Mathematical and Computational Applications, Vol. 18, No. 3, pp. 392-398, 2013

A COMPARATIVE STUDY OF ARTIFICIAL NEURAL NETWORKS AND LOGISTIC REGRESSION FOR CLASSIFICATION OF MARKETING CAMPAIGN RESULTS

Ali Aydın Koç and Özgür Yeniay

Department of Statistics, Hacettepe University, 06800, Beytepe, Ankara, Turkey
aaydin08@hacettepe.edu.tr and yeniay@hacettepe.edu.tr

Abstract- In this study, we focus on Artificial Neural Networks which are popularly used as universal non-linear inference models and Logistic Regression, which is a well known classification method in the field of statistical learning; there are many classification algorithms in the literature, though. We briefly introduce the techniques and discuss the advantages and disadvantages of these two methods through an application with real-world data set related with direct marketing campaigns of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit or not after campaigns.

Key Words- Artificial Neural Networks, Logistic Regression, Classification, Marketing

1. INTRODUCTION

In a classification problem, we try to determine to which class a data point belongs to, based on a labeled data set $D = [(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)]$. Here, n is the number of patterns (subjects) in the data set, x_is are data points, which are called features or input variables with known class memberships y_i. In the literature there are many well-known and mostly used classification methods, such as Logistic Regression (LR), Artificial Neural Networks (ANNs), Discriminant Analysis, Decision Trees, Support Vector Machines and Genetic Algorithms. Each one has its own purposes and capabilities.

Among all these classifiers, ANNs are being widely used in areas of prediction and classification, the areas where statistical methods have traditionally been used. Although ANNs originated in mathematical neurobiology, the rather simplified practical models currently in use have moved steadily towards the field of statistics. Since the last decade, ANNs are being used as an alternative to traditional statistical techniques and gain popularity in recent years. This has led to a number of studies comparing the traditional statistical techniques with neural networks in a variety of applications. A number of researchers have illustrated the connection of neural networks to traditional statistical methods. Gallinari *et al.* [1] study the relations between discriminant analysis and multilayer perceptrons used for classification problems. Richard and Lippmann [2] show that neural networks can provide estimates of Bayesian posterior probabilities. Cheng and Titterington [3] make a detailed analysis and comparison of various neural network models with traditional statistical methods. Sarle [4] translates neural network jargon into statistical jargon, and shows the relationships between neural networks and statistical models such as generalized linear models, maximum redundancy analysis, projection pursuit, and cluster analysis

Vach *et al.* [5] and Schumacher *et al.* [6] make a detailed comparison between feedforward neural networks and LR. The similarities and differences between the two methods are also analyzed [7]. Hervas and Martinez-Estudillo [8] combine logistic regression and neural network models to solve binary classification problems. The error function considered and the neural network framework used there was designed specifically for two-class classification problems.

In this study, we focus on ANNs and LR and we discuss the advantages and disadvantages of these two methods through an application with real-world data set.

2. ARTIFICIAL NEURAL NETWORKS

ANNs are computer models inspired by the structure of biologic neural networks and one of the developing machine learning techniques to solve the complex nonlinear systems in the real life. ANNs have been extensively used in many research areas from marketing to medicine. In most cases a neural network is an adaptive system that changes its structure during a learning phase. Neural networks can identify and learn correlated pattern between inputs and corresponding outputs. Fig. 1 depicts a simple example of an ANN.

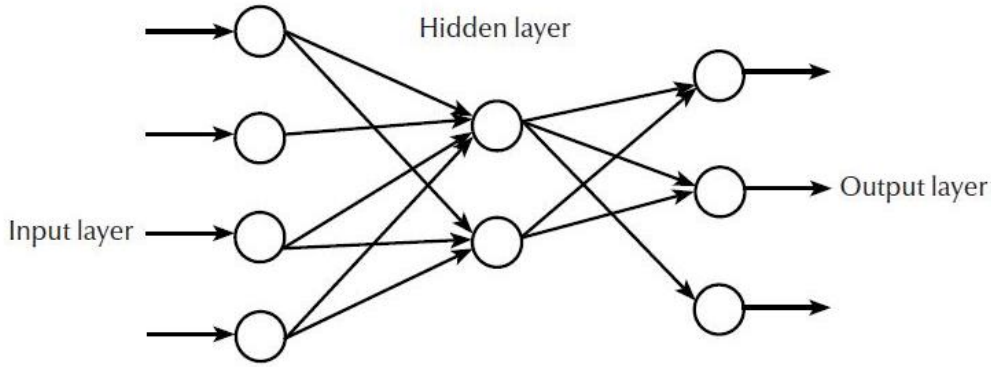

Figure 1. Typical multilayer perceptron model with 4 neurons in the input layer, 2 in the hidden layer and 3 in the output layer and with no direct connection from input to output layers

