ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Biomedical Informatics 35 (2002) 352–359

Journal of Biomedical Informatics

www.elsevier.com/locate/yjbin

Methodological Review

Logistic regression and artificial neural network classification models: a methodology review

Stephan Dreiseitl[a,*] and Lucila Ohno-Machado[b]

[a] *Department of Software Engineering for Medicine, Upper Austria University of Applied Sciences, Hagenberg, Austria*
[b] *Decision Systems Group, Brigham and Women's Hospital, Division of Health Sciences and Technology, Harvard Medical School and Massachusetts Institute of Technology, Boston, MA, USA*

Received 7 February 2003

Abstract

Logistic regression and artificial neural networks are the models of choice in many medical data classification tasks. In this review, we summarize the differences and similarities of these models from a technical point of view, and compare them with other machine learning algorithms. We provide considerations useful for critically assessing the quality of the models and the results based on these models. Finally, we summarize our findings on how quality criteria for logistic regression and artificial neural network models are met in a sample of papers from the medical literature.

Keywords: Artificial neural networks; Logistic regression; Classification; Model comparison; Model evaluation; Medical data analysis

1. Introduction

Predictive models are used in a variety of medical domains for diagnostic and prognostic tasks. These models are built from "experience", which constitutes data acquired from actual cases. The data can be preprocessed and expressed in a set of rules, such as it is often the case in knowledge-based expert systems, or serve as training data for statistical and machine learning models. Among the options in the latter category, the most popular models in medicine are logistic regression (LR) and artificial neural networks (ANN). These models have their origins in two different communities (statistics and computer science), but share many similarities.

In this article, we show that logistic regression and artificial neural networks share common roots in statistical pattern recognition, and how the latter model can be seen as a generalization of the former. We briefly compare these two methods with other popular classification algorithms from the machine learning field, such as k-nearest neighbors, decision trees, and support vector machines.

There are now several implementations of predictive modeling algorithms readily available, both as free and commercial software. The quality of the results obtained using these models mainly depends on three factors: the quality of the data set employed in model-building, the care with which adjustable model parameters were chosen, and the evaluation criteria used to report the results of the modeling process.

It is imperative that these details be presented in papers using predictive modeling, as otherwise the validity of the claims in the papers cannot be assessed by the reader. We therefore analyze the model-building process of logistic regression and neural network models in some detail, and point out which factors need to be considered when judging research results using predictive models.

To gauge the current state of reporting results in the literature, we sampled 72 papers comparing both logistic regression and neural network models on medical data sets. We analyzed these papers with respect to several criteria, such as size of data sets, model parameter

* Corresponding author. Fax: +43-7236-3888-2099.
E-mail address: Stephan.Dreiseitl@fh-hagenberg.at (S. Dreiseitl).

1532-0464/02/$ - see front matter © 2003 Elsevier Science (USA). All rights reserved.
doi:10.1016/S1532-0464(03)00034-0

selection scheme, and performance measure used in reporting model results.

2. The data classification task

The task of classifying data is to decide class membership y' of an unknown data item x' based on a data set $D = (x_1, y_1), \ldots, (x_n, y_n)$ of data items x_i with known class memberships y_i. For ease of discussion, we consider only dichotomous classification problems, where the class labels y are either 0 or 1. The x_i are usually m-dimensional vectors, the components of which are called *covariates* and *independent variables* (in statistics parlance) or *input variables* (by the machine learning community). In most problem domains, there is no functional relationship $y = f(x)$ between y and x. In this case, the relationship between x and y has to be described more generally by a probability distribution $P(x, y)$; one then assumes that the data set D contains independent samples from P. From statistical decision theory, it is well known that the optimal class membership decision is to choose the class label y that maximizes the posterior distribution $P(y|x)$ [1].

