



## 一种基于决策树终端换机间隔的研究

崔羽飞, 魏进武, 潘思宇, 张溶芳, 王笑, 贾子翔  
(中国联合网络通信有限公司研究院, 北京 100176)

**摘要:** 随着移动互联网的快速发展, 运营商人口红利逐渐饱和, 用户移动终端的换机节点成为运营商关注的重点。主要根据用户历史的消费信息和移动终端的更换信息, 通过决策树算法来预测用户的换机频率, 为移动终端的精准营销、用户的存量经营提供数据支撑。

**关键词:** 手机终端; 决策树; 精准营销

**中图分类号:** TP301

**文献标识码:** A

**doi:** 10.11959/j.issn.1000-0801.2019098

## Research on terminal switch interval based on DTL

CUI Yufei, WEI Jinwu, PAN Siyu, ZHANG Rongfang, WANG Xiao, JIA Zixiang  
Research Institute of China United Network Communications Co., Ltd., Beijing 100176, China

**Abstract:** With the rapid development of mobile internet, the population dividend of operators is gradually saturated, and the switch node of user mobile terminal has become the focus of operators' attention. The switching frequency of users was predicted by decision tree algorithm based on the consumption information of user history and the replacement information of mobile terminals, which provided data support for the accurate marketing of mobile terminals and the stock management of users.

**Key words:** mobile terminal, decision tree, precise marketing

### 1 引言

随着互联网的发展, 移动数据流量呈现爆发式增长。传统以语音和短信为主要业务的三大运营商, 受到来自互联网公司的挑战。

工业和信息化部数据显示, 2015 年移动电话总数达 13.06 亿部, 4G 用户占比超过 1/4。2016 年 9 月, 移动电话用户总数达 13.16 亿户。4G 用户总数达到 6.86 亿户, 占移动电话用户的 52.1%。从数据上来看, 全国移动终端的数量增长速度缓

慢, 趋于饱和。而 4G 用户的占比呈增长的趋势。

随着互联网的发展及互联网应用的不断普及, 三大运营商在语音和短信上的业务受到冲击。工业和信息化部数据显示, 在 2015 年, 全国移动电话去话通话时长为  $2.85 \times 10^{12}$  min, 同比下降了 2.6%; 在 2016 年, 全国移动电话去话通话时长为  $2.81 \times 10^{12}$  min, 同比下降了 1.4%。从数据中可以看出, 受到移动互联网应用的打击, 三大运营商面临着严峻的挑战。

虽然移动运营商受到移动互联网应用的打



击,但 4G 用户的占比增加了。由于移动用户趋于饱和,增加用户所带来的效益不断减少。抓住用户的换机时间窗口进行精准营销成为运营商存量经营的一个重要方面。本文通过移动大数据分析用户数据和行为,利用决策树对数据进行建模,分析移动用户终端换机间隔,为移动终端的精准营销提供数据依据。

## 2 运营商基于存量经营的终端营销业务场景

近年来,随着移动互联网的高速发展,运营商在语音、短信和彩信等方面的业务受到了极大的挑战。从手机终端的发展情况来看,在 2016 年,手机终端数量的发展速度趋于饱和。为了应对移动互联网对运营商的挑战,运营商只能从其他地方弥补在语音、短信和彩信等方面业务上受到的损失。存量经营成为目前运营商势在必行的发展方向。

存量经营即通过利用运营商在过去积累的庞大用户体系,为用户提供更加优质的服务,以提高运营商的价值。随着大数据和数据挖掘的崛起,数据价值已经成为衡量价值的因素之一。运营商拥有丰富的用户信息,运营商可以通过对这些信息进行处理和分析,得到用户的诸多标签属性、生活习惯、消费行为以及通过数据分析出来的潜在用户属性等。利用用户信息,为用户提供个性化的服务,是将来发展的一个方向。例如,如果了解一个用户在特定的日子会有为他人送花的习惯,可以根据时间信息、用户的消费水平、用户对花的喜爱程度和节日的性质等为用户提供花店的位置、花的样式和送花服务等。

近年来,随着科技的不断发展,手机终端更新换代的速度不断加快。终端数量的增长速度已经趋于饱和,而手机终端的更换速度还拥有很大的发展空间。运营商在终端的营销是一个很好的选择。

