基于主成分分析和支持向量机的个人信用评估

肖 智,李文娟
(重庆大学 经济与工商管理学院,重庆 400030)

摘 要:本文针对信用评估指标维数较高的问题,运用主成分分析与支持向量机理论建立了一个新的个人信用评估预测模型。为反映该模型在信用评估分类方面的优越性,又分别建立了基于神经网络、K 近邻判别分析等多种理论的信用评估模型,并用同一组数据对不同的模型分别进行训练,然后比较其预测分类正确率。实验结果表明,基于主成分分析与支持向量机理论的个人信用评估模型具有较优的预测分类正确率。

关键词:主成分分析;支持向量机;预测正确率;个人信用评估
中图分类号:F830.5 **文献标识码**:A **文章编号**:1002-980X(2010)03-0069-04

1 研究背景

近年来,商业银行等金融机构的消费信贷业务不断扩大,有效的信贷管理工作越来越重要。为了降低信用风险,各授信机构都在积极地开展对贷款申请人的信用评估工作。个人信用评估也就是授信者根据贷款申请人的可知信用信息,利用各种信用评估模型,对可能引起信用风险的因素进行定性分析、定量计算,以期得到贷款申请人的还款概率,然后据此决定是否授信以及授信额度的过程。其最终目的就是为授信决策提供依据。目前,个人信用评估主要采用"分类"方法,也即是根据贷款申请人的信用资料,通过信用评估模型评估,将其分为正常类(可以按期还款者)和违约类(不能按期还款者)。

由此可见,对于授信者而言,如何能够在现有信用环境下选取科学、高效的信用评估方法,从而对贷款申请人做出有效的信用评估,就显得尤为重要。

最初的个人信用评估主要是由评价人依据个人审核的经验和判断事务的能力对贷款申请人的还款能力进行主观评价的,例如传统的分析法:3C 和 5C 评价原则。这些定性分析方法主观因素太多而指标较少,具有很大的主观随意性。

为了降低信用评估中的主观因素,使得授信者能够对贷款申请人的个人信用状况进行全面有效的评估,大量统计、运筹方法开始被应用到个人信用评估中[1],包括判别分析、回归分析、线形规划等。这些模型具有较好的可解释性和简明性,但这些方法的假设前提往往与现实中的一些情形相违背,例如:利用判别分析法的前提要求是变量要服从多元正态分布,而消费信贷数据通常具有高维性,并且含有很多定性变量,它们是非连续指标,因而不服从正态分布,这使得这类模型在应用中产生很多问题。

随着计算机技术的不断发展,各种非参数统计方法和人工智能方法开始被应用到个人信用评估中,包括分类树[2]、K 近邻判别分析[3]、神经网络[4]、支持向量机[5]等。与统计、运筹方法相比,基于这类方法建立的信用评估模型具有较好的预测能力,但同时也存在其自身的缺陷,即不能量化解释指标的重要程度。在分类树中没有参数。K 近邻判别分析是按照数据样本之间的距离或相关系数来度量亲属关系的,因而受少数异常数据影响非常大。在神经网络中没有参数解释。

到目前为止,神经网络的预测精度是各种方法中较好的,但其训练过程在黑箱中进行,摒弃了许多行业经验和专家经验。而且使用神经网络理论也会遇到一些棘手的问题,例如神经网络机构不好确定、参数选取不当时产生过学习与欠学习及容易陷入局部极小等。

支持向量机是一种基于结构风险最小化的非线性辨识方法,其推广能力很强,在解决小样本、非线性及高维模式识别问题中表现出许多特有的优势,已在很多领域得到广泛使用[6-8]。因此本文中我们建立起基于支持向量机的个人信用评估模型。支持向量机尽管分类效果很好,但是它是通过升维来达

收稿日期:2010-01-04
作者简介:肖智(1961—),男,重庆人,重庆大学经济与工商管理学院教授,博士生导师,研究方向:经济管理智能方法及统计分析;李文娟(1986—),女,河南商丘人,重庆大学经济与工商管理学院硕士研究生,研究方向:经济管理智能方法及统计分析。

到很好的分类效果的,输入数据的维数很高时将导致数据计算量的急剧增加,训练速度会很慢,因此本文中我们采用主成分分析法(principal component analysis,PCA)这种特征提取方法对评估指标进行降维,然后对提取的特征数据应用支持向量机(support vector machine,SVM)理论建立个人信用评估模型,这样会大幅度降低支持向量机的训练时间。

