Computer Science and Application 计算机科学与应用, 2017, 7(12), 1278-1286
Published Online December 2017 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2017.712143

Hans汉斯

The Credit Scoring Model for Individuals Based on Cost-Sensitive SVM Model

Jiajun Shu

School of Mathematics and Statistics, Beijing Institute of Technology, Beijing
Email: shujiajunxx@126.com

Received: Dec. 7th, 2017; accepted: Dec. 20th, 2017; published: Dec. 28th, 2017

Abstract

In this paper, we mainly focus on the influence of the amount of credit in personal credit scoring. At first, we divided the customers into two parts based on their credit amount, and give different weights to the wrong-judging of different parts of customers; Then we used the cost-sensitive SVM model to establish a new personal credit evaluation model. Results suggest that this new model can decrease the total cost of personal credit evaluation, and improve the accuracy of judgment of those large customers who will break the contract.

Keywords

Credit Scoring Model for Individuals, Amount of Personal Credit, Cost-Sensitive, Support Vector Machine

基于代价敏感SVM的个人信用评估模型

束加俊

北京理工大学，数学与统计学院，北京
Email: shujiajunxx@126.com

收稿日期：2017年12月7日；录用日期：2017年12月20日；发布日期：2017年12月28日

摘　要

本文主要研究了客户借贷规模在个人信用评估中的影响，首先按照客户借贷规模对所有借贷客户进行分类，对不同类别的客户的错判所导致的损失赋予不同的权重；然后使用代价敏感支持向量机，构建了一种新的个人信贷评估模型。实证检验说明，此模型可以降低信贷公司在个人信用评估过程中的总损失，同时提高对借贷规模相对较大的客户的判别准确率。

文章引用: 束加俊. 基于代价敏感 SVM 的个人信用评估模型[J]. 计算机科学与应用, 2017, 7(12): 1278-1286.
DOI: 10.12677/csa.2017.712143

关键词

个人信用评估模型，客户借贷规模，代价敏感，支持向量机

1. 引言

信用风险是信贷机构在日常工作中需要规避的主要风险之一，主要是指借贷用户在借款之后未能按时还款给机构带来损失的风险。为了减小这个风险，借贷机构需要有效的把借贷用户按照是否会违约分成两部分，建立个人信用评估模型就是常用的方法之一。

个人信用评估模型建立的目标是极小化对已有的贷款客户样本错误分类个数。用这个模型判断一个新的贷款申请者是否会违约，根据判断的结果决定是否对本次贷款申请者的申请授权。国外的个人信用评估模型建立已经非常成熟，早期主要有依靠经验判断的专家打分法，1941 年杜兰特·戴维首次将数学方法运用到个人信用评估中，使用线性判别判定个人信贷风险。之后如 Z-score、ZETA 等判别分析法也相继被运用到个人信用评估模型建立中。近几年随着大数据和数据挖掘技术的发展，国内外学者也会使用非参数统计和机器学习来构建个人信用评估模型。如宋云鹏等人梳理了数据挖掘的操作流程和关键点以及常见的技术问题[1]，姜明辉等人对个人信用评分的主要模型及其发展进行了归纳[2]。沈翠华等人把支持向量机算法运用在个人信用评估中并和 K 近邻法相比，说明了支持向量机可以得到更好的结果[3]，肖文兵等人在使用支持向量机建立个人征信模型时使用网格 5-折交叉确认来寻找不同核函数的最优参数[4]，陆爱国等人首次提出基于三变量的 SVM 学习算法，并将其应用到个人信用评分模型中[5]。

在这个过程中，学者发现不同的错判方式给信贷机构带来不同损失，因此在模型的建立中，需要引入“代价敏感”的概念，如 2015 年段薇基于“‘把会违约的客户判为不会违约的客户’这种错误判断比‘把不会违约的客户判为会违约的客户’这种错误判断带来的损失更高”这个假设，提出一种利用熵值法来建立代价敏感支持向量机的方法[6]。