There are two different approaches to data classification: the first considers only a dichotomous distinction between the two classes, and assigns class labels 0 or 1 to an unknown data item. The second attempts to model $P(y|x)$; this yields not only a class label for a data item, but also a probability of class membership. The most prominent representatives of the first class are support vector machines. Logistic regression, artificial neural networks, k-nearest neighbors, and decision trees are all members of the second class, although they vary considerably in building an approximation to $P(y|x)$ from data. Some details on these models, including a comparison on their respective advantages and disadvantages, are given below.

Currently, logistic regression and artificial neural networks are the most widely used models in biomedicine, as measured by the number of publications indexed in MEDLINE: 28,500 for logistic regression, 8500 for neural networks, 1300 for k-nearest neighbors, 1100 for decision trees, and 100 for support vector machines.

2.1. Support vector machines

These models are algorithmic implementations of ideas from statistical learning theory [2], which concerns itself with the problem of building *consistent estimators* from data: how can the performance of a model on an unknown data set be estimated, given only characteristics of the model, and performance on a training set?

Algorithmically, support vector machines build optimal separating boundaries between data sets by solving a constrained quadratic optimization problem [3,4]. By using different *kernel functions*, varying degrees of nonlinearity and flexibility can be included in the model. Because they can be derived from advanced statistical ideas, and bounds on the generalization error can be calculated for them, support vector machines have received considerable research interest over the past years. Performances on par with or exceeding that of other machine learning algorithms have been reported in the medical literature.

The disadvantage of support vector machines is that the classification result is purely dichotomous, and no probability of class membership is given.

2.2. k-Nearest neighbors

Classification based on the k-nearest neighbor algorithm differs from the other methods considered here, as this algorithm uses the data directly for classification, without building a model first [5,6]. As such, no details of model construction need to be considered, and the only adjustable parameter in the model is k, the number of nearest neighbors to include in the estimate of class membership: the value of $P(y|x)$ is calculated simply as the ratio of members of class y among the k nearest neighbors of x. By varying k, the model can be made more or less flexible (small or large values of k, respectively).

The advantage that k-nearest neighbors have over other algorithms is the fact that the neighbors can provide an explanation for the classification result; this *case-based explanation* can provide an advantage in areas where black-box models are inadequate.

The major drawback of k-nearest neighbors lies in the calculation of the case neighborhood: for this, one needs to define a metric that measures the distance between data items. In most application areas, it is not clear how to, other than by trial and error, define a metric in such a way that the relative (but unknown!) importance of data components is reflected in the metric.

2.3. Decision trees

This algorithm repeatedly splits the data set according to a criterion that maximizes the separation of the data, resulting in a tree-like structure [7,8]. The most common criterion employed is *information gain*; this means that at each split, the decrease in entropy due to this split is maximized. The estimate of $P(y|x)$ is the ratio of y class elements over all elements of the leaf node that contains data item x.

A major disadvantage of decision trees is given by the greedy construction process: at each step, the combination of single best variable and optimal split-point is selected; however, a multi-step lookahead that considers *combinations* of variables may obtain different (and better) results. A further drawback lies in the fact that

continuous variables are implicitly discretized by the splitting process, losing information along the way.

Compared with the other machine learning methods mentioned here, decision trees have the advantage that they are not black-box models, but can easily be expressed as rules. In many application domains, this advantage weighs more heavily than the drawbacks, so that these models are widely used in medicine.

2.4. Logistic regression and artificial neural networks

These models differ from the three algorithms above in the sense that they both provide a functional form f and parameter vector α to express $P(y|x)$ as

$$P(y|x) = f(x, \alpha).$$

The parameters α are determined based on the data set D, usually by maximum-likelihood estimation. As the functional form of f differs for logistic regression and artificial neural nets, the former is known as a *parametric method*, whereas the latter is sometimes called *semi-parametric* or *non-parametric*. This distinction is important because the contribution of parameters in logistic regression (coefficients and intercept) can be interpreted, whereas this is not always the case with the parameters of a neural network (weights).