本文结构如下:首先介绍关于主成分分析和支持向量机的理论基础;其次说明样本数据来源、数据处理及实验设计方案;第三部分为实验结果分析;最后是结束语。

2 理论背景

2.1 主成分分析法

主成分分析法是 1933 年由 Hotelling 首先提出的,它是一种把多指标转化为少数综合指标的多元统计分析方法。它既可以降低数据维数,综合数据信息,还可以降低计算的复杂度,因此该方法越来越受到人们的重视和广泛应用[9-12]。

设个人信用评估指标集 $X = \{X_1, X_2, \cdots, X_k\}$,主成分分析的过程如下:

1)对评估指标进行标准化处理。

2)计算评估指标的相关矩阵 R 以及 R 的特征值和单位特征向量。

由 $|R - \lambda I = 0|$ 解方程得到 k 个非负的特征值 λ_1、λ_2、…、λ_k($\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq 0$)及各特征值对应的单位特征向量 U_1、U_2、…、U_k。此时,λ_1、λ_2、…、λ_k 实际上分别是主成分 Y_1、Y_2、…、Y_k 的方差。

3)计算主成分。

$Y_i = X^T U_i (i = 1, 2, \cdots, k)$。

4)主成分选取。

第 i 个主成分的特征贡献率为 $\alpha_i = \frac{\lambda_i}{\sum_{j}^{k}\lambda_j} (i = 1, 2, \cdots, k)$;主成分 Y_1、Y_2、…、Y_i 的累计贡献率为 $E = \frac{\sum_{j}^{i}\lambda_j}{\sum_{j}^{k}\lambda_j}$,一般取得 l 个主成分($l < k$),使得 $E \geq$ 85 %。

2.2 支持向量机分类原理

支持向量机的基本思想是将输入空间的样本通过非线性变换映射到高维特征空间,然后在特征空间中求取把样本线性分开的最优分类面。

设训练样本集 $D = \{(x_i, y_i)\} (i = 1, 2, \cdots, m, x_i \in R^n, y_i \in \{+1, -1\}$,$y_i$ 为输出。把这 m 个样本点看作是 n 维空间中的点,如果存在一个分类超平面:

$$\sum_{i}^{m} w_i \cdot x_i + b = 0。 \quad (1)$$

这个超平面能将 m 个样本分为两类,而且能使分类间隔($2/\|w\|^2$)最大,这样的超平面称为最优分类面。要使分类间隔最大就等价于使 $\|w\|^2/2$ 最小,寻求最优分类面的问题就转化为求解下面的最优化问题:

$$\min \frac{1}{2}\|w^2\|;$$

$$s.t.\ y_i[w^T x + b] \geq 1 (i = 1, 2, \cdots, m)。 \quad (2)$$

根据优化理论,可得线性可分条件下的分类决策树:

$$f(x) = \mathrm{sgn}\{\sum_{i}^{m} \alpha_i^* y_i (x_i^{\mathrm{T}} x) + b^*\}。 \quad (3)$$

其中:b^* 是分类阈值;α_i 是每个样本对应的 Lagrange 乘子,α_i 不为零时所对应的样本就是支持向量(support vector)。

对于线性不可分情况,通常要引进核函数来解决。只要采用的内积核函数适当,就可以将低维输入空间中的非线性可分问题转化为高维特征空间中的线性可分问题。应注意的一点是,引入的核函数 k 应满 Mercer's 条件。此时需要在目标函数式(2)中增加松弛变量 ξ_i 和惩罚函数 C,式(2)转化为:

$$\min \frac{1}{2}\|w^2\| + C\sum_{i}^{m}\xi_i,$$

$$s.t.\ y_i[w^T x + b] \geq 1 (i = 1, 2, \cdots, m)。 \quad (4)$$

所得分类决策函数为:

$$f(x) = \mathrm{sgn}\{\sum_{i}^{m} \alpha_i^* y_i K(x_i, x) + b^*\}。 \quad (5)$$

2.3 个人信用评估的实现

首先,利用 PCA 特征提取技术对个人信用评估的指标进行降维,然后应用 SVM 对通过降维变换后的数据样本进行训练,找出最优超平面,建立个人信用评估模型。最后利用已建立好的模型实现对新的贷款申请人的信用评估。

3 样本数据、数据处理及实验设计

3.1 样本数据

本文利用德国某商业银行的消费信贷数据作为研究样本集。该样本集中共有 1000 个样本:700 个正常样本和 300 个违约样本。每个样本用 21 个变量描述,其中含有 20 个属性指标变量和一个目标变量。为了检测模型的泛化能力,我们分别从两种样本数据中随机抽取 75 %组成训练样本集(525 个正常样本和 225 个违约样本),其余的组成测试样本集

