Credit Scoring Analysisi

by Tatik Widiharih

Submission date: 25-Jan-2020 06:09PM (UTC+0700)
Submission ID: 1246225033
File name: B316.pdf (503.48K)
Word count: 3015
Character count: 15726

Credit scoring analysis using kernel discriminant

IOP Conf. Series: Journal of Physics: Conf. Series **1025** (2018) 012124 doi:10.1088/1742-6596/1025/1/012124

Credit scoring analysis using kernel discriminant

T Widiharih, M A Mukid, Mustafid

Department of Statistics, Diponegoro University, Semarang, Indonesia

E-mail: widiharih@gmail.com

Abstract. Credit scoring model is an important tool for reducing the risk of wrong decisions when granting credit facilities to applicants. This paper investigate the performance of kernel discriminant model in assessing customer credit risk. Kernel discriminant analysis is a non-parametric method which means that it does not require any assumptions about the probability distribution of the input. The main ingredient is a kernel that allows an efficient computation of Fisher discriminant. We use several kernel such as normal, epanechnikov, biweight, and triweight. The models accuracy was compared each other using data from a financial institution in Indonesia. The results show that kernel discriminant can be an alternative method that can be used to determine who is eligible for a credit loan. In the data we use, it shows that a normal kernel is relevant to be selected for credit scoring using kernel discriminant model. Sensitivity and specificity reach to 0.5556 and 0.5488 respectively.

1. Introduction

Effective credit risk assessment has become a very important factor for obtaining many advantages in credit market which can help financial institutions to grant credit to credit worthy customers and reject non-creditworthy customers [1]. The decision makers need some help to decide whether to grant credit or not for a credit applicant from some efficient and feasible tools [2]. Credit scoring is the most widely used techniques that help them to make credit granting decision. Credit scoring models play an important role in contemporary risk management practice. They contribute to the key requirement in loan approval process, which is to accurately and efficiently quantify the level of credit risk associated with a customer [3]. Credit scoring models help credit institutions evaluate credit applications with respect to customer characteristics such as age, income, and marital status [4]. Technically, credit scoring models classify loan clients to either good credit or bad credit [5].

A wide range of classification techniques have already been proposed in the credit scoring literature, including statistical methods, such as linear discriminant analysis and logistic regression and non-parametric models, such as decision trees, k-nearest neighbor, and nonparametric discriminant [6]. These models are categorized into parametric and non-parametric or data mining models [1]. Briefly, a parametric model presumes that the form of the model is known except for finitely many unknown parameters, whereas a non-parametric models only assumes that the model belongs to some infinite dimensional collection of functions [7]. Generally, linear discriminant analysis and logistic regression are categorized into parametric model and decision trees, k-nearest neighbor, and kernel discriminant are classified into nonparametric one.

In a discriminant analysis, the optimal Bayes rule is used to assign an object to the class with the largest posterior probability. This probability is the product of the prior and the density function of the input. But, the density functions are usually unknown in practice, and can be estimated from the training data set either parametrically or nonparametrically. In parametric approaches, the underlying

population distributions are assumed to be known except for some unknown parameters. Consequently, the performance of a parametric discrimination rule largely depends on the validity of those parametric models [8]. Nonparametric classification techniques, however, are more flexible in nature and free from such parametric model assumptions. Kernel density estimation is a famous method for constructing nonparametric estimates of population densities. The use of kernel density estimates in discriminant analysis is quite popular in the existing literature ([9], [10], [11], [12]). The application of kernel methods for scoring credit scoreing analysis is still rare. We are interested in using this method to classify credit customers based on the characteristics of the previous borrower.

The rest of this paper is organized as follows. In Section 2, we give a brief overview of kernel discriminant. The data used in this paper describe in Section 3. An empirical study of a credit from a financial institution and its results are presented in Section 4. Finally, conclusions are offered in Section 5.

