

# 基于 Group Lasso 的多源电信数据离网用户分析

孙良君<sup>1</sup>, 范剑锋<sup>2</sup>, 杨琬琪<sup>2</sup>, 史颖欢<sup>2</sup>, 高 阳<sup>2</sup>, 周新民<sup>3</sup>

(1. 中博信息技术研究院有限公司, 江苏 南京 210012)  
(2. 南京大学计算机软件新技术国家重点实验室, 江苏 南京 210046)  
(3. 江苏省公安厅物证鉴定中心, 江苏 南京 210024)

**[摘要]** 随着行业竞争愈演愈烈, 电信企业的客户流失情况越来越严重, 给电信企业造成了巨大损失. 通过电信企业的数据来做离网用户的预测, 从而进一步作出挽留客户的正确决策, 成为电信企业日益关注的问题. 面对电信后台汇总的多源数据, 经分析发现其呈现天然的组结构. 为了选择对于离网类别最具判别性的特征, 本文使用了一种基于 Group Lasso 的组特征选择方法, 在此基础上用交叉验证法选择适当的特征组, 最终将选择出的少量组特征用于预测离网和停机的宽带用户. 实验表明, 在江苏某地级市电信离网用户分析数据中取得了比其他特征选择方法的精度平均高至少 10% 的预测性能.

**[关键词]** 电信企业, 客户流失, 多源数据, 特征选择, Group Lasso

**[中图分类号]** TP181 **[文献标志码]** A **[文章编号]** 1672-1292(2014)04-0077-07

## Group Lasso-Based Feature Selection for Off-network Analysis in Multisource Teledata

Sun Liangjun<sup>1</sup>, Fan Jianfeng<sup>2</sup>, Yang Wanqi<sup>2</sup>, Shi Yinghuan<sup>2</sup>, Gao Yang<sup>2</sup>, Zhou Xinmin<sup>3</sup>

(1. Zhongbo Information Technology Research Institute Company, Nanjing 210012, China)  
(2. State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210046, China)  
(3. Forensic Center of Jiangsu Province Public Security Bureau, Nanjing 210024, China)

**Abstract:** With the intensified competition in the industry, customer churn analysis is becoming one of the most significant tasks for the telecom companies, which might lead great financial loss to them. Thus, using the data to predict potential off-network customers and then making business decisions to retain these customers, have drawn lots of attention nowadays. In this paper, we present a Group Lasso-based feature selection method to predict the latent off-network customers by analyzing the corresponding multisource teledata. Specifically, we utilize the cross-validation strategy to choose the optimal sets of feature groups. Extensive experiment results show that the proposed approach has the superior performance (the Precision value is 10% higher than the other methods) on a real telecom dataset derived by a certain city in a prefectural city of Jiangsu.

**Key words:** telecom companies, customer churn, multisource data, feature selection, Group Lasso

互联网宽带业务的普及在城市已达饱和, 电信企业为了争取用户、拓宽市场所推出的业务和套餐也是日新月异、层出不穷, 这导致了用户的流动性的增加. 对于企业来讲, 固定用户数量相应减少. 有研究表明<sup>[1]</sup>, 如果将用户的流失率降低 5%, 利润就能增加 25% ~ 85%. 某国际公司的调查数据表明, 开发一个新客户的费用是维持一个老客户成本的 4 ~ 5 倍<sup>[1]</sup>. 由此可见, 若能很好地预测离网用户, 电信企业可有效提高决策能力, 更好地挽留离网用户, 从而降低运营成本, 增加利润收益.

本文所分析的江苏某地级市的宽带用户, 用户总数约为 80 万, 每个月的离网用户有近两万之多, 而停机的用户更是有十几万, 这给运营商带来了很大的困扰. 在近年宽带业务市场饱和, 增量客户发展速度减

收稿日期: 2014-07-20.

基金项目: 国家自然科学基金(61035003、61175042、61021062、61305068)、江苏省科技厅项目(BK2011005、BK20130581)、新世纪人才项目(NCET-10-0476)、江苏省医疗专项(BL2013033)、江苏省高校研究生科研创新计划项目(CXZZ13\_0055).

通讯联系人: 高阳, 教授, 博士生导师, 研究方向: 强化学习、智能 Agent、智能应用. E-mail: gaoy@nju.edu.cn

缓,众企业正逐渐由“生产型”企业向“利润型”企业过渡的趋势中,挽留离网用户、维系存量客户更加成为企业决策的重中之重<sup>[2]</sup>.

