i>Thyroid disease</i>										
DEV	230.939	224.558	169.566	1.0973	1.0488	0.7821	61.373	0.0000	54.9920	0.0000
MSE	0.0335	0.0326	0.0257	0.1348	0.1310	0.1131	0.0078	0.0001	0.0069	0.0001
C	0.9474	0.9506	0.9761	0.0161	0.0156	0.0087	-0.0287	0.0013	-0.0255	0.0023
ER	4.20	4.20	3.54	0.2007	0.2007	0.1848	0.6637	0.5422	0.6637	0.5422
<i>Cardiac arrhythmia</i>										
DEV	222.731	225.661	154.423	2.9576	2.9781	0.9546	68.3080	0.0319	71.2380	0.0271
MSE	0.2368	0.2383	0.1952	0.2387	0.2422	0.1990	0.0416	0.0039	0.0431	0.0047
C	0.6568	0.6564	0.7337	0.0460	0.0471	0.0433	-0.0769	0.0066	-0.0773	0.0149
ER	32.84	33.58	25.37	0.4696	0.4723	0.4351	7.4669	0.2262	8.2089	0.1803
<i>Prostate cancer</i>										
DEV	229.625	226.961	218.017	1.2201	1.1630	0.9365	11.608	0.0166	8.944	0.0536
MSE	0.2530	0.2510	0.2440	0.2433	0.2340	0.2062	0.009	0.0456	0.007	0.1058
C	0.5760	0.5760	0.5600	0.0393	0.0380	0.0393	0.016	0.0313	0.016	0.0954
ER	39.62	38.37	40.25	0.4891	0.4860	0.4904	-0.63	0.9999	-1.88	0.8184

* *P*-values for differences are based on the following: paired *t*-test for DEViance and MSE; a non-parametric test described in DeLong *et al.* (1988) [32] for C-index; chi-square with 1 d.f. for ER.

improvement ranges from 2 per cent (BUPA liver disorder and heart disease databases) to 31 per cent (cardiac arrhythmia database). For six of the ten data sets, namely Pima Indian diabetes, Wisconsin breast cancer, contraceptive methods choice, thyroid disease, cardiac arrhythmia, and prostate Cancer, the improvement of the BP-ANN with respect to full model LR appears to be also 'statistically significant', as shown in columns 8 and 10. For five of these data sets, all but the Pima Indian diabetes, the comparison with stepwise LR is also favourable to the BP-ANN, see column 10. Interestingly, the results of the comparisons on the other criteria are different. While the MSE and the C-index follow closely the pattern of the deviance, one notable exception occurs for the trauma data, where the BP-ANN-based predictor performs significantly worse than both LRs according to the C-index, and significantly better than stepwise LR according to the MSE. On the other hand, the Error Rate of the BP-ANN-based predictor improves on both LRs for only four data sets, namely the Pima Indian diabetes, the contraceptive method choice, thyroid disease and the cardiac arrhythmia databases; of these improvements, which are, respectively, of 7.4, 16.5, 15.7 and 22.4 per cent, only the second is (marginally) significant.

From the above, it appears that the results of the comparison depend, to a certain extent, on the evaluation criterion chosen. Therefore we have considered slight variations in our training procedure, in the hope of improving the performance of the BP-ANN on a pre-selected criterion. For example, if we want to have an optimal ER and care less about the deviance, it seems reasonable to select the regularization parameter and the number of hidden units so as to minimize the cross-validated ER instead of the cross-validated deviance. With this in mind, we have repeated the training varying the criterion for choosing λ and n_h . In general this variation did not improve the selected criterion substantially and, indeed, introduced greater inconsistency in the results (see Table III for partial results).

4.2. Some details for the Pima Indian diabetes database

In Figure 2 we show the logistic regression initialization of the BP-ANN for the PIMA Indian diabetes database, and the same network after training. It can be seen that the trained BP-ANN is quite different from the initial one and that no simple patterns emerge from an inspection of the connection weights.

Figure 3 shows the ROC curves for both the BP-ANN and the LR predictor, from which can be seen the improvement of the former over the latter.