(175 个正常样本和 75 个违约样本)。

由于属性指标的数据类型不统一,既有定量指标也有定性指标,因此在使用之前要将其全部转化为数值型。本文中的处理方法是:对于定量指标,采用其实际数值;对于定性指标,对其分类后再量化。以"是否外籍劳工"指标为例,分为两类:"是外籍劳工"和"不是外籍劳工",为 0 表示"是外籍劳工",为 1 表示"不是外籍劳工"。其他同理。

此外,因各个指标的量纲不同,通常不具有可比性,因此要对它们进行标准化处理,从而转化为可以同度量的指标。

3.2　基于 PCA 的特征提取

由于在该评估系统指标体系中存在的指标个数较多,因此有必要对其进行综合分析,在尽量不减少信息丢失的前提下减少指标的个数,即在保障评估结果的前提下降低信用评估指标的维度。针对上述 20 个信用评估指标,我们采用 SPSS16.0 软件实现主成分分析,从 20 个属性指标中提取 15 个主成分,使累积贡献率达到 88.0%。

3.3　实验设计

在构建支持向量模型时要注意核函数以及核函数参数的选取,因为 SVM 算法中采用不同的核函数及核函数对样本数据的预测能力产生不同的效果。常用的核函数有:线性核函数、多项式核函数、Sigmoid 函数和径向基核函数,而最常用的是径向基核函数。本文中我们采用径向基核函数作为支持向量的内积核函数。为了使模型有较好的预测能力,我们采用网格与交叉验证来选择径向基核函数的最优参数 σ 和惩罚因子 C。

此外,为了较好地反映基于主成分分析与支持向量机的个人信用评估模型的分类正确率,在本文中我们分别建立起了基于 PCA-SVM、SVM、MLP(多层感知器)、PCA-MLP、RBF(径向基神经网络)、PCA-RBF、KNN(K 近邻判别分析)以及 PCA-KNN 的信用评估模型,并用同一组数据对上述不同的模型进行训练,然后比较其对测试样本的预测分类正确率。

在整个实验中,PCA 特征提取、神经网络模型及 K 近邻判别分析都用 SPSS16.0 软件实现,支持向量机用 LIBSVM-2.89 实现。

4　实证结果分析

表 1、表 2 和表 3 中所有数据均采用的是测试样本的预测分类正确率(%)。

1) 对比表 1 和表 2 的实验结果可知,σ = 0.001、C = 100 时,基于主成分分析和支持向量机的个人信用评估最高预测正确率达到 80.4%;而在没有经过主成分特征提取的前提下,直接建立基于支持向量机的个人信用评估模型,其最高分类正确率是 80%(σ = 0.001、C = 100)。一方面,主成分特征提取技术降低了评估数据的维数,从而可以提高评估过程的运算效率;另一方面,与仅用支持向量机理论建立起的个人信用评估模型的预测分类正确率相比,基于主成分分析和支持向量机的信用评估模型也在一定程度上提高了分类正确率,从而更有助于信贷管理工作的开展。

表 1　基于主成分和支持向量机的个人信用评估结果

σ / C	C = 100	C = 300	C = 500	C = 700	C = 900
σ = 0.001	80.4	79.2	77.6	77.2	77.2
σ = 0.01	75.6	72.4	70.4	68.4	67.6
σ = 0.05	68	69.6	69.6	69.6	69.6
σ = 0.1	72.8	72.8	72.8	72.8	72.8

表 2　基于支持向量机的个人信用评估结果(%)

σ / C	C = 100	C = 300	C = 500	C = 700	C = 900
σ = 0.001	80	78.4	78.8	77.6	78
σ = 0.01	74	72.8	69.6	68	67.6
σ = 0.05	70.4	70.4	70.4	70.4	70.4
σ = 0.1	75.6	75.6	75.6	75.6	75.6

表 3　利用神经网络和 K 近邻判别分析方法的评估结果(%)

分类方法	MLP	PCA-MLP	RBF	PCA-RBF	KNN	PCA-KNN
分类正确率	70.4	79.7	76	76.5	67.5	76.6

2) 如表 3 所示,利用 MLP、RBF 和 K 近邻方法的预测分类争取率可分别达到 70.4%、76% 和 67.5%,将该评估结果与基于主成分分析和支持向量机的信用评估模型分类正确率相比较,可知 PCA-SVM 信用评估模型的预测效果要优于基于神经网络或 K 近邻判别分析方法的信用评估模型,这也突显出了联合使用主成分分析和支持向量机理论在信用评估分类中的优越性。

