

基于多维多粒度分析的电信网用户行为模式挖掘

程晓涛, 吉立新, 黄瑞阳, 于洪涛, 杨奕卓

(国家数字交换系统工程技术研究中心, 河南 郑州 450002)

摘 要: 为了更好地理解电信网用户行为规律, 以大规模电信网用户通信详细记录 (CDR, call detail record) 数据为研究对象, 运用混合概率模型与特征工程方法, 从用户群体与个体的角度分析了用户呼叫中的通话时长、通话频次、通联关系等多维度特征, 并从小时、天、周等不同时间粒度上进一步细化, 实现了对不同用户群体通话行为模式的有效发现。通过混合概率模型对用户行为中的分布特性进行建模, 解决了用户通话时长、频次等分布特征难以刻画的问题。实验中采用某地区电信网的真实数据作为数据集, 对比了决策树、朴素贝叶斯、SVM 等常见分类算法的实验效果, 证明了所提用户行为特征的有效性 & 计算可行性; 并以快递、航班、银行等服务号码为例, 对比了不同群体通信行为模式的差异性。

关键词: 电信网; 多维度; 多粒度; 高斯混合模型; 行为模式挖掘

中图分类号: TP393

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2018083

User behavior pattern mining method based on multi-dimension and multi-granularity analysis in telecom networks

CHENG Xiaotao, JI Lixin, HUANG Ruiyang, YU Hongtao, YANG Yizhuo

National Digital Switching System Engineering & Technological R&D Center, Zhengzhou 450002, China

Abstract: In order to better understand the behavior of users in telecom networks, it takes CDR (call detail record) data of large-scale telecom network as the research object. By using the mixed probability model and feature engineering method, the multi-dimension characteristics of the call time, call frequency and connections are analyzed from the perspective of user groups and individuals. It is further refined from different time granularities such as hour, day, and week to realize effective discovery of call behavior patterns for different user groups. The distribution characteristics of user behavior are modeled by mixed probability model, which solves the problem of describing the distribution characteristics such as user's call time and frequency. Based on the dataset of a regional telecom network, the performance of decision tree, naive Bayes and SVM classification algorithm are compared. It proves the validity and computational feasibility of the proposed method. The differences in communication behavior patterns of different groups are also compared by taking the service numbers like express, flight and bank as examples.

Key words: telecom network, multi-dimension, multi-granularity, mixture-of-Gaussian, behavior pattern mining

收稿日期: 2018-09-16; 修回日期: 2018-09-28

通信作者: 程晓涛, chengxt90@mail.com

基金项目: 国家自然科学基金创新群体基金资助项目 (No.61521003); 国家自然科学基金资助项目 (No.61601513)

Foundation Items: The Foundation for Innovative Research Groups of the National Natural Science Foundation of China (No.61521003), The National Natural Science Foundation of China (No.61601513)

1 引言

移动通信是人们最常用的交流方式之一,是记录人们通信行为的重要载体。深入理解人类个体用户的交互模式,对于控制信息传播^[1]与预测用户行为^[2]具有重要意义。但由于全网通信数据规模庞大且获取不易,目前针对大规模电信网数据的分析挖掘研究不足,尤其是针对特殊群体用户通信模式及异常用户行为检测^[3-4]方面还不够深入,不同的职业人群中,往往用户的通信模式有不同的体现^[5]。如果电信网运营商能通过分析用户的通信行为特征,准确识别用户类别,则对运营商开展个性服务和安全监控有着重要的指导价值。并且随着近年来电信诈骗活动的日益泛滥,不法分子利用传统电信网的信任关系进行违法犯罪行为,深入开展电信网通信数据挖掘,研究如何从通信数据中找出诈骗分子等特殊群体的异常行为模式也成为一项亟待解决的问题。

然而,目前针对电信网用户行为分析还存在明显不足:一是电信网的通信数据相对于互联网的通信数据私密性更强,对其分析涉及用户隐私等敏感话题,各国政府对该领域的公开研究成果较少,可借鉴的国外经验少;二是针对电信网用户行为特性的分析主要服务于电信运营商产品推广和营销,而较少考虑网络安全和信息获取的需求;三是该领域的公开数据少,研究相对小众,主要以用户群体作为研究对象,针对单个用户行为特性的研究较少。

