服务号码捆绑特征在离网预测系统中的应用

张正卿[1] 朱奕健[1] 白瑞瑞[2] 黄一清[2] 严建峰[2]
(中国联合网络通信有限公司上海市分公司 上海 200050)[1]
(苏州大学计算机科学与技术学院 苏州 215006)[2]

摘 要 用户流失问题是电信运营商面临的亟待解决的问题,针对不同的场景,业界研究开发了多个用户离网预测系统。服务号码捆绑指用户在使用运营商服务期间,与银行、电商、便利店等第三方服务提供商通过绑定手机号产生联系。通过研究发现用户在服务存续期间普遍会绑定多种第三方服务提供商,这些商家会不定时给用户推送短信,当用户即将流失时,多数用户会逐渐取消这类服务的绑定。因此,服务号码捆绑特征对于离网用户的甄别起到了重要的作用。采用随机森林算法构建离网预测模型,利用逻辑回归算法对服务号码捆绑特征进行降维,并加入模型,进行离网用户分析,从而辅助决策者制订相应的客户维挽策略,降低客户离网率。实验结果表明,服务号码软捆绑特征能够提高系统的分析预测能力。

关键词 客户流失,随机森林,服务号码捆绑,逻辑回归,离网预测系统
中图法分类号 TP391 **文献标识码** A

Application of Service Bundling in Churn Predict System

ZHANG Zheng-qing[1] ZHU Yi-jian[1] BAI Rui-rui[2] HUANG Yi-qing[2] YAN Jian-feng[2]
(China Unicom Co. Ltd. Shanghai Branch, Shanghai 200050, China)[1]
(College of Computer Science and Technology, Soochow University, Suzhou 215006, China)[2]

Abstract Customer churn problem is one of the biggest problems that telco operators are faced with, and it need to be solved. For different scenarios, operators have developed many churn prediction systems. Service number bundling means that customers have relations with 3rd-service support companies such as bank, E-commercials, supermarkets etc. during service available period through bundling phone number. After researching, we found that customers will have service number bundling behaviors with kinds of 3rd-service support companies during service available period. These companies will send SMSs to customers at any time. For customers who are about to churn, they will cancel the service bundling gradually. Thus, service bundling feature plays an important role in predicting potential churners. In order to help decision-making manager formulate corresponding customer retention campaigns and drop the churn probability, we developed the churn prediction model with random forest algorithm, and used logistic regression algorithm to reduce service number features. Experiment results show that service number soft bundling can improve the analysis and prediction performance of churn prediction system.

Keywords Customer churn, Random forest, Service bundling, Logistic regression, Churn prediction system

1 引言

近年来,随着国内外电信市场竞争的加剧,客户选择产品及服务的余地愈加广泛,客户的频繁流失是困扰运行商的一大难题,而且这种状况还在进一步恶化,运营商企业面临着巨大挑战。研究表明,客户离网率减少 5%,能给企业带来 30%~85%的利润增长;发展新客户的成本是挽留客户的5~7倍,而维挽客户的成功率却是发展新客户成功率的 16 倍[1]。因此,如何有效地预测未来潜在的离网用户,降低客户离网率,提升客户维挽工作效率,成为一个重要的研究课题。

客户离网预测是通过对客户的基本信息与历史行为等数据进行深入分析,提炼出已离网客户在离网前的一些规律,建立客户离网预测模型。通过模型预测近期离网的用户名单,企业可以根据客户的通话特征和习惯采取针对性的维系挽留活动,例如套餐计划或改善服务,来减少或避免高价值客户的流失。数据挖掘技术是离网预测的一项关键技术。数据挖掘主要依靠人工智能、模式识别、机器学习、统计学等理论知识,

本文受江苏省科技支撑计划重点项目(BE2014005-4)资助。
张正卿(1978-),男,硕士,工程师,主要研究方向为移动互联网技术研究、业务开发、大数据技术发展研究、平台构建以及大数据的应用性研究;朱奕健(1975-),男,硕士,工程师,主要研究方向为移动通信增值产品、移动互联网技术研究、业务开发、物联网技术、大数据技术发展研究、平台构建以及大数据的应用性研究;白瑞瑞(1989-),女,博士生,主要研究方向为机器学习与数据挖掘;黄一清(1989-),男,博士生,主要研究方向为机器学习与数据挖掘;严建峰(1978-),男,副教授,硕士生导师,主要研究方向为机器学习,E-mail:yanjf@suda.edu.cn(通信作者)。

