



电信行业基于种子用户群扩展技术的定向营销研究与应用

张旭, 刘洋, 胡磊, 赵晓东, 张海滨

(华为软件技术有限公司, 江苏 南京 210012)

摘要: 在数据业务爆发式增长的时代, 掌握海量用户数据的电信运营商占有价值优势。通过大数据技术对数据价值进行洞察识别和探索已经成为运营商发展增值业务的必要手段。对于传统的营销方式, 各省市公司通过一些业务指标筛选目标客户或者单独建模的方式进行粗放营销, 数据获取周期长, 营销接触率不高。针对上述痛点, 提出在小样本数据情况下构建相似群扩展服务(Lookalike)。该服务结合电信行业的数据及业务特点, 集成了度量学习、深度学习等方法, 有效支撑电信运营商实现精准营销。Lookalike 服务减少了人工参与, 自动实现营销目标客户的输出, 大大提高了营销活动的工作效率及成功率, 在多次实际电信项目中得到验证 TTM (time to market) 取得了由月到日的提升, 营销成功率明显提高。

关键词: 智能数据集成; 种子相似用户群自动扩展技术; 华为运营商大数据平台

中图分类号: TP311.13

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2018029

Application and research of marketing using automatic similar users extension technology in telecom industry

ZHANG Xu, LIU Yang, HU Lei, ZHAO Xiaodong, ZHANG Haibin

Huawei Software Technologies Co., Ltd., Nanjing 210012, China

Abstract: With the explosive development of data service, TSP (telecommunication service provider) has taken advantages of owing massive user data. It is a necessity for TSP to develop value added business using big data techniques which explores the data value. The traditional way of marketing utilizes rules or supervised classification method meets the challenge of low success rate and long data acquisition period. Therefore, a new service named Lookalike was proposed. The service supports precision marketing by integrating metric learning and deep learning efficiently based on telecom operators' data characteristics. The Lookalike service decreases the artificial participation and enhances the efficiency and success rate of marketing activity. The enhancement of TTM (time to market) has been improved in many real programs and the success rate of marketing has increased absolutely.

Key words: smart data integration, automatic similar users extension technology, Huawei FusionInsight platform

1 引言

随着社会信息化的飞速发展, 人们已经迈入

大数据时代。Gartner 统计数据认为从 2011 年开始, 大数据已经处于期望快速提高的阶段。各行各业产生的数据呈现出 3V 的特性, 即数量大、数

收稿日期: 2017-12-01; 修回日期: 2017-12-20

据种类多,数据响应速度要求高。与互联网相比,电信运营商积累了庞大的数据资源,每天都有数以亿计的用户在运营商管道中留下电信行为数据^[1]。工业和信息化部统计数据显示,2016年我国三大运营商电话用户数达15.3亿户,移动用户数为13.2亿户^[2],固话用户数为2.07亿户^[3]。数据源种类丰富多样,运营商的数据包括通话数据、位置数据、用户属性数据、用户上网数据、手机消费数据、终端数据等,数据类型涵盖结构化的用户基本信息数据、半结构化的用户访问日志数据、非结构化的流媒体数据等。为了应对信息化运营商的转型,各大国际运营商在大数据产品方向大力投入,如西班牙电信(Telefónica)率先成立名为“动态洞察”的大数据业务部门;Verizon成立了基于大数据分析的精准营销部门(Precision Marketing Division);新西兰电信(Telecom New Zealand)成立独立的大数据子公司Qrious^[4]。

精准营销一直是电信运营商业务最重要的部分,传统的营销方式是业务专家根据规则筛选目标用户,随着机器学习的兴起,也将监督的分类方法融入模型中。机器学习的方法对数据分析人员的技术背景及能力有较高要求,数据预处理及