目前运营商在终端营销的存量经营业务场景

主要包括:终端手机的换机预测、用户流量包业务推荐和终端换机导购。

### 2.1 终端手机的换机预测

终端手机的换机预测即利用用户的各项属性及其历史行为,预测用户的换机时间,从而指导运营商的终端营销,提升用户体验。

本文侧重研究预测终端换机间隔,从而为终端换机预测提供数据依据。

### 2.2 用户流量包业务推荐

随着人们生活水平的提高,物质文明已经不能满足人们日益发展的需求。精神文明成为这个时代的主题,人们的消费不断升级。移动互联网的发展,更是提高了人们的消费水平。视频、点播和直播的业务成为用户关注的重点。因此,数据流量业务成为运营商发展新用户、维系老用户、发展数据及信息业务的一个重要切入点。如何根据用户的消费情况和用户的历史流量使用情况为用户快速找到匹配的流量包类型,并及时有效地推荐,已经成为运营商精准营销的重要方向之一。

### 2.3 终端换机导购

作为移动业务的载体,终端是存量市场的维系,是数据及信息业务发展的一个重要方面。根据用户终端偏好,包括用户的换机频率、喜爱的品牌、喜欢的手机大小和喜欢的手机内存等信息,及时快速地发现终端换机的潜在用户并进行精准营销,为运营商提供价值。

