http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 10 (11): 2237-2241, 2011
ISSN 1812-5638 / DOI: 10.3923/itj.2011.2237.2241

The Comparisons of Personal Credit Evaluation Models

[1]Huanzhuo Ye, [1]Nan Li, [1]Hao Feng and [2]Yonghua Wang
[1]School of Information and Safety Engineering, Zhongnan University of Economics and Law, China
[2]China Merchants Bank, Wuhan Branch, China

Abstract: The financial institutions concern on the customer credit evaluation. Highly effective and accurate evaluation models can help the financial institutions to reduce the risk and loss. There are several commonly used evaluation models, for example, the logistic regression, decision tree, support vector machine, artificial neural network, etc. We use the same data set to test these models and give the advantages and disadvantages of these models. Among these models, there is not a unified view that which model is the best because eacxh one has their advantages. The logistic regression is the most stable model while the decision tree is the lowest. in stability, MLP-ANN has the better accuracy rate than other models.

Key words: Logistic regression, decision tree, SVM, ANN, models comparisons

INTRODUCTION

With the rapid development of our country's economy, the credit almost affects all crafts. It is especially important for the financial industry. With the growth in credit card transactions, there has been an increase in credit fraud. So how to construct an efficient, accurate and steady credit evaluation model appears particularly urgent.

In a well-developed financial system, the goal of credit risk management is to predict customers's credit risk and reduce the loss of the financial institutions. There are four steps to construct a credit evaluation model. First, collect data and process it. We must make use the object of study and consult the domestic and foreign relevant works. Then we can select the important characters for the concrete customer. But after this operation, we should unify the format of the data, delete the wrong data, repair the incomplete data and clean the duplicate data. Second, set up the personal credit evaluation index system. When selecting indexes, there are some principles to follow, such as available principle, independent principle and practical principle. Thirdly, set up the credit evaluation model. On the basis of the front two steps, we can choose the appropriate model for training and prediction. Now, there are several commonly used evaluation models, for example, the Decision Tree, BP neural network, Logistic Regression, Support Vector Machine and Genetic Algorithm to choose. Also, we can use a hybrid model which includes several different methods. Finally, estimate the model and forecast the data. After setting up the model, we must validate the model. The commonly used measuring methods are Misclassification rate and ROC (Receiver Operating Characteristic) curve. When the model is mature, it can be used to predict the inspection samples.

LITERATURE REVIEW

Now, a range of different techniques have been used in credit evaluation models, such as statistical methods and data mining methods. Among the statistical methods, it is common to use logistic regression and linear discriminant analysis to establish credit evaluation models. There are several data mining techniques used in credit evaluation, such as decision tree, support vector machines. Besides, artificial neural network of artificial intelligence has been employed in credit evaluation models.

A logistic regression model is built in Dong *et al.* (2010) and Qingyan and Yunhui (2004). In these two articles, specific data is used to train and test the model. A logistic regression model based on random coefficient is set up in Dong *et al.* (2010) which requires the coefficient to obey multivariate normal distribution. C5.0 algorithm is applied in credit evaluation (Sulin and Jizhang, 2009). SVM model is contrasted with logistic regression model, linear discriminative analysis and K-neighbor model (Bellotti and Crook, 2009). A same credit data is used to carry out the experiments. The author uses the ROC curves to estimate the different models. The SVM model based on linear function and Gaussian function has a better result compared with the traditional method. There are a lot of literatures carrying on a study of SVM (Zhu *et al.*, 2010; Luo *et al.*, 2009; Chen *et al.*, 2009). Neural network relies on the complexity

Corresponding Author: Huanzhuo Ye, School of Information and Safety Engineering, Zhongnan University of Economics and Law, China

of the system. It achieves the purpose of information processing by adjusting the connection relationship of the internal nodes. Some papers have examined neural networks (Lee and Chen, 2005; Liu, 2010; Khashman, 2010; Angelini *et al*., 2008).

METHODOLOGY

We use Logistic regression, Decision tree, SVM, Artificial Neural network to analyze the selected data respectively. Here, we introduce these techniques briefly.

Logistic regression: Wiginton used Logistic regression in credit evaluation firstly. It was proposed in 1980. In credit evaluation, we usually choose binary logistic regression. It studies the relations of the dependent variables and independent variables in a binary classification. Logistic regression method is different from the linear discriminative method which requests data to obey normal distribution.

There are several advantages of the Logistic regression. First, the stability is strong, the interpretation of the model is good and the use is simple. Second, it can be used to analyze both quantitative variables and qualitative variables. Thirdly, it's strong to expand function. Finally, it is applicable to calculate all kinds of the customers's default probability which have a data base.