3. Logistic regression vs. artificial neural network models

For the following, let all data vectors x_i contain an additional component 1. This will facilitate notation in allowing us to write a simple dot product $\alpha \cdot x$ for a linear combination of vector components instead of the more cumbersome $\alpha \cdot x + \alpha_0$.

Generally, a logistic regression model calculates the class membership probability for one of the two categories in the data set:

$$P(1|x, \alpha) = \frac{1}{1 + \mathrm{e}^{-(\alpha \cdot x)}},$$

and $P(0|x, \alpha) = 1 - P(1|x, \alpha)$. Here, we write $P(1|x, \alpha)$ to make the dependence of the posterior distribution on the parameters α explicit. It can be shown that this model is correct when both the class-conditional densities $p(x|1)$ and $p(x|0)$ are multinormal with equal covariance matrices [6].

The hyperplane of all points x satisfying the equation $\alpha \cdot x = 0$ forms the *decision boundary* between the two classes; these are the points for which $P(1|x, \alpha) = P(0|x, \alpha) = 0.5$. A logistic regression model that includes only the original covariates is called a *main effects model*; including *interaction terms* such as products makes the model nonlinear in the covariates, and therefore more flexible. Although higher flexibility may be desirable in general, it carries with it a higher risk for model overfitting ("memorizing the training cases"), which can potentially reduce a model's accuracy on previously unseen cases. In predictive modeling, fitting the training cases is just part of the task: correctly classifying new cases is the most important goal.

Maximum likelihood estimation of the optimal parameter values α requires the maximization of $\prod_{i=1}^{n} P(y_i|x_i, \alpha)$. Although the functional forms for logistic regression and artificial neural network models are quite different, a network without a hidden layer is actually identical to a logistic regression model if the logistic (sigmoidal) activation function is used [9,10].

Artificial neural networks are aggregations of perceptrons. For multi-layer feedforward networks, the output is

$$o_{\mathrm{N}} = \frac{1}{1 + \mathrm{e}^{-(\beta \cdot o_{\mathrm{H}} + \beta_0)}},$$

and this output is again taken as $P(1|x, \beta, \beta_0, \alpha)$. Here, o_{H} is a vector of perceptron outputs, each with its own α parameters; these perceptrons are usually called *hidden neurons*. Due to the nonlinearity in these hidden neurons, the output o_{N} of an artificial neural network is a nonlinear function of the inputs. In a classification context, this means that the decision boundary can be nonlinear as well, making the model more flexible compared to logistic regression. In Section 4, we summarize a sampling of publications from the biomedical field to assess whether this higher degree of flexibility results in improved classification accuracy on real-world data sets.

3.1. Parameter estimation techniques

For both logistic regression and artificial neural networks, the model parameters are determined by maximum likelihood estimation, i.e., the parameters α are chosen to maximize $\prod_{i=1}^{n} P(y_i|x_i, \alpha)$. Usually, it is easier (and equivalent) to minimize $-\sum_{i=1}^{n} \log P(y_i|x_i, \alpha)$. A variety of numerical optimization algorithms, from simple gradient descent to more complicated second-order methods, can be used to determine the optimal parameter values [11].

Artificial neural networks are usually trained by minimizing an error function; an appropriate choice of such a function for binary classification problems is the *cross-entropy error*

$$E = \sum_{i=1}^{n} y \log o_{\mathrm{N}} + (1 - y) \log(1 - o_{\mathrm{N}}).$$

Given only a limited size data set D, *any* model for $P(y|x)$ based on this data set will be influenced by the particular choice of D. The model-building challenge is to abstract the underlying distribution from the particular instance D of samples. The problem of memorizing the data set instead of identifying the underlying

distribution is known as *overfitting*. Various methods to avoid overfitting have emerged over the years; these can be categorized as either restricting model complexity, or restricting the influence of the data on the model parameters.