高度智能地分析海量数据，做出归纳性推理，从中挖掘出潜在的模式，从而为决策者调整市场策略、做出正确决策提供科学的依据。

国内外学者已经围绕着客户离网行为分析、离网预测算法实现和改进、客户挽留策略实施和改进等许多方面进行了研究。文献[2]将7种分类器应用在用户离网预测模型中，这7种分类器分别是逻辑回归(Logistic Regression)、线性分类器(Linear Classifications)、朴素贝叶斯(Naive Bayes)、决策树(Decision Trees)、多层感知器神经网络(Multilayer Perceptron Neural Networks)、支持向量机(Support Vector Machines)和演变的数据挖掘算法(the Evolutionary Data Mining Algorithm)。文献[3]将社交网络信息应用到用户离网预测系统中，通过实验表明其可以提高维系挽留的利润。文献[4]等基于预算限制和客户挽留价值最大化，构建了客户的挽留模型。文献[5]等研究了用户间多相似度协同过滤推荐算法，通过用户间对不同项目类型的多个评分相似度来计算用户对未评分项目的预测评分。

从运营商角度，客户状态可以分为在网和离网，在网即依然使用运营商为其提供的服务，离网即不再使用。因此，判断客户状态属于典型的二分类变量问题。预测客户离网主要分为两个步骤，首先是从原始数据中抽取特征，形成特征表作为训练集和测试集，然后选择分类算法，用带有标签(是否离网)的训练集进行训练，得到分类器，使用测试集来评价分类器的性能。常用的分类算法包括：逻辑回归[6,7]、决策树[7,8]、boosting算法[9,10]、随机森林[11,12]、神经网络[13,14]及支持向量机[1,15,16]等。

本文以上海联通的客户离网问题为研究对象，通过观察用户的行为数据，发现98%的客户都与银行、电商、便利店等第三方服务提供商通过绑定手机号码产生联系，会不定时收到服务商提供的推送短信。因此，我们推测捆绑有服务号码的客户比没有捆绑服务号码的客户的离网率低。基于文献[11]的工作，利用随机森林算法构建离网预测系统，将服务号码分为9大类，利用逻辑回归算法对服务号码特征降维，加入特征库进行研究。本文第2节对逻辑回归模型和随机森林算法进行介绍；第3节进行实验仿真，并进行了结果分析；最后对全文进行总结。

2 逻辑回归模型和随机森林算法

2.1 逻辑回归模型

逻辑回归(Logistic Regression)是一种常用的机器学习方法，用于估计某种事物的可能性。逻辑回归延伸了多元线性回归的思想，数据集为$\{x_1,x_2,\cdots,x_n;y\}$，其中，自变量x_k为特征值；y为类别，即因变量，当y取两个值时，则为二分类，取多个值时，则为多分类。设逻辑回归模型的最终结果为0,1分类结果，其中，1表示属于该类，0表示不属于该类别。逻辑回归方程可以表示如下：

$$f(x,\beta)=\frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n}} \tag{1}$$

其中，$\beta_0,\beta_1,\beta_2,\cdots,\beta_n$表示特征属性参数值。构建逻辑回归模型的主要目的是求解逻辑回归方程中的参数$\beta_0,\beta_1,\beta_2,\cdots,\beta_k$，一般是基于最大似然估计来计算的。为描述方便，假设一个样本属于正类的概率值为$f(x,\beta)$，那么，该样本属于负类的概率值为$1-f(x,\beta)$。所以在样本和概率已知的条件下，该样本属于正类的概率可以表示为：

$$P(y|x,\beta)=\begin{cases}f(x,\beta), & \text{if } y=1\\ 1-f(x,\beta), & \text{if } y=0\end{cases} \tag{2}$$

从式(2)可以推断出在参数为β且逻辑回归模型已知的条件下，训练集X发生的可能性以及对数可能性可以表示为：

$$L(X,x,\beta)=\prod_{i=1}^{N}f(x_i,\beta)^{y_i}(1-f(x_i,\beta))^{1-y_i} \tag{3}$$

$$\ln L(X,y,\beta)=\sum_{i=1}^{N}(y_i\ln(f(x_i,\beta))+(1-y_i)\ln(1-f(x_i,\beta))) \tag{4}$$

其中，x_i表示训练集X的第i个样本，y_i表示第i个样本的分类结果，N表示样本的数目。在训练样本以及样本的分类结果已知的条件下，可以通过计算最大对数$\ln L(X,y,\beta)$求解逻辑回归模型参数。

2.2 随机森林

随机森林是由美国科学家Leo Breiman将其在1996年提出的Bagging集成学习理论与Ho在1998年提出的随机子空间方法相结合，于2001年发表的一种机器学习算法[17]。随机森林是一个集成学习模型，构成它的基础分类器称作决策树。