本文以某地区约 1×10^7 通信用户的省际长途CDR数据为对象,重点分析了用户通信中通话频次、通信时长等特征在群体与个体层面的分布特性,并对几类典型公共服务类号码群体的通信模式进行深入分析。针对电信网用户行为分析中的行为模式挖掘问题,提出一种基于多维多粒度分析的电信网用户行为模式挖掘方法。

本文的主要贡献如下。

1) 借鉴特征工程的思想,提出了一种基于多维度、多粒度分析的大规模电信网用户行为模式挖掘方法,通过对用户呼叫行为与通联关系的抽象与量化,形成具有较好区分性与可计算的用户行为特征,对于用户群体分类与行为模式挖掘具

有重要意义。

2) 针对电信网用户呼叫行为中通话分布特性难以准确量化的难题,采用混合概率模型建模用户行为,将时间序列数据中的用户呼叫分布特征转化为概率曲线匹配问题,利用模型参数刻画用户分布特征,形成用户呼叫分布的量化指标值。

3) 本文以几种典型服务类号码为数据集,通过仿真实验证明了本文所提电信网用户行为特征的有效性,也进一步说明本文提出的电信网用户行为模式分析方法可以有效实现对电信网异常用户的检测。

2 相关研究

要实现大规模电信网数据挖掘,分析用户行为模式是一个关键问题。用户行为模式^[6]是人们在通信过程中所体现出的个体或群体重复出现的通信特征集合。特定群体的通话记录中隐含了许多内部员工的工作性质、日常行为习惯、生活作息时间等有关信息,如何充分挖掘这类信息并实现有效利用是研究者十分关注的问题。

基于通信内容的数据挖掘方法主要集中在邮件通信与互联网通信领域。李全刚等^[7]研究分析了邮件网络中面向事件检测的用户行为模式,对比了域内通信与域外通信信息完整性的差异,并引入模元的概念,将常见的二元对应关系(特征-模式)转化为三元对应关系(特征-模元-模式),并从模元的角度对用户模式进行统一描述。郝秀兰等^[8]采用话题识别与追踪方法,实现了对互联网论坛杂乱无章内容的安全监测与有害信息识别。但在真实情况中,由于通信隐私问题无法获取通话内容,导致该方法研究在电信网领域存在较大政策与法规上的局限性。

由此,一些学者从通信行为所形成的网络结构方面开展广泛研究,挖掘通信网用户行为的内在规律。以针对手机通信数据的研究为例,人类通话网络的节点度分布满足幂律分布,幂指数随着统计样本和统计指标的不同而不同。Aiello等^[9]得到的一天中有向长途电话数据的入度与出度的幂指数均为2.1。Onnela等^[10]得到的双向呼叫无向网络的节点度分布幂指数为8.4,通话时长幂指数是1.9。如果考虑节点的地理位置,则两节点间

连线的概率以及主被叫之间联系的强度（时长和次数）随节点间距离增大而减小。Jiang 等^[11]对通话数据研究发现人们拨打电话的时间间隔只有 3.46% 的用户满足幂律分布，73.34% 的用户是韦伯分布。余晓平等^[12]对手机通话中的网络度进行了全面分析，分别从号码度、通话度、时长的角度分析了用户通信的度分布情况，总体上度分布满足幂律分布的特点，并实证统计了日尺度与不同时段尺度下的用户通信模式，揭示了绝大多数用户每日只拨打 1 个号码的电话，节假日期间拨打电话的用户数、次数、时长减少，但平均通话时长增大的特征。文献[13-14]中采用概率模型的方法实现对 Twitter 或 Blog 中的用户行为建模与事件检测。文献[15-16]中采用大规模图分析的方法对社会媒体中的转发行为进行分析挖掘，以上都为本文的研究提供了思路借鉴。

综合以上分析，本文提出一种基于多维度、多粒度分析的电信网用户行为模式挖掘方法，主要采用特征工程的思想，对通信网信令数据进行细致分析，避免因分析通话内容而导致的隐私问题，引入呼叫离散度、呼损原因、通联关系等特有属性，实现对通信网络中不同职业群体的分类识别与应用。针对电信网络数据的特点，第 4 节将重点介绍本文提出的相关特征提取方法。

3 电信网用户行为模式分析

电信网涵盖用户广，用户种类多，集团号码、客服号码、个人号码等每类用户具有不同的行为特征。同时，不同运营商之间对于漫游号码的处理方式也存在差异，呼叫转移、语音信箱等特殊业务的通信信息往往还需要进行真实主被叫号码的提取与还原，这些实际问题都给电信网数据的分析挖掘带来严峻挑战。