In logistic regression, the model complexity is already low, especially when no or few interaction terms and variable transformations are used. Overfitting is less of an issue in this case. Performing variable selection is a way to reduce a model's complexity and consequently decrease the risk of overfitting. As mentioned before, this may cause a loss in the model's flexibility. Compared to logistic regression, neural network models are more flexible, and thus more susceptible to overfitting. Network size can be restricted by decreasing the number of variables and hidden neurons, and by pruning the network after training. Alternatively, one can require the model output to be sufficiently smooth. This can be achieved by *regularization*; in a neural network context this is called *weight decay*. Weight decay, as the name implies, limits the magnitude of the weights and is a method that is analogous to logistic regression's *shrinkage* [10,12]. Weight decay and shrinkage make decision boundaries smoother. Sufficiently smooth decision boundaries are not as flexible as unrestricted decision boundaries, so that they cannot adapt to the particularities of a data set. For weight decay, one needs to empirically determine a weight reduction factor; this is usually done by cross-validation or bootstrapping (see Section 3.3).

The alternative to restricting model complexity is to only partially adapt the model to the data set. This can be achieved by *early stopping*, when parameter adaptation is terminated before the maximum-likelihood estimate is found. The use of early stopping requires a subset of the training data to be used as a *holdout* set, to terminate training when adaptation shifts away from the data generator to the particular instance of data set.

The *Bayesian framework* provides an alternative to maximum-likelihood parameter estimation, and thus to the problem of overfitting. In this framework, one does not calculate a single best parameter vector α_{ML}, but rather a distribution $P(\alpha|D)$ over the parameters as

$$P(\alpha|D) = \frac{P(D|\alpha)P(\alpha)}{P(D)}.$$

In this equation, the denominator does not depend on α and can therefore be ignored. In the remaining term, $P(D|\alpha) = \prod_{i=1}^{n} P(y_i|x_i, \alpha)$ is the likelihood, and $P(\alpha)$ the prior distribution over the parameters. For large data sets, the posterior $P(\alpha|D)$ becomes sharply peaked around α_{ML}, so that the choice of prior distribution has little effect on the calculation. For smaller data sets, this influence is more pronounced, and can be used to incorporate prior knowledge into the model [9,13]. Weight decay can be seen as representative of this reasoning, as smooth decision boundaries correspond to a preference for smaller weights.

In-depth discussions of the topics mentioned here can be found in the books of Bishop [9] and Ripley [6] for artificial neural networks, Neal [14] for a Bayesian perspective on neural network training, and Hosmer and Lemeshow [15] and Harrell [16] for logistic regression.

3.2. Variable selection

In many application domains, it is not only important to be able to separate two data sets, but also to determine which variables are the most relevant for achieving this separation. On the one hand, the removal of superfluous variables can lead to more accurate models; on the other, money, time and effort can be saved by dropping unnecessary tests or asking only relevant questions.

For logistic regression models, it is possible to test the statistical significance of the coefficients in the model [15]; these tests can be used to build models incrementally. The three most common approaches are to start with an empty model and successively add covariates (*forward selection*), to start with the full model and remove covariates (*backward selection*), or to both add and remove variables (*stepwise selection*).

Due to the nonlinear nature of artificial neural networks, the statistical tests for parameter significance that are used in logistic regression cannot be applied here. Instead, one can use *automatic relevance determination* [9] or *sensitivity analysis* [17] to heuristically assess the importance of input variables for the classification result.

3.3. Model evaluation

The two criteria to assess the quality of a classification model are *discrimination* and *calibration*. Discrimination is a measure of how well the two classes in the data set are separated; calibration determines how accurate the model probability estimate $f(x, \alpha)$ is to the true probability $P(y|x)$. To provide an unbiased estimate of a model's discrimination and calibration, these values have to be calculated from a data set not used in the model building process. Usually, a portion of the original data set, called the *test* or *validation set*, is put aside for this purpose. In small data sets, there may not be enough data items for both training and testing. In this case, the whole data set is divided into n pieces, $n-1$ pieces are used for training, and the last piece is the test set. This process of *n-fold cross-validation* builds n models; the numbers reported are the averages over all n test sets [18,19]. The extreme case of using only one data item for testing is known as *leave-one-out cross-validation*.

