Journal of Business Economics and Management
ISSN 1611-1699 / eISSN 2029-4433
2017 Volume 18(2): 224–240
doi:10.3846/16111699.2017.1280844

MODELING CREDIT APPROVAL DATA WITH NEURAL NETWORKS: AN EXPERIMENTAL INVESTIGATION AND OPTIMIZATION*

Chi GUOTAI[1], Mohammad Zoynul ABEDIN[2, 3], Fahmida–E–MOULA[4]

[1, 2, 4]Faculty of Management and Economics, Dalian University of Technology, Dalian 116024, China
[3]Department of Finance and Banking, Hajee Mohammad Danesh Science and Technology University, Dinajpur- 5200, Bangladesh
E-mails: [1]chigt@dlut.edu.cn; [2]abedin@mail.dlut.edu.cn (corresponding author); [4]fahmidamz@yahoo.com

Received 27 March 2016; accepted 08 January 2017

Abstract. This study proposes an investigation and optimization of Multi-Layer Perceptron (MLP) based artificial neural networks (ANN) credit prediction model, combine with the effect of different ratios of training to testing instances over five real-world credit databases. As an outcome from the alteration procedure, three different types of hidden units [K = 9 (ANN–1), K = 10 (ANN–2), K = 23 (ANN–3)] are chosen through the pilot experiments and execute, therefore, 45 (5×3×3) unique neural models. Experimental results indicate that "the neural architecture with ten hidden units" is proposed as an optimal approach to classifying the credit information. With these contributions, therefore, we complement previous evidence and modernize the methods of credit prediction modeling. This study, however, has realistic implications for bank managers and other stakeholders to delineate the risk profile of the credit customers.

Keywords: credit prediction, neural networks, Multi-Layer Perceptron, hidden neurons, alteration experiments, investigation and optimization.

JEL Classification: G21, C45, C51, C52.

Online supplementary material: Supporting information for this paper is available as online supplementary material at https://doi.org/10.3846/16111699.2017.1280844

Introduction

Credit prediction is a key application in statistical modeling and plays an important function in contemporary financial risk management practice. It gives to the key element in credit approval process, which is to precisely and effectively quantify the degree of uncertainty associated with a creditor. The degree of the credit risk of a creditor is connected with the probability of default, i.e. an event of not paying back the approved loan

* Authors contributions are equal.

over a given period. The credit prediction classifier task is however to make partition between the ones who do default and the ones who do not, i.e. between good and bad loan clients in terms of their creditworthiness. Discriminatory power is the main symbol of classifier optimality, and hence the higher the discrimination ability, the more feasible the credit prediction model will be.

Multi-Layer Perceptron (MLP), a stylish credit prediction model comes out as an important alternative, among all neural networks (NN) available, and draws attention from numerous modelers with its high forecast accuracy, from the last two decades. NN depends primarily on mathematical transferring the operations of the human brain to the computer systems. Contrasting with statistical methods, NN doesn't need prior assumptions, can generalize, can approximate continuous function, can properly infer the hidden part of a sample, and in study about credit approval, for many years, authors supported the supremacy of MLP based NN model over lots of optimization and statistical methods (Khashei *et al.* 2012).

Recently, Zhao *et al.* (2015) demonstrated that the classification performance of MLP based NN methods could significantly improve by changing the ratio of sample composite mixture (SCM) of training and testing instances, the number of hidden neurons, and the training iterations. It also depends on chosen real world databases for training and validating the trained neural model, Khashman (2011) added. If careful attention is paid to the earlier studies, however, one can notice the lack of original databases; i.e., their database for experiential investigation was to some extent imperfect; fixed SCM ratios, defective selection of hidden neurons in NN models that can hinder its performance. For example, Lee (2007), Min and Lee (2005), Kim and Ahn (2012) and Shin *et al.* (2005) applied MLP and SVM to Korean credit prediction and bankruptcy prediction and drew the conclusion that the best classification accuracy was found for MLP when the number of hidden nodes was 10; MLP (88.16%) was better than that of SVM (88.01%); MLP once more, was better approach as opposed to ordinary statistical approaches; while SVM was better to learn a small size of data patterns in the 1st, 2nd, 3rd and 4th study, respectively but they use an imbalance training (80%) and testing (20%) sets, far smaller datasets with fewer features. However, Li *et al.* (2006) and Li *et al.* (2016) found that MLP–trained model obtained satisfactory accuracy rate for consumer credit and SME credit databases, respectively but their results were also based on a small sample size with imbalance training–testing ratio. In contrast with the above work, Khashman (2010) trained three MLP neural networks on German credit dataset based on nine learning schemes with different training to testing ratios and different number of hidden neurons. That study concluded that the learning scheme with 40% training and 60% testing dataset and 23 hidden neurons performed best. More recently, however, adjacent to the above study, Zhao *et al.* (2015) concluded that models with nine hidden units performed best out of 34 MLP neural network models on the similar database. In the study of Khashman (2009), seven learning schemes with different training to testing dataset was investigated on Australian credit and concluded that neural model with and 43.5% training to 56.5% testing set with 9 hidden neurons performed best, on the other hand, an emotional neural network outperformed the conventional neural network in the study of Khashman (2011) on the same training to testing ratio with ten hidden neurons

on an identical dataset. Similarly, in Jeong *et al.* (2012), the tuned NN model with four hidden units outperformed the non–tuned NN model for a Korean bankruptcy database.

One major constraint of existing studies is that the most relevant studies (Khashman 2009, 2010, 2011; Zhao *et al.* 2015) simply use one dataset, small number of sample sizes and with fewer features for system validation than would be used by a financial institution. Another problem in most cases when using neural networks is the use of either balance or imbalance training to testing ratio. Furthermore, no experimental investigation is followed, except a few studies, to select optimum number hidden units. Therefore, for the competitive performance of NN model, versatile databases with different ratios of SCM in training-testing examples, an optimum selection of hidden neurons during the model construction phase must be cautiously refined.