本文数据集提取自某地区电信运营商一个月内的省级长途通话数据。用户 CDR 数据中所使用的具体信息字段包括：主叫号码、被叫号码、呼叫发起时间、通话开始时间、通话结束时间、呼叫结束时间、通话时长、呼损原因、号码归属地等信息。考虑到现有数据的特点，从群体与个体 2 个角度分析用户通信行为规律。个体呼叫行为

特征中，本文主要从通话时长、通话频次、通联关系等多维度特征入手；并按小时、天、周等不同时间粒度特征展开挖掘分析。针对用户行为模式挖掘问题，归纳起来就是根据用户行为提取行为特征，每一种特征都从某一方面反映了用户特性，其单一特征的识别方法存在利用信息不充分的问题，为了提高识别准确率，采用多维多粒度特征综合的方法进行用户行为模式挖掘。

3.1 通信行为群体特征

若用户在一定时间内通话频次为 k ，其概率密度函数为 $P(k)$ ，通话频次满足幂律分布的表达式为

$$P(k) = Ck^{-\gamma} \quad (1)$$

对式(1)取对数，有

$$\ln P(k) = \ln C - \gamma \ln k \quad (2)$$

由式(2)可知，若在双对数坐标系下用户的概率密度分布近似为一条直线，则用户该特征满足幂律分布。对数据集中网络用户一天内通话频次的统计结果如图 1(a)所示，横坐标代表用户一天内的通话频次，纵坐标代表该通信频次的人数占总人数的比例。由图 1(b)可知，用户群体通话频次符合幂律分布，幂指数为 1.694，幂律分布是社会与自然界中普遍存在的现象。

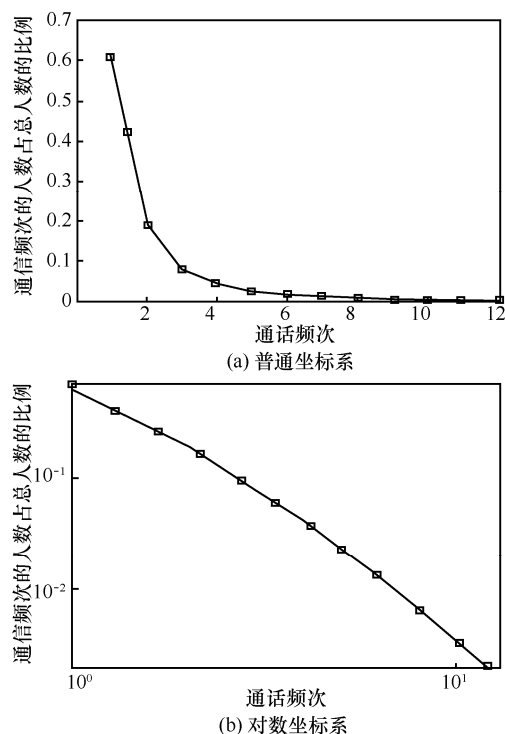


图 1 用户通话频次的概率密度分布

3.2 通信行为个体特征

针对单个用户行为特征, 需要从数据集中分别统计单个用户的通信频次分布、通话时长分布、通联关系等特征, 定义如下特征度量指标。

定义 1 呼叫离散度。借鉴信息论知识, 该值越大, 表明用户拨打的范围越大, 每个人之间越平均, 该值越小, 表明用户拨打的范围越小, 可能只是少数几个朋友。

$$\phi_i = \frac{-\sum_{j=1}^{k_i} \log p_{ij}}{\log k_i} \quad (3)$$

$$p_{ij} = \frac{n_{ij}^{ij}}{n_{ij}^i} = \frac{n_{ij}^{ij}}{\sum_{j=1}^{k_i} n_{ij}^{ij}} \quad (4)$$

其中, k_i 是用户 i 呼出的不同电话的数量, n_{ij}^{ij} 是用户 i 拨打给用户 j 的呼叫次数, n_{ij}^i 是用户 i 所有拨打的用户呼叫数量。