目前,关于电信离网用户分析的研究有一些进展<sup>[1-4]</sup>. 王雷和田玲<sup>[1,2]</sup>等人分别使用客户细分和用户满意程度的调查信息提出了客户流失预警模型,建立客户流失预警系统. Richter 等人<sup>[3]</sup>利用用户之间的通话记录建立用户网络,找出潜在的离网群体以及群体之间的隐含关系,从而对手机用户的离网倾向进行分析. Idris 等人<sup>[4]</sup>针对离网用户分析中多源数据的异构性使用(minimum redundancy and maximum relevance, mRMR)特征选择的方法和基于 RotBoost 的 Ensemble 方法建立分类预警模型,该方法不但通过降维提取了有足够区分度的特征,也提高了学习效率. 这些方法大都是基于用户的前台数据(如年龄、性别、业务类型、消费状况等). 然而,由于访问权限和用户隐私等原因,前台数据提取困难.

本文使用的是电信企业的后台汇总数据,包括用户上下线记录、线路稳定性等,如表 1 所示. 它们来自于电信的各个后台维护部门,不同的数据具有完全不同的特征意义. 如何从这些多源高维的大数据中预测离网用户,是本文所要解决的问题. 高维数据对数据挖掘任务造成了问题,一方面它降低了数据的可理解性,增加了训练和预测的时间;另一方面由于数据的真实维度是未知的,对数据处理不当易导致预测性能下降<sup>[5]</sup>. 由于高维数据存在这些问题,特征选择的方法一直是人们关注的问题. 针对数据成组的特性,本文使用了一种基于 Group Lasso 的组特征选择方法. 实验证明,该方法在江苏某地级市电信离网用户分析数据中取得了比其他特征选择方法的准确率(Precision)平均高至少 10% 的预测性能.

表 1 多源数据项概念定义  
Table 1 Definition of multisource teledata

| 数据项        | 描 述  |
|------------|--|
| 端口速率信息     | 每天都有一条记录,包括如下字段:平均上传和下载带宽,最大上传和下载带宽,上传和下载速度以及平均上传和下载衰减. 上传和下载衰减是根据特定的后台算法计算出来的值范围在 0 ~ 10 的离散值,值越大代表衰减越严重. |
| 线路稳定信息     | 每天都有一条记录,只有一个字段:线路的稳定状态,包括离散值 0、1、2、3、9,0 ~ 3 是从极稳定到极不稳定的状态标识,而 9 表示状态无法收集.                                |
| 掉线信息       | 每天一条记录,包括如下字段:异常掉线次数、正常掉线次数和线路是否正常.  |
| 用户的硬件信息    | 包括用户的终端机型号.  |
| 在线信息       | 包括每天的在线次数.   |
| 申告信息       | 用户投诉的记录,包括投诉用户的号码、用户投诉的内容以及用户投诉的时间.  |
| 用户每天上下线的信息 | 每个用户上下线一次就会有一条记录,主要字段有:上下线的时间、期间上传和下载的包和字节的数量.   |

1 问题背景分析

本文的数据来自于电信后台各个部门的汇总数据,其中一项数据是用户基本信息,包括如下字段:宽带用户唯一标识、用户状态、用户套餐类型、开户下行带宽、开户上行带宽. 其中用户状态是后台通过相关算法给出的该宽带用户该月是否活跃的标识. 每个月数据库中都会生成这些字段,若用户记录不出现在表格中,则视为离网;若用户状态是 0,表示该用户停机;若是 1,表示该用户处于活跃状态. 因为停机用户存在更大的离网倾向,离网用户分析主要是针对离网和停机用户进行预测.

对于客户流失,分为主动流失和被动流失. 被动流失是指由于客户的拖欠、欺诈和滥用服务等行为停止对客户的服务,而主动流失包括客户自己停用服务. 而主动流失的原因又有蓄意流失:受到企业的产品和服务的影响和无意流失,由于居住地变迁和各种非业务原因停用业务. 由于被动流失数量极少,所以离网用户分析主要针对的是主动流失,至于是由于蓄意还是无意流失,本文并无区分,也即假设流失的用户和其后台数据有关.

2 特征抽取

由于所用数据包含多种特征类型,根据领域知识,本文对不同的特征采用不同的特征提取方法.

2.1 硬件相关特征提取

端口速率信息、线路稳定信息和掉线信息 3 个字段都是电信后台对用户硬件的监控信息,和硬件的状况相关. 对于宽带用户,硬件状态的变更很少,通过观察可以发现同一个用户这个几个字段的信息在一个

月的每天基本没有变化,所以对于这些字段,提取每个月的均值作为特征.

