



安徽联通企业级大数据平台构建及应用实践

吴涛, 刘韬, 王斌

(中国联合网络通信有限公司安徽分公司, 安徽 合肥 230069)

摘要: 完整地总结了中国联通网络通信有限公司安徽分公司建设大数据面临的挑战和机遇以及通过快速构建 B/O 域融合的智慧运营平台累积的跨域数据多维度、多层次融合治理经验。此外, 安徽联通还对用户离网预测进行了深入研究, 通过大数据平台构建离网预测模型, 实现离网预测模型查准率 90%、查全率 9% 的突破, 并通过提升用户感知等应用深入推广大数据在各领域的融合应用, 对内提升了运营效率, 实现了 NOC(network operations center, 网络运维中心) 向 SOC(service operations center, 业务运营中心) 的转型。

关键词: 运营商大数据; B/O 域数据融合; 离网预测大数据应用; NOC 向 SOC 的转型

中图分类号: TP319

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2018036

Construction and application of Anhui Unicom enterprise big data platform

WU Tao, LIU Tao, WANG Bin

Anhui Branch of China United Network Communications Co., Ltd., Hefei 230069, China

Abstract: The challenges and opportunities faced by Anhui Unicom of big data construction, and the experience of Anhui Unicom rapidly constructing B+O convergent smart operation platform which realized multi-dimensional and multi-level fusion management of cross-domain platform were summarized. In addition, the customer off-grid prediction was also deeply studied by Anhui Unicom. Achieving the off-grid prediction model through the big data platform which made the precision ratio get 90% and the recall ration get 9%. On the other hand, the big data platform has been popularized to other departments by integrated applications, such as, user awareness enhancement etc. Internal operation efficiency was improved and NOC (network operations center) was transformed to SOC (service operations center).

Key words: carrier big data, B+O convergent scenario, big data applications of off-grid prediction, transform from NOC to SOC

1 引言

2015 年中国政府提出“互联网+行动计划”, 互联网化已经成为各个传统行业升级和创新商业模式最重要的议题。“互联网+”时代的到来, 不

仅在改善和提升用户体验上发挥作用, 也将诞生海量的数据资产。根据爱立信的报告, 2020 年大数据的产生量将接近目前的 10 倍, 大数据的激增以及各类大数据的融合使得大数据的大规模商业化和落地成为可能^[1]。

收稿日期: 2017-12-10; 修回日期: 2018-01-10



2 运营商发展大数据的优势

电信运营商发展大数据具有天然的优势，主要体现为数据资源的丰富性、完整性和连续性。

- 丰富性：主要指运营商拥有的数据涉及范围广，不但涉及财务收入、业务发展量等结构化数据，也会涉及图片、文本、音频、视频等非结构化数据，数据量级已从 TB（1 012 GB）发展至 PB 乃至 ZB。
- 完整性：主要指运营商拥有的数据涵盖全业务、全用户和全渠道，数据信息完整。
- 连续性：主要指运营商拥有的数据记录周期长，数据延续性好，覆盖用户从入网到离网前的全生命周期^[2]。

3 安徽联通建设大数据的现状与挑战

大数据给电信运营商带来机遇的同时，也给运营商相关技术带来极大挑战^[3]。

中国联合网络通信有限公司安徽分公司（以下简称安徽联通）拥有庞大的用户资源，随着实名制工作的开展，数据会更加翔实，并通过分析移动互联网生成的多样化数据，了解人们生活的方方面面，具备非常高的价值纯度。安徽联通还具备丰富的 IDC（integrated data center，集成数据中心）资源，能够保证海量数据拥有足够的存储空间，从而保证数据分析的时间连续性，以把握数据趋势。但是，安徽联通传统的建设模式基本采用封闭、独立、各系统自成一体的建设方式，导致大量的数据分布在各个底层应用系统中，没有统一的数据标准，各自私有的数据结构，且存在大量的冗余建设，如何有效地实现数据的各域、各系统的数据统一采集、治理、共享，将是安徽联通面临的巨大挑战，归纳起来，有以下 5 点。

（1）接口复杂覆盖面广

智慧运营平台的数据来源于安徽联通各业务系统，涉及计费、网管、核心网、无线等跨领域业务。

不同系统间数据结构、接口千差万别，这就对大数据平台的采集能力提出了很高的要求。如何保质保量地将不同系统的数据准确采集过来，是智慧运营平台数据整合首先要解决的问题。例如，实时数据通过 SDTP（serial data transport protocol，串行数据传输协议）进行采集，B 域的数据通过 SFTP（secure file transfer protocol，安全文件传输协议）进行采集，流日志数据通过 syslog 方式进行采集；各系统的数据接口格式也存在千差万别，例如 4G MR（mobile router，移动路由器）数据是 XML（extensible markup language，可扩展标记语言）报文格式，CBSS 数据文件格式是固定长度记录格式，ECS 是特殊字符串分割方式，大数据平台需要支持各种差异化接口、格式，实现数据的异构。

（2）数据标准不一致

不同业务系统对同一个事物的理解存在差异，同样的事物可能具有不同的命名规则、不同的度量单位。如果不进行有效的数据治理，统一数据标准，大数据只能是一堆杂乱数据的集合，实现统一建模、统一数据标准，是智慧运营平台面临的最大挑战。例如，在 CBSS 系统中流量单位是比特，但是，在 BSS 系统中流量单位为 KB，两者相差一个数量级。如果不进行标准统一，就会导致数据失真，在大数据挖掘的过程中错误会指数级别放大。

（3）异厂商技术壁垒

大数据平台数据来源于各应用厂商，各应用厂商系统接口通常采用私有协议，具有特殊的数据结构，且不对外提供数据解析能力，导致大数据平台采集数据后，需要协调各厂商提供数据解析支撑，并做好各厂商数据的异构适配。因此，大幅度地提升了大数据平台的数据采集难度，降低数据采集效率。例如，2G/3G 用户 MR 数据，中兴通讯、华为、诺基亚西门子 3 个厂商数据格式各不一样，任何两个厂商之间数据都无法相互解析，大数据平台进行数据采集后，需要根据业务理解对各厂商数据进行定制化的转化，实现数

据的统一建模,对外提供统一的数据格式。

(4) 烟囱发展各自为政

由于历史发展,各业务系统各自为政,形成一个个数据烟囱,烟囱长短不一、数据质量参差不齐,各业务分析支撑只能基于单系统数据展开,缺少一个全量数据的整合平台,影响数据价值的深度挖掘,智慧运营平台将致力于抹平烟囱,实现全量数据的统一共享。