In order to gain a more detailed understanding of the way the two predictors operate, we examined the details of the predictions given by both BP-ANN and LR, and compared them with the actual outcome for the test set. There are 23 diabetes negative patients who are classed positives by both predictors; on the other hand, the BP-ANN properly recognizes 3 negative patients that LR wrongly classifies as positives. Similarly, there are 40 diabetes positive patients that are missed by both predictors; 8 positives that are correctly classified by the BP-ANN and missed by the LR; and, finally, 5 positives that are correctly classified by LR and missed by the BP-ANN. An inspection of the covariates of the 17 patients for which the two predictors disagree indicates that the BP-ANN predictor performs better than the LR predictor for patients with values of plasma glucose, serum insulin and diastolic blood pressure higher than the sample average, while the LR is better than BP-ANN for values of the same variables lower than the sample average.

Table III. Average performance of logistic regression (LR) versus back-propagation artificial neural networks (BP-ANN).

Seed	Number of cases	Deviance		MSE		C-index		Error rate (ER)	
		Full LR	BP-ANN	Full LR	BP-ANN	Full LR	BP-ANN	Full LR	BP-ANN
<i>(a) BUPA liver disorder</i>									
1	123	144.591	137.593	0.2008	0.1879	0.7467	0.7769	33.33	25.20
2	109	164.294	149.095	0.2540	0.2329	0.6059	0.6587	40.37	31.19
3	100	127.477	120.061	0.2218	0.2062	0.6993	0.7399	31.00	29.00
4	106	138.141	140.094	0.2227	0.2183	0.6768	0.6982	29.25	33.02
5	109	135.051	128.197	0.2152	0.1969	0.7163	0.7759	35.78	27.52
6	119	134.797	133.105	0.1917	0.1868	0.7814	0.7877	25.21	26.05
7	114	162.033	150.746	0.2506	0.2272	0.6177	0.6750	41.23	31.58
8	121	145.385	137.942	0.2081	0.1922	0.7213	0.7608	29.75	28.10
9	126	147.981	139.633	0.2001	0.1841	0.7787	0.8243	24.60	23.02
10	112	138.400	137.503	0.2077	0.2108	0.7395	0.6773	27.68	31.25
Mean		143.815	137.397	0.2172	0.2043	0.7083	0.7375	31.82	28.59
SD		11.824	9.014	0.0208	0.0175	0.0603	0.0567	5.81	3.21
<i>T</i> -test*		4.01		4.89		2.55		1.96	
<i>P</i> -value		0.0030		0.0009		0.0311		0.0816	
<i>(b) Pima Indian diabetes</i>									
1	264	265.858	256.026	0.1627	0.1590	0.8228	0.8306	21.59	21.97
2	252	243.409	228.413	0.1547	0.1473	0.8364	0.8555	21.83	21.43
3	272	283.142	267.953	0.1713	0.1625	0.8198	0.8403	25.00	23.90
4	258	232.841	225.531	0.1444	0.1410	0.8540	0.8590	19.77	21.32
5	236	237.068	230.944	0.1613	0.1602	0.8263	0.8263	23.73	24.58
6	271	275.790	273.192	0.1657	0.1664	0.8063	0.8129	24.35	22.88
7	269	269.045	263.347	0.1590	0.1567	0.8193	0.8252	24.16	21.93
8	263	289.298	280.511	0.1821	0.1751	0.8070	0.8141	26.62	26.24
9	262	267.329	257.093	0.1683	0.1622	0.8166	0.8249	25.19	23.66
10	239	243.589	241.074	0.1681	0.1676	0.8124	0.8153	25.10	23.85
Mean		260.737	252.408	0.1638	0.1598	0.8221	0.8304	23.73	23.18
SD		20.044	19.739	0.0101	0.0098	0.0144	0.0164	2.07	1.57
<i>T</i> -test*		5.94		3.87		4.02		1.49	
<i>P</i> -value		0.0002		0.0038		0.0030		0.1717	

*Paired *t*-statistics for comparing LR with BP-ANN over ten random partitions into training and test set.

4.3. Average performance on ten test sets for the PIMA Indian diabetes and the BUPA liver disorder databases

Table III (a) and (b) shows the results of the evaluation of the average performance of BP-ANN, full model and stepwise LR for the two databases Pima Indian diabetes and BUPA liver disorders. The former was chosen because previous work by Ripley suggests that it is difficult to improve on LR for these data, the latter because it is a case where small sample size may be at the root of the disappointing performance of the BP-ANN-based predictor.