2. Kernel Discriminant Analysis

This study aims to apply the kernel discriminant method for the purpose of classification of a prospective borrower as a good or bad borrower. Let $\mathbf{x}_1,\ldots.\mathbf{x}_{n_t}$ is a random sample from population Π_t , and $\mathbf{x}$ is an observation from population Π_t which has an unknown probability density function $f_t(\mathrm{x})$. In this paper $f_t(\mathrm{x})$ was estimated by $\hat{f}_t(\mathbf{x}) = \frac{1}{n_t}\sum_{i=1}^{n_t} K_t(\mathrm{x}-\mathrm{x}_i)$, where $K_t(\mathbf{x})$ is a function defined by $\mathbf{x}$, a vector d dimensions. $K_t(\mathbf{x})$ is called a kernel function of the population t [13]. Let $\mathrm{z} = \mathrm{x} - \mathrm{x}_i$, h is bandwidth value, d is the number of explanatory variables, $\mathbf{V}_t$ represent a covariance matrix of population t, and $t = 1, 2, \ldots, g$. Below are some kernels that frequently used:

a. Kernel Normal (mean $\mathbf{0}$, varian $h^2 \cdot \mathbf{V}_t$)

$$K_t(\mathbf{z}) = \frac{1}{c_{0(t)}}\exp\left(-\mathbf{0.5}\mathbf{z}^{\mathrm{T}}\mathbf{V}_t^{-1}\mathbf{z}/h^2\right)$$

where $c_{0(t)} = (2\pi)^{\frac{d}{2}} h^d |\mathbf{V}_t|^{\frac{1}{2}}$

b. Kernel Epanechnikov

$$K_t(\mathbf{z}) = \begin{cases} c_{1(t)}\left(1-\mathbf{z}^{\mathrm{T}}\mathbf{V}_t^{-1}\mathbf{z}/h^2\right), & \text{if } \mathbf{z}^{\mathrm{T}}\mathbf{V}_t^{-1}\mathbf{z} \le h^2 \\ 0 \quad , & \text{elsewhere} \end{cases}$$

where $c_{1(t)} = \left(1+\frac{d}{2}\right)v_{h(t)}$

$$v_{h(t)} = \frac{\pi^{\frac{d}{2}} h^d}{\Gamma\left(\frac{d}{2}+1\right)}|\mathbf{V}_t|^{\frac{1}{2}}$$

c. Kernel Biweight

$$K_t(z) = \begin{cases} c_2(t)\left(1-\mathbf{z}'\mathbf{V}_t\mathbf{z}/h^2\right)^2, & \text{if } \mathbf{z}'\mathbf{V}_t\mathbf{z} \le h^2 \\ 0 \quad , & \text{elsewhere} \end{cases}$$

where $c_2(t) = \left(1+\frac{d}{4}\right)c_1(t)$

d. Kernel Triweight

$$K_t(z) = \begin{cases} c_3(t)\left(1 - \mathbf{z}'\mathbf{V}_t\mathbf{z}/h^2\right)^3, & \text{if } \mathbf{z}'\mathbf{V}_t\mathbf{z} \le h^2 \\ 0, & \text{elsewhere} \end{cases}$$

where $c_3(t) = \left(1 + \frac{d}{6}\right)c_2(t)$

An observation will be classified to population t if the posterior probability value in that population is greatest when compared to the posterior probability value in the other populations. The posterior probability of an observation $\mathbf{x}$ in population t is $P(\Pi_t \mid \mathbf{x}) = \frac{p_t \hat{f}_t(\mathbf{x})}{\sum_{t=1}^{g} p_t \hat{f}_t(\mathbf{x})}$, where p_t is the prior probability which is defined by $p_t = \frac{n_t}{\sum_{t=1}^{g} n_t}$.

3. Data And Methods

The data used in this paper come from a financial institution in Indonesia consisting 2075 clients of which 409 clients are categorized as bad customer which are debtors in July 2017. The financial institution disburse a loans for purchasing a motorcycles. For the data set, a bad customer was defined as someone who had missed three consecutive months of payments. The data involve 8 continuous explanatory variables including amount principal, working experince, total income, price, down payment, installment, long repayment, and rate. The description of each variables and its unit of measure is shown in Table 1.