## 2.2 申告和硬件信息特征提取

由于产生申告行为的用户很少,与离网用户数量相比也很少,故对申告内容不加细分,仅仅统计用户在5个月中每个月申告的次数.对于每个用户,硬件信息变化极少,所以直接提取月份中最后一个月的数据作为特征.

## 2.3 基于多尺度直方图统计的上网时间趋势特征提取方法

用户上下线的数据量很大,包含的信息很多,图1所示为50个用户在5月份的每日上网时间.

由于该时间序列并非传统意义上典型的时间序列,简单的距离度量和信号处理的方法并不适用.本文借鉴了 Cong 等人的工作<sup>[6]</sup>,提出了针对用户上网时间变化特征的多尺度直方图统计方法:对于用户每个月的上网时间,将时间序列两两做差值,得出上网时间每天的变化信息如图2所示(以半个月举例).

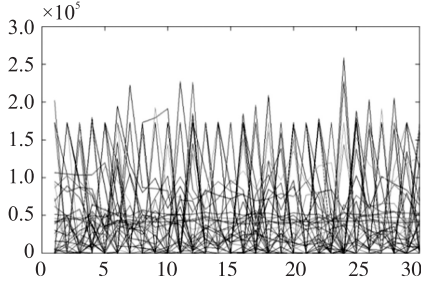


图1 用户上网时间序列

Fig.1 Time series of users' on line record

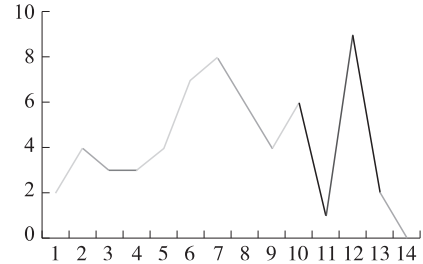


图2 上网时间差值的分类

Fig.2 Types of the difference between different periods

首先算出该差值序列的绝对值均值,然后用该值作为阈值,将时间差值细分为上网时间增加或减少“显著”和“不显著”以及上网时间不变的分组,并对这些组进行统计,得出直方图特征,如图3所示.可以发现,本文对上网时间的变化采用了不同的衡量“尺度”——“显著”或者“不显著”,且对上网时间差值的增加、减少或者不变这样的“方向”信息进行了多层次的提取.将若干月的直方图拼在一起,即可作为该用户的上网时间趋势特征.

不难发现,这些特征呈现组的结构,即对于同一组特征,它们在语义上更加相似且在数值上相关,在训练过程中,同一组的特征表征能力也相似.

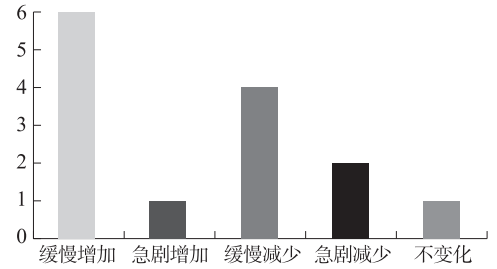


图3 上网时间直方图

Fig.3 Histograms of on line time series

## 3 基于 Group Lasso 的组特征选择

### 3.1 Lasso 方法

Lasso 方法是 Tibshirani<sup>[7]</sup>于1996年提出的稀疏特征选择的方法.该方法可以写成以下形式:

$$\min_{\beta} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1, \quad (1)$$

其中,  $X \in \mathbf{R}^{m \times n}$ ,  $Y \in \mathbf{R}^m$  分别表示数据矩阵和类标向量;  $\lambda \geq 0$  是正则参数;  $\beta \in \mathbf{R}^n$  是特征选择参数向量. 式(1)第一项为损失函数项,第二项为罚函数项.

Lasso 方法引起了广泛的研究兴趣. Fan 等人<sup>[8]</sup>在2001年提出了另一种惩罚项 (smoothly clipped absolute deviation, SCAD), 该方法效率比传统方法高,且通过恰当的选择正则化系数,即使在噪声很大的数据集上也可获得很好的预测性能. Tibshirani 等人<sup>[9]</sup>在2005年进一步提出了融合的 Lasso,在特征可序列化且特征数远大于样本数量的前提下,对参数向量做局部平滑的约束,获得了不错的分类和回归效果. Zou<sup>[10]</sup>在2006年针对 Lasso 在某些情况下特征选择结果不一致的问题,提出了 Adaptive Lasso 的方法,自适应地选择参数向量的权重,该方法在很多情况下预测性能和数据的真实模型相仿.

