Expert Systems with Applications 42 (2015) 3508–3516

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Investigation and improvement of multi-layer perceptron neural networks for credit scoring

Zongyuan Zhao [a], Shuxiang Xu [a], Byeong Ho Kang [b], Mir Md Jahangir Kabir [a], Yunling Liu [c,*], Rainer Wasinger [a]

[a] School of Engineering and ICT, University of Tasmania, Launceston, Tasmania, Australia
[b] School of Engineering and ICT, University of Tasmania, Hobart, Tasmania, Australia
[c] College of Information and Electrical Engineering, China Agricultural University, Beijing, China

ARTICLE INFO

Article history:
Available online 10 December 2014

Keywords:
Credit scoring
Neural networks
Back propagation

ABSTRACT

Multi-Layer Perceptron (MLP) neural networks are widely used in automatic credit scoring systems with high accuracy and efficiency. This paper presents a higher accuracy credit scoring model based on MLP neural networks that have been trained with the back propagation algorithm. Our work focuses on enhancing credit scoring models in three aspects: (i) to optimise the data distribution in datasets using a new method called Average Random Choosing; (ii) to compare effects of training–validation–test instance numbers; and (iii) to find the most suitable number of hidden units. We trained 34 models 20 times with different initial weights and training instances. Each model has 6 to 39 hidden units with one hidden layer. Using the well-known German credit dataset we provide test results and a comparison between models, and we get a model with a classification accuracy of 87%, which is higher by 5% than the best result reported in the relevant literature of recent years. We have also proved that our optimisation of dataset structure can increase a model's accuracy significantly in comparison with traditional methods. Finally, we summarise the tendency of scoring accuracy of models when the number of hidden units increases. The results of this work can be applied not only to credit scoring, but also to other MLP neural network applications, especially when the distribution of instances in a dataset is imbalanced.

1. Introduction

Credit scoring, or credit rating, is the set of decision models and their underlying techniques that help lenders judge whether an application of credit should be approved or rejected (Thomas, Edelman, & Crook, 2002). Although credit scoring systems cannot themselves forecast profits, they have been widely used in banks and other financial institutions (Jensen, 1992). So far, these systems have shown the ability to decrease credit risk and reduce "bad" loans. From this point of view, a highly accurate scoring system can save on costs and in this way increase profits.

Credit scoring systems can generally be divided into two kinds: new credit application judgement and prediction of bankruptcy after lending. The first kind uses personal information and the financial status of a loan applicant as inputs to calculate a score. If the score is higher than a "safe" level, the applicant will have a high probability of exhibiting good credit behaviour. On the contrary, a low score means high risk for the loan, so the lender needs to take careful consideration of the application. The second kind of credit scoring focuses on the credit record of existing customers. From the payment history of a customer, a financial institution can predict a customer's payment ability and alter his/her credit level. This paper only focuses on new credit application judgement scoring systems.

Compared with traditional credit scoring, which is calculated by professional bank managers, automatic scoring has some obvious advantages: it saves cost and time for evaluating new credit applications; and it is consistent and objective (Marques, Garcia, & Sanchez, 2013). However, some current computational approaches are not as capable as experienced loan experts on judgement accuracy. As the accuracy of scoring can greatly affect the interests of financial institutions, researchers continually strive to improve and enhance scoring accuracy rates. In recent years, artificial intelligence has shown its advantages in credit scoring in comparison with linear probability models, discriminant analysis, and other statistical techniques (Saberi et al., 2013).

* Corresponding author.
E-mail addresses: Zongyuan.Zhao@utas.edu.au (Z. Zhao), Shuxiang.Xu@utas.edu.au (S. Xu), Byeong.Kang@utas.edu.au (B.H. Kang), mmjkabir@utas.edu.au (M.M.J. Kabir), lyunling@163.com (Y. Liu), rainer.wasinger@utas.edu.au (R. Wasinger).

http://dx.doi.org/10.1016/j.eswa.2014.12.006
0957-4174/

Among all of the computational methods available, MLP models are widely utilised (Jensen, 1992; Oreski, Oreski, & Oreski, 2012; Zhong, Miao, Shen, & Feng, 2014) because they provide competitive prediction ability against other methods (West, 2000; Šušteršič, Mramor, & Zupan, 2009). In Werbos (1975), the back-propagation (BP) algorithm was developed and has now been widely used in training MLP feed-forward neural networks. Memetic pareto artificial neural networks (MPANN) have been used to optimise the BP algorithm using a multi-objective evolutionary algorithm and a gradient based local search (Abbass, 2003). This training method can reduce training time and at the same time enhance classification accuracy. That paper also presented a self-adaptive version called SPANN, which was faster than BP and able to largely reduce computational complexity. Many tests showed that RBF, LS-SVM and BP classifiers yielded very good performance with eight credit scoring datasets (Baesens et al., 2003). At the same time however, some linear classifiers such as LDA and LOG also generated good results. This indicated that the performance differences between some models were not obvious (Marqués, García, & Sánchez, 2012). Another test received similar results by testing the accuracy of several automatic scoring models using the German, Australian, and Japanese credit datasets (Bache & Lichman, 2013). It reported that compared with BP, the C4.5 decision tree performed a little better for credit scoring but both of them could achieve high accuracies. Also, Nearest Neighbour and Naïve Bayes classifiers appeared to be the worst in their tests.

