A Hybrid Model Based on ANFIS and Nonlinear Feature Selection for Credit Risk Evaluation

Zhi-bin XIONG
School of Mathematical Sciences, South China Normal University, Guangzhou City, Guangdong Province, 510631. China

Keywords: Credit risk evaluation, ANFIS, Feature selection, Gebelein’s maximal correlation (GMC).

Abstract. Credit risk evaluation is an important decision process to financial institutions. Feature (variable) selection is a key step to many credit evaluation problems and it is often used as a dimension reduction technique to process credit data. However, the traditional correlation-based feature selection (CFS) is a linear analysis method when calculating the correlation coefficient and it cannot deal efficiently with nonlinearly correlated variables. This paper presents an improved approach of nonlinear correlation-based feature selection—Gebelein’s maximal correlation-based feature selection (GCFS), based on analysis of CFS and Gebelein’s maximal correlation (GMC), to realize the data reduction. Furthermore, an integrated model, GCFS-ANFIS model, is presented combined GCFS with Adaptive Neuro Fuzzy Inference System (ANFIS). The proposed model has been applied to credit evaluation based on the data collected from a set of Chinese listed corporations, and the results indicate that the performance of the GCFS-ANFIS model is much better than the ones of the other classic methods.

Introduction

Credit risk evaluation is an important to financial institutions which provide loans to businesses and individuals. A variety of credit evaluation models have been extensively used for the credit admission evaluation with the rapid growth in the credit industry, including traditional statistical methods, such as logistic regression (LR), linear discriminate analysis (LDA), logit and probit; and artificial intelligence methods, such as neural networks (NN), fuzzy logic (FL), genetic algorithm (GA), support vector machines (SVM), etc. [1]. A large of evidences found in numerous recent studies have shown that artificial intelligence methods can obviously improve the accuracy of statistical methods without the reliance on restrictive assumptions [2].

In the past decades, artificial intelligence models, especially NNs, are developed and applied quickly to credit risk prediction because of the advantages of treating non-linear data with adaptation, learning and approximation [3, 4]. Even though NNs showed a great deal of excellent experimental result in many studies, the shortcomings of NNs are also significant due to a “black box” syndrome and the difficulty in dealing with qualitative information, and the generated rules from NN are hard to understand [5]. On the other hand, fuzzy system as a rule-based development in artificial intelligence can not only tolerate imprecise information, but also make a framework of approximate reasoning. In recent years, ANFIS (Adaptive Neuro Fuzzy Inference System) has been used widely to generate nonlinear models of processes to determine input-output relationships because it is benefited from both the features of NN and fuzzy system [6]. Therefore, we believe that ANFIS would provide and valuable insights for researchers, financial institutions and policy makers who are interested in the field of credit risk evaluation.

However, despite the fact that ANFIS has outstanding performance, its evaluation performance and generalization ability is often influenced by its number of input variables (or feature variables). In the credit risk evaluation applications, high dimensional feature variables often lead to high computational, high complexity, instability or lack of predictive accuracy for most models [1]. So, identifying the dependent variables can help avoid the above defects by selecting and using only relevant variables in building evaluation models for a target variable. There are many ways for feature selection, including forward selection (FS), backward elimination (BE), stepwise regression (SR) and
correlation-based feature selection (CFS), etc. CFS is currently one of the most widely used methods for features selection in the field of economics and management because the selected features by CFS are the ones with a strong correlation with the target variable and low correlation with each other [7]. Pearson correlation coefficient is often employed the correlation measure generally for measuring the correlation among random variables in above feature selection methods. However, Pearson correlation coefficient is a linear measure and need to meet some specific assumptions, so it cannot deal efficiently with nonlinearly correlated variables. Gebelein’s maximal correlation (GMC) is the most general measure of correlation, because it avoids making statistical assumptions about the correlation and can detect linear or nonlinear correlation between two random features [8].

Therefore, this study presents an improved approach of nonlinear feature selection, Gebelein’s maximal correlation–based feature selection (GCFS), to realize the dimension reduction. Furthermore, a hybrid model, GCFS-ANFIS model, combined GCFS with ANFIS, is proposed for credit risk prediction. The experimental results show the proposed model can achieve a better performance than the models based traditional approach. By this way, the research results also provide a new idea and useful reference for credit risk management and credit decision making.