在移动互联网大展拳脚的时代,运营商必须根据自身优势找到发力点。存量经营是运营商势在必行的发展方向。利用运营商积累下来的用户资源,为运营商创造价值。

## 3 模型应用系统架构

### 3.1 模型应用架构

图 1 为模型应用的系统架构,该架构包括数据源、数据挖掘与分析、业务场景和模型应用。

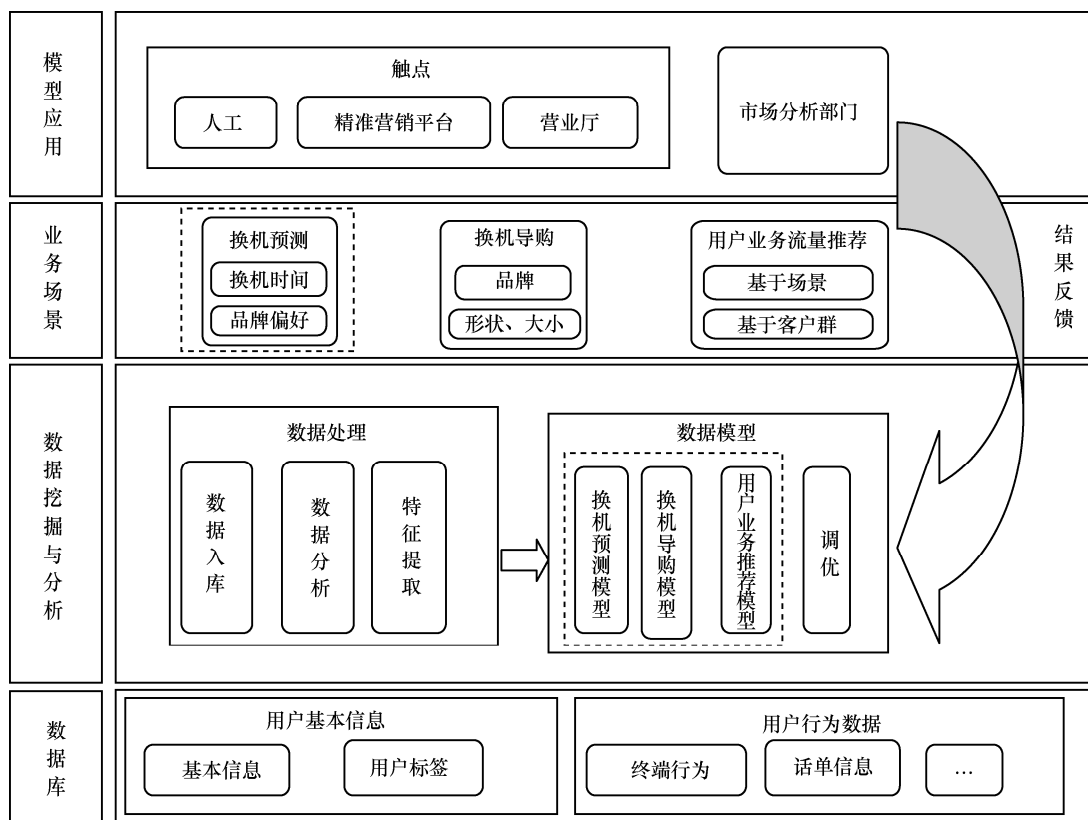


图1 中国联通大数据平台架构

### （1）数据源

数据源主要包括用户的基本信息和行为信息。用户的基本信息包括用户的基本信息和用户的基本标签信息，而用户的行为数据包括用户终端的行为、话单信息以及消费行为等。这些是做数据处理的基本依据，为业务层面的精准营销提供保障。

### （2）数据挖掘与分析

数据挖掘与分析主要包括数据处理和数据模型。数据处理主要是数据入库、数据分析和特征提取，为建模提供数据基础。数据模型是基于业务需求为业务提供基本保障。

### （3）业务场景

业务场景是存量经营中的实际应用，这里主要针对运营商终端营销方面的业务场景。终端营销方面的业务场景主要包括用户终端换机预测、用户终端导购及用户流量业务推荐等。业务场景为模型应用提供基础。

### （4）模型应用

模型应用是模型效果的展现方式，主要体现为精准营销平台、营业厅业务推荐等。精准营销平台为企业营销提供支撑，是企业价值得到最大化的体现。营业厅业务推荐可以根据用户的消费情况为用户提供更加合理的套餐。模型应用的实际效果能为数据挖掘与分析提供依据。

## 3.2 技术方案

本文根据具有相似消费行为的用户拥有一些共同的消费水平来预测用户的换机间隔，根据用户换机的开始时间和结束时间分别计算用户换机间隔的月份数，根据不同参数的限制条件逐层分裂树枝直到叶子节点，确定用户的换机间隔时间。

本文的技术方案主要分为4个部分，包括模型数据生成模块、模型的选择模块、模型结果分析模块和推荐模块，如图2所示。

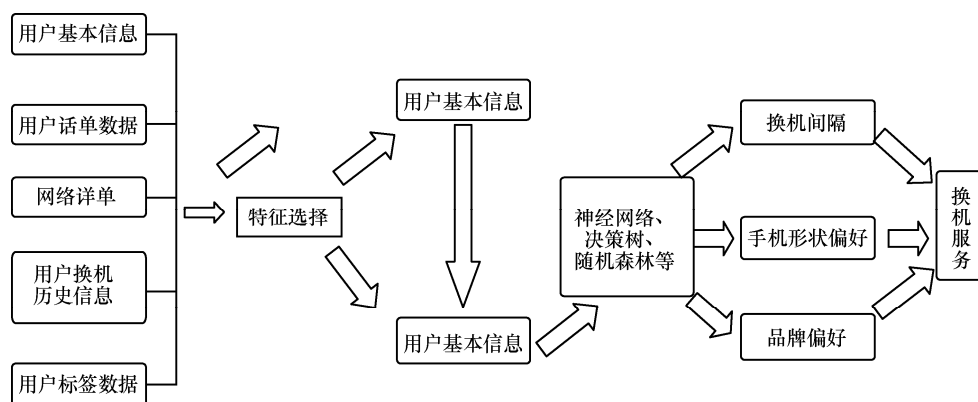


图2 换机预测技术方案

**数据生成模块：**根据中国联通的架构，对数据集进行处理最终形成用户不同维度的反应。根据用户不同维度的数据，以终端换机间隔的需求为依托进行特征的选择，生成模型需要的数据。

**模型选择模块：**依据数据的特点，单独选择决策树作为算法进行建模。

**模型结果分析模块：**根据建模的输出结果对模型进行分析，同时指导模型的选择和调参。模型结果的分析包括换机间隔、手机类型偏好和品牌偏好的分析。本文着重对换机间隔部分进行重点分析。

**推荐模块：**根据模型的结果，为用户合理地推荐机型，提高用户的满意度，提高企业的价值。

## 4 用户终端换机间隔预测模型

### 4.1 数据准备及分析

本文采用中国联通某省份公司某年的数据，结合业务需求，选取这一年内存存在手机终端更换行为的用户。

用户的基本统计信息如图3所示。**count**为总样本数，**mean**为平均值，**std**为标准差，**min**为最小值，**max**为最大值，25%、50%和75%对应1/4、1/2和3/4分位数。从图3可以看出，总样本数为30 000个，其中age列存在空值。