决策树算法是一种以实例为基础的归纳学习算法，着眼于从一组无次序、无规则的事例中归纳并推断出以决策树表示的分类规则。决策树可视为一个树状预测模型，它是由节点和有向边组成的层次结构，图1显示的是二叉树形式的决策树的结构。树中包含3种节点：根节点、内部节点、终节点(又称叶子节点)。决策树只有一个根节点，即全体训练数据集合。树中的每一个内部节点都是一个分裂问题，利用信息论中的信息增益寻找到达该节点的样本集中具有最大信息量的属性字段，根据该属性字段的不同取值建立树的分枝，然后在每个内部节点重复递归并建立树的下一个内部节点和分枝。每个终节点(即叶子节点)是带有分类标签的数据集合。从决策树的根节点到叶节点的每一条路径都形成一个分类。

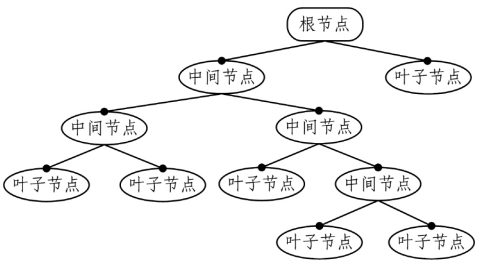

图1 二叉树形式的决策树结构图

单个决策树往往有分类精度不高、容易出现过拟合等问题，从而导致分类器的泛化能力较弱。一种很容易理解的方法就是采用集成方法将单个决策树聚集起来，形成一片森林，当输入待分类的样本时，最终的分类结果由单个决策树的分类结果投票决定。

设随机森林是由K个决策树$\{h(X,\theta_k),k=1,2,\cdots,K\}$组成，其中$\{\theta_k,k=1,2,\cdots,K\}$是一个随机变量序列，它由随机

森林的两大随机化思想决定的[18]。

(1)Bagging 思想:从原样本集 X 中有放回地随机抽取 K 个与原样本集同样大小的训练样本集。设原样本集的样本容量大小为 N,每次有放回抽取的训练样本大小也为 N,那么每个样本未被抽中的概率约为 $(1-\frac{1}{N})^N$,当 N 很大时,这个概率值趋于 $1/e\approx0.368$,表明每次约有 37% 的样本未被抽中,每个训练集 T_k 构造一个对应的决策树。

(2)特征子空间思想:在对决策树每个节点进行分裂时,从全部属性中等概率随机抽取一个属性子集,通常取 $\lfloor \log_2 M \rfloor+1$ 个属性,其中 M 为特征总数,从这个子集中选择一个最优属性来分裂节点。

由于构建每棵决策树时,随机抽取训练样本集和属性子集的过程都是独立的,且总体都是一样的,因此 $\{\theta_k, k=1, 2, \cdots, K\}$ 是一个独立同分布的随机变量序列。

训练随机森林的过程就是训练各个决策树的过程,由于各个决策树的训练是相互独立的,因此随机森林的训练可以通过并行处理来实现,这将大大提高生成模型的效率。随机森林中第 k 棵决策树 $h(X,\theta_k)$ 的训练过程如图 2 所示。

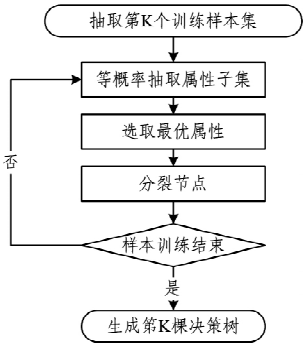

图 2 第 k 棵决策树的生成过程

将以同样的方式训练得到的 K 个决策树组合起来,就可以得到一个随机森林。当输入待分类的样本时,随机森林输出的分类结果由每个决策树的输出结果进行简单投票(即取众数)决定。

随机森林是一种有效的分类预测方法,它有很高的分类精度,对于噪声和异常值有较好的稳健性,且有较强的泛化能力。此外,随机森林是由数据驱动的一种非参数化分类方法,只需要通过给定样本,学习训练分类规则,并不需要分类的先验知识。

3 离网预测系统

跨行业数据挖掘标准流程(Cross-industry SandardProcess for Data Mining,CRISP-DM)模型为一个 KDD 工程提供了一个完整的过程描述,该过程分为 6 个阶段:商业理解(Business Understanding)、数据理解(Data Understanding)、数据准备(Data Preparation)、建立模型(Modeling)、评估(Evaluation)、部署(Deployment)。这个数据挖掘的程序模型为数据挖掘项目的生命周期提供了一个综合的描绘。它描述了一个数据挖掘项目所要经历的各个阶段、各阶段的任务以及这些任务之间的相互关系。本文从利用现有数据挖掘技术解决实际问题的角度出发,将构造离网预测系统分为 5 个步骤:明确问题、数据准备(包括数据抽取和特征提取)、制定标签、选择挖掘算法训练模型分类器、评价模型分类性能。图 3 示出离网预测系统的结构框架。