### 3.2 Group Lasso 方法

在本文的实际问题中,特征具有组结构,而 Lasso 仅适用于单一特征的选择,并不适用于组特征的选

择. 所以当 Lasso 直接应用于具有组结构的模型中时, 其倾向于选择出单个特征, 破坏了特征的组结构. 且 Meier 等人指出<sup>[11]</sup>, 对于 LR 分类器来说, 特征的微小变化也会对最终的预测结果造成很大的影响. Yuan 等人<sup>[12]</sup>提出的 Group Lasso 方法通过引入对罚函数的扩展很好地解决了这两个问题. 因此, Group Lasso 是对 Lasso 的一种推广, 对组特征的选择进行研究, 从而改进了 Lasso 在组结构数据上的缺陷.

Group Lasso 方法可以形式化为下式:

$$\hat{\beta}_{\lambda} = \underset{\beta}{\operatorname{argmin}} (\|Y - X\beta\|_2^2 + \lambda \sum_{g=1}^G \|\beta_{I_g}\|_2), \quad (2)$$

其中,  $\beta \in \mathbf{R}^n$  表示特征选择向量;  $X$  是  $m \times n$  的数据矩阵;  $Y \in \mathbf{R}^n$  是数据的标签数组;  $I_g$  是  $g$  组的特征下标,  $g = 1, \dots, G$  ( $G \in \mathbf{N}^+$  是组的个数);  $\lambda \geq 0$  是正则参数.

不难看出, 该方法的一个基本假设就是特征聚为  $G$  个分组. 在罚函数中, 不同于 Lasso 方法将每个特征的系数项的绝对值进行加总, 这里所加总的是每个组系数的 L2 范式, 在优化求解过程中, 导致某些组组内的稀疏都趋向于 0, 而另一些重要的相关组被选择出来了. Bach 等人<sup>[13]</sup>给出了 Group Lasso 和异构数据源多核学习一致性的充分必要条件, 从而证明了 Group Lasso 对多源异构数据的学习能力.

对于组特征内部的稀疏, Friedman 等人<sup>[14]</sup>提出了更一般的稀疏 Group Lasso 罚函数, 该罚函数能使组内每个特征都得到稀疏性. 由于在本文的问题中, 组内特征是同质特征, 实验证明表达能力类似, 所以对该方法不予探讨.

对于该优化问题, 本文采用 Beck 等人提出的 FISTA (fast iterative shrinkage-thresholding algorithm) 方法<sup>[15]</sup>, 该方法保持了 ISTA (iterative shrinkage-thresholding algorithm) 方法的计算简洁性, 且在此基础上增加了全局的收敛系数, 在理论上和实际应用中均获得了很好的效果.

本文使用 SLEP 工具箱<sup>[16]</sup>实现该算法, 具体来说, 该工具箱解决  $\ell_1/\ell_q$  范式的正则化最小平方差问题:

$$\min_x \frac{1}{2} \|Y - X\beta\|_2^2 + \lambda \sum_{g=1}^G w_g \|\beta_{I_g}\|_q, \quad (3)$$

其中,  $X \in \mathbf{R}^{m \times n}$  是数据集;  $Y \in \mathbf{R}^m$  是类标;  $\beta \in \mathbf{R}^n$  是特征选择向量, 它分成了  $G$  个不重叠的组  $\beta_{I_1}, \beta_{I_2}, \dots, \beta_{I_G}$ ;  $w_g$  表示第  $g$  个组的权值;  $q$  表示正则项的范数. 在实验中, 将  $q$  值设为 2. 可以看出, 在组内采取的是 2 范式的约束, 而在组间通过 L1 范式进行稀疏约束. 该方法的前提假设是特征的采纳与否是以组为单位的, 而采用的组的个数要尽可能的少.

### 3.3 离网用户分析方法框架

本文所建立的离网用户分析模型遵循的框架如图 4 所示.

首先训练集上针对不同的源数据提取相应特征, 通过交叉验证进行特征选择, 在选出的特征组上建立分类器. 对于测试集数据, 采用相应的特征抽取方法, 将特征输入分类器, 得到测试数据的分类结果.

## 4 实验结果与分析

### 4.1 训练集和测试集的划分

针对 2013 年 5 月份 ~ 2014 年 2 月份的数据, 训练集和测试集的划分如图 5 所示.

本文在 5 ~ 9 月份的数据上提取特征, 然后将该用户在 9 ~ 11 月份是否离网 (用户是否出现在 10 ~ 12 月份清单中) 作为标签. 对于测试集, 从 7 ~ 11 月份的数据中提取特征, 将该用户在 11 ~ 1 月份是否离网作为标签. 对于停机用户, 特征提取和标签获取方法相同.

