基于随机森林模型的电信运营商外呼推荐系统

朱奕健[1] 张正卿[1] 黄一清[2] 白瑞瑞[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 在电信运营商领域,外呼推荐是一种重要的推荐产品和服务的途径。实现了一种基于运营商大数据的自动外呼推荐系统,该系统能够挖掘用户的行为特征并且使用机器学习的方法预测用户对于被推荐产品的接受可能性。传统推荐系统使用的模型算法为矩阵分解、大规模稀疏特征分类、神经网络等。采用随机森林算法的主要原因是随机森林具有并行化程度高、训练速度快、生成的决策树可解释等诸多优点,适合于基于电信业数据的推荐系统。该外呼推荐系统基于 Hadoop、Impala 和 Spark 等大数据处理平台及工具,使用随机森林分类器作为核心算法,将用户最近的行为特征回归为接受外呼推荐产品的可能性。在线测试表明使用该系统与当前部署的人工随机外呼相比,能够提升约 41%的用户接受率;同时,根据模型算法输出特征的重要性,进一步给出了两类用户的特征分析。

关键词 外呼,推荐系统,随机森林,电信运营商
中图法分类号 TP181 **文献标识码** A

Random Forest Based Telco Out-calling Recommendation System

ZHU Yi-jian[1] ZHANG Zheng-qing[1] HUANG Yi-qing[2] BAI Rui-rui[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 Out-calling recommendation is widely used in recommending products and services to customers by telecommunication (telco) operators. In this paper, we developed an automatic out-calling recommendation system which relies on telco big data. This system uses data-mining methods to extract customer behaviors and machine-learning algorithm to predict the acceptance probabilities when customers are recommended to certain products. Different from most recommendation systems which use matrix factorization (MF), sparse features classification, neural network and etc, this paper used random forest, not only because the algorithm is easy to be parallel implemented and has fast training speed, but also the rules from the resulting decision trees are easy to explain. These characteristics make the random forest suitable for telco recommendation system. Our system is implemented on the top of Hadoop, Impala and Spark. Random forest is used as the core algorithm to calculate the acceptance probability when a user is recommended to a product based on user behavior features. Online testing shows that the proposed system can achieve 41% improvement compared with the current deployed random out-calling recommendation method. We also gave the customer behavior analysis according to the feature importance from the outputs of random forest.

Keywords Out-calling, Recommend system, Random forest, Telco operator

1 引言

近年来,推荐系统[1,2]已经被广泛应用于各个领域,例如电子商务、音视频网站、社交网络、个性化阅读等,并已经融入人们的生活。对于用户来说,推荐系统丰富和改善了用户的各项体验,让用户能够从当今的信息爆炸中获取自己需要的信息或产品。对于商家来说,推荐系统作为一项新的技术,能够提升用户兴趣与关注度,从而为他们带来更多的商机。因此,关于推荐系统的研究成为了最近的热点。

如何正确有效地向用户推荐产品是目前很多商家和企业面临的问题。一次成功的推荐能够激发用户对于推荐产品的兴趣,进而推动用户购买所需的产品或服务。电信行业对推荐系统的需求也日渐增多,以电信运营商为例,电信运营上有两大显著特点:1)用户数众多;2)产品与服务数众多。这两大固有特点很好地符合推荐系统的研究与应用场景,运营商希望用户能够尽可能地选择他们所需要的产品与服务,这就需

本文受江苏省科技支撑计划重点项目(BE2014005),国家自然科学基金(61572339)资助。
朱奕健(1975-),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、物联网技术、大数据技术发展研究、平台构建以及大数据应用;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;黄一清(1989-),男,博士生,主要研究方向为机器学习与数据挖掘;白瑞瑞(1989-),女,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)。

要有自动的方法能够为海量用户推荐其最可能感兴趣的产品和服务。

运营商向用户推荐产品和服务的渠道有多种：1）通过短信形式向用户发送推荐产品的信息；2）在运营商的门户网站上向用户推荐，也可在用户登录网站后做个性化的推荐；3）采用外呼的形式，直接通过运营商客户服务热线向用户拨打电话，与用户沟通推荐产品和服务。其中，外呼推荐形式由于是直接向用户拨打电话，比起短信推荐与门户网站推荐来说更容易造成用户的反感，因此外呼推荐的产品与服务的准确性会对运营商的客户关系管理造成较大的影响。客户如果多次接听到运营商推荐自己所不感兴趣的产品和服务的电话，会对运营商的服务质量产生不满情绪，进而投诉运营商甚至可能选择更换运营商。

除了积极主动地向用户推荐新的产品与服务外，在运营商的客户关系管理模型中，推荐系统还能够起到维系挽留将要离网用户的作用。用户流失是多数公司企业都会遇到的问题，运营商也不例外。随着电信运营行业的迅猛发展，各运营商之间的竞争日益激烈。用户有多种选择，如果对一家运营商的服务体验不满意，会选择使用另一家的服务，这样势必会造成用户流失。上海某运营商 2014 年的预付费用户月均离网率为 10%，后付费用户月均离网率为 5%，针对客户流失问题，我们研发了一套用户离网预测模型[3,4]，前 5 万最可能离网用户的预测结果精度约为 96%。根据预测结果，运营商会采取一些营销措施尝试维系挽留即将离网的用户，其本质依旧是通过外呼和短信的形式给不同的用户推荐不同的产品和服务。若推荐的产品和服务能够抓住使用户产生离网倾向的痛点，用户便会接受该产品和活动，继续留在网内使用运营商的各项服务。成功推荐维挽营销活动给即将离网的用户能够降低运营商的客户流失率，在如今运营商获取新客户成本远高于挽留用户成本的现状之下，降低运营商的客户流失率无疑能够进一步巩固客户关系，增加企业利润。相较于短信这种弱接触方式，外呼维系挽留能增强用户的感受，使维系挽留的效果更好。这就要求为每个外呼提供准确的产品和服务推荐。所以推荐系统是运营商产品推荐和用户维挽等业务所需的核心技术。