Decision tree: The decision tree is a simple but extensively used classification technology. It was proposed by Hunt in 1966. Many decision tree study algorithms that appear afterwards are improvements of the original algorithm. The ID3 algorithm is the most influential algorithm in many of the improved algorithms. In addition, C4.5, C5.0 algorithm and CART (Classification And Regression Tree) algorithm are also the improved algorithms.

The decision tree algorithms don't need any prior assumption and don't assume attributes to obey certain probability distribution. Once the decision tree was established, the unknown sample can be classified quickly. In the worst situation, the complication of time is O(w) which is the depth of the tree. In addition, the result of the decision tree is easy to explain and redundant attributes will not affect its accuracy.

SVM: SVM (Support Vector Machine) has a solid statistical theoretical basis. SVM are learning systems that use a hypothesis space of linear functions in a high-dimensional space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory. SVM separates binary classified data by a hyperplane so that the margin width between the hyperplane and the examples is maximized. It is widely used in the digit identification, text classification, credit evaluation, etc. Now, more and more scholars begin to pay close attention to SVM technique.

SVM technique is widely used in different fields because it has many advantages. SVM learning problems can be expressed as a convex optimization problem. So, it can use the known effective algorithm to find the global minimum of the objective function. Maximizing the margin width can reduce the complexity of the model and reduce the general risk of error.

Artificial neural network: The artificial neural network is constituted by a lot of neurons which are contacting and interacting with each other. This network relies on the complexity of the system and adjusts the connection relationships of the internal nodes to achieve the purpose of processing information.

There are several characteristics of the artificial neural network. The multilayer neural network is a kind of universal approximator which can be used to approximate any function. It can process the redundancy features, because the weights in the training process will be adjusted automatically. The weights of the redundancy features are very small. Training ANN is a process that consumes much time, especially when there are many hidden nodes. However, classification of the test samples is very fast.

EXPERIMENT

We analyze personal credit data using these four techniques, respectively and then the experiment results.

Data preparation: We use a Germany bank's personal credit data set obtained from the internet (Table 1). The number of instances is 1000, each record having 21 fields. The front 20 fields are the descriptions of the customers' credit information. The 20 fields include: Status of existing checking account, duration in month, credit history, the loan purpose, credit amount, saving account/bonds, present employment since, installment rate in percentage of disposable income, personal status and sex, other debtors/guarantors, present residence since, property, age, other installment plans, housing, number of existing credits at this bank, job, number of people being liable to provide maintenance for, telephone, foreign worker. The last filed is the customer's credit lever which is defined by the bank. The customers are divided into two types: "Good" and "Bad". "Good" customer is the one that the credit agencies are willing to provide consumption credit for. The agencies believe they can pay the principle and interest on time. "Bad" customer is the one that the credit agencies are not willing to provide consumption credit

Table 1: The fields of the Germany credit data

Field ID	Field name	Type	Values	Value instruction
A1	Status of existing checking account	Discrete	A11, A12, A13, A14	A11: <0 DM. A12: <200 DM. A13: ≥ 200 DM/ salary assignments for at least 1 year. A14: No checking account.
A2	Duration in month	Continuous	(4, 72)	
A3	Credit history	Discrete	A30, A31, A32, A33, A34	A30: No credits taken/all credits paid back duly. A31: All credits at this bank paid back duly. A32: Existing credits paid back duly till now. A33: Delay in paying off in the past. A34: Critical account/other credits existing (not at this bank).
A4	Purpose	Discrete	A40, A41, A42, A43, A44, A45, A46, A47, A48, A49, A410	A40: Car (new). A41: Car (used). A42: Furniture/equipment.A43: Radio/television. A44: Domestic appliances. A45: Repairs. A46: Education. A47: Vacation. A48: Retraining. A49: Business. A410: Others.
A5	Credit amount	Continuous	(250, 18424)	
A6	Savings account/bonds	Discrete	A61, A62, A63, A64, A65	A61: <100DM. A62: ≥100DM and <500DM. A63: ≥500DM and <1000DM. A64: ≥1000DM. A65: Unknown/no savings account.
A7	Present employment since	Discrete	A71, A72,A73, A74, A75	A71: Unemployed. A72: <1 year. A73: ≥1 year and <4 years. A74: ≥4 years and <7 years. A75: ≥7 years.
A8	Installment rate in percentage of disposable income	Continuous	(0,1)	
A9	Personal status and sex	Discrete	A91, A92, A93, A94, A95	A91: Male (divorced/separated). A92: female (divorced/separated/married). A93: Male (single). A94: Male (married/ widowed). A95: Female (single).
A10	Other debtors/guarantors	Discrete	A101, A102, A103	A101: None. A102: Co-application. A103: Guarantors.
A11	Present residence since	Continuous	(1, 10)	
A12	Property	Discrete	A121, A122, A123, A124	A121: Real estate. A122: If not A121, Building society saving agreement. A123: If not A121/A122, car or other, not in. A124: Unknown/no property.
A13	Age in years	Continuous	(19, 75)	
A14	Other installment plans	Discrete	A141, A142, A143	A141: Bank. A142: Stores. A143: None.
A15	Housing	Discrete	A151, A152, A153	A151: Rent. A152: Own. A153: For free.
A16	Number of existing credits at this bank	Continuous	(1,4)	
A17	Job	Discrete	A171, A172, A173, A174	A171: Unemployed/unskilled (non-resident). A172: Unskilled-resident. A173: Skilled employee/official. A174: Management/self-employed/highly qualified employee/officer.
A18	No. of people being liable to provide maintenance for	Continuous	(1,4)	
A19	Telephone	Discrete	A191, A192	A191: None. A192: yes, registered under the customer's name.
A20	Foreign worker	Discrete	A201, A202	A201: Yes. A202: No.