Improvements to neural networks include altering the ratios of training and testing datasets, the number of hidden nodes, and the training iterations. A nine learning schemes with different training-to-validation data ratios was investigated and got the implementation results with the German dataset (Khashman, 2010). That paper concluded that the learning scheme with 400 cases for training and 600 for validation performed best with an overall accuracy rate of 83.6%. An emotional neural network (Khashman, 2008) is a modified BP learning algorithm. It has additional emotional weights that are updated using two additional emotional parameters: anxiety and confidence. When comparing emotional neural networks with conventional networks for credit risk evaluation, experimental results have shown that both models are effective, but that the emotional models outperform the conventional ones in decision making speed and accuracy (Khashman, 2011). Another enhancement is the artificial metaplasticity MLP, which is especially efficient when fewer patterns of a class are available or when information inherent to low probability events is crucial for a successful application. This model achieved an accuracy of 84.67% for the German dataset, and 92.75% for the Australian dataset (Marcano-Cedeño, Marin-de-la-Barcena, Jimenez-Trillo, Piñuela, & Andina, 2011). Fuzzy numbers can replace crisp weights and biases to overcome uncertainties and complexities in financial datasets (Khashei, Rezvan, Hamadani, & Bijari, 2013). Results here have shown that this hybrid classification model outperformed the traditional ANNs and are also better than SVM, KNN (k-Nearest Neighbours), and others.

There are some hybrid systems that have implemented neural networks as part of a larger whole construction of a combination system. In (Lee & Chen, 2005), a two-stage hybrid modelling procedure with ANN and multivariate adaptive regression splines (MARS) is presented. After using MARS in building the credit scoring model, the obtained significant variables then served as the input nodes of the ANN. However, the improvements were not obvious. In fact, ensemble systems like this one performed better only in one of the three datasets in the experiments of Tsai and Wu (2008). The authors compared MLP with multiple classifiers or classifier ensembles and concluded that the ability of hybrid systems was not better than usual methods, which meant that all methods needed to be considered when making financial decisions.

As to other methods, support vector machine (SVM) and genetic algorithms (GA) are also used for credit rating with good performance. In Hens and Tiwari (2012), a SVM model was refined by reduction of features using an F score and took a sample instead of a whole dataset to create the credit scoring model. Test results showed that this method was competitive in the view of accuracy as well as computational time. In Chi and Hsu (2012), they selected important variables for use by a GA to combine a bank's internal behavioural rating model and an external credit bureau model. This dual scoring model underwent more accurate risk judgment and segmentation to further discover the parts which were required to be enhanced in management or control from a mortgage portfolio. Other than SVM and GA, Clustering-Launched Classification (CLC) is also available and may perform better than SVM (Luo, Cheng, & Hsieh, 2009). A multi-criteria quadratic programming (MCQP) model was proposed based on the idea of maximising external distance between groups and minimising internal distances within a certain group (Peng, Kou, Shi, & Chen, 2008). This model could solve linear equations to find a global optimal solution and obtained the classifier and at the same time used kernel functions to solve nonlinear problems. Compared to SVM it seemed more accurate and scalable to massive problems. Decision tree (DT) is another good alternative method. A dual strategy ensemble trees was developed based on bagging and random subspace (Wang, Ma, Huang, & Xu, 2012). This DT model reduced influences of noise data and redundant attributes of data to get relatively higher classification accuracy.

Recently, imbalanced datasets (i.e. where instances belonging to one class heavily outnumber instances in other classes) have attracted attention, and some work has shown that appropriate ratios of different kinds of instances can augment classification accuracy. In Brown and Mues (2012), they used five real-world datasets to test the effect of good/bad credit instance ratio. Results showed that linear discriminant analysis (LDA) and logistic regression (LOG) performed acceptable rating accuracy with both slightly imbalanced datasets and highly imbalanced ones. To avoid the effect of imbalance data distribution, a dynamic classifier and dynamic ensemble selection of features were added in the scoring model. This resulted in better performance compared to ensemble static classifiers (Xiao, Xie, He, & Jiang, 2012).

As there are usually some irrelevant attributes in credit datasets, optimisation of input attribution is especially important for SVM systems. Two feature selection methods – a probabilistic classifier with a proper prior structure, and multiple kernel learning using a specific kernel calculation strategy – showed good performance with choosing parts of features in the German dataset (Gonen, Gonen, & Gurgen, 2012). Ping and Yongheng (2011) designed a hybrid SVM-based model using a neighbourhood rough set to select input features and using a grid search to optimise RBF kernel parameters. This model showed its ability when selecting input features and performed better in scoring compared with other hybrid classifiers. A novel feature selection model was proposed based on a rough set and scatter search method called RSFS (Wang, Hedar, Wang, & Ma, 2012). In RSFS, conditional entropy is regarded as heuristic to search optimal solutions. Tests on the Australian dataset and the German dataset demonstrated its competitive performance in saving computational costs and improving classification accuracy. In Yu, Yao, Wang, and Lai (2011), they not only optimised input features by the design of experiment (DOE) method, but also proposed a weighted least squares SVM that emphasised the importance of different classes. Tests in the German dataset and the Australian dataset showed that this combination method had acceptable accuracy with less computation time.

In general, a neural network with back propagation algorithm can score credit applications with high accuracy. Consistent exper-