运营商拥有大量用户数据，包括用户详细信息、账单信息、通话详单、套餐订购情况和网络使用情况等。这些数据是设计推荐系统的关键。因为通过系统地分析这些用户的历史数据，可以挖掘出很多隐藏在这些数据之后的信息，这些信息刻画了用户的兴趣爱好、行为习惯、使用偏好等个性化的特征，进而可以更好地用于外呼推荐系统的建模。

电信外呼推荐系统是基于大数据的推荐，它处理的数据满足大数据的 4V（Volume，Variety，Value，Velocity）特性：Volume 指数据量大。众所周知，运营商在全国各地用户数众多，业务每时每刻频繁地发生，产生的数据量非常可观，据不完全统计，上海某中型规模运营商一天的数据生成量在 5TB 左右。Variety 指数据来源的多样性。以部署本文系统的运营商为例，从大类上分为业务支持数据（OSS）与运维支持数据（BSS），其中，BSS 富含丰富的用户信息，如用户基本信息、用户各种通讯行为、用户上网记录、用户账单等信息；OSS 富含丰富的网络侧及信令数据，包含用户体验数据如掉话率、接通率、上网时延等，测量报告如用户通话和上网时的各项电平信号数据，连接的基站与备选基站数据等，数据来源多种多样。Value 指数据背后蕴藏巨大的价值。现如今，各家运营商都在寻求大数据的利用方法，大致可以有 3 个方向：用户洞察、网络洞察与数据变现。Velocity 指时效性，具体是指对于大数据建模来说，如果数据源的更新速度越快，模型用到越新的数据，给模型带来的精度提升就越明显，简而言之，时效性是指数据的新旧更替速度，运营商的数据只要有业务发生就会更新。

外呼推荐包含两个主要步骤，第一步需要从运营商掌握的数据中构造出能充分表述用户的有用的特征，第二步是构建一个有效的分类算法。推荐本质上是一个排序过程，用户对于推荐的产品与服务都会有一个接受强度，代表用户愿意接受所推荐的产品的意愿强度，用户和产品的匹配程度就可以通过接受强度排序计算得到。抽象来看，推荐过程就是一个分类的过程，分类器将用户分为接受与不接受两类并同时给出分类强度。我们首先从运营商的若干数据表里抽取有用的特征组成大宽表，用于生成训练、测试数据集，然后选择分类器算法训练带有标签的训练数据集得到分类器模型。之前的相关工作对于解决推荐系统的问题，普遍采用基于内容推荐[1]、协同过滤推荐[2]、基于规则的推荐与基于知识的推荐等。这些方法在处理电信大数据时都存有弱点，基于内容时推荐和协同过滤推荐无法处理稀疏数据问题，无法解决新用户的推荐问题；基于规则推荐的方法对于规则的抽取很困难，尤其是对于大数据的应用场景，由于规则是全局性的，无法做到个性化推荐；基于知识的推荐是静态推荐。由于以上常用的推荐系统方法对于大数据的应用场景或多或少存在缺陷，对于运营商的外呼推荐建模，我们采用的是集成学习算法中的随机森林分类器算法。

本文提出了一种全新架构的推荐系统——基于随机森林模型的外呼推荐算法，并在上海某运营商部署了该套系统。选择随机森林作为外呼推荐的分类器算法最主要的原因是其具有训练速度快、能够处理稀疏高维数据的特点，并且随机森林分类器的输出可以是[0，1]之间的实数，这种软分类输出结果能够方便地进行排序。在实验部分给出了使用随机森林算法和其它常用算法如逻辑回归（LR）、矩阵分解（LibFM）和迭代决策树（GBDT）的对比。实验表明，在电信外呼推荐场景下，随机森林在预测精度上有着明显的优势。因为电信数据满足大数据的 4V 特性，所设计的推荐系统构建在 Hadoop、Impala 和 Spark 等大数据处理平台及工具基础上。通过真实上线对比测试，本系统大幅提高了运营商外呼推荐需求的接通率与转化率。

2 系统架构与实现

2.1 基于大数据平台的四层架构

整个外呼推荐模型的系统架构一共包含 4 层，它们分别是数据存储层、数据分析层、模型算法层、业务应用层。各层向上层开放功能接口，为上层提供服务。下层架构依托于大数据平台，能够直接对运营商的一切数据进行高效、合理的存放及使用。图 1 展示了整个系统的层结构框架。

