终端数据挖掘技术研究

Research of Terminal Data Mining Technology

路　玮，李轶群，李佳俊，王蕴实（中国联通网络技术研究院，北京 100048）
Lu Wei，Li Yiqun，Li Jiajun，Wang Yunshi（China Unicom Network Technology Research Institute，Beijing 100048，China）

摘　要：
随着5G时代的到来，通信行业进入快速发展阶段。扩大终端销售规模成为运营商拓展市场的战略重心，同时终端换机需求量越来越大，终端换机数量呈增长趋势。通过数据挖掘技术对终端换机业务进行数据分析，运用决策树算法预测用户换机变量因子，挖掘潜在换机用户需求，为终端销售提供准确依据。

关键词：
终端换机；数据挖掘；决策树
doi：10.12045/j.issn.1007-3043.2019.10.014
中图分类号：TN929.5
文献标识码：A
文章编号：1007-3043（2019）10-0062-04

Abstract：
With the advent of 5G，the communication industry has entered a rapid development stage. The expansion of terminal sales scale has become the strategic focus for operators to expand the market. At the same time，the demand for terminal replacement is increasing，and the data volume of terminal replacement is increasing. Data mining technology is used to analyze the data of terminal replacement business. Decision tree algorithm is used to predict the variable factors of user replacement，which can excavate the potential replacement user demand，and provide accurate basis for terminal sales.

Keywords：
Terminal replacment；Data mining；Decision tree

引用格式：路玮，李轶群，李佳俊，等. 终端数据挖掘技术研究[J]. 邮电设计技术，2019(10)：62-65.

0 引言

随着移动互联网终端不断增加，移动互联网业务需求呈爆炸式增长，运营商转型之路必将围绕海量数据所带来的商机做深度挖掘分析。本文利用深度学习算法，同时结合LTE网络大数据的分析挖掘，发现隐藏在庞大数据背后的业务规律，通过大数据分析挖掘找出数据存在的关联关系，分析现网用户换机情况预测用户未来需求，通过数据间的规律，预测业务新需求，并将其转化为新业务和新产品。

1 终端换机数据挖掘关键技术

随着4G移动技术不断走向成熟，移动终端用户数量急剧增加，运营商运用大数据挖掘技术对现网终端出账数据进行深度挖掘，探索用户换机潜在因素。数据挖掘是以人工智能为基础的业务信息处理技术，通过对数据库中的大量业务数据进行抽取、转换、分析和其他模型化处理，发现大量数据中的潜在信息，获取有利于业务决策的关键性数据。数据挖掘包括业务理解、数据理解、数据准备、数据预处理、数据建模、模型评估等6个环节。如图1所示，这6个环节之间是可以相互交互的，例如在数据理解阶段如果发现现有数据无法解决业务理解阶段提出的问题，则需要重新调整定义业务问题或者采集更丰富的原始数据去论证问题；如果在建模阶段发现数据无法满足建模需求，则需要重新处理数据直至满足建模要求；如果在模型评估阶段发现模型预期结果不理想，则重新回到业务理解阶段审视问题合理性并进行调整。

收稿日期：2019-07-26

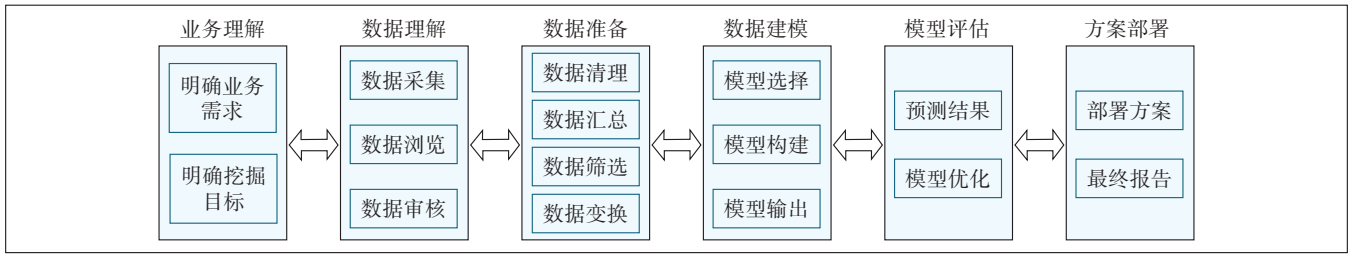

图1　数据挖掘框架

本文采用IBM SPSS Modeler工具进行数据挖掘操作，终端数据处理过程主要有以下几个关键步骤。

a）首先在业务理解阶段需要从海量数据中发现有用的信息从而明确业务目标和数据挖掘目标，制定详细的项目计划。移动终端换机技术研究依托现网每月出账数据，对现在市场中终端2G/3G/4G网络制式换机情况，以及2G/3G/4G换机终端的型号、品牌、功能支持情况、3G/4G终端驻留网络等内容进行挖掘分析。在充分理解终端业务后，对不同客户群体进行挖掘，并了解现网用户终端使用情况。