传统的代价敏感信用评估模型都是只考虑客户的待划分类别(是否违约)对误判的敏感性不同。但是在对个人信用进行评估时，信贷机构会面临一个非常重要的问题：客户借贷规模，例如一位借款 100 万元的客户和一位借款 1 万元的客户，其违约与否给信贷机构带来的收益或损失是有巨大差别的，因此在建立个人信用评估模型时需要考虑客户借贷规模这一因素。但如果还是仅仅按照最小化错判个数和对是否违约代价敏感来建立模型，无法满足信贷机构的上述需求。

本文主要针对不同借贷规模的客户对于公司的影响不同这一实际问题，对不同借贷规模的客户进行分类，同时考虑客户借贷规模和违约与否这两个因素对于误判的敏感性，使用代价敏感的支持向量机建立个人信用评估模型，对客户进行分类判别，以此来降低借贷公司在客户信用评估方面的总损失。

2. 支持向量机

支持向量机(Support Vector Machine, SVM)是数据挖掘领域一种常用的监督学习模型，主要用于二分类问题，其基本定义是在特征空间中使得不同类别之间间隔最大的线性分类器，由于核技巧的引入，使得它也可以成为非线性分类器。

假设特征空间上的数据集为 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，$N$ 为样本点的个数，x_i 为第 i 个样本点的特征向量，y_i 为第 i 个样本点的类别，对于二分类问题，假设 $y_i\in\{+1,-1\}$，为方便描述，下文中我们把 $y_i=+1$ 的样本点称为正类样本，把 $y_i=-1$ 的样本点称为负类样本。

支持向量机的本质是一个分割特征空间中样本点的超平面，模型建立的过程可以转化成对超平面参数的求解过程。当数据集上的样本点线性可分时，支持向量机的目标就是找到一个超平面将两类样本点分开，并且使它们的间距最大。此种方法对数据要求过高(完全线性可分)，因此在模型实际建立过程中，需要适当放松条件，允许模型存在误分类的情况。此时需引入松弛变量 $\xi_i\geq 0$ 和惩罚因子 C。惩罚因子 C 表示对分类错误的惩罚程度，C 值越大说明对误分类的惩罚越大，C 值通常被用来平衡模型复杂度和误判损失之间的关系。相应的模型如下式[7] [8]：

$$\begin{aligned}&\min_{w,b,\xi}\ \frac{1}{2}\|w\|^2+C\sum_{i=1}^{N}\xi_i\\&\text{s.t.}\quad y_i(w\cdot x_i+b)\geq 1-\xi_i,\quad i=1,2,\cdots,N\\&\qquad \xi_i\geq 0,\quad i=1,2,\cdots,N\end{aligned}\tag{1}$$

在模型建立过程中，当发现对于待判别类型的误判方式不同，误判带来的损失也有巨大差别时，就需要引入“代价敏感”的概念。具体到模型中，就是要对不同的样本点赋予不同的惩罚因子，来表明不同样本被误判的惩罚不同。目前通用的代价敏感支持向量机模型，其敏感性都是从被误判样本的分类类别(即 y_i 的取值)的角度考虑的，对“正类样本被判为负类”和“负类样本被判为正类”施加不同的惩罚因子。对应的代价敏感模型如下：

$$\begin{aligned}&\min_{w,b,\xi}\ \frac{1}{2}\|w\|^2+C_1\sum_{i\in y^+}\xi_i+C_2\sum_{i\in y^-}\xi_i\\&\text{s.t.}\quad y_i(w\cdot x_i+b)\geq 1-\xi_i,\quad i=1,2,\cdots,N\\&\qquad \xi_i\geq 0,\quad i=1,2,\cdots,N\end{aligned}\tag{2}$$

其中 y^+,y^- 分别表示样本点属于正类和负类，C_1,C_2 为模型中对正类点和负类点判错所赋予的不同大小的惩罚因子。w,b 即为最优分离超平面的参数，由 w,b 可以得到分离超平面以及分类决策函数[9] [10]。

3. 考虑客户借贷规模的代价敏感支持向量机

在本文中，我们希望建立一个考虑了客户借贷规模的个人信用评估模型，这个模型不仅可以对分类类别做到代价敏感，也可以对客户借贷规模的大小做到代价敏感。首先需要做的就是对客户借贷规模进行分类。有关客户借贷规模的分类，根据方法和判定标准的不同可以产生多种方式，本文仅以客户借款金额数目作为衡量客户借贷规模的标准，把客户分成大小两类，举例说明模型建立的过程。