This typical multilayer perception model, constructed with layers of units, in the figure consists of 4 neurons in the input layer, 2 in the hidden layer and 3 in the output layer and with no direct connection from input to output layers. Since interconnections do not loop back or skip other neurons, the network is called feed-forward [9]. In such networks, there are two functions governing the behavior of a unit in a particular layer and affect the generalization of the model. One of them is input function and the other one is output function, which can be generally called activation function. The input function is normally given by Eq. (1),

$$y = x_1 w_1 + x_2 w_2 + \cdots + x_n w_n \quad + b(bians) \tag{1}$$

The activation function of a node can be defined as the output of that node given an input or set of inputs. A number of nonlinear functions have been used in the literature as activation functions. However, the most common one is sigmoid function,

just because it can show both linear and nonlinear property. Sigmoid function can be expressed as in Eq. (2),

$$f(x) = \frac{1}{1 + \exp(-x)} \tag{2}$$

Neural networks are becoming very popular with data mining practitioners, particularly in medical research, finance and marketing. This is because they present the main advantage of not being based on "*a priori*" assumptions and of allowing detection of links between factors that conventional statistical techniques such as LR may not be able to detect [10]. Comparing ANN models with standard statistical generalized linear models such as LR is an important step in the development procedure [11]. If the results show that the gain of using a non-linear model, such as the ANN, is limited, one should usually go for the less complicated model.

3. MULTIPLE LOGISTIC REGRESSIONS

LR, which is one of the multivariate statistical techniques, is a type of regression analysis, utilized for predicting the outcome of a categorical dependent variable based on one or more independent variables [12-13]. Generally, it calculates the class membership probabilities for the categories in the data set, by using the following equation,

$$p_i = \frac{1}{1 + \exp\left[-\left(\sum_{j=0}^{K} B_j X_j\right)\right]}, \qquad i = 1,2,\ldots p \tag{3}$$

Where $\sum_{j=0}^{K} B_j X_j = B_0 + B_1 X_1 + \cdots + B_K X_K$ is the logistic regression model, K is the number of independent variable, p is the number of categories in dependent variable and B_0 is the constant.

In this study we use two widely-used performance measures for logistic regression to answer the question "How well does my model fit the data?". The first one is The Hosmer–Lemeshow test (HL) which is a statistical test for goodness of fit. A high HL statistic is related to a small p-value and indicates a lack of fit. Therefore, the bigger p-value, the better is the fit, with a perfectly calibrated model having an p-value of greater than 0.05 (%95 confidence interval). The second one is The Nagelkerke R^2 which is a measure of predictive power, that is, how well you can predict the dependent variable based on the independent variables. It attempts to quantify the proportion of explained variance in the logistic regression model, similar to the R^2 in linear regression [14].

A maximum likelihood estimation procedure can be used to obtain the parameter estimates. In this study, SPSS Clementine is used to do the stepwise logistic regression, whose estimates are obtained by an iterative procedure [15].

4. AN APPLICATION ON MARKETING CAMPAIGNS OF BANKS IN PORTUGAL

The data used in this study is related with direct marketing campaigns of a Portuguese banking institution, from May 2008 to November 2010 [16]. The marketing campaigns were based on phone calls. The classification goal is to predict if the client will subscribe a term deposit or not after campaigns. Thus, the success of these campaigns can be evaluated. The detailed descriptions of input variables and the output variables can be seen in Table 1 and Table 2.

Table 1. Description of input variables

Input variables	Type of data
age	numeric
job : type of job	categorical: "admin.","unknown","unemployed","management", "housemaid","entrepreneur","student", "blue-collar", "self-employed", "retired", "technician", "services"
marital status	categorical: "married", "divorced", "single"; note: "divorced" means divorced or widowed)
education	categorical: "unknown", "secondary", "primary", "tertiary"
default: has credit in default?	binary: "yes", "no"
balance: average yearly balance, in euros	numeric
housing: has housing loan?	binary: "yes", "no"
loan: has personal loan?	binary: "yes", "no"
contact: contact communication type	categorical: "unknown", "telephone", "cellular"
day: last contact day of the month	numeric
month: last contact month of year	categorical: "jan", "feb", "mar", ..., "nov", "dec"
duration: last contact duration, in seconds	numeric
campaign: number of contacts performed during this campaign and for this client	client (numeric, includes last contact
pdays: number of days that passed by after the client was last contacted from a previous campaign	numeric, -1 means client was not previously contacted
previous: number of contacts performed before this campaign and for this client	numeric
poutcome: outcome of the previous marketing campaign	categorical: "unknown", "other", "failure", "success"