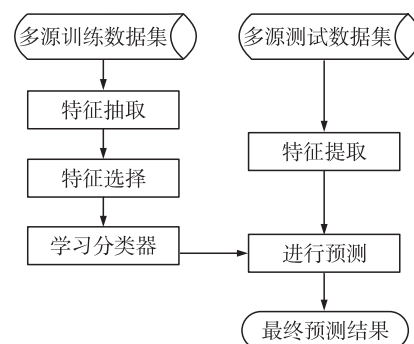


图 4 离网用户分析框架

Fig. 4 Framework of customer churn analysis

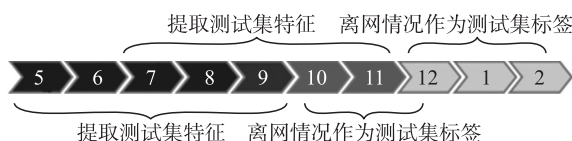


图 5 训练集和测试集的划分

Fig. 5 Splits of the training and testing sets



## 4.2 参数设置

对于  $\lambda$  值的调整,分别取  $(5 \times 10^{-5}, 5 \times 10^{-4}, 5 \times 10^{-3}, 0.05, 0.1, 0.5$  和  $0.9)$  7 组值,在测试集上分别对离网和停机用户标签进行学习。

## 4.3 评价指标

由于挽留客户是需要代价的,故对离网和停机用户的预测,更注重 Precision (预测为离网用户且命中的用户数在离网用户总数中的比重),保证被采取挽留措施的用户确实存在流失倾向。所以对于交叉验证和测试集上的训练结果,主要以准确率 (Precision) 作为组特征选择的判据。对于召回率 (Recall) 是预测为离网用户且命中的用户在所有离网用户中的比重。而 F-measure 是通过  $\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$  计算出来的两个指标的综合值,表示预测的综合性能。

## 4.4 组内特征观察

对停机用户  $\lambda$  取值为  $5 \times 10^{-5}$  的各组的  $x$  进行观察,如表 2 所示,可以发现组内参数值的差异很小,说明同组特征的表征能力是相似的。表中加粗了参数中的部分绝对值相似的值(每组参数可能存在超过一组的相似参数值,以第 0 组为例,其中部分参数集中在 0.002 0 左右,而另一部分集中在 0.000 5 左右)。

表 2 各组  $x$  值  
Table 2 The values in  $x$

| 组编号 | 组内 $x$ 值         |                  |                  |                  |                  |                  |                  |                  |                  |                  |                      |
|-----|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|----------------------|
| 0   | 0.000 4,         | 0.001 2,         | 0.000 1,         | <b>0.002 7,</b>  | <b>-0.028 2,</b> | <b>0.025 3,</b>  | 0.000 1,         | <b>-0.002 2,</b> | <b>-0.023 5,</b> | 0.001 2,         | ..., <b>0.001 9</b>  |
| 1   | 0.016 6,         | <b>-0.007 7,</b> | <b>-0.006 9,</b> | <b>0.010 7,</b>  | 0.000 8,         | <b>-0.011 1,</b> | -0.002 5,        | 0.034 4,         | <b>-0.005 3,</b> | 0.000 7,         | ..., <b>0.007 0</b>  |
| 2   | <b>0.001 3,</b>  | <b>0.001 7,</b>  | -0.002 7,        | -0.004 1,        | 0.000 2,         | <b>0.000 8,</b>  | 0.002 8,         | <b>-0.001 4,</b> | <b>0.001 2,</b>  | -0.000 7,        | ..., <b>0.001 8</b>  |
| 3   | <b>-0.006 4,</b> | 0.000 0,         | -0.000 1,        | <b>-0.006 6,</b> | 0.000 0,         | -0.000 4,        | <b>-0.005 0,</b> | -0.000 1,        | -0.000 1,        | -0.000 7,        | ..., <b>0.031 4</b>  |
| 4   | <b>-0.000 0,</b> | <b>-0.000 0,</b> | <b>-0.000 1,</b> | <b>-0.000 1,</b> | <b>-0.000 0,</b> | <b>-0.000 0,</b> | <b>-0.000 1,</b> | <b>-0.000 0,</b> | <b>-0.000 0,</b> | <b>-0.000 0,</b> | ..., <b>-0.000 1</b> |
| 5   | -0.000 8,        | <b>-0.000 2,</b> | <b>-0.000 3,</b> | <b>0.000 2,</b>  | <b>-0.000 1,</b> | <b>-0.000 1,</b> | 0.001 0,         | -0.000 5,        | <b>-0.000 1,</b> | 0.000 8,         | ..., <b>0.000 3</b>  |
| 6   | <b>-0.000 0,</b> | <b>0.000 2,</b>  | <b>0.000 2,</b>  | <b>0.000 4,</b>  | 0.002 0,         | -0.000 8,        | <b>-0.000 1,</b> | <b>-0.000 5,</b> | 0.001 8,         | -0.000 7,        | ..., 0.002 9         |
| 7   | 0.014 9,         | 0.006 2,         | -0.001 5,        | <b>-0.000 8,</b> | 0.007 7,         | <b>-0.000 5,</b> | -0.003 6,        | <b>-0.000 7,</b> | <b>0.000 6,</b>  | -0.000 0,        | ..., 0.000 2         |
| 8   | -0.000 7,        | <b>-0.004 2,</b> | <b>0.002 5,</b>  | <b>-0.002 4,</b> | -0.000 6         |                  |                  |                  |                  |                  |                      |
| 9   | <b>-0.004 0,</b> | 0.001 6,         | <b>-0.003 4,</b> | 0.000 6,         | <b>-0.004 4,</b> | <b>-0.005 5,</b> | <b>0.004 6,</b>  | <b>0.005 0,</b>  | 0.017 7,         | 0.017 7,         | ..., -0.022 8        |

