基于错分代价的用户换手机的分类器阈值和预期风险研究

王超发 孙静春
(西安交通大学管理学院,西安 710049)

摘要:传统的分类算法将正确预测和错误预测平等看待,忽略了人的主观因素,不能很好地对错误率进行控制。本研究基于某移动通讯公司西安分公司的用户消费数据,用引入错分代价后的 Logistic 模型研究了预测用户换手机的阈值及预期风险,研究发现:引入错分代价后的 Logistic 模型具有较好的分类效果;不同的错分代价对应不同的最优阈值,但预测准确率基本一致;用传统的阈值 0.5 进行分类不但降低了预测准确率还增加了预期风险;随着正负类别间的分类代价差异越大,分类器预测所面临的预期风险会上升;最优分类器的取值、最优阈值和预期风险三者之间具有动态平衡和相互制约关系。因而,该结论不但为数据挖掘人员提供多维度的分析框架,而且也为制造商和销售商提供决策参考。

关键词:错分代价;算法;手机用户;阈值

引 言

根据研究机构 Digitimes Research 的调查,2013 年中国市场智能手机销量为 3.2 亿台。如此大份额的市场,无论是对制造商还是销售商都具有重要的战略意义。因此,预测用户是否换手机不但能提高竞争力还能有效调配资源。事前分析的一个重要指标就是事件的发生概率,该指标的参考依据是分类算法的阈值,该阈值也是本研究对用户是否换手机概念的界定。本研究以某移动通讯公司西安分公司 2014 年的用户数据为样本,在引入错分代价后的 Logistic 模型的基础上分析了判定用户换手机的概率阈值(分类器阈值)和预期风险(错分代价的期望值)的变化特征。

国外关于分类算法及其应用研究主要表现在:David 等人从物流模式角度研究了外伤导致的病人死亡率,指出虽然 Logistic 模型能在一定程度上分离出有效的影响因素,但分类效果较差[1]。Farquad 和 Indranil 在非平衡数据集上比较了多层感知器、逻辑回归、随机森林和基于训练数据目标价值最大化的支持向量机等分类算法,认为支持向量机能有效处理数据不平衡带来的预测准确度下降问题[2]。然而,由于该模型较为复杂且不能对预测准确率进行有效控制,从而失去了实际应用价值[3]。Miwa 等人将 AdaBoost 算法和人脸检测算法相结合,以矩阵特征为弱学习器,最终得到一个很强的人脸检测器[4]。该方法的缺点在于将随机生成的样本整合到少数样本中去容易出现"并列现象"和过拟合现象[5]。造成这些现象的原因在于分类算法只重视聚类样本,而对次子集的描述和分析较少[6]。基于此,Fanyong 等人通过将代价敏感性学习纳入到生产商的分类研究中提升了分类效果[7]。Cuihuan 等人构建了基于 AdaBoost 算法的 Skin 模型并将其应用到了人脸检测中[8]。

国内关于分类算法及其应用研究具有代表性的有:张婷婷等人运用一种基于动态分类器集成选择的不完整数据分类方法 DOES-ID,分别在 UCI 客户分类数据集以及某券商用户数据集上进行了实证分析[9],该方法虽然有一定的实用性,但每个特征的权重赋值具有随意性。赵宇等人将半正定特征和支持向量机结合起来,将特征选择整合在数据分类过程中探讨了多维数据分类问题。然而,由于该方法仅仅涉及了单变量子集特征组合的情况,是否适应两种特征空间子集情形还有待讨论[10]。王雷等人针对突发暴恐事件的分级问题,提出了基于和声搜索算法优化的支持向量机的分级模型,指出该方法能有效优化支持向量机的参数,从而提高分类准确度[11]。商丽媛和谭清美利用支持向量机研究了地震灾害分级问题,认为支持向量机能够准确且有效

收稿日期:2016-09-22
基金项目:国家自然科学基金面上项目(71372164)。
作者简介:王超发(通讯作者),西安交通大学管理学院博士研究生;孙静春,西安交通大学管理学院教授,博士生导师,博士。

的为决策者提供分级决策依据[12]。以上方法的缺点在于分类算法过多地关注了量较小的样本,少数样本中的有价值信息没有得到充分挖掘。针对以上缺点,应维云虽然在平衡随机森林算法的基础上考虑了代价敏感[13],但该方法的泛化能力和效率比较差,造成这种现象的原因在于如何确定代价矩阵是这类研究的关键所在。邹鹏等人将代价敏感纳入到决策树模型中研究了不平衡数据的分类问题,并指出代价敏感决策树模型对用户细分具有很好的分类效果[14]。

通过文献调研发现,当前分类算法及其应用研究存在的问题主要表现在:(1)当前的分类算法将正确预测和错误预测平等看待且忽略了人的主观因素;(2)没有对分类器阈值和预期风险变化进行深入分析,不能很好地对错误率进行控制。造成这种现象的原因在于主观因素不仅体现在数据的主观性上还表现在错分代价设定的主观性。实际上,用户换手机概率不仅受客观因素的影响也受主观因素的影响,而主观因素较难把握。因此,将错分代价纳入到数据集的分类中,并基于最小化总体错分代价原理来设计分类器具有一定的实际意义。