b）明确业务目标后需要对各个数据源的数据进行整理，在数据理解阶段主要完成对数据资源的初步认识和清理，收集原始数据，并对数据进行梳理和描述，对数据进行探索性分析形成数据质量报告。终端数据是来自各个省份的原始数据、市场部数据以及网络平台数据。收集各个数据源的终端数据，统一各种数据类型的格式，并对各种信息进行筛选、过滤、剔除等处理。在处理数据之前，需要明确数据来源。通过数据审核节点的统计数据可以发现数据的异常和极端数据。

c）数据预处理准备阶段是将同数据源或不同数据库中的数据表进行整合，生成可以建立数据挖掘模型的数据集。在数据准备过程中需要对数据进行清理，确定需要的有效数据，调整或剔除不符合实际情况的数据，然后对相关数据进行合并处理或重构成新的字段或数据。数据预处理过程主要包括数据清洗、数据汇总、数据转化、数据筛选。例如各个省份的原始终端数据格式不统一，有的省份提供TAC信息，有的省份提供的IMSI信息，为了统一处理，需要将TAC信息和IMSI信息进行转化。有的省份提供的数据包含很多无效数据，例如空白数据、无效字符等，需要筛选和删除这些数据，并去除重复数据才能得到有效数据。根据业务需求和数据变量间的相关性，将原始数据派生成新的变量，例如派生成换机标识用于判决用户终端是否换机。

d）模型建立是数据挖掘的核心阶段，首先需要选择建模模型，通过对模型的假定和要求来对模型技术进行评估，并对模型效果进行检验。初步建立模型后可根据实际情况调整模型的各个参数，并对模型使用进行评价。不同的数据挖掘模型有不同的挖掘算法，不同的技术方案产生的模型差异也很大。数据模型有分类、聚类、关联规则、神经网络等模型，其中决策树方法和神经网络模型最为常用。本文采用决策树模型，通过输入终端2G/3G/4G网络制式支持情况、终端品牌、价格、终端3G/4G网络附着使用情况等，对终端换机情况进行分类预测，判断影响终端用户流失的关键因素。

e）模型评估可以从技术角度对模型效果进行评价，也可从业务角度对模型在现实业务环境中的适用性进行评估，从而筛选出被认可的数据挖掘模型。评估模型是否达到预期效果的指标有很多种，其中结果准确率是一个重要指标。通过模型评估，可以将数据挖掘的结果运用到实际业务中。

f）在方案部署阶段对预测结果方案进行部署，同时形成最终的报告。通过预测换机变量的重要性，可以对换机的用户终端采取不同的营销方案，对不同用户换机需求采取不同的优化部署。

2 换机业务挖掘模型

IBM SPSS Modeler具有丰富的数据挖掘算法，如图2所示，通过数据库之间的数据和模型的交互，使数据在各个节点间的流动，形成1条或多条数据流，然后通过执行数据流完成数据分析任务。在数据分析过程中可以对数据节点进行调整和修改，通过数据收集、数据预处理、模型建立、模型评估等环节，将不同省份的原始数据或其他源数据进行分析处理以满足不同业务需求。