图4为异常值检测的箱型图，根据实际情况取出income不合理的数据。由于age中空值所占比例比较少，取出age为空值的数据。

|       | USERID         | AGE            | GENDER         | INCOME         | TRAFFIC_UP \   |
|-------|----------------|----------------|----------------|----------------|----------------|
| count | 30 000.000 000 | 29 730.000 000 | 30 000.000 000 | 30 000.000 000 | 30 000.000 000 |
| mean  | 15 000.500 000 | 38.142 281     | 1.483 800      | 2 598.876 633  | 12.167 786     |
| std   | 8 660.398 374  | 15.731 888     | 0.499 746      | 1 883.244 348  | 59.452 805     |
| min   | 1.000 000      | 10.000 000     | 1.000 000      | 0.000 000      | 0.000 061      |
| 25%   | 7 500.750 000  | 25.000 000     | 1.000 000      | 1 000.000 000  | 0.464 749      |
| 50%   | 15 000.500 000 | 38.000 000     | 1.000 000      | 2 542.000 000  | 2.858 611      |
| 75%   | 22 500.250 000 | 50.000 000     | 2.000 000      | 3 708.000 000  | 10.671 205     |
| max   | 30 000.000 000 | 69.000 000     | 2.000 000      | 8 000.000 000  | 2 773.796 715  |

|       | TRAFFIC_DOWN   | FREQUENCY      | TIMES          | CLASS          |
|-------|----------------|----------------|----------------|----------------|
| count | 30 000.000 000 | 30 000.000 000 | 30 000.000 000 | 30 000.000 000 |
| mean  | 82.221 175     | 82.624 500     | 2.091 800      | 3.778 200      |
| std   | 308.205 392    | 84.203 103     | 0.288 748      | 2.231 331      |
| min   | 0.000 000      | 0.000 000      | 2.000 000      | 2.000 000      |
| 25%   | 1.728 565      | 14.000 000     | 2.000 000      | 2.000 000      |
| 50%   | 13.706 488     | 52.000 000     | 2.000 000      | 2.000 000      |
| 75%   | 60.394 395     | 129.000 000    | 2.000 000      | 6.000 000      |
| max   | 7 144.217 633  | 364.000 000    | 3.000 000      | 8.000 000      |

图3 数据基本统计信息

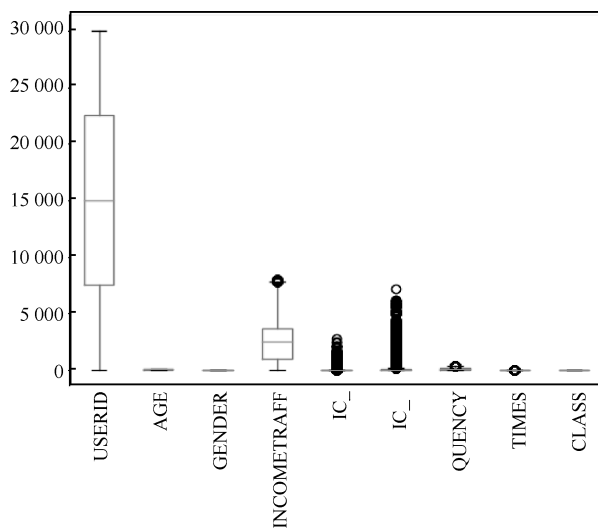


图4 异常值检测的箱型图

本文主要研究手机终端的更换间隔，以便为企业终端的精准营销提供数据依据。通过数据筛选、数据采样、数据探索、数据清洗和数据预处理，得到23 852个用户数据，其中：男性12 342个，占51.74%；女性11 510个，占48.26%。

