营销科学学报
第 2 卷第 4 辑:1－7

Journal of Marketing Science
Vol. 2 No. 4, December 2006:1－7

夏国恩①,金炜东②

摘　要　在客户流失预测过程中不可避免地会出现“拒真纳伪”两类错误。本文通过对“拒真纳伪”两类错误在客户流失预测中不同影响的分析比较,采用支持向量机(SVM)作为预测模型,并利用某电信公司实际数据对两类错误的平衡控制进行了研究。实验结果表明,选取一个适当的损失比例系数,预测模型能在控制两类错误的前提下有效地减少期望损失函数值,这在实际应用中具有反映问题本质的现实意义。

关键词　两类错误,客户流失,支持向量机

客户流失预测中两类错误的平衡控制研究

0　引言

客户流失预测是客户关系管理中一项重要的基础性工作。为了有效地预测未来潜在的流失客户,学者们主要提出了以下两类预测模型:第一类是传统预测模型,如决策树(Chih 和 Chiu, 2002)、Logistic 回归(Logistic Regression)(Kim 和 Yoon,2004;Rosset 和 Neumann,2003)、贝叶斯分类器(Naive Bayesian Classifiers)(Nath, 2003)、聚类分析(Clustering)(Yi Ming 等, 2004);第二类是人工智能预测模型,如人工神经网络(Artificial Neural Network, ANN)(Yan 等,2001)、自组织映射(Self Organizing Maps, SOM)(Ultch, 2002)和进化学习(Evolutionary Leaning,EL)算法(Au 等,2003)。在使用模型进行预测的过程中,不可避免地会犯“拒真纳伪”两类错误。这两类错误对于客户保持的影响有很大的差别。然而,无论是传统预测模型还是人工智能预测模型,都无法直接控制两类错误的分布并且其模型评价标准也没有考虑两类错误的关系,从而影响了它们在实际应用中的效果。

针对上述问题,笔者以支持向量机(Support Vector Machine,SVM)为预测工具,讨论了客户流失预测中的两类错误的平衡控制问题。文中引进了一个损失比例系数 Y 用其来调整第一类错误与第二类错误间惩罚系数的比率,在此基础上定义了评价预测模型的客户流失期望损失函数。通过对美国某电信公司未来潜在的流失客户进行预测,发现在当前和未来时间条件下,SVM 可以在不同的错误类别上采用不同的惩罚系数,从而有效控制“拒真纳伪”两类错误的分布,并且以客户流失期望损失函数值为预测模型的评价标准比模型总错误率更具有实际意义。

1　客户流失预测中的两类错误

1.1　假设检验中的“拒真纳伪”两类错误

在统计学的假设检验中由于样本的随机性,在进行判断时可能犯两类错误。一是当假设 H_0 为真时,拒绝了它,称为犯第一类错误发生的概率或称拒真概率,记作

$$\alpha = P\{X \in W \mid H_0 \text{ 真}\} \qquad (1)$$

① 夏国恩,西南交通大学经济管理学院博士研究生,E-mail:gandlf007711@163.com

② 金炜东,西南交通大学经济管理学院教授、博士生导师,E-mail:WDJin@vip.sina.com

其中，W 是一个检验的拒绝域。二是当假设 H_0 为非真时，而接受了它，称为犯第二类错误发生的概率或称纳伪概率，记作

$$\beta = P\{X \notin W \mid H_0 \text{ 非真}\} \quad (2)$$

犯这两类错误所造成的影响常常很不一样。以电信业根据客户的可能流失情况判断是否给对客户进行客户保持为例，原假设为 H_0（该客户可能流失）。此时，犯第二类错误会使企业多花费用于客户保持的费用，但犯第一类错误可能导致客户流失。企业希望根据历史样本的检验结果做出的预测使犯两类错误的概率都尽可能小，但实际上是不可能的。由于两类错误之间存在制约关系（蔡越江，1999）：当其他条件不变时 α 减小必导致 β 增大；反之，β 减小则 α 增大。因此，在样本容量一定的情况下，不能同时控制犯两类错误的概率大小。在统计检验中，一般采取限制第一类错误的概率，即选一个正数作为 α 的上限，这个正数通常称为检验水平或显著水平。其他常用的解决方法是：增大样本容量，尽量采用单边检验等。在实际应用时，必须根据客观事物的背景，恰当选取合适的 α 或合适的 β。