In these contexts, this study proposes an investigation and optimization of MLP based NN credit prediction model, combine with the effect of different ratios of training to testing datasets. Therefore, we use three different types of balance/imbalance mixtures of training and testing instances, 40%:60%, 50%:50% and 90%:10%, respectively to determine the most optimal one. The training data is utilized to train the network while the test data is used to validate the network's performance upon completion of training. In addition to the above aspects, the number of hidden neurons (K) can have a large impact on the performance of the network architecture. The optimal number of hidden neurons, however, was selected after several trials involving the alterations of the number of hidden neurons from one to fifty neurons, with maintaining the following criteria: it should have the lowest root mean square error (RMSE), largest percentage of overall accuracy rate and the lowest type II error. As an outcome from the alteration procedure, three different types of hidden units [K = 9 (ANN–1), K = 10 (ANN–2), K = 23 (ANN–3)] are chosen through the pilot experiments (see, e.g., Supplementary information Tables A1–A5). In fact, we compare 45 (5×3×3) unique neural models over the five databases with different number of hidden neurons; on different SCM ratios, to get the model with the best accuracy and effectiveness. Experimental results indicate that ANN–2, the neural architecture with ten hidden units, is proposed as an optimal approach to classify the credit information. With these contributions, therefore, we complement previous evidence and modernize the methods of credit prediction modeling.

1. Experimental design

1.1. Real-world credit database

We focus on five real-world credit datasets, e.g., the Australian, German and Japanese are from UCI machine learning database repository (Lichman 2013), and have been extensively used as a benchmark in many prediction models. The Chinese credit, a project dataset, provided by a Chinese commercial bank, while SPSS credit modeling dataset is from Vukovic *et al.* (2012). The datasets comprise example of non–default and default creditors with a binary target variable, illustrated by a set of risk drivers which capture information from the creditor application form. A summary of the five datasets is presented in Table 1.

Table 1. Description of databases used in the experiment

	Total cases	Non–default /default cases	No. of attributes
Australian credit	690	307/383	14
German credit	1000	700/300	20
Japanese credit	690	307/383	15
Chinese credit	3111	3040/71	81
SPSS credit	700	517/183	9

1.2. Neural network architecture

We used popular neural network architecture, namely Multi-Layer Perceptron (MLP) where all neurons and data flow assembled through hidden units in a feedforward manner. The opening layer is called the input layer which is composed of pieces of credit information from the external environment. In this study, different input neurons used for the different databases, e.g., for Australian credit, the NN input layer has 14 neurons, according to the number of credit applicant's features in the database. The final layer is called the output layer where the network produces the target output, *Y*, by defining a credit customer either non-default or default. Any layers between these two are called hidden layers those have no contact with external credit information and can only receive responses from the connected layers. Therefore, the final decision can be gained by evaluating target output, *Y*, with a threshold, generally set at 0.5, thereby reaching a decision of non-default if $Y > 0.5$; otherwise it will be classified as a default. For reducing the computational complexity, single hidden layer containing *K* neurons used in this study for all networks. Figure 1 depicts the architecture of ANN classifier.

The number of neurons, *K*, in hidden layer was first chosen as 9 (ANN–1), then was changed to 10 (ANN–2) and 23 (ANN–3) during subsequent experiments for comparison of system performance which have a large impact on the performance of the network architecture. The huge number of iterations needed in the training phase for

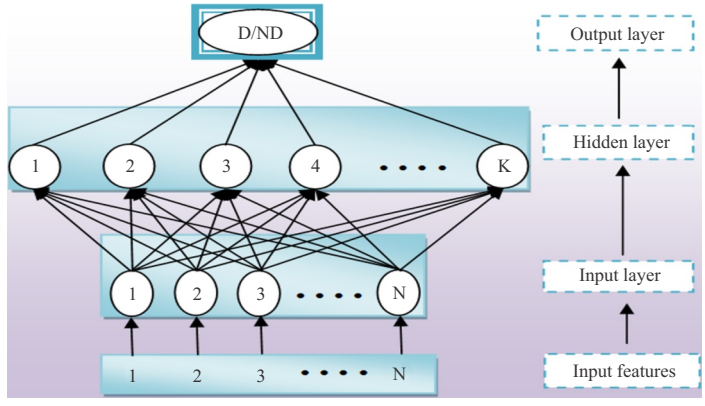

Fig. 1. Illustration of feed – forward MLP neural network

the huge number of hidden nodes. It is vital not to over-fit the network with a huge number of hidden units than needed until it can memorize the training set. As a result, different rules of thumb are proposed in the literature for selecting the optimal number of hidden nodes, but their optimality is not ensured. For instance, Salchenberger *et al.* (1992) suggested that the number of hidden units should be 75% of the number of features. Tang and Chi (2005) advocated that number of hidden units in a three layer MLP network should be $m/2$, $2m/3$, m, $m+1$, and $2m+1$where m is the number of features in the respective database. Moreover, a random selection of hidden layer neurons was followed in Tsai and Wu (2008). Later, the more popular method was cited by Jeong *et al.* (2012) who quoted that hidden units of MLP architecture optimized by means of a Cross-Validation (CV) procedure. But some evidence (Tommi 2014; Arlot, Celisse 2010; Barrow, Crone 2016) advised that in practice CV undergoes from two major shortcomings. The first shortcoming is that when it is employed to choose between two/more networks the approximate network accuracy of CV tends to be higher than the accurate network accuracy, and this propensity becomes more pronounced as the number of networks tested raises. The second and associated difficulty is that, usually, the more networks that are tested, the higher the probability that CV will unsuccessful to choose the best available neural architecture.