LogitBoost 算法和考虑错分代价后的 Logistic 模型的损失函数

分类问题中的提升算法是统计学习中的常用方法,其中具有代表性的为 LogitBoost 算法。该算法通过改变训练样本的权重,在多个分类器上进行学习,最后将这些分类器进行线性组合得出一个强分类器,该强分类器能有效提高分类效果[15]。下面先介绍 LogitBoost 算法,然后给出考虑错分代价后的 Logistic 模型的损失函数。

1、LogitBoost 算法的构建

LogitBoost 算法的具体构建过程如下:

首先,选取样本集$(x_{i1},\cdots,x_{ik},y_i)$,$i=1,2,\cdots,M$ 和预测变量 $x=(x_{i1},\cdots,x_{ik})$。$Y_i=+1$ 表示正类(换手机),$Y_i=-1$ 表示负类(不换手机);

其次,对每个样本赋予同一权重:$w_i^{(0)}=1/M$,$i=1,2,\cdots,M$ 并基于该权重构建判别模型;$h_1(x)=0$;概率估计为 $\pi^{(0)}(x_i)=1/2$;取基础分类器 $g(x_i)=\mathrm{Ln}\dfrac{\pi(x_i)}{1-\pi(x_i)}$,并在判别模型中回代样本;

再次,不断更新样本权重:

(1)取作业应变量

$$A_i^{(n)}=\frac{Y_i^*-\pi^{(n-1)}(x_i)}{\pi^{(n-1)}(x_i)(1-\pi^{(n-1)}(x_i))}$$

和每个样本权重

$$B_i^{(n)}=\pi^{(n-1)}(x_i)(1-\pi^{(n-1)}(x_i))$$

这里,n 为迭代轮数,$n=1,2,\cdots,N$,$Y_i^*=(Y_i+1)/2$,$\pi(x)$ 为后验概率,$\pi(x)=\pi(Y=+1\,|\,x)$,$1-\pi(x)=\pi(Y=-1\,|\,x)$。

(2)基于加权最小二乘法和样本权重 $B_i^{(n)}$ 拟合出弱分类器

$$g^{(n)}(x)=\arg\min_g\sum_{i=1}^{T}B_i^{(n)}\,(A_i^{(n)}-g(x_i))^2$$

(3)计算每一次的 $\pi^{(n)}(x_i)$ 和 $h_1^{(n)}(x_i)$

$$\frac{e^{h_1^{(n)}}(x_i)}{e^{h_1^{(n)}}(x_i)+e^{-h_1^{(n)}(x_i)}}\to\pi^{(n)}(x_i)$$

$$h_1^{(n-1)}(x_i)+\frac{1}{2}g^{(n)}(x_i)\to h_1^{(n)}(x_i)$$

直至 N 次结束,N 代表回代的轮数。

(4)输出最终的分类器 $h_1(x)$

$$h_1(x_i)=\sum_{n=1}^{N}g^{(n)}(x_i)$$

根据 $h_1(x_i)$ 的正负对第 i 个样本进行类别划分

$$sign[h_1(x_i)] = sign[\sum_{n=1}^{N} g^{(n)}(x_i)]$$

2、考虑错分代价后的 Logistic 模型的损失函数

在利用分类器对数据集进行预测时，通过引入错分代价矩阵，一方面可以提高重要样本的识别率，另一方面有利于降低预测错误带来的损失[16]。下面在传统 Logistic 模型的基础上引入错分代价，并探讨错误分类方面的损失。

假设用户换手机的驱动力不仅受到客观因素的影响还受主观因素的影响，则引入代价矩阵后的损失函数（以下简称为错分代价）可表示为：

$$L(c_Y, Y, h_2(x)) = c_Y \cdot \mathrm{Ln}(1 + e^{-Yh_2(x)})$$

其中，c_Y 表示样本类别为 Y 时的错分代价；$c_Y = \begin{cases} \mathrm{cost}_+, Y=1 \\ \mathrm{cost}_-, Y=-1 \end{cases}$，这里，$\mathrm{cost}_{-+} - \mathrm{cost}_{++} = \mathrm{cost}_+$，$\mathrm{cost}_{+-} - \mathrm{cost}_{--} = \mathrm{cost}_-$，$\mathrm{cost}_+$ 和 cost_- 分别代表数据集由正类和负类组成时对应的正类、负类的误分类代价；$h_2(x)$ 表示分类器算法。现定义代价矩阵如下：

$$\begin{pmatrix} \mathrm{cost}_{++} & \mathrm{cost}_{+-} \\ \mathrm{cost}_{-+} & \mathrm{cost}_{--} \end{pmatrix}$$

