Credit risk Evaluation by hybrid data mining technique

Weimin Chen*, Guocheng Xiang, Youjin Liu, Kexi Wang

Hunan University of Science and Technology, Xiangtan 411201, China

Abstract

Most studies have concentrated on building an accurate credit scoring model to decide whether or not to grant credit to new applicants and the efforts to build more accurate credit scoring model seems to be not significant. In this paper, we proposed a hybrid data mining technique which contains two processing stages. In the clustering stage, the samples of the accepted and new applicants are grouped into homogeneous clusters, the isolated samples are deleted and inconsistent samples are relabeled. In the classification stage, support vector machines used samples with new labels to build the scoring model. The difference from the other credit scoring model is that the samples were classified into three or four classes, rather than two the good and the bad credit classes. Experimental results based on the credit data set provided by a local bank in China showed that by choosing a proper cut-off point, super classification accuracy of the good and the bad credit is obtained. Risk management strategies are developed according to the characteristic of each class.

© 2011 Published by Elsevier Ltd. Selection and peer-review under responsibility of Desheng Dash Wu.

Keywords: Credit scoring; Data mining; clustering; Support vector machines; Credit-Risk evaluation; financial engineering

1. Introduction

The major functions of credit cards include consumer credit, balance, deposit and withdrawal of cash and bank transfer etc. Credit cards with such advantages as simple and quick application, high-usage of fund, a flexible loan, etc., seems much more economical and practical than a simple loan application or withdrawal of the time deposits before maturities for those who need money for contingencies, even if they cannot enjoy the interest-free loan funds. With the rapid growth in the credit industry, credit scoring models have been extensively used for the credit admission evaluation. The credit industry can benefit from improving cash flow, insuring credit collections, reducing possible risks and implementing better managerial decisions and credit cards have become an important source of income for banks. But credit card fraud also makes great losses for banks. With the widespread use of credit cards, personal credit scoring is getting increasingly important, more and more attention has been paid to credit scoring.

With the emergence of the Basel II requirement, the issue with granting loans or credit card becomes more important for financial institutions in risk management. BASEL II, which was implemented in 2006, requires more flexibility and risk sensitivity. It allows banks to measure credit and operational risk using an internal rating...
Credit scoring is the term used to describe formal statistical methods used for classifying applicants for credit into ‘good’ and ‘bad’ classes\textsuperscript{[5]}. The purpose of credit scoring is to classify the applicants into two types—applicants with good credit and applicants with bad credit. The former class has great possibilities to repay financial obligations, and the latter have high possibilities of defaulting. Those who are assigned to the ‘bad credit’ group whose application will be denied. In fact, the credit accuracy is quite significant in credit scoring. That the accuracy of applicants with bad credit increases only 1% may retrieve a great loss for the bank.

Many credit scoring models have been developed by bankers and researchers for the credit admission decision. At first personal credit scoring was evaluated subjectively according to personal experiences, and later it was based on 3Cs—Character, Capacity and Collateral. In recent years, the traditional statistic methods and contemporary data mining methods are adopted in credit scoring, such as linear discriminant analysis (LDA), logistic regression, classification and regression tree (CART), multivariate adaptive regression splines (MARS), expert system, genetic programming models, neural network and its latest development—support vector machines (SVM)\textsuperscript{[6-7]}. A number of effective credit scoring models of hybrid technique have been proposed in recent years. Lee et al. proposed a hybrid neural discriminant technique integrating BP neural network and the discriminant analysis, which outperform better than the BP neural network and discriminant analysis in convergence rate and credit scoring accuracy\textsuperscript{[8]}; Lee and Chen proposed a two-stage hybrid procedure with artificial neural networks and multivariate adaptive regression splines\textsuperscript{[9]}; Huang, Chen, and Wang integrated support vector machines, genetic algorithms and F-score\textsuperscript{[10]}. There are also other hybrid techniques, such as\textsuperscript{[11-12]}. In order to solve classification and decrease the error of credit scoring model, reassigning models have been proposed\textsuperscript{[12-13]}.

The efforts to build more accurate credit scoring model seems to be not significant. These results are summarized in\textsuperscript{[6-7, 14]}. The reason may be the flat maximum, which means significant changes in the weights around the optimal model have relatively little effect on its performance. Due to the highly nonlinear characteristic of credit data and the impact of economic conditions, even the scoring models are accurate, some misclassification patterns could arise, such as Type I error and type II error.

