Neural Networks vs Logistic Regression: a Comparative Study on a Large Data Set

Paulo J.L. Adeodato1,2, Germano C. Vasconcelos1,2, Adrian L. Amaud1, Roberto A.F. Santos1, Rodrigo C.L.V. Cunha1, Domingos S.M.P. Monteiro2

1 Computational Intelligence Group, Federal University of Pernambuco (UFPE), Recife, Pernambuco, Brazil
Email: {pjla,gcv,ala2,rafs,rchvc}@cin.ufpe.br
2 NeuroTech Ltd, Recife, Pernambuco, Brazil
Email: {paulo,germano,domingos}@neurotech.com.br

Abstract

Neural networks and logistic regression have been among the most widely used AI techniques in applications of pattern classification. Much has been discussed about if there is any significant difference in between them but much less has been actually done with real-world applications data (large scale) to help settle this matter, with a few exceptions. This paper presents a performance comparison between these two techniques on the market application of credit risk assessment, making use of a large database from an outstanding credit bureau and financial institution (a sample of 180,000 examples). The comparison was carried out through a 30-fold stratified cross-validation process to define the confidence intervals for the performance evaluation. Several metrics were applied both on the optimal decision point and along the continuous output domain. The statistical tests showed that multilayer perceptrons perform better than logistic regression at 95% confidence level, for all the metrics used.

1. Introduction

Some years ago, when artificial neural networks were booming and becoming very popular, researchers started questioning if such systems would bring any significant improvement in performance when compared to the more traditional statistical classification techniques. This question was of particular concern to researchers, professionals and companies interested in efficiently solving the real-world problem of credit risk assessment of consumers [1]. At that time, training time was a matter of real concern among neural network practitioners and represented an actual constraint in carrying out a thorough comparison of the technique with the others available. Nowadays, the increase in computing power has turned data mining into a reality, and made possible the effective use of neural networks as one of its most important tools.

Despite the increase both in volume of data and in computing power available, the experimental comparisons among different techniques have yet much to benefit from this evolution.

A relevant approach has been carried out on large samples of synthetic data [2]. Kiang has compared five among the most widely used classification techniques on large samples of data. The control of the data features yielded the generation of data sets to either comply with some of the techniques’ hypotheses or to violate them, thus highlighting their deficiencies and the effects of each type of hypothesis violation.

This paper, in particular, aims at casting some light and reaching some conclusions on the comparison between neural networks and logistic regression. A simulation framework with rigorous statistical basis is conducted to evaluate these techniques on a data set that combines the feature of being composed of real world examples with the characteristic of having very large size (180,000 examples). The following 7 Sections describe this work.

2. Credit Risk Assessment

When a person applies for credit there are several ways to evaluate the request. Financial institutions define different cost functions to be optimized by the decision system (maximize their profit, minimize the risk of non-payment, minimize the delay in repayment etc.) Nevertheless, the main goal is to decide whether or not to give credit to the applicant, turning the task of credit risk assessment into a classification problem of only two classes: bad or good payers/payments (binary decision).

In the context of risk assessment here, the credit applicant supplies information such as social security number, annual income and address through an application form. More information is available for the decision if the applicant has already had a previous relationship with the institution and has fulfilled his/her obligations (behavioral information). Additional financial information for that social security number is also obtained by the institution from credit security database bureaux (e.g. Equifax).

All the information about the applicant (input variables) is fed into the credit concession system which awards a scalar score to the application request. If the score is above a certain threshold defined by the institution, the application is approved, otherwise, it is refused (the rejection score interval is not the focus of this paper). After the credit concession, the
institution keeps track of the credit payments and, according to the institution’s criteria, labels the payment as good or bad.

In this paper, credit is granted to the applicants through the guarantee of their cheques in commercial transactions. A payment is labeled as “bad” if it delays more than 60 days beyond the due date. With many more good payers than bad payers, this binary decision problem presents highly unbalanced a priori probabilities for each class and also very different costs for each type of decision errors.

3. Techniques Compared

The logistic regression and neural networks systems considered are those typically used for the application selected. The architectures and algorithms are the most commonly reported in the literature [3, 4] (nearly a standard). Both techniques have the common feature, which is essential for the evaluation criteria defined in Section-6: the scalar output for a binary classification problem. The techniques map the multidimensional input space into a unidimensional continuous output space, the latter playing the role of representing the “risk” of defaulting payment in the credit risk assessment domain.