**Materials and Methods**

**Adaptive Neuro Fuzzy Inference System (ANFIS)**

ANFIS combined the fuzzy inference system with NN, it can be thought of as five-layer architecture each having a distinct functionality. For illustrating the system, it is assumed that the inference system has two inputs \(x, y\), and one output \(z\), and the ANFIS model uses Gaussian functions for fuzzy sets and linear functions for the rule outputs. The architecture of ANFIS is shown in Fig. 1.

![Figure 1. The architecture of ANFIS network.](image)

Layer 1: This layer is the input layer and each node \(I\) in this layer generates a membership grade of a linguistic label. The input and output in this layer are formulated as follows:

\[
O_{1i} = \mu_{A_i}(x), \quad i = 1, 2, \quad \text{Eq. (1)}
\]

where \(\mu_{A_i}(x)\) is fuzzy membership function (Gaussian membership function).

Layer 2: Each node in this layer is a fixed node labeled II which multiplies the incoming signals and sends the product out by Eq. (2)

\[
O_{2i} = \mu_{A_i}(x) \mu_{B_i}(y), \quad i = 1, 2, \quad \text{Eq. (2)}
\]

Layer 3: Every node in this layer is fixed node labeled N. The \(i\)-th node calculates ratio of the \(j\)-th rule’s firing strength to the sum of all rules firing strength.

\[
O_{3i} = w_i = \frac{\mu_{A_i}(x) \mu_{B_i}(y)}{\mu_{A_1}(x) \mu_{B_1}(y) + \mu_{A_2}(x) \mu_{B_2}(y)}, \quad i = 1, 2, \quad \text{Eq. (3)}
\]
Layer 4: The nodes in this layer is a square node with node function (as Eq. (4)), parameters in this layer will be referred to as the consequent parameters.

\[
O_{4j} = w_j f_i = w_j (p_i x + q_i y + r_j)
\]

where \( w_j \) is the output of layer 3, and \( \{p_i, q_i, r_j\} \) is the parameter set.

Layer 5: The single node in this layer is a fixed node labeled \( \Sigma \), which computes the overall output as the summation of all incoming signals.

\[
O_{5j} = \sum w_j f_i = \sum w_j f_i
\]

Correlation-based Feature Selection (CFS) Method

CFS is a filtering method that orders a set of features using a correlation-based heuristic evaluation function. It can find subsets of features that are strongly correlated with the target variable and weakly correlated amongst each other. The function is expressed as follows:

\[
D_M = \frac{k r_{yi}}{\sqrt{k + k(k - 1) r_{yi}}}
\]

Where \( D_M \) is the score of a feature subset \( M \) containing \( k \) features, \( r_{yi} \) is the average feature-target correlation \( (i \in M, y \in Y, Y \text{ is target variable set}) \), \( r_{yi} \) is the average feature-feature inter-correlation. The difference between normal filter method and CFS is that while normal filter provide scores for each feature independently, CFS gives a heuristic “merit” of a feature subset and reports the best subset it finds.

Proposed Model

As mentioned earlier, many feature selection methodologies use Pearson correlation coefficient to measure the correlation between two random variables. However, Pearson correlation coefficient is a linear measure, Gebelein’s maximal correlation (GMC) is a genuine measure of dependence and can reflect linear or nonlinear dependences. Based on the above analysis, this study introduces GMC into CFS method firstly, and constructs a non-linear feature selection method, Gebelein’s maximal correlation based feature selection (GCFS). Then, this study proposes a hybrid model which combines GCFS with ANFIS, named GCFS-ANFIS model. Before going into detail of GCFS-ANFIS, we will first review GMC.