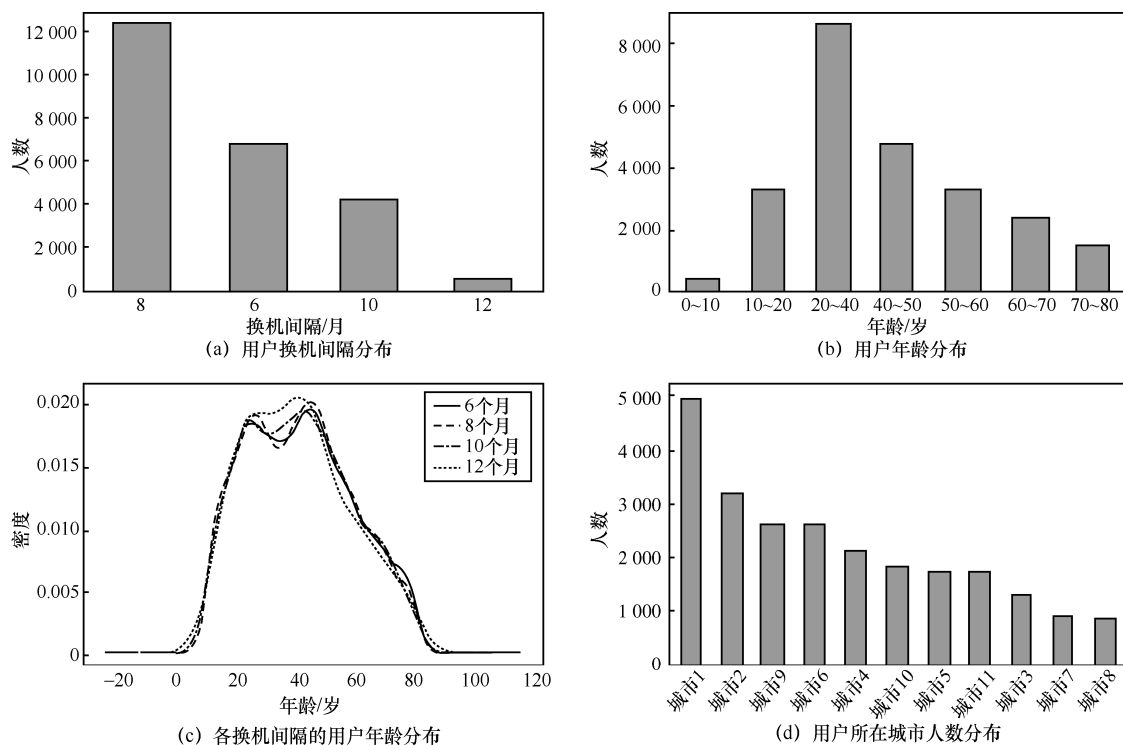


图5 基于中国联通某省分公司数据的统计分析

通过程序对用户进行进一步的处理,统计信息如图5所示。

从图5可以得到,用户主要分布在用户换机间隔为8个月的区间,用户的年龄主要集中在20~40岁,城市1居住的用户相对较多。通过上述分析,用户的年龄、所在城市可能对用户的换机间隔存在影响。

## 4.2 决策树算法流程

目前机器学习算法包括监督学习和非监督学习。机器学习包括决策树、随机森林、逻辑回归、支持向量机、朴素贝叶斯、 $k$ 近邻、 $k$ 均值、Adaboost算法、神经网络和马尔可夫算法。根据前面筛选的特征以及这批数据的特点,本文采用决策树作为分析的算法。

### (1) 决策树基本概念

决策树是经典机器学习算法中的分类算法。决策树是一种树形结构,决策树通过提取数据中的规则来表达数据中所蕴含的知识信息。决策树是自顶向下的递归方法,其基本思想是以信息熵