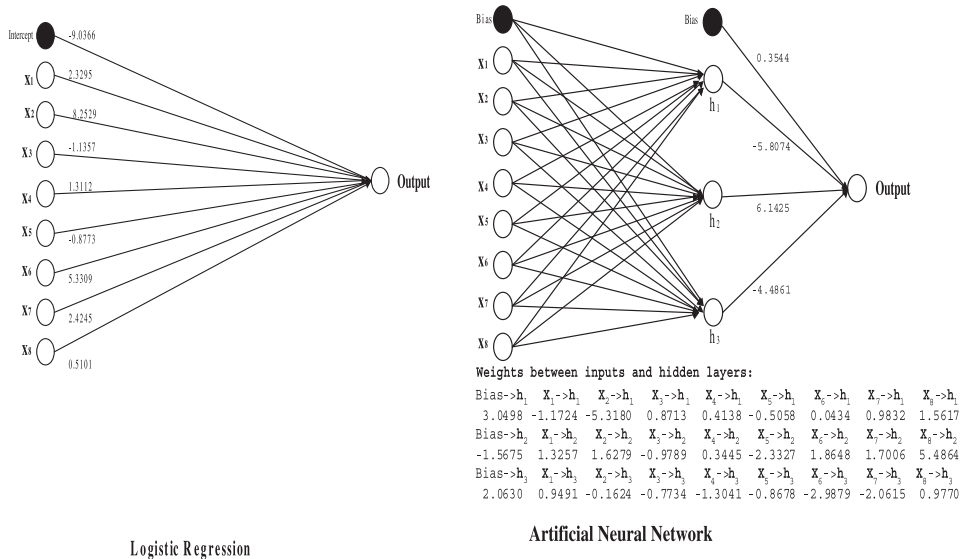


Figure 2. Comparison of trained artificial neural networks (ANN) and estimated logistic regression (LR), Pima Indians diabetes

For the Pima Indian diabetes database, the average performance reproduces reasonably well the results of the single test set performance evaluation. Using the deviance, the C-index and the MSE as evaluator of performance, the BP-ANN-based predictor is significantly better, on the average, than both the LR-based predictors. Furthermore, the BP-ANN is consistently better than (or as good as) the two LR on nine of the ten replications; the only exception appears in replication 6, where the MSE of the BP-ANN is slightly higher than that of the two other predictors. Similarly, using ER as evaluator, BP-ANN also outperforms both full model and stepwise LR, on the average, but the difference (the BP-ANN yields, on the average, an improvement in the ER of 2.3 per cent with respect to the best LR) is not significant; however, the significance depends here on the number of replications, and the results suggest that the *P*-value may decrease below 0.05 by increasing the latter.

For the BUPA liver disorder database, we find that on the average, even with as few as ten replications, BP-ANN performs better than both LR when the evaluation criteria are deviance, MSE and C-index, though with lesser consistency than for the former database. The variation is even more evident when using ER as evaluation criterion: in this case the BP-ANN still outperforms LR on the average, but we do not reach statistical significance (however, the best *P*-value is 0.08, not far from marginal significance). This is in contrast with the single test performance evaluation, where BP-ANN did not represent a significant improvement, no matter the criterion of evaluation used.

4.4. Advantage of regularization versus early stopping

In order to perform a limited comparison of early stopping with our regularization based approach, we trained a BP-ANN on each of the ten databases by early stopping (ES). There are