参数调整都相对复杂。此外,在运营商的很多实际营销场景中,由于没有积累营销反馈数据,难以准确定义正负样本。客户希望根据现网存量用户,分析他们的共性且独特的特征,然后找出最相似的目标用户群,作为最可能转化的潜在用户。只利用现网存量的种子用户进行潜在目标用户的挖掘,可能面临种子用户数量较少、种子用户与未知用户数量相差悬殊的问题。可见,传统的基于正负样本的分类模型不再适用,需寻找或开发适用于无法准确定义正负样本且能较好解决种子样本所占比例较低的问题。因此本文提出在小样本数据情况下构建相似群扩展服务(Lookalike),可以更好地发挥数据价值,最大限度地提高数据分析人员的工作效率,模型构建从2个月缩短至2周,降低人员专业要求,无需数据分析专家,只需业务人员参与,快速支撑融合营销目标客户群选取。

2 Lookalike 技术方案及关键技术

2.1 大数据平台预测分析服务框架

Lookalike 是华为运营商大数据平台预测分析服务产品中的一部分,预测分析服务整体方案如图1所示。

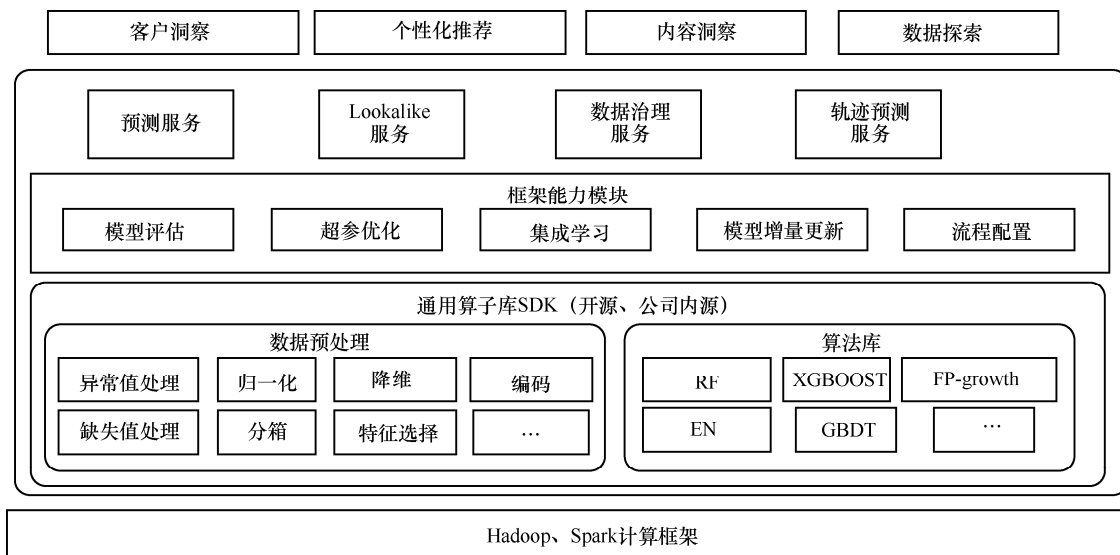


图1 预测分析服务框架



华为运营商大数据平台的预测分析服务具备以下特点。

- 机器学习自动化：提供超参优化、特征选择、集成学习、各种评估指标等自动化能力，提供开箱即用的分析服务。
- 分析服务插件化：支持第三方开发服务插件，插件接口灵活，提供服务自动发布、计算平台自动对接等工程能力，开发者仅需关注核心分析服务逻辑。
- 扩展算法库：集成 MLlib、TensorFlow、Sparkling Water 等业界先进算法平台，提供丰富的算法库，降低构建分析服务难度，提升服务准确性。

2.2 Lookalike 分析服务

Lookalike 服务的定位是客户只需要提供少量的种子用户，自动输出潜在目标客户群，可以支撑潜在客户挖掘、用户拉新、精准营销、销售提升以及在线广告等大数据分析应用。Lookalike 具有“COURSE”（control, on-demand, unified, rule-free, scalable, efficient）六大亮点。

- control（扩展规模可选）：支持扩展群规模灵活选择。
- on-demand（按需触发）：根据业务需求随时触发。
- unified（通用场景）：通用 Lookalike 服务，

适用各种精准营销场景。

- rule-free（不依赖规则）：无需维护庞大规则库。
- scalable（可伸缩）：最小支持 100 种子，最大可支持百万级种子。
- efficient（有效性）：百万级种子条件下可在小时内完成服务响应。