在模型建立过程中，样本客户数据集还是 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，根据是否违约分为正负两类，分别用 y^+,y^- 表示。同时，借贷公司根据数据集中客户借款金额的数据，划定区分客户借贷规模大小的界限，以界限为标准将客户分为两类。

结合之前的客户是否违约，我们把客户分成以下四类：

$$\begin{cases}y_b^+:\text{不会违约的大客户}\\y_s^+:\text{不会违约的小客户}\\y_b^-:\text{会违约的大客户}\\y_s^-:\text{会违约的小客户}\end{cases}\tag{3}$$

对这四种类型的客户样本点赋予不同的惩罚因子，即得到如下的代价敏感支持向量机模型：

$$\begin{aligned}&\min_{w,b,\xi}\ \frac{1}{2}\|w\|^2+C_1\sum_{i\in y_b^+}\xi_i+C_2\sum_{i\in y_s^+}\xi_i+C_3\sum_{i\in y_b^-}\xi_i+C_4\sum_{i\in y_s^-}\xi_i\\&\text{s.t.}\quad y_i\left(w\cdot x_i+b\right)\geq 1-\xi_i,\quad i=1,2,\cdots,N\\&\qquad \xi_i\geq 0,\quad i=1,2,\cdots,N\end{aligned}\tag{4}$$

其中 C_1,C_2,C_3,C_4 是与(3)中四类样本点一一对应的大小不同的惩罚因子。

为了求解上述优化问题，首先需要确定 C_1,C_2,C_3,C_4 这四个参数的值。C_1,C_2,C_3,C_4 的确定，一方面依赖于不同类别之间的损失值的对比，即对大客户与小客户的误判损失比和对会违约和不会违约的客户的误判损失比。另一方面也要考虑 $C_1\sum_{i\in y_b^+}\xi_i+C_2\sum_{i\in y_s^+}\xi_i+C_3\sum_{i\in y_b^-}\xi_i+C_4\sum_{i\in y_s^-}\xi_i$ 作为模型复杂度的表达式，其与误判损失 $\frac{1}{2}\|w\|^2$ 之间的平衡程度。在实际使用模型的过程中，这几个参数需要根据信贷机构掌握的真实数据和相应需求来确定。

当 C_1,C_2,C_3,C_4 这些参数确定之后，就可以对模型进行求解。优化问题(4)的拉格朗日函数为：

$$L\left(w,b,\xi,\alpha,u\right)=\frac{1}{2}\|w\|^2+C_1\sum_{i\in y_b^+}\xi_i+C_2\sum_{i\in y_s^+}\xi_i+C_3\sum_{i\in y_b^-}\xi_i+C_4\sum_{i\in y_s^-}\xi_i-\sum_{i=1}^{N}\alpha_i\left(y_i\left(w\cdot x_i+b\right)-1+\xi_i\right)-\sum_{i=1}^{N}u_i\xi_i\tag{5}$$

其中 $\alpha_i\geq 0,u_i\geq 0$。

求 $L\left(w,b,\xi,\alpha,u\right)$ 对 w,b,ξ 的极小，得到

$$\begin{aligned}&\nabla_w L=w-\sum_{i=1}^{N}\alpha_i y_i x_i=0\\&\nabla_b L=-\sum_{i=1}^{N}\alpha_i y_i=0\\&\nabla_{\xi_i}L=\begin{cases}C_1-\alpha_i-u_i=0,&i\in y_b^+\\C_2-\alpha_i-u_i=0,&i\in y_s^+\\C_3-\alpha_i-u_i=0,&i\in y_b^-\\C_4-\alpha_i-u_i=0,&i\in y_s^-\end{cases}\end{aligned}\tag{6}$$

即

$$w=\sum_{i=1}^{N}\alpha_i y_i x_i,\ \sum_{i=1}^{N}\alpha_i y_i=0,\ \begin{cases}C_1=\alpha_i+u_i,&i\in y_b^+\\C_2=\alpha_i+u_i,&i\in y_s^+\\C_3=\alpha_i+u_i,&i\in y_b^-\\C_4=\alpha_i+u_i,&i\in y_s^-\end{cases}\tag{7}$$