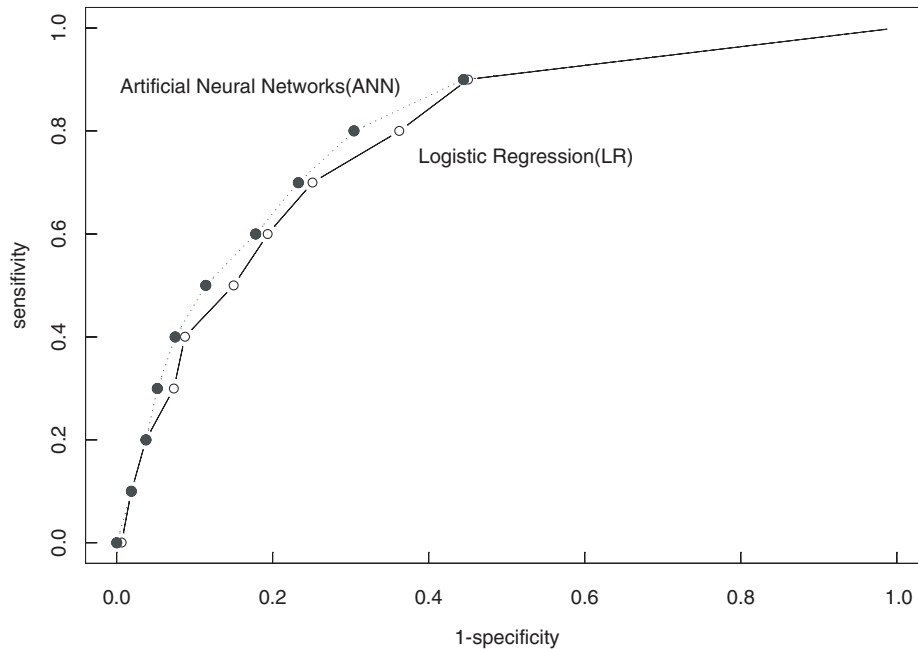


Figure 3. Receiver operating characteristic curves of LR and ANN, Pima Indians Diabetes.

many possible variants of ES; here we used the variant described in Duh *et al.* [15] (training is done at 25, 150, 250 and 500 iterations, and the result retained is the one corresponding to the minimum of the criterion *on the training set*). However, we introduced a simple modification to make our approach and ES comparable. On the one hand, we could have used the training sets to determine where to stop training; this, however, would have given biased results in favour of ES, since the test sets would be used both for ES and final evaluation. On the other hand, we could have divided the learning set into two, one for training and the other to choose where to stop; this, however, would have resulted in some bias against ES, since our approach uses the learning sets more efficiently, owing to 10-fold CV. Therefore we have determined where to stop by 10-fold cross-validation. We also introduced another obvious modification with the aim to make the approaches comparable; we have used the same initialization for both ES and our variant of regularization – Duh *et al.* [15] used random initialization. The results are presented in Table IV. The table clearly shows that our approach has a considerable empirical advantage over ES. Indeed, in only one case (BUPA liver disorder) the deviance of the final model obtained by ES is lower than the result obtained by our approach, and then only by a relative small amount (4.2 per cent). Also, ES tends to give poorer results than logistic regression.

We should stress that we do *not* recommend the ES procedure used here; we simply wanted to show that our procedure improves on the specific ES procedure of Duh *et al.* [15], which, with all its limitations, has the advantage of having being published on a biomedical journal and therefore can be considered a bench mark in our context. An extensive comparison of ES with regularization would indeed be useful but is beyond the scope of this paper.

Table IV. Comparison of early stopping (ES) and regularization (Reg.) for training BP-ANN.

Data sets	Deviance on the test set	
	BP-ANN (ES)	BP-ANN (Reg.)
Pima Indians diabetes	265.59	246.89
Wisconsin breast cancer	136.17	63.90
Haberman's survival	134.47	119.30
BUPA liver disorder	150.14	156.79
Trauma data	72.42	46.04
Heart disease	90.82	71.59
Contraceptive method choice	613.48	610.22
Thyroid disease	225.21	169.57
Cardiac arrhythmia	175.54	154.42
Prostate cancer	241.38	218.02

4.5. Advantage of our initialization procedure

Table V (a) and (b) shows some details of our calculations to illustrate the advantage of our initialization procedure over a totally random initialization. The two parts (a) BUPA liver disorder and (b) Pima Indian diabetes show the 'best' and the 'worst' case. In panel (a), we see that the LR initialization gives the two smallest values of the deviance even when compared with 100 random initializations; in panel (b) we see that we need 100 and 25 random initializations to improve on those provided by LR. Considering also the means and the standard deviations reported in the table, it is clear that our procedure has *on the average* a computational advantage on random initialization.