Table 1. Variables used for building The credit scoring model

Variable	Definition
Amount principal	Amount principal of applicant in Rupiahs
Total income	Monthly income in Rupiahs
Working experience	Working experience of the applicant in years
Price	Price of motorcycle
Down payment	Down payment for the purchase of motorcycles in Rupiahs
Installment	Installment in Rupiahs
Long repayment	Long repayment in month
Rate	Rate in percent per month

For this empirical study, we split the data into training set consisting of 80% and testing set for the rest. After specifying a kernel and executing the algorithm of the method, the accuracy of the method was calculated using sensitivity, specificity, percentage of correctly classified (PCC), and apparent error rate (APER). The formulas of the four measures of accuracy refer to the paper of Zhou, Lai, and Yu [1].

$$\text{Sensitivity} = \frac{GG}{GG + GB}$$

$$\text{Specificity} = \frac{BB}{BB + BG}$$

$$\text{PCC} = \frac{GG + BB}{GG + GB + BB + BG}$$

$$\text{APER} = \frac{GB + BG}{GG + GB + BB + BG}$$

where GG reperesent the number of good customers that were classified as good by the classifier; GB represent the number of good customer that were mistakenly classified as bad; BB represent the number of correctly classified customer that belongs to class bad; BG represent the number of observed bad customer that were were mistakenly classified as good. Good models should have high value on both sensitivity and specificity. But it is often not fulfilled, because there is a trade off between the values of both, when a model has a high sensitivity, it usually has low specificity. We chose a model whose sensitivity and specificity values did not differ large.

4. Results And Discussions

Tables 2 is statistical summaries of each good and bad customers based on 8 explanatory variables. The tables tell that the average of each variable on good customers shows tend to be better than bad customers. For example, the average of total monthly income of good customers reaches Rp 3785948.379 which is greater than the total revenue of bad customers whose value only reaches Rp 3390931.540. Furthermore, the mean of down payment from good customers reach Rp 3656439.226 which is greater than bad customers. Similar condition happen to other variables that indicate good customers has a positif individual performance. Tables 2 and 3 also show that there were a different variations in the independent variables involved in the analysis on both types of customers. Overall, variability of variables from good customers is greater than bad ones.

Table 2. Statistical summary of both good and bad customer

Variables	Good Customers		Bad Customers	
	Mean	Standart Deviation	Mean	Standart Deviation
Amount principal	9268090.09	5379189.651	8609974.98	4799958.397
Total income	3785948.379	3302309.184	3390931.54	3126797.906
Working experience	8.659	7.426	7.71	6.843
Price	11993838.73	6589531.057	10704745.36	5633475.18
Down payment	3656439.226	2864128.995	3029413.203	1823309.02
Installment	635646.459	458660.991	563264.059	241175.733
Long repayment	20.993	10.863	21.191	9.963
Rate	21.367	5.759	21.913	5.58

In this paper we examine the performance of several kernels including normal, epanechnikov, biweight, and triweight in classifying credit status of customers. The main parameter of a kernel is a smoothing parameter or sometimes called a bandwidth. Each kernel has an optimal performance at a certain bandwidth value. In order to know the optimal bandwidth, we do by trying some values at a certain interval. Computations of nonparameteric discrimination using kernel were conducted using SAS. A good classifier should has a high value on both sensitivity and specificity.

Table 3 shows the performance of nonparametric discriminant using kernel Normal in assessing credit worthines. We tried some bandwidth values at intervals (0,1]. It appears that at a bandwidth value of 0.4, nonparamateric discriminant models with kernel Normal produce a sensitivity and specificity of 0.5556 and 0.5488 respectively. It means that customers whose credit status is good, by the model are classified as good reach to 55.56% while customers whose credit status is bad, by the model are classified as bad reach to 54.88%. The PCC and APER at bandwidth 0.4 respectively 0.5542 and 0.4458. For bandwidth values greater than 0.4 will produce a sensitivity greater than

0.5556 but the specificity value is lower than 0.5488. We did not try a bandwidth value greater than 1, because it would widen the gap between the sensitivity and specificity of the model.

Table 3. Accuration of nonparametric discriminant using kernel normal

Bandwdith	Sensitivity	Specificity	PCC	APER
1.0	0.2492	0.8659	0.3711	0.6289
0.6	0.4565	0.6829	0.5012	0.4988
0.5	0.5015	0.6341	0.5277	0.4723
0.4	0.5556	0.5488	0.5542	0.4458
0.3	0.6637	0.5000	0.6313	0.3687
0.2	0.7958	0.3537	0.7084	0.2916
0.1	0.8739	0.1951	0.7398	0.2602