1.2 客户流失预测中分析两类错误的重要性

构建一个适用的客户流失预测模型是企业进行客户关系管理的有力保障。这里将需要评判的客户分为：流失客户和非流失客户。影响模型性能的主要因素之一是误判率，即：将流失客户评判为非流失客户和将非流失客户评判为流失客户这两类错误所引起的误判比率。根据“尽量使后果严重的错误成为第一类错误”的原则，将流失客户评判为非流失客户称为第一类错误，将非流失客户评判为流失客户称为第二类错误。

客户流失预测中出现的两类错误是统计学中两类错误在具体应用中的一种表现，因此具有上述两类错误的基本性质：由于样本的随机性及样本容量的有限性，无法同时控制犯两类错误的概率 α 和 β 都很小，大多数客户流失研究一味地强调模型整体的准确率，却忽视了两类错误对企业客户保持的不同影响，以至于实际应用效果并不理想，因此有必要对这两类错误进行深入的分析和探讨。就本文所讨论的问题而言，企业犯第二类错误至多增加一笔进行客户保持的费用，而犯第一类错误则面临着客户流失的巨大风险，因此第一类错误的危害性要远比第二类错误严重。Bhattacharya 等（1998）的研究指出：在客户流失预测中，第一类错误造成的损失为第二类错误损失的 5 倍～6 倍。两类错之间的制约关系及对实际问题的不同影响，要求企业在进行客户流失分析时，除了尽可能地规避风险较大的第一类错误，还要减少由两类错误所带来的期望损失。

2 基于改进支持向量机的客户流失预测模型

根据预测问题的不同，支持向量机（SVM）可分为支持向量分类机（Support Vector Classifier，SVC）和支持向量回归机（Support Vector Regression，SVR）。在标准的 C-SVC（Cortes 和 Vapnik，1995）中，$C>0$ 是一个常数，它控制对错分样本惩罚的程度，并且对于错分样本的惩罚相同。因此，没有对第一类错分和第二类错分的样本分别进行统计。为了刻画客户流失预测中两类错误存在的差异，考虑在标准 C-SVC 目标函数中，对它们分别采用不同的惩罚系数 C_1 和 C_2（$C_1, C_2>0$），并通过对 C_1、C_2 的调整来控制两类错误的分布。

设样本集为 (x_i, y_i)，$i=1,2,\cdots,n$，$x_i \in R^n$，$y_i \in \{1,-1\}$ 是类别标号。为了对第一类错分和第二类错分的样本类别采用不同的惩罚参数 C，令 C_1 是第一类错误的惩罚系数，即对正类点集的惩罚系数，C_2 是第二类错误的惩罚系数，即对负类点集的惩罚系数。此时 C-SVC 的原始问题形式变为

$$\min\phi(w) = \frac{1}{2}\|w\|^2 + \sum_{y_i=1} C_1\xi_i + \sum_{y_i=-1} C_2\xi_i \quad (3)$$

$$\text{s.t. } y_i[(w \cdot x_i)+b] \geqslant 1-\xi_i, \xi_i \geqslant 0, i=1,2,\cdots,n \quad (4)$$

引入 Lagrange 系数 $a_i \geqslant 0$，$\beta_i \geqslant 0$，构造 Lagrange 函数

$$L=\frac{1}{2}\|w\|^2+\sum_{y_i=1}C_1\xi_i+\sum_{y_i=-1}C_2\xi_i-\sum_{i=1}^{n}[a_i(y_i(w\cdot x_i+b)-1+\xi_i)+\beta_i\xi_i] \tag{5}$$

计算偏微分,并令偏微分等于零,得

$$\frac{\partial L}{\partial w}=w-\sum_{i=1}^{n}a_iy_ix_i=0 \tag{6}$$

$$\frac{\partial L}{\partial b}=-\sum_{i=1}^{n}a_iy_i=0 \tag{7}$$

$$\frac{\partial L}{\partial \xi_i}=\begin{cases}C_1-a_i-\beta_i=0,y_i=1\\C_2-a_i-\beta_i=0,y_i=-1\end{cases} \tag{8}$$