5. DISCUSSION

This work was motivated by the variability of results and opinions on the effectiveness of BP-ANN in biostatistics. Negative opinions about BP-ANN usually compare these with logistic regression and the results reported in the literature are indeed, on occasions, rather disappointing, with logistic regression *outperforming* at times BP-ANN (for example, Ennis *et al.* [8], Duh *et al.* [15], Manel *et al.* [16], Bertels *et al.* [17]). What appeared puzzling to us was that BP-ANN cannot *in theory* do worse than logistic regression, since one can always build a BP-ANN which is practically equivalent to *any given* LR. Therefore we devised an approach which guarantees that, starting from a specific LR model, the trained BP-ANN will be at least as good as the starting point *on the training set*: this can be achieved by simply initializing the BP-ANN with initial conditions corresponding to the specific LR model. Now, nothing can guarantee that the advantage of the BP-ANN thus trained will be maintained, in general, on the test set. However, by grafting onto our basic approach a technique that attempts to improve the generalizability of the BP-ANN predictors, we could reasonably hope that, in many situations, the advantage, if real, would be maintained on the test set. The technique proposed in this paper is a combination of *regularization* and cross-validation.

Table V. Estimated deviances of BP-ANN with weights initialized by randomly selection and logistic regression coefficients: illustrative examples.

<i>(a) BUPA liver disorder data, 221 training observations</i>					
<i>Units = 11 $\lambda = 0$</i>					
Initial weights by coefficients of LR					
Procedure	Delta	Deviance	Rank		
(i)	0	0.0026	1		
(ii)	0	100.5526	9		
(iii)	0.0001	0.0147	2		
(iv)	0.0001	9.1489	7		
Randomly initial weights					
Initial times		Min		Mean DEV	SD of DEV
5		10.7424	8	40.1655	25.9960
10		2.6560	4	36.1050	22.2122
25		8.1404	6	38.1858	18.6733
50		3.1083	5	40.2900	19.7436
100		0.0325	3	37.4654	20.7070
<i>Units = 13 $\lambda = 0.001$</i>					
Initial weights by coefficients of LR					
Procedure	Delta	Deviance	Rank		
(i)	0.1	1.0648	3		
(ii)	0	1.1492	3		
(iii)	0	1.1979	5		
(iv)	0.0001	0.7580	1		
Randomly initial weights					
Initial times		Min		Mean DEV	SD of DEV
5		2.1101	9	16.7498	13.864
10		1.5265	7	16.1895	16.3653
25		1.9312	8	14.0940	13.2906
50		1.2560	6	13.0750	12.3589
100		1.0423	2	13.0803	11.7400
<i>(b) Pima Indians diabetes, 523 training observations</i>					
<i>Units = 11 $\lambda = 0$</i>					
Initial weights by coefficients of LR					
Procedure	Delta	Deviance	Rank		
(i)	0.1	117.6598	9		
(ii)	0.0001	81.2028	5		
(iii)	0.0001	87.7464	7		
(iv)	0.0001	60.0836	2		
Randomly initial weights					
Initial times		Min		Mean DEV	SD of DEV
5		86.4673	6	118.4901	39.0434
10		102.0752	8	133.1727	24.3364
25		74.7851	3	113.7233	26.7911
50		78.1598	4	123.5483	24.3911
100		52.0669	1	118.7338	25.5399
<i>Units = 13 $\lambda = 0.001$</i>					
Initial weights by coefficients of LR					
Procedure	Delta	Deviance	Rank		
(i)	0.0001	49.9360	6		
(ii)	0.0001	32.9683	4		
(iii)	0.1	73.8939	9		
(iv)	0.1	67.1244	8		
Randomly initial weights					
Initial times		Min		Mean DEV	SD of DEV
5		58.6293	7	65.3886	7.6123
10		38.9228	5	65.9354	14.5628
25		30.7601	3	60.1752	13.8929
50		25.3404	2	61.6327	15.2476
100		19.4320	1	61.5154	16.1762

Delta is the value of the initial weights which are indicated by a dashed line in Figure 1.

