一种 F-scores 和 SVM 结合的客户分类方法①

段刚龙，黄志文，王建仁

(西安理工大学 经济与管理学院，西安 710054)

摘 要：为了克服现有客户分类方法在假设前提、准确度、泛化能力等方面的不足，提出了一种 F-scores 和 SVM 算法相结合的客户分类方法，并把该方法应用到银行信用卡客户分类问题中予以验证。实证分析表明：该方法最终的模型验证准确率可达 95%以上，学习和分类能力良好。

关键词：SVM；F-scores；属性选择；客户分类

A Method Combined of Support Vector Machine and F-scores for Customer Classification

DUAN Gang-Long, HUANG Zhi-Wen, WANG Jian-Ren

(Economics and Management School of Xi'an University of Technology, Xi'an 710054, China)

Abstract: A method combined of F-scores and support vector machine for customer classification was proposed, which can overcome the shortages of the existing customer classification method such as strict hypothesis, poor generalization ability, low prediction accuracy and low learning rate etc., and was applied to the problem of bank credit card customer classification. Empirical results show the validation accuracies of the final model can achieve 95% or more, which concludes that learning and generalization abilities of this model are excellent.

Keywords: support vector machine; F-scores; attribute selection; customer classification

1 引言

支持向量机(SVM)是基于统计学习理论的机器学习方法，也是数据挖掘算法研究的热点之一。SVM 能够较好的解决小样本，非线性，高维数识别和局部极小点等问题，在模式识别等数据挖掘领域应用广泛。详细说来，可以应用于如下领域：人脸检测，故障诊断，分类，回归，聚类，时间序列预测，系统辨识，金融工程，生物医药信号处理，生物信息，文本挖掘，自适应信号处理，剪接位点识别，手写体相似字识别，岩爆预测，缺陷识别，计算机键盘用户身份验证，视频字幕自动定位于提取，说话人的确认等[1-5]。

客户分类问题是当今客户关系管理及信息化技术高度发展形势下的热点问题，这类问题属于数据挖掘技术的应用问题。目前数据挖掘技术在客户分类问题上已有许多解决方法，但是这些方法普遍存在着学习稳定性差、分类准确率不高、泛化能力不强等问题。

本文应用 SVM 算法理论，结合 F-scores 方法进行属性筛选，提出了一种基于 F-scores 和支持向量机结合的客户分类方法，能够解决客户分类中的很多实际问题，弥补了一些现有方法的不足。

2 SVM基本原理

支持向量机(Support Vector Machine，SVM)是 Vapnik 等人提出的一种新型机器学习方法，它遵循结构风险最小原则和有限样本假设，克服了传统机器学习(如神经网络)过学习、局部收敛、高维灾难等问题，具有较好的学习能力和推广能力。SVM 的基本思想就是根据结构风险最小化原理,构造一个目标函数将两类模式尽可能地区分开来，通常分为两类情况来讨论：(1) 线性可分，(2) 线性不可分。对于线性可分问题，假定总体 D，$D=\{x_i,y_i \mid i=1,2,...,n\}(x\in R^P, y\in R^q)$，能被超平面 H：$w\cdot x+b=0$ 正确分开，且分类间隔最大。对于线性可分样本 $(x_i,y_i),i=1,2,...,n$ 构成造其最优分类超平面可以用如下凸二次规划描述[6]：

① 收稿时间:2010-05-11;收到修改稿时间:2010-06-11

$$\min_{w,b}\frac{1}{2}\|W\|=\min_{w,b}\frac{1}{2}W^TW \tag{1}$$

$$\text{s.t. } y_i\left[(w\cdot x_i)+b\right]-1\geq 1 \quad (i=1,2,...,n) \tag{2}$$