However, an alteration experiment was used in the most relevant studies (Khashman 2009, 2010, 2011; Zhao *et al.* 2015) of this work to determine the number of neurons in the hidden layer. In addition, this procedure was also used in other application area, e.g., project selection in portfolio management (Costantino 2015), for nonlinear time series forecasting (Zhong, Enke 2017), for predicting soil distribution (Falamaki 2013). Considering their successful experimentation, the optimal number of hidden neurons therefore, was selected after several trials involving the alterations of the number of hidden neurons from one to 50, maintaining the following criteria: it should have the lowest RMSE, largest percentage of overall accuracy and the lowest type II error (see, e.g., Supplementary information Tables A1–A5). The learning rate is set to 0.4, and the momentum term is to 0.9, while the learning of NN stopped when predetermined number of epochs was reached. It uses the gradient decent method to control the speed of training. The activation function in the hidden layer and the output layer is the hyperbolic tangent.

1.3. Training and testing sub-sets

No rules of thumb are suggested in the literature for designing training to testing subset ratios (Khashman 2010). Therefore, in the current experiments, we use three different types of balance/imbalance mixtures of training and testing instances, 40%:60%, 50%:50% and 90%:10%, respectively to determine the most optimal one. First two ratios were chosen as their closer/equal to 50%:50%, and for the third case, although inclusion of 80%:20% could be sensible, but it is assumed that the training accuracy of the model increased by increasing the training examples, supporting that the adoption of 90%:10% is more sensible thereby enhancing the network's predictability (Zhao *et al.* 2015). Though it is reasonable to adopt the mixture of all possibilities, i.e. the mixture of

10%:90%, 20%:80%, 30%:70%, and so on, until the last mixture of 90%:10%. But due to space constraint, it is less possible to adopt all possible mixtures over five different databases. However, the training data is utilized to train the network while the test data is used to validate the network's performance upon completion of training.

1.4. Performance evaluation

In order to evaluate the NN-based credit evaluation system, three standard measures are used which are originated from a 2 × 2 confusion matrix as that given in Table 2, where *tp* are true positive, *fp* false positive, *fn* false negative, and *tn* true negative counts. The evaluation measures are described respectively as following:

$$Accuracy = (tp + tn) / (tp + fn + tn + fp); \quad (1)$$

$$Type - I\ error = fn / (tp + fn); \quad (2)$$

$$Type - II\ error = fp / (tn + fp). \quad (3)$$

Table 2. The confusion matrix for classification problem

		Predicted observations	
		Predicted positive	Predicted negative
Actual observations	Actual positive	*tp*	*fp*
	Actual negative	*fn*	*tn*

1.5. Cost of credit prediction errors

We summarize the costs of credit prediction errors, type I and type II errors, and their impact on classifier selection. Several evidence (Abellán, Castellano 2017; Lee, Chen 2005; West 2000) advise that adding these costs into the prediction models can guide to better and more precise results. It is marked that the costs related to type I errors (a creditor being non-default is misclassified as default) and type II errors (a creditor being default is misclassified as non-default) are notably different. Usually, the misclassification costs related to type II errors, P_{12} are much higher and more detrimental than those related to type I errors, P_{21}. It is vital, in this aspect, to assess the credit investigation neural network algorithms with their associated cost, described as following, rather than relying on the overall accuracy.

$$Cost = P_{12} * \pi_2 * (s_2/S_2) + P_{21} * \pi_1 * (s_1/S_1). \quad (4)$$

So as to determine the cost function of the credit prediction models, the ratio of misclassification (MC) costs proposed by Dr. Hofmann, associated with type II and type I, is 5:1 (West 2000). The stress is not only on this relative cost ratio at 5:1, but also it offers a sensitivity analysis using higher cost ratios at e.g. 7:1, 10:1, 12:1, 15:1, respectively and in current study, therefore, we consider five different levels of MC cost for each database. For the turmoil financial situation, particularly, it is expected that the higher

cost ratio might be more suitable and Kao *et al.* (2012), however, suggested that the relative cost ratio can range from 5:1 to 20:1. Determination of the cost function also requires an estimation of the prior probabilities of non-default credit, π_1 and default credit, π_2 in the applicant pool of the credit prediction model. These prior probabilities are estimated from the actual ratios of non-default and default credit in the empirical databases. The ratios s_2/S_2 and s_1/S_1 in Equation (4) compute the probability of making type II errors and type I errors, respectively.

2. Results and discussion

2.1. Model prediction

We use three different types of hidden units 9 (ANN–1), 10 (ANN–2), 23 (ANN–3), those are picked through pilot studies and execute, therefore, 45 (5×3×3) unique neural models. The effects of the number of hidden units on accuracy rate and error rate, the different levels of MC cost ratios over the five credit databases with three SCM ratios are summarized in Tables 3–7.

From the experimental results shown in Table 3, for the Australian credit, we can find that ANN–2 model, under 50%:50% SCMR has the highest averages of overall credit prediction rate of 94.25%, achieving 94.88% training and 90.63% testing accuracies whereas ANN–1, under 90%:10% SCMR, closely following model, yields an overall accuracy of 93.10% with 93.33% training while 86.96% testing rates, attaining the smallest RMS error of 13.41%. ANN–2 model, under 90%:10% SCMR however, gets the highest accuracy of 94.12% by considering the testing dataset but it is under an imbalance SCMR.

Table 3. Performance obtained by the Australian database

SCM Ratio (%)	Neural network	Accuracy (%)			RMS error	Error (%)		Expected misclassification cost (%)				
		Tr–dataset	Te–dataset*	Overall		Type – I	Type – II	5:1	7:1	10:1	12:1	15:1
40:60	ANN – 1	91.29	80.00	89.34	0.1468	8.96	13.01	4.79	4.84	4.87	4.88	4.90
	ANN – 2	92.20	84.21	90.86	0.1426	8.87	9.56	5.42	5.40	5.39	5.38	5.38
	ANN – 3	88.32	82.81	87.32	0.1520	7.81	19.33	9.59	9.90	10.16	10.24	10.38
50:50	ANN – 1	86.77	87.01	86.81	0.1519	10.73	16.30	8.39	8.58	8.73	8.78	8.87
	ANN – 2	94.88	90.63	94.25	0.1352	7.22	3.49	2.16	2.11	2.06	2.05	2.03
	ANN – 3	92.31	83.53	90.60	0.1459	7.60	11.83	6.08	6.21	6.33	6.36	6.43
90:10	ANN – 1	93.33	86.96	93.10	0.1341	6.46	7.45	3.95	4.00	4.05	4.06	4.09
	ANN – 2	89.45	94.12	89.57	0.1429	7.96	13.33	6.80	6.97	7.11	7.15	7.23
	ANN – 3	89.89	78.26	89.47	0.1441	7.83	13.75	6.99	7.17	7.32	7.36	7.44

Note: * Tr–dataset and Te–dataset refer to training and testing datasets, respectively.