将(7)代入拉格朗日函数(5)中，再对(5)求 α 的极大，即可得到优化问题(4)的对偶问题：

$$\begin{aligned}&\max_{\alpha}\ -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K\left(x_i,x_j\right)+\sum_{i=1}^{N}\alpha_i\\&\text{s.t.}\quad \sum_{i=1}^{N}\alpha_i y_i=0\quad i=1,2,\cdots,N\\&\qquad 0\leq\alpha_i\leq C_1\quad i\in y_b^+\\&\qquad 0\leq\alpha_i\leq C_2\quad i\in y_s^+\\&\qquad 0\leq\alpha_i\leq C_3\quad i\in y_b^-\\&\qquad 0\leq\alpha_i\leq C_4\quad i\in y_s^-\end{aligned}\tag{8}$$

其中 $K\left(x_i, x_j\right)$ 为相应的核函数[11]。

通过求解凸约束优化问题的对偶问题(8)得到原问题(4)的解，即为此代价敏感支持向量机模型的对应参数，继而可以得到此个人信用评估模型的决策函数。

4. 类别划分和参数确定

本节我们将用一个具体的案例来说明新的个人信用评估模型如何建立，包括客户借贷规模大小的划分和模型参数 C_1, C_2, C_3, C_4 的确定。在本次试验中，我们选择 UCI 数据集中的 German credit 数据集(http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data))作为试验数据。German credit 数据集中共包含 1000 个客户样本，样本特征包括客户的违约情况(300 人违约，700 人未违约)、借贷金额数目、历史借贷情况、个人拥有资产、婚姻状况等信息。为了方便使用数据建立机器学习的模型，数据集中已经对原始数据进行了相关处理，包括缺失值、异常值的筛选处理，以及类别变量数值化，部分连续数据离散化，特征选择和拆分等。进一步的，在本次试验中，因为需要使用该数据集建立代价敏感支持向量机模型，为了避免不同变量间因为量纲不同导致某变量对超平面的建立影响过大的问题，还需要对特征变量做了标准化处理，本文使用的是 min-max 标准化方法，即对原始数据做线性变换，将其映射到[0,1]范围之间。

新的个人信用评估模型需要将训练数据分为四类，其中正负类别可以根据数据集中的违约情况记录进行分类，而对于客户借贷规模的类别划分，我们直接选择 German credit 数据集中的借款金额(credit amount)这一特征，将借款金额作为衡量客户借贷规模大小的标准。German credit 数据集中 1000 个借贷客户的借款金额的箱型图和概率分布图如图 1 和图 2 所示。

从图 1 和图 2 中可以看出，在客户的借款金额方面，约有 80%的客户借款数目相对集中，而剩余的约 20%客户借款数目相对较大(借款金额 > 4700)，因此将借款数目较大的前 19%客户作为可能会给借贷公司带来较大影响的客户，而其余借款数目相对较少的 81%客户作为小型客户处理。这里选择前 19%的客户而非前 20%的客户作为大客户，是为了避免借款金额相同的客户被分到不同类别当中。完整的客户样本分类情况见表 1。

通过计算大型客户的平均借款数额和小型客户平均借款数额的比值，得到大型客户平均借款数额约 8100，小型客户平均借款数额约为 2100，近似比值为 4:1，因此我们把大型客户与小型客户错判损失比定为 4:1，对于违约与否的错判损失比，本文直接使用 German credit 数据集中给定的比值，即违约错判:未违约错判 = 5:1。综上可得 $C_1 : C_2 : C_3 : C_4 = 4 : 1 : 20 : 5$。另一方面，考虑 $C_1 \sum_{i \in y_b^+} \xi_i + C_2 \sum_{i \in y_s^+} \xi_i + C_3 \sum_{i \in y_b^-} \xi_i + C_4 \sum_{i \in y_s^-} \xi_i$ 和 $\frac{1}{2}\|w\|^2$ 之间的平衡关系，在默认误判损失 $\frac{1}{2}\|w\|^2$ 前的系数为 1 的情况下，我们选取原始支持向量机中的惩罚系数 $C = 1$，相应的，在新的模型中，为了能够和原模型进行对比，我们设定