其中，cost_{ij} 代表将 j 类样本预测为 i 类样本的分类代价。该矩阵的具体生成为：随机生成一个 k（类别数目）维的正实数作为等式 $w_1/w_2 = \mathrm{cost}_{+-}/\mathrm{cost}_{-+}$ 的解（这里 w_i 表示第 i 类的样本的权重），再随机生成一个正实数作为 cost_{ij}，而 cost_{ji} 的值可根据该等式得到，其中所有的随机数在区间 $[1,k]$ 内。另外，根据 Bayes 决策理论，最优决策应最小化预期风险分类代价[17]，即给定样本 $x = x(x_{i1}, x_{i2}, \cdots, x_{ik})$，若式

$$\mathrm{cost}_{++}\pi(x) + \mathrm{cost}_{+-}(1 - \pi(x)) \leq \mathrm{cost}_{-+}\pi(x) + \mathrm{cost}_{--}(1 - \pi(x))$$

成立，则预测其类为正，否则，预测其类为负。

引入错分代价后的 Logistic 模型的阈值和预期风险分析

利用阈值分类产生的预期风险的大小是判别分类器是否合适的一个重要衡量指标[18]。下面通过分析引入错分代价后的 Logistic 算法分类器来进一步选取最佳阈值。若记 $\pi(+|x)$ 和 $\pi(-|x)$ 分别为正类和负类的先验概率，则可得给定样本空间后的分类器 $h_2(x)$ 对该样本空间进行预测产生的预期风险：

$$E[L(c_Y, Y, h_2(x))] = \pi(+|x) \cdot \mathrm{cost}_{+-} \cdot \mathrm{Ln}(1 + e^{-Yh_2(x)}) + \pi(-|x) \cdot \mathrm{cost}_{-+} \cdot \mathrm{Ln}(1 + e^{Yh_2(x)})$$

对上式分别求关于函数 $h_2(x)$ 的一、二阶导数得：

$$\frac{\partial E[L(c_Y, Y, h_2(x))]}{\partial h_2(x)} = \pi(+|x) \cdot \mathrm{cost}_{+-} \cdot \frac{-1}{e^{h_2(x)} + 1} + \pi(-|x) \cdot \mathrm{cost}_{-+} \cdot \frac{e^{h_2(x)}}{1 + e^{h_2(x)}} \tag{1}$$

$$\frac{\partial^2 E[L(c_Y, Y, h_2(x))]}{\partial h_2^2(x)} = \pi(+|x) \cdot \mathrm{cost}_{+-} \cdot \frac{e^{h_2(x)}}{[1 + e^{h_2(x)}]^2} + \pi(-|x) \cdot \mathrm{cost}_{-+} \cdot \frac{e^{h_2(x)}}{[1 + e^{h_2(x)}]^2} > 0$$

根据最小化总体错分代价原理，再令（1）式为零且由函数性质可知，预期风险损失在函数 $h_2(x)$ 的定义域内存在最小值。那么，根据（1）式可得引入错分代价后 Logistic 算法的最优分类器（可使预测损失最小）和错分代价分别为：

$$h^* = \mathrm{Ln}\frac{\mathrm{cost}_{+-} \cdot \pi(+|x)}{\mathrm{cost}_{-+} \cdot (1 - \pi(+|x))} \tag{2}$$

$$L(c_Y, Y, h^*(x)) = c_Y \cdot \mathrm{Ln}[1 + e^{-Y\mathrm{Ln}\frac{\mathrm{cost}_{+-} \cdot \pi(+|x)}{\mathrm{cost}_{-+} \cdot (1-\pi(+|x))}}]$$

另外，若将 $E[L(c_Y, Y, h_2(x))]$ 看做 $\pi(+|x)$ 的函数，则

$$E[L(c_Y, Y, h_2(x))] = \pi(+|x) \cdot \mathrm{cost}_{+-} \cdot \mathrm{Ln}(1 + \frac{1-\pi(+|x)}{\pi(+|x)}) + (1-\pi(+|x)) \cdot \mathrm{cost}_{-+} \cdot \mathrm{Ln}(1 + \frac{\pi(+|x)}{1-\pi(+|x)})$$

对上式分别求关于概率 $\pi(+|x)$ 的一、二阶导数得：

$$\frac{\partial E[L(c_Y,Y,h_2(x))]}{\partial \pi(+|x)} = -\cos t_{+-}\cdot \mathrm{Ln}(\pi(+|x)) - \cos t_{+-} + \cos t_{-+}\cdot \mathrm{Ln}(1-\pi(+|x)) + \cos t_{-+}$$

$$\frac{\partial^2 E[L(c_Y,Y,h_2(x))]}{\partial \pi^2(+|x)} = -\frac{\cos t_{+-}}{\pi(+|x)} - \frac{\cos t_{-+}}{1-\pi(+|x)} < 0 \qquad (3)$$

由(3)式可知,预期风险在 π 的定义域内存在最大值。再令 $\cos t_{+-} = \lambda \cos t_{-+}$,这里 λ 为分类代价间的关系系数(以下简称关系系数)。那么,可得最优阈值 π^* 和该关系系数满足 $1-\pi^* = \pi^{*\lambda} e^{\lambda-1}$。因此,若不考虑错分代价或认为代价相同时(λ=1):①阈值 $\pi^* = 0.5$ 对应最优分类器的预期风险最大;②由(2)式可知,此时的最优分类器为零,即此时的最优分类器不能合理区分样本类别。若不考虑错分代价,预期风险的大小由最优阈值和关系系数共同决定,即根据代价矩阵得到的最优阈值点不一定为预期风险最大点。现绘制不同最优阈值下的无代价、正类大于负类和负类大于正类的预期风险曲线,见图 1:

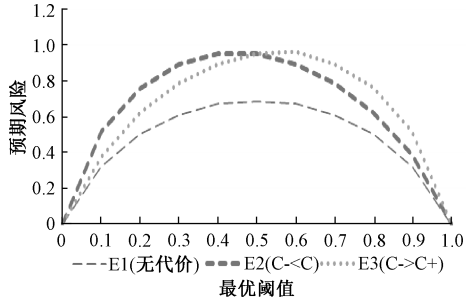

图 1　引入错分代价后预期风险比较

由图 1 可看出,无代价情形下的预期风险对应的最优阈值点在 0.5 处;当正类样本代价高于负类时,较高的预期风险对应较小的分类阈值;当负类样本代价高于正类时,较高的预期风险对应较大的分类阈值。

实证研究

1、数据来源、变量设置和说明

本研究所用的数据来自某移动通讯公司西安分公司的内部资料。由于原始数据存在数据量大、属性繁多、数据缺失和交互作用等,因此在对数据进行分析之前有必要对原始数据中的变量进行精简,剔除冗余变量。鉴于此,首先本研究采用专家评判法挑选出主要变量(具体方法为:先对每个变量定出评价等级,每个等级的标准用分值表示;然后以此为基准,由专家对评价对象进行分析和评价,确定各个变量的分值,采用加法评分法求出评价对象的总分值,将总分值低的变量剔除);其次,对筛选的变量进行相关性分析,检验依存关系较强的变量(具体方法为:对第一步筛选出的变量做正态性检验,再做散点图,初步判断变量之间是否具有相关性,然后对相关性较强的变量以数据的完整性为基本原则剔除掉数据不完整的变量);最后通过相关性分析确定了在 0.01 置信水平下的相关系数均没有超过 0.75 的 17 个备选变量,共计30 935个用户数据样本。Logistic 模型是将用户换手机的概率问题转化成通过一定时期内用户的消费数据、在用手机的一些特征数据(手机屏幕尺寸、是否智能、屏幕分辨率和操作系统等)以及用户个人的主观因素变量(手机品牌、价格偏好和手机品牌忠诚度)预测将来换手机的概率。其基本思想是设变量 x_{ij} 为用户 i 的第 $j(1\leq j\leq k$,这里 k 表示选取的变量个数)个消费变量的取值,那么预测该用户换手机的模型为:

$$\mathrm{Ln}\frac{\pi_i(Y=1|x_{i1},x_{i2},\cdots,x_{ik})}{1-\pi_i(Y=1|x_{i1},x_{i2},\cdots,x_{ik})} = \alpha+\beta_1 To+\beta_2 Fr+\beta_3 Av+\beta_4 Ef+\beta_5 Fe+\beta_6 Al+\beta_7 Sc+\beta_8 Ne$$
$$+\beta_9 In+\beta_{10} Re+\beta_{11} Pt+\beta_{12} Me+\beta_{13} Op+\beta_{14} Br+\beta_{15} Pf+\beta_{16} Lo+\beta_{17} Pe \qquad (4)$$

这里,$\pi_i = (Y=1|x_{i1},x_{i2},\cdots,x_{ik})$ 为用户 i 换手机的概率,α 和 β_j 分别为模型系数(可通过数据拟合得到),(4)式中的变量定义和赋值如表 1 所示。另外,为了体现人的主观因素,本研究用在用手机品牌、价格偏好和手机品牌忠诚度作为衡量用户换手机的主观因素。首先,根据慢慢买网(http://www. Manmanbuy. com/distop_57. aspx)统计手机品牌排行榜的前八位主要手机品牌作为本研究认定的手机品牌(按照销量从大到小依次分别为苹果,荣耀,小米,魅族,华为,三星,oppo 和 vivo)。具体的影响因素和其反映的实际意义如表 2 所示。

表1 变量定义和赋值

变量名	变量说明
Y	1表示用户换手机,-1表示用户不换手机;
To	总通话时长(分钟/月);
Fr	总通话轮数(通话轮数/月);
Av	平均通话时间(分钟/次);
Ef	平均总流量(兆/月);
Fe	平均流量费用(元/月);
Al	平均总费用(元/月);
Sc	在用手机屏幕大小(寸);
Ne	平均联网时长(月);
In	是否智能(1表示智能手机,0表示非智能手机);
Re	屏幕分辨率(1表示高分辨率,0表示低分辨率),(像素大于240×320为高分辨率,小于240×320为低分辨率);
Pt	平均赠送流量(兆/月);
Me	平均套餐费用(元/月);
Op	操作系统(1表示Android(谷歌)、iOS(苹果),2表示windows phone(微软)、Symbian(诺基亚),3表示BlackBerry OS(黑莓)、windows mobile(微软));
Br	手机品牌(1表示用户在用手机品牌为苹果,荣耀和小米,2表示用户在用手机品牌为魅族,华为和三星,3表示用户在用手机品牌为oppo和vivo);
Pf	价格偏好(元);
Lo	手机品牌忠诚度(1表示用户一直采用同一手机品牌,2表示用户偶尔变换(变换三次以内)手机品牌,3表示经常变换(变换四次以上)手机品牌);
Pe	平均换机周期(年);