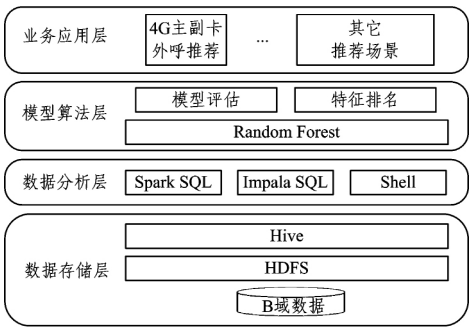

图1　推荐系统模型架构图

2.1.1　数据存储层

本系统所用的数据基于上海某运营商的大数据平台。每天该大数据平台都会生成大约2.3TB的数据，这些数据分为两大类，BSS(Business Support System)业务支持系统数据及OSS(Operation Support System)运维支持数据。通常BSS数据包含有用户基本信息、用户行为、账单信息、语音详单、短信详单及上网详单等，这些数据是刻画用户使用习惯和兴趣爱好等特征的关键。而对于运营支持系统，它的数据主要包含3部分，分别是CS(Circuit Switch)数据、PS(Packet Switch)数据及MR(Measurement Report)数据，其中CS数据是指通话连接质量数据，例如通话掉话率和通话连接成功率等；PS数据是指用户手机上网行为数据，包括手机上网速度、连接网络成功率和移动搜索等；MR数据是采自无线网络控制器，测量报告中蕴含着用户的地理位置信息。BSS数据与用户密切相关，也称为B域数据，包含有140多张数据表。在数据管理这一层面，我们选择使用Apache Hadoop分布式结构来存储和管理数据。Hadoop的分布式文件系统HDFS可以处理PB级别的超大数据，并且支持流式访问数据；同时Hadoop的数据仓库工具Hive可以提供SQL查询功能，并且可以将SQL语句转化为MapReduce任务分布式运行，容易实现并行化，为上层应用特别是Spark和Impala建立了良好的底层环境支撑。

2.1.2　数据分析层

数据分析层主要配置并部署了Spark SQL与Impala SQL，用于对数据的分析与处理。Spark是UC Berkeley AMP实验室开源的类HadoopMapReduce的通用的并行计算框架，Spark基于MapReduce算法实现分布式计算。Spark的分布式计算效率很高，如今普遍用于大数据的分析处理。Spark SQL支持在Spark中使用SQL关系型查询表达式。Impala是一个在Hadoop集群上运行的本地SQL查询引擎，提供原始HDFS数据的简单查询访问。Impala只依赖Hive的元数据，不使用MapReduce的方式执行任务，中间结果存储于内存中，因此运行速度快于Hive，可以很方便地用作数据预处理工作。在数据分析层，底层的原始业务数据(BSS数据和OSS数据)会经过一系列的数据预处理工作，为随后的特征工程做准备。特征工程是基于Spark SQL及一些广泛使用的非监督/监督学习算法来完成的，包括PageRank[5]、Label Propagation[6]、主题模型[7]等。项目涉及的所有原始数据都存储于HDFS中，然后用Spark SQL的关联和聚合操作提取原始数据表中的有用字段做成临时表，这些临时表可以重复使用，最后将所有的临时表特征字段通过数据表连接的方式形成一张大宽表，表内每一行记录就表示一个用户特征。外呼推荐模型涉及到的特征主要分为5大类：用户基本信息特征、CS特征、PS特征、基于通话图与短信图的特征及二次特征。用户基本信息特征主要是从BSS数据中抽取，包括年龄、入网时长、性别、卡归属地、账单、余额、通话频率、通话时长、短信频率、短信时长、上网频率、上网时长、投诉频率、充值金额等。而CS及PS特征则是从OSS数据中提取，CS特征是用户通话质量方面的特征，主要包括掉话率、信号电平值等，他们可以评估用户语音服务的质量；而PS特征是网络服务相关特征，主要的数据指标有数据包吞吐量、丢包率、时延等，可以用来评估用户数据流量服务的质量。基于图的特征是通过PageRank和Label Propagation算法从通话图、短信图中抽取的，PageRank的特征代表用户在他的通话关系图中的重要程度，Label Propagation的特征代表用户受到其所在的通话关系图中其它被推荐用户的接受活动情况对其造成的影响值。而二次特征则是通过LIBFM[8]将建模时特征重要性排名前5位的特征两两相乘而得。最后将所有的特征在Spark计算框架中拼接成全量数据的特征向量，每个用户的特征向量表示为 $X_m=[x_1,\cdots,x_i,\cdots,x_j,\cdots,x_N]$，其中 x_i 表示用户 X 的第 i 个特征。

2.1.3　模型算法层

模型算法层包含整套架构所使用的核心模型算法，数据从上一层的数据分析层流入该层，这一层主要使用随机森林作为分类器模型进行外呼推荐模型的训练、预测。上一层流入的数据分为两部分，带标签的训练数据与不带标签的预测数据。模型训练完成后，会对模型进行评估，评价分类器性能的好坏和及早调整模型参数使其始终能够维持在较高的预测精度。同时，获得训练后的模型的特征重要性排名，重要的特征能够指导分类器的进一步改进与优化。

2.1.4　业务应用层

业务应用层是整个系统的顶层架构，它基于各基础层，构建具体的业务应用场景。运营商由于用户数众多，自身提供的产品和服务种类也很繁多，如何向用户进行产品营销或者活动推荐是运营商十分重视的一个问题。以往的做法是进行全网的推荐或营销，比如将新产品的相关信息发布于运营商的门户网站、向全体用户推送产品描述的短信、向用户拨打电话进行人工推荐等。运营商向客户推荐产品的渠道多种多样，而且都掌握在自身手中，整个推荐行为的下游已畅通无阻，如果能将上游的产品信息做到个性化的推荐，最终得到的营销效果一定会有质的飞跃。业务应用层就是为了解决这一问题而设计的，在这一层中有各种不同的业务模块来完成不同的个性化推荐功能，本文将侧重于外呼推荐应用的研究。