Logistic Regression is a traditional statistical inference technique [3] that relates a binary output (such as good/bad, yes/no, 1/0 etc.) with a set of input variables \(x_1, x_2, \ldots, x_n\). The dependent variable is the logarithm of the ratio of the probabilities of the two possible outcomes of the output variable, \(\log \frac{p}{1-p}\), and can be turned into:

\[
P(Y = 1) = \frac{1}{1 + \exp(-\beta_0 - \beta_1 x_1 - \beta_2 x_2 - \ldots - \beta_n x_n)}
\]

where \(Y\) is the dichotomic output variable, \(\beta_0\) is the intersection and \(\beta_i\) (\(i=1, \ldots, k\)) are the coefficients corresponding to each explanatory variable \(x_i\). The widespread maximum likelihood method is used here for estimating the binary response having its solution found through Newton-Raphson’s iterative process. In the experimental setup in this paper, all the 30 numerical input variables were statistically significant at the level of 5% on the data sets and have been used in all the 30 experiments with the logistic regression technique.

Multilayer perceptrons (MLPs) trained with the backpropagation algorithm [4] have been the neural network models most frequently used in pattern classification. Amongst its attractive features, the excellent generalization capacity, the simplicity of operation and the ability to perform universal function approximation [5] can be regarded as the MLP’s most important characteristics. In the experimental setup here, the neural networks were multilayer perceptrons with a single hidden layer with 3 neurons (to constrain the possible sources of variations to a minimum). They were trained by the standard error backpropagation algorithm at a learning rate of 0.005, having the minimum squared error on the cross-validation set as the training stopping criterion. The cross-validation set represented 1/3 of the 174,000 examples.

4. Data Set Used

The data used as test bed for the comparison carried out in this paper has been collected from an operation of guarantee of cheque payments, of a outstanding credit information bureau. The credit concession decision that originated the database was made by a credit experts’ rule-based system which yielded a default rate of 2.0% on the guaranteed cheques. A payment is labeled as “bad” if it delays more than 60 days beyond the due date.

The data set consists a sample of 180,000 examples with the default payment rate of 2.0% observed in the credit operation. Human experts on the application domain selected the 30 most important numerical variables (mostly indicative of the consumers’ behavioral profile) from those present in the data-mart, as inputs for both decision making techniques under comparison.

5. Preprocessing

The data used in the experiments required few of the levels of data preparation needed in data mining: treatment of missing data, outlier removal, data transformation and normalization [6].

The reliability of the data is assured because most of the variable fields come from behavioral data kept in the data-mart and the few variables from the cheques have been verified for reimbursement of occasional payment default.

Missing data was treated through Statistical analysis to fill in the gaps in the database where appropriate. Statistical analysis also helped on the removal of the few outliers found. Fields representing time (e.g., years and months) needed to be combined for expressing their information in a more appropriate way (months). Numerical data were all normalized to fall in the interval between zero (0) and one (1).

6. Performance Evaluation and Methodology

As the main purpose of this research was to carry out a thorough comparison between neural networks and logistic regression, the formalism presented in this section is crucial for the relevance of the paper. Both the data set partitioning and the evaluation criteria were chosen according to their adequacy to the matter, to their assertive-ness and to the accuracy of the resulting assertions.

6.1. Stratified 30-fold cross-validation

The \(k\)-fold cross-validation method is a widely accepted way of dividing a single sample [7], for producing \(k\) experimental results on guaranteed statistically independent test sets (a partition). Class stratification in each of the \(k\) sets [8] is particularly important here due to the very unbalanced representation of each class. For \(k=30\), this approach produces confidence intervals for the paired performance measures on
each evaluation criterion described below, complying with the normality hypothesis.

Summing up, the original sample was divided for the stratified 30-fold cross-validation process with 2.0% of the sets' examples as default payments from a total of 6,000 examples. Both classification techniques were submitted to the same data sets (results paired for the comparison): trained 30 times on 174,000 gradually different examples and tested on 6,000 statistically independent examples.

6.2. Evaluation criteria

For techniques that produce a continuous output, the binary decision is reached through a threshold which depends on several actual practical factors such as the relative cost of the two types of decision error and the impact of the decision on the profitability, default payments rate, market size, among other criteria. For generality, this work presents a comparison throughout the whole continuous domain of decision (the score range) for each technique. This comparison was based on 4 performance metrics for binary classification based on continuous output, namely; Kolmogorov-Smirnov Curves and ROC Curves, and their maximum value and minimum distance to optimum, respectively. These metrics represent similar forms of performance evaluation with slightly different points of view. Nevertheless, they are compatible with the binary decision problem and widely accepted in this application domain. For each of these metrics, the results of the 30 test sets are averaged for each value on the X-axis of the plot, through threshold averaging [3].