For the German credit, as the experimental results show in Table 4, to sum up, ANN–2 classifier, under 90%:10% SCMR generate the best results with an overall accuracy of 79.00% with 78.85% training while 80.46% testing accuracies. In contrast, ANN–3 classifier in 90%:10% SCMR yields the highest accuracy considering test dataset alone. The minimum RMSE error of 12.40% then comes from ANN–3 classifier, under 50%:50% SCMR. For the same dataset, in contrast to Zhao *et al.* (2015), in the setting of 70%:30% approved/rejected instances with nine hidden units; Zhao's model achieved 87% accuracy. There are two possible reasons: the database designed by a novel data distribution method, namely, "Average Random Choosing", the first reason; and their models trained with training, validation, and test datasets, the second reason. However, as indicated in Table 5, the prediction results for the Japanese credit display some of the similar patterns discussed for the German credit. ANN–3 classifier in 90%:10% SCMR outperforms the other neural models with regard to test dataset accuracy. ANN–3 classifier, under 40%:60% SCMR conversely, show the best classification capability in terms of overall accuracy of 95.28% with 97.09% training while 87.50% testing accuracies including the minimum RMSE error of 14.05%.

For the Chinese credit, the experiential results presented in Table 6 reveal that ANN–2 credit classifier in 90%:10% SCMR produces the best results, with the highest overall accuracy of 78.83% through 80.77% training while 58.01% testing accuracies together with a minimum RMSE error of 15.39%. ANN–3 classifier in 90%:10% SCMR, on the contrary, acquire the maximum ability to predict the default creditors pertaining to test dataset accuracy of 76.19%. However, there is a remarkable accuracy rate gap in between of training, and testing examples as the database is highly unbalanced.

Table 4. Performance obtained by the German database

SCM Ratio (%)	Neural network	Accuracy (%)			RMS error	Error (%)		Expected misclassification cost (%)				
		Tr–dataset	Te–dataset*	Overall		Type – I	Type – II	5:1	7:1	10:1	12:1	15:1
40:60	ANN – 1	77.32	74.44	75.65	0.1384	22.76	34.11	11.18	10.95	10.75	10.69	10.59
	ANN – 2	83.16	73.80	77.58	0.1328	20.49	31.52	10.27	10.07	9.89	9.85	9.76
	ANN – 3	80.15	70.52	74.71	0.1441	20.79	40.37	12.52	12.42	12.33	12.31	12.26
50:50	ANN – 1	71.07	75.68	73.36	0.1380	22.77	44.00	13.66	13.54	13.45	13.42	13.37
	ANN – 2	69.25	73.92	71.43	0.1479	28.01	45.16	14.56	14.31	14.09	14.03	13.93
	ANN – 3	75.73	72.20	73.96	0.1240	22.81	41.57	13.05	12.91	12.78	12.75	12.69
90:10	ANN – 1	77.60	78.07	77.66	0.1427	19.87	31.73	10.25	10.07	9.92	9.87	9.79
	ANN – 2	78.85	80.46	79.00	0.1374	18.21	30.53	9.76	9.61	9.48	9.45	9.38
	ANN – 3	70.63	81.25	71.66	0.1456	25.64	44.68	14.16	13.97	13.81	13.77	13.69

Note: * Tr–dataset and Te–dataset refer to training and testing datasets, respectively.

Table 5. Performance obtained by the Japanese database

SCM Ratio (%)	Neural network	Accuracy (%)			RMS error	Error (%)		Expected misclassification cost (%)				
		Tr–dataset	Te–dataset*	Overall		Type – I	Type – II	5:1	7:1	10:1	12:1	15:1
40:60	ANN – 1	90.78	85.92	89.94	0.1466	16.00	5.77	3.87	3.71	3.57	3.54	3.47
	ANN – 2	92.23	89.23	91.67	0.1412	9.46	7.50	4.19	4.19	4.20	4.20	4.20
	ANN – 3	97.09	87.50	95.28	0.1405	4.25	5.05	2.67	2.71	2.74	2.75	2.77
50:50	ANN – 1	90.46	95.16	91.14	0.1529	6.40	10.51	5.37	5.50	5.61	5.64	5.69
	ANN – 2	91.78	88.68	91.38	0.1512	10.00	7.63	4.29	4.29	4.28	4.28	4.28
	ANN – 3	90.31	87.27	89.87	0.1549	15.25	5.56	3.71	3.56	3.44	3.40	3.34
90:10	ANN – 1	88.80	71.88	88.00	0.1477	14.43	9.94	5.70	5.67	5.64	5.63	5.62
	ANN – 2	88.53	88.89	88.55	0.1458	15.56	7.85	4.80	4.70	4.62	4.59	4.55
	ANN – 3	88.25	95.83	88.54	0.1443	13.49	9.77	5.55	5.53	5.51	5.51	5.50

Note: * Tr–dataset and Te–dataset refer to training and testing datasets, respectively.