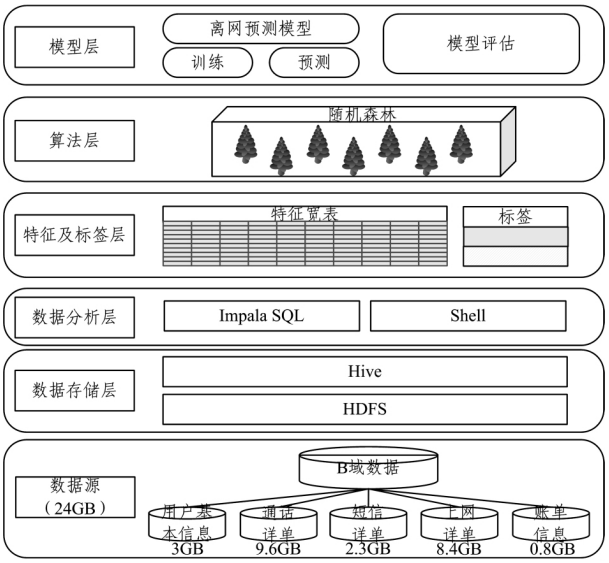

图 3 离网预测系统的结构框架

3.1 明确问题

电信企业实施数据挖掘,首先要明确当前数据挖掘所需要解决的实际问题,这是提出合理任务、实现预期目标和取得理想效果的基础。上海联通针对客户流失现象,提出离网预测的需求,目标是从当前在网的用户中挑选出即将离网的联通用户,对其进行维系挽留等一系列的营销措施,降低离网率。

3.2 数据准备

3.2.1 数据抽取

从数据源中所得到的历史数据存在着量大、属性繁多等特性。需要从海量的数据中选择与用户离网有关的适合分析的数据。由于离网用户在不同的时期有着不同的离网特征,不能用两年前所建立的模型对两年后的数据进行预测,这样会导致比较大的偏差。本文的离网预测系统是通过选取连续 3 个月的训练数据集训练模型,用生成的模型对这 3 个月之后 1 个月的数据进行预测分析。例如,如果需要对 2015 年 7 月份的用户做离网预测分析,则需抽取 2015 年 4 月份到 2015 年 6 月份之间的数据作为训练样本集,训练离网模型,然后预测 2015 年 7 月份的客户在 8 月份离网的可能性。

3.2.2 特征提取

数据库中的数据往往都对应着大量的属性,但是并不是每一个属性都是可用的,如果将不相关的属性或关联性很小的属性用于建模之中,将会使得计算代价呈几何级的倍数增加,并且会降低模型精确性。上海联通离运营的大数据平台每天都会产生近 2.3TB 大小的数据,包括 BSS(Business Support System)数据以及 OSS(Operation Support System)数据。BSS 又称为业务支持系统,通常 BSS 数据包括用户基本信息、用户行为、账单信息、语音数据、短信数据及通话详单等,每天会产生大约 24GB 的数据量。我们的离网预测模型针对 2/3G 预付费全量用户,基于全量 B 域数据,从 30 多张数据表中选出 11 张表作为模型的数据来源,这 11 张表分别是:用户基本信息日表、用户基本信息月表、用户行为日表、用户行为月表、套餐表、用户余额日表、用户语音通话详单表、用户短信收发详单表、账单月表、终端表和充值表。

模型中的特征主要包括两类:第一类是基本特征,即直接

从表中抽取的字段，例如通话时长、上网时长、充值金额、余额、终端类型等；第二类是二次处理特征，包括利用 PageRank 算法和 Label Propagation 算法从通话图和短信图中提取出来的特征，以及计算余额和账单之间的差值，通过身份证判断客户是否为上海本地人等。然后将所有的特征拼接起来形成特征向量，$X_m=[x_1,x_2,\cdots,x_N]$ 则代表用户 m 的特征，其中 x_i 是该用户的第 i 个特征。将用到的原始数据存储在 hadoop 分布式文件系统 HDFS 中，然后通过使用 Spark SQL 或者 Impala SQL 的关联操作或聚合操作将特征字段做成临时表，这样可以方便重复使用，最后将所有的特征临时表拼接成一张大宽表，表中的每一行代表一个用户的特征向量。目前我们的特征宽表中有 55 维特征，可以参看文献[11]。