Lookalike 服务总体流程如图 2 所示，主要分为不平衡数据集处理、数据预处理、核心算法处理和模型选择、模型应用 4 个部分，下面对 Lookalike 中的几个关键技术进行介绍。

2.2.1 数据预处理

每个电信运营商都具有丰富的数据资源，传统的方式对于不同的业务场景都需要业务专家对特征维度进行多次筛选，多域的数据特征近千维，给业务专家带来较大挑战。为了减轻业务专家工作量，Lookalike 服务融入了特征降维的模块。大多数业务场景中，用户没有提供相关的负样本，有监督的特征选择方法会带入计算误差，因此优先选用 RBM（restricted Boltzmann machine，受限玻尔兹曼机）降维方法。

RBM 是由 Hinton 和 Sejnowski 于 1986 年提出的一种生成式随机神经网络（generative stochastic neural network），该网络由一些可见单元（visible unit，对应可见变量，亦即数据样本）和

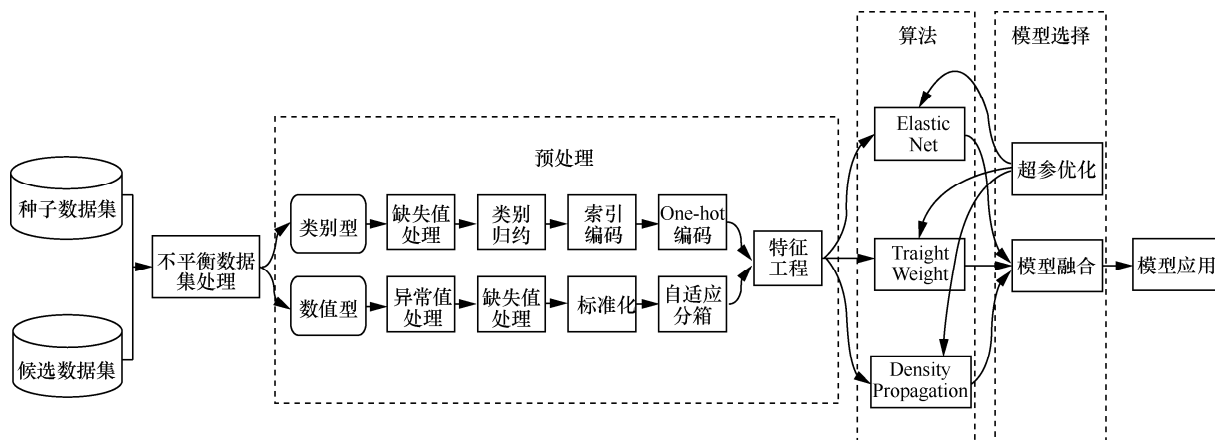


图 2 Lookalike 分析服务流程

一些隐藏单元(hidden unit, 对应隐藏变量)构成。整个网络是一个二部图, 只有可见单元和隐藏单元之间才会存在边, 可见单元之间以及隐藏单元之间都不会有边连接^[5]。

RBM的核心在于定义一个合理的能量函数, 通过能量定义状态的概率, 然后通过极大似然估计求解模型的参数, 最后得到每一个输入样本的隐式表达, 通过输出层可以获取希望得到的降维后的特征。RBM构建示意如图3所示。

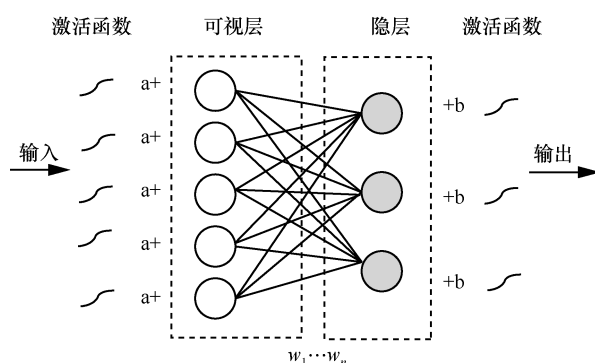


图3 RBM构建示意

2.2.2 DensityPropagation 算法

基于密度传播的相似用户扩展方案, 可以降低噪声影响。密度传播算法分为两个步骤, 首先通过KNN计算用户之间相似性, 为种子和候选集分配不同的初始能量, 通过种子和候选集之间的递更新各自能量值。模型训练示意如图4所示。