Table 2: Two type errors

	Classified as good customer	Classified as bad customer
Good customer	p_{11}	$1-p_{11}$ (Type a error)
Bad customer	$1-p_{22}$(Type B error)	p_{22}

for. Because, the agencies consider that they can't pay the principle and interest on time.

Usually, we use two kinds of error rates to estimate the model. If the good customer is classified as bad customer, we called this error A. We called the opposite situation error B. The two errors are shown in Table 2. In Table 2, $1-p_{11}$ is the probability of type A error and $1-p_{22}$ is the probability of type B error. Under the general situation, we think the error B causes greater losses. So, we suppose the cost of error B is an integer which is 1 or greater than 1, the cost of A is 1.

We divide the 1000 records into two parts. One part is used to train the model, called training sample. The other part is used to test the model, called inspection sample. We use the 70% of the samples to train the model and the rest 30% to test the model. The data has been dited and several indicator variables added to make it suitable for algorithms which cannot cope with categorical variables.

Experiment and results analyses: We use the SPSS (Statistical Product And Service Solution) Clementine to do Logistic regression experiment and can get the results in Table 3.

We use the C5.0 algorithm of Decision Tree to do the experiment. We consider the total error and error B rate synthetically. When the cost of error B is 2, the model is better. Table 4 shows the result of the experiment.

We use Matlab (Matrix Laboratory) to do the SVM experiment, result is shown in Table 5.

Table 3: Results of logistic regression

Type	Error A rate (%)	Error B rate (%)	Total error rate (%)
Training data	6.714	14.143	20.857
Test data	7.333	13.333	20.667

Table 4: Results of five different values of the error B cost

Type	Cost of error B	1%	2%	3%	4%	5%
Training	Total error rate	1.857	0.143	1.714	9.286	22
data	Error A rate	0	0.143	1.714	9.286	22
	Error B rate	1.857	0	0	0	0
Test	Total error rate	24	26.333	28.667	33	37.333
data	Error A rate	8	17.333	21.667	27.333	33
	Error B rate	16	9	7	5.667	4.333

Table 5: The result of the simple SVM model

Type	Error A rate (%)	Error B rate (%)	Total error rate (%)
Training data	5.286	7.571	12.857
Test data	8	13	21

Table 6: Results of MLP and RBF

	Type	Error A rate (%)	Error B rate (%)	Total error rate (%)
MLP	Training data	9.571	14	23.571
	Test data	7.333	12.667	20
RBF	Training data	4.286	22.714	27
	Test data	5	21	26

We use the MLP (Multi-layer Perceptron) and the RBF (Radial Basis Function) methods in the Matlab to do the experiment. Table 6 shows the results of the experiment.

It can be seen from the experiments above that every model has its own advantages. Now, there is not a unified view that which model is the best. We use the error rate to evaluate different models. In other words, we compare the accuracy and misclassification rate of every model, when using the model to classify the training samples and test samples. But total error rate is not the only evaluation criteria. As mentioned above, there are two kinds of error rates: Error A rate and error B rate. In practical problems, the damage of the two kinds of errors is difficult to estimate. So, besides the total error rates, the rates of error A and error B are also compared. Figure 1 shows the comparative results of different models with training data and Fig. 2 shows the comparative results of different models with test data.