表2 影响因素和其反映的实际意义

影响因素	反映的实际意义
总通话时长、总通话轮数、平均通话时间	反映了用户通过语音方式获取信息时用手机的强度;
平均总流量、平均流量费用、平均总费用、平均联网时长	反映了用户通过非语音方式获取信息时用手机的强度;
在用手机屏幕大小	屏幕尺寸小于等于2.5寸为小屏,屏幕尺寸介于2.5和4.5寸为中屏,屏幕尺寸大于等于4.5寸为大屏。屏幕大小反映了用户在用手机的外形和功能特征;
屏幕分辨率	反映了手机屏幕的清晰程度;
平均赠送流量、平均套餐费用	反映了运营商根据用户的实际利用资源情况所设置的优惠标准和收费标准;
是否智能	反映了将手机、音视频播放及其他个人数据处理等安装应用功能的整合能力;
操作系统	反映了手机处理信息的性能和效率;
手机品牌、品牌忠诚度	反映用户对手机的感知或手机对用户的吸引力;
价格偏好	反映用户对手机价值的主观预期或价值取向;
平均换机周期	反映手机使用年限的长度;
响应变量	反映用户是否换机的概率;

2、变量的描述性分析

除赋值的五个虚拟变量外(见表1),对剩余12个变量进行描述性分析,其结果见表3。从表3知各变量的变异系数均小于0.1,这说明数据值较集中。平均联网时长和在用手机屏幕大小的变异系数较小且取值分布较为接近,这可能与该地区用户的手机接受程度和屏幕大小基本相同的手机进入该地区市场的步调基本一致有关。而平均赠送流量、平均通话时间、平均总流量和平均套餐费用的变异系数差异较大,这与用户个人联网和通讯需求量有关。比较平均值和中位数发现,除总通话时长外,其他变量的中位数均接近于其平均数,说明这些数据值基本集中在同一区域。以上分析表明,本研究选取的变量测定值均在可接受的范围内且用户联网和通讯需求差异较大,这些数据具有一定的特殊性和代表性。

表 3　用户获取网络资源和支付费用的连续型变量的描述性统计

变量	取值范围	平均值	变异系数	中位数	10 百分位数
To	(1.01, 5932.81)	6535.74	0.0084	518.23	1115.04
Fr	(9, 45524.44)	2976.35	0.0090	2285.00	511.50
Av	(0.1, 19.21)	2.34	0.0347	2.19	1.47
Ef	(0, 6097.96)	130.87	0.0143	81.68	0.01
Fe	(0, 735.94)	25.95	0.0105	19.37	0.50
Al	(0, 9358.15)	100.30	0.0090	80.16	27.47
Sc	(1.7, 8)	3.85	0.0020	3.70	3.20
Ne	(1.5, 47.51)	14.52	0.0062	11.50	5.50
Pt	(0, 10289.67)	197.37	0.0226	112.50	0.83
Me	(0, 226.67)	9.09	0.0119	5.83	10.22
Pf	(0,6800)	1824	0.0058	1903	877.91
Pe	(0,8)	1.671	0.0075	1.435	0.883

注:为了消除各变量的量纲和平均数所不同对数据变异程度比较的影响,本研究使用中位数和变异系数作为比较指标。

3、关系系数敏感度分析和算法判别结果比较

从前文分析可知:(1)错分代价取值具有一定的主观性且实际中的最优阈值不可能是一个固定的点,而是在一个区间内变动;(2)将换手机的用户分到不换手机的用户群体造成的损失大于将不换手机的用户分到换手机用户群体造成的损失。因而,关系系数取值区间为 (0,1)。为了进一步分析分类代价的设定对最优阈值和预测结果的影响机理,本研究结合前文对最优阈值和预测准确率的分析过程给出不同(本研究取以 0.1 为步长)关系系数(在这里用关系系数代替正负类错分代价有利于降低分析维度)变化下的最优阈值和预测准确率变化趋势,具体结果如表 4 所示。

表 4　不同关系系数变化下的最优阈值和预测准确率取值

关系系数	最优阈值	预测准确率(%)
0	0.63	88.3
0.1	0.25	88.5
0.2	0.11	88.1
0.3	0.07	88.9
0.4	0.04	88.3
0.5	0.01	88.9
0.6	0.13	88.8
0.7	0.27	88.1
0.8	0.31	88.4
0.9	0.43	88.8
1.0	0.50	88.7

从表 4 可知,随着关系系数的增大最优阈值先减小后增大,而预测准确率基本保持不变。当关系系数等于 0.5 时最优阈值取最小值 0.01,预测准确率为 88.9%。基于该结论,本研究取关系系数为 0.5 分析分类器阈值、预期风险和最优分类器的变化特征,在此不妨取 $cost_+ = 2, cost_- = 1$。下面基于该分类代价,通过对数据进行 Logitboost 算法、Logistic 模型和考虑错分代价后的 Logistic 模型的判别效果进行分析,进一步确定需要选取的分类算法和阈值。