本文利用Moderler工具进行导入、统计、分析、预测等操作。采集数据来源于省份提取的每月全网出账用户终端数据，包含终端的归属地（市）、终端移动

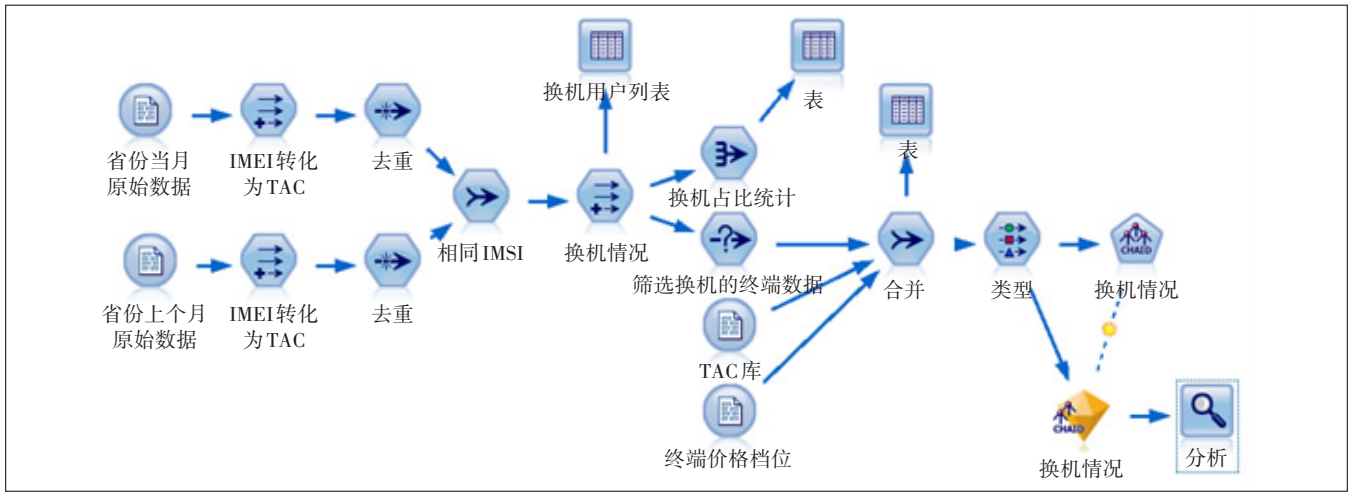

图2 数据挖掘处理思路

设备国际识别码(IMEI——International Mobile Equipment Identity)、用户编号、国际移动用户标识(IMSI——International Mobile Subscriber Identity)、4G网络附着标识、3G网络使用标识等基本字段信息。终端TAC库体现终端基本属性信息,包含TAC号、终端品牌(Marketing Name)、终端厂商(Manufacturer)、支持的频段(Bands)、2G标识、3G标识、4G标识、双卡、设备类型等基本字段信息。终端价格表信息体现不同终端价格档位基本信息,包括终端ID、终端型号、厂家编码、厂家名称以及终端价格档位等字段信息。通过Modeler导入省份终端数据,浏览数据内容,对数据进行过滤保留有用字段;确认字段存储类型,例如将IMSI号和IMEI号存储类型从字符串变更为整数;变更变量类型,例如将2G/3G/4G标识变更为连续类型。通过数据审核节点统计数据可以筛选出异常、极端数据。

对于异常数据需要在数据预处理阶段进行修改或删除。为了提取有效数据,在数据预处理环节需要对数据进行清洗、转化、加载等一系列处理,将省份提供的话单数据转化为标准格式,将一些不完整的数据信息或错误数据进行删除,以便后期分析加工和处理。首先需要将原始数据中缺失值进行替换或删除。缺失值是指空白数据或不合理数据。采用函数@BLANK(@FIELD),在@FIELD字段中填写IMEI、Manufacturer、2G标识、3G标识、4G标识等需要检查的字段,并将字段中的空值改为0;采用函数@NULL(@FIELD)将系统中缺失值$null$替换为0。对数据记录缺失值的处理可以减少分析结果偏差。然后利用函数intof('IMEI号'/1000000)将IMEI值转化为TAC值,并与TAC库对比可获得终端基本信息。对IMEI进行去重,保证用户终端的唯一性。终端IMEI是终端设备唯一标识,它与每台手机终端是一一对应关系。换机是对比同一用户IMSI号下不同IMEI号的终端信息,IMEI_pre表示换机前终端的IMEI信息,IMEI_cur表示为换机后终端的IMEI信息,如果IMEI_pre=IMEI_cur,则表示没有换机,换机标识为0;如果IMEI_pre≠IMEI_cur,则表示用户换机,换机标识为1。在建模前需要将换机标识为1的用户终端筛选出来,通过对这些换机用户的行为特征进行分析得出影响用户换机的重要因素。利用选择节点筛选换机标识=1的终端数据。通过一系列数据流处理过程,将省份上个月和当月的终端换机业务数据统计出来汇总在一张数据表中。

3 预测模型及结果

根据数据分析结果得出A省有300多万用户更换了手机,本文运用决策树CHAID模型预测出换机变量的重要性排序。决策树是通过数据学习,依据输入的数据变量推测输出变量的分类取值,对数据对象进行分类预测,清晰显示每个字段的重要性。CHAID模型优点是可产生多个分支,从统计角度可以确定分支变量和分割值从而优化分支过程。用户换机可能考虑的因素有终端价格档位、用户对终端品牌和型号的爱好、终端网络制式的变更(如2G/3G终端用户变更使用4G终端)、终端属性(如是否是双卡终端)等。结合以上用户换机因素,用换机业务模型对A省连续2个月的63万条终端出账数据进行分析,选取终端换机的变量参数有以下3类。

a)现网出账用户基本信息数据:终端的IMSI、

IMEI、60天内登录过4G网络标识、3G网络使用标识等（60天内登录过4G网络标识为1的表示终端未换机，标识为0表示终端已经换机；3G网络使用标识为1表示终端未换机，如果标识为0表示终端已换机）。

b）终端基本属性信息数据：终端厂商、终端品牌、支持2G/3G/4G网络标识等。

c）终端价格基本信息数据：终端价格档位等。

通过用户行为分析筛选出10个重要建模变量，其中将换机情况作为目标预测结果，其他参数变量设为输入变量，如图3所示。

为了提升分析模型准确度，加入“分区”节点，将数据分为50%训练数据和50%的测试数据。训练模型通过50%的换机数据进行模型预测，评估模型参数来以确定最合适的预测模型。