定义 2 日通话频次分布。根据用户 CDR 数据, 提取用户在一个月内每天各小时内的通话次数, 然后计算每个小时的通话次数占总通话次数的比例。由此可以得到用户在 24 h 内通话次数的分布特征, 表示为

$$X_n = \{x_0, x_1, \dots, x_{23}\} \quad (5)$$

其中, x_i 表示用户一段时间内每天 24 h 中第 i h 到 $i+1$ h 内的通话次数占总通话次数的比例。用户一天中通话次数的分布反映了用户在一天中各时间段的通话活跃度情况, 间接表明了用户的工作、休息等生活习惯。

定义 3 日通话时长分布。根据用户每小时内通话时长占总通话时长的比例, 可以计算得到用户 24 h 内的通话时长分布情况, 表示为

$$X_t = \{t_0, t_1, \dots, t_{23}\} \quad (6)$$

其中, t_i 表示用户一段时间内在 24 h 中第 i h 到 $i+1$ h 内的通话时长占总通话时长的占比。用户的通话时长分布特性反映了用户的主要通话特征, 如图 2 所示。一般而言, 具有长通话时长的对象为关系亲密且稳定的联系人。

定义 4 周通信频次、时长分布。根据用户一周内各天的通话频次及时长情况, 确定用户的通话分布情况。

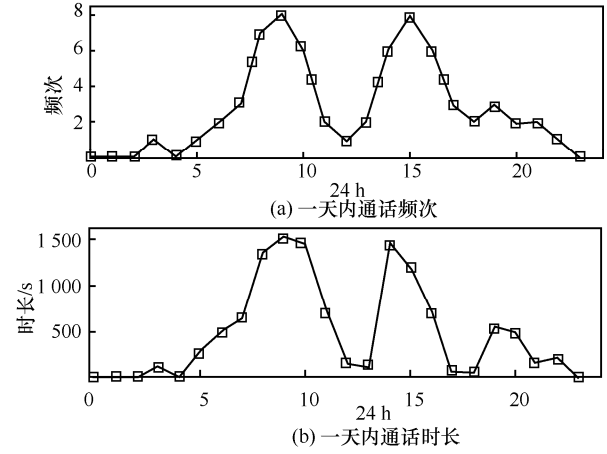


图 2 通话频次、时长一天内的变化

$$\begin{aligned} X_w &= \{w_0, w_1, \dots, w_6\} \\ X_d &= \{d_0, d_1, \dots, d_6\} \end{aligned} \quad (7)$$

用户在一周内的通话频次及时长分布情况, 反映了用户在一周内的工作、休息情况。这类用户号码的日常通话一般呈现明显的潮汐效应, 即工作日与非工作日的通话量有明显变化, 对于识别判定用户号码性质 (工作、生活) 具有重要价值。

定义 5 拨打空号比。定义用户拨打空号比为某用户在一段时间内拨打的空号占总呼叫个数的比例。该指标反映了用户发起呼叫的正确程度。一般正常用户呼叫都是基于通信录的, 空号很少, 而推销或骚扰用户的呼叫可能使用相关自动拨号软件, 空号比例较高。

设用户 A 在时间 t 内发起的呼叫个数为 N_t , 拨打空号的个数为 N_e , 则用户拨打空号比为

$$U_{\text{caller}}^A = \frac{N_e}{N_t} \quad (8)$$

定义 6 主被叫占比。定义主被叫占比为某用户号码在一段时间通话中发起呼叫次数与用户接听呼叫次数的比值。根据经验可知, 正常用户的发起呼叫次数与接听呼叫次数应该相对均衡, 比值近似于 1; 而异常骚扰或诈骗用户可能会发起大量主叫, 而被叫次数很少。

设用户 A 在时间 t 内发起的呼叫个数为 N_t , 用户 A 在时间 t 内的被叫次数为 $N_{t-\text{called}}$, 则用户主被叫占比为

$$U_{\text{called}}^A = \frac{N_t}{N_{t-\text{called}}} \quad (9)$$

表 1 不同类别号码的 top-5 联系人通话频次占比及其累加和

类别	号码	Top-1	Top-2	Top-3	Top-4	Top-5	累加和
私人号码	185****6507	0.161 62	0.141 41	0.101 01	0.101 01	0.060 61	0.565 66
	138****0696	0.283 87	0.103 23	0.070 97	0.070 97	0.064 52	0.593 55
服务类号码	95558	0.021 53	0.016 28	0.009 75	0.009 72	0.009 59	0.066 87
	95566	0.007 06	0.007 06	0.005 29	0.005 29	0.005 29	0.029 98
	95339	0.025 86	0.017 24	0.008 62	0.008 62	0.008 62	0.068 97
	95530	0.008 25	0.005 50	0.004 58	0.003 67	0.003 67	0.025 66