The main limitation of our approach is that we have used as initial LR model two variants of a linear, no-interaction model: one contains all variables, and the other containing only those variables which are retained after an automatic stepwise selection is applied. This does not correspond to biostatistical practice at its best. Certainly it would have been preferable to define a starting model as the one obtained through a more thoughtful strategy reflecting the state-of-the-art. However, it would be difficult to reach a consensus on the definition of the state-of-the-art in biostatistical modelling. As stated in the method section, we acknowledge that this limits the scope of our comparison. We have only empirically compared the predictive accuracy of a *given LR model* with that of a BP-ANN based on this model; we have found that the latter consistently outperforms the former. We make *no claim* of having compared LR with BP-ANN in general. The interest of this result lies elsewhere. Indeed, BP-ANN theory guarantees that such result can be achieved, but it does not provide prescriptions on how to achieve it in practice. We are not aware of any published work that develops and evaluates a strategy, such as the one we propose, which empirically achieves what theory promises.

Another limitation is that we make no attempt to compare predictors on grounds other than predictive accuracy. It would be indeed useful to try and quantify the trade-off between gain in predictive accuracy and actual cost of model estimation. However, such a comparison would heavily depend on the definition of cost: it would be extremely difficult to reach a consensus on how to evaluate cost in this context. Furthermore, any cost attributed to computations would quickly become obsolete as computing power evolves at an ever-increasing rate. Indeed it may appear that a single BP-ANN predictor is obtained by fitting as many models as there are initialization procedures; in actual fact our proposed procedure *reduces* initialization costs since, as we have shown, one needs fewer 'intelligent initialization' than random ones. In any case, our limited experience suggests that proper programming of our procedure would almost completely hide its computational burden to potential users.

The reasonable performance of the proposed approach is established on the basis of the empirical evidence presented here. Indeed in *all* of the ten databases analysed in this paper, we have seen that BP-ANN outperforms the two LR models used for its construction, when the evaluation is based on the criterion used for training, namely deviance; this is true, obviously, on the training set, but also, much less obviously, on the test set. On the other hand, and not surprisingly, we cannot always establish the statistical significance of this improvement, especially when the size of the database is small.

As for the clinical significance of the observed improvement, we leave it to the reader to judge. From our point of view, at this stage of the development of BP-ANN technology, it is urgent to be concerned with statistical criteria for assessing and improving generalizability of non-linear predictors. There is a need of tools for detecting, with reasonable assurance, even relatively small improvements of BP-ANN predictors with respect to more familiar ones. Only when these tools are readily available and understood will we be able to decide, on the basis of experience, whether BP-ANN can truly improve our understanding of clinical data. We hope that the attempt presented here shows that the development of the proper tools is both possible and fruitful.

ACKNOWLEDGEMENTS

Partially supported by the Canadian National Science and Engineering Research Council (NSERC).