Table 1. The classification of sample customer
表 1. 客户样本分类表

类型	未违约客户数/人	违约客户数/人	总计
大型客户数/人	110	80	190
小型客户数/人	590	220	810
总计	700	300	1000

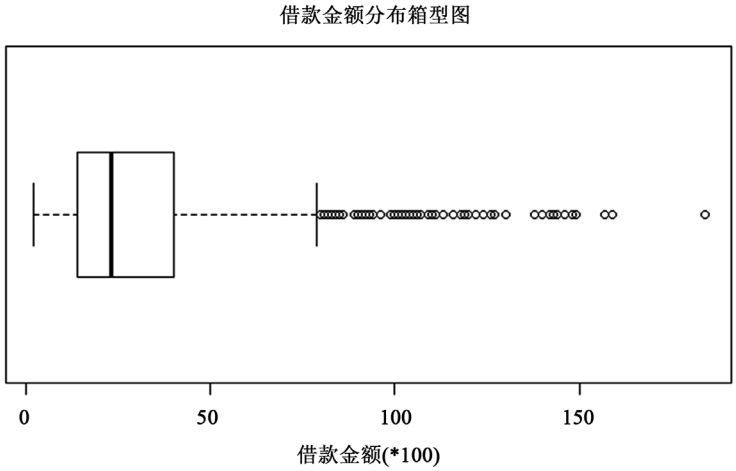

Figure 1. The box-plot of the distribution of credit amount
图 1. 借款金额分布箱型图

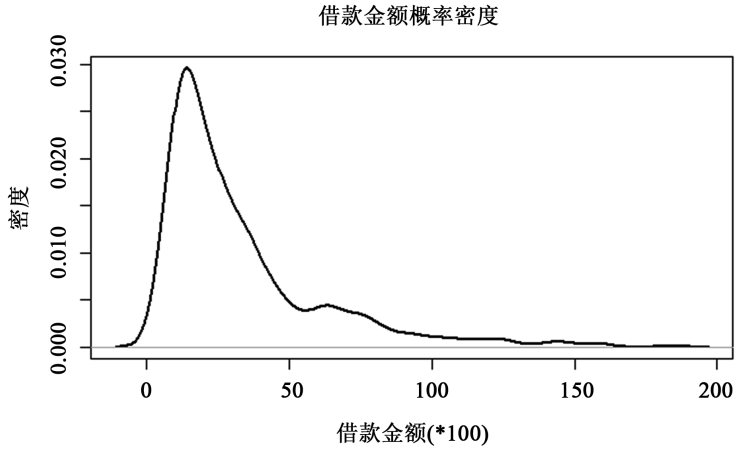

Figure 2. The probability density of credit amount
图 2. 借款金额概率密度图

$$\begin{cases} C_1 : C_2 : C_3 : C_4 = 4:1:20:5 \\ C_1 + C_2 + C_3 + C_4 = C * 4 = 4 \end{cases}$$

得到最终的惩罚系数 C_1, C_2, C_3, C_4 依次为 0.53，0.13，2.67，0.67。在类别划分和参数确定完毕之后，即可把数据带入模型中进行求解。

5. 结果对比

为了验证新的个人信用评估模型的性能，我们将实验数据分别代入原始的支持向量机模型和新的代价敏感支持向量机模型中，通过判断结果对两个模型的性能进行分析对比。评判的指标主要为客户整体以及各类别的准确率和客户整体误判损失。其中准确率定义为本类别中被模型正确判别的样本数/本类别中所有样本数，而整体误判损失定义为整体样本点中所有被错误判断的样本点的损失之和，特别的，我们需要定义每个点被错判后的损失值。因本次试验主要是对两个模型进行对比，因此不妨直接设定 $y_b^+, y_s^+, y_b^-, y_s^-$ 各类别的点被错判的损失分别为 4，1，20，5。通过对对应类别的错判的点的损失值进行加和，即可得到整体误判损失这一指标。