将式(6)、式(7)、式(8)代入式(5)得到对偶表达式

$$\min Q(a)=\frac{1}{2}\sum_{i,j=1}a_ia_jy_iy_j(x_i\cdot x_j)-\sum_{i=1}a_i \tag{9}$$

$$\text{s.t.}\quad \sum_{i=1}^{n}y_ia_i=0;\quad 0\leqslant a_i\leqslant C_1,y_i=1;\quad 0\leqslant a_i\leqslant C_2,y_i=-1 \tag{10}$$

一般情况下,该优化问题解的特点是大部分 a_i 将为零,其中不为零的 a_i 所对应的样本为支持向量。根据 KKT(Cortes 和 Vapnik,1995)条件,在鞍点有

$$a_i[y_i(w\cdot x_i+b)-1+\xi_i]=0,i=1,\cdots,n \tag{11}$$

$$(C_1-a_i)\xi_i=0,y_i=1 \tag{12}$$

$$(C_2-a_i)\xi_i=0,y_i=-1 \tag{13}$$

于是可得 b 的计算式

$$y_i\Big(\sum_{j=1}^{l}a_jy_j(x_j\cdot x_i)+b\Big)-1=0 \tag{14}$$

因此,可以通过任意一个支持向量求出 b 的值。最后得到最优分类函数为

$$f(x)=\text{sign}\Big\{\sum_{i=1}^{n}a_iy_i(x_i\cdot x)+b\Big\} \tag{15}$$

对于非线性问题,Vapnik 引入了核空间理论:将低维的输入空间数据通过非线性映射函数映射到高维属性空间,将分类问题转化到属性空间进行。可以证明,如果选用适当的映射函数,输入空间线性不可分问题在属性空间将转化为线性可分问题。这种非线性映射函数被称之为核函数(Vapnik,2004)。从理论上讲,满足 Mercer 条件的对称函数 $K(x,x')$都可以作为核函数。引入核函数后,最优分类函数为 $f(x)=\text{sign}\Big\{\sum_{i=1}^{n}a_iy_iK(x_i\cdot x)+b\Big\}$。

3 客户流失预测中两类错误的平衡控制实证研究

3.1 指标体系的建立

本文以电信业客户流失预测为例,来对两类错误的平衡控制进行研究。目前在电信业客户流失中采用的指标可分为:客户基本特征、客户通话行为、客户接触信息、客户签约信息、产品特征等。根据 Wei Yu 等的 Delta 客户流失管理战略模型(Wei 和 Jutla,2005),并综合考虑客户流失的各影响因素和数据的可获得性,选取产品特征、客户方案和客户信息三类指标共 52 个(见表 1)。

3.2 样本数据处理

本文数据来源于美国 DUKE 大学 TERADATA 客户关系管理中心,该数据包括 6 个月的客户数据,因此本研究将前 4 个月的客户数据作为输入指标,第 6 个月的客户状态为输出指标,中间预留 1 个月作为时间延迟。首先对有 30%以上缺失值的指标进行删除;然后通过指标均值对缺失项进行处理;最后进行稳健性处理,选用两倍、三倍标准差检验进行异常数据剔除。通过抽样,最终获得 5 617 个训练样本数据(其中 2 849 个为非流失客户,2 768 个为流失客户),3 107 个与训练样本相同时间下的测试样本数据(其中 2 705 个为非流失客户,402 个为流失客户),2 484 个比上述测试样本数据晚 2 个月~3 个月的未来时间下的未来客户样本数据(其中2 191 个为非流失客户,293 个为流失客户)。通过利用 SPSS11.5 对训练样本的数据进行因子分析,并用方差最大正交旋转法(VARIMAX),在特征值大于 1 的情况下,当前样本数据条件下的解释因子数为 10,其累计方差贡献为 86.16%,从而指

标被分成标准使用因子、生命周期使用因子、付费意愿因子、额外开销因子、使用时间长度因子、客户关怀因子、无线语音使用因子、产品价值因子、产品质量因子、使用变化因子(见表2)。