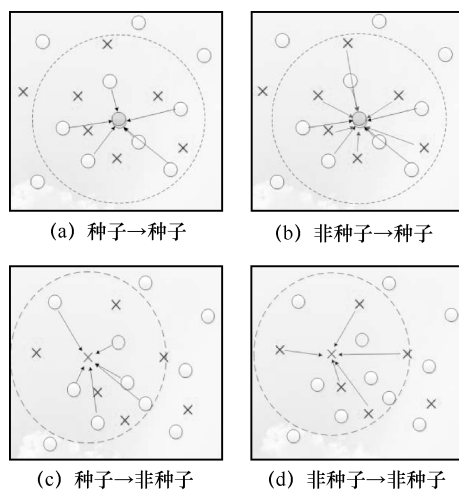


图4 DensityPropagation 算法示意

DensityPropagation 模型训练流程具体如下。

(1) 训练样本包含种子集 S 和候选集 U , S_i 和 U_j 分别代表种子和非种子。

(2) 设置所有种子的初始能量 E_s , 设置所有非种子的初始能量 E_u 。

(3) 对于每个种子, 选取它的 K 个近邻, 每个近邻把自身初始能量的 $1/K$ 传递给这个种子, 能量加上种子的初始能量就是种子的新能量 ES_{si} , 得到种子的最终能量。

(4) 对于每个非种子, 选取它的 K 个近邻, 近邻中的种子用户把更新后的能量的 EF_{si} 的 $1/K$ 传递给这个候选用户, 更新候选用户能量 ES_{uj} 。

(5) 对于每个非种子, 选取它的 K 个近邻, 近邻中的非种子用户把能量 ES_{uj} 的 $1/K$ 传递给候选用户, 得到新的能量为 EU_{uj} 。

(6) 非种子的最终能量为 $EF_{uj} = (E_s / \tanh(E_s \cdot 9)) \cdot \tanh(E_u + ES_{uj} + EU_{uj})$ 。

如果能量传递后种子的能量升高, 说明种子周围的种子数量较多, 该种子为正样本的权重越高; 如果候选用户的能量升高, 说明该候选用户周围的种子用户数量较多, 由此可初步推断该候选用户的营销成功概率较高。经过多次迭代, 输出候选集中每个用户营销成功率列表并按能量值进行排序。客户根据业务需求确定阈值, 该阈值以上的用户即潜在目标用户。算法采用KNN计算用户之间相似性, 为了更好地描述不同特征, 将度量学习融入算法中^[6]。

电信领域数据具有数据量大、维度高、稀疏性强等特点, 其固有数据分布空间可能不是欧氏空间。采用简单的欧氏距离来度量样本点之间的相似性, 往往无法保证得到令人满意的结果。通过训练数据学习出的度量距离对比欧氏距离, 能够更好地描述样本的本征相似性。因此通过KNN算法计算用户之间相似性时, Lookalike 算法采用度量学习方法, 在最小化同类样本对的马氏距离平方和的同时, 约束不同



类样本对的马氏距离,使其大于给定阈值。使在新的度量空间中相似对更加紧凑,而非相似对更加分离^[6],如图5所示。

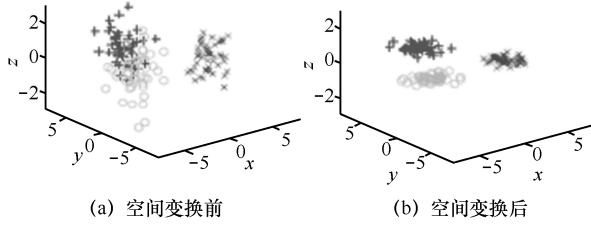


图5 度量空间变换前后示意

度量学习目标函数如式(1)所示:

$$\begin{aligned} \min_{\mathbf{A}} \sum_{(x_i, x_j) \in S} \|x_i - x_j\|_{\mathbf{A}}^2 \\ \text{s.t.} \sum_{(x_p, x_q) \in D} \|x_p - x_q\|_{\mathbf{A}} \geq 1, \mathbf{A} \succeq 0 \end{aligned} \quad (1)$$