An alternative to cross-validation is *bootstrapping*, a process by which training sets are sampled with

replacement from the original data sets [20]. When using the complete data set as testing data, the estimate of generalization error will be too low, as data items are used for both training and testing. It is, however, possible to estimate the bias, and thus to adjust an overly optimistic generalization error estimate. Bootstrapping was shown to be superior to cross-validation on many data sets [16,21].

Common measures of discrimination are *sensitivity*, *specificity*, *accuracy* and the *area under the ROC curve* (or, equivalently, the *c-index*). For all these measures, there exist statistical tests to determine whether one model exceeds another in discrimination ability [22,23].

Calibration is a measure of how close the predictions of a given model are to the real underlying probability. Almost always, the true underlying probability is unknown and can only be estimated retrospectively by verifying the true binary outcome of the data being studied. Calibration thus measures the similarity between two different estimates of a probability. One of the ways to assess calibration is to take the difference between the average observation and the average outcome of a given group as a measure of discalibration. A more refined way to measure calibration requires dividing the sample into smaller groups sorted by predictions, calculating the sum of predictions and sum of outcomes for each group, and determining whether there are any statistically significant differences between the expected and observed numbers by a simple χ^2 method [15].

4. Logistic regression and artificial neural network comparisons in the literature

We reviewed 72 papers that compare the classification performance of artificial neural networks with logistic regression models. The references were obtained as a sample from PUBMED and chosen for ease of availability; a general literature review is beyond the scope of this paper.

The objective of this sampling was to determine the overall standard of publications reporting results based on logistic regression and artificial neural network modeling. We focused on those papers that use both methodologies to see whether one of them consistently outperforms the other on medical data sets.

For this study, we analyzed the 72 papers with respect to the following criteria: whether details of the model building process are given (variable selection scheme for logistic regression, parameter selection and overfitting avoidance for artificial neural networks), whether unbiased estimates of the generalization error are reported (by using test sets, cross-validation, or bootstrapping), whether measures of discriminatory power were given (and statistical testing using these measures), and whether calibration information is included.

Every paper was rated in each of these five categories as either giving details or not. The latter was the case when no details were reported in the paper, or when the methodology used in the paper was questionable (such as not taking overfitting avoidance into account, or reporting a model's superiority over another without statistical testing).

The results of this survey are summarized in Table 1. It is interesting to note that details on model building are given more often for logistic regression than for artificial neural networks. This may be due to the fact that forward, backward, and stepwise variable selection schemes are implemented in standard logistic regression software, and thus easily used and reported. It takes more effort and considerations on the part of the user to achieve the same level of sophistication with artificial neural networks, as many advanced methods are not available in all software packages. These model building details may also be considered not important for publication by authors, although they help to assess the quality of the findings obtained with the model.

Since all the papers surveyed compare the performance of logistic regression with artificial neural networks in discriminating two data sets, it is understandable that only a quarter of them give calibration information.

The results of comparing the discriminatory power of logistic regression and artificial neural network models are summarized in Table 2. It can be seen that both models perform on about the same level more often than not, with the more flexible neural networks generally outperforming logistic regression in the remaining cases.

Table 1
Percentage of papers (out of 72) satisfying five quality criteria

	Details given (%)	Details not given (%)
LR model building details	76	24
ANN model building details	51	49
Generalization error estimate	89	11
Statistical discriminant testing	61	39
Calibration information	25	75

Table 2
Summary of comparing the discriminatory power of artificial neural networks with logistic regression models, as percentage of 72 papers

	ANN better (%)	LR better (%)	No difference (%)
Stat. testing	18	1	42
No stat. testing	33	6	0

5. Discussion

An increasingly large number of data items are collected routinely, and often automatically, in many areas of medicine. It is a challenge for the field of machine learning and statistics to extract useful information and knowledge from this wealth of data.