Table 6. Performance obtained by the Chinese database

SCM Ratio (%)	Neural network	Accuracy (%)			RMS error	Error (%)		Expected misclassification cost (%)				
		Tr–dataset	Te–dataset*	Overall		Type – I	Type – II	5:1	7:1	10:1	12:1	15:1
40:60	ANN – 1	72.60	45.77	54.50	0.1789	1.30	97.89	1.84	1.87	1.90	1.90	1.91
	ANN – 2	60.09	58.28	59.07	0.1782	2.24	98.53	2.01	2.00	1.99	1.99	1.98
	ANN – 3	57.54	56.39	56.90	0.1785	2.03	98.46	1.97	1.97	1.97	1.97	1.97
50:50	ANN – 1	79.73	58.66	69.67	0.1669	1.90	95.97	1.91	1.91	1.91	1.92	1.92
	ANN – 2	71.34	71.72	71.52	0.1642	0.50	93.94	1.65	1.71	1.75	1.77	1.79
	ANN – 3	50.75	55.61	52.43	0.1813	1.64	96.16	1.89	1.90	1.91	1.92	1.92
90:10	ANN – 1	76.32	78.76	76.43	0.1547	0.84	96.80	1.75	1.80	1.84	1.85	1.87
	ANN – 2	80.77	58.01	78.83	0.1539	1.43	95.29	1.82	1.84	1.86	1.87	1.87
	ANN – 3	65.43	76.19	66.00	0.1687	1.16	96.55	1.80	1.83	1.86	1.87	1.88

Note: * Tr–dataset and Te–dataset refer to training and testing datasets, respectively.

Conversely, SPSS credit modeling database results presented in Table 7 reveal that ANN–1 credit prediction classifier in 40%:60% SCMR illustrates the best extrapolative performance through 91.22% overall accuracy with 92.55% training while 85.92% testing accuracies together with a minimum RMSE error of 11.47%. ANN–2 classifier in 90%:10% SCMR in contrast, shows the best predictive performance in connection with testing dataset accuracy of 90.91%.

Table 7. Performance obtained by the SPSS credit modeling database

SCM Ratio (%)	Neural network	Accuracy (%)			RMS error	Error (%)		Expected misclassification cost (%)				
		Tr–dataset	Te–dataset*	Overall		Type – I	Type – II	5:1	7:1	10:1	12:1	15:1
40:60	ANN – 1	92.55	85.92	91.22	0.1147	10.81	8.24	3.12	2.88	2.67	2.61	2.49
	ANN – 2	83.15	79.45	82.39	0.1492	25.00	16.45	6.65	6.05	5.56	5.41	5.13
	ANN – 3	78.99	75.00	78.27	0.1654	33.33	20.59	8.58	7.77	7.09	6.90	6.51
50:50	ANN – 1	91.42	86.25	90.43	0.1264	13.64	8.48	3.52	3.19	2.92	2.84	2.68
	ANN – 2	74.79	80.22	75.88	0.1586	25.00	24.11	8.31	7.80	7.37	7.25	7.00
	ANN – 3	77.01	73.45	76.23	0.1556	75.00	23.30	14.32	12.24	10.51	10.01	9.03
90:10	ANN – 1	88.53	82.14	88.29	0.1431	20.39	9.14	4.50	3.97	3.52	3.40	3.14
	ANN – 2	80.67	90.91	81.17	0.1537	26.21	17.47	7.02	6.40	5.88	5.73	5.43
	ANN – 3	73.68	81.82	74.09	0.1681	42.86	25.73	10.87	9.82	8.94	8.69	8.19

* Tr–dataset and Te–dataset refer to training and testing datasets, respectively.

2.2. Type I and type II errors with their corresponding EMC cost

Tables 3–7 also summarize the type I and type II errors of the neural models across five credit databases with their corresponding expected MC (EMC) costs. According to the results from Tables 3–7, for type I and type II errors, ANN–2 in 50%:50% SCMR reduces two indicators into 7.22% and the most competitive, 3.49% for the Australian credit dataset; into the best, 0.50% and the worst, 93.94% for the Chinese credit dataset. It is mentioned earlier that the later dataset is the most imbalanced (the ratio is about 43:1) and produce the worst type II error rate. For the German (Japanese) credit dataset, ANN–2 in 90%:10% (ANN–3 in 40%:60%) SCMR has the most competitive rate of type II error 30.53% (5.05%) and a significantly low (the lowest) rate of type I error 19.87% (4.25%) in associations with the other neural models. In addition, ANN–1 in 40%:60% SCMR decreases two indicators into the most competitive, 10.81% and 8.24% for SPSS credit modeling dataset. These results, however, are consistent with the overall accuracy of the models for all except Chinese, databases.

In case of prediction cost, for the Chinese and Australian credit databases ANN–2 in 50%:50% SCMR neural model is the best amongst all classifiers in all databases at MC

ratio of 5:1. Being an extensive investigation to incorporate MC ratio ratios of 7:1 to 15:1, ANN–2 in 50%:50% SCMR is still the best producing the minimum MC costs. For the Japanese credit, amongst all neural classifiers, the lowest EMC with all MC ratios, 5:1 to 15:1 is for ANN–3 in 40%:60% model; for the SPSS credit, is for ANN–1 in 40%:60% SCMR model; for the German credit, is for ANN–2 in 90%:10% SCMR model. These results are extremely significant for the decision makers, being noteworthy for them to achieve an appropriate balance between both error types so as not to lose potentially non-default creditors.

2.3. Selecting the optimal SCM ratio

We average the performance indicators in Figures 2–4 across all training to testing ratios over the five credit datasets to assess their influences for modeling the credit approval data. Other than above aspects, we then display the performance indicators in Figures 5–7 across five credit datasets to further review the neural based credit prediction models.

As illustrated by the Figures 2–4, it is possible to see that a SCM ratio of 90%:10% performs best in ANN–1 and ANN–2 models in regards to accuracy rate but it is an imbalance ratio which is biased toward majority class. There is no exact "winner" then for the type I error indicator. For example, ANN–1 performs the best in 50%:50% SCM ratio; following ANN–2 in 90%:10%, ANN–3 in 40%:60% SCM ratios. ANN–1 and ANN–2 once more, perform best considering type II error in 40%:60% SCM ratio. Different SCM ratios, therefore, show the different results on different indicators. Taking a closer look at Figures 5–7, in addition, the predictive performance of ANN–1 credit approval model is the best considering all indicators which means that the neural model with nine hidden units seems more stable.