Most studies have concentrated on building an accurate credit scoring model to decide whether or not to grant credit to new applicants. To strengthen customer behavioral management for existing credit card customers, Hsieh proposed an integrated mining and behavioral scoring model to manage the existing credit card customers\textsuperscript{[15]}. The purposes of this paper are to mine the bank database and classify the credit card holder into different groups, and serve as a tool for establishing better bank risk management. The proposed credit scoring model has two processing stages. In the clustering stage, the samples of accepted and new applicants are grouped into homogeneous clusters, and the isolated samples are deleted and inconsistent samples are relabeled. In the classification stage, support vector machines use samples with new labels to build the scoring model. One of the differences from other credit scoring models is that the samples were classified to three or four class, rather than good and bad credit.

The organization of this paper is as follows. Section 2 offers a hybrid data mining credit scoring model. Section 3 empirically tests the hybrid model using two real world data sets. Finally, conclusions and future research directions are discussed in Section 4.

2. Proposed model

The proposed model, as shown in Fig. 1, is a two-stage approach: k-means cluster, support vector machines classification and computation of feature importance.

2.1. K-means cluster

Building a good performance model for customer credit management is difficult because the unrepresentative samples exist. The purpose of credit scoring is to classify the applicants into two types—applicants with good credit and applicants with bad credit. The data set from which the model parameters are estimated are always
those applicants who have been accepted in the past. The accepted samples with known classes, which are usually two classes: ‘good credit’ and ‘bad credit’, may include isolated or inconsistent cases. The clustering of the data set will improve the prediction efficiency by treating the heterogeneous borrowers separately \[16-17\].

![Fig. 1 Procedure of proposed reject inference model](image)

In this phase, the clustering process generated clusters of representative samples belonging to the class. In order to improve the performance of the reject inference model, the data set are grouped into homogeneous cluster, and isolated samples are eliminated. Meanwhile, inconsistent samples should be investigated closely to prevent the loss of valuable information and some of them should be relabeled.

2.2. Support vector machines classification

Support vector machines are taken as the best one in dealing with small sample classification and regression. There are three main problems we may encounter when applying SVM to treat classification and regression: 1) selecting the optimal feature; 2) the choice of kernels; and 3) the determination of the kernel’s parameter. In this paper, the accepted samples and the new applicant samples processed by 2.1 are regarded as the input feature. RBF can handle the case when the relation between class labels and attributes is nonlinear, often obtains satisfactory results and is most often used, so it is chosen to be the kernel. The parameters should be optimized to improve the classification accuracy. In this paper, we adopt grid-search to obtain the optimal parameter. Consider a grid space of \((C, \gamma)\) with \(\log_2 C \in \{-5,-3,-1,\ldots,13\}\) and \(\log_2 \gamma \in \{-13,-11,-9,\ldots,5\}\), for each pair of parameters, conduct 5-fold cross validation on the training set. In the neighborhood of the parameter \((C, \gamma)\) that leads to the lowest CV error classification rate, choose a fine grid, and repeat this step. Choose parameter \((C, \gamma)\) that leads to the lowest CV error classification rate to create a model as the classifier.

In order to obtain valuable information for risk management, the samples of credit cards are classified into three or four types, rather than two types—applicants with good credit and applicants with bad credit.

3. Empirical analysis

3.1. Data set
In order to verify the feasibility and effectiveness of the proposed hybrid SVM model, one credit card dataset contains 2000 samples provided by a local bank in China is used in this study. Each bank client in the dataset contains fifteen predictor variables: Country/State, certificate, gender, native place, age, marital status, number of dependents to support, educational level, telephone, residential status, occupation, ownership of employment establishment, annual income, a client of the bank or not, and credit limits. And the response variable is the credit status of the customer—good or bad credit. Two thousand datasets with respect to the ratio of good and bad credits were randomly selected and then used to build the credit scoring models. Among them, 80% of the datasets will be used as the training set and the remaining 20% will be reserved as the testing set.

For nominal attributes, we use m-dimensional vector to represent an m-category attribute according to; for numeric attributes, we linearly scale them to the range [-1,1].

3.2. k-means cluster to process the credit set

The difficulty of developing a credit scoring model is the lack of samples from which the parameters are obtained. Clustering samples before attributes contributes to the network’s achieving a higher accuracy. Hence, k-means cluster algorithm is employed to obtain homogeneous groups.