定义 7 Top- k 呼叫占比。根据一段时间内用户通话联系人频次或时长进行排名，前 K 个主要联系人的通话频次或时长占总通话频次或时长的比例。

$$R_{\text{top-}k} = \frac{N_{\text{top-}k}}{N_{\text{total}}} \quad (10)$$

公共服务类号码由于其面向大众，其联系人数量理论上应该相当庞大，且没有固定的几个联系人。以 2 个个人号码与常见的 3 类公共服务号码作为对比，分别对其作为主叫和被叫时的 Top- k 联系人通话频次占比进行统计。表 1 所示的统计分析结果证实了本文这一猜测，即私人号码其与 Top-5 联系人的通话占了很大比例，而公共服务类号码其与 Top-5 联系人的通话占总通话频次的比例较小。

定义 8 呼损原因。即用户呼叫失败原因，本文 CDR 数据记录有 fail_reason 字段，代表不同的呼叫失败原因，具体字段说明如表 2 所示。

表 2 呼损原因字段说明

呼损码	失败原因
0	呼叫成功应答(主叫先挂机)
1	拨号中途挂机
2	振铃中途挂机
3	空号
4	用户市话忙
5	用户长话忙
6	呼叫异常释放
7~14	其他
15	呼叫故障
16	呼叫成功应答(被叫先挂机)

3.3 用户通联关系特征

在电信用户呼叫网络中，单个用户 u 与联系人 u_{neighbor} 之间的通联关系在一定时间范围具有稳定性。但随着时间的推移，用户 u 的联系人集合 $\{u_{\text{neighbor}}\}$ ，即他们之间的通联统计关系可能会发生相应的变化。因此本文采用时序通联特征对单个用户的通联行为进行表示。首先在时间轴上选取一系列等距的时间点 $T = \{t_0, t_1, \dots, t_{n-1}, t_n\}$ ，然后在每个时间区间 $\Delta t_i = t_{i+1} - t_i$ 上提取用户 u 与其邻居节点之间的通联关系特征，如图 3 所示。

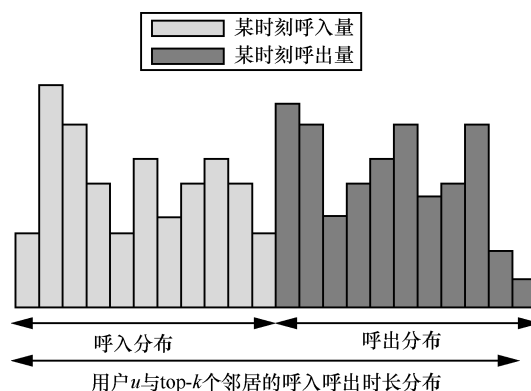


图 3 单个目标号码在时间轴上的特征表示示意

在图 4 中，目标号码节点 A 有 B、C、D、E 和 F 等 5 个邻居节点。节点 A 和邻居节点的呼叫关系表示为有向图，节点 A 作为主叫的呼出表示为实线箭头，节点 A 作为被叫的呼入表示为虚线箭头。箭头和连接线的权值表示节点之间通话的频次高低，频次越高，权值越大。

通过对电信网单个用户的呼叫特征分析可知，虽然在较长时间内用户 u 的联系人集合大小 $\|u_{\text{neighbor}}\|$ 可能很大，但是与用户 u 经常联系的邻