## 4.5 交叉验证结果分析

组的个数、各自的维度以及特征意义如表 3 所示。

表 3 特征组的描述  
Table 3 Description of the feature groups

| 特征组     | 特征组描述               | 维度 | 编号 |
|---------|---------------------|----|----|
| 用户基本信息  | 包括用户的业务类型           | 47 | 0  |
| 硬件信息    | 用户终端类型、接入类型、线路类型等   | 17 | 1  |
| 线路稳定状态  | 线路的稳定状况             | 25 | 2  |
| 掉线情况    | 线路连接的掉线的情况          | 15 | 3  |
| 在线时间    | 用户每个月在线的时间          | 15 | 4  |
| 在线时间差值  | 用户每个月在线时间的差值        | 30 | 5  |
| 在线次数    | 用户每个月的日均在线次数以及差值    | 15 | 6  |
| 上下行速率   | 最近一个月的上下行速率信息       | 8  | 7  |
| 申告信息    | 用户的 5 个月中的申告次数      | 5  | 8  |
| 时间序列直方图 | 从用户每日上网时间差值提取的直方图特征 | 25 | 9  |

对于停机用户,在不同的参数下得到的实验结果如表 4、表 5 所示。由于采用的 Group Lasso 方法对组内特征没有稀疏约束,所以每个组内的  $x$  多数不为 0。可以发现,其中  $\lambda$  值越大,稀疏约束的权值越大,得到的  $x$  越稀疏。

表 4 停机用户参数调节结果

| Table 4 Parameter selection of service suspended user |                |
|---|----------------|
| $\lambda$ 值   | 对应的特征组编号       |
| $5 \times 10^{-5}$                                    | 0 ~ 9          |
| $5 \times 10^{-4}$                                    | 0 ~ 3, 6, 7, 9 |
| $5 \times 10^{-3}$                                    | 0 ~ 3, 6, 9    |
| $5 \times 10^{-2}$                                    | 0, 9           |
| 0.1   | 0, 9           |
| 0.5   | 0, 9           |
| 0.9   | 0, 9           |

表 5 离网用户参数调解结果

| Table 5 Parameter selection of off-network user |                |
|---|----------------|
| $\lambda$ 值                                     | 对应的特征组编号       |
| $5 \times 10^{-5}$                              | 0 ~ 9          |
| $5 \times 10^{-4}$                              | 0 ~ 3, 5 ~ 9   |
| $5 \times 10^{-3}$                              | 0 ~ 3, 6, 7, 9 |
| $5 \times 10^{-2}$                              | 0 ~ 1, 9       |
| 0.1   | 0, 9           |
| 0.5   | 9              |
| 0.9   | 9              |

在训练集上,根据针对不同的  $\lambda$  值选出的用户组,用 LR 的学习方法(该结果和 C45 决策树结果类似),采用了十折交叉验证,得到了停机和离网用户的结果如表 6、表 7 所示(表 6、表 7 为不同  $\lambda$  值所对应的交叉验证的结果,对最好的 Precision 值加粗表示),从而选出相应的组特征。