To obtain the representative homogeneous samples, the procedure of clustering was tried several times. Unrepresentative samples prevent accurate predictions. The isolated samples should thus be deleted and inconsistent clusters investigated in order to prevent loss of valuable information and improve the discriminative ability. Credit status of applicants in the inconsistent cluster may have more of an opportunity to change. The training set of credit card dataset is grouped into two, three and four clusters. When the samples are grouped into three clusters, the results seem to be satisfactory. The results are summarized in Table 1. There are 176 isolated samples to be deleted. Cluster-1 has 23 samples of good credit and 371 samples of bad credit. Since the proportion of samples having good credit is small, these samples can be ignored or relabeled with bad credit label. In this paper, the 23 samples of good credit were relabeled. Cluster-2 is an inconsistent cluster, which contains 183 samples of good credit and 168 samples of bad credit. Cluster-3 has 672 samples of good credit and 51 samples of bad credit. These 51 samples were relabeled. Fig. 3 depicts the distribution of the credit data.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Number of samples</th>
<th>Credit status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster -1</td>
<td>394</td>
<td>23(1), 371(-1)</td>
</tr>
<tr>
<td>Cluster -2</td>
<td>351</td>
<td>183(1), 168(-1)</td>
</tr>
<tr>
<td>Cluster -3</td>
<td>723</td>
<td>672(1), 51(-1)</td>
</tr>
<tr>
<td>isolated</td>
<td>176</td>
<td>153(1), 23(-1)</td>
</tr>
</tbody>
</table>

3.3. Classification and suggestion

Until now, most existing credit scoring models have concentrated on building an accurate credit scoring model to decide whether or not to grant credit to new applicants. Due to the highly nonlinear characteristics of credit data and the impact of economic conditions, even if the scoring models are accurate, some misclassification patterns could be emerged, such as Type I error and type II error. Especially Type II error, which means a customer with bad credit is misclassified as a customer with good credit, incurs great losses for banks. Thus, by choosing moderate cut-off point, the test samples are classified as more groups. The purpose is to obtain more homogeneous class through the proposed technique, and develop strategies according to the characteristics of each class. Three and four classes are tried by the trained SVM classifier, the results are summarized in Table 3 and Table 4 respectively.

The cut-off point was set as follows: in the case where 3 classes are tried, the first point is chosen as 0.76 so that the predictive value of test samples greater than this point are always good credit; the other point is chosen as 0.34, the predictive value of test samples lower than this point are always bad credit; and the samples evaluated between the two cut-off points are categorized into another class; in the case where 4 classes are tried, the first
and the last cut-off points are the same as those in 3-classes case; the in-between group is once again divided into 2 classes with the cut-off point at the mid-point within the class.

For the three class case, samples in class-1 are those with good credit, bank could give them higher credit limit and offering other credit products; samples in class-3 are those with bad credit, their credit card should be canceled at once; samples in class-2 are those may have more of an opportunity to change their credit status, their credit limit should be decreased even to zero, and more attention should be paid.

For the four class case, class-1 and class-4 are similar to class-1 and class-3 in three class case. Samples in class-2 have more opportunity to repay and their credit limit should be kept. Samples in class-3 have more opportunity to delay and their credit limit should be decreased.

4. Conclusions and future works
Constructing the credit scoring models with a credit database can be taken as a task of data mining. Credit scoring is engineering and not pure science. Until now, most existing credit scoring models have concentrated on building an accurate credit scoring model to decide whether or not to grant credit to new applicants. Due to the highly nonlinear characteristics of credit data and the impact of economic conditions, even when the scoring models are accurate, some misclassification patterns could occur, such as Type I error and Type II error.

In this paper, we proposed a hybrid data mining technique which contains two processing stages. In the clustering stage, the samples of accepted and new applicants are grouped into homogeneous clusters; the isolated samples are deleted and inconsistent samples are relabeled. In the classification stage, support vector machines use samples with new labels to build the scoring model. The difference from the other credit scoring models is that the samples were classified into three or four classes, rather than the two good and bad credit. Experimental results based on the credit data set provided by a local bank in China showed that by choosing a proper cut-off point, super classification accuracy of good and bad credit is obtained. Risk management strategies are developed according to the characteristics of each class.

Future research can be made to investigate more suitable cut-off point and the rule between the variable and the credit status of the credit card holders.

5. Copyright

All authors must sign the Transfer of Copyright agreement before the article can be published. This transfer agreement enables Elsevier to protect the copyrighted material for the authors, but does not relinquish the authors' proprietary rights. The copyright transfer covers the exclusive rights to reproduce and distribute the article, including reprints, photographic reproductions, microfilm or any other reproductions of similar nature and translations. Authors are responsible for obtaining from the copyright holder permission to reproduce any figures for which copyright exists.

Acknowledgements

This research was partially supported by Major Bidding Program of National Social Science Foundation of China (Grant No. 09&ZD041), Projects of the National Social Science Foundation of China (Grant No. 07BYJ110), Philosophies and Social Sciences Foundation of Hunan Province (Grant No. 2010YBB127, 08YBB282), The Open Fund Project of Key Research Institute of Philosophies and Social Sciences in Hunan Universities (Grant No. 10FEFM02), Scientific Research Foundation for Doctor (Grant No. E51022). Special thanks to the local bank for providing the credit card data that made this research possible.

References