表1 客户流失预测指标

产品特征	手机价格、目前手机使用的时间、网络掉线语音通话次数、网络阻塞语音通话次数、网络掉线、阻塞语音呼叫总次数
客户方案	无应答语音通话次数、语音通话主叫次数、接听语音通话次数、成功的语音通话次数、客户关怀通话次数、客户关怀通话使用的取整时间长度、客户关怀通话使用的非取整时间长度、少于一分钟的被叫通话次数、成功的语音通话使用的非取整时间长度、接听语音通话使用的非取整时间长度、国际无线语音通话次数、国际无线语音通话使用的非取整时间长度、国内无线语音通话次数、国内无线语音通话使用的非取整时间长度、峰值语音通话次数、峰值语音通话使用的非取整时间长度、非峰值语音通话次数、非峰值语音通话使用的非取整时间长度、通话主叫次数、成功的通话次数、通话等待次数、总的服务月份数、客户生命周期内总的通话次数、客户生命周期内总的通话时间长度、客户生命周期内调整的总的通话时间长度、客户生命周期内调整的总的通话次数、客户生命周期内月均通话时间长度、客户生命周期内月均通话次数、过去3个月月均通话时间长度、过去3个月月均通话次数、月均通话费用、月均通话时间长度、月均连续消费费用、查号呼叫次数、月均超时使用时间长度、过去3个月内月均通话时间长度改变的百分数、过去6个月月均通话时间长度、过去6个月月均通话次数
客户信息	第一家庭成员的年龄、总的通话收益、客户生命周期内调整的总的通话收益、客户生命周期内月均通话收益、过去3个月月均通话收益、月均超额收益、月均语音通话超额收益、过去3个月内月均收益改变的百分数、过去6个月月均通话收益

表2 因子分析结果

因子	指标	因子载荷	累积方差贡献率(%)
标准使用因子	无应答语音通话次数	0.77	25.75
	语音通话主叫次数	0.87	
	接听语音通话次数	0.82	
	成功的语音通话次数	0.86	
	少于一分钟的被叫通话次数	0.82	
	接听语音通话使用的非取整时间长度	0.61	
	峰值语音通话次数	0.81	
	峰值语音通话使用的非取整时间长度	0.58	
	非峰值语音通话次数	0.81	
	通话主叫次数	0.87	
	成功的通话次数	0.86	
	通话等待次数	0.59	
	客户生命周期内月均通话次数	0.61	
	过去3个月月均通话次数	0.76	
	过去6个月月均通话次数	0.70	
生命周期使用因子	总的服务月份数	0.59	38.62
	客户生命周期内总的通话次数	0.86	
	客户生命周期内总的通话时间长度	0.82	
	客户生命周期内总的通话收益	0.79	
	客户生命周期内调整的总的通话收益	0.79	
	客户生命周期内调整的总的通话时间长度	0.82	
	客户生命周期内调整的总的通话次数	0.85	

续表

因子	指标	因子载荷	累积方差贡献率(%)
付费意愿因子	月均连续消费费用	0.73	47.21
	客户生命周期内月均通话收益	0.85	
	月均通话收益	0.72	
	过去 3 个月月均通话收益	0.73	
	过去 6 个月月均通话收益	0.76	
额外开销因子	月均超时使用时间长度	0.94	55.11
	月均超额收益	0.94	
	月均语音通话超额收益	0.94	
使用时间长度因子	客户生命周期内月均通话时间长度	0.61	62.50
	成功的语音通话使用的非取整时间长度	0.66	
	非峰值语音通话使用的非取整时间长度	0.69	
	月均通话时间长度	0.57	
	过去 3 个月月均通话时间长度	0.61	
	过去 6 个月月均通话时间长度	0.61	
客户关怀因子	客户关怀通话次数	0.82	68.37
	客户关怀通话使用的取整时间长度	0.96	
	客户关怀通话使用的非取整时间长度	0.93	
无线语音使用因子	国际无线语音通话次数	0.57	74.13
	国际无线语音通话使用的非取整时间长度	0.65	
	国内无线语音通话次数	0.78	
	国内无线语音通话使用的非取整时间长度	0.84	
产品价值因子	手机价格	0.64	79.22
	目前手机使用的时间	−0.74	
产品质量因子	网络阻塞语音通话次数	0.92	82.72
	网络掉线和阻塞语音呼叫总次数	0.72	
使用变化因子	过去 3 个月内月均通话时间长度改变的百分数	0.87	86.16
	过去 3 个月内月均收益改变的百分数	0.87	