(5) 开放及安全防护

大数据的重要特点就是数据的开放和互通,随着大数据的广泛应用,数据共享和开放势在必行,但大数据平台数据包含了大量的用户隐私数据,如用户行为偏好、位置轨迹、消费数据等隐私数据,如何在侵犯个人隐私、保障数据安全的情况下,做好数据开放实现互利共赢,是智慧运营平台需要面对的一大挑战。智慧运营平台通过网络隔离、数据加密、去隐私化等方式保障数据的安全^[4]。

4 安徽联通技术+管理的应对措施

面对建设大数据遇到的挑战,安徽联通迎难而上,自上而下明确了整体的工作思路:在组织机制保障下,借助智慧运营平台相关建设,实现大数据采集、建模、开放及安全管理。

(1) 组织保障

安徽联通调整组织架构,成立了智慧运营专(兼)职运营团队,负责智慧运营平台建设及日常运营、需求受理等,并基于大数据挖掘的应用场景使用。

(2) 系统建设

安徽联通智慧运营平台是业内领先的企业级大数据平台,包括分布式存储计算平台、大数据应该管理平台、流处理平台三大资源池,为大数据采集和整合奠定了坚实基础。

(3) 数据整合

建立全新的四层融合数据模型,通过数据采集、数据建模、数据挖掘、数据服务和数据开放

等过程,实现了大数据的处理、加工、应用和价值变现。

(4) 数据安全

在流程方面建立严格的数据管理流程,数据的接入和获取需要逐层审批,并实现不同账号的分权分域;在平台方面实现数据隔离,做到“进不来、看不见、拿不走”的三“不”要求;在对外合作方面做到数据脱敏,保障用户信息的安全。

(5) 数据管理

依托智慧运营平台的资产管理平台,建立一系列领先的大数据管理能力,实现对智慧运营平台的数据以及数据生产的全生命周期进行的全面管控。

5 安徽联通智慧运营平台的构建

安徽联通于2016年底成功构建基于华为FusionInsight解决方案的B域、O域数据融合的企业级大数据平台——智慧运营平台,目的是对内实现智慧运营、精准营销,对外实现价值变现。

智慧运营平台基于“ $M+1+N$ ”的理念构建,即 N 个可扩展的数据源、1个企业级大数据融合平台、基于平台实现的多(M)个对内、对外应用,如图1所示。遵循“ $M+1+N$ ”的建设理念,安徽联通在搭建智慧运营平台架构时首要考虑的是对融合数据的统一采集、统一存储、跨域数据的共享以及跨域数据的精品应用^[5]。

5.1 融合数据统一采集

安徽联通现网有多个烟囱状系统,这些系统的数据相互独立、类型互不相同,如BSS、ECS、Gn上网日志、CS信令、基站小区基本信息、网络感知系统、客服系统、号线资源系统等,通过梳理涉及的接口共197个。为了实现对任何系统各类数据的统一采集,安徽联通智慧运营平台摒弃传统ETL(无法支持非结构化文本,不支持流式数据处理),而采用云化ETL(BDI)技术。BDI