其中, x_i, x_j 为样本对, $\mathbf{A}$ 称为度量矩阵。

$$\|x_i - x_j\|_{\mathbf{A}} = \sqrt{(x_i - x_j)^T \mathbf{A} (x_i - x_j)} \quad (2)$$

通过式(1)计算得到度量矩阵 $\mathbf{A}$ 后,优化后的样本间距离表达式见式(2)。

2.2.3 TraightWeight 算法

TraightWeight 算法是基于 FP-growth 获取关联规则分类的思想。由于没有明确的负样本,该服务中采用关联规则得到种子关键特征进一步计算相似性的方法来输出潜在客户群。与 C4.5、SVM 算法相比,基于频繁项的分类方法近几年在很多的的研究中都得到了较好的结果^[7]。FP-growth 与 Apriori 算法相比,只需要两次扫描数据库,采用分治的策略有效降低搜索开销。TraightWeight 算法的流程如图6所示,具体介绍如下。

(1) 首先对数值型数据进行分箱,类别型的

属性做索引。

(2) 出于性能考虑,根据种子用户抽样比例参数的设置,对候选集用户进行抽样。

(3) 基于 FP-growth,采用关联规则分类的方法,挖掘频繁项规则,当产生的规则的数量小于设定的最小规则生成数量时,规则的最小支持度阈值减半。直至从种子集和全量集中产生的规则的数量都满足要求或达到迭代次数限制。

(4) 根据频繁项在种子和全量集合的不同,获取关键关联规则的权重。

(5) 累加某未知用户包含的所有规则的权重,将其作为该未知用户的得分。得分越高,成为目标用户的概率越大。

2.2.4 WeightedElasticNet 算法

利用正类样本和无标记样本进行学习 (learning from positive and unlabeled example) 称为 PU 或 LPU 学习^[8],是一种半监督的二元分类模型,通过标注过的正样本和大量未标注的样本训练出一个二元分类器。由于正样本集合 P 中难免会有一些噪声以及正负样本的数据倾斜,采用 WeightedElasticNet 算法^[9],可以解决样本不平衡的问题,WeightedElasticNet 算法的目标函数如式(3)所示:

$$\begin{aligned} \arg \min_{\beta} \frac{1}{2} \|y - \sum_{j=1}^p x_j \beta_j\|_2^2 + n \lambda_1 \sum_{j=1}^p w_j |\beta_j| + \\ \frac{n}{2} \lambda_2 \sum_{j=1}^p |w_j \beta_j|^2 \end{aligned} \quad (3)$$

2.2.5 模型融合

为了提高推荐结果的准确度,系统通常会应用多个子模型(多个算法)产生多个推荐候选集。然后通过一定的模型融合策略,将多个候选集融

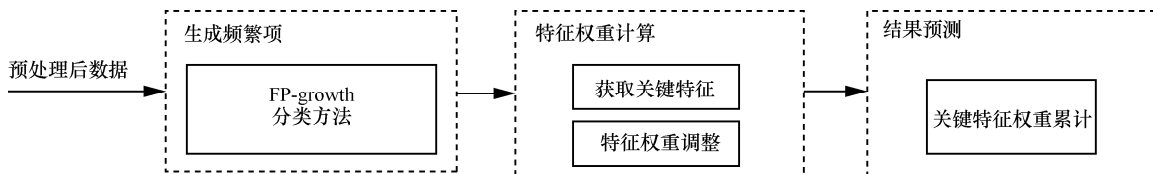


图6 TraightWeight 示意

合成最终的结果展示给用户,以达到“1+1>2”的效果^[10]。Lookalike 融合模型如图 7 所示。

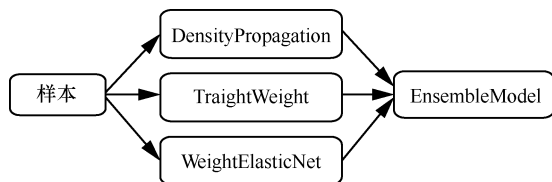


图 7 Lookalike 融合模型

本文采用加权平均法对模型进行集成:

$$H(x) = \sum_{i=1}^T w_i h_i(x) \quad (4)$$

其中, w_i 是个体学习器 h_i 的权重, 通常要求 $w_i \geq 0$, $\sum_{i=1}^T w_i = 1$, 不同算法的权重是从训练集中学习得到的。