外呼推荐是运营商向用户开展营销的一条重要途径。运营商的客户除上网卡客户之外，基本都具备语音服务功能，通过运营商的客户热线向用户拨打电话，直接由外呼工作人员向用户解释并推荐产品或服务。本文实验部分选取的外呼场景为4G主副卡外呼推荐，属于新产品。4G主副卡这一新产品的概念是多张副卡可以共用主卡的套餐，包括语音通话分钟数、流量以及短信等套餐包含的服务，满足单用户多卡的需求。由于并不是全网用户都会对这项新推出的业务感兴趣，因此在系统的业务应用层实现了4G主副卡外呼推荐模块，用于提高对于4G主副卡这一特定业务进行外呼营销的成功率。

2.2 系统实现

图 2 为外呼推荐系统的实现流程。将以 4G 主副卡推荐这一业务场景来还原整套系统的实现流程。系统的数据存储层包含有运营商的所有数据，对于 4G 主副卡推荐场景，只需要使用 4G 用户的相关数据。

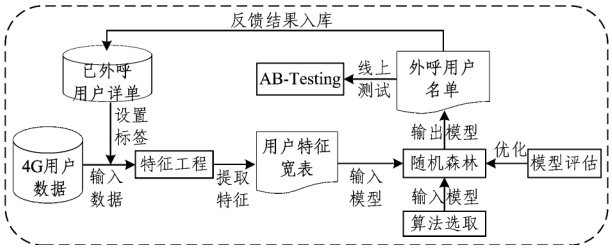

图 2 模型详细架构图

2.2.1 数据准备

使用 Impala 操纵 HDFS 上的用户数据，通过特征工程的方法决定模型的输入所包含的特征，这一步主要在数据分析层完成。借助于 Impala 的高性能数据库操作，对全网 4G 用户的数据记录进行清洗，生成中间表便于管理，使用 Spark SQL 将中间表合并成用户特征大宽表，存储于 HDFS 中，并在 Impala 中建立对应的数据表。因为我们使用的模型为随机森林分类器，该算法是有监督学习算法的，所以必须为训练数据准备标签。运营商对每次外呼营销都会有结果记录，能够判断用户是否接受了营销活动，通过外呼工作人员提供的工单数据就可以得到训练需要的标签。

2.2.2 标签

在 4G 主副卡外呼推荐模型构建过程中，需要用已知的外呼推荐活动接受和未接受的数据作为标签来训练模型，即需要提供随机森林算法训练集中每条样本是属于接受（正样本）还是未接受（负样本），标记样本分类的过程称为数据标记，标记的结果直接当作训练标签使用。

每次外呼推荐任务结束后，客服系统会接收外呼工作人员录入的工单数据，从这些工单数据中得到用户对于外呼活动的反馈行为，如表 1 所列，我们把所有未接通、接通但未接受活动的用户标记为负样本，把接通并接受活动的用户标记为正样本。由于最终模型的作用是尽可能得到会接受外呼推荐的活动的用户群体，因此我们把接通但没有明确表达接受活动的用户归为负样本，只有最后明确表达接受活动的用户归为正样本。

标签的时间粒度以月计，这符合运营商做营销活动的周期，一般为 20 天至 30 天，尽量安排在一个自然月中。将第 n 个月的标签标记为 L_n。

表 1 用户进入充值期充值时间分布

标签	行为
正样本	接通并接受活动
负样本	未接通、接通但未接受活动

2.2.3 用户特征

用户特征的构建是整个系统架构中比较关键的步骤，特征选取的好坏直接决定了模型的预测精度。在搭建整个系统之前，我们详细研究并整理了运营商所掌握的数据，在 4G 主副卡推荐这一具体系统应用上，我们侧重于对 B 域数据的使用，原因是 4G 主副卡属于一项新的业务，B 域由于是业务支持数据，能够很好地刻画用户对当前业务的使用情况，其将要推荐的新业务的相关性也较 O 域数据强。

训练模型时用到的主要用户特征大致可以分为 8 类：用户账户级数据，包含主产品标识（套餐号）、积分值、信用等级和入网时长；税额数据，包含营业税公允后收入、增值税公允后收入等；用户账单数据，包含出账费用、基本月租费、套餐月租费、功能月租费、本地通话费、省际长途费、国内漫游费、数据流量费等；用户数据流量详情，包含总上网次数、免费上网次数、收费上网次数、总流量、免费流量、免费流量占比、收费流量、上行流量、下行流量、总上网时长（秒）、总流量费用等；用户语音通话详情，包括总通话时长、总通话基本计费时长、总通话次数、本地通话次数、本地通话次数占比、本地通话基本计费时长、本地通话基本计费时长占比、长途通话基本计费时长、漫游通话基本计费时长、被叫通话总次数、主叫通话总次数等；用户短信收发详情，包括总短信条数、发送短信条数、接收短信条数、集团短信条数、其它短信条数等；用户基本信息，包括性别、年龄等；用户余额信息，包括现金余额、预存款余额、往月欠费金额等。