Mistakes in model building and evaluation can have disastrous consequences in some medical applications. Special care must therefore be taken to ensure that the models are validated, preferably by using an external data set and checking the model's plausibility by surveying a panel of experts in the domain [24,25].

The latter is possible only for so-called *white-box models* that allow an interpretation of model parameters. Examples of such algorithms are decision trees (which may be expressed as a set of rules), *k*-nearest neighbors (which provides exemplars similar to cases to be classified), and logistic regression (where coefficients' sizes determine their relative importance for the classification result).

Black-box models, such as support vector machines or artificial neural networks, do not allow such an interpretation, and can only be verified externally. Contrasting views on the role of artificial neural networks as predictive models are given in [26,27]. Nevertheless, their discriminating power is often significantly better than that of white-box models, which may explain their popularity in domains where classification performance is more important than model interpretation.

Most of the papers summarized in Section 4 have shown logistic regression and artificial neural networks to work well on a wide variety of data sets. Their performance is generally better, at least on continuous data, than that of decision trees and *k*-nearest neighbors. This may be explained by the fact that the decision tree algorithm does not construct a decision boundary between classes per se, but rather splits the data set optimally at each tree node. As explained in Section 2, this may result in suboptimal classification results. The performance of *k*-nearest neighbors is generally worse on high-dimensional data because, when the relative importance of dimensions is not weighted, the data from spurious and irrelevant dimensions may negatively influence the distance calculation [28].

Support vector machines, on the other hand, have shown comparable performance in the few studies on medical data sets [29,30]. They are not as widely used yet as logistic regression and artificial neural networks, in part because few easy-to-use software implementations are available, and the kernel functions and kernel function parameter settings have to be estimated from the data (mostly by cross-validation or bootstrapping).

In short, the widespread use of logistic regression and artificial neural network models seems to be motivated by the fact that they have lower generalization error than decision trees and *k*-nearest neighbors, while being easier to build than support vector machines.

The following points should be kept in mind when using logistic regression and artificial neural networks as data classification tools; pitfalls to avoid when comparing classifiers are given in [31].

5.1. Logistic regression

With anything more than a few covariates, a variable selection scheme should be used to remove spurious covariates. If computationally feasible, one should include interaction terms to make the model more flexible. A variable selection scheme can then be used to remove unnecessary interaction terms. The p value for statistical testing of variable significance for inclusion in and exclusion from the model is generally set to 0.05, but this threshold should be modified given expert opinion.

5.2. Artificial neural networks

One layer of hidden neurons is generally sufficient for classifying most data sets. The number of neurons in the hidden layer needs to be set empirically, e.g. by cross-validation or bootstrapping. One should avoid the use of plain backpropagation or backpropagation with momentum, as these minimization algorithms are slower to convergence than second-order algorithms such as conjugate gradients or quasi-Newton methods. It is imperative to not overfit the network during training; this can be achieved either by restricting the topology of the network (i.e., decreasing the number of nodes), by early stopping, or by using weight decay. If computationally possible, one should consider the use of a Bayesian approach that averages over several plausible networks.

5.3. Estimate of generalization error

A classification result may be overly optimistic if performance cannot be measured on a data set not used for model building. In the ideal case, testing on a separate data set will provide an unbiased estimate of generalization error. If the original data set is too small for this approach, the recommended strategy is to use cross-validation or bootstrapping to make the best possible use of the limited amount of data.

A discussion of model evaluation, especially as it pertains to medical data and the use of logistic regression and artificial neural network models, can be found in [32–35].

