ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.5,No.7,March 2009, pp.1661-1663

E-mail: eduf@cccc.net.cn
http://www.dnzs.net.cn
Tel:+86-551-5690963 5690964

支持向量机在信用卡信用评估的应用

郭振亚

(厦门大学,福建 厦门 36100)

摘要:信用卡业务现在是银行很重要的资产业务,构建一个适用的个人信用评估模型十分重要。基于近年来在智能学习系统领域发展起来的新理论,引入小样本学习的通用学习算法——支持向量机(SVM),建立了个人信用评估模型,通过与神经网络模型的比较,证实了该方法用于信用卡个人信用评估的有效性及优越性。

关键词:信用卡个人信用评估;支持向量机;分类

中图分类号:TP301 **文献标识码**:A **文章编号**:1009-3044(2009)07-1661-03

Using Support Vector Machine for the Credit Evaluation

GUO Zhen-ya

(Xiamen University, Xiamen 36100, China)

Abstract: Credit card business is an important asset business in the bank,to construct a suitable personal credit evaluation model is very important.Based on the recent development in the field of intelligent system of the new theory,introduced the general learning small sample learning algorithm: support vector machine (SVM) to establish the individual credit evaluation model,through the comparison with the neural network model,This method proves to be used to evaluate the personal credit card superiority and effectively.

Key words: credit card evaluation; SVM; Classification

1 引言

信用卡已经成为市场上不容忽视的重要的投资品种和理财方式。在人人几乎都有信用卡的今天,信用卡个人信用评估显的尤为重要,怎样采用科学方法,合理评估个人信用成为银行和研究者关注的焦点之一。

个人信用评估问题其实是一个分类问题,国内外对信用评估约有 50 多年的历史,发展了统计评估和非统计评估两大类方法。统计评估方法主要包括判别分析(MDA)、线性回归、非线性回归、Logit 模型[1]以及非参数统计中的 k 一近邻判别分析方法等。非统计评估方法包括线性规划、整数规划、人工神经网络、进化算法、专家系统等.MDA 最大优点是具有较好的解释性和简明性,但需满足实际中难以满足的正态、等协方差的条件。尽管二次判别分析(ODA)模型可解决等协方差阵问题,但却不满足正态性假定;并且,当数据样本少、维数高时 ODA 的性能明显下降;而样本少、维数高是目前信用数据的显著特点。Logit 模型无需假定任何概率分布,也不要求等协方差性,但当样本点存在完全分离时,模型参数的最大似然估计可能不存在。另外,该方法对中间区域的判别敏感性较强,导致判别结果的不稳定。卜近邻判别分析方法是一种常用的非参数模式识别方法,它不要求数据正态分布,主要特点在于它的非参数的特点使得在特征变量空间上对于不规则变量的建模成为可能。然而,当数据维数比较高时,会存在所谓的"维数灾"问题。而且,即使样本量很大,其散落在高维空间中仍显得非常稀疏,绝大多数点附近根本没有样本点,这使得 k—近邻法的思想很难体现。人工神经网络是一种对数据分布无任何要求的非线性技术,它能有效解决非正态分布、非线性的信用评估问题,但由于神经网络算法采用的是经验风险最小化原则,容易陷人局部极小点,收敛速度慢,并且其结构难以确定,以及易出现过学习现象,限制了其在实际中的应用。支持向量机(SVM)[1]是由 Cones 和 Vapnik 于 1995 年首先提出来的。它具有良好的泛化能力和较好的分类精确性,在解决模式识别中小样本、非线性及高维识别问题中表现出独特的优势和良好的应用前景。另外,SVM 采用的是结构风险最小化原则,整个求解过程转化为一个凸二次规划问题,解是全局最优的和唯一的,因此,正成为继模式识别和神经网络研究之后机器学习领域新的研究热点[2]。本文将 SVM 用于建立信用卡个人信用评估模型,取得了较好的效果。

2 模型与参数

2.1 SVM 算法

统计学习理论(Statistical Learning Theoyr)是 Vapnik 等人提出的一种专门研究小样本情况下机器学习规律的理论,是传统统计学的重要发展和补充。该理论针对小样本统计问题建立了一套新的理论体系,在这种体系下的统计推理规则不仅考虑了对渐近性能的要求,而且追求在现有有限信息的条件下得到最优结果[3]。在这一理论的基础上发展了一种新的通用的学习方法—SVM(Support Vector Machine) SVM 是从线性可分情况下的最优分类超平面发展而来的[4],基本原理如下:假定训练数据$(x_1,y_1),\ldots,(x_l,y_l),x=R^n,y_i\in\{-1,+1\}$可以被一个超平面$xw-b=0$没有错误地分开,与两类样本点距离最大的分类超平面会获得最佳的推广能力。最优超平面将由离它最近的少数样本点(称为支持向量)决定,而与其它样本无关。用如下形式描述与样本间隔为 Δ的分类超平面:

$$\begin{cases} wx-b=0,\|w\|=1 \\ y=1,若\ wx-b\geq\Delta \\ y=-1.若\ wx-b\leq\Delta \end{cases}$$