Tables 4 describe the performance of Epanechnikov, biweight, and triweight kernels on bandwidth values at intervals [4,9]. At the interval, the kernel discriminant model shows unsatisfactory performance. For those kernels with bandwidth values at intervals [4,9] yields very low sensitivity and high specificity. This means that the three kernel functions on the bandwidth [4,9] are incapable of predicting a good consumer in paying credit as a good consumer according to the kernel discriminant model. In contrast, the kernel discriminant model using the epanechnikov, biweight, and triweight kernels has a high specificity value which indicates that it is able to predict poor consumers in paying credit as a bad consumer as well. Furthermore, the combination of kernel types and the bandwidth values also result in low PCC and high APER values. We did not try a bandwidth value that is less than 4, because it produces an unidentifiable classification of whether the customer is in good or bad credit status. We also did not try a bandwidth wich is larger than 9, because the difference between sensitivity and specificity is getting bigger.

Table 4. Accuration of nonparametric discriminant using kernel epanechnikov, biweight, and triweight

	Bandwidth	4	5	6	7	8	9
Epanechnikov	Sensitivity	0.1652	0.1201	0.0781	0.0601	0.039	0.042
	Specificity	0.9146	0.939	0.9512	0.9634	0.9756	0.9756
	PCC	0.3133	0.2819	0.2506	0.2386	0.2241	0.2265
	APER	0.6867	0.7181	0.7494	0.7614	0.7759	0.7735
Biweight	Sensitivity	0.1862	0.1291	0.0961	0.0751	0.0511	0.042
	Specificity	0.878	0.9268	0.939	0.9512	0.9878	0.9756
	PCC	0.3229	0.2867	0.2627	0.2482	0.2361	0.2265
	APER	0.6771	0.7133	0.7373	0.7518	0.7639	0.7735
Triweight	Sensitivity	0.2102	0.1622	0.1141	0.0841	0.0691	0.0511
	Specificity	0.8902	0.9268	0.9268	0.939	0.9634	0.9878
	PCC	0.3446	0.2529	0.2747	0.253	0.2458	0.2361
	APPER	0.6554	0.7471	0.7253	0.747	0.7542	0.7639

5. Conclussions

Credit scoring has become an important task as financial industries can increase their benefits. This paper apply kernel discriminant model for evaluating credit applicants worthiness. We investigate the performance of several kernels in order to find a nonparametric discriminant model which good accuration in classification. In our case, the results show that kernel discriminant can be an alternative

method that can be used to determine who is eligible for a credit loan. In the data we use, it shows that a normal kernel is relevant to be selected for credit scoring using kernel discriminant model. Sensitivity and specificity reach to 0.5556 and 0.5488 respectively.This such model is very required by a financial institutions for reducing the risk of wrong decisions when granting credit facilities. The objective of this model is to separate a credit applicants who is eligible and who isn't.

Acknoledgment
We would like to thanks to research, technology, and high education ministry of Republic of Indonesia for the financial support based on contract number 007/SP2H/LT/DRPM/IV/2017.

References

[1] L. Zhou, K. K. Lai, and L. Yu, "Least squares support vector machines ensemble models for credit scoring" *Expert Sytems with Applications*, vol. 37, pp. 127-133, 2010.
[2] D. Zhang, X. Zhou, S. C. H. Leung, and J. Zheng, "Vertical bagging decisions trees model for credit scoring" *Expert Sytems with Applications*, vol. 37, pp. 7838-7843, 2010.
[3] N. Nicolic, N. Zarkic-Joksimovic, D. Stojanovski, and I. Joksimovic, "The application of brute force logistic regression to corporate credit scoring models: Evidence from Seerbian financial statements", *Expert Sytems with Applications*, vol. 40, pp. 5932-5944, 2013.
[4] M. C. Chen, and S. H. Huang, "Credit scoring and rejected instances reassigning through evolutionary computation techniques", *Expert Sytems with Applications*, vol. 24, pp. 433-441, 2003.
[5] F. Louzada, P. H. Ferreira-Silva, and C. A. R. Diniz, "On the impact of disproportional samples in credit scoring models: An application to a Brazilian band data", *Expert Sytems with Applications*, vol. 39, pp. 8071-8078. 2012.
[6] I. Brown, and C. Mues, "An experimental comparison of classification algorithm for imbalanced credit scoring data set", *Expert Sytems with Applications*, vol. 39, pp. 3446-3453, 2012.
[7] R. L. Eubank, Nonparametric regression and spline smoothing, Maercel Dekker Inc, New York, 1988.
[8] A. K. Gosh, and P. Chaudhuri, "Optimal smoothing in kernel discriminant analysis", *Statistica Sinica*, vol. 14, pp. 457-483, 2004.
[9] D. J. Hand, Kernel Discriminant Analysis, Wiley, Chichester, 1982.
[10] D. Coomans, and I. Broeckaert, Potential Pattern Recognition in Chemical and Medical Decision Making. Research Studies Press, Letchworth, 1986.
[11] B. W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman and Hall, London, 1986.
[12] P. Hall, and M. P. Wand, "On nonparametric discrimination using density differences", *Biometrika* vol. 75, pp. 541-547, 1988.
[13] R. Khattree, and D. N. Naik, Multivariate Data Reduction and Discrimination with SAS® Software, Cary, NC: SAS Institute. Inc, 2000.