Table 2. Description of output variable

Output variable	Type of Data
y - has the client subscribed a term deposit?	binary: "yes","no"

We have disproportional class labels in response variable. The data set consists of a total number of 45211 clients, which 39911 of these clients give the

answer “yes”, and the rest, 5289 customers give the answer “no”. 45211 observations (clients) with 16 independent variables (inputs) are considered in this data set. The input variables consist of numeric, categorical and binary variables. The numeric data used as predictors in the model are the client’s age, the client’s average yearly balance (in Euros), last contact duration (in seconds), number of contacts performed during this campaign for the client, last contact day of the month and so on. The categorical inputs are the client’s job, the client’s marital status, the client’s education, last communication type (unknown/telephone/cellular) and last contact month of the year. As binary inputs, the client’s usage of housing loan and personal loan before the campaign, outcome of the previous marketing campaign and the client’s credit in default are chosen.

Before the analysis, as one can see, the data set is highly unbalanced. In order to overcome this problem, Data Balancing option in SPSS Clementine has been applied to the data set. Thus, we get 10013 patterns, equally for the two classes in the output variable. After then, we proceed data pre-processing, which is an important step in the data mining process. Data pre-processing includes cleaning the outliers with a proper method, normalizing and scaling the continuous features as well as feature extraction and selection.

Classification models in machine learning are evaluated for their performance by common performance measures [17]. In this study we use accuracy, sensitivity and specificity by using confusion matrix and then we plot Return On Investment (ROI) graph, which is the benefit (return) of the selection of the classification models is divided by the cost of the modeling; the result is expressed as a percentage or a ratio (see Fig. 2).

After all this procedure, as the result of the analysis, ANN classifies “yes” answers with the accuracy %87.8 and “no” answers with the accuracy %81.14. Logistic Regression classifies “yes” answers with the accuracy %82.8 and “no” answers with the accuracy %84.4 with the HL statistic p-value and the Nagelkerke R^2, 0.064 and 0.607, respectively. However, when we check overall accuracy, we see that ANN classifies %84.4 of the whole data correctly, while Logistic Regression classifies %83.63 of the whole data correctly.

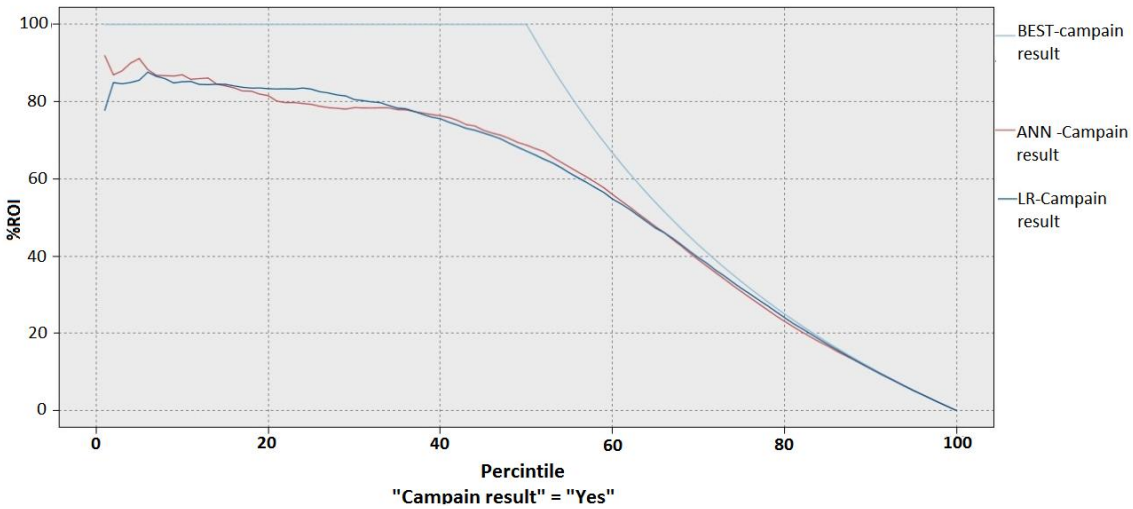

Figure 2. Return on Investment graph

One of the most important aspects of an algorithm is how fast it is. It is often easy to come up with an algorithm to solve a problem, but if the algorithm is too slow, it's not possible to use that classification algorithm. Thus, in this study, we also wanted to measure the runtime how long it would take for the algorithms to run if it were given the whole data. As a result, it is seen that LR takes 54 seconds and ANN takes 11 seconds. Thus, we can say that in order to minimize loss of time, using ANN is efficient.

6. CONCLUSION

In this study, we want to evaluate the success of the marketing campaigns of a Portuguese banking institution. Thus, we investigate the predictive accuracy of two mostly-used and well-known classification algorithms ANN and LR with 16 binary, numeric and categorical input variables. Our output is whether the client has subscribed a term deposit or not after these marketing campaigns. The results showed that these two algorithms achieved the identical overall accuracy. However, when we check the runtime for a big data like the one we have, ANN is much faster than LR. Thus, we can propose that with more data and higher-dimensional feature space, using ANN will be more efficient.

7. REFERENCES

1. P. Gallinari, S.Thiria, F. Badran, F. Fogelman-Soulie, On the relations between discriminant analysis and multilayer perceptrons, *Neural Networks* **4(3)**, 349–360,1991.
2. M.D. Richard, R.P. Lippmann, Neural network classifiers estimate Bayesian a posteriori probabilities, *Neural Computation* **3**, 461–483, 1991.
3. B. Cheng, D. M. Titterington, Neural networks: A review from a statistical perspective, *Statistical Science* **9(1)**, 2–30, 1994.
4. W. S. Sarle, Neural networks and statistical models, in: *Proceedings of the 19th Annual SAS Users Group International Conference*, Dallas, Texas, 1538–1550, 1994.
5. W. Vach, R. Robner, M. Schumacher, Neural networks and logistic regression: Part II. *Computational Statistics & Data Analysis* **21(6)**, 683–701,1996.
6. M. Schumacher, R. Robner, W. Vach, Neural networks and logistic regression: Part I. *Computational Statistics & Data Analysis* **21(6)**, 661–682, 1996.
7. M. Paliwal, U. A. Kumar, Neural networks and statistical techniques: A review of applications, *Expert Systems with Applications* **36**, 2–17, 2009.
8. C. Hervas, F. J. Martinez-Estudillo, Logistic regression using covariates obtained by product-unit neural network models, *Pattern Recognition* **40**, 52–64, 2007.
9. T. Ayer, J. Chhatwal, O. Alagoz, C.E. Kahn, R.W. Woods, E.S. Burnside, Comparison of logistic regression and artificial neural network models in breast cancer risk estimation, *Radio Graphics* **30**,13–22, 2010.
10. C. M. Bishop, Neural networks for pattern recognition, Oxford University press, New York, 2005.
11. S. Dreiseitl, L. Machado, Logistic regression and artificial neural network classification models: a methodology review, *Journal of Biomedical Informatics* **35**, 352–359, 2002.

12. D. Hosmer, S. Lemeshow, *Applied logistic regression*, Wiley, (2nd ed.), New York, 2000.
13. E. Steyerberg, F. Harrell, P. Goodman, Neural networks, logistic regression, and calibration, *Medical Decision Making* **18**, 349–350, 1998.
14. K. J. Ottenbacher, P. M. Smith, S. B. Illig, R. T. Linn, R. C. Fiedler, C. V. Granger, Comparison of logistic regression and neural networks to predict rehospitalization in patients with stroke, *Journal of Clinical Epidemiology* **54**, 1159–1165, 2001.
15. SPSS Clementine® 12.0 User's guide, SPSS Inc., Chicago, 2008.
16. S. Moro, R. Laureano, P. Cortez, Using data mining for bank direct marketing: an application of the CRISP-DM methodology. in P. Novais et al. (Eds.), *Proceedings of the European Simulation and Modelling Conference - ESM'2011*, 117-121, Guimaraes, Portugal, October, 2011. EUROSIS.
17. J. H. Song, S. S. Venkatesh, E. A. Conant, P. H. Arger, C. M. Sehgal, Comparative analysis of logistic regression and artificial neural network for computer-aided diagnosis of breast masses, *Academic Radiology* **12**, 4, 487-495, 2005.