A Study on Assessment Customer Defaults of Mortgage Loans

Yanyou Hao[1, 2], Hongyang Bao[2, 3], Zhongxian Chi[1] and Bo Wang[1]

[1] Department of Computer Science and Engineering,
Dalian University of Technology,
Dalian, 116024,China
[2] Dalian Branch of China Construction Bank,
Dalian, 116011,China
[3] School of Management, Dalian University of Technology,
Dalian, 116024,China
haoyanyou@yahoo.com.cn, haoyanyou@126.com

Abstract. Credit risk is the primary source of risk to financial institutions. As part of the credit risk assessment, the New Basel Accord suggests more granularity in risk-rating classes than currently exists. Credit scoring is one of important tools help financial institutions to hedge the credit risks. Support Vector Machine (SVM) is a new machine learning method based on the idea of VC dimension and Statistical Learning Theory (SLT). It is a good classifier to solve binary classification problem and the learning results possess stronger robustness. In this paper default prediction model of the housing mortgage loan is established by using SVM. We use grid-search method adjusts these penalty parameters to achieve better generalization performances in our application.

Keywords: Credit scoring; Credit risks; Support Vector Machine (SVM); Grid-search; Hoausing mortgage loan.

1 Introduction

Credit risk is the primary risk facing financial institutions. With the proposed guidelines under the New Basel Accord, financial institutions will benefit from better assessing their risks [1]. Credit risk is commonly defined as the loss resulting from failure of obligors to honor their payments. Arguably a cornerstone of credit risk modeling is the probability of default (PD) [2]. The housing mortgage is an important component of bank loan. However, there are huge credit risks that banks take back the principals and interest of loan with default of the consumers. Managers in the bank require the ability to predict the proportion of mortgages that will be defaulted. Many methods have been used to estimate default. Standard & Poors (S&P) and Moody have employed accounting analytic and migration analysis to predict the probability of default. Statistical methods for forecasting default risk include linear discriminant analysis [3], [4], the logistic regression model [5], [6]. The above methods typically require large data to build the forecasting model. However, there are not large data to use in real-life.

Advances in Artificial Intelligence.
Research in Computing Science 26, 2006, pp. 73-81

Received 02/06/06
Accepted 03/10/06
Final version 10/10/06

There is a credit scoring system for consumer mortgage loan application to produce an internal rating. It is a traditional approach based an empirical model, which takes into account various quantitative as well as subjective factors, such as the consumers' age, household income, interest rate, etc. Through this scoring system, analyzing all the pieces of information in your credit record and summarizing them in a number calculate a credit score of a consumer. A company named Fair, Isaac & Co. (FICO) developed a mathematical way to look at factors in your credit record that may affect your ability and willingness to repay a debt [7]. The problem with this approach is of course the subjective aspect of the prediction, which makes it difficult to make consistent estimates. The credit scoring problems can transform to classification than predict the probabilities of defaults. Recent researches have shown that Artificial Intelligence (AI) methods achieved better performance than traditional statistical methods [8], [9].

This paper applies SVM to estimate the mortgage default. In general, SVM has good generalization performance. However, there are some cases that the numbers of data in different classed are imbalance. The over-fitting of classifier affects the generalization performance of model. The kernel parameters γ and upper bound C control the generalization of SVM. Chang and Lin give a grid-search method to find the best parameter for SVM kernel and upper bound C [10]. We extend and use this method that adjusts these penalty parameters to achieve better generalization performances in our application.

This paper is organized as follows. Section 2 introduces the Support Vector Machine. In section 3, we described the methodology and procedure in assessment mortgage default, and defined default accuracy and normal accuracy to measure the performance of prediction model. In section 4, a real-life mortgage data set is to test the prediction model. Finally, we summarize the work in section 5.