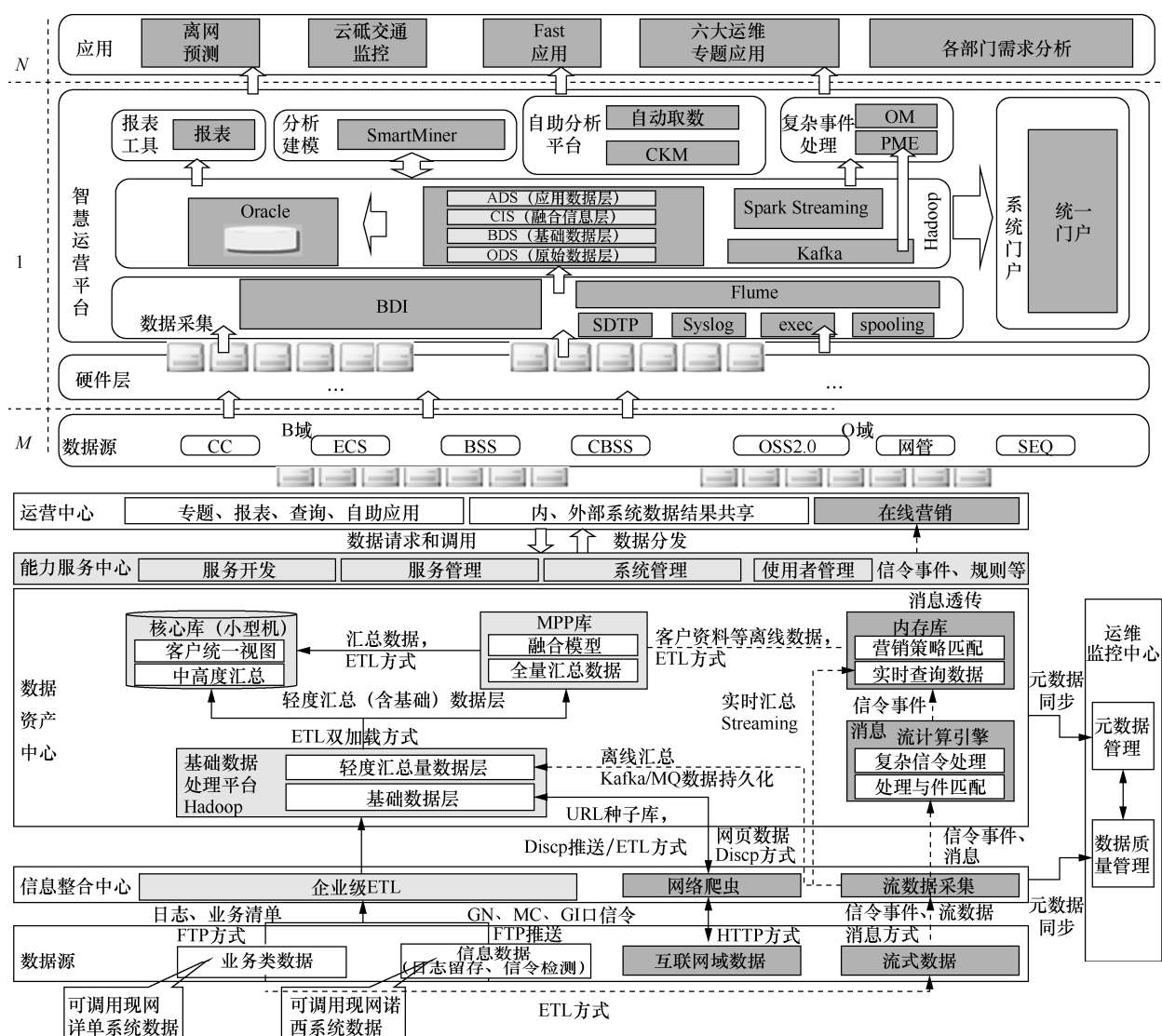


图1 安徽联通智慧运营平台“M+1+N”架构

实现融合数据统一采集 & BDI 功能视图如图 2 所示。

云化 ETL (BDI) 技术的数据采集功能支持采集多种类型数据, 并将数据源不同的字段类型进行统一。如, OGG 采集 B 域增量数据、FTP/SFTP 对文本文件进行抽取、Flume/Kafka 对 O 域数据实时采集、RDB 和 MPPDB 对数据库直接采集等。数据转换功能对异构数据源的数据进行清洗、转换、统一标准, 包括过滤、分组、查找、去重、路由、合并、拆分、排序、转换、联合、统计、加密等, 最终实现维度命

名统一、数据数值统一等。数据加载/分发功能将融合数据加载分发到不同的数据源, 支持异构数据源并行加载、加载前过滤、加载前路由以及不落地加载等。

5.2 融合数据统一存储/统一处理

安徽联通智慧运营平台采用 FusionInsight Hadoop 平台作为融合数据统一存储和处理中心, 如图 3 所示, 采用分布式共享存储, 不需要额外的磁阵和双机软件, 实现现网 B/O/M 三域一份数据的统一集中化存储, 确保 B 域全部、O 域全部和 M 域核心价值数据的统一与一致性存储, 并通

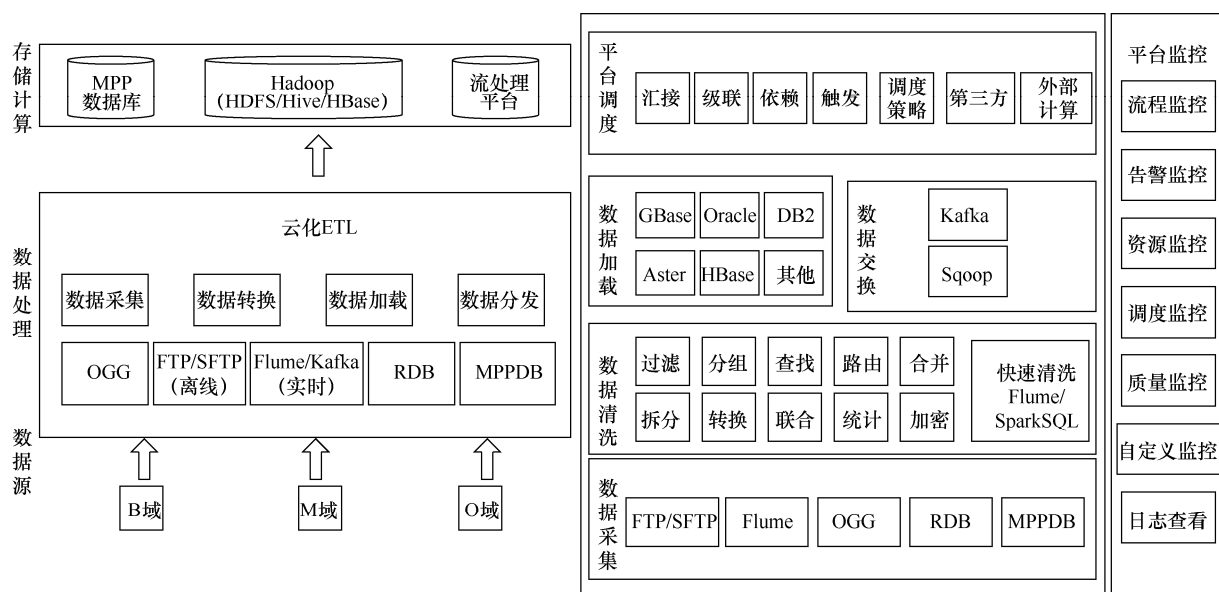


图2 BDI实现融合数据统一采集 & BDI功能视图

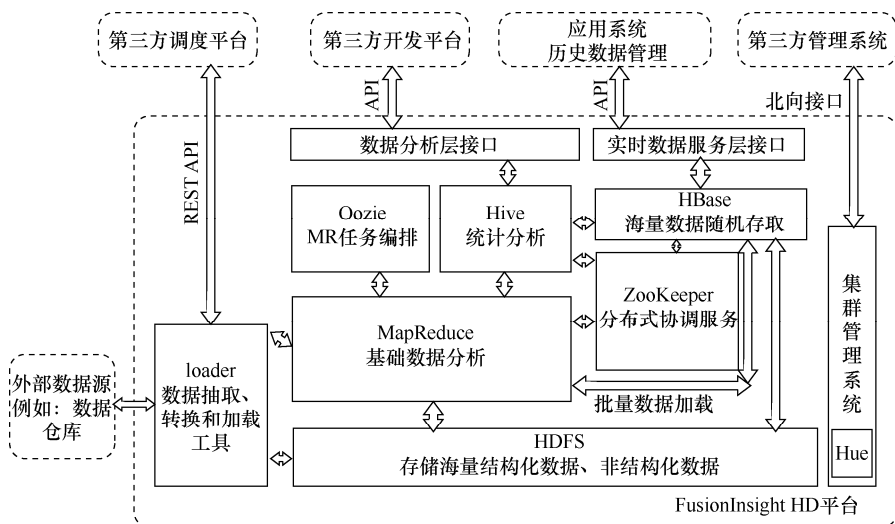


图3 FusionInsight Hadoop 逻辑架构

过云化 ETL 平台向数据仓库、实时查询库和流处理平台进行统一分发^[6]。

5.3 跨域数据共享

安徽联通智慧运营平台通过强大的多租户能力（在一套服务器上运行某个应用实例，它为多个租户提供服务，租户之间不感知）进行权限控制和资源隔离，解决部门间大数据平台资源共享与数据共享困难，实现跨域数据的共享。多租户逻辑架构如图4所示。

5.4 跨域数据精品应用^[7]