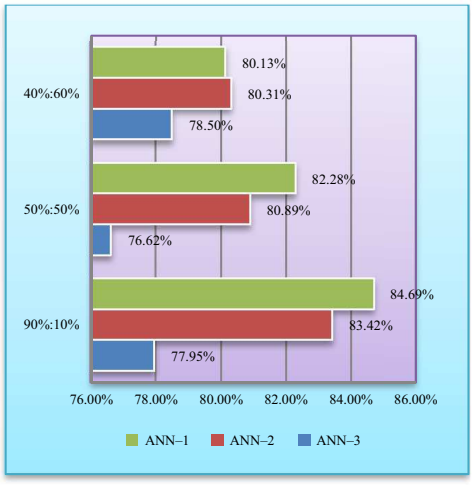

Fig. 2. Average results of accuracy on different Tr/Te sets*

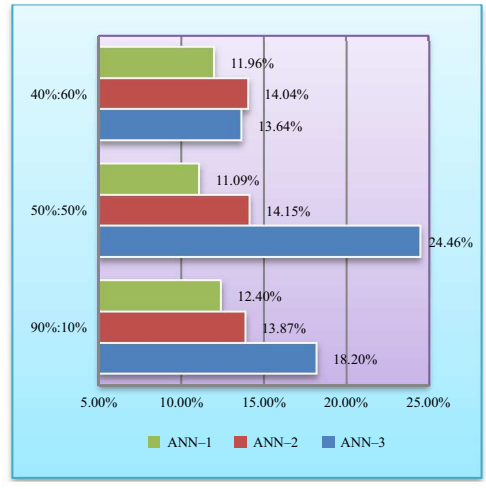

Fig. 3. Average results of type–I error on different Tr/Te sets*

Note: * Tr/Te sets refer to training and testing datasets, respectively.

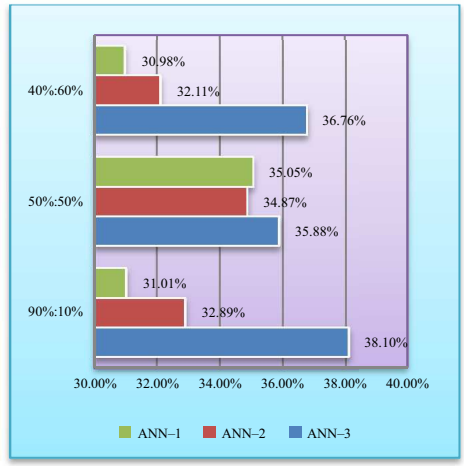

Fig. 4. Average results of type–II error on different Tr/Te sets*

Note: * Tr/Te sets refer to training and testing datasets, respectively.

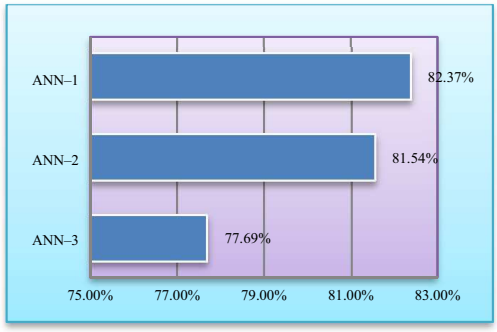

Fig. 5. Average results of overall accuracy

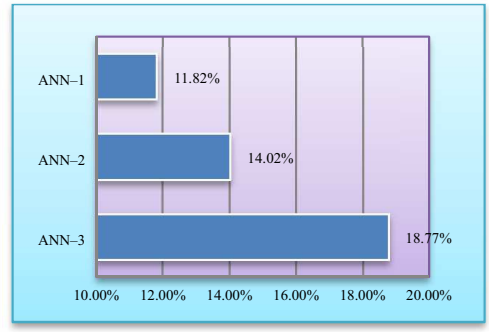

Fig. 6. Average results of type – I error

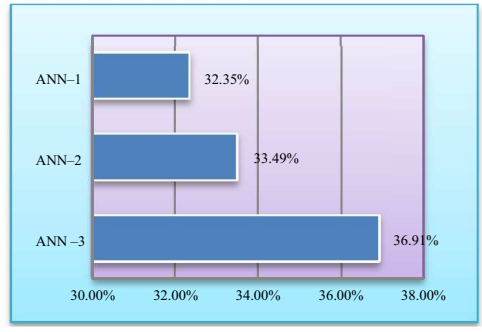

Fig. 7. Average results of type – II error

2.4. Comparison with the most perfect models

In order to see the reliability of the findings, we used a non-parametric Wilcoxon signed-ranks (WSR) test, establishing the significance level at $p = 0.01/0.05$, to fix on statistically significant performance differences between the neural based credit approval classifiers. All credit classifiers (Model Y) are verified for significant dissimilarities with the most perfect classifier (Model X) in the database. The null hypothesis is "Model X's overall accuracy/type I /II errors = Model Y's overall accuracy/type I/II errors" while the reverse is the alternative hypothesis. The column "improvement" presents the comparative improvement in overall accuracy (type I/II errors) that model X earns over model Y. The results are summarized in Supplementary information Tables B1–B5.

Evidences from Supplementary information Tables B1–B5 that ANN–2 in 50%:50% for the Chinese credit, following ANN–1 in 50%:50% for SPSS credit acquires a massive improvement with comparing to other classifiers considering overall accuracy criterion. For type I error, ANN–1 in 50%:50% for SPSS credit yields more than 81% improvement while for type II error, ANN–2 in 50%:50% for Australian credit obtains more than 78% improvement. It is obvious from the Supplementary information Tables B1–B5 that all improvements in all databases, except Chinese with a few in others, are statistically significant with respect to the best performing neural classifiers.

It is worth noting the following explanations from the investigation of the results presented in Tables 3–7 with Supplementary information Tables B1–B5.