Gebelein’s Maximal Correlation (GMC)

As we know, Pearson’s correlation coefficient \( \text{corr}(X, Y) \) is a linear measure to describe of dependence between two random variables, and obviously the equality \( \text{corr}(X, Y) = 0 \) does not imply independence of \( X \) from \( Y \). Gebelein has introduced the mixing (weak dependence) coefficient:

\[
\rho(X; Y) = \sup \text{corr}(f(X); g(Y)),
\]

where, the supremum is taken over all functions \( f, g \) with finite variance:

\[
0 < E[|f(X)|^2] < \infty; 0 < E[|g(Y)|^2] < \infty,
\]

and where \( \text{corr}(X; Y) \) is the classic (Pearson) correlation between random variables \( X \) and \( Y \). GMC, denoted \( \rho \), is a genuine measure of dependence also in that it satisfies all the seven postulates of Renyi (1959). According to Renyi’s study, only GMC satisfies all the seven postulates of the six
dependence measures (more specific description can be found in the literature [9]). Moreover, the definition of GMC given in Eq. (7) is also equivalent to ([10]):

\[ \rho(X;Y) = \text{sup } \text{corr}(f(X); g(Y)), \]

s.t. \[ E[f(X)] = E[g(Y)] = 0; \]
\[ E[f(X)^2] = E[g(Y)^2] = 1; \]

where, in this formulation the supremum is taken over all functions f, g with zero mean and unit variance.

In this paper, we use kernel canonical correlation analysis (KCCA) to approximate GMC, more detailed design of KCCA can be found in the literature [11].

**Construction of GCFS-ANFIS Model**

In this paper, we build a hybrid model, named GCFS-ANFIS model. The building steps of the model are as follows:

Step 1: Employ KCCA to calculate the GMC between variables (features) \( \rho_{ij} \);

Step 2: Introduce the GMC into CFS method, build a nonlinear feature selection method, named GCFS. At this point, Eq. (6) can be rewritten as Eq. (10):

\[ D_y = \sqrt{\frac{k \rho_{ij}}{k + k(k - 1) \rho_{ii}}} \]  

According to the method, the best subset of features is obtained;

Step 3: The selected features (variables) are used as input to the ANFIS model, and the output of the model is the final evaluation result.

The research framework of proposed model is shown in Fig. 2.

**Empirical Results and Comparison Analysis**

**Data Source**

The data of samples used in this paper are selected from the listed corporations of China. Those sample corporations can be divided into two categories: ST (special treatment) corporations and the normal corporations. The main reason that listed corporations become ST corporations is due to the bad financial status. So, the ST corporations denote the corporations in bad credit and the normal corporations denote the corporations in good credit in this study. This study selects 760 observations which include 190 ST and 570 normal corporations from 2004 to 2012. The data set is divided into two subsets: one is a training sample set with 560 corporations including 140 ST corporations and 420 normal corporations; another is a test sample set with 200 corporations including 50 ST and 150 normal corporations to test the performance of proposed model.
In this paper, 10 financial variables are selected in this data according to the prior studies [3,4], which include: net profit to total assets (X1), ratio of main business profit (X2), return on equity (X3), total liabilities to total assets (X4), quick ratio (X5), interest coverage ratio (X6), working cash to total liability (X7), turnover of total assets (X8), turnover of accounts receivable (X9) and growth ratio of main business income (X10).

**Empirical Results**

In this study, the first step for credit evaluation is feature (variable) selection with training sample set by GCFS method. Then, In order to provide evidence that the subset of features selected by GCFS performs well in evaluation accuracy, this study compares the accuracy results based GCFS selected variables with the results based on variables selected by classic methods, including forward selection (FS), backward elimination (BE), stepwise regression (SR) and PCFS (CFS based Pearson correlation coefficient). So, five credit risk models were built and the classification results of the corresponding testing samples were summarized in Table 1. From the results in Table 1, we can observe that the highest average correct classification rate of GCFS-ANFIS is 96.00%, while the average correct classification rate for the other four models are 94.00%, 94.00%, 93.50% and 93.00% respectively. The GCFS-ANFIS model performs well in the case of China listed corporations data. Not only is the correct rate higher but also the number of the features is fewer when comparing results from other models. From the results in Table 1, the proposed model’s number is least of five methods, only 3, while the numbers of the features for the four models are 4, 6, 4 and 4. In short, GCFS-ANFIS model performs well in terms of correct rate and number of selected features.