安徽联通智慧运营平台整合 B/O 域多种数据源，在如图5所示的六大实体数据（参与人、服务、资源、账务、营销、企业管理）和事件数据基础上，通过表1的九大主题域统一建模，以“数据→信息→知识→智慧”为核心将跨域数据关联的价值提炼。目前，采用的融合建模方法主要有维度建模（基于静态信息，一般是“人”和“物”）和事件+规则建模（基于动态信息，即时间轴模型，

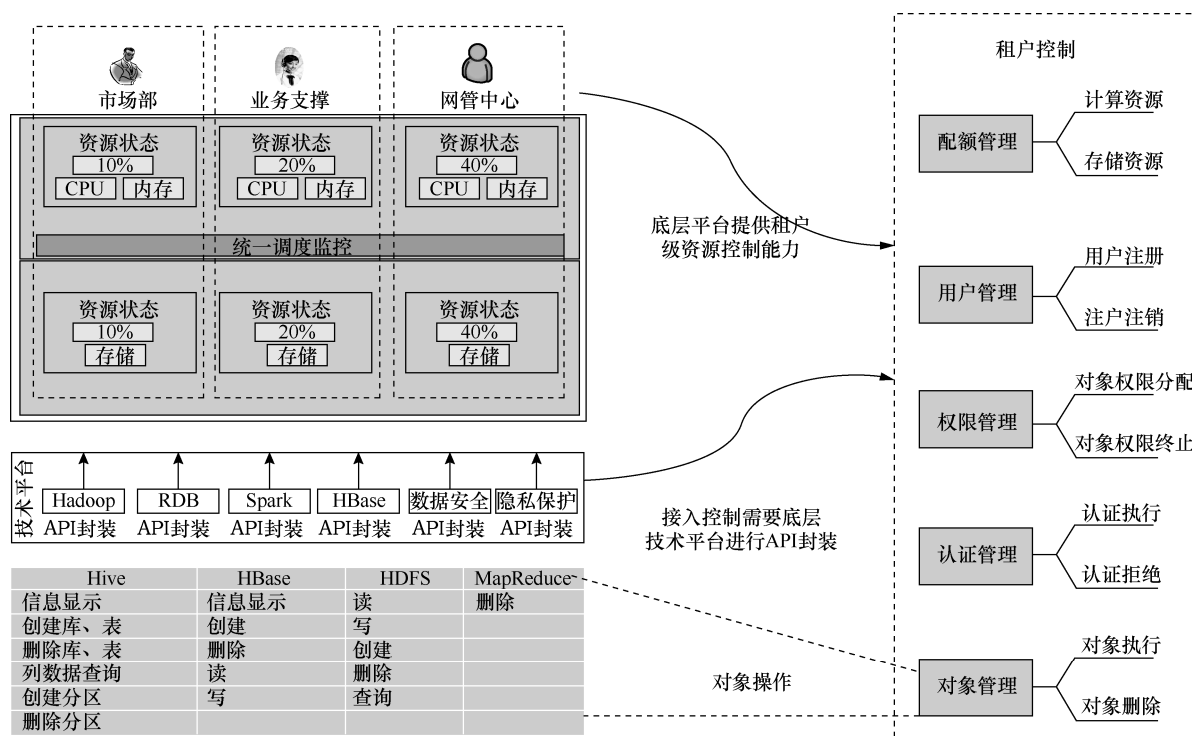


图4 多租户逻辑架构

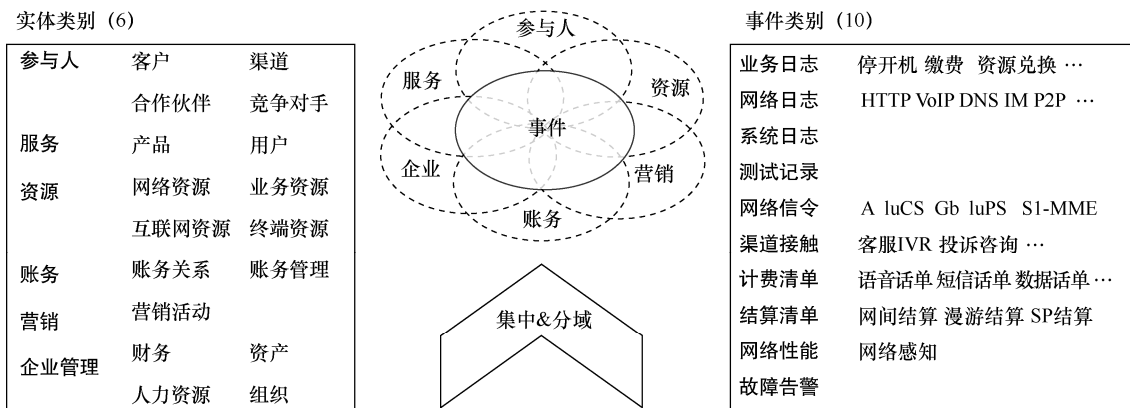


图5 六大实体数据和事件数据

表1 九大主题域

主题域	主题子域
个人客户统一视图	个人客户基本信息、个人客户业务使用、个人客户产品订购、个人客户收入
集团客户统一视图	集团客户基本信息、集团客户业务使用、集团客户产品使用、集团客户收入
家庭客户统一视图	家庭客户基本信息、家庭客户业务使用、家庭客户产品订购、家庭客户收入
竞争对手统一视图	竞争对手用户基本信息、竞争对手业务量、竞争对手交往圈
产品统一视图	产品基本信息、产品使用、产品销售
用户统一视图	用户基本信息、用户业务使用、用户渠道接触、用户位置、用户产品订购、用户交往圈、用户账务、用户终端使用、用户互联网使用
终端统一视图	终端基本信息、终端使用、终端销售
渠道统一视图	网络基本信息、网络使用、网络质量
营销信息统一视图	营销基本信息、营销效果

如离散模型、步进模型、连续模型和恒定模型等), 已经在现网实现的跨域数据应用主要有离网预测、云砥交通监控、2G 退网、用户感知等精品应用开发等。

6 安徽联通智慧运营平台主要技术和能力

大数据具备 4V 特点: volume (大量)、variety (多样)、velocity (高速)、value (价值), 第一, 数据体量巨大, 从 TB 级别跃升为 PB 级别; 第二, 数据结果类型繁多; 第三, 数据处理速度快, 可以从各种类型数据中快速获得高价信息; 第四, 合理利用数据并进行正确的分析, 将会获取高价值回报。