3.3 支持向量机模型的构造

根据上述分析，构造样本集(x,y)。其中，x的维数为 10；y 是样本的类别属性，对于流失客户，$y=1$，对于非流失客户，$y=-1$。对于核函数和参数的选择，通过在 MATLAB6.5 上进行实验分析并与线性核函数、多项式核函数进行对比发现，采用参数 $\sigma=1.11$ 的径向基核函数 $K(x_i,x)=\exp\left\{-\frac{|x-x_i|^2}{2\sigma^2}\right\}$来构造 SVC 模型，在客户流失率第一个 10 分位上能取得较好效果(这里利用支持向量回归机对客户流失概率进行预测，模型也采用径向基核函数，其参数为 $\sigma=0.24$，$C=12$，$\varepsilon=0.02$)。对于 SVC 惩罚参数，选取适当的 C_1，C_2 来控制两类错误的分布，对惩罚系数相同($C_1=C_2$)及不同($C_1=YC_2$，$Y=0.2\sim2$)的情况进行了建模，其中 $Y(Y>0)$称为损失比例系数。

3.4 实证结果分析

表 4 对两类错误平衡控制下的测试集内客户流失率在第一个 10 分位上预测结果进行了对比。模型评价标准由表 3 可得：第一类错误率＝B/(A＋B)；第二类错误率＝C/(C＋D)；总的错误率＝(B＋C)/(A＋B＋C＋D)。另外，期望损失函数定义为(Joos 等，1998)

表3 分类矩阵 单位:个

样本中客户状态	预测流失	预测非流失
实际流失	A	B
实际非流失	C	D

表4 两类错误平衡控制下的预测结果对比

损失比例系数(Y)	当前时间				未来时间			
	第一类错误率	第二类错误率	总错误率	期望损失函数值	第一类错误率	第二类错误率	总错误率	期望损失函数值
0.2	0.972	0.000	0.339	143.759	0.974	0.010	0.157	144.359
0.4	0.528	0.040	0.210	79.308	0.921	0.067	0.198	138.254
0.6	0.454	0.198	0.287	73.170	0.790	0.252	0.335	124.507
0.8	0.514	0.429	0.494	89.071	0.447	0.471	0.468	80.439
1	0.065	0.703	0.481	30.999	0.342	0.667	0.617	70.872
1.2	0.028	0.832	0.552	29.451	0.184	0.795	0.702	51.398
1.4	0.009	0.886	0.581	28.283	0.105	0.885	0.766	42.451
1.6	0.009	0.931	0.610	29.652	0.053	0.914	0.782	35.643
1.8	0.009	0.965	0.632	30.686	0.026	0.938	0.798	32.379
2	0.009	0.980	0.642	31.143	0.026	0.962	0.819	33.109

$$EC = P_1 * L_1 * T_1 + P_2 * L_2 * T_2 \quad (16)$$

其中,EC 为损失的期望,P_1 为训练集中"1"类所占比率,P_2 为训练集中"-1"类所占比率,L_1 为将一个"1"类错分为"-1"类所造成的损失(文中为300美元),L_2为将一个"-1"类错分为"1"类所造成的损失(文中为50美元)(Athanassopoulos,2000),T_1 为第一类错误率,T_2 为第二类错误率。从表4中可以看出,在预测当前时间下的潜在流失客户时,随着损失比例系数的增加,第一类错误率逐渐下降,第二类错误率逐渐上升,总的错误率有上升的趋势,期望损失函数值起初逐渐减少,当减少到一定值时(当前时间为28.283,$Y=1.4$;未来时间为32.379,$Y=1.8$),到达最小,随后又逐渐增加。上述现象说明,在SVM预测模型中,通过调整两类错分的惩罚系数,可以控制两类错误的分布率,并且使总错误率达到最小的损失比例系数,却不能使期望损失函数值达到最小,即采用总错误率来评价预测模型是不妥的。因此,考虑实际应用背景,较理想的客户流失预测模型应尽可能地避免第一类错误引起客户流失带来的巨大风险,同时也要考虑到第二类错误引起客户保持费用的增加。