表 6 停机用户交叉验证结果

| Table 6 Cross validation of service suspended user |              |        |           |
|--|--------------|--------|-----------|
| $\lambda$ 值  | Precision    | Recall | F-measure |
| $5 \times 10^{-5}$                                 | 0.891        | 0.706  | 0.788     |
| $5 \times 10^{-4}$                                 | 0.888        | 0.707  | 0.787     |
| $5 \times 10^{-3}$                                 | 0.892        | 0.705  | 0.787     |
| $5 \times 10^{-2}$                                 | <b>0.899</b> | 0.699  | 0.786     |
| 0.1  | <b>0.899</b> | 0.699  | 0.786     |
| 0.5  | <b>0.899</b> | 0.699  | 0.786     |
| 0.9  | <b>0.899</b> | 0.699  | 0.786     |

表 7 离网用户交叉验证结果

| Table 7 Cross validation of off-network user |             |                       |           |
|--|-------------|-----------------------|-----------|
| $\lambda$ 值                                  | Precision   | Recall                | F-measure |
| $5 \times 10^{-5}$                           | 0.217       | $4.07 \times 10^{-4}$ | 0.001     |
| $5 \times 10^{-4}$                           | 0.307       | 0.002                 | 0.004     |
| $5 \times 10^{-3}$                           | 0.316       | 0.003                 | 0.006     |
| $5 \times 10^{-2}$                           | <b>1.00</b> | $8.15 \times 10^{-5}$ | 0.500     |
| 0.1  | <b>1.00</b> | $8.15 \times 10^{-5}$ | 0.500     |
| 0.5  | 0           | 0                     | 0         |
| 0.9  | 0           | 0                     | 0         |

4.6 测试集实验结果

本文采用了离网用户分析领域经常使用的决策树模型,该模型可以配合 SQL 语句直接在数据库上得到结果.决策树模型具有很好的非线性拟合能力,且可通过参数调节很好地应对过拟合的问题.本文采用 C45 决策树来建立最终的预测模型.将选出的特征在测试集上验证,并使用逻辑回归和决策树模型做对比.

表 8 中 P 代表 Precision 指标,R 代表 Recall 指标,这都是对于离网或停机用户预测而言的.对于特征选择,还试验了基于皮尔森相关系数的 Filter 方法<sup>[17]</sup>,采用 0.1 为阈值,在离网和停机问题上分别筛选了 69 和 74 维特征.对于离网用户分析,由于样本具有不平衡性;离网用户只占 1/40,所以将非离网用户随机分成  $n$  组,使其数量和离网用户数量相当,然后和离网用户合并求得相关系数,最终对相关系数进行加权平均.在 C45 算法离网用户预测中,Group Lasso 方法预测的 Precision 值比其他方法高 40 个百分点;在停机用户预测中,平均高出 10 个百分点.LR 方法在 Group Lasso 上的预测性能也普遍比其他特征要好.

表 8 结果对比 1

| Table 8 Comparison one of the feature selection methods |       |                       |       |       |       |       |                                     |                       |             |                       |
|---|-------|-----------------------|-------|-------|-------|-------|-------------------------------------|-----------------------|-------------|-----------------------|
|   | 0     |                       | 9     |       | 所有特征  |       | Yu 的方法 <sup>[17]</sup><br>69/74 维特征 |                       | Group Lasso |                       |
|   | P     | R                     | P     | R     | P     | R     | P                                   | R                     | P           | R                     |
| 离网(C45)   | 0     | 0                     | 0     | 0     | 0.265 | 0.064 | 0.256                               | 0.015                 | 0.451       | 0.015                 |
| 离网(LR)  | 1.00  | $4.31 \times 10^{-4}$ | 0     | 0     | 0.286 | 0.005 | 0.065                               | $6.63 \times 10^{-5}$ | 1.00        | $4.31 \times 10^{-4}$ |
| 停机(C45)   | 0.879 | 0.696                 | 0.745 | 0.590 | 0.753 | 0.720 | 0.822                               | 0.707                 | 0.877       | 0.694                 |
| 停机(LR)  | 0.876 | 0.752                 | 0.708 | 0.613 | 0.869 | 0.704 | 0.885                               | 0.696                 | 0.880       | 0.699                 |

将 Lasso 方法和 Group Lasso 方法的结果进行对比,对于 Lasso 方法的不同参数值,选取最好的参数结果.对比结果如表 9 所示,Group Lasso 特征上的结果比 Lasso 单独提取的特征的预测性能平均高出 10 个百分点.