REFERENCES

1. Tu JV. Advantages and disadvantages of using neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology* 1996; **49**(11):1225–1231.
2. Nielsen M, Schukken YH, Brand A, Haring S, *et al.* Comparison of analysis techniques for on-line detection of clinical mastitis. *Journal of Dairy Science* 1995; **78**(5):1050–1061.
3. Selker HP, Griffith JL, Patil S, Long WJ, D'Agostino RB. A comparison of performance of mathematical predictive methods for medical diagnosis: identifying acute cardiac ischemia among emergency department patients. *Journal of Investigative Medicine* 1995; **43**(5):468–476.
4. Marchevsky AM, Shah S, Patel S. Reasoning with uncertainty in pathology: artificial neural networks and logistic regression as tools for prediction of lymph node status in breast cancer patients. *Modern Pathology* 1999; **12**(5):505–513.
5. Tu JV, Weinstein MC, McNeil BJ, Naylor CD, *et al.* Predicting mortality after coronary artery bypass surgery: what do artificial neural networks learn? *Medical Decision Making* 1998; **18**(2):229–235.
6. Penny W, Frost D. Neural networks in clinical medicine. *Medical Decision Making* 1996; **16**(4):386–398.
7. Lapuerta P, L'Italien GJ, Paul S, Hendel RC, *et al.* Neural network assessment of perioperative cardiac risk in vascular surgery patients. *Medical Decision Making* 1998; **18**(1):70–75.
8. Ennis M, Hinton G, Naylor D, Revow M, Tibshirani R. A comparison of statistical learning methods on the GUSTO database. *Statistics in Medicine* 1998; **17**(21):2501–2508.
9. Carpenter GA, Markuzon N. ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases. *Neural Networks* 1998; **11**(2):323–336.
10. Curram SP, Mingers J. Neural networks, decision tree induction and discriminant analysis: an empirical comparison. *Journal of the Operational Research Society* 1994; **45**(4):440–450.
11. Jefferson MF, Pendeton N, Lucas SB, Horan MA. Comparison of genetic algorithm neural network with logistic regression for predicting outcome after surgery for patients with non-small cell lung carcinoma. *Cancer* 1997; **79**(7):1338–1342.
12. Michael WK, Robert JB. Artificial neural networks for medical classification decisions. *Archives of Pathology & Laboratory Medicine* 1995; **119**(8):672–677.
13. Bostwick DG. Practical clinical application of predictive factors in prostate cancer: a review with an emphasis on quantitative methods in tissue specimens. *Analytical & Quantitative Cytology and Histology* 1998; **20**(5):323–342.
14. Jain BA, Nag BN. Performance evaluation of neural network decision models. *Journal of Management Information System* 1997; **14**(2):201–216.
15. Duh MS, Walker AM, Pagano M, Kronlund K. Prediction and cross-validation of neural networks versus logistic regression: using hepatic disorder as an example. *American Journal of Epidemiology* 1998; **147**(4):407–413.
16. Manel S, Dias JM, Ormerod SJ. Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. *Ecological Modelling* 1999; **120**(2–3):337–347.
17. Bertels K, Jacques JM, Neuberger L, Gatot L. Qualitative company performance evaluation: linear discriminant analysis and neural network models. *European Journal of Operational Research* 1999; **115**(3):608–615.
18. Venables WN, Ripley BD. *Modern Applied Statistics with S-PLUS*. Springer-Verlag: New York, 1997.
19. Smith JW, Everhart, JE, Dickson WC, Knowler WC, Johannes RS. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Symposium on Computer Applications and Medical Care*. IEEE Computer Society Press: 1988; 261–265.
20. Prechelt L. PROBEN1 – a set of neural network benchmark problems and benchmarking rules. Technical report 21/94, 1994; ftp.ira.uka.de.
21. Ripley BD. *Pattern Recognition and Neural Networks*. Cambridge University Press: Cambridge, 1996.
22. Wolberg WH, Mangasarian OL. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the National Academy of Sciences, U.S.A.* 1990; **87**(23):9193–9196.
23. Haberman SJ. Generalized residuals for log-linear models. *Proceedings of the 9th International Biometrics Conference*, 1976, Boston, 104–122.
24. Landwehr JM, Pregibon D, Shoemaker AC. Graphical methods for assessing logistic regression models (with discussion). *Journal of the American Statistical Association* 1984; **79**(385):61–83.
25. Christensen R. *Log-Linear Models and Logistic Regression*. 2nd edn. Springer-Verlag: New York, 1997.
26. Gennari JH, Langley P, Fisher D. Models of incremental concept formation. *Artificial Intelligence* 1989; **40**(1):11–61.
27. Aha DW, Kibler D, Albert MK. Instance-based learning algorithms. *Machine Learning* 1991; **6**(1):37–66.
28. Lim TS, Loh WY, Shih YS. *A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms*. Kluwer Academic Publishers: Boston, 1999 (<http://www.stat.wisc.edu/~llimt/compare.ps>).

29. Harrell FE. Comparison of strategies for validating binary logistic regression models. Technical reports from the Division of Biostatistics and Epidemiology, Virginia University, 1998 (<http://hesweb1.med.virginia.edu/biostat/reports/logistic.val.pdf>).
30. Tourassi GD, Floyd CE. The effect of data sampling on the performance evaluation of artificial neural networks in medical diagnosis. *Medical Decision Making* 1997; **17**(2):186–192.
31. Schwarzer G, Vach W, Schumacher M. On the misuse of artificial neural networks for prognostic and diagnostic classification in oncology. *Statistics in Medicine* 2000; **19**(4):541–561.
32. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a non parametric approach. *Biometrics* 1988; **44**(11):837–845.