3) 对比表 1、表 2 和表 3 的实验结果,如果用 R(x)表示一种模型的预测分类正确率,x 是分类方法,则不难发现有如下规律:R(PCA-SVM) > R(SVM),R(PCA-MLP) > R(MLP),R(PCA-RBF) > R(RBF),R(PCA-KNN) > R(KNN)。

这也就是说,无论是建立基于支持向量、神经网

络还是 K 近邻判别分析方法的信用评估模型,利用 PCA 对评估指标首先进行特征提取然后建立基于各种方法的评估模型的分类正确率都高于未使用 PCA 特征提取技术而直接建立基于各种方法的评估模型的分类正确率,特别是对 MLP 神经网络模型而言(直接建立的 MLP 信用评估模型分类正确率是 70.4%,而 PCA-MLP 信用评估模型分类精确率可以达到 79.7%)。该规律仅限于本文的实验结果,但可以说明的是,在评估指标维数很高时,使用 PCA 特征提取技术对评估指标进行降维,确实能够起到很好的效果。

5 结束语

随着消费信贷规模的扩大,如何提高信用评估准确率已经成为信贷行业的一个重要问题,因为信用评估准确率哪怕只有很少的提高都会给信贷机构带来很大的利益收获。本文利用主成分分析与支持向量机理论建立了一个新的个人信用评估预测方法,以期能在一定程度上提高信用评估准确率。通过对比多种方法的实验结果发现,基于 PCA-SVM 的个人信用评估模型取得了较好的预测分类正确率。

参考文献

[1] ALTMA E I, SAUNDERS A. Credit risk measurement: developments over the last 20 years[J]. Journal of Banking & Finance, 1998, 21(11/12): 1721-1724.

[2] CHEN T Y, POON P L. Construction of classification trees via the classification-hierarchy table[J]. Information and Software Technology, 1997, 39(13): 889-896.

[3] 姜明辉,王雅林,赵欣,等. K-近邻判别分析在个人信用评估中的应用[J]. 数量经济技术经济研究, 2004, 21(2): 143-149.

[4] 吴冲,吕静杰,潘启树,等. 基于模糊神经网络的商业银行信用风险评估模型研究[J]. 系统工程理论与实践, 2004, 24(11): 1-7.

[5] 吴冲,夏晗. 基于支持向量机集成的电子商务环境下客户信用评估模型研究[J]. 中国管理科学, 2003, 16(7): 362-367.

[6] 扬海军,太雷. 基于模糊支持向量机的上市公司财务困境预测[J]. 管理科学学报, 2009(3): 102-109.

[7] 向小东,宋芳. 基于核主成分与加权支持向量机的福建省城镇登记失业率预测[J]. 系统工程理论与实践, 2009, 29(1): 73-80.

[8] 纪延光,徐启华,韩之俊. 基于支持向量机的 R&D 项目过程质量度量[J]. 中国管理科学, 2004, 12(6): 62-67.

[9] 李建平,徐伟宣. 消费者信用评估中的 PCALWM 方法研究[J]. 中国管理科学, 2004, 12(2): 17-21.

[10] 戚湧,李千目,孙海华. 基于主成分神经网络和聚类分析的高校创新能力评估[J]. 科学学与科学技术管理, 2009(10): 112-117.

[11] 吴开亚,何琼,孙世群. 区域生态安全的主成分投影评估模型及应用[J]. 中国管理科学, 2004, 12(1): 106-109.

[12] 张晓利,贺国光. 基于主成分分析和组合神经网络的短时交通流预测方法[J]. 系统工程理论与实践, 2007, 27(8): 167-171.

Personal Credit Scoring Based on PCA and SVM

Xiao Zhi, Li Wenjuan

(College of Economics and Business Administration, Chongqing University, Chongqing 400030, China)

Abstract: This paper attempts to build up a new personal credit scoring model based on principal component analysis (PCA) and support vector machine (SVM). In order to present the superiority of this model in consumer credit scoring, it also establishes several other personal credit scoring models based on these theories such as neural networks, K-neighbor discriminate analysis and so on, and compares the forecasting accuracy of these model through training by a same set of data. The experiment results show that the forecasting accuracy of the personal credit scoring model based on PCA-SVM is superior.

Key words: principal component analysis; support vector machine; forecasting accuracy; personal credit scoring