3.3 制定标签

一个手机号码的生命周期包括 5 个阶段：预开户、有效期、充值期、锁定期和销户，如图 4 所示。预开户状态是指该号码可供使用，处于被激活状态。有效期是指该手机号码在正常使用中。充值期是指该手机号码处于欠费状态。锁定期是指该手机号码的欠费状态持续时间超过 2 个月（一般情况）。销户代表该手机号码生命周期结束。一般的，手机号码进入充值期后，继续充值即可恢复到有效期状态；或者处于锁定期的号码，继续充值也可恢复到有效期状态；当手机号码处于锁定期超过 2 个月（一般情况）时，该号码被销户。

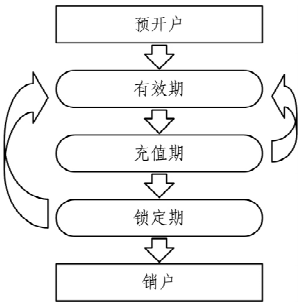

图 4 一个手机号码的生命周期

在我们的数据库中，有直接的字段标注手机号码当前的状态，这个属性的取值有 4 个：11（正常开机），12（有效期），23（锁定期），32（充值期）。表 1 是统计的过去 9 个月中用户处于充值期中进行充值的比例分布。从表中可以看到，有 84% 的人在进入充值期 10 天之内都会充值，因此，打标签策略是用户进入充值期 10 天之内继续充值，则视为在网用户，其余都视为离网用户。这也说明打标签是延后的，例如，需要判断 201508 月的用户真实离网情况，需要根据 20150910 这一天的数据来判断。

表 1 用户进入充值期充值时间分布

天数	0～5	6～10	11～15	15 以上
充值人数比例	65%	19%	9%	7%

3.4 分类器

根据主题问题及数据样本集的特点选择合适的算法，并根据具体问题对这些算法进行适当的组合。在算法选定的同时，必须根据企业现存的环境与技术发展趋势，提出具体的算法实现技术，这种实现技术必须是高效开放的，而且是能够在现存企业环境中实现的。上海联通离网预测模型选择随机森林算法训练分类器。采用分布式运行方式，同时使用 8 台服务器训练，每台服务器设置 100 棵树，模型整个生成 800 棵树，待输入预测样本时，取每棵树预测结果的平均值判断该样本的离网概率。同时，限制当叶子节点的样本个数小于 100 时，则停止分裂，目的是防止过拟合。

3.5 模型评估

衡量预测模型优劣常用的指标有正确率、错误率、均方差、提升度、置信度和支持度等。上海联通离网预测模型使用查全率（Recall）、查准率（Precision）和 AUC 3 个指标来评价模型。式（2）和式（3）分别对应离网预测模型中查全率及查准率的定义。

$$Recall=\frac{U_{True}}{Total_{True}} \tag{2}$$

$$Precision=\frac{U_{True}}{U} \tag{3}$$

其中，根据分类器的预测结果，即预测离网率，对用户进行降序排列，取出前 U 位用户，U_{True} 表示在选择的 U 位用户里面实际离网的用户数，$Total_{True}$ 表示测试样本集中所有实际离网用户数。一般的，当 U 取定某个具体的值时，Recall 和 Precision 值越大表明模型的预测性能越好；当 U 的取值不断增大，相应的 Recall 和 Precision 值也会不断增大。

AUC 是指 ROC（Receiver Operating Characteristic）曲线与横轴（False Positive Rate，FPR）之间的面积，可以通过式（4）来计算。

$$AUC=\frac{\sum_{c\text{为实际离网用户}} rank_c-\frac{P\times(P+1)}{2}}{P\times N} \tag{4}$$

其中，P 表示实际的离网用户数，N 表示实际的在网用户数，$rank_c$ 表示用户 c 在给出的离网概率名单中的排名，其中概率值最高的赋值为 n，次高为 $n-1$，依次类推。AUC 值越高说明模型性能越好。

4 服务号码捆绑特征在离网预测模型中的应用实现

4.1 特征生成过程

服务号码特征生成过程包括 4 部分，分别是生成服务号码维表、计算特征向量、训练分类器和生成特征，如图 5 所示。

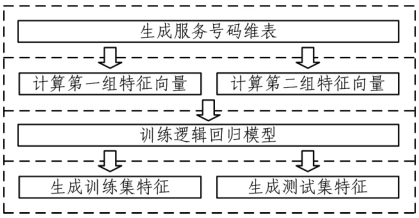

图 5 服务号码特征生成过程

4.1.1 生成服务号码维表

针对所有用户统计 201501 到 201508 期间，收到的非 13 位短信号码的总次数，将排名前 500 的短信号码抽出来，经过筛选得到一张服务号码维表，其中共有 288 个服务号码。本文认为用户接收到该维表中号码的短信，则具有捆绑行为。举个例子，若是某一个用户经常会收到铁道部火车票订购号码（服务号码是 12306）的短信通知，说明他将自己的手机号码和 12306 订票系统进行了捆绑。