The KS statistical method is a traditional non-parametric tool used for measuring the adherence of a cumulative distribution function (CDF) to the cumulative representation of the actual data [10]. In binary decision systems, this metric is applied for assessing the lack of adherence between the data sets from the 2 classes, having the score as independent variable. The Kolmogorov-Smirnov Curves are the difference between the CDFs of the data sets of the two classes and the higher the curve, the better the system. The point of maximum value is particularly important for the performance evaluation in credit risk analysis. Another important metric applied is the area under the curve which indicates the classes' separability throughout the whole score range.

6.2.2. ROC curves and minimum distance to optimum.

The Receiver Operating Characteristic Curve (ROC Curve) [11] is a widely used tool whose plot represents the compromise between the true positive and the false positive example classifications based on a continuous output along all its possible decision threshold values (the score). The closer the ROC curve is to the upper left corner (optimum point), the better the decision system is. The minimum distance of the curve to this point is an important metric. To assess the performance throughout the whole X-axis range, the metric is the area under the ROC curve (AUC) [9]. The bigger the area, the closer to optimum the decision system is.

If the ROC curve of a classifier stays above that of another along the entire domain of variation, the former system is better than the latter.

7. Results

The simulations were run according to the experimental setup described above for each of the two techniques, resulting in a pair scores for each of the 6,000 labeled examples of the 30 test sets, all statistically independent and in a number of 30 for guaranteeing the normality hypothesis to hold on the paired t-tests. Also, for each of the test sets, the four performance metrics were calculated as defined in Section-6: the maximum KS2, the area under the KS2 curve, the minimum ROC distance to the optimum point and the area under the ROC curve.

<table>
<thead>
<tr>
<th>Table 1. Summary results of the two models.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neural Networks</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>KS max</td>
</tr>
<tr>
<td>KS area</td>
</tr>
<tr>
<td>ROC min</td>
</tr>
<tr>
<td>ROC area</td>
</tr>
</tbody>
</table>

Table-1 presents a summary where the best results in performance are highlighted in boldface and the values marked by an asterisk (*) represent statistically significant differences. The paired t-tests for all the 4 metrics state that neural networks performed better than logistic regression at a 95% confidence level.

![Figure 1. Average KS2 curves.](#)
The KS2 plots (Figure 1) show that the multilayer perceptron performs better than logistic regression because its curve lies above that of logistic regression for all but a small part of the graph where the score range is not relevant for this thresholded decision making process since it affects less than 5% of the examples in either class.

The ROC curves (Figure 2) show that multilayer perceptron performs better than logistic regression for all their scores' range of variation.

![Figure 2. Average ROC curves.](image)

8. Concluding Remarks

This paper has presented a thorough performance comparison between logistic regression and neural networks (multilayer perceptron), carried out on a sample of 180,000 examples through a stratified 30-fold cross-validation process with 95% confidence intervals on the performance measures.

This research used several performance metrics to compare the techniques and the results show that there is a statistically significant difference favoring neural networks. Furthermore, the KS2 and ROC curves show that neural networks were superior to logistic regression throughout the whole score domain. The experimental results on large scale data sets, presented in this paper for a real-world application and those reported on artificially generated data [2], both agree that multilayer perceptron performs better than logistic regression, even considering the experimental setups less favorable for neural networks.

This difference can not be explained by the fact that the multilayer perceptron is a universal function approximator [5] while logistic regression is not, since this advantage holds only for multilayer perceptrons with an infinite number of units in the hidden layers. The most plausible explanation comes from the type of non-linearity of the solutions developed by each technique. While logistic regression applies non-linearity to each individual variable followed by a linear multivariate transformation, multilayer perceptrons apply a non-linear transformation in a truly multidimensional space. Research is analyzing this explanation through experimental verification on artificially generated data, applying Kiang’s approach [2].

This paper has presented relevant results in terms of statistical methodology and real-world large size of the data sets involved in the comparison of AI techniques. Future directions in this research are the comparison of other AI techniques based on very different mathematical functions with actual world data as Kiang [2] has done with artificially generated data and, then, the combination of complementary techniques through the method of stacking.

9. References