Vapnik 给出了一个关于Δ—间隔分类超平面 VC 维上界的定理:如果向量 x 属于一个半径为 R 的球中,那么 Δ—间隔分类超平面集合的 VC 维 h 有下面的界:

$$h\leq \min\left(\left[\frac{R^2}{\Delta^2}\right],n\right)+1$$

收稿日期:2009-01-04

由此 SVM 首先保证了一个小的经验风险 (在训练样本可分时就是零), 并通过选择边缘最大的超平面的方式控制了函数集的 VC 维,这正是 SRM 原则所要求的。

目前,在解决二分类问题的 SVM 分类算法主要有两种,分别为 C_SVC 系列和 V_SVC 系列,下面将介绍这两种算法的基本思想。

2.2 C_SVC 算法

C_SVC 支持向量机分类算法是最经典的支持向量机形式。对于训练向量 $x_i \in R^{(N)}, i=1,...,l$,属于两类,即 $y_i \in \{1,-1\}$,初始问题为:

$$\min_{w,b,\varepsilon} \frac{1}{2} w^T w + C\sum_{i=1}^{l} \varepsilon_i$$
$$s.t. y_i(w^T \varphi(x_i)+b) \geq 1-\varepsilon_i$$
$$\varepsilon_i \geq 0, i=1,2,...,l$$

其中 C 为惩罚参数,C 越大表示对错误分类的惩罚越大,它也是算法中唯一可以调节的参数。为了区别,称之为 C_ SVM 算法。采用拉格朗日乘子法求解这个具有线性约束的二次规划问题,得到的对偶问题为:

$$\min_{a} \frac{1}{2} \alpha^T Q\alpha - e^T\alpha$$
$$s.t. o \leq \alpha_i \leq C, i=1,2...,l$$
$$y^T\alpha = 0$$

其中 e 是单位向量,C>0 是上界,Q 是一个 l×l 正半定义矩阵, $Q_{ij} = y_i y_j K(x_i, x_j)$。

2.3 V_SVC 算法

V_SVM 支持向量机分类算法使用一个新参数 v 来控制支持向量的数目和误差。其初始问题为:

$$\min_{w,b,\xi_i,\rho} \frac{1}{2} w^T w - vp + \frac{1}{l}\sum_{i=1}^{l} \xi_i$$
$$s.t. y_i(w^T \varphi(x_i)+b) \geq \rho - \xi_i$$
$$\xi_i \geq 0, i=1,2,...l$$
$$\rho \geq 0$$

比较起来,它与 C_SVM 相比区别就比较大了。这里不含参数 C,但是却有另一个参数 V,其含义是,对于 l 个训练样本,若被错分的样本数为 R,则 $v \geq \frac{R}{l}$。若支持向量的个数为 S,则 $v \geq \frac{S}{l}$。另外,上式子中还多出一个变量 ρ,这个变量的几何意义是:当 $\xi = 0$ 时,从约束条件知道,两类样本点以 $2\rho/\|w\|$ 的间隔被分开。其对偶问题为:

$$\min_{a} \frac{1}{2} \alpha^T Q\alpha$$
$$s.t. 0 \leq \alpha_i \leq \frac{1}{l}, i=1,2,...l$$
$$y^T\alpha = 0$$
$$e^T\alpha \geq v$$

决策函数为: $f(x) = sign\left(\sum_{i=1}^{l} y_i\alpha_i K(x_i, x) + b\right)$

3 SVM 的信用风险评估模型

3.1 个人信用特征数据

在信用个人信用评估中,个人信用特征数据的选择至关重要。根据银行客户信用特征数据评估体系的要求,可考虑选择特征数据的范围是:客户自然状况、客户消费行为习惯、客户道德行为状况等。借鉴银行的客户信用数据指标,综合考虑指标体系中主要数据指标,以及数据度量的可操作性,设计了新的客户信用特征数据。该特征数据与银行客户信用评估指标体系有很好的一致性[5]。

根据银行个人信用评级指标类别,可分为五大方面,即:

自然情况、偿债能力、信用情况、其它、加分因素等,其具体指标数据如下:

自然情况:年龄、性别、文化程度、职业、婚姻状况、健康状况。

偿债能力:家庭月收入。

信用情况:银行卡记录、代发工资情况。

其它: 债率、存款余额、其它借款情况。

因此,根据常用准则选取了 8 个客户信用评估指标:1)性别,2)年龄,3)月收人,4)月支出,5)所居住地区的平均收人,6)所居住地区的失业率,7)所拥有的信用卡级别,8)未偿还的贷款余额。