5.4. Measuring the discriminatory power of a model

The most commonly used measures of discriminatory power are the area under the ROC curve (AUC), sensitivity, specificity, and accuracy. While sensitivity and

specificity of a classifier are reported for a single threshold (mostly taken as classifier output $o = 0.5$), the area under the ROC curve represents a common measure of sensitivity and specificity over all possible thresholds. One should be aware, though, that the AUC measure remains the same when classifier outputs are transformed monotonically. This means that models may exhibit good discrimination (as measured by AUC), but may be poorly calibrated.

Accuracy is the only discrimination measure influenced by the class distribution in the data set. This measure must therefore be treated with caution when the case distribution in the training set is different from the case distribution of the population on which the classifier is used.

5.5. Assessing claims in the literature

Most studies on the use of classification algorithms in biomedicine focus on one of two questions:

- Is it possible to distinguish one class of data items from another, based on some set of measurements (features)?
- Is it possible to build a decision-support system that helps in the diagnosis/prognosis of unknown cases?

Although both approaches use the same methodology for model building, the use of performance indicators is different: for the first, the question of discrimination is more important, whereas for the second, good calibration is essential. For both of them, the appropriate performance indicators need to be published to substantiate any claims of model performance. More studies currently published are motivated by answering the first question, as can be seen from the relatively low number of papers that report calibration information (see Section 4).

When assessing the model-building process reported in a paper, one should check whether the following questions were addressed in a satisfactory manner:

- How is the choice of classifier motivated?
- How were the parameters of the classifier, and the parameters of the training process chosen?
- How is the performance of the classifier evaluated?

For logistic regression and artificial neural network models, the points to consider in evaluating the description of a model-building process are given in Section 3.

Answering the questions above allows the reader to determine the overall quality of the result reported in a paper, and to distinguish between overly optimistic claims (such as when performance is reported on the training set) and needlessly pessimistic ones (when model parameters are chosen in a suboptimal manner). The latter is especially common in studies that promote "new" algorithms. Needless to say, some articles do not even report comparisons and instead just report the performance of a single method.

6. Conclusion

In this methodology review, we explained the use of logistic regression and artificial neural network models for biomedical data classification. We outlined the common foundations of both models in statistical pattern recognition, and briefly compared these models with other classification algorithms. We showed how to build logistic regression and artificial neural network models, how to evaluate them, and which performance indices to report.

We surveyed papers that compare both models to determine the current level of publication standard, and noticed that the information relevant for measuring the methodological soundness of a paper is reported more often for logistic regression models. We conjecture that this is due to the fact that the model-building process is easier for logistic regression, and may be considered too detailed and not worthy of publication for artificial neural networks. This greatly limits the readers' ability to reproduce the reported results.

We discussed the application areas, relative merits and common pitfalls of classification algorithms in biomedicine. So far, there is no single algorithm that performs better than all other algorithms on any given data set and application area. For logistic regression, the popularity may be attributed to the interpretability of model parameters and ease of use; for artificial neural networks, this may be due to the fact that these models can be seen as nonlinear generalizations of logistic regression, and thus at least as powerful as that model. The evidence summarized in Section 4 shows that of the tasks where performance was compared statistically, there was a 5:2 ratio of cases in which it was not significantly better to use neural networks. It remains to be seen whether newer machine learning algorithms, such as support vector machines and other kernel-based algorithms, can prove to be significantly better than both logistic regression and artificial neural networks.

Until further studies are conducted and some guidelines for predictive modeling evaluation are utilized, there may continue to exist a publication bias in favor of the newer machine learning methods, often with disregard to proper evaluation of the results. This may mislead readers into thinking that the new methods are not subject to the pervasive trade-offs between flexibility and overfitting that are typical of classical models such as logistic regression and artificial neural networks.

References

[1] Duda R, Hart P, Stork D. Pattern classification. 2nd ed. New York: Wiley/Interscience; 2000.
[2] Vapnik V. The nature of statistical learning theory. 2nd ed. New York: Springer; 2000.