4 结束语

"拒真纳伪"两类错误是许多实际应用领域研究的重要问题,如何控制两类错误要视具体问题具体对待。就本文讨论的客户流失预测问题,引入了可以对两类错误进行控制的 C-SVC,并定量地研究它们之间的关系。实证结果表明:在SVM预测模型中,通过调整两类错分的惩罚系数,可以控制两类错误的分布率;采用文中定义的期望损失函数来评价预测模型更有实际意义。在上述研究中,其预测模型的范围仅限国外电信业,在后续工作中,可以考虑将其推广到国内电信业和其他客户信息数据的分类预测,这样,两类错误的平衡控制研究就会更有意义。

参考文献

[1] 蔡越江.论假设检验中的两类错误[J].数理统计与管理,1999,18(3):30-35.

[2] VAPNIK V N.统计学习理论[M].许建华,张学工,译.北京:电子工业出版社,2004.

[3] ATHANASSOPOULOS A D. Customer satisfaction cues to support market segmentation and explain switching behavior[J]. Journal of Business Research, 2000,47 (3):191-207.

[4] AU W H, CHAN K C C, YAO X. A novel evolutionary data mining algorithm with applications to churn prediction [J]. Evolutionary Computation, IEEE Transactions,2003,7(6):532-545.

[5] BHATTACHARYA C B. When customers are members: customer retention in paid membership contexts [J]. Journal of the Academy of Marketing Science,1998,26 (1):31-44.

[6] CHIH P W,CHIU I T. Turning telecommunications call details to churn prediction: a data mining approach [J]. Expert Systems with Applications, 2002,23(2):103-112.

[7] CORTES C,VAPNIK V. Support vector networks [J]. Machine Learning,1995,20(3):273-297.

[8] JOOS P,VANHOOF K,OOGHE H,et al. Credit classification: a comparison of logit models and decision trees:ECML 1998:Application of machine learning and data mining in finance: European Conference on Machine Learning, 1998 [C]. Chemnitz (Germany):1998.

[9] KIM H S,YOON C H. Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market [J]. Telecommunications Policy, 2004,28(9):751-765.

[10] NATH S V. Data warehousing and mining:customer churn analysis in the wireless industry [D]. Boca Raton,Florida:Florida Atlantic University,2003.

[11] ROSSET S, NEUMANN E. Integrating Customer Value Considerations into Predictive Modeling: ICDM 2003:Third IEEE International Conference on Data Mining,2003[C]. Florida (USA):2003.

[12] ULTCH A. Emergent self-organizing feature maps used for prediction and prevention of churn in mobile phone markets [J]. Journal of Targeting,2002,4(10):401-425.

[13] WEI Y, JUTLA D N, SIVAKUMAR S C. A churn-strategy alignment model for managers in mobile telecom: CNSR 2005: Networks and Services Research Conference, Proceedings of the 3rd Annual 16-18 May 2005[C]. Novia Scotia (Canada):2005.

[14] YAN L, MILLER D J, MOZER M C. et al. Improving prediction of customer behavior in nonstationary environments:IJCNN 2001. Proc. of the International Joint Conference on Neural Networks, July 15-19, 2001 [C]. Washington, DC:2001.

[15] YI M, Wan H, LI L, et al. Multi-dimensional model-based clustering for user-behavior mining in telecommunications industry: ICLMC 2004: proceeding of the third international conference on machine learning and cybernetics, August 26-29, 2004[C]. Shanghai:2004.

Tradeoff of Errors of Two Types in Customer Churn Prediction

Xia Guo-en, Jin Weidong

(School of Economics & Management, Southwest Jiaotong University)

Abstract Two types of errors about "rejecting true and accepting false" are inevitable in customer churn prediction. In this paper, a model based on SVM was used to predict customer churn. Through the analysis of different effects by two types of errors in customer churn prediction, issues related to the tradeoff between the errors in an American telecommunication carrier is studied. The results show that by adjusting the "loss-ratio-coefficient", the model is efficient in reducing expectation loss function value on the condition that errors of the two types can be controlled. This reflects the real essence of the problem and can be a powerful decision tool in customer churn.

Key Words Two Types of Errors, Customer Churn, Support Vector Machine