2.2.6 超参优化

Lookalike 服务中涉及许多可以调整的配置参数, 有些参数很难预置在软件开发中, 需要根据实际业务数据进行调整, 为了减轻人工配置的工作量, 本文在框架中置入超参优化模块, 可以识别算法的关键超参及其有效范围, 克服超参优化的维度灾难和非凸性。本文选择贝叶斯优化算法对 Lookalike 模型中涉及的参数进行优化。贝叶斯优化是一种联合优化超参数的强力工具, 并且最近也越来越流行。它能自动调节超参数以提升产品质量和人类生产力。贝叶斯优化是一种近似逼近的方法, 采用在函数方程不知的情况下根据已有的采样点预估函数最大值的思想^[11]。贝叶斯优化算法的伪代码如下。

算法 1 贝叶斯优化算法

```

for  $n=1,2,\dots,\text{do}$ 
    select new  $x_{n+1}$  by optimizing acquisition
    function  $\alpha x_{n+1} = \arg \max_x \alpha(x; \mathcal{D}_n)$ 

    query objective function to obtain  $y_{n+1}$ 
    augment data  $\mathcal{D}_{n+1} = \{\mathcal{D}_n, (x_{n+1}, y_{n+1})\}$ 

    update statistical model
  
```

end for

3 应用效果

Lookalike 服务经过多个项目的验证, 在提高营销成功率的同时还可以缩短 TTM, 获取含数据与分析时长由月到日的转变。

3.1 X 省机顶盒加装项目

X 省电信运营商推广第二部机顶盒加装业务, 匿名特征 198 维度, 将采用随机森林算法与 Lookalike 算法做比较。其中随机森林参数设置: 树的棵数为 100 棵, 树深 10 层。实验环境为 2 台华为 RH2288 (单台 24 核 300 GB 内存)、Hadoop2.7.2 和 Spark1.5.1。其实验数据和实验结果分别见表 1 和表 2。

表 1 X 省机顶盒加装项目实验数据

	种子用户数	候选用户数	特征维度
训练集	5 949	66 332	198
测试集	2 429	28 428	198

表 2 X 省机顶盒加装项目实验结果

指标	Lookalike	分类模型 (随机森林)
AUC	0.782	0.635

从上述结果可以看出, 监督的分类模型自动将非种子集合 (unlabeled data) 认为是负样本代入效果误差, Lookalike 的 AUC 与随机森林相比提升 23%。

3.2 Y 省流量套餐营销

Y 省运营商为了培养全省的客户的流量使用习惯, 通过短信进行“任我享”流量包营销。采用 Lookalike 服务输出营销潜在客户名单与运营商实际的营销结果做 AB 对比测试, 此应用案例采用 Top N 的查全率和查准率进行比较。

用户的接入特征有基本特征、账单列表、订购套餐、流量特征等, 都做了匿名处理, 分别为 F1, F1, ..., F128。实验环境为 5 台 E9000 (单台 6 核, 96 GB 内存)、Hadoop2.7.2 和 Spark1.5.1。