1) The number of hidden units affects the accuracy of the classifier. Across most ranges of situations, ANN–2 with ten hidden units, a credit modeling classifier gets the higher prediction accuracy, lower type I and type II errors, and the most competitive EMCC in all except SPSS, databases and which is also justified by non-parametric WSR test. As illustrated in the Figures 5–7, in contrast, it is possible to see that the predictive performance of ANN–1 with nine hidden units over the five databases is the best considering average accuracy (82.37%), type I error (11.82%), and type II error (32.35%). The neural architecture (ANN–2) with ten hidden units is therefore proposed as a feasible approach to classifying the credit data which is consistent with the findings of Khashman (2011).
2) Although different ratios of training to testing set usually result in different performances, we found that a SCM ratio of 90%:10% gives better prediction accuracy, suggesting that increasing the training sample size has a positive influence on the classification accuracy.
3) Combining the numbers of creditors in each databases (Table 1), and the result of the best accuracy of database from Tables 3–7, an interesting result seems to be found that the lower the number of creditors with a reasonable feature set, the higher the classification accuracy with lower the type–II error. For the Australian and Japanese credit, 94.25% and 95.28% overall accuracy achieved, respectively, with 690 cases each (Tables 3 and 5); for the SPSS, it is 91.22% with 700 creditors (Table 7). A 79.00% overall accuracy earned with 1000 creditors in German (Table 4) while it is 78.83% with 3111 creditor and 81 features in Chinese case (Table 6), suggesting that when sample sizes are limited with a reasonable feature

set it appears that more reliable and accurate results could be attained by network architecture. But when a dataset is larger with vastly incomplete features (here it is Chinese), it may contain noisy/outliers information with redundant and irrelevant features, resulting in a low performance of the trained network.

Conclusions

During the history of credit prediction investigation, the neural networks have occupied an important position. For many years, existing literature supported the supremacy of NN classifier over a versatile optimization and statistical methods. However, three challenges have encountered the prestige of NN classifiers. First, the use of insufficient number of databases and small number of instances with fewer features has been a common problem in most relevant studies. Another problem in most cases when using neural networks is the use of either balance or imbalance training to testing ratio. More example set used in training means less used in testing, which may result in low accuracy. The number of hidden units is the third challenge since too many hidden nodes can easily cause an over–training of the network and insufficient units cannot attain optimal accuracy.

Therefore, we set out an experimental investigation to optimize MLP feed forward neural network for modeling credit approval data over the Australian, Chinese, Germen, Japanese and SPSS, five different databases, which can help to overcome the aforementioned challenges. We focused on the effect of hidden units in the hidden layers and versatile databases with different ratios of SCM in training-testing examples to boost up the performance of suggested investigation. In our experiment, we trained 45 unique neural models on three SCM ratios over five different databases. These models vary in the proportion of the number of credit approval examples used for training, against those used for testing. Having compared the experimental results of the 45 neural models; based on the performance appraisal criterions, it can be concluded that ANN–2; the neural architecture with ten hidden units, outperforms the remaining neural models. Besides the above aspects, a SCM ratio of 90%:10% gets better prediction accuracy, thus this combination is most fitting for modeling an acceptable neural architecture. In addition to, we find a motivating relationship between the sample sizes and the overall accuracy: when a network is trained with the limited sample sizes with a reasonable feature set, it may achieve the higher classification accuracy. With these contributions thus, we complement previous evidence and modernize the methods of credit prediction modeling in several domains. First, our classifier outperforms most of the relevant studies in terms of predicting ability. Second, the findings of our study are compared against an extensive set of substitute methods than relevant articles do. Third, our investigation is easier and, offer an apparent visualization of the complex neural architecture.

This study, however, has realistic implications for bank managers and other stakeholders to delineate the risk profile of the credit customers. From the managerial point of view, the defensive measures can be different in the short, medium, or long term based on the prediction of creditors' status, that is, in the group to which the creditor belongs. Similarly, the profitability value of more accurate credit predictions is an important

concern. Therefore, it is vital to judge whether the findings that we observe from this study simplifies to real-world applications, and to what extend their implementation would add to profit. These queries are much more debated in the literature and from this study, we can add some points to the debate. This investigation could also be extended to include other financial products by collecting more important features that will improve the prediction ability. We hope that these attempts would be taken in other regions by many modelers.

Acknowledgements

We are very grateful to anonymous reviewers for their substantial contribution because with their assistance the manuscript has been significantly improved. The authors would also like to thank Sanja Vukovic, Boris Delibasic, Ana Uzelac, and Milija Suknovic for providing SPSS credit modeling database. The research is supported by the National Natural Science Foundation of China (71171031 and 71471027), National Social Science Foundation of China (16BTJ017), Social Science Foundation of Liaoning Province of China (L16BJY016), Credit Risks Rating System and Loan Pricing Project of Small Enterprises for Bank of Dalian (2012–01) and Credit Risks Evaluation and Loan Pricing For Petty Loan Funded for the Head Office of Postal Savings Bank of China (2009–07). We thank the organizations mentioned above.

Disclosure statement

The authors declare that there is no conflict of interests regarding the publication of the paper.