[3] Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge: Cambridge University Press; 2000.
[4] Schölkopf B, Smola A. Learning with kernels: support vector machines, regularization, optimization, and beyond. Cambridge, MA: MIT Press; 2002.
[5] Dasarathy B. Nearest neighbor pattern classification techniques. Silver Spring, MD: IEEE Computer Society Press; 1991.
[6] Ripley B. Pattern recognition and neural networks. Cambridge: Cambridge University Press; 1996.
[7] Breiman L et al. Classification and regression trees. Belmont, CA: Wadsworth; 1984.
[8] Quinlan R. C4.5: programs for machine learning. Los Altos, CA: Morgan Kaufmann; 1993.
[9] Bishop C. Neural networks for pattern recognition. Oxford: Oxford University Press; 1995.
[10] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. New York: Springer; 2001.
[11] Press W et al. Numerical recipes in C. 2nd ed. Cambridge: Cambridge University Press; 1993.
[12] Copas J. Regression, prediction and shrinkage (with discussion). J Roy Stat Soc B 1983;45:311–54.
[13] Gelfand A, Sahu S, Carlin B. Efficient parametrisations for generalized linear mixed models. In: Bernardo J et al., editors. Bayesian statistics, vol. 5. Oxford: Oxford University Press; 1996. p. 165–80.
[14] Neal R. Bayesian learning for neural networks. New York: Springer; 1996.
[15] Hosmer D, Lemeshow S. Applied logistic regression. 2nd ed. New York: Wiley; 2000.
[16] Harrell F. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer; 2001.
[17] Zurada J, Malinowski A, Cloete A. Sensitivity analysis for minimization of input dimension for feedforward neural networks. In: Proc IEEE Int Symp Circuits Systems, vol. 6; 1994. p. 447–50.
[18] Stone M. Cross-validatory choice and assessment of statistical predications. J Roy Stat Soc 1974;36:111–47.
[19] Allen D. The relationship between variable selection and data augmentation and a method of prediction. Technometric 1977;16:125–7.
[20] Efron B, Tibshirani R. An introduction to the bootstrap. London: Chapman & Hall; 1993.
[21] Efron B. Estimating the error rate of a prediction rule: some improvements on cross-validation. J Am Stat Assoc 1983;78:316–31.
[22] Hanley J, McNeil B. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:839–43.
[23] DeLong E, DeLong D, Clarke-Pearson D. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–45.
[24] Altman D, Rayston P. What do we mean by validating a prognostic model? Stat Med 2000;19:453–73.
[25] Vergouwe Y et al. Validity of prognostic models: when is a model clinically useful. Semin Urol Oncol 2002;20:96–107.
[26] Schwarzer G, Vach W, Schumacher M. On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. Stat Med 2000;19:541–61.
[27] Lisboa P. A review of evidence of health benefit from artificial neural networks in medical intervention. Neural Networks 2002;15:11–39.
[28] Mitchell T. Machine learning. New York: McGraw-Hill; 1997.
[29] Dreiseitl S et al. A comparison of machine learning methods for the diagnosis of pigmented skin lesions. J Biomed Inform 2001;34:28–36.
[30] Chang R et al. Support vector machines for diagnosis of breast tumors on US images. Acad Radiol 2003;10:189–97.
[31] Salzberg S. On comparing classifiers: pitfalls to avoid and a recommended approach. Data Min Knowl Disc 1997;1:317–28.
[32] Harrell F, Lee K, Mark D. Multivariable prognostic models: issues in developing models, evaluation assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361–87.
[33] Hilden J. Neural networks and the roles of cross validation. Med Decis Making 1998;18:122–4.
[34] Steyerberg E, Harrell F, Goodman P. Neural networks, logistic regression, and calibration. Med Decis Making 1998;18:349–50.
[35] Steyerberg E, Harrell F, Goodman P. Neural networks, logistic regression, and calibration: a rejoinder. Med Decis Making 1998;18:445–6.