Credit Scoring Analysisi

ORIGINALITY REPORT

15%	10%	12%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Haojie Chen, Minghui Jiang, Xue Wang. "Bayesian ensemble assessment for credit scoring", 2017 4th International Conference on Industrial Economics System and Industrial Security Engineering (IEIS), 2017 Publication	1%
2	www.thesai.org Internet Source	1%
3	www.thinkmind.org Internet Source	1%
4	Anil K Ghosh, Probal Chaudhuri, Debasis Sengupta. "Classification Using Kernel Density Estimates", Technometrics, 2006 Publication	1%
5	Hong-Liang Dai. "Imbalanced Protein Data Classification Using Ensemble FTM-SVM", IEEE Transactions on NanoBioscience, 2015 Publication	1%
6	Anil K. Ghosh. "Kernel Discriminant Analysis Using Case-Specific Smoothing Parameters",	1%

IEEE Transactions on Systems Man and Cybernetics Part B (Cybernetics), 2008
Publication

7	U Bhuvaneswari, P. James Daniel Paul, Siddhant Sahu. "Financial risk modelling in vehicle credit portfolio", 2014 International Conference on Data Mining and Intelligent Computing (ICDMIC), 2014 Publication	1%
8	Dinh, T.H.T.. "A credit scoring model for Vietnam's retail banking market", International Review of Financial Analysis, 2007 Publication	1%
9	Silvia Botelho. "Prediction of Protein Secondary Structure Using Nonlinear Method", Lecture Notes in Computer Science, 2006 Publication	1%
10	M.Sagrario Sánchez, Luis A. Sarabia. "GINN (Genetic Inside Neural Network): towards a non-parametric training", Analytica Chimica Acta, 1997 Publication	1%
11	www.mcduplessis.com Internet Source	1%
12	www.inf.unibz.it Internet Source	1%

13	ANIL KUMAR GHOSH, SMARAJIT BOSE. "FEATURE EXTRACTION FOR CLASSIFICATION USING STATISTICAL NETWORKS", International Journal of Pattern Recognition and Artificial Intelligence, 2011 Publication	1%
14	etds.lib.ncku.edu.tw Internet Source	1%
15	www.marceljirina.cz Internet Source	1%
16	Submitted to Florida Virtual School Student Paper	<1%
17	Submitted to University of Wolverhampton Student Paper	<1%
18	www.wikicoursenote.com Internet Source	<1%
19	Xu, X.. "Credit scoring algorithm based on link analysis ranking with support vector machine", Expert Systems With Applications, 200903 Publication	<1%
20	bulletin.pan.pl Internet Source	<1%
21	Submitted to University of Louisiana, Lafayette Student Paper	<1%

22	Submitted to Curtin University of Technology Student Paper	<1%
23	imtc.ctgroup.co.in Internet Source	<1%
24	www.yumpu.com Internet Source	<1%
25	usir.salford.ac.uk Internet Source	<1%
26	WEI-WEN WU. "IMPROVING CLASSIFICATION ACCURACY AND CAUSAL KNOWLEDGE FOR BETTER CREDIT DECISIONS", International Journal of Neural Systems, 2011 Publication	<1%
27	"Emerging Intelligent Computing Technology and Applications", Springer Science and Business Media LLC, 2013 Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		

Credit Scoring Analysisi

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7