对于预测结果,电信企业对所预测的离网名单采取挽留措施.由于挽留措施需要一定代价,所以该问

表 9 结果对比 2

| Table 9 Comparison two of the feature selection methods |       |                       |              |                       |
|---|-------|-----------------------|--------------|-----------------------|
|   | Lasso |                       | Group Lasso  |                       |
|   | P     | R                     | P            | R                     |
| 离网(C45)   | 0.363 | 0.017                 | <b>0.451</b> | 0.015                 |
| 离网(LR)  | 0.059 | $3.00 \times 10^{-4}$ | <b>1.00</b>  | $4.31 \times 10^{-4}$ |
| 停机(C45)   | 0.815 | 0.732                 | <b>0.877</b> | 0.694                 |
| 停机(LR)  | 0.881 | 0.701                 | 0.880        | 0.699                 |

题更加注重 Precision 指标.

## 5 结语

本文对多源电信数据提取特征,使用 Group Lasso 方法进行组特征提取,在此基础上建立离网用户预测模型.实验结果表明,本文的组特征比各个组的单独作为特征、使用所有特征、Yu 方法<sup>[17]</sup>筛选出来的特征以及直接使用 Lasso 方法得出的特征都要好.

本文未来的工作可以进一步探索 Group Lasso 组特征内部的稀疏在该问题上的特性.在 Lasso 模型和决策树的模型上,采取代价敏感的策略,使系统的预测性能向 Precision 进一步倾斜.

## [参考文献](References)

- [1] 王雷,陈松林,顾学道.客户流失预警模型及其在电信企业的应用[J].电信科学,2006,22(9):47-51.  
Wang Lei, Chen Songlin, Gu Xuedao. Analysis and application for national telecoms of customer churn alarm models[J]. Telecommunication Science, 2006, 22(9): 47-51. (in Chinese)
- [2] 田玲,邱会中,郑莉华.基于神经网络的电信客户流失预测主题建模及实现[J].计算机应用,2007,27(9):2 294-2 297.  
Tian Ling, Qiu Huizhong, Zheng Lihua. Telecom churn prediction modeling and application based on neural network[J]. Journal of Computer Application, 2007, 27(9): 2 294-2 297. (in Chinese)
- [3] Richter Y, Yom-Tov E, Slonim N. Predicting customer churn in mobile networks through analysis of social groups[C]//Proceedings of SIAM International Conference on Data Mining. Columbus, 2010: 732-741.
- [4] Idris Adnan, Asifullah Khan, YeonSoo Lee. Intelligent churn prediction in telecom: employing mRMR feature selection and RotBoost based ensemble classification[J]. Applied Intelligence, 2013, 39(3): 659-672.
- [5] Guyon, Isabelle, André Elisseeff. An introduction to variable and feature selection[J]. The Journal of Machine Learning Research, 2003(3): 1 157-1 182.
- [6] Cong Y, Yuan J S, Liu J. Sparse reconstruction cost for abnormal event detection[C]//Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition. IEEE Conference Society, 2011: 3 449-3 456.
- [7] Tibshirani R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1996, 58(1): 267-288.
- [8] Fan Jianqing, Li Runze. Variable selection via nonconcave penalized likelihood and its oracle properties[J]. Journal of the American Statistical Association, 2001, 96(456): 1 348-1 360.
- [9] Tibshirani R, Saunders M. Sparsity and smoothness via the fused lasso[J]. Journal of the Royal Statistical Society: Statistical Methodology, 2005, 67(1): 91-108.
- [10] Zou Hui. The adaptive lasso and its oracle properties[J]. Journal of the American Statistical Association, 2006, 101(476): 1 418-1 429.
- [11] Meier L, Van De Geer S, Bühlmann P. The group lasso for logistic regression[J]. Journal of the Royal Statistical Society: Statistical Methodology, 2008, 70(1): 53-71.
- [12] Yuan Ming, Lin Yi. Model selection and estimation in regression with grouped variables[J]. Journal of the Royal Statistical Society: Statistical Methodology, 2006, 68(1): 49-67.
- [13] Francis R Bach. Consistency of the group lasso and multiple kernel learning[J]. The Journal of Machine Learning Research, 2008(9): 1 179-1 225.
- [14] Friedman J, Hastie T, Tibshirani R. A note on the group lasso and a sparse group lasso[J/OL]. [2014-07-20] <http://arxiv.org/abs/1001.0736>.
- [15] Beck A, Teboulle M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems[J]. SIAM Journal on Imaging Sciences, 2009, 2(1): 183-202.
- [16] Liu Jun, Ji Shuiwang, Ye Jieping. SLEP: Sparse Learning with Efficient Projections[M]. TEMPE: Arizona State University, 2009.
- [17] Yu L, Liu H. Feature selection for high-dimensional data: a fast correlation-based filter solution[C]//Proceedings of the Twentieth International Conference on Machine Learning (ICML 2003). Washington DC, 2003.

[责任编辑:严海琳]