引入 Lagrange 系数 a_i，其对偶形式为：

$$\max_a\sum_{i=1}^{n}a_i-\frac{1}{2}\sum_{i,j=1}^{n}a_ia_jy_iy_j(x_ix_j) \tag{3}$$

$$\text{s.t.}\sum_{i=1}^{n}y_ia_i=0 \quad (a_i>0;i=1,2,...,n) \tag{4}$$

求解得到决策函数：

$$f(x)=\text{sgn}\left\{\sum_{i=1}^{n}a_iy_i(x\cdot x_i)+b\right\} \tag{5}$$

其中 *sgn*(・)为符号函数，*b* 为分类域值。

对于非线性问题，可通过非线性变换转化为在高维特征空间求取最优分类面。通常采用满足 Mercer 条件的核函数 $K(x, x_i)$来实现这一非线性变换，引入规则化常数 C (C>0)，最优分类问题转换为求解二次规划[1][5]：

$$\max_a\sum_{i=1}^{n}a_i-\frac{1}{2}\sum_{i,j=1}^{n}a_ia_jy_iy_jK(x_ix_j) \tag{6}$$

$$\text{s.t.}\sum_{i=1}^{n}y_ia_i=0 \quad (0\leq a_i\leq C;i=1,...,n) \tag{7}$$

求解得到相应的决策函数：

$$f(x)=\text{sgn}\left\{\sum_{i=1}^{n}a_iy_iK(x\cdot x_i)+b\right\} \tag{8}$$

3　基于SVM的客户分类

3.1 客户分类问题描述

客户分类就是把客户根据某种特征分成若干类，为了描述清楚，我们以两分类为例，这里具体指商业银行信用卡瑕疵客户分类问题。用 $\mathrm{Y}=[y_1,...,y_m]^T$ 来表示 m 个样本的类标号，$y_i=k$ 用来表示样本 i 属于第 k 类客户，其中，k=1 和-1。k 是样本的类别属性，用 k= 1 表示“第一类客户”，用 k= -1 表示“第二类客户”。 x_{ij} 来表示第 i 个样本中的第 j 个属性的表达值，j=1,2,…,n。所有的属性的表达式 $X=(x_{ij})_{m,n}$ 可以表示为

$$\begin{pmatrix} \text{属性1} & \text{属性2} & \cdots & \text{属性26} \\ x_{11} & x_{12} & \cdots & x_{1,26} \\ x_{21} & x_{22} & \cdots & x_{2,26} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{m26} \end{pmatrix} \tag{9}$$

这里，用 $x_1,\ldots, x_m$ 来分别表示 m 个样本，此处 $x_i=[x_{i1},x_{i2},...,x_{in}]$。

运用当中，每一个样本都可以用一个超平面来进行划分，该优化超平面满足将所有的训练数据最大化区分开来的要求。当用该超平面来划分当前的训练集时，可以获得最低的分类错误率。该超平面可以用公式（10）来建立模型。

$$f(x)=(\sum_{i=1}^{L}y_i\alpha_iK(x_i,x)+b) \tag{10}$$

此处 α_i 表示权重，每一个支持向量都有一个 α，b 为超平面的偏置项，所有的支持向量 1,…,L 运算之和即 $\sum_{i=1}^{L}y_i\alpha_iK(x_i,x)$ 构成超平分类面[4][7]。

3.2 基于 F-scores 的输入变量选择

给定训练样本 $x_k\in R_n$，$k=1,2,\cdots,l$，其中属于正类和负类的样本个数分别为 n_+和 n_-。则训练数据的第 i 个属性的 F-score 定义为[2][8]：

$$F(i)=\frac{(\bar{x}_i^{(+)}-\bar{x}_i)^2+(\bar{x}_i^{(-)}-\bar{x}_i)^2}{\frac{1}{n_+-1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)}-\bar{x}_i^{(+)})^2+\frac{1}{n_--1}\sum_{k=1}^{n_+}(x_{k,i}^{(-)}-\bar{x}_i^{(-)})^2} \tag{11}$$

其中，$\bar{x}_i,\bar{x}_i^{(+)}$ 和 $\bar{x}_i^{(-)}$ 分别为第 i 个属性在整个数据集上的平均属性值，在正类数据上的平均属性值和在负类数据上的平均属性值；$x_{k,i}^{(+)}$ 为第 k 个正类样本点上第 i 个属性的属性值；$x_{k,i}^{(-)}$ 为第 k 个负类样本点上第 i 个属性的属性值。F 值越大，表明此属性的辨别力越大。选择 F 值大于平均值的属性作为输入属