其实验数据见表 3。

表 3 Y 省流量套餐营销实验数据

种子用户数	候选用户数	特征维度
42 000	19 080 000	128

全量目标用户是用户对全量 1 908 万用户营销的结果, 查全率比较结果如图 8 所示。

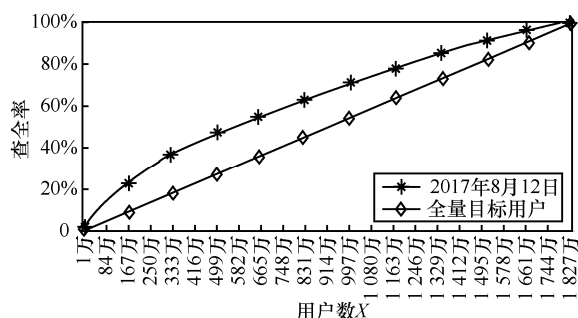


图 8 查全率比较

从查全率来看, 全量目标用户截止至 12 日共订购用户 149 457 个, 其中 23 048 个在 Lookalike 结果得分靠前的 100 万用户中, 查全率为 15.4%。

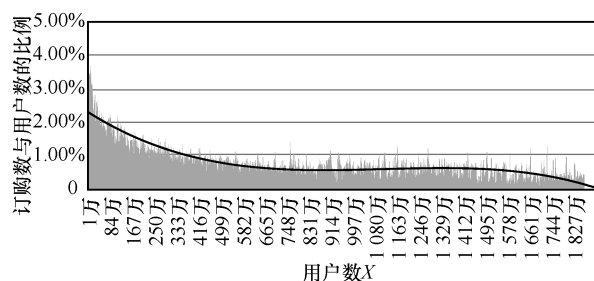


图 9 命中率变化趋势

命中率为根据 Lookalike 选定 X 个用户, 该 X 个用户中订购的个数与 X 的比例, 命中率变化趋势如图 9 所示。从命中率来看, 截止至 12 日的全量目标命中率为 0.78%, 算法圈定的得分靠前的 100 万用户命中率为 2.30%, 前 10 万用户命中率为 3.64%, 命中率有显著的提高。

Lookalike 服务产品对目标客户有较强的识别能力, 若需达到 2.3 万订购目标, 使用 Lookalike 只需的营销为 100 万用户, 传统方式需营销 300 万用户, 将为运营商节约 2/3 的营销资源。

3.3 H 省 4G 终端营销

H 省运营商 4G 客户渗透率已达到 60%, 但仍有 1 700 万客户未使用 4G 终端, 以低端客户为主, 遇到发展瓶颈, 各市公司通过 DOU/ARPU 等指标筛选目标客户或单独建模的方式进行粗放营销, 数据获取周期长, 营销接触率低, 营销成功率不高。因此该公司尝试用 Lookalike 同时开展营销活动, 实验环境 2 台华为 RH2288 (单台 12 核, 256 GB 内存), 实验数据和实验结果见表 4 和表 5。

表 4 H 省 4G 终端营销实验数据

种子用户数	候选用户数	特征维度
47 000	2 200 000	196

从表 5 来看, 接通率提升了 23.04%, 外呼成功率提升了 12.77%, 从原来的月建模时间缩短为 2 h 输出潜在客户名单, 同时保证了较高的营销成功率。

4 结束语

针对电信行业营销场景中负样本难以精确定义和搜集的特点, 本文提出了一种基于小样本种子用户群扩展技术来快速、准确识别潜在客户, 该技术将客户发现问题抽象成无监督或半监督机器学习任务进行建模并对预测结果进行有效融合。经多个局点和多个实际业务场景验证发现, 与传统方法相比, 种子用户群扩展技术提高了目标客户发现的准确率, 降低了一线服务或业务实施人员的使用门槛, 还可适用于电信行业

表 5 H 省 4G 终端营销实验结果

外呼项目	渠道	外呼总量	外呼接通量	外呼成功量	接通率	外呼成功率
4G 终端营销 (2G、3G 迁转 4G)	Lookalike	2 318	1 554	601	67.04%	38.67%
	人工筛选	500	173	41	34.60%	23.70%

多个业务营销场景,在帮助电信运营商大大提升营销活动 E2E 转化率的同时也有效节约了营销支撑成本。

在业务部署过程中,也发现了一些还未得到有效解决的技术问题,包括:用户建模问题,即如何科学选择有效特征或属性来描述用户,才能保证后续种子用户群算法的覆盖率和精准度;持续优化问题,即如何及时搜集营销反馈数据,并根据反馈持续优化机器学习模型,形成算法和营销活动的闭环。