4.1.2 计算特征向量

将维表中的 288 个服务号码细分为 9 大类，如图 6 所示。系统中其他的 55 组特征，时间窗口是一个月，例如通话时长，我们是按月取平均通话时长作为一个特征。服务号码特征的时间窗口设置为 4 个月。

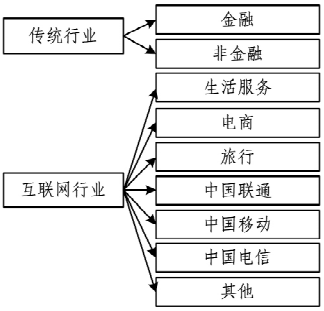

图 6　服务号码分类

首先，计算第一组特征向量。为了方便表述，用 I^k 表示第 k 类服务号码的集合，用 m 代表一个用户的手机号码。通过 ImpalaSQL 语言对数据库进行操作，统计在 4 个月中，手机号码 m 收到包含在集合 I^k 中的服务号码的短信总数 n_k，手机号码 m 第一次收到包含在集合 I^k 中的服务号码短信时间 B_k，手机号码 m 最后一次收到包含在集合 I^k 中的服务号码短信时间 E_k，手机号码 m 的入网时间为 IN_m，如图 7 所示。最后在每一类服务号码下，手机号码 m 得到向量 $F_{km}=[n_k,|B_k-IN_m|,|E_k-B_k|,|T-E_k|]$，其中 T 是第 5 个月中的一天(我们定为 15 号)，$|B_k-IN_m|,|E_k-B_k|,|T-E_k|$ 分别代表 B_k 和 IN_m，E_k 和 B_k，T 和 E_k 之间相差的天数。

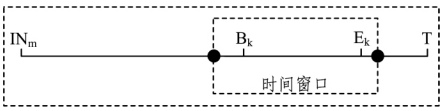

图 7　IN_m,B_k,E_k,T 的时间轴表示

第二组特征主要是计算熵特征。对于集合 I^k，设时间窗口中，每个月手机号码 m 收到包含在集合 I^k 中的服务号码的短信总数 $n_k^a(a=1,2,3,4)$，然后通过式(5)计算手机号码 m 的第 k 类的熵值 h_k，并通过式(6)来计算手机号码的熵值 h(针对所有类别的服务号码)。这样得到的第二组特征向量为 $F_m=[h_1,h_2,\cdots,h_k,h]$。

将第一组特征$[F_{1m},F_{2m},\cdots,F_{9m}]$和第二组特征 F_m 拼接起来得到手机号码 m 的一个 1×46 的行向量 $F^m_{1\times46}$。

$$h_k=-\frac{n_k^1}{\sum_{a=1}^{4}n_k^a}\log\frac{n_k^1}{\sum_{a=1}^{4}n_k^a}-\frac{n_k^2}{\sum_{a=1}^{4}n_k^a}\log\frac{n_k^2}{\sum_{a=1}^{4}n_k^a}-\frac{n_k^3}{\sum_{a=1}^{4}n_k^a}\log\frac{n_k^3}{\sum_{a=1}^{4}n_k^a}-\frac{n_k^4}{\sum_{a=1}^{4}n_k^a}\log\frac{n_k^4}{\sum_{a=1}^{4}n_k^a} \tag{5}$$

$$h=-\frac{\sum_{a=1}^{4}n_1^a}{\sum_{k=1}^{9}n_k}\log\frac{\sum_{a=1}^{4}n_1^a}{\sum_{k=1}^{9}n_k}-\frac{\sum_{a=1}^{4}n_1^a}{\sum_{k=1}^{9}n_k}\log\frac{\sum_{a=1}^{4}n_1^a}{\sum_{k=1}^{9}n_k}+\cdots-\frac{\sum_{a=1}^{4}n_9^a}{\sum_{k=1}^{9}n_k}\log\frac{\sum_{a=1}^{4}n_9^a}{\sum_{k=1}^{9}n_k} \tag{6}$$

4.1.3　训练逻辑回归模型

对于得到的 46 维的训练数据集，需要先打标签，标注是否离网，0 代表在网，1 代表离网。例如，时间窗口确定为 201503－201507，打标签是按照 20150810 这一天的数据进行标注。然后训练模型，使用的是 sklearn.linear_model.LogisticRegression 包生成模型，参数使用默认值。