把数据代入模型重复试验 50 次，每次将 1000 个样本点随机分为训练集和测试集，其中训练集 800 个样本点，测试集 200 个样本点。为了控制其他条件相同，我们把原始支持向量机模型和本文的改进模型中的核函数都选定为线性核。通过模型判断，得到 50 次实验平均结果如下(SVM 表示原始的 SVM 模型，C-SVM 表示本文改进的代价敏感 SVM 模型)：(表 2)。

通过上述数据结果和对比图可以得出，改进的 SVM 模型虽然比原始的 SVM 模型准确率略低(约 4%)，但是在总体损失方面，改进的模型损失也比原始的模型低(约 35%)。

进一步的，我们可以得到如下各组数据 50 次试验的对比图。

从图 3~图 6 中我们可以得出，改进模型的准确率下降的主要原因是模型对不会违约的小客户的判断准确率有了较大幅度的下降，而在对于会违约的大客户这一损失占比最大的类别的判断上，新模型的准确度要远远高于原始的模型。可以认为，相比于原始的 SVM 模型，改进的 SVM 模型可以降低因为误判带来的总损失，同时也可以提高对大型客户尤其是会违约的大型客户的判别准确率。

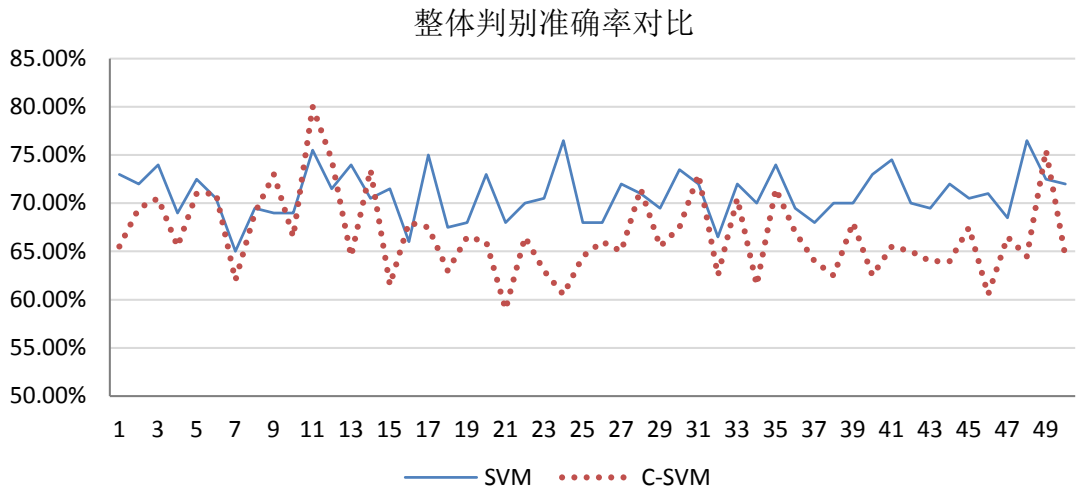

Figure 3. The comparison of accuracy for all samples
图 3. 整体判别准确率对比图

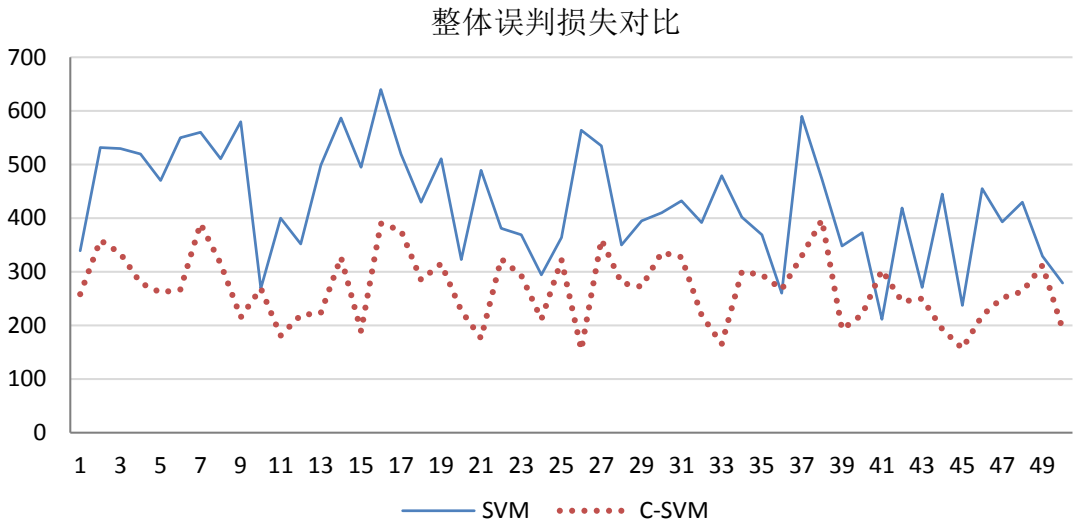

Figure 4. The comparison of misjudgment loss
图 4. 整体误判损失对比图

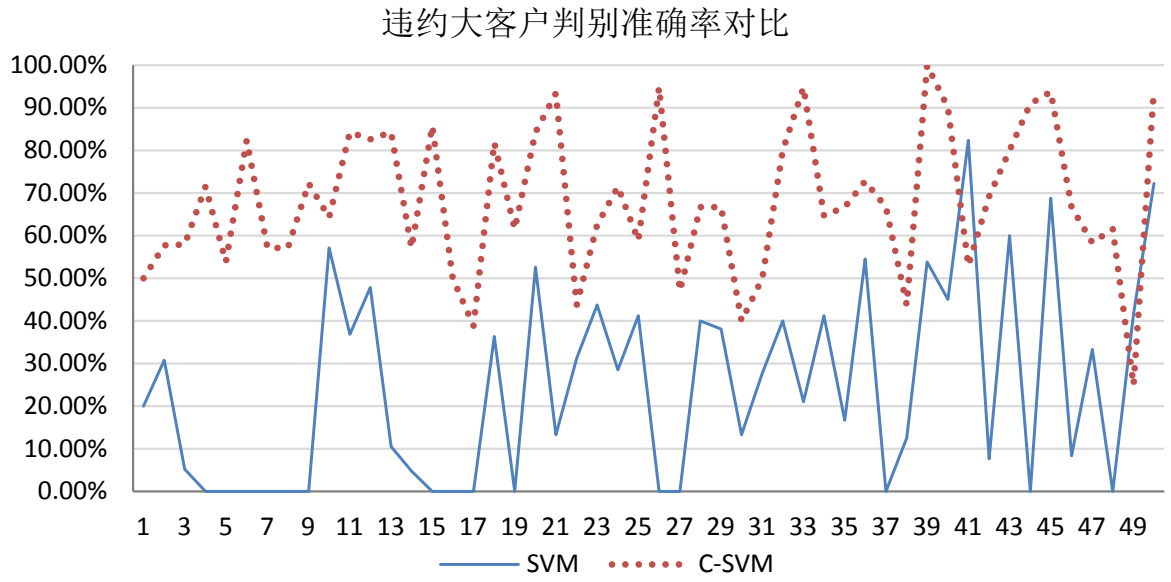

Figure 5. The comparison of accuracy for contract-broken and big-amount customers
图 5. 违约大客户判别准确率对比图