2 Support Vector Machine and Parameter Selection

In this section, we give a brief review of SVM classification. SVM is a novel-learning machine first developed by Vapnik. It is based on the Structural Risk Minimization (SRM) principle from computational learning theory [11]. We consider a binary classification task with input variables $x_i (i = 1,...,l)$ having corresponding labels $y_i = \pm 1$. SVM finds the hyperplane to separate these two classes with a maximum margin. This is equivalent to solving the following optimization problem:

$$\text{Minimize: } \frac{1}{2} w^T \cdot w$$
$$\text{Subject to: } y_i (w \cdot x_i + b) \geq 1 \qquad (1)$$

By introducing Lagrange multiples α_i for the constraints in the (1), the problem can be transformed into its dual form

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j (x_i \cdot x_j) - \sum_{j=1}^{l} \alpha_j$$

$$\textit{Subject to}: \sum_{i=1}^{l} y_i \alpha_i, \qquad \alpha_i \geq 0, i = 1,...,l \tag{2}$$

Fig. 1 is a sample linearly separable case. Solid points and circle points represent two kind of sample separately. H is the separating hyperplane. H_1 and H_2 are two hyperplane through the closest points (the Support Vectors, SVs). The margin is the perpendicular distance between the separating hyperplane H_1 and H_2.

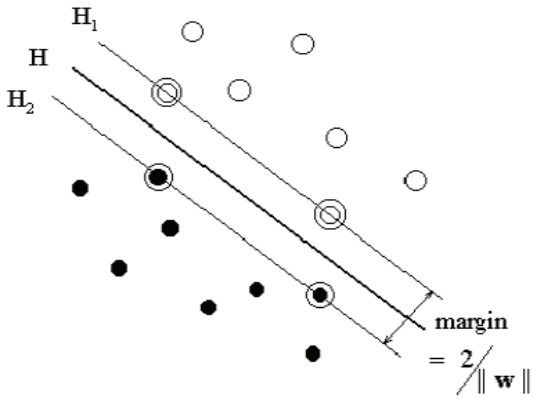


Fig. 1: Optimal Separation Hyperplane

To allow some training errors for generalization, slack variables ξ_i and penalty parameter *C* are introduced. The optimization problem is re-formulated as

$$\text{Minimize } \frac{1}{2} w^T \cdot w + C\sum_{i=1}^{l} \xi_i$$

$$\text{Subject to } y_i (w \cdot x_i + b) \geq 1 - \xi_i \tag{3}$$

The purpose of $C\sum_{i=1}^{l} \xi_i$ is to control the number of misclassified samples. The user chooses the parameter *C* so that a large *C* corresponds to assigning a higher penalty to errors term [12]. In addition, for some classification problems, numbers of data in different classes are imbalanced. The imbalance may come from the unequal proportion of samples between the different classes or the unequal density of the clusters in the feature space even if the populations are the same. Thus, a method using different penalty C_+ and C_- for each class to adjust the penalties on the false positive and false negative [13]. Finding this hyperplane can be translated into the following optimization problem:

$$\text{Minimize } \frac{1}{2}w^T \cdot w + C_+ \sum_{i:y_i=1} \xi_i + C_- \sum_{i:y_i=-1} \xi_i$$

$$\text{Subject to } \quad y_i(w \cdot x_i + b) \geq 1 - \xi_i \tag{4}$$

For nonlinear case, we map the input space into high dimension feature space by a nonlinear mapping. With a suitable choice of kernel the data can become separable in feature space despite being non-separable in the original input space. Here are three kinds of kernel function which are most commonly used:

Polynomial: $K_{poly}(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0.$

RBF: $K_{rbf}(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2), \gamma > 0.$

Sigmoid: $K_{sig}(x_i, x_j) = \tanh(\gamma x_i^T x_j + r).$

Here, γ, r and d are kernel parameters. Every kernel has its advantages and disadvantages [14]. The kernel type, kernel parameters and the penalty parameter C control the generalization of Support Vector Machines. The best choice of kernel of C depends on each other and the art of researcher.

3 Methodology

The estimation of mortgage defaults is a two-class classification task. Accuracy is the typical performance measure for two-class classification schemes. However, two learning algorithms can have the same accuracy, but the one which groups the errors near the decision border is the better one.

In order to appraise the performance of the classifier, Default Accuracy and Normal Accuracy are selected as standard criteria. We define the default accuracy and normal accuracy as:

$$\text{Default Accuracy} = \frac{\text{default samples classified}}{\text{total default samples}} \tag{5}$$

$$\text{Normal Accuracy} = \frac{\text{normal samples classified}}{\text{total normal samples}} \tag{6}$$

The advantage of the default accuracy and normal accuracy is that that they are a good indicator of whether the errors are close to the decision border or not. Given two classifiers with the same accuracy, the one with high default accuracy and normal accuracy is the better one. This definition is equivalent to the definitions of False Alarm and Miss Rate in [15].