4.1.4　生成训练集和测试集的特征

准备两组数据集，例如，训练集时间窗口为 201504－201508，测试集时间窗口为 201505－201509。然后，将两组数据集放入模型中，分别得到两组实验结果，我们称该实验结果分别为训练集和测试集的特征。

4.2　结果展示和分析

本文一共进行 3 组实验，每组实验的时间窗口如表 2 所列。在对照组实验中，首先确定离网预测模型的训练集，训练集包括 3 个月的数据，分别是用 20150410 给 201503 标注标签作为训练子集 1；用 20150510 给 201504 标注标签作为训练子集 2；用 20150610 给 201505 标注标签作为训练子集 3。3 组训练子集构成随机森林分类器的训练集，测试集是 201506 的数据。该训练集和测试集拥有 55 维特征。因为在不同时间窗口的训练集和测试集的数据分布都会有不同，所以有些月份的离网预测精度会高一些，有些月份的离网精度会低一些，但是都是在可接受范围内。然后，在实验组中加入软捆绑特征，验证软捆绑特征对于提高离网模型的有效性。用 201501－201504 之间的服务号码软捆绑特征向量来训练逻辑回归模型，将训练集(201502－201505)和测试集(201503－201506)放入模型，得到两个数据集的特征，将这一维特征也加到离网模型的训练集和测试集中，此时训练集和测试集拥有 56 维特征。实验结果如表 3－表 5 所列。

表 2　三组实验的时窗口

第一组实验		
模型	训练集时间窗口	测试集时间窗口
离网预测模型	训练集 1：201503 训练集 2：201504 训练集 3：201505	201506
逻辑回归模型	201502－201505	201503－201506
第二组实验		
模型	训练集时间窗口	测试集时间窗口
离网预测模型	训练集 1：201507 训练集 2：201508 训练集 3：201509	201510
逻辑回归模型	201506－201509	201507－201510
第三组实验		
模型	训练集时间窗口	测试集时间窗口
离网预测模型	训练集 1：201510 训练集 2：201511 训练集 3：201512	201601
逻辑回归模型	201509－201512	201510－201601

表 3　第一组实验结果

实验组(56 维特征)			对照组(55 维特征)		
Total	2110914		Total	2110914	
Churn	225878		Churn	225878	
AUC	0.8206		AUC	0.7951	
Top	recall	precision	Top	recall	precision
50000	0.1493	0.6246	50000	0.1108	0.6032
100000	0.2460	0.5557	100000	0.2195	0.5331
150000	0.3187	0.4400	150000	0.5439	0.4185
200000	0.3799	0.3990	200000	0.6173	0.3622

表 4　第二组实验结果

实验组(56 维特征)			对照组(55 维特征)		
Total	1848213		Total	1848213	
Churn	111915		Churn	111915	
AUC	0.9218		AUC	0.9023	
Top	recall	precision	Top	recall	precision
50000	0.3757	0.8309	50000	0.3469	0.8096
100000	0.5561	0.6223	100000	0.5231	0.5934
150000	0.6525	0.4669	150000	0.6329	0.4326
200000	0.7178	0.4017	200000	0.6962	0.3847

表 5　第三组实验结果

实验组(56 维特征)			对照组(55 维特征)		
Total	1713356		Total	1713356	
Churn	96088		Churn	96088	
AUC	0.9194		AUC	0.8913	
Top	recall	precision	Top	recall	precision
50000	0.3918	0.6929	50000	0.3514	0.6753
100000	0.5729	0.4505	100000	0.4394	0.4222
150000	0.6654	0.3763	150000	0.5439	0.3484
200000	0.7284	0.32	200000	0.6173	0.2966

实验结果表明，加入服务号码特征之后，系统的预测精度有明显的提高。一般的，若是一个用户多次捆绑手机号码，那这些捆绑行为就会从一定程度降低其离网概率，通过本文实验也很好地证明了这一点。

结束语　客户流失问题一直是学术界和工业界广泛关注的一个重要问题。良好的离网预测模型，可以辅助运营商对潜在离网用户进行提前干预，通过提供有效的营销策略降低客户离网率，最大程度提高企业利润增长。本文以服务号码软捆绑行为为研究对象，利用逻辑回归模型从中提取特征，加入到离网预测模型的特征库，通过实验证明，服务号码软捆绑特征的确可以提高系统预测精度。

参 考 文 献

[1] 蒋国瑞，司学峰. 基于代价敏感 SVM 的电信客户流失预测研究[J]. 计算机应用研究，2009，26(2)：521-523

[2] Huang B，Kechadi M T，Buckley B. Customer churn prediction in telecommunications[J]. Expert Systems with Applications，2012，39(1)：1414-1425