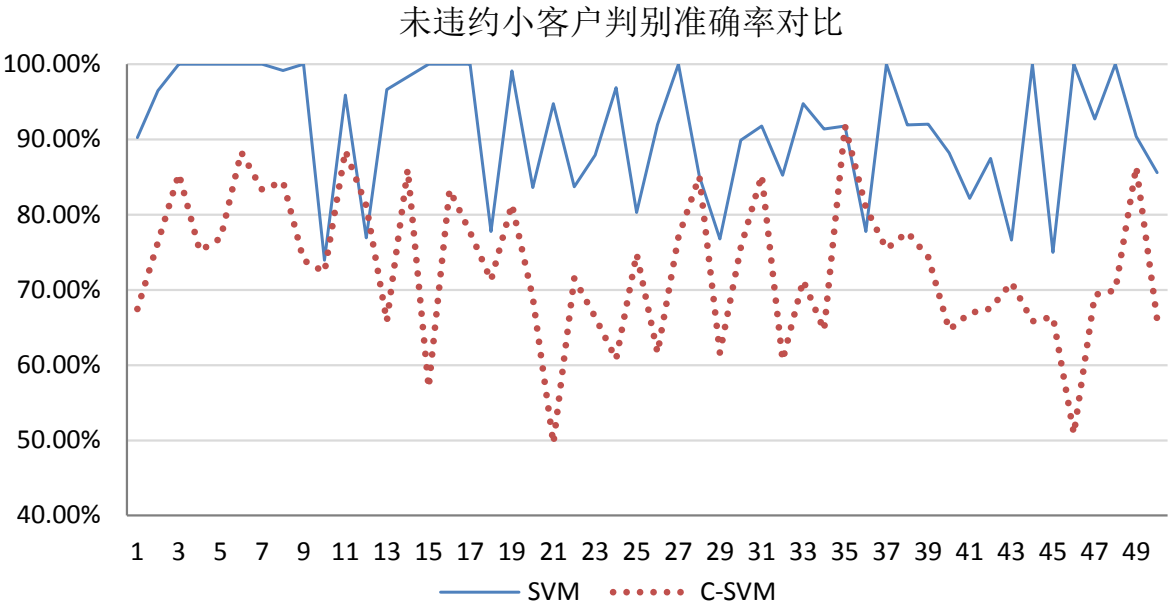

Figure 6. The comparison of accuracy for contract-unbroken and small-amount customers
图 6. 未违约小客户判别准确率对比图

Table 2. The comparison of the result of two models
表 2. 模型结果对照表

模型	准确率					总损失
	总计	y_b^+	y_s^+	y_b^-	y_s^-	
SVM	70.9%	87.70%	91.42%	24.77%	21.13%	427.12
C-SVM	66.76%	57.32%	73.13%	68.03%	53.33%	270.04

6. 结论

本文在建立信用评估模型时同时考虑了客户借贷规模大小和违约与否两个因素，对客户类型加以细分，设置不同的权重，进一步地对传统的 SVM 模型进行相应的改进，建立新的个人信用评估模型。实验证明，改进后的模型对于提高对大型客户的判别准确率，减小客户整体的误判损失有着良好的效果。下一步需要研究的是关于权重的对比定量问题和客户借贷规模的分层标准问题，即如何使用更科学的方法对客户借贷规模进行分类并赋予不同类别合适的权重值。

致　谢

本篇论文前后共花费约三个月时间，感谢在这过程中，本人导师杨国孝老师的指导与帮助。同时感谢本文参考文献中列举出的诸位学术工作者们和原始数据的提供和加工者们，为本文的研究和本人思路的开拓所提供的帮助。

参考文献 (References)

[1] 宋云鹏, 武钰. 数据挖掘技术在信用评分中的应用研究[J]. 征信, 2013(10): 24-28.

[2] 姜明辉, 许佩, 任潇, 车凯. 个人信用评分模型的发展及优化算法分析[J]. 哈尔滨工业大学学报, 2015, 47(5): 40-45.

[3] 沈翠华, 邓乃扬, 肖瑞彦. 基于支持向量机的个人信用评估[J]. 计算机工程与应用, 2004(23): 198-199.

[4] 肖文兵, 费奇. 基于支持向量机的个人信用评估模型及最优参数选择研究[J]. 系统工程理论与实践, 2006(10): 73-79.

[5] 陆爱国, 王钰, 刘红卫. 基于改进的 SVM 学习算法及其在信用评分中的应用[J]. 系统工程理论与实践, 2012, 32(3): 515-521.

[6] 段薇, 路向阳. 基于代价敏感支持向量机的银行信用风险评估模型[J]. 江西科技师范大学学报, 2015(6): 75-79.

[7] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012: 95-130.

[8] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 23-44, 121-137.

[9] 郑恩辉, 李平, 宋执环. 代价敏感支持向量机[J]. 控制与决策, 2006, 21(4): 473-476.

[10] 郑恩辉, 李平, 宋执环. 基于支持向量机的代价敏感挖掘[J]. 信息与控制, 2006, 35(3): 294-298.

[11] 解可新, 韩健, 林友联. 最优化方法[M]. 天津: 天津大学出版社, 2013: 132-166.