Since there are three measures of SVM performance, searching for the optimal solution of SVM parameters is a multi-objective programming. We use RBF kernel to map the input space into high dimension feature space, and use different upper bounder C_+ and C_- for different classes.

The grid-search method use cross-validation on C and γ. Basically pairs of (C, γ) are tried and the one with the best cross-validation accuracy is picked. The

grid-search is straight forward but seems stupid. However, there are two motivations why we prefer the simple grid-search approach. One is that psychologically we may not feel safe to use methods which avoid doing an exhaustive parameter search by approximations or heuristics. The other reason is that the computational time to find good parameters by grid-search is not much more than those by advanced methods since there are only two parameters. Furthermore, the grid-search can be easily parallelized because each (C, γ) is independent. The grid-search method in [10] just is used in training data sets; we extend to use the method in training and test data sets. The best parameters not only are of the training data sets but also of test data sets.

4 Experiments Results and Analysis

The available default estimation data set in housing mortgage loan is provided by a major commercial bank of China. This data set contains 18960 samples from January 1998 to December 2004. We defined two classes: "good" and "bad" customers. The "bad" customer is the borrower with at least one defaulted instalment, or more than 3, the one that did not pay one instalment over a period of three months. The "good" customer is the borrower who repayment on time. The data is typically from a sample of applicants who have been granted credit already. The data is imbalanced; the "bad" class is rare class.

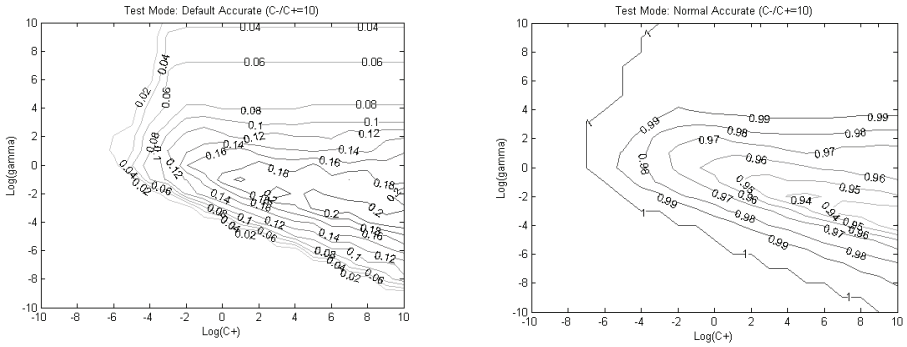
All customers are the bank's consumer who had applied to the bank for mortgage loan. The Data sets consisted of the customer information and the other information of the loan application form. There are 3 categorical attributes and 11 numerical attributes. Application characteristics available in the data set are summarized in Table 1. The "good" customers are labeled "1" and the "bad" customers are labeled "–1".

Table 1. The input variables

Index	Indicators
1	Customer Age
2	Educational Level
3	Vocation
4	Working at Industry
5	Years at current work
6	Household income
7	Price of the House
8	Area of House
9	House value at purchase
10	Monthly Payment
11	First Payment Ratio
12	Amount of Contract
13	Loan Terms
14	Balance of Loan

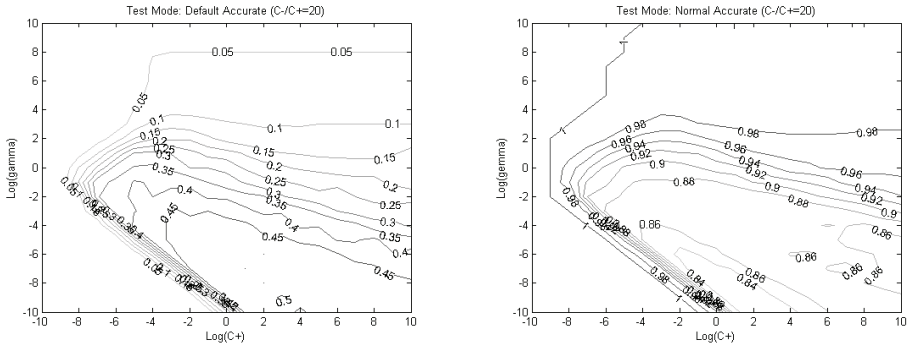
We spitted the data into two sub-datasets evenly, one half for training and the other half for testing, this is referred to as 50-50 split experiment. The training set includes 9480 samples with 241 "bad" customers and 9239 "good" customers. The test set includes 9480 samples with 240 "bad" customer and 9240 "good" customers. The data set is imbalanced between the "bad" and "good" class.