Based on the experiment data, we can see from the two figures:

- The models we choose have a good resolution. If we don't use the model to classify the "Good" customers and the "Bad" customers, the classification rate is only 50%
- Among the models, logistic regression is the most stable. The difference between the total error rate of training data and test data is only 0.21%. On the contray, decision model, SVM model, MLP-ANN model and RBF-ANN model have a bigger difference

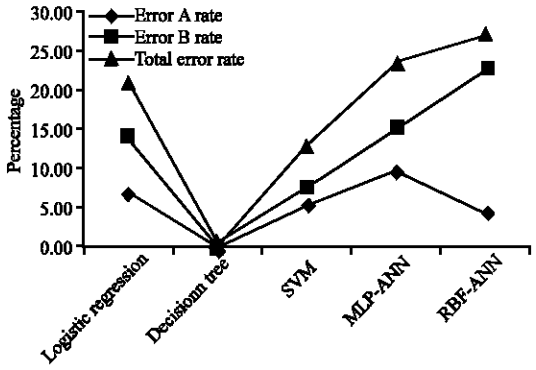

Fig. 1: Comparative results of different models with training data

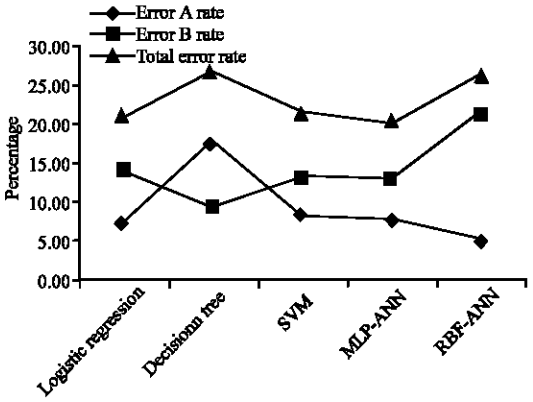

Fig. 2: Comparative results of different models with test data

- The total error rate of test data is higher than the total rate of training data in decision tree model and SVM model. So, we can not only use the error rate of training data to evaluate models. Using the error rate of test data is a better way
- The stability of the decision tree is low. It must be used carefully. Otherwise, it will cause great losses

When using the models, we must choose the right model according to actual situations. If the total error rate is more important to the financial institutions, the MLP-ANN model can be chosen.

If the financial institutions think the rate of error B is the most important, they can choose the decision tree model or MLP-ANN model.

CONCLUSIONS

This study compares four techniques which are Logistic regression, Decision tree, SVM and Artificial Neural network in personal credit evaluation. The dataset

analyzed consist of 1000 instances which is from a Germany bank's personal credit data set. Seven hundred instance is for training data and 300 is for test data. The results show that every model has its own advantages and disadvantages. There is not a unified view that which model is the best. Among these four, logistic regression is the most stable while the decision tree is the lowest in stability, MLP-ANN has the best accuracy rate. There are also some limits in this study. First, the data is not comprehensive, some important attributes maybe ignored; Second, the parameters in above four techniques are not unchanged, difference parameters have marked impact. The next step work is using more comprehensive data and more suitable parameters to compare these techniques, integrate these techniques and gain better model for customer credit evaluation.

ACKNOWLEDGMENT

This study was supported by National Natural Science Foundation of China (70972138). And it is a partial result of the research project of the Educational Commission of Hubei Province of China (2009b080). It is also a partial result of the project of 2009 Graduate Education Innovation Foundation of Zhongnan University of Economics and Law, China (2009SGL03).

REFERENCES

Angelini, E., G. di Tollo and A. Roli, 2008. A neural network approach for credit risk evalution. Q. Rev. Econ. Finance, 48: 733-755.

Bellotti, T. and J. Crook, 2009. Support vector machines for credit scoring and discovery of significant features. Expert Syst. Appl., 36: 3302-3308.

Chen, W., C. Ma and L. Ma, 2009. Mining the customer credit using hybrid support vector machine technique. Expert Syst. Appl., 36: 7611-7616.

Dong, G., K.K. Lai and J. Yen, 2010. Credit scorecard based on logistic regression with random coefficients. Procedia Comput. Sci., 1: 2463-2468.

Khashman, A., 2010. Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. Expert Syst. Appl., 37: 6233-6239.

Lee, T.S. and I.F. Chen, 2005. A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. Expert Syst. Appl., 28: 743-752.

Liu, Y., 2010. Improving the RBF neural network under stein loss and its application to the credit score. Int. Conf. Adv. Comput. Theory Eng., 5: V5-335-V5-337.

Luo, S.T., B.W. Cheng and C.H. Hsieh, 2009. Prediction model building with clustering-launched classification and support vector machines in credit scoring. Expert Syst. Appl., 36: 7562-7566.

Qingyan, S. and J. Yunhui, 2004. The comparative study of various personal credit score models in China. Stat. Res., 6: 43-47.

Sulin, P. and G. Jizhang, 2009. The application of C5.0 classification algorithm in individual credit scoring. Syst. Eng. Theory Practice, 29: 5-103.

Zhu, C., Y. Zhan and S. Jia, 2010. Credit risk identification of bank client basing on supporting vector machines. Proceedings of the 3rd International Conference on Business Intelligence and Financial Engineering, Aug. 13-15, Hong Kong, pp: 62-66.