性，其余属性的分辨力较弱，可舍去不同，提高运算效率[3,9]。

对于商业银行信用卡瑕疵客户分类问题，在银行数据仓库中，信用卡客户的数据包括信用卡顾客编号、瑕疵客户、申请书来源、逾期、呆帐、借款余额、退票、拒往记录、强制停卡记录、张数、频率、户籍、都市化程度、性别、年龄、婚姻、学历、职业、个人月收入、个人月开销、住家、家庭月收入、月刷卡额、宗教信仰、人口数、家庭经济、血型、星座等 26 个属性。其中“瑕疵客户”是客户分类的目标属性，其于属性为输入决策属性。

为了提高处理效率，通过 F-scores 方法对属性进行属性筛选，根据公式(3)在训练集上计算所有属性的 F 值，并得到平均值为 0.214131 将其设为门阀值。然后取得分大于平均值的属性：强制停卡记录(1.287545)、逾期(0.949445)、退票(0.917371)、借款余额(0.908946)、呆帐(0.825485)、拒往记录(0.663530)作为输入变量，在此将其命名为特征属性。

3.3 数据预处理与归一化

数据预处理就是将连续属性值离散化，归一化处理是把所有属性的数值都缩放到已知的同一个数量级上，本文将现有数据归一到[0,1]上。用 X 表示输入空间，即由每一个样本输入变量值组成，Y 表示输出域，即由每个样本对应的类别属性值组成。$(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)$ 表示样本。$X \in R_n, y\{-1,+1\}$。为了获得更好的分类效果，对于每一列输入样本使用如下公式归一化：

$$x_{ij}' = \frac{x_{ij} - x_{\min j}}{x_{\max j} - x_{\min j}}$$

$$(i = 1,2...,l;\ j = 1,2,...,26) \qquad (12)$$

其中，x_{ij} 表示第 j 个向量的 i 个样列。x_{ij}' 记录 x_{ij} 归一化后的值，L 记录样本个数，$x_{\min j}$ 指该向量中的最小值。$x_{\max j}$ 指该向量中的最大值。归一化后 $x_{ij}' \in [0,1]$ 例如年龄、个人月收入、个人月开销、家庭月收入、月刷卡额均等为连续型变量，将其离散化。并对属性根据公式(4)进行归一化处理。

3.4 核函数确定

对于 SVM 算法，常用的核函数有以下四种[1][8]：

(1) Dot 函数：$K(x, x_i) = x \cdot x_i$

(2) Polynomial 函数：$K(x, x_i) = [(x \cdot x_i) + 1]^d$

(3) Neural 函数：$K(x, x_i) = \tanh[(ax \cdot x_i) + b]$

(4) Radial 函数：$K(x, x_i) = \exp(-\gamma \|x - x_i\|^2)$

在客户分类 SVM 模型的训练过程中，从这四个核函数中选择最合适的核函数和其它相关参数，得到最优的客户分类 SVM 模型，其决策函数为：

$f(x) = \mathrm{sgn}\left\{\sum_{i=1}^{n} a_i y_i K(x \cdot x_i) + b\right\}$。这里采取对比实验的方法，分别将以上四个核函数代入模型实验，比较结果，找出最合适的核函数。

4 实验分析

论文从某商业银行信用卡客户数据仓库中提取 43750 条输入输出完整的历史数据作为实验分析对象总样本，选择 2.2 中所选出的关键属性变量，并对变量值进行预处理与归一化后，选取 10000 条作为训练数据，33750 条作为测试数据。