Our SVM for default prediction training code is a modification of LIBSVM [16]. The training tools of LIBVM can set different weight on the parameter *C* of two classes. Since doing a complete grid-search is a time-consuming work, so we parallel do this experiments on 3 PC servers at same time. We give the range of *C* and γ in $2^{-10} \sim 2^{10}$ and $C_- / C_+ = \{10, 20, 30\}$. We get 3 group experiment results and there are 421 records in every group. The model' performance of generalization attracts our more attentions. The Fig. 2 to Fig. 3 show a part of the experiment results.



(a) (b)

Figure 2. Where C_- / C_+=10, test mode: (a) the highest Default Accuracy is 0.20; (b) the lowest Normal Accuracy is 0.95.



(c) (d)

Figure 3. Where C_- / C_+=20, test mode: (c) the highest Default Accuracy is 0.50; (b) the lowest Normal Accuracy is 0.84.

Table 2 shows a group of detail results in training sets and test sets.

The predicting total accuracy is around 70% both on training and test phases especially a similar results in Default Accuracy and Normal Accuracy where C_- / C_+ = 30 and Log (C_+) = 0 and Log (γ) = -6. We consider that the model is under-fitting on the training set with those parameters. When the penalty parameter *C* becomes too small the error term impact will be decrease on the training process. In general, the real life application is one of nonlinear-classifiable case; we need a relatively big penalty parameter *C* in fact.

Table 2. The grid-search results

C_- / C_+	Log (C_+)	Log (γ)	Training sets			Test sets		
			Total Accuracy	Default Accuracy	Normal Accuracy	Total Accuracy	Default Accuracy	Normal Accuracy
10	**10**	**-2**	**98%**	**95%**	**98%**	**93%**	**22%**	**95%**
10	6	-3	95%	63%	96%	92%	**22%**	94%
10	10	-4	95%	70%	96%	91%	**22%**	93%
10	7	-3	95%	71%	96%	92%	**21%**	93%
10	9	-4	95%	65%	96%	92%	**21%**	94%
10	8	-3	95%	77%	96%	91%	**21%**	93%
20	**3**	**-9**	**84%**	**54%**	**85%**	**83%**	**50%**	**84%**
20	1	-7	84%	53%	85%	83%	**50%**	84%
20	2	-8	84%	54%	85%	83%	**50%**	84%
20	2	-6	85%	61%	86%	86%	**50%**	87%
20	0	-5	86%	57%	87%	86%	**50%**	87%
20	8	-10	85%	56%	85%	85%	**50%**	86%
30	**0**	**-6**	**74%**	**76%**	**74%**	**70%**	**69%**	**70%**
30	2	-8	73%	76%	73%	68%	**69%**	68%
30	3	-8	74%	74%	74%	71%	**69%**	71%
30	0	-7	73%	74%	73%	67%	**68%**	67%
30	1	-7	74%	76%	73%	69%	**68%**	69%
30	-3	-4	75%	74%	75%	70%	**68%**	70%
30	-1	-6	73%	73%	73%	68%	**68%**	68%

The predicting total accuracy is above 90% both in training sets and test sets where C_- / C_+ = 10 and Log (C_+) = 10 and Log (γ) = -2, but the Default Accuracy is very poor in test phase. Obviously we must to provide more detailed analysis why the results are so poor. We discuss the problem with those experts of the commercial bank. They give some explanations about the imbalance data set. Due to the housing

mortgage loan belong to a long-term loan; the period of repayment of credit is between 5 and 30 years in general. In China the loans are developed just before few years ago. The default period is not coming so soon. There is a reason for the poor result that credit information of customers is incomplete and some dirty data in the data sets.