表 5　LogitBoost 算法、Logistic 模型和考虑错分代价后的 Logistic 模型效果比较

指标	LogitBoost 算法			Logistic 模型	考虑错分代价的 Logistic 模型
	10N	40N	70N		
正确率	0.655	0.721	0.673	0.249	0.862
换手机的用户不换手机的可能性	0.611	0.716	0.666	0.231	0.811
不换手机的用户换手机的可能性	0.688	0.689	0.677	0.235	0.724
灵敏度	0.644	0.738	0.612	0.277	0.813
特异度	0.664	0.734	0.605	0.313	0.845
Kappa 值	0.623	0.691	0.652	0.356	0.788

从表 5 可以看出,应用 LogitBoost 算法进行分类时,随着迭代轮数的增加其最终预测模型的正确率、换手机的用户不换手机的可能性、不换手机的用户换手机的可能性、灵敏度、特异度以及一致性指标(Kappa 值)先

增加后减小。当迭代到 40 轮时,各值达到最高,继续迭代判别效果反而下降,这说明 LogitBoost 重复轮数过多会造成过度拟合。另外,从表 5 也可以看到,Logistic 模型对用户换手机的预测较为不准确,LogitBoost 算法比较准确,考虑错分代价的 Logistic 模型最准确。因此,本研究选取考虑错分代价后的 Logistic 模型作为分类模型。

4、考虑错分代价后的 Logistic 模型阈值分析

由于代价敏感损失设计具有一定的准则[17]:(1)考虑错分代价的代价敏感损失函数需满足 Bayes 一致性;(2)代价敏感决策函数对应的条件代价敏感风险在 Bayes 分类边界$\{x \mid \pi(+\mid x)=\frac{cost_-}{cost_+ + cost_-}\}$处取得最大值。基于以上准则,本研究仍取 $cost_+=2$,$cost_-=1$ 作为下文的参数值,在默认阈值 $\pi^*=0.5$ 处输出的受试者工作特征曲线(ROC)并得出曲线与横轴所围成的面积 AUC 值为 0.7378(该值大于标准判别值 0.7),这说明考虑错分代价的 Logistic 模型具备较好的分类能力。另外,对于分类器而言,预测准确率包含正类样本的预测准确率与负类样本的预测准确率两部分。在实际应用中,研究者常常希望在负类样本预测准确率不低的情况下正类样本的预测准确率越高。基于此,本研究应用约登指数(Youden's indx)方法[19],即选取 TPR-FPR 的最大值所对应的阈值作为最优阈值,这里 TPR 是指正类样本被预测正确的数量占所有正类样本的比例,FPR 是指负类样本被预测为正类样本的数量占所有负类样本的比例。通过以 0.05 为步长在 ROC 曲线上移动阈值,得到预测模型不同类别的正确率(由 FPR 的定义可知 1-FPR 表示负类分类正确率)以及 TPR-FPR,其结果如图 2 所示。

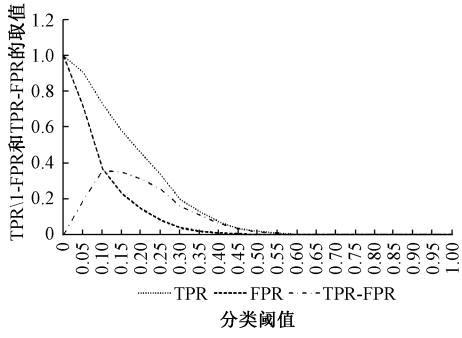

图 2 TPR、1-FPR、TPR-FPR 取值在不同阈值下的变化

从图 2 可以看出 TPR-FPR 取值在阈值区间(0.1, 0.2)内取最大值,此时模型的分类效果最好。为了进一步确定具体的最优阈值,再在区间(0.1,0.2)内以 0.01 为步长确定具体的最优阈值,其结果如表 6 所示。

表 6 阈值在区间(0.1,0.2)内变化时的 TPR、1-FPR、TPR-FPR 取值

p_0	TPR	1-FPR	TPR-FPR
0.10	0.734	0.266	0.337
0.11	0.718	0.282	0.339
0.12	0.703	0.297	0.342
0.13	0.688	0.312	0.344
0.14	0.674	0.326	0.346
0.15	0.660	0.340	0.346
0.16	0.647	0.353	0.348
0.17	0.632	0.368	0.344
0.18	0.618	0.382	0.346
0.19	0.605	0.395	0.346
0.20	0.592	0.408	0.345

由表 6 可知,当 $\pi^*=0.16$ 时,分类器的分类效果最优。若以该最优阈值点作为判断标准,以正样本准确率优先于负样本准确率为准则输出的混淆矩阵表明换机用户预测准确率为 75.6%,非换机用户预测准确率为 8.7%,模型总体预测准确率达到了 88.2%。所以,由 Bayes 分类边界和两类错分代价关系可知,当最优阈值在区间(0.1,0.2)内变化时,分类代价比值在区间(1/9,1/4)内时模型的分类效果较好。下面在该代价区间内分析预期风险及其最优分类器的变化趋势(分别见图 3 和图 4)。