图3　字段类型定义

基于训练样本集的模型参数对测试样本集数据进行数据分析可得出如图4所示的结论。

从模型分析结果来看，2G/3G/4G标识、终端品牌、终端价格、60天内登录过4G网络的标识、3G网络使用标识、双卡等参数变量对模型构建起关键作用。其中用户换机时对网络制式变更的需求最大，其次是终端品牌。

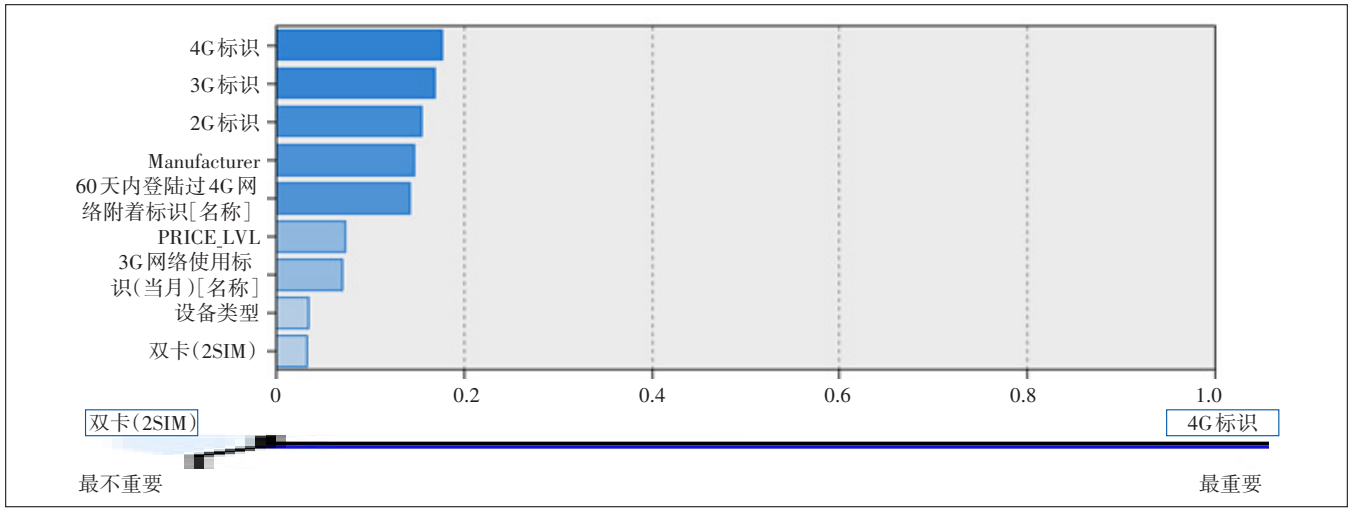

图4　预测变量重要性

对数据模型进行评估，分析数据节点的准确率，通过对模型测试集的准确率分析可以判断模型的准确性。通过对换机情况测试集的分析结果进行统计，发现其预测正确率为90.67%，表明结果非常理想。

终端换机预测只是数据挖掘应用的一部分，通过此模型还可以挖掘出终端在市场营销、客户服务等多方面的应用。

4　结束语

利用数据挖掘技术对终端换机行为的深入研究发现，终端换机因素既与用户使用爱好习惯有关，又与用户消费情况有关，每个指标对用户换机的影响程度不尽相同。通过预测换机变量重要性，可以提高运营商和终端厂商的营销精准度，有利于开拓市场，提升营销业绩；同时为运营商和终端厂家提供用户喜好规律，为终端销售制定生产计划提供依据。

参考文献：

[1]　宋春涛，张帆，曹振. 电信运营商的终端大数据分析及应用[J]. 邮电设计技术，2016(8)：41-46.

作者简介：

路玮，工程师，主要从事无线通信相关业务和技术研究等工作；李轶群，高级工程师，主要从事无线通信相关业务和技术研究等工作；李佳俊，高级工程师，主要从事无线通信相关业务和技术研究等工作；王蕴实，工程师，主要从事无线通信相关业务和技术研究等工作。