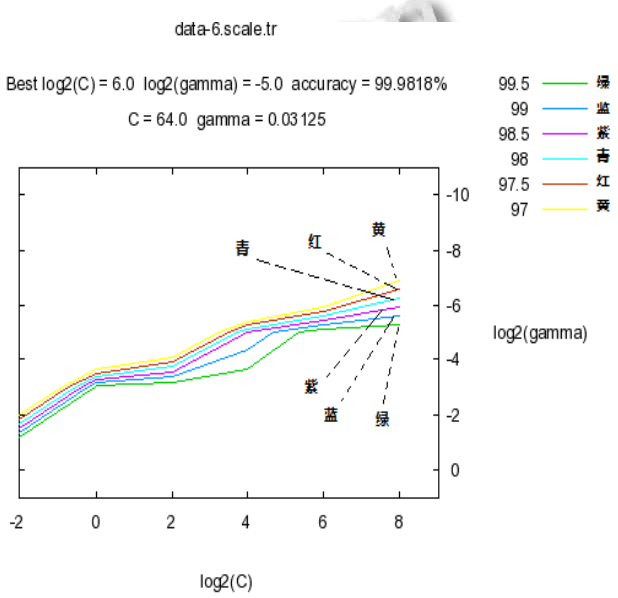

图 1 寻参(C, g)结果

这里应用 libsvm[7][10]通过训练集对支持向量机的

参数 C 和 g 进行寻优，本文采用 grid-search 和交叉验证的方法。(1)根据 grid-search 方法，不妨假定 $C\in\{2^{-2},2^{-1},...,2^{8},2^{9}\},g\in\{2^{1},2^{0},...,2^{-10},2^{-11}\}$，然后在训练集上对每个参数对($C$, g) 进行 5-fold 交叉验证。(2)选取平均验证误差最低的参数对(C ,g) 作为最优参数。寻参结果如图 1 所示，最终采用 C=64，g=0.03125 作为训练模型的参数。

选取四类不同的核函数，分别对训练数据进行建模验证，实验结果见表 1。

表 1　四种不同核函数测试结果比较

	线性核函数	多项式核函数	高斯径向基核函数	S-核函数
5-fold 交叉验证的平均准确率	89.9091%	95.5364%	95.9818%	80.7364%
迭代次数	3648	2710	14537	17502
支持向量总数	2279	1034	1342	1938
测试准确率	91.6552%	96.8815%	96.873%	84.1818%

表 1 列出了分别采用不同核函数的 SVM 模型的验证结果及其对比情况，结果表明采用多项式核函数及高斯径向基核函数均得到了很好的分类效果，两者相比，选择多项式核函数时模型最优；同时，也说明试验选取的六个属性具有很强的特征性，能代表全体数据集的绝大多数信息。

5　结论

论文提出了一种 F-scores 和支持向量机结合的客户分类方法，并将其在银行信用卡瑕疵客户分类问题中进行了应用验证。实证分析结果表明：该方法最终的模型验证准确率可达 95%以上，说明其学习能力及泛化能力较强。此外，该模型还很好地克服了样本假设难以满足的缺陷。总之，该方法可以满足银行信用卡瑕疵客户分类问题的实际应用需求，同时可以把该方法推广到其它的分类问题当中。

参考文献

1 Burges CJC. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery. 1998,2: 121－167.

2 杨立才,李金亮,姚玉翠,吴晓晴.基于 F-score 特征选择和支持向量机的 P300 识别算法.生物医学工程学杂志,2008, 25(1):23－26.

3 谢娟英,王春霞,蒋帅,张琰.基于改进的 F-score 与支持向量机的特征选择方法.计算机应用,2010,30(4):993－996.

4 郑启鹏,李秀,刘文煌,李兵.支持向量机在银行贷款客户分类中的应用研究.微计算机信息,2005,21(11-3):68－70.

5 李红莲,王春花,袁保宗,朱占辉.针对大规模训练集的支持向量机的学习策略.计算机学报,2004,27(5):715－719.

6 Vapnik V. The nature of statistical learning theory. New York: Springer-Verlag, 1995.

7 Chen YW, Lin CJ. Combining SVMs with Various Feature Selection Strategies. [2009-12-21]. http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/#5/features.pdf.

8 Axelberg PGV, Irene Gu YH, Bollen MHJ. Support Vector Machine for Classification of Voltage Disturbances. IEEE Trans. on Power Delivery, 2007,22 (3):1297－1303.

9 陈启买,陈森平.支持向量机的一种特征选取算法.计算机工程与应用,2009,45(23):49－51.

10 Lin CJ. LIBSVM: a library for support vector machines (Version2.6). [2010-03-18]. http://www.csie.edu.tw/~cjlin/papers/libsvm.pdf.