传统数据仓库在 PB 级数据处理上已经力不从心, 在大数据领域中, Hadoop 目前已经是大数据平台中应用效率最高的技术, 特别是针对文本、社交媒体订阅及视频等非结构化数据, 支撑多种并行处理架构, 如 MapReduce、Spark、Storm; 关注 NoSQL 数据库, 解决传统关系型数据库引擎瓶颈; 基于内存的并行实时分析, 使实时流式数据分析成为可能。

安徽联通智慧运营平台采用华为公司 FusionInside-Hadoop + Universe 解决方案实现, 该技术架构具备 Hadoop 所有的特性功能, 并在 Hadoop 的基础制定专业的应用解决方案, 使智慧运营平台成为一个安全、开放、便捷的大数据平

台, 主要具备以下功能特点。

6.1 多样化的数据采集

智慧运营平台支持对表、文件、消息等多种数据的实时增量数据采集 (使用 Flume、消息队列、Ogg 等技术) 和批量数据分布式采集等能力 (Sqoop、FTP VOER HDFS), 比基于传统 ETL 的采集性能有数量级的提升, 采集能力在业界处于领先地位。另外, 还支持与 DB2、Oracle、Hadoop (Hive)、Hadoop (Impala)、Hadoop (HBase)、Hadoop (Spark)、MPP (Aster)、MPP (GBase)、Sqlfire、GemfireXD 以及 IBM Streams 共 11 类数据库技术平台的对接, 真正地实现了跨平台开发和数据管理。

6.2 统一建模、多域数据融合

智慧运营平台将数据模型框架分为原始数据层 (original data store, ODS)、基础数据层 (basic data store, BDS)、融合数据层 (convergence information store, CIS) 和应用数据层 (application data store, ADS) 4 个层次。原始数据层数据结构与源系统完全一致, 基础数据模型、融合数据模型和应用层是平台重点建设的数据模型。数据模型架构如图 6 所示。

在智慧运营平台总体设计中, 遵循高层级可以从低层级获取数据、低层级不允许从高层级获取数据的整体原则, 见表 2。

表 2 数据模型框架

	DWD	整合层	统一视图层
目标	标准化、原生态的企业级细节数据中心	定义数据模型, 清晰数据关系, 通过横向数据整合, 减少系统冗余度	面向业务应用的抽象数据模型, 经过简单加工快速支持上层应用
原则	面向主题, 完成结构的转换、数据标准化, 编码转换	面向主题, 主题域内的整合与轻度汇总	面向应用, 提炼共性需求, 进行广泛主题域的汇总
模型设计方法	表结构与原系统保持一致, 但做命名的规范化、标准化	星型模型设计, 反规范化的设计, 降低系统冗余度	星型、雪花模型设计, 反规范化的设计, 适当冗余
重点	主题定义 框架设计 标准化 减少或控制冗余	基础数据定义 派生数据设计 数据关系管理	在于对业务需求的抽象 考虑生成的性能 考虑查询的性能
用途	便捷查询, 可以回答任务问题	便捷数据组织, 尽量避免数据跨域访问	支持上层应用 支持最终使用者进行即席查询

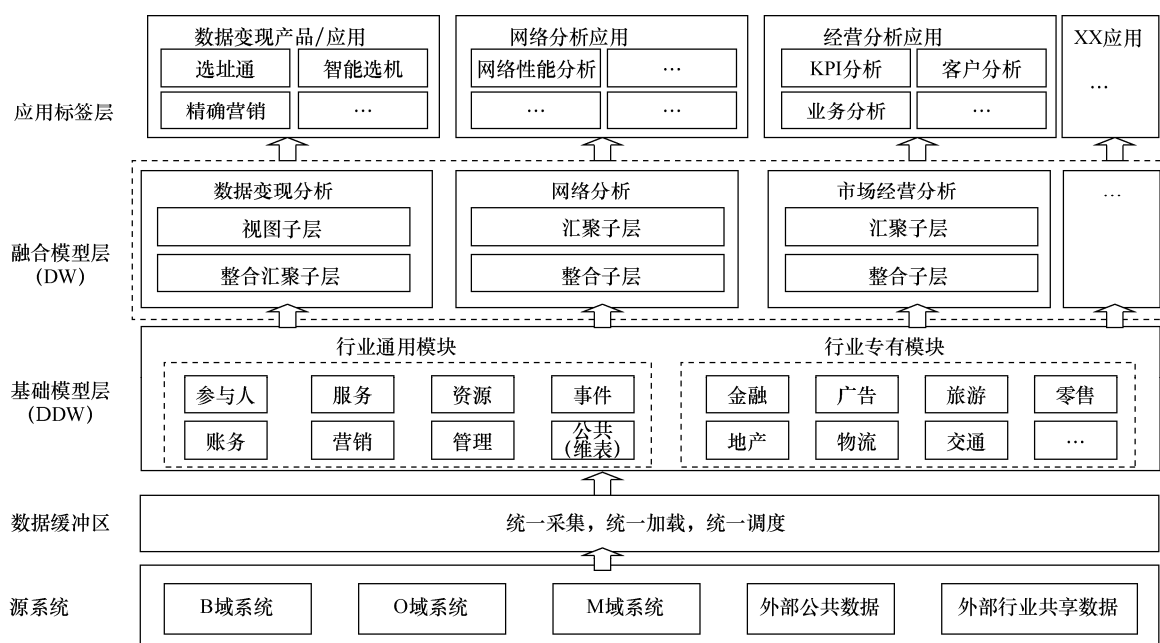


图6 数据模型框架

智慧运营平台在数据融合方面，目前实现B域、O域数据融合，共采集八大系统300多个接口数据，日存储处理B域数据500GB，O域数据10TB，数据源见表3。

表3 数据源范围

数据域	系统	数据内容
B域	BSS	用户信息、消费信息、计费话单、投诉、终端信息、渠道、营销活动等数据
	CBSS	
	ECS	
	客服	
O域	SEQ	A/IUCS、Iups、S1-MME、GN、GB、S1-U、MR、基站等数据
	博瑞德	
	永鼎	
	综合网管	

在数据准确性方面，智慧运营平台在建模阶段开始介入，制定严格数据校验规则及数据异常处理规则，确保模型加载数据的准确性。针对跨域同类数据，安徽联通独创单用户流程数据准确性比较方案，取得平台与网元及KPI对比差异小于1%，单日流量差异小于0.2%，当天存在流量差异用户比例小于30%的成果。