References

Abellán, J.; Javier G. Castellano, J. G. 2017. A comparative study on base classifiers in ensemble methods for credit scoring, *Expert Systems with Applications* 73: 1–10. https://doi.org/10.1016/j.eswa.2016.12.020

Arlot, S.; Celisse, A. 2010. A survey of cross-validation procedures for model selection, *Statistics Surveys* 4: 40–79. https://doi.org/10.1214/09-SS054

Barrow, D. K.; Crone, S. F. 2016. Cross-validation aggregation for combining autoregressive neural network forecasts, *International Journal of Forecasting* 32: 1120–1137. https://doi.org/10.1016/j.ijforecast.2015.12.011

Costantino, F.; Gravio, G. D.; Nonino, F. 2015. Project selection in project portfolio management: an artificial neural network model based on critical success factors, *International Journal of Project Management* 33: 1744–1754. https://doi.org/10.1016/j.ijproman.2015.07.003

Falamaki, A. 2013. Artificial neural network application for predicting soil distribution coefficient of nickel, *Journal of Environmental Radioactivity* 115: 6–12. https://doi.org/10.1016/j.jenvrad.2012.06.008

Jeong, C.; Min, J. H.; Kim, M. S. 2012. A tuning method for the architecture of neural network models incorporating GAM and GA as applied to bankruptcy prediction, *Expert Systems with Applications* 39: 3650–3658. https://doi.org/10.1016/j.eswa.2011.09.056

Kao, L. J.; Chiu, C. C.; Chiu, F. Y. 2012. A Bayesian latent variable model with classification and regression tree approach for behavior and credit scoring, *Knowledge Based Systems* 36: 245–252. https://doi.org/10.1016/j.knosys.2012.07.004

Khashei, M.; Hamadani, A. Z.; Bijari, M. 2012. A novel hybrid classification model of artificial neural networks and multiple linear regression models, *Expert Systems with Application* 39: 2606–2620. https://doi.org/10.1016/j.eswa.2011.08.116

Khashman, A. 2009. A neural network model for credit risk evaluation, *International Journal of Neural Systems* 19(4): 285–294. https://doi.org/10.1142/S0129065709002014

Khashman, A. 2010. Neural networks for credit risk evaluation: investigation of different neural models and learning schemes, *Expert Systems with Applications* 37: 6233–6239. https://doi.org/10.1016/j.eswa.2010.02.101

Khashman, A. 2011. Credit risk evaluation using neural networks: emotional versus conventional models, *Applied Soft Computing* 11: 5477–5484. https://doi.org/10.1016/j.asoc.2011.05.011

Kim, K. J.; Ahn, H. 2012. A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach, *Computers & Operations Research* 39: 1800–1811. https://doi.org/10.1016/j.cor.2011.06.023

Lee, T. S.; Chen, I. F. 2005. A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines, *Expert Systems with Application* 28: 743–752. https://doi.org/10.1016/j.eswa.2004.12.031

Lee, Y. C. 2007. Application of support vector machines to corporate credit rating prediction, *Expert Systems with Application* 33(1): 67–74. https://doi.org/10.1016/j.eswa.2006.04.018

Li, S. T.; Shiue, W.; Huang, M. H. 2006. The evaluation of consumer loans using support vector machines, *Expert Systems with Application* 30(4): 772–782. https://doi.org/10.1016/j.eswa.2005.07.041

Li, K.; Niskanen, J.; Kolehmainen, M.; Niskanen, M. 2016. Financial innovation: credit default hybrid model for SME lending, *Expert Systems with Application* 61: 343–355.

https://doi.org/10.1016/j.eswa.2016.05.029

Lichman, M. 2013. *UCI machine learning repository* [online], [cited 20 August 2015]. Available from Internet: http://archive.ics.uci.edu/ml

Min, J. H.; Lee, Y. C. 2005. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters, *Expert Systems with Applications* 28: 603–614.

https://doi.org/10.1016/j.eswa.2004.12.008

Salchenberger, L.; Cinar, E. M.; Lash, N. A. 1992. Neural networks: a new tool for predicting thrift failures, *Decision Sciences* 23: 899–916. https://doi.org/10.1111/j.1540-5915.1992.tb00425.x

Shin, K. S.; Lee, T. S.; Kim, H. J. 2005. An application of support vector machines in bankruptcy prediction model, *Expert Systems with Applications* 28 (1): 127–135.

https://doi.org/10.1016/j.eswa.2004.08.009

Tang, T. C.; Chi, L. C. 2005. Neural networks analysis in business failure prediction of Chinese importers: a between-countries approach, *Expert Systems with Applications* 29: 244–255.

https://doi.org/10.1016/j.eswa.2005.03.003

Tsai, C. F.; Wu, J. W. 2008. Using neural network ensembles for bankruptcy prediction and credit scoring, *Expert Systems with Applications* 34(4): 2639–2649. https://doi.org/10.1016/j.eswa.2007.05.019

Tommi, K. 2014. On cross-validation for MLP model evaluation, in P. Franti, G. Brown, M. Loog, F. Escolano, M. Pelillo (Eds.). *Lecture Notes in Computer Science.* Finland: Springer, 291–300.

West, D. 2000. Neural network credit scoring models, *Computers & Operations Research* 27: 1131–1152. https://doi.org/10.1016/S0305-0548(99)00149-5

Vukovic, S.; Delibasic, B.; Uzelac, A.; Suknovic, M. 2012. A case-based reasoning model that uses preference theory functions for credit scoring, *Expert Systems with Applications* 39: 8389–8395. https://doi.org/10.1016/j.eswa.2012.01.181

Zhao, Z.; Shuxiang, X.; Byeong, H. K.; Kabir, M. M. J.; Yunling L.; Rainer, W. 2015. Investigation and improvement of multi-layer perceptron neural networks for credit scoring, *Expert Systems with Application* 42: 3508–3516. https://doi.org/10.1016/j.eswa.2014.12.006

Zhong, X.; Enke, D. 2017. Forecasting daily stock market return using dimensionality reduction, *Expert Systems with Applications* 67: 126–139. https://doi.org/10.1016/j.eswa.2016.09.027

Chi GUOTAI is Professor of Finance, Doctoral Adviser, Dalian University of Technology, Dalian 116024, China. His research interest includes asset-liability management, financial risk management, credit rating, etc. He is a visiting Professor at various universities in China and has successfully managed various national sponsored research projects and grants.

Mohammad Zoynul ABEDIN is an Investment Theory Doctor graduate student, Dalian University of Technology, Dalian 116024, China. His research interest includes financial risk management, credit analysis, data mining, artificial intelligence.

Fahmida–E–MOULA is an Investment Theory Doctor graduate student, Dalian University of Technology, Dalian 116024, China. Her research interest includes financial risk management, credit analysis.