为度量构造一棵熵值下降最快的树。

例如你正在认真工作,突然弹出一个QQ信息。如何处理这条信息呢,无需处理、稍后处理还是紧急处理?图6给出了一条QQ信息的处理流程,这就是决策树构造过程。

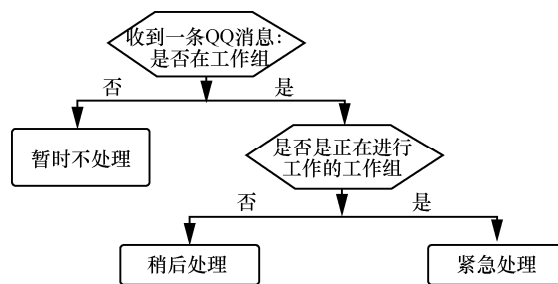


图6 QQ信息处理的流程

### (2) 信息增益

信息增益以信息熵为依据。信息是一个事件不确定性的体现,而熵描述了事件的不确定性。熵可以表示数据集的无序程度,定义如式(1)所示。

$$H(X) = -\sum_{i=1}^n p(x_i) \lg p(x_i) \quad (1)$$



信息增益表示得知一部分信息后而使目标信息的不确定性减少的程度。信息增益的定义如式(2)所示。

$$H(X) = -\sum_{i=0}^n p(x_i) \lg p(x_i) \quad (2)$$

### (3) 数据预处理

为了提高算法的准确度,根据第4.1节对数据进行预处理。

- 年龄划分为小于20岁、20~40岁、40~60岁、60岁以上4个部分。
- 用户的换机时间为6个月、8个月、10个月和12个月。
- 上行流量和下行流量相加作为总流量,流量数据将单位转化为GB。
- 收入划分为<2 000元、2 000~5 000元、>5 000元3个部分。

### (4) 决策树算法分类及流程

根据衡量信息不确定性减少的程度的不同,决策树分为ID3、C4.5和CART树。其中ID3以信息增益为衡量标准,C4.5以信息增益率为衡量标准,CART树以Gini系数为衡量标准。决策树构建树的过程就是通过衡量标准不断选择最佳特征,根据最佳特征划分数据集,直到不能划分数据集或信息增益减少到0为止。

决策树构建的伪代码如下。

Function 创建决策树

```

if 数据集中所有样本分类一致
    创建携带类标签的叶子节点
else
    寻找最好的特征
    根据最好的特征划分数据集
    for 每个划分的数据集:
        创建决策树
  
```

### (5) 决策树在实际工作中形成的流程

**步骤1** 在年龄、性别、收入、流量4个维度中寻找最好的特征,在本文中年龄是最好的特

征,按照年龄小于20岁、20~40岁、大于40岁将数据划分成3个数据集;

**步骤2** 对每个数据子集重复步骤1,直到不能进行分割为止;

**步骤3** 每个子集都分割完成,形成叶子节点,叶子节点即换机的周期结果。

4个特征的得分为:[age, gender, income, traffic] = [0.027 884 64 0.002 5147 7 0.055 972 71 0.001 786 72]。

4个维度的特征的重要性顺序为:收入>年龄>性别>流量。具体的决策树结果如图7所示,实际树的分支很多,图7仅画出一部分。

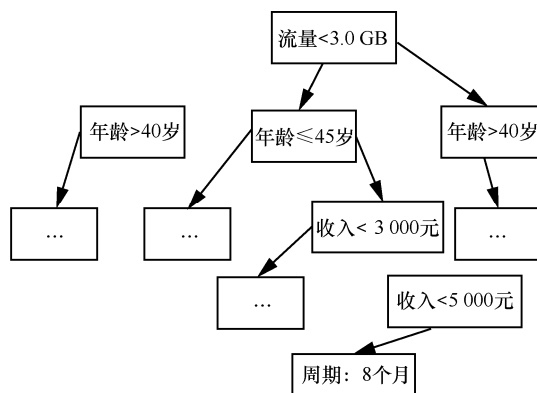


图7 数据形成的决策树流程

## 4.3 模型结果

结合用户特征分析,将用户的年龄、性别、上行流量、下行流量和用户所在城市作为模型的输入。采用Python中的sklearn算法包,从数据集中随机选出20%数据作为测试集,作为交叉验证使用。

交叉验证结果如下:

feat importance=[0.027 884 64 0.002 514 77 0.055 972 71 0.001 786 72]

dot: graph is too large for cairo-renderer bit-maps. Scaling by 0.0942371 to fit

决策树模型的正确率为: 0.445 6

正确的概率为 44.56%。

对数据进行聚类分析,将整体数据分为5类,如图8所示。

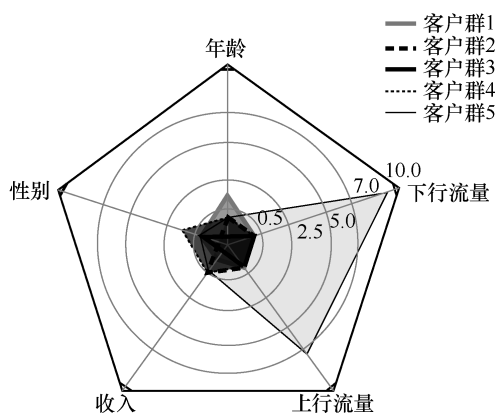


图8 各种因素对客户群的影响

从图8中可以看出，客户群1主要受到年龄因素的影响，客户群2主要受到性别和收入因素的影响，客户群3主要受到流量和性别因素的影响，客户群4受到5个因素影响的程度比较均衡，客户群5则主要受到流量的影响。