从图 3 可见,随着分类代价比值的增大,预期风险(同一阈值下)增大;同一分类代价比值下的预期风险,随着阈值的增大表现出先增大后减小的趋势;不同错分代价下的预期风险最大值远远偏离了传统的阈值 0.5,这说明采用传统的对阈值的赋值方法不能正确预测用户是否换手机;当以较低的阈值使更多的正类样本被预测正确时,分类器也面临较高的预期风险;随着正负类别间的分类代价差异的增加,分类器预测所面临的预期风险会上升;当最佳阈值在(0.1,0.3)区间内时,错分代价对预期风险的影响显得尤为重要。

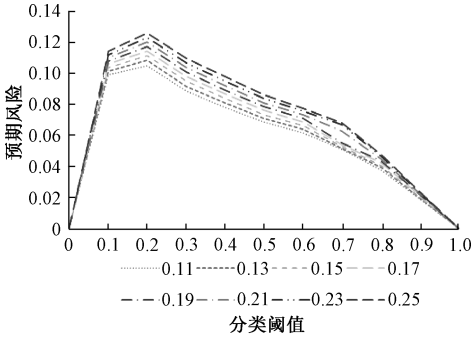

图 3　不同分类代价比值下的预期风险的变化趋势

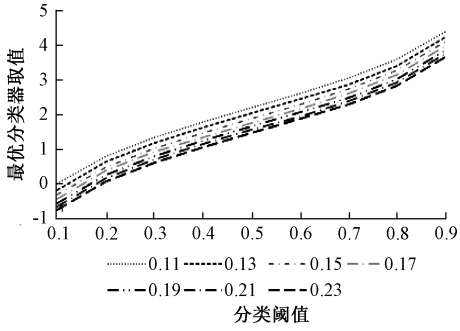

图 4　不同分类代价比值下的最优分类器的变化趋势

从图 4 可见,随着分类代价比值的增加,最优分类器取值(同一阈值下)变小;同一分类代价比值下的分类器取值,随着阈值的增加而增加;随着分类器取值不断地向 0 以下延伸,分类器越来越注重负类样本的预测,使得更多的样本被预测为负类;因而,最优分类器的取值、最优阈值和预期风险三者之具有动态平衡和相互制约的关系。

结　论

传统的分类算法将正确预测和错误预测平等看待,忽略了人的主观因素,不能很好地对错误率进行控制。本研究在 Logistic 模型的基础上引入错分代价机制,以某移动通讯公司西安分公司的手机用户消费数据为基础,将当前较为流行的分类算法——LogitBoost 算法作为参考标准,从引入错分代价的 Logistic 模型出发,对引入错分代价后的分类器的阈值和预期风险进行了分析,研究结果表明:(1)在判定用户是否换手机时,LogitBoost 算法的分类效果介于传统 Logistic 模型和引入错分代价后的 Logistic 模型之间;(2)通过设定不同错分代价虽然能改变 Logistic 模型判别的最优阈值,但不能明显改变预测准确率,所以仅用传统的阈值 0.5 进行分类不但降低了预测准确率还增加了预期风险;(3)当正类代价高于负类代价时,较高的预期风险对应较小的分类阈值,反之亦然;(4)当以较低的阈值使更多的正类样本被预测正确时,分类器也面临较高的预期风险,随着正负类别间的分类代价差异越大,分类器预测所面临的预期风险会上升;(5)从模型可以看到,不但用户是否换手机受在用手机特征和主观认识的双重影响,而且最优分类器的取值、最优阈值和预期风险三者之间还具有动态平衡和相互制约的关系。

以上结论不但为数据挖掘人员提供多维度的分析框架,而且也为制造商和销售商提供了决策参考。针对这些结论,为了更好地提高手机制造商和销售商的竞争力和资源调配能力,本研究提出以下建议:

(1)在对用户进行分类时采用引入错分代价后的 Logistic 模型较为妥当,分析时应考虑分类器阈值、预期风险和最优分类器的取值来综合评价模型的性能,这样才能避免将换手机的用户错误划分到不换手机的用户群体中去,从而造成重要用户的流失,同时也防止了将不换手机的用户错误划分到换手机的用户群体中去造成的用户保持和发展成本的增加。(2)制造商、销售商要明确识别自己和竞争对手之间的差异,在手机特征方面,应重视手机屏幕尺寸、智能化、屏幕分辨率和操作系统等方面的设计;在用户关系经济性方面,企业可基于结论(3)和(4)在保证预期风险较低的情况下力求识别用户的需求特征,然后再关注对企业价值最大的高端群体。(3)借助结论(4)的变化机制,生产商和销售商可在原有用户规模的基础上能较多地关注那些换手机频繁和有潜在换手机趋势的用户,也可最大限度地识别用户对手机的特征和价值需求,从而保证有限资源最大限度的分配给这些用户,实现企业客户关系价值最大化。(4)为了节省资源,制造商和销售商可针对不换手机的群体进一步采取对策,例如,制造商和销售商可将不换手机的群体根据换手机的概率大小划分为接近换手机、较接近换手机和不换手机三个层次,然后进行有针对性的研发和销售。(5)在影响用户主观认识