3.2 样本数据处理

本文所采用的实验数据来自于福建商业银行的部分财务数据,通过对关系数据库模型进行分析,确定所需的相应字段,并通过 SQL 语句对字段中的数据进行清洗,使得数据符合数据挖掘应用的要求,即将数据归一划成[-1,1]区间。最后分析得出,该银行数据中拥有 4500 名具有贷款资格的客户,而办理过贷款业务的客户只有 827 名,其中 792 名客户的信誉度良好,银行对其发放贷款的风险较小,记为“好”客户;其余的 31 名客户信誉度较差,若给予贷款,其违约的可能性较大,记为“坏”客户。因此,从 7%名“履约”客户中抽取 415 名,与 31 名“违约”客户构成了一个规模为 450 名客户的样本集。

3.3 SVM 模型构造

根据上述分析,构造了样本集(x,y),其中二的维数为 8,y 是样本的类别属性,对于“履约”客户 y=1,对于“违约”客户 y=-1。在 SVM 算法中不同的内积核函数对数据样本的预测能力具有不同的效果,目前主要有:多项式核函数(Ploy: nomial Kernel);径向基核

函数(Redial Bais Kernel) ;Sigmoid 函数 3 种。手写数字识别实验表明采用这三种不同核函数的 SVM 得到相近的结果，且支持向量的分布差别不大。对于具体问题,如何选择核函数,目前还没有一般性的方法。因此,选取了多项式和径向基这两种核函数。对 SVM 分类算法,采用了交叉验证方法解决两类样本数据不均衡的问题 [6]，通过两种 SVM 分类算法,即 C_SVC 和 V_SVC,与两种核函数的结合同神经网络算法进行实验分析比较。

表 1 不同客户信用评估模型的判别结果

分类算法	样本正确分类率/%
C_SVM(POLY)	93.47
C_SVM(RBF)	93.56
V_SVM(POLY)	93.45
V_SVM(RBF)	93.52
神经网络（BP）	88.65

4 结果分析

表 1 列出 SVM 模型的结果,包括在采用不同 SVM 分类算法的和核函数的判别结果。同时,和神经网络所建模型的结果进行了比较。神经网络使用的是 BP 算法,目标误差及隐层的个数也是采用交叉验证的方法得到的(目标误差为 0.1,隐层个数为 15)。由于神经网络方法并不是一种稳定的方法，故表中神经网络的结果是 9 次平均的结果。从表 1 可以看出,SVM 在整体准确率较高可达 93.44%,明显好于神经网络模型的 88.65%。在设计神经网络过程中,有效利用自己的经验和先验知识是至关重要的。因此,神经网络模型的优劣是因人而异的,而支持向量机具有严格的理论和数学基础,可以有效克服神经网络中人为因素影响的不足。在本实验过程中,对 SVM 模型的不同分类算法及核函数的选取的影响也作了分析比较,实验数据表明,无论采用哪种 SVM 模型,均可以达到很高的分类精度。但由于 C_SVC 分类算法中唯一可调节的参数 C 没有直观解释,从而导致在实际应用中合适的参数值难以选择。而 V_SVC 分类算法恰好弥补了 C_SVC 算法的这一不足,它是通过用参数,取代参数 C 来控制支持向量的数目和误差的,较参数 C 易于选择。

5 结束语

SVM 是一种基于小样本学习理论的通用学习算法,具有严格的理论基础,能较好地解决小样本、非线性、高维数和局部极小点等实际问题。通过采用不同的核函数以及选取不同的参数提出了一种有效的、基于 SVM 的银行客户信用评估模型。通过与神经网络模型的比较,发现 SVM 模型在选取较优的核函数及参数后能有效地提高预测准确率,模型本身的鲁棒性也较强,具有较好的发展前景,值得深入研究。未来的工作可从以下二方面开展:首先,本文研究的仅限于一个两类的分类问题,可以考虑将 SVM 推广到更为复杂的、多类的信用等级评估问题上,以更好地反映持卡人的信用情况,为银行的贷款决策提供更有力更细致的辅助工具;其次,由于实际样本数据类别分布不均匀,对样本数量较多的类进行了重新采样,随机地剔除了一些样本,造成信息的浪费。如何在最大化利用样本信息的前提下,对样本类别分布不均匀的数据构造 SVM 模型也是未来的一个方向。

参考文献:

[1] 邓乃扬,田英杰.数据挖掘中的新方法—持向量机[M].北京:科学出版社,2004.
[2] 王春峰,万海晖,张维.组合预测在商业银行信用风险评估中的应用[J].管理工程学报,1999,13(1):5–8.
[3] 林成德,叶武.基于神经网络的企业信用等级评估[J].系统工程学报,2002,17(6):570–575.
[4] Vapnik V N. 统计学习理论的本质[M].张学工,译.北京:清华大学出版社,1999.
[5] Burges C J C. A tutorial on support vector machines for pattern recognition[J].Data Mining and Knowledge Dis2covery,1998,2(2):955–974.
[6] 建华,吴今培.样本数目不对称时的 SVM 模型[J].计算机科学,2003(30):165–167.

郭振亚(1982–),男 ,河南驻马店人,硕士,
主要研究方向:数据挖掘技术。