本文通过决策树分析了影响用户换机行为的各种因素。研究结果显示：流量对用户换机行为影响比较大；用户终端换机的年龄段主要集中在20~40岁；男性用户较女性用户更容易换机。这些都为运营商存量经营中的终端精准营销提供数据参考，具有很强的指导意义。

## 5 结束语

在移动互联网高速发展的今天，运营商需要适时改变业务类型，迎合时代的发展，提出新型的发展模式。数据和用户是运营商在以往经营中积累下的一个优势。存量经营是运营商业务发展的一个趋势。本文就存量经营的终端换机间隔进行研究，为运营商的精准营销提供了数据支撑。

后续通过对用户换机品牌偏好和用户手机类型的偏好进行预测，以更好地支撑用户终端换机预测。聚类分析的数据可以作为用户几个因素生成的个性化特征，可以把这部分加入用户特征工程的选择中。在算法的选择方面，可以多尝试几个算法以提高预测的准确性。

## 参考文献:

- [1] 魏进武, 张云勇. 电信运营商的大数据价值创造[J]. 信息技术, 2013, 7(6): 10-14.  
WEI J W, ZHANG Y Y. Big data value creation for telecom operators[J]. Information and Communications Technologies, 2013, 7(6): 10-14.
- [2] 陈天池, 王振, 陈开恒. 基于决策树 C5.0 算法的 4G 终端换机模型[J]. 信息通信, 2016(10): 261-263.  
CHEN T Q, WANG Z, CHEN K H. 4G terminal converter model based on decision tree C5.0 algorithm [J]. Information & Communications, 2016(10): 261-263.
- [3] 李福东. 面向移动互联网的位置营销关键技术与方法研究[J]. 信息技术, 2015(5): 91-96.  
LI F D. Research on location-based marketing key techniques and methods for mobile internet [J]. Information and Communications Technologies, 2015(5): 91-96.
- [4] 王雪琼, 熊珺洁, 姚晓辉. 基于大数据挖掘的终端换机模型[J]. 电信科学, 2016, 32(12): 43-52.  
WANG X Q, XIONG J J, YAO X H. Terminal replacement model based on big data mining[J]. Telecommunications Science, 2016, 32(12): 43-52.
- [5] 沈磊, 虞慧群, 范贵生, 等. 基于随机森林算法的推荐系统的设计与实现[J]. 计算机科学, 2017(11): 170-173, 192.  
SHEN J L, YU H Q, FAN G S, et al. Design and implementation of recommender system based on random forest algorithm[J]. Scientific Journal of Computer Science, 2017(11): 170-173, 192.
- [6] 宋燕辉, 陈霖. 基于大数据挖掘的终端感知与换机应用研究[J]. 湖南邮电职业技术学院学报, 2015(4): 13-16.  
SONG Y H, CHEN L. Application research on terminal perception and replacement based on big data mining[J]. Journal of Changsha Telecommunications and Technology Vocational College, 2015(4): 13-16.
- [7] 刘畅. 基于 Cox 回归模型的用户终端换机研究[J]. 电子科学技术(北京), 2016, 3(4): 418-421.  
LIU C. Customers' cell phone replacement study: based on cox regression model [J]. Beijing: Electronic Science & Technology, 2016, 3(4): 418-421.
- [8] 宫夏屹, 李伯虎, 柴旭东, 等. 大数据平台技术综述[J]. 系统仿真学报, 2014, 26(3): 489-496.  
GONG X Y, LI B H, CHAI X D, et al. Survey on big data platform technology [J]. Journal of System Simulation, 2014, 26(3): 489-496.
- [9] 张第, 张鹏, 杨宇. 音乐粉丝用户的大数据挖掘方法[J]. 电信技术, 2016(3).  
ZHANG D, ZHANG P, YANG Y. Big data mining method for music fan users[J]. Telecommunications Technology, 2016(3).
- [10] AMIN R K, INDWIARTI, SIBARONI Y. Implementation of decision tree using C4.5 algorithm in decision making of loan