6.3 支持多框架，实现统一调度

智慧运营平台内置多种计算框架

（MapReduce、Spark、Storm），能够高效地处理海量非结构化、半结构化数据。同时，满足批处理、实时处理等多种计算场景需求；可支持个性化、灵活多样的调度策略编排，实现多种并行处理框架任务（如MapReduce、Spark）的调度、关系型数据库存储过程、shell脚本等调度。

6.4 集成挖掘算法，快速实现建模

智慧运营平台基于分布式计算架构，整合业界先进的挖掘算法，实现基于Hadoop的并行计算改造，充分利用大数据平台强大的计算能力及丰富的数据资源，快速实现原始数据的二次挖掘，提升数据价值。目前，已经集成的建模算法包含朴素贝叶斯、稀疏线性矩阵、决策树、随机森林、逻辑回归、K-means、社交网络推荐、影响力传播、协调过滤、线性回归等一系列的挖掘算法。整个建模过程采用图形化操作界面，大大降低了数学建模门槛，并支持系统自动建模能力。

6.5 图形化开发，降低大数据应用门槛

智慧运营平台通过一整套的图形化的开发和维护工具，支持图形化拖拽式开发，免代码编写，降低大数据应用的开发难度，业务需求响应周期

由周级别降低到小时级别,效率成倍提升。

6.6 基于大数据的网络维挽工具,实现 NOC 向 SOC 转型^[8]

6.6.1 重点区域 KQI/KPI 实时监控能力

根据业务需求,快速制定对指定区域的 KQI/KPI 监控任务,对指定区域进行实时 KQI/KPI 监控,实现 5 min 粒度监控自定义区域 KQI/KPI 状态并上报告警,并可以通过地图直观呈现自定义区域中的异常区域,对网络中关键指标进行实时监控,制定保障方案。

6.6.2 移动宽带网络可视化能力

基于 MBB 流量分析,实现现网流量的多维分析统计(网元、协议、APN、用户、接入网、网站、位置、终端等),并提供多维实时的详细报表,实现移动宽带网络可视化。

6.6.3 端到端故障分析能力

打通 UE (user equipment, 用户设备) 到 SP (signaling point, 信令点) 链路所有网元,实现全网元特定业务的质量分析,发现业务质量问题及根因,实现端到端故障定界,并可通过历史数据的分析,发现潜在问题和未来趋势。

6.6.4 开放的定制化能力

智慧运营平台采用开放式架构,基于智慧运营平台可以实现应用的百花齐放,目前已经实现 CSFB (circuit switched fallback, 电路域交换回落) 专题、Web 质差小区专题、高铁专题、2G/3G/4G 专题、OTT 用户行分析、校园用户分析六大专题应用。

- CSFB 专题: 分析对指定时间段内、指定区域的 CSFB 业务指标进行监控,快速发现并上报问题,对呼损类问题,使用呼损分析方式形象展示出问题发生次数和所在位置,帮助运营商问题定界。
- Web 质差小区专题: 结合 TCP 管道指标,识别影响用户感知的无线质差小区。
- 高铁专题: 对高铁用户进行精准识别,针对性地进行问题处理。

- 2G/3G/4G 专题: 分析 2G/3G/4G 用户分布情况,高流量 2G 小区下用户终端类型、用户业务量情况,为 2G 退网提供数据支撑。
- OTT 用户分析专题: 挖掘高价值用户行为,为市场营销提供数据支撑。
- 校园分析专题: 分析校园用户感知情况、结合终端分布及用户行为分析为市场进行校园营销提供数据支撑。

7 安徽联通智慧运营平台用户离网预测大数据的应用实践

7.1 用户离网定义

用户离网是指用户因为某种原因与运营商解除服务合同行为,即用户停止消费目前正在使用的电信产品或服务,选择其他运营商的产品或服务,或者选择该运营商的其他替代性电信产品或服务。

用户的离网表现主要分为两大类: 竞争性用户离网和非竞争性用户离网。竞争性用户离网主要是由于其他运营商通过主动方式向用户提供更优惠的资费、服务导致用户的离网; 非竞争性离网是由于运营商自身的产品替代、电信政策调测或用户自身行为对通信产品使用的调整造成的离网。

7.2 用户离网因素

根据用户的离网意愿可分为主动离网和被动离网,被动离网又分为无意离网和蓄意离网,从现网实际数据分析,蓄意离网占整个离网用户的 80% 以上,这部分用户主要受到企业产品或服务的技术原因、经济原因等影响离网,是本次大数据应用研究预测的对象。

7.3 离网预测应用原理

7.3.1 离网预测算法

安徽联通离网模型采用随机森林算法,该算法是一个包含多个决策树的分类器。每个决策树的形成采用了随机方法,各决策树之间无关联、彼此独立,所有决策树训练都是采用同样的参数,但是每棵树的训练集不同。随机森林对输入的数



据要进行和列的随机采样, 并保证采样过程的随机性, 所以不容易出现 over-fitting, 且该算法实现相对简单、运行速度快, 容易基于 Hadoop 实现并行计算等特点, 所以本次离网预测课题, 采用随机森林算法进行建模^[9-16]。

7.3.2 离网预测模型

本离网预测模型采用隔月预测未来 3 月离网方式进行建模打标, 目的在于提前发现准离网用户,

并拉大维系挽留动作实施窗口, 该预测方式建模更具备实际使用价值, 建模打标方式如图 7 所示。

本离网课题与其他课题项目在数据选择上有明显区别, 其他课题数据均来自 BSS 系统数据, 本课题数据结合 CBSS、ECS、客服、O 域用户网络感知数据。在数据周期上, 本课题采用数据周期较长, 数据周期大于 6 个月, 通过用户消费趋势变化数据, 更直观反映用户的离网情况, 具体指标参数见表 4。