| Credit model | Features number | Credit risk prediction results | Average correct classification rate |
|--------------|-----------------|---------------------------------|-----------------------------------|
| GCFS-ANFIS   | 3               | 45/50 (90.00%)                  | 147/150 (98.00%)                 | 192/200 (96.00%)                |
| PCFS-ANFIS   | 4               | 43/50 (86.00%)                  | 145/150 (96.67%)                 | 188/200 (94.00%)                |
| BE-ANFIS     | 6               | 42/50 (84.00%)                  | 146/150 (97.33%)                 | 188/200 (94.00%)                |
| FS-ANFIS     | 4               | 42/50 (84.00%)                  | 145/150 (96.67%)                 | 187/200 (93.50%)                |
| SR-ANFIS     | 4               | 42/50 (84.00%)                  | 144/150 (96.00%)                 | 186/200 (93.00%)                |

Here a class 1 corporation is defined as a corporation with bad credit (ST corporations) while a class 2 corporation is the one with good credit (normal corporations).

**Conclusions**

In this paper, we introduce GMC into CFS method firstly, and construct a non-linear feature selection method, Gebelein’s maximal correlation based feature selection (GCFS). Then, a hybrid model for the credit risk prediction has been proposed. This model combines ANFIS and GCFS, named as GCFS-ANFIS model. In this model, we use KCCA method to measure the GMC between two random variables and integrate GMC into CFS, build a nonlinear feature selection method, named GCFS. The best subset of features is obtained based on the GCFS method. Next, the selected variables are used as input to the ANFIS model and the output (the final evaluation result) is obtained. In the end of this study, we investigate the performances of the PCFS-ANFIS model, BE-ANFIS model, FS-ANFIS model and SR-ANFIS model on credit risk evaluation based on a set of financial data selected from China listed corporations from 2004 to 2012. Experimental results show that the performance of GCFS-ANFIS is much better than other models and show the validity and effectiveness of the proposed model.
Acknowledgement

This research is supported by Ministry of Education in China Project of Humanities and Social Sciences Planning Fund (Project No. 16YJA790053)

References

[1] K.P. Ravi, V. Ravi, Bankruptcy prediction in banks and firms via statistical and intelligent techniques—A review, European Journal of Operational Research. 180 (2007) 1-28.

[2] Ning Chen, Bernardete Ribeiro, An Chen, Financial credit risk assessment: a recent review, Artif Intell Rev. 45 (2016) 1-23.

[3] Patricia Jackson, Wikkiam Perraudin, Regulatory implications of credit risk modeling, Journal of Banking and Finance. 24 (2000) 1-14.

[4] G. Thomas, Calderon, J.J. Cheh, A roadmap for future neural networks research in auditing and risk assessment, International Journal of Accounting Information Systems. 3 (2002) 203-235.

[5] A.F. Shapiro, The merging of neural networks, fuzzy logic, and genetic algorithms, Insurance, Mathematics and Economics. 31 (2002) 115-131.

[6] LY. Wei, A hybrid ANFIS model based on empirical mode decomposition for stock time series forecasting, Appl. Soft Comput. 42 (2016) 368-376.

[7] M. Duma, Bhekiisipho Twala, V. Fulufhelo, Nelwamondo and Tshilidzi Marwala, Partial imputation to improve predictive modelling in insurance risk classification using a hybrid positive selection algorithm and correlation-based feature selection, Current Science. 103 (2012) 697-705.

[8] C.O. Sakar, O. Kursun, A Method for Combining Mutual Information and Canonical Correlation Analysis: Predictive Mutual Information and Its Use in Feature Selection, Expert Systems with Applications. 39 (2012) 3333~3344.

[9] A. Renyi, On Measures of Dependence, Acta Mathematica Hungaricae. 10 (1959) 441~451.

[10] W. Bryc, A. Dembo, A. Kagan, On the Maximum Correlation Coefficient[J], Theory Probab. Appl. 49 (2005) 132~138.

[11] S. Akaho, A Kernel Method for Canonical Correlation Analysis, International Meeting of Psychometric Society (2001).