- application by debtor[C]//2015 3rd International Conference on Information and Communication Technology (ICICT), May 27-29, 2015, Nusa Dua, Bali. Piscataway: IEEE Press, 2015.
- [11] KAMBATLA K, KOLLIAS G, KUMAR V, et al. Trends in big data analytics[J]. Journal of Parallel & Distributed Computing, 2014, 74(7): 2561-2573.
- [12] DUAN L, XIONG Y. Big data analytics and business analytics[J]. Journal of Management Analytics, 2015, 2(1): 1-21.
- [13] 吴道, 刘启兵, 汪飞. 决策树算法在实际中的应用[J]. 工业控制计算机, 2017(12): 120-121.  
WU X, LIU Q B, WANG F. Application of decision tree algorithm in practice[J]. Industrial Control Computer, 2017(12): 120-121.
- [14] 韩艳敏, 郑卫荣, 张杨, 等. 基于 C5.0 决策树的在线协作学习因素与路径分析[J]. 教育现代化, 2017(49).  
HAN Y M, ZHENG W R, ZHANG Y, et al. Analysis of online collaborative learning factors and path based on C5.0 decision tree[J]. Education Modernization, 2017(49).
- [15] 谭俊璐, 武建华. 基于决策树规则的分类算法研究[J]. 计算机工程与设计, 2010, 31(5): 1017-1019.  
TAN J L, WU J H. Classification algorithm of rule based on decision-tree[J]. Computer Engineering and Design, 2010, 31(5): 1017-1019.
- [16] MA B, WANG D, CHENG S, et al. Modeling and analysis for vertical handoff based on the decision tree in a heterogeneous vehicle network[J]. IEEE Access, 2017: 1.

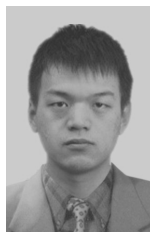
## [作者简介]



崔羽飞 (1990- ), 男, 中国联合网络通信有限公司研究院大数据研究中心大数据工程师, 主要研究方向为大数据处理、Web 研发等。



魏进武 (1978- ), 男, 博士, 中国联合网络通信有限公司研究院大数据研究中心主任、高级工程师, 主要研究方向为云计算、大数据以及 IT 架构的技术研究、系统设计和研发验证等。



潘思宇 (1994- ), 男, 中国联合网络通信有限公司研究院大数据研究中心机器学习算法工程师, 主要研究方向为数据挖掘、算法研究、数学建模等。



张溶芳 (1991- ), 女, 中国联合网络通信有限公司研究院大数据研究中心机器学习算法工程师, 主要研究方向为数据挖掘、建模、数据可视化等。



王笑 (1986- ), 女, 中国联合网络通信有限公司研究院大数据研究中心大数据工程师, 主要研究方向为前端开发、界面设计等。

贾子翔 (1991- ), 男, 中国联合网络通信有限公司研究院大数据与 IT 技术研究中心数据工程师, 主要研究方向为大数据平台研究开发以及数据加工、可视化等前后台开发。



国之重器出版工程

网络强国建设



## 5G移动通信： 面向全连接的世界

5G Mobile Communication:  
A Fully Connected World

刘光毅 方敏 关皓 李云岗 孙程君 编著

人民邮电出版社

书号：ISBN 987-7-115-50792-1

定价：169.00元

出版时间：2019年3月

### 内容简介

本书从5G需求与愿景出发，深入揭示了5G产生的背景、频率需求、候选频率及其无线传播特性，围绕5G相关的主要标准化组织及其研究的最新进展对5G标准的未来进行了展望；围绕5G的关键技术，分别从多天线的演进、非正交与多址、全双工、编码与链路自适应、网络架构、用户无感知的移动性管理、以用户为中心的自治网络、毫米波系统设计、超密集网络及小区虚拟化以及物联网的优化等方面详细介绍了5G的关键技术，最后对3.5 GHz 5G样机的设计与外场试验情况进行了介绍，并展望了5G应用与部署。

本书全方位地系统阐述5G的来龙去脉以及未来主要发展方向，是5G研发、无线通信技术教学以及工程实施人员有益的参考书。

ISSN 1000-0801



邮发代号：2-397 国外代号：M841 定价：68.00元