运营商的数据量大而全面，涵盖各个粒度，有以汇总数据形式呈现的月表（基础信息月表、账单月表、用户行为月表等），有以业务发生当天的日表记录（如用户行为日表、通话详单日表、短信详单日表等）。由于运营商用户的数据存在一定的月周期性，比如月初开账、月末出账，因此我们将各个粒度的数据通过数据层的处理，使最终的用户大宽表的粒度以月计，即每个用户每个月会产生一条汇总的特征向量，我们将生成第 n 个月的特称宽表标记为 F_n。

2.2.4 外呼推荐模型设置

经过数据分析层对存储数据进行处理之后，能够得到每个月的用户特征宽表与用户标签，将宽表数据与标签输入随机森林模型，便可以训练相应的外呼推荐模型。

在本文的实验设置中，每次的实验需要一组连续 4 个月的数据，具体参照图 3。对于训练数据集，需要特征向量与对应的标签数据，选取第 $N-1$ 月的大宽表特征数据，同时关联第 N 月的标签数据；然后将关联后的训练数据集输入随机森林分类器模型中进行模型的训练；预测及评估模型时，将第 N 月的特征数据输入模型，让模型输出预测结果，将结果与 $N+1$ 月的真实标签进行比对计算模型的精度，模型精度的评估方法会在下文介绍；最后，若模型精度符合预期要求，则上线对即时的数据进行预测，将 $N+1$ 月的用户特征数据输入给已经训练好的分类器模型，模型的输出结果即 $N+2$ 月的预测结果。

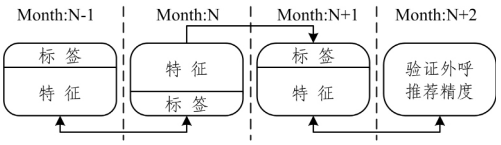

图 3 4G 主副卡推荐模型实验设置

做的 4G 主副卡外呼推荐在本质上需要获得的结果是用户对外呼产品的接收率的降序排序，是一个排序问题。我们选取排序列表中排名靠前的用户，对他们进行外呼推荐，因此，需要让外呼推荐活动的用户预测接收率越高的人排在列表最前端。最后选取排序列表的前 5%－10% 的用户，给外呼任务团队进行相应的外呼推荐。系统还会结合模型的特征

重要性排名，通过 LIBFM 算法得到能够优化模型的二次相乘特征，使模型能够自我动态调整，以适应变化的数据特征。

3 随机森林与评价指标

在模型选择上，通过对比主流的有监督机器的学习算法如逻辑回归[9-11]、决策树[11-12]、boosting 算法[13]、随机森林[14,19]、神经网络[15,16]、支持向量机[17]后，最终选择随机森林分类器作为 4G 主副卡外呼推荐系统的核心模型。由于目标属于二元分类问题，因此选择的是随机森林分类器模型。

3.1 随机森林分类器

在模型选择上，通过对比主流的有监督机器学习算法，最终选择随机森林分类器作为 4G 主副卡外呼推荐系统的核心模型。随机森林是美国学者 Breiman 提出的集成机器学习算法，是有监督的集成机器学习算法，算法会对输入的训练模型分别构建不同的决策树，构建过程中采用 Bootstrap 采样与特征的采样，通过这两个维度上的分别采样，保证每棵树的生成都尽可能不一样，每一棵树都相当于某一个领域的专家，许多棵树共同对最后的分类结果进行投票或使用 bagging 的方法产生最终的结果。由于生成树的过程是独立的，因此随机森林算法可并行化处理，同时分类精度高、训练速度快，是目前学术界与工业界研究和采用得最多的分类器算法。

设随机森林是由 K 个决策树$\{h(X,\theta_k),k=1,2,L,K\}$组成，其中$\{\theta_k,K=1,2,L,K\}$是一个随机变量序列，它由随机森林的两大随机化方法决定[4]：1) Bootstrap：从原样本 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。可以保证每次约有 37%的样本未被抽中，每个抽取的训练子集 T_k 构造一棵对应的决策树。2)特征子空间：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，通常取值为$\sqrt{M}$个特征数，M 为数据总特征个数，每次从这个子集中选择一个最优属性对当前节点中的剩余样本进行分裂。预测时，可以通过式(1)得到测试样本为正例的概率值，

$$y=\frac{1}{T_n}\sum_{t=1}^{T}f_t(x) \tag{1}$$

其中，y 是样本 x 倾向于接受推荐活动的概率值，f_t 是每棵树给出的分类结果，T_n 为决策树的集合，一共 n 有棵决策树组成了随机森林。

随机森林分类器对于 4G 主副卡推荐这一场景来说，是一个非常合适的分类器，它有很高的分类精度，对于噪声和异常值有较好的稳健性，而且有较强的泛化能力。更重要的是，随机森林不仅可以输出分类的类别结果，还可以输出属于不同类别的预测概率，这对于后续的排序来说至关重要。