[3] Verbeke W，Martens D，Baesens B. Social network analysis for customer churn prediction[J]. Applied Soft Computing，2014，14(1)：431-446

[4] 罗彬，邵培基，罗尽尧，等. 基于预算限制和客户挽留价值最大化的电信客户流失挽留研究[J]. 管理学报，2012，9(2)：280-288

[5] 范波，程久军. 用户间多相似度协同过滤推荐算法[J]. 计算机科学，2012，39(1)：23-26

[6] Stripling E，Antonio K，Baesens B，et al. Profit maximizing logistic regression modeling for customer churn prediction[C]// IEEE International Conference on Data Science and Advanced Analytics (DSAA)，2015. IEEE，2015：1-10

[7] Nie G，Rowe W，Zhang L，et al. Credit card churn forecasting by logistic regression and decision tree[J]. Expert Systems with Applications，2011，38(12)：15273-15285

[8] Bin L，Peiji S，Juan L. Customer churn prediction based on the decision tree in personal handyphone system service[C]// 2007 International Conference on Service Systems and Service Management. IEEE，2007：1-5

[9] Lu N，Lin H，Lu J，et al. A customer churn prediction model in telecom industry using boosting[J]. IEEE Transactions on Industrial Informatics，2014，10(2)：1659-1665

[10] Lemmens A，Croux C. Bagging and boosting classification trees to predict churn[J]. Journal of Marketing Research，2006，43(2)：276-286

[11] Huang Y，Zhu F，Yuan M，et al. Telco churn prediction with big data[C]// Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. ACM，2015：607-618

[12] Xie Y，Li X，Ngai E W T，et al. Customer churn prediction using improved balanced random forests[J]. Expert Systems with Applications，2009，36(3)：5445-5449

[13] Hung S Y，Yen D C，Wang H Y. Applying data mining to telecom churn management[J]. Expert Systems with Applications，2006，31(3)：515-524

[14] Tsai C F，Lu Y H. Customer churn prediction by hybrid neural networks[J]. Expert Systems with Applications，2009，36(10)：12547-12553

[15] Farquad M A H，Ravi V，Raju S B. Churn prediction using comprehensible support vector machine：An analytical CRM application[J]. Applied Soft Computing，2014，19(2)：31-40

[16] Xia G，Jin W，et al. Model of customer churn prediction on support vector machine[J]. Systems Engineering-Theory & Practice，2008，28(1)：71-77

[17] Breiman L. Random forests[J]. Machine Learning，2001，45(1)：5-32

[18] 董师师，黄哲学. 随机森林理论浅析[J]. 集成技术，2013，2(1)：1-7

(上接第 574 页)

好的交易平台。此外，软件的功能设计完善，可做到更高效、更便捷地处理高校大学生的闲置物品，减少浪费现象，提高物品利用率。

参 考 文 献

[1] 韩敬海，丁春强. Android 程序设计[M]. 北京：电子工业出版社，2012：10-11

[2] 韩义波，宋莉，宋俊杰. Alax 技术结合 XML 或 JSON 的使用比较[J]. 电脑知识与技术，2009，5(1)：101-103

[3] Porting your apps from Django 0.96 to 1.0[EB/OL]. [2016-05-25]. http://docs.djangoproject.com/en/dev/releases/1.0-porting-guide

[4] E2EColud 工作室. 深入浅出 Google Android[M]. 北京：人民邮电出版社，2009：8-12

[5] 周绪宏，梁阿磊，戚正伟. 基于嵌入式 Linux 的智能手机系统软件的设计与实现[J]. 计算机应用与软件，2008，25(3)：59-61

[6] 孟岩. Android 组件模型评析(上)[J]. 程序员，2008(1)：49-51

[7] 乜聚虎，周学海，余艳玮，等. Android 安全加固技术[J]. 计算机系统应用，2011，20(10)：74-77

[8] 雷刚跃. 基于 XML 的异构数据库间数据交互技术研究[J]. 科学技术与工程，2006，6(23)：36-50

[9] 尹文刚，杨斌. Android 应用程序中的内存泄露与规避方法[J]. 单片机与嵌入式系统应用，2012，16：4-6

[10] 丁锐. 基于多级缓存的内存管理方案[J]. 杭州电子科学大学学报，2011，31(5)：25-28

[11] 公磊，周聪. 基于 Android 的移动终端应用程序开发与研究[J]. 计算机与现代化，2008，17(11)：86-89

[12] 王健. 疯狂升级的 Android 系统[J]. 电脑爱好者，2012(23)：83-84