驱使营销活动成功的因素较多,算法和模型只是其中一部分,以最小的代价获取最大受益,是营销活动永恒的追求。在后续的研究和优化过程中,或可考虑将营销成本和代价等因素构建到模型当中,打造完全基于数据驱动的智能营销系统。

参考文献:

- [1] 刘春, 邹海峰, 向勇. 大数据环境下电信数据服务能力开放研究[J]. 电信科学, 2014, 30(3): 156-161.
LIU C, ZOU H F, XIANG Y, et al. Research on telecom data service open ability under the environment of big data [J]. Telecommunications Science, 2014, 30(3): 156-161.
- [2] 廖建新. 大数据技术的应用现状与展望[J]. 电信科学, 2015, 31(7): 7-18.
LIAO J X. Big data technology: current applications and prospects [J]. Telecommunications Science, 2015, 31(7): 7-18.
- [3] 黄勇军, 冯明, 丁圣勇, 等. 电信运营商大数据发展策略探讨[J]. 电信科学, 2013, 29(3): 6-11.
HUANG Y J, FENG M, DING S Y, et al. Big data development strategy for telecom operators[J]. Telecommunications Science, 2013, 29(3): 6-11.
- [4] 梁灿, 何阳, 韩涵, 等. 数据中心联盟大数据发展促进委员会电信工作组[S]. 电信大数据应用白皮书. 2017.
LIANG C, HE Y, HAN H, et al. Data center alliance major data development promotion committee telecom working group[S]. White Book of Telecommunication Big Data Application. 2017.
- [5] 张春霞, 姬楠楠, 王冠伟. 受限波尔兹曼机[J]. 工程数学学报, 2015(2): 159-173.
ZHANG C X, JI N N, WANG G W. Restricted Boltzmann machines[J]. Chinese Journal of Engineering Mathematics, 2015(2): 159-173.

- [6] XING E P, NG A Y, JORDAN M I, et al. Distance metric learning, with application to clustering with side-information[C]// International Conference on Neural Information Processing Systems, Dec 9-14, 2002, Vancouver, British Columbia, Canada. Cambridge: MIT Press, 2002: 521-528.
- [7] WAND J, KARYPIS G. Harmony: efficiently mining the best rules for classification[C]//M The Fifth SIAM International Conference on Data Mining, Apr 21-23, 2005, Newport Beach, California, USA. Piscataway: IEEE Press, 2005: 205-216.
- [8] LEE W S, LIU B. Learning with positive and unlabeled examples using weighted logistic regression[C]// Machine Learning, Proceedings of the Twentieth International Conference, Aug 21-24, 2003, Washington DC, USA. Palo Alto: Aaai Press, 2003: 448-455.
- [9] HONG D, ZHANG F. Weighted elastic net model for mass spectrometry imaging, processing[J]. Mathematical Modelling of Natural Phenomena, 2010, 5(3): 115-133.
- [10] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.
ZHOU Z H. Machine learning[M]. Beijing: Tsinghua University Press, 2016.
- [11] SHAHRIARI B, SWERSKY K, WAND Z, et al. Taking the human out of the loop: a review of bayesian optimization[J]. Proceedings of the IEEE, 2015, 104(1): 148-175.

作者简介



张旭(1976-), 男, 博士, 华为软件技术有限公司大数据首席分析建模专家, 主要研究方向为数据挖掘与分析、用户偏好分析、用户建模、用户相似性分析与个性化推荐。

刘洋(1989-), 女, 华为软件技术有限公司大数据科学实验室高级工程师, 主要研究方向为数据挖掘与分析、个性化推荐、用户建模。

胡磊(1981-), 男, 华为软件技术有限公司大数据科学实验室高级工程师, 主要研究方向为数据挖掘和机器学习。

赵晓东(1981-), 男, 华为软件技术有限公司大数据科学实验室高级工程师, 主要研究方向为数据挖掘、数据集成与治理。

张海滨(1983-), 男, 博士, 现就职于华为软件技术有限公司大数据科学实验室, 主要研究方向为数据挖掘与机器学习、时间序列分析、服务计算等。