方面,除手机品牌外(虽然制造商不能随意更换手机品牌,但销售商可在多个品牌之间进行选择),还应重视用户对手机品牌忠诚度和价格偏好等,例如,可在实践中进行调查问卷或者市场调研搜集用户对哪些手机品牌比较青睐以及大部分群体对手机价格的偏好等信息,然后可制定有针对性的生产和销售策略。

参考文献:

[1] David E. C., Edward L., Hannan C. W. Predicting Risk-Adjusted Mortality for Trauma Patients: Logistic Versus Multilevel Logistic Models[J]. Journal of the American College of Surgeons, 2010,211(2):224-231

[2] Farquad M., Indranil B. Preprocessing Unbalanced Data Using Support Vector Machine[J]. Decision Support Systems, 2012,53(1):226-233

[3] Zhou T., Shan H., Banerjee A. Kernelized Probabilistic Matrix Factorization: Exploiting Graphs and Side Information[C]. Proceedings of the 12th SIAM International Conference on Data Mining. Philadelphia: SIAM, 2012

[4] Miwa S., Hirai T., Sumi K. Robust Face Detection Using One-classestimation and Real AdaBoost[J]. Electronics and Communications in Japan, 2014,97(7):39-47

[5] Mierswa I. Controlling Overfitting with Multi-objective Support Vector Machine[C]. Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation. In GECCO, 2007

[6] Fanyong C., Jing Z., Cuihong W., et al. Large Cost-sensitive Margin Distribution Machine for Imbalanced Data Classification[J]. Neurocomputing, 2017,224(8):45-57

[7] Fanyong C., Jing Z., Cuihong W. Cost-Sensitive Large Margin Distribution Machine for Classification of Iimbalanced data[J]. Pattern Recognition Letters, 2016,80(1):107-112

[8] Cuihuan D., Hong Z., Liming L., et al. Face Detection in Video Based on AdaBoost Algorithm and Skin Model[J]. The Journal of China Universities of Posts and Telecommunications, 2013,20(1):6-24

[9] 张婷婷,贺昌政,肖进. 基于动态分类器集成选择的不完整数据客户分类方法实证研究[J]. 管理评论, 2012,24(6):83-123

[10] 赵宇,黄思明,陈锐. 数据分类中的特征选择算法研究[J]. 中国管理科学, 2013,21(6):38-46

[11] 王雷,王欣,赵秋红. 基于和声搜索算法优化支持向量机的突发暴恐事件分级研究[J]. 管理评论, 2016,28(8):125-132

[12] 商丽媛,谭清美. 基于支持向量机的突发事件分级研究[J]. 管理工程学报, 2014,28(1):119-123

[13] 应维云. 随机森林方法及其在客户流失预测中的应用研究[J]. 管理评论, 2012,24(2):140-145

[14] 邹鹏,莫佳卉,江亦华,等. 基于代价敏感决策树的客户价值细分[J]. 管理科学, 2011,24(2):20-29

[15] Marc G. LogitBoost Autoregressive Networks[J]. Computational Statistics and Data Analysis, 2017,112(3):88-98

[16] Cao P., Zhao D. Z., Zaiane O. An Optimized Cost-sensitive SVM for Imbalanced Data Learning[C]. Advances in Knowledge Discovery and Data Mining Lecture Notes in Computer Science. Berlin: Springer, 2013

[17] Guoqing Z., Huaijiang S., Zexuan J., et al. Cost-sensitive Dictionary Learning for Face Recognition[J]. Pattern Recognition, 2016,60(7):613-629

[18] Cao P., Zhao D. Z., Zaiane O. A PSO-based Cost-sensitive Neural Network for Imbalanced Data Classification[C]. Trends and Appli-canons in Knowledge Discovery and Data Mining Lecture Notes in Computer Science. Berlin: Springer, 2013

[19] Jingjing Y., Lili T. Joint Inference about Sensitivity and Specificity at the Optimal Cut-off Point Associated with Youden Index [J]. Computational Statistics and Data Analysis, 2014,77(14):1-13

Research of Classifier Threshold and Expected Risk of Mobile Phone Replacement Based on the Misclassification Cost

Wang Chaofa and Sun Jingchun

(School of Management, Xi'an Jiaotong University, Xi'an 710049)

Abstract: The traditional classification algorithm treats the correct prediction and the error prediction equally, ignores the subjective factors and can't control the error rate well. Based on the users' consumption data selected from Xi'an branch of a mobile communication company, this paper studies the threshold and expected risk of forecasting mobile phone replacement by using Logistic model with misclassification cost. We find that: the Logistic model with misclassification cost has a good classification effect; different misclassification costs correspond to different optimal thresholds, but the prediction accuracy is basically the same; classification with a traditional threshold of 0.5 not only reduces the accuracy of the forecast but also increases the expected risk; the greater the difference in classification costs between positive and negative categories, the higher the expected risk for the classifier to predict; there is a dynamic equilibrium and mutual restraint between the optimal classifier's value, the optimal threshold and the expected risk. Thus, these results not only provide a multi-dimensional analysis framework for data mining researchers, but also provide a decision-making reference for manufacturers and vendors.

Key words: misclassification cost, algorithms, mobile users, threshold