3.2 评价指标

对于一个分类问题来说，常用的评价指标是查全率(Recall)、查准率(Precision)和 AUC。针对 4G 主副卡推荐这一业务场景，对评价方法稍作改进，采用排序后的 Top U 的各项指标来衡量，U 为预测的样本数。通常情况下，U 选取越大，查全率越高，查准率越低。因此，最终需要选取合适的 U 进行外呼推荐。式(2)和式(3)分别对应离网预测模型中对查全率及查准率的定义：

$$\text{Recall}=\frac{U_T}{Total_T} \tag{2}$$

$$\text{Precision}=\frac{U_T}{U} \tag{3}$$

其中，U_T 是在列表排名前 U 个用户中，真实标签为接受 4G 主副卡推荐营销活动的用户人数，即真正样本人数。$Total_T$ 是测试集中所有的真正样本的人数。除了使用了常规的 AUC 指标之外，还引入了 PR-AUC 指标[18]，这是由于训练集与测试集的正负样本比重不平衡，正样本与负样本的比值约为 1∶10，PR-AUC 在样本不平衡的数据集上能够更合理地刻画模型的预测准确性。

4 实验结果

4.1 模型精度

按照 2.2.4 小节的实验设置，选取 8 月的用户特征匹配 9 月的外呼标签训练模型。然后用 9 月的特征输入训练好的模型，得到 10 月的预测输出结果，用 TOP U 的评估方式评估模型，在进行了模型调优后，得到的模型精度评估结果如表 2 所列。

表 2 4G 外呼推荐系统精度报告

TopU	Recall	Precision	AUC	PR-AUC
10000	0.0822	0.1823		
20000	0.1493	0.1655		
30000	0.1868	0.1381	0.7912	0.3815
40000	0.2159	0.1197		
50000	0.2329	0.1033		

在排序结果的前 1 万名预测用户中，所提分类器的精度为 0.1823，也就是说，这 1 万名用户中大约有 1823 人接受了外呼推荐的营销活动。在 10 月份，总共接受外呼推荐产品的用户数为 22177 人，召回率为 0.0822。相应的 AUC 和 PR-AUC 指标分别为 0.7912 和 0.3815。由于前 1 万人的预测精度已经高出不用任何模型情况下随机外呼的转化率 0.061 的两倍以上，可以初步断定模型对于 4G 主副卡外呼推荐这一项业务是有效的。

4.2 随机森林与其它主流分类器对比

在最终决定使用随机森林算法之前，把随机森林算法与主流的机器学习分类算法进行了横向比较，比较结果如图 4 所示。

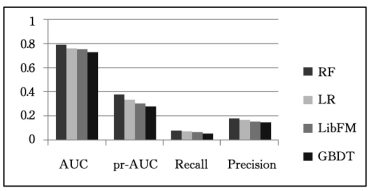

图 4 随机森林算法与其它分类器算法的对比

将随机森林算法、逻辑回归算法、矩阵分解算法与迭代决策树算法进行了比较，比较指标有 AUC、pr-AUC、Recall 和 Precision。值得注意的是，这里的 Recall 和 Precision 取的是按输出结果从大到小排序后 Top U(U=10000)用户的预测结果的统计值。从结果中可以看到，4 项指标基本保持一致的趋势，随机森林的表现最为突出，4 项对比指标分别较 GBDT 提升了 8.67%(AUC)，37.38%(pr-AUC)，43.46%(Recall)，20.63%(Precision)。在算法运行速度上，由于随机森林算法自然支持并行，并行的策略是在每个节点上构建少量的树，这样能够保证在总共构建的树的棵数不变的情况下精

度完全没有损失。实际系统被部署在 20 个节点上，每个节点运行 100 棵分类与回归树(CART)，结果采用 bagging 的方法求均值。采用这样配置下的随机森林算法，整个训练与预测过程总时间约为 5 分钟。

4.3 模型上线表现

在运营商现场部署了 4G 主副卡外呼推荐模型，并将模型的预测结果直接交付给外呼工作人员，经过近一个月的外呼作业，在月底获取了外呼工单，通过分析工单结果验证模型上线的实际效果。在实际外呼过程中发现用户会因为各种原因无法被叫，主要原因有欠费停机、关机等，将接通率也一并考虑在内，测试结果如表 3 所列，其中实际外呼组为 9 月份未经模型优化后的情况，模型外呼组为 10 月份使用模型输出结果的前 13.5 万用户的外呼情况。

表 3 模型外呼组与实际外呼组对比结果

外呼	用户数	接通数	转化数	接通率	转化率
模型	135000	79413	9332	0.5882	0.0691
实际	206527	107208	10099	0.5191	0.0489

外呼组分为实际外呼组和模型外呼组，实际外呼组是没有使用模型进行外呼推荐的结果，模型外呼组是使用了模型后的外呼结果。从结果中可以看出，在实际外呼组人数比模型外呼组人数多 52.98%的情况下，模型组的接通率比实际组的接通率提高了 13.32%，同时，转化率提高了 41.31%。使用模型后的外呼效果有了大幅提升。

图 5 为模型外呼组与实际外呼组在接通率上的对比结果。图表的横坐标代表外呼人数，纵坐标代表接通人数占比，从结果中可以发现，在 14 万人的外呼数量级上模型外呼组的接通率明显优于实际外呼组的接通率。可见随机森林模型在训练时同时考虑了用户对于运营商服务的使用情况，若用户经常不使用运营商的服务，或者经常欠费停机，自然在训练集中就会倾向于表现为负样本；相反，经常使用运营商服务的用户，接通率会显著提升。随着人数段的加大，接通率呈现递减趋势，下降趋势缓和，到达 14 万人的时候，模型外呼组的接通率依然比实际外呼组的接通率高 6.7%。

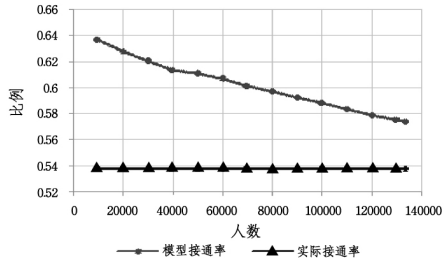

图 5 模型外呼组与实际外呼组的接通率对比

与接通率相比，更为重要的一项指标是转化率，转化率是衡量外呼推荐活动的最终效果好坏的重要指标，因为只有用户成功接受并且同意使用新产品或新服务，这一次的营销活动才会给运营商带来收益。图 6 展示的就是模型外呼组与实际外呼组的转化率对比情况。从图中可以看出，自然转化率大约在 6%，而模型组排名最靠前的 1 万个用户的转化率高达 17.7%，随着人数段的增加，模型组的转化率也同样呈现下降趋势，下降较接通率略微陡峭，更快地逼近实际转化率。

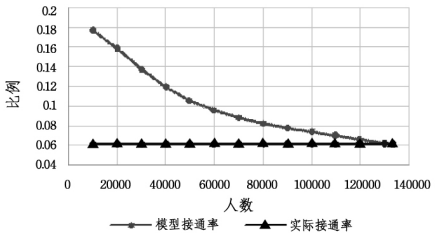

图 6 模型外呼组与实际外呼组的转化率对比

4.4 重要特征分析

随机森林模型中，每一个分支节点上都会选取一个特征进行分裂，选择的依据常用的是计算基尼不纯度，当分裂过后两个叶子节点的基尼不纯度的和与分裂节点基尼不纯度的差最大时，选择该特征及分裂点进行分裂。因此分裂点上保留了自己分裂时的特征，通过计算所有树的分裂特征的基尼不纯度差值，可以得出所有特征的重要性排名。用此方法计算出 4G 主副卡的前 5 个重要特征分别是预存款余额、套餐月租、每月出账费用、积分值和总使用流量。得出的这些重要特征与我们的先验知识相吻合。4G 主副卡业务如本文最开始描述，是一项对于主卡的分享业务，业务的需求用户往往是那些原本套餐内的语音通话、短信、流量等就很多的用户。因为套餐包含服务多，所以区别于包含服务少的套餐，包含服务少的套餐的内容基本上集中在上述几项特征值中。

为了验证我们的观点，分别对这 5 项重要的特征作了分析，限于篇幅，只给出预存款余额与套餐月租的分析结果。对于预存款余额，如图 7 所示，从用户接不接受 4G 主副卡业务的推荐在用户预存款余额的分布可以看出，接受用户在预存款余额为 1000 元到 2000 元区间及 2000 元到 5000 元区间内的占比较高，分别为 21.6%和 41.1%。预存款较高的用户能够在短中期时间内支付得起较高月租费的套餐，因此结论合理。

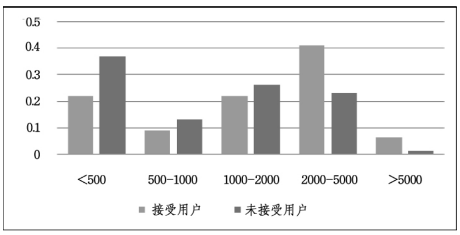

图 7 预存款余额特征分析

排名第二的特征是套餐月租，如图 8 的分析结果所示，在 100 元到 200 元、200 元到 300 元、300 元到 400 元区间段内的接受用户比重较高，分别为 37.2%、25.9%和 26.4%。与之前的结果一致，高消费用户对于 4G 主副卡的业务比较欢迎。

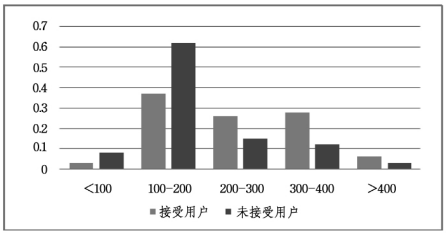

图 8 预存款余额特征分析

结束语 本文介绍了一种使用随机森林分类器模型实现外呼推荐系统应用的方法，从电信运营商的实际需求入手，引入 4G 主副卡外呼推荐的业务案例，逐层介绍了电信运营商

的数据挖掘与机器学习系统的平台构成。在系统顶层的业务应用层部署了 4G 主副卡外呼推荐应用，从模型训练、结果评估到对比测试，详细介绍了 4G 主副卡外呼推荐模块的实现。最后，用查准率、查全率、AUC、PR-AUC 等常用评估指标刻画了模型的精确度，分析了模型的重要特征的实际分布情况，总结了模型对于外呼推荐的适用性。

参 考 文 献

[1] Sarwar B, Karypis G, Konstan J, et al. Application of dimensionality reduction in recommender system-a case study[R]. Minnesota Univ Minneapolis Dept of Computer Science, 2000

[2] Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems[J]. Computer, 2009 (8): 30-37

[3] Huang Yi-qing, Zhu Fang-zhou, Yuan Ming-xuan, et al. Telco churn prediction with big data[M]. SIGMOD, 2015

[4] Yuan Ming-xuan, Deng Ke, Zeng Jia, et al. OceanST: A distributed analytic system for large-scale spatiotemporal mobile broadband data[C]// VLDB (Demo). 2014: 1561-1564

[5] Page L, Brin S, Motwani R, et al. The PageRank Citation Ranking: Bringing Order to the Web[R]. Stanford InfoLab, 1999

[6] Zhu X, Ghahramani Z. Learning from labeled andunlabeled data with label propagation[R]. Technical Report CMU-CALD-02-107, Carnegie MellonUniversity, 2002

[7] Zeng J, Cheung W K, Liu J. Learning topic modelsby belief propagation[J]. IEEE Trans. Pattern Anal. Mach. Intell., 2013, 35 (5): 1121-1134

[8] Rendle S. Scaling factorization machines to relational data[C]// PVLDB. 2013: 337-348

[9] Neslin S, Gupta S, Kamakura W A, et al. Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models[J]. Social Science Electronic Publishing, 2006, 43(2): 204-211

[10] Hadden J, Tiwari A, Roy R, et al. Computer assisted customer churn management: State-of-the-art and future trends[J]. Computers & Operations Research, 2007, 34(10): 2902-2917

[11] Lima E. Domain knowledge integration in data mining using decision tables: case studies in churn prediction[J]. Journal of the Operational Research Society, 2009, 60(8): 1096-1106(11)

[12] Verbeke W, Martens D, Mues C, et al. Building comprehensible customer churn prediction models with advanced rule induction techniques. [J]. Expert Systems with Applications, 2011, 38 (3): 2354-2364

[13] Jinbo S, Xiu L, Wenhuang L. The Application ofAdaBoost in Customer Churn Prediction[C]// 2007 International Conference on Service Systems and Service Management. IEEE, 2007: 1-6

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

[15] Datta P, Masand B R, Mani D, et al. Automated Cellular Modeling and Prediction on a Large Scale[J]. Artificial Intelligence Review, 2000, 14(6): 485-502

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

[17] Burez J, Van den Poel D. Handling class imbalance in customer churn prediction[J]. Dirk Van den Poel, 2008, 36(3): 4626-4636

[18] Davis J, Goadrich M. The Relationship Between Precision-Recall and ROC Curves[C]// ICML '06: Proceedings of the 23rd International Conference on Machine Learning. 2006

[19] 方匡南，吴见彬，朱建平，等. 随机森林方法研究综述[J]. 统计与信息论坛，2011，26(3)：32-38

(上接第 553 页)

从图中可以看出，Anti-windup 算法能进行很好的控制。

结束语　本文通过在 MATLAB 环境下建立四旋翼飞行器的非线性模型，分别设计了常规 PID 控制器模型和 Anti-windup PID 控制器模型，并在软件平台下对这两种控制器进行仿真实验。仿真结果表明，Anti-windup PID 控制器在动态性能及稳定性能上均优于常规 PID 控制器，Anti-windup PID 控制器能更好地实现对四旋翼飞行器的控制。

参 考 文 献

[1] Bouabdallah S. PIDvs LQ Control Techniques Applied to an Indoor Micro Quadrotor[C]// IEEE/RSJ International Conference on Intelligent Robots and Systems. 2004, 3: 2451-2456

[2] Argentim L M, Rezende W C, et al. PID, LQR and LQR-PID on a Quadcopter Platform[J]. Informatics, Electronics & Vision (ICIEV), 2013: 1-6

[3] Salih A L, Moghavvemi M, et al. Modelling and PID Controller Design for a Quadrotor Unmanned Air Vehicle[J]. Automation Quality and Testing Robotics (AQTR), 2010: 1-5

[4] Erginer B, Altug E. Modeling and PD Control of a Quadrotor VTOL Vehicle[J]. Intelligent Vehicles Symposium, 2007: 894-899

[5] 胡锦添，舒怀林. 基于 Adams 与 Matlab 的四旋翼飞行器控制仿真[J]. 自动化与信息工程，2012(5)：25-28

[6] Gheorghita D, Vintu I, et al. Quadcopter Control System[J]. Control and Computing(ICSTCC), 2015: 421-426

[7] 吴建德. 基于频域辨识的微小型无人直升机的建模与控制研究[D]. 杭州：浙江大学，2007

[8] Ly D, et al. Modeling and Control of Quadrotor MAV Using Vision based Measurement[J]. International Forum on Strategic Technology, 2010, 33(4): 70-75

[9] 刘金坤. 先进 PID 控制 MATLAB 仿真[M]. 北京：电子工业出版社，2004

[10] 王树刚. 四旋翼直升机控制问题研究[D]. 哈尔滨工业大学，2006

[11] 聂博文. 微小型四旋翼无人直升机建模及控制方法研究[D]. 长沙：国防科学技术大学，2006

[12] 吴中华，贾秋玲. 四旋翼几种控制方法研究[J]. 现代电子技术，2013(15)：88-90

[13] 李钟慎，郭辉，张磊，等. PID 控制系统抗饱和方法的对比研究[J]. 信息技术与信息化，2013(4)：68-71

[14] Bresciani T. Mdelling, Identification and Control of a Quadrotor Helicopter[D]. Department of Automatic Control Lund university, 2008

[15] 杨帆. 微型四旋翼飞行器的建模与控制系统研究[D]. 太原：太原理工大学，2014