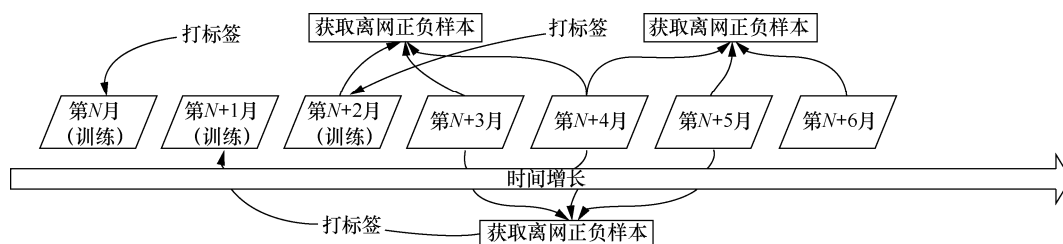


图 7 离网打标原理

表 4 离网用户关键数据属性字段

建模字段	建模字段	建模字段	建模字段
用户状态	套餐内语音时长	近 3 月 ARPU 均值	闲时通话次数占比
手机号码	上行点对点短信条数	当月欠费金额	工作日通话次数占比
用户 ID	上行网内点对点短信条数	近 3 月欠费金额均值	周末通话次数占比
客户 ID	上行网间点对点短信条数	零售渠道充值次数	工作日忙时通话次数占比
省分编码	国际漫出点对点短信条数	当月充值次数	忙时流量占比
城市代码	短信计费量	社会渠道接触次数	闲时流量占比
性别	下行点对点短信条数	近 3 月充值金额均值	工作日流量占比
客户年龄	套餐内短信条数	零售渠道充值金额	周末流量占比
用户信用度	上行点对点彩信条数	近 3 个月套餐更换次数总和	工作日忙时流量占比
网龄	上行网内点对点彩信条数	近 3 月欠费次数总和	忙时平均每次通话时长
本地网外时长	点对点彩信次数	近半年通话次数标准差	闲时平均每次通话时长
长途主叫时长	下行点对点彩信条数	近半年通话时长变异系数	工作日平均每次通话时长
漫游主叫时长	总流量	月收入环比	周末平均每次通话时长
本地被叫时长	套餐内流量	欠费金额环比	近 6 个月通话时长趋势
漫游被叫时长	服务类短信号码种类	欠费次数环比	近 6 个月工作日忙时通话次数趋势
网间通话时长	服务类短信条数	通话时长环比	近 6 个月工作日通话次数趋势
忙时通话时长	拨打客服电话次数	通话次数环比	近 6 个月流量趋势
周末通话时长	本月总收入	总流量环比	近 6 个月工作日忙时流量趋势
通话次数	流量费用	上网时长环比	近 6 个月工作日流量趋势
长途语音通话次数	短信_点对点短信费	缴费金额环比	近 3 个月通话次数标准差
主叫次数	套餐价格	缴费次数环比	近 3 个月通话时长变异系数
语音通话计费时长	余额	异网客服拨打次数	近 3 个月通话时长趋势
主叫计费时长	缴费金额	忙时通话时长占比	近 3 个月工作日忙时通话次数趋势
本地网内时长	上个月 ARPU 值	闲时通话次数占比	近 3 个月工作日通话次数趋势
本地通话时长	过缴费期天数	工作日通话时长占比	近 3 个月流量趋势
长途通话时长	余额与上月账单差值	周末通话时长占比	近 3 个月工作日忙时流量趋势
漫游通话时长	社会影响力	忙时通话次数占比	近 3 个月工作日流量趋势

7.3.3 模型构建

本预测应用构建工具采用华为公司大数据 Universe 平台的 SmartMiner 工具进行建模。SmartMiner 是华为大数据平台中的一个支持图形化开发的专业的数据挖掘工具, 基于 Spark 的分布式内存处理框架, 集成了大量的数据挖掘算法, 借助大数据平台强大数据并行处理能力, 批量、快速、便捷地实现数据建模挖掘工作。

本次应用建模使用 SmartMiner 中的随机森林算法实现, 共建 200 棵随机树, 使用 SQR 采样方法, 树的最大深度 15 层, 叶子最小样本数 100 个, 最大分箱数 32。

本次建模采用两种方式: 方式一采用使用 4 个月历史数据, 次月给前一个月打标标签, 采用当前月数据预测次月离网情况; 方式二采用 7 个月历史数据, 使用隔月后未来 3 个月数据该月打标标签, 使用当前月数据隔月后预测未来 3 个月用户离网情况。

7.3.4 模型评估

本次课题采用查准率和查全率 2 个指标来评估模型的预测结果, 其计算式如下。

查准率 = 预测为离网且最终离网的用户/预测离网用户数据 (1)

查全率 = 预测为离网且最终离网的用户/实际离网用户数 (2)

采用近期历史数据进行建模和评估, 得到表 5、表 6、表 7 所示的评估数据。

表 5 预付费用户预测次月离网结果 (2016 年 12 月)

预测人数/人	查准率	查全率
25 000	90.60%	8.80%
50 000	81.00%	15.80%
100 000	69.20%	27.00%
150 000	61.50%	36.00%
200 000	55.50%	43.40%
250 000	50.50%	49.30%
300 000	46.30%	54.30%
350 000	42.90%	58.60%
400 000	40.00%	62.40%

表 6 预付费用户隔月预测次 3 月离网结果 (2017 年 2 月)

预测人数/人	查准率	查全率
25 000	71.30%	7.00%
50 000	66.00%	12.90%
100 000	58.30%	22.80%
150 000	53.10%	31.10%
200 000	48.90%	38.20%
250 000	45.50%	44.50%
300 000	42.70%	50.10%
350 000	40.40%	55.20%
400 000	38.30%	59.80%

表 7 后付费用户隔月预测次 3 月离网结果 (2017 年 2 月)

预测人数/人	查准率	查全率
25 000	54.80%	11.80%
50 000	41.70%	18.00%
100 000	31.30%	27.10%
150 000	26.60%	34.50%
200 000	23.90%	41.20%
250 000	21.80%	47.10%
300 000	20.10%	52.10%
350 000	18.80%	56.80%
400 000	17.70%	61.00%

根据业内的统计数据, 国内运营商的离网预测模型查准率达到 40%, 查全率达到 60%, 就可算作优秀的离网预测模型, 根据表 5、表 6、表 7 的结果可以得出, 采用随机森林算法的预测模型在查准率方面有非常大的优势。

7.4 离网根因分析

分析历史离网用户与非离网用户属性分布, 从而确定对应属性特征阈值, 然后, 分析预测用户的对应属性的均值和标准差, 计算对应属性偏好值 = (属性 - 属性平均值) / 方差, 从而确定属性特征阈值 (0.5 或 -0.5), 确定该用户的离网根因, 共分析总结 11 类离网根因, 见表 8。

经实践数据验证, 符合离网根因用户离网率的实际离网率, 相对于全网用户的离网率有明显提升, 离网根因应用效果见表 9、表 10。



表 8 离网根因

序号	根因类别
1	资费原因
2	合约感知原因
3	社会交往影响原因
4	终端换机原因
5	地域变更原因
6	服务质量原因
7	通信质量原因
8	弃卡原因
9	多号类用户
10	新入网三低用户
11	其他原因

表 9 2017 年 1 月非三无用户未来 3 个月离网率 10.4%

特征描述	符合离网根因 用户离网率	相对全网 离网率提升
对资费不满	13.20%	1.27%
流量压抑型	17.81%	1.71%
通话下降	27.29%	2.62%
流量下降	15.25%	1.47%
漫游用户	13.26%	1.27%
对通信质量不满	16.25%	1.56%
新入网三低用户	29.92%	2.88%
异网社交圈比例	26.95%	2.59%
社交圈减小	35.54%	3.42%
多号客户	18.61%	1.79%

表 10 2017 年 1 月预测 Top40 万用户在后 3 个月
实际离网 21.6 万, 查准率 54.18%

特征描述	符合离网根因 用户离网率	相对全网 离网率提升
对资费不满	35.95%	3.46%
流量压抑型	68.98%	6.63%
合约即将到期	69.71%	6.7%
通话下降	59.44%	5.71%
流量下降	65.33%	6.28%
漫游用户	61.23%	5.89%
对服务质量不满	42.74%	4.11%
对通信质量不满	46.46%	4.47%
新入网三低用户	67.49%	6.49%

8 结束语

安徽联通紧抓大数据发展的时代脉搏, 敢于创新、敢于挑战, 在全国运营商中率先实现技术、架构先进的企业级融合大数据平台, 完成安徽联通 B/O/M 域数据的接入及治理, 实现安徽联通的数字化转型及全业务流程的智慧运营, 并基于平台的开放架构, 与融合大数据领域合作伙伴合作实现大数据对外应用的百花齐放。

企业级融合大数据平台是个不断迭代的过程, 未来, 安徽联通智慧运营平台还将朝着“更大、更全、更准”的方向不断迈进, 并探索专业化的团队建设, 通过 PaaS 化功能完善, 实现对内、对外应用的百花齐放。

参考文献:

- [1] 童晓渝, 张云勇, 房秉毅, 等. 大数据时代电信运营商的机遇[J]. 信息通信技术, 2013(1): 5-9.
TONG X Y, ZHANG Y Y, FANG B Y, et al. Opportunities and strategies to adopt big data for telecom operators [J]. Information and Communications Technology, 2013(1): 5-9.
- [2] 左超, 耿庆鹏, 刘旭峰. 基于大数据的电信业务发展策略研究[J]. 邮电设计技术, 2013(10): 1-4.
ZUO C, GENG Q P, LIU X F. Study on operator's development strategy for big data business [J]. Designing Techniques of Posts and Telecommunications, 2013(10): 1-4.
- [3] 丁俊发. 大数据时代的机遇与挑战[J]. 硅谷, 2013(1): 9-10.
DING J F. Opportunities and strategies in big data era [J]. Silicon Valley, 2013(1): 9-10.
- [4] 冯登国, 张敏, 李昊. 大数据安全与隐私保护[J]. 计算机学报, 2014, 37(1): 246-258.
FENG D G, ZHANG M, LI H. Big data security and privacy protection [J]. Chinese Journal of Computers, 2014, 37(1): 246-258.
- [5] 韩晶, 张智江, 王健全, 等. 面向统一运营的电信运营商大数据战略[J]. 电信科学, 2014, 30(11): 154-158.
HAN J, ZHANG Z J, WANG J Q, et al. The unified-operation-oriented big data strategy for telecom operators [J]. Telecommunications Science, 2014, 30(11): 154-158.
- [6] 于鹏. 数据仓库与大数据融合的探讨[J]. 电信科学, 2015, 31(3): 166-170.
YU J. Discussion on integration of data warehouse and big data [J]. Telecommunications Science, 2015, 31(3): 166-170.
- [7] 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.
- [8] 王帅, 汪来富, 金华敏, 等. 网络安全分析中的大数据技术应用[J]. 电信科学, 2015, 31(7): 145-150.
- WANG S, WANG L F, JIN H M, et al. Big data application in network security analysis [J]. Telecommunications Science, 2015, 31(7): 145-150.
- [9] ALMANA A M, AKSOY M S, ALZHRANI R. A survey on data mining techniques in customer churn analysis for telecom industry[J]. Journal of Engineering Research and Applications, 2014, 4(5): 165-171.
- [10] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.
- [11] AU W, CHAN K, YAO X. A novel evolutionary data mining algorithm with applications to churn prediction[J]. IEEE Transactions on Evolutionary Computation, 2003, 7(6): 532-545.
- [12] OSEMAN K B, BINTI S, SHUKOR M, et al. Data mining in churn analysis model for telecommunication industry[J]. Journal of Statistical Modeling and Analytics, 2010 (1): 19-27.
- [13] JUNQUE D, MARTENS D, PROVOST F. Predictive modeling with big data: is bigger really better? [J]. Big Data, 2013, 1(4): 215-226.
- [14] KIM N, JUNG K H, YONG S K, et al. Uniformly subsampled ensemble (use) for churn management: theory and implementation [J]. Expert Systems with Applications, 2012, 39(15): 11839-11845.
- [15] COUSSEMENT K, POEL D. Churn prediction in subscription services: an application of support vector machines while comparing two parameter-selection techniques[J]. Expert Systems with Applications, 2008, 34(1): 313-327.

- [16] GUYON I, LEMAIRE V, VOGEL D, et al. Analysis of the KDD cup 2009: fast scoring on a large orange customer database[J]. ACM SIGKDD Explorations Newsletter, 2009, 11(2): 1-22.

[作者简介]



吴涛（1973-），男，中国联合网络通信有限公司安徽分公司副总经理、大数据项目总负责人、数据分析顾问，主要研究方向为数据分析与用户感知提升、精准化营销分析与数据挖掘、用户消费行为。



刘韬（1976-），男，中国联合网络通信有限公司安徽分公司计划建设与资产运营部数据分析师，主要研究方向为大数据平台架构及搭建、B/O 域基础数据融合、数据分析、行业大数据应用。



王斌（1975-），男，中国联合网络通信有限公司安徽分公司计划建设与资产运营部副总经理，主要负责大数据平台架构、数据分析顾问、用户消费行为数据分析、行业大数据应用研究等工作。