We compare the respectively-penalty SVM (RP-SVM) in this paper with classic SVM and other machine learning algorithms. We select the classic SVM, BP Neural Network and Decision Tree algorithms to training the same data set. Then we predict the same test data set using their classification model respectively. Table 3 shows the comparison of predicting result of those algorithms.

Table 3. The comparison of predicting result

Classification Algorithm	Normal Accuracy	Default Accuracy	Total Accuracy
RP-SVM	95.31%	22.45%	93.47%
Classic SVM	95.60%	16.63%	93.60%
BP Neural Network	95.10%	12.05%	92.50%
Decision Tree	95.29%	14.96%	93.26%

It is obvious in the Table 3 that Default Accuracy of RP-SVM is better than other methods under similar total predicting accuracy, which the predicting accuracy can increase just 5%. Our future direction of the research would focus on how to improve the Default Accuracy especially in the testing data set.

5 Conclusions

As we have shown, SVM are capable of estimation defaults from real life mortgage data. We can adjust just few parameters to obtain the results not very obvious at first glance. This makes SVM particularly well suited as an underlying technique for credit rating and risk management methods applied by financial institution. Further research focus on how to improve the Default Accuracy and Normal Accuracy in the test data set. We believe that deeper data preprocessing and more suitable parameters selection will contribute to improve the performance of generalization. Extending the two-class classification to multi-class classification and introducing fuzzy SVM are also our future research work.

References

1. Allen M. Featherstone, Laura M. Roessler,and Peter J. Barry.: Determining the Probability of Default and Risk-Rating Class for Loans in the Seventh Farm Credit District Portfolio. Review of Agricultural Economics, Volume 28 Page 4 - March 2006
2. Til Schuermann and Samuel Hanson.: Estimating Probabilities of Default. Federal Reserve Bank of New York Staff Reports, no. 190, July 2004
3. Altman, E. I. (1968).: Financial Ratio Discriminant Analysis and the Prediction of Corporate Bankruptcy. Journal of Finance, 23(3), 589-609
4. Altman, E. I., Marco, G., and Varetto, F. (1994).: Corporate Distress Diagnosis: Comparisons Using Linear Discriminant Analysis and Neural Networks (The Italian Experience). Journal of Banking and Finance, 18, 505-529
5. Martin, D. (1977).: Early warring of Bank Failure: A Logic Regression Approach. Journal of Banking and Finance, 1, 249-276
6. Smith, L. D., and Lawrence, E. C. (1995).: Forecasting Losses on a Liquidating Long-term Loan Portfolio. Journal of Banking and Finance, 19, 959-985
7. www.fairisaac.com
8. Hand D.J. and Henley W.E. (1997): Statistical Classification Methods in Consumer Credit Scoring: a Review. Journal of the Royal Statistical Society, A, 160, 523-541
9. Zan Huang and Hsinchun Chen.: Credit rating analysis with support vector machines and neural networks: a market comparative study. Decision Support Systems 37(2004) 543-558
10. Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin.: A Practical Guide to Support Vector Classification. Technical Report, National Taiwan University, Taiwan (2001)
11. C. Cortes and V. Vapnik.: Support-vector networks, Machine Learning. vol 20, no.3, 273-297, 1995
12. Osuna E., Freund R., Girosi F.: Support vector machines: Training and applications. Massachusetts Institute of Technology, AI Memo No. 1602. 1997
13. K. Morik, P. Brockhausen, and T. Joachims.: Combining statistical learning with a knowledge-based approach, a case study in intensive care monitoring, in Proc. 16th International Conference on Machine Learning, Bled, Slovenia, 1999
14. Smits, G.F., Jordaan, E. M. Improved svm regression using mixtures of kernels. Proceedings of the 2002 International Joint Conference on Neural Networks, Vol. 3, 2002, Pages:2785-2790
15. Drucker H., Wu D., V. Vapnik.: Support Vector Machines for Spam Categorization. IEEE Transactions on Neural Networks. Vol. 10, Issue 5, Sep 1999
16. Chih-Chung Chang, Chih-Jen Lin.: LIBSVM: a Library for Support Vector Machines. URL: http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf