Statistical Learning And Benchmarking: Credit Approval Using Artificial Neural Networks

Constantin Renato Ivanescu

Department of Computer Science, Faculty of Sciences, University of Craiova
ivanescurenato@yahoo.com

Abstract

Taking into consideration the world's economic situation, an important issue nowadays is predicting whether a client is suitable for a credit loan or card. Credit approval can be predicted using many machine learning methods. In this study we aim to investigate the potential contribution of three types of neural networks in business intelligence. The algorithms have been tested and applied on a publicly available dataset. An extensive statistical analysis has been performed to validate our experiment. The results show that NN should be used verifying whether a client can reimburse a certain credit loan.

Keywords: *machine learning, neural networks, statistics*.

1. Introduction

Nowadays an explosion of data and information surrounds us. The human brain cannot simply process this huge amount of information, thus the role of Artificial Intelligence (AI) and Machine Learning (ML) has increased exponentially. An important problem that needs to be solved fast and elegant is credit approval. Economic crises can begin if clients are granted loans that they cannot pay off. The use of AI in predicting the whether a client is suitable for a credit loan is on the rise.

Several methods have been used in the business intelligence, such as: neural networks (NN)[8], an evolutionary regressor selection in ARIMA model, [3], an optimized NN in [14],termite colony optimization [16], biometric technologies [17].

The goal of this paper is to investigate three types of NNs in predicting the suitability of granting a credit loan or card to a client: 1) a three-layer perceptron (MLP), 2) a radial basis function network (RBF), and 3) an extreme leaning machine algorithm (ELM). The statistical benchmarking has been done with the use of power analysis, normality tests, equality of variances, *t*-test for independent samples, comparing proportions, confidence intervals and cross-validations. The paper is organized as follows: in the next section, section 2 we briefly pass through the mentioned NNs and the statistical tool used for validating their results. In section 3 the dataset is described. Section 4 presents the results of the study together with the statistical analysis. Finally, the paper ends with the conclusions comments.

2. Methods

2.1 Neural Networks Algorithms

This study purpose is to establish a statistical benchmarking for four NN algorithms. In this section we will briefly describe the three NN methods.

In the last years, NNs have been used more and more to process information. NNs mimic how the human brain makes decisions. Using a large number of interconnecting processing units, *neurons*, NNs can process data that is too complex to be detected by humans.

One of the most frequently used types of NN is the *multi-layer perceptron* (MLP). There are two types of MLP, defined by the number of hidden layers: 3-MLP and 4-MLP. Due to Kolmogorov's theorem, which states that a MLP with only two-hidden layers is theoretically sufficient to model any real-life problem, in this study we shall consider only the 3-MLP algorithm. The MLP uses hyper-planes to divide the problem space. The activation function of a MLP is the non-linear activation of the scalar product obtained by multiplication of the input and synaptic weights. MLP have been used is various domains, from business intelligence [8] to medicine [10], [11]

Another way to differentiate between classes is to divide the problem space using hyper-spheres, instead of hyper-planes. Centers and radii characterize the hyper-spheres. The NNs that uses this approach are the *radial basis function* (RBF). In this case, the distance between the input and a prototype vector computes the activation of a hidden unit. The architecture of a RBF contains only a hidden layer, thus making the training of the network very fast. The speed of the training is an advantage, but it is contra balanced by the fact that RBFs are very sensitive to the curse of dimensionality. RBF have been successfully applied in economics [15], medicine [9], [12]

Few years ago, a debate concerning the *extreme learning machine*(ELM) NN rose in the academic community. ELM is a single hidden feed-forward NN that has a very fast learning algorithm [4]. The learning process is based on the random choice of hidden nodes and the analytical computation of the output weights. ELM is used in different domains due to its robustness, fast learning speed and it's ability to generalize, [2], [5], [6].

Briefly, the whole concept behind the "black-box" inside an NN is that by training it, the NN is trying to associate the output with the input patterns. In order to do that, the datasets are divided into two subsets: *training set* and *testing set*. During the training phase, the training subset is used to establish the hyperparameters of the network that define each model particularly. After the hyperparameters are set, the NN model is tested on the testing subset.

2.2. Statistical Benchmarking

In this study, we are going to conduct the statistical benchmarking using the following techniques:

2.2.1 Power Analysis

As stated before, the goal of this paper is to assess and compare the diagnosing performances of the three NN algorithms. Thus, an *a priori* statistical power analysis needs to be performed to determine which is the most suitable sample size in order to achieve adequate power [1].

2.2.2 Normality Tests

The *Kolomogorov-Smirnovone-sample test* for normality is used in benchmarking. The test is based on the maximum difference between the hypothesised cumulative distribution and the sample cumulative distribution. The hypothesis that the distribution is normal is rejected, if the corresponding *D* statistic is significant. The probability values that are stated are valid, if the mean and standard deviation of the normal distribution are known *a priori*, and not computed from the data. If that is not case, the *Lilliefors* significance level should be utilized instead.

Another method for testing the normality is the *Shapiro-WilkW* test. *W* is the correlation between the data and its corresponding normal scores. The hypothesis that the distribution is normal is rejected if the *W* statistic is significant.

The good news is that even if the results reported after using these tests is not satisfactory, we can always apply the *Central Limit Theorem* [1], which states that the distribution of an average tends to be normal as the sample size increases, regardless of the distribution from which the average is taken. Thus is the sample size is big enough, the distribution is 'nearly' normal.

2.2.3. Equality of Variances

Besides the normality tests, one must verify the equality of variances. In this study we have used the *F-test*. The *F-test* is based on the statement that under the null hypothesis, two normal distributed samples have equal variances, thus the ratio of the variances has an *F* distribution [1].

2.2.4. T-test For Independent Samples

The *t-test* for independent samples is based on Student's *t* distribution. Having independent samples, we are concerned about the mean difference between the two samples and the variability between the observations. We cannot use the *t-test* if the variables are not normally distributed. This is why we need to perform the normality tests before, or simple use the result of the Central Limit Theorem.

2.2.5 Comparison of the Two Population Proportions

We are interested in testing the differences between two population proportions, or percentages. For resolving this issue we shall use the *z*-test statistic. The *p*-value is computed using the *z*-value for the respective comparison [13].

2.2.6. Confidence interval

The confidence interval is the range of values for the true mean, given with a probability P. Mathematically speaking, we can compute the 95% confidence interval by using the following formula (mean-$1.96 \times SE$, (mean+$1.96 \times SE$), where the SE is the standard error of the sample mean. According to [1], the true population mean is included in the 95% confidence interval with a probability of 95%.

2.2.7 Cross-validation

The simplest way to compare models is to use the n-fold cross-validation technique. The technique may come in handy when there is no test subset available and the whole dataset is too small. Practically, the dataset is randomly divided into *N* subsets of equal sizes and run the method *N* times. Each time we leave out one of the *N* subsets as testing set, the rest of (*N*-1) subsets being the training sample. It has been proven, empirically, that if we use the 10-fold cross-validation we achieve good results [7].

3. Datasets

The dataset contains credit cards applications. We cannot refer to specific attribute names, due to the fact that they have been changed in order to protect data confidentiality.

The dataset is a combo of continuous and nominal attributes. Even the nominal attributes have random values for the same reason as the attribute's names. The data is composed of 690 instances and 15 predictive attributes. The classes are encoded as well with (+) 307 cases (44.5%) and (-) 383 cases (55.5%). The dataset can be found at the UCI Repository, at the following URL: https://archive.ics.uci.edu/ml/datasets/Credit+Approval.

4. Results

In order to compare the prediction performances of the three ML algorithms, we performed an *a priori* statistical power analysis test to determine the appropriate sample size in order to achieve adequate power. Thus, we have considered a sample of 100 different computer runs for each ML method, which provided 99% with type I error $\alpha = 0.05$. Having 103 different computer runs, and applying the Central Limit Theorem, we can conclude that the normality presumption is achieved. Basically, we have run each NN method 103 times in a complete 10-fold cross-validation cycle. After each run, the corresponding training and testing accuracy has been recorded.

The NNs' performances have been evaluated by computing: 1) the *training performance*, the proportion of cases that were classified correctly in the training phase and 2) the *testing performances*, the proportion of cases that were classified correctly in the testing phase.

Taking the evaluation even further we have computed an overall and per-class classification statistics, establishing the number of correctly and incorrectly classified cases, and the number of cases that actually belonged to each class, and classified as belonging to the other class. The number of hidden neurons that were used in the architecturesis also presented along with the mean, standard deviation (Std.), the 95% confidence interval and the AUC score.All the results are in the Table 1.

We can depict from Table 1, that the number of processing units is 5 for the MLP, 30 for the RBF and 6 for the ELM. This means that we can achieve good accuracies even if the architecture is simple, around 85% for MLP and ELM. We can see that the classification performance for the MLP and ELM is excellent (AUC > 0.9).

Table 1. NN's performances in terms of average, standard deviation, 95% confidence interval and AUC score over 103 runs.

NN type	Mean/Std no hidden neurons	Mean/Std training performance(%)	Mean/Std testing performance (%)	95% Confidence Interval testing	AUC
MLP	5/3	88.22 / 4.97	84.46 / 3.34	(77.78, 91.14)	0.93
RBF	30/10	68.80 / 4.95	69.90 / 3.30	(63.30, 76.50)	0.80
ELM	6	87.56 / 5.02	85.20 / 4.03	(77.14, 93.26)	0.92

From Fig.1 we can see the performance of all classifiers in terms of the AUROC. We can see that the MLP and ELM perform the same, whereas the RBF performs poorly.

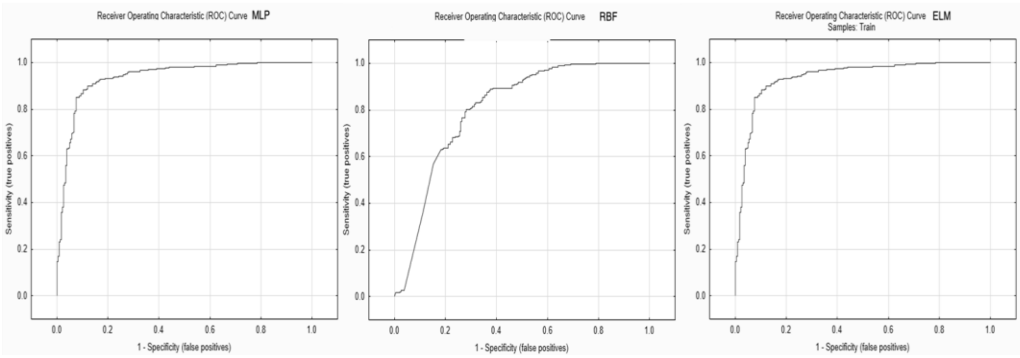

Fig. 1. Comparing ROC curves for each NN classifier

Table 2. Overall and per-class statistics MLP

RBF	**Class (-)**	**Class (+)**	**All**
Total	261	223	484
Correct	224	203	427
Incorrect	37	20	57
Correct (%)	85.82	91.03	88.23
Incorrect (%)	14.17	8.96	11.77

Table 3. Overall and per-class statistics RBF

RBF	**Class (-)**	**Class (+)**	**All**
Total	261	223	484
Correct	259	74	333
Incorrect	2	149	151
Correct (%)	99.23	33.18	68.80
Incorrect (%)	0.76	66.81	31.19

Table 4. Overall and per-class statistics ELM

RBF	**Class (-)**	**Class (+)**	**All**
Total	261	223	484
Correct	220	200	420
Incorrect	41	23	64
Correct (%)	84.29	89.68	86.77
Incorrect (%)	15,70	10.31	13.22

From Table 2, 3 and 4, we can see that the MLP and RBF classify correctly cases from both classes, around 84% for the (-) class and around 90% for the (+), whereas the RBF classifies correctly 99% cases from the (-) class and only 33% cases from the (+) class. It is important to mention that all the NNs perform good despite the mixt nature of the attributes.

For each NN method, for the testing phase we have performed the *Kolmogorov-Smirnov* and *Shapiro-Wilk*W tests. The results are presented in Table 5. Regardless of the used method, the distributions are not normal. Still, the sample size is 103, over 100, thus we can presume that the sample is nearly Gaussian.

Table 5. Testing normality of NN's diagnostic testing performances

Variable	Kolmogorov-Smirnov		Shapiro Wilk	
	K-S max D	Lillifors p	S-W W	p-value
MLP	0.149	0.01	0.968	0.000
RBF	0.110	0.01	0.951	0.007
ELM	0.134	0.01	0956	0.000

The equality of variances has been tested using the classical *F*-test. We can see from Table 6, that the MLP and ELM have equal variances, whereas MPL and RBF and ELM and RBF have different variances (p <0.05).

Table 6. Testing equality of variances

Variables	Equality of variance
	F-test/ p-level
MLP (testing) vs. RBF (testing)	4.45 / 0.00
MLP (testing) vs. ELM (testing)	1.00 / 0.94
RBF (testing) vs. ELM (testing)	4.44 / 0.00

Using the *t*-test for independent variables, we are interested in the mean difference between the testing performances of the NN models, taking into account the variability of over 100 computer runs. Because the sample size is greater than 100, we can perform the *t*-test. From Table 7, we can see that there are significant difference between MLP and RBF, ELM and RBF (*p*-level < 0.05), whereas between the MLP and ELM there are no significant differences.

Table 7. Comparing testing performances (*t*-test)

Variables	Comparison test
	t-test/ p-level
MLP vs. RBF	5.45 / 0.000
MLP vs. ELM	-11.20 / 0.5
RBF vs. ELM	5.98 / 0.000

5. Conclusions

In this paper, we demonstrated that the use of NNs in predicting the suitability of a client to be granted a credit loan is useful. We noticed that simple architectures of MLP and ELM perform very good, even excellent, AUC > 0.9, whereas a RBF does not.

The statistical benchmarking showed no significant differences between the MLP and ELM, whereas significant difference have been spotted between MLP and RBF, ELM and RBF.

References

[1]. Altman, D.G., 1991. Practical statistics for medical research. New york: Chapman and Hall.

[2]. E. Cambria , G. -B. Huang, L. Lekamalage, et al., "Extreme Learning Machines [Trends & Controversies],"IEEE Intelligent Systems, vol. 28, no. 6, pp. 30-59, 2013.

[3]. R. Stoean, C. Stoean, A. Sandita, Evolutionary Regressor Selection in ARIMA Model for Stock Price Time Series Forecasting. In: Czarnowski I., Howlett R., Jain L. (eds) Intelligent Decision Technologies 2017. IDT 2017. Smart Innovation, Systems and Technologies 73 (2018), Springer, Cham.

[4]. G. -B. Huang, L. Chen and C. -K. Siew, “Universal Approximation Using Incremental Constructive Feedforward Networks with Random Hidden Nodes,” IEEE Trans. Neural Networks, vol. 17, no. 4, pp. 879-892, 2006.

[5]. Extreme Learning Machine 2013: Algorithms and Applications. Adaptation, Learning, and Optimization 16, F. Sun, K. -A. Toh, M. Romay and K. Mao, Eds., Springer, 2014.

[6]. F. Gorunescu, S. Belciug, , Boosting backpropagation algorithm by stimulus-sampling: Application in computer-aided medical diagnosis, Journal of Biomedical Informatics, vol. 63, pp. 74-81, 2016

[7]. R. Kohavi, 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. Fourteenth international joint conference on artificial intelligence. 2(12), 1137–1143.

[8]. S. Belciug, A.Sandita, Business Intelligence: Statistics in predicting stock market, Annals of the University of Craiova, Mathematics and Computer Science, Science Series, Vol. 44, No. 2, pp. 292-298, 2017.

[9]. F.Gorunescu, M. Gorunescu, E. El-Darzi, S.Belciug, A statistical framework for evaluating neural networks to predict recurrent events in breast cancer, International Journal of General Systems, Vol. 39, No. 5, pp. 471-488, 2010.

[10]. S.Belciug, M.Lupsor, A multi-layer based procedure for detecting liver fibrosis, Annals of the University of Craiova, Mathematics and Computer Science Series, Vol. 36, No. 1, pp. 64-70, 2009.

[11]. S.Belciug, M.Lupsor, R.Badea, Features selection approach for non-invasive evaluation of liver fibrosis, Annals of the University of Craiova, Mathematics and Computer Science Series, Vol. 35, pp. 15 - 28, 2008.

[12]. F. Gorunescu, M. Gorunescu, E. El-Darzi, S.Belciug (Gorunescu), A statistical evaluation of neural computing approaches to predict recurrent events in breast cancer, Proceedings 4th International IEEE, Conference on Intelligent Systems - IS08, Varna, Bulgaria 6-8.09.2008, pp. 38-43, 2008.

[13]. R. Schumacker, A. Akers., 2001. Understanding statistical concepts using S-plus. Mahwah, NJ:Lawrence Erlbaum Associates, Inc.

[14]. M. Qiu, Y. Song, Predicting the direction of Stock market index movement using an optimized artificial neural network model, https://doi.org/10.1371/journal.**pone**.0155133, Plos one (2016).

[15]. Nekoukar, V., Beheshti, MTH, A local linear radial basis function neural network for financial time-series forecasting, Applied Intelligence, 33 (3) 352-356, 2010.

[16]. Radwa Ali, Neveen I. Ghali, Taghrid A. Imam, A Comparative Study for Optimizing Retail Inventory Market Using Termite Colony Optimization, Egyptian Computer Science Journal Vol. 38 No. 1 January 2014

[17]. Silvia Parusheva, A comparative study on the application of biometric technologies for authentication in online banking, Egyptian Computer Science Journal Vol. 39 No. 4 September 2015.