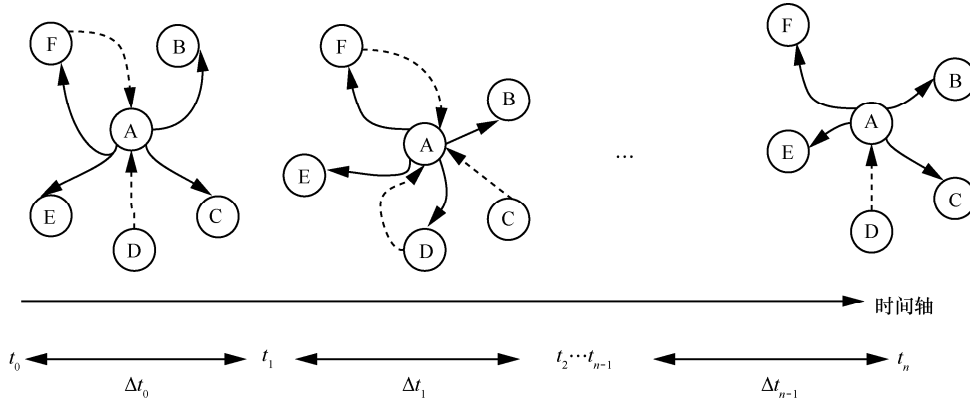


图 4 目标号码 A 在时间轴上与邻居节点的通联关系

居节点数量并不多。基于上述分析和假设,选取与用户 u 的通话频次(包括呼出和呼入,即出度 d_{out} 和入度 d_{in} 之和 d) 占有通联频次比例大于指定阈值 τ 的前 k_u 个邻居节点,统计他们之间的通联特征。具体地,计算用户 u 与所有邻居节点 u_{neighbor} 之间的通话次数,记为 $C_{\text{fig}} = d_u = d_{u,\text{out}} + d_{u,\text{in}}$;按照与用户 u 的通话频次,对集合 $\{u_{\text{neighbor}}\}$ 排序,并选取前 k_u 邻居节点组成集合 $\{u_{\text{neighbor}}^{k_u}\}$,使 k_u 个节点与用户 u 之间通话的频次比例 $r_{u..} \geq \tau$,即

$$r_{u..} = \frac{\sum_{v \in \{u_{\text{neighbor}}^{k_u}\}} d_{v,u,\text{in}} + d_{v,u,\text{out}}}{C_{\text{fig}}} \geq \tau \quad (11)$$

对于不同类型的电信用户 r 的 k_u 值可能不同。为了保证对所有用户的特征表示等长,每个用户最大的 Top- k 邻居节点数设置为 N 。对于 Top- k 个邻居数小于 N 的用户 u ,不足的部分设置为 0;对于 Top- k 个邻居数大于 N 的用户 u ,至多选取 N 个邻居节点进行统计。对于每个用户 u 的 Top- k 个邻居节点,按照通话频次排序,累计每个邻居节点 $v \in \{u_{\text{neighbor}}\}$ 与 u 的通话时长。为区分呼入和呼出的统计特性,通联关系特征中划分呼出特征分布和呼入特征分布 2 个不同。最终得到的通联特征 $u_{u,\text{conn}}$ 示意如图 3 所示。

通过在时间轴上的平移,可以得到用户 u 在不同时间段内的时序通联关系特征 $F = \{f_{u,\text{conn}}, \Delta t_0, f_{u,\text{conn}}, \Delta t_1, \dots, f_{u,\text{conn}}, \Delta t_{n-1}\}$ 。获取用户时序通联关系特征后,通过频繁模式挖掘算法^[17]即可得到用户不同时段内的频繁主叫与频繁被叫信息。

4 电信网用户行为模式挖掘

电信网用户通话行为具有显著的规律性与差异性,不同用户群体具有不同的呼叫行为模式^[18]。为了验证本文所提取的电信网用户行为特征的有效性,基于现有电信网 CDR 数据,本文选取其中部分服务类号码(快递、航班、银行等)和个人用户号码,提取用户呼叫频次、呼叫时长、通联关系等方面的特征构建用户特征向量,通过模型训练选取效果较好的分类器进行示范应用。对于用户呼叫行为中的通话频次、时长等分布特征,采用高斯混合概率模型进行刻画,模型参数作为用户呼叫行为分布的特征。

4.1 电信网用户特征提取

电信网用户行为特征提取,即是对电信网用户呼叫行为进行量化统计,制定合理的特征度量用以表征用户的呼叫行为特点^[19-20],为用户分类及模式挖掘奠定基础。根据现有数据特点及电信用户的呼叫习惯,提取用户 4 个方面共计 21 项特征,用于电信网用户群体的行为模式挖掘,如表 3 所示。

本文采用 k 维的特征向量表示用户 u 的呼叫行为特征,即

$$P(u) \in \mathbb{R}^k \quad (12)$$

所提取的用户特征向量构成用户的特征空间,本文旨在通过分析不同用户群体的行为特征,判别用户号码类别或发现其中的异常用户行为。该方法适用于大规模数据的处理,特征计算中主要为统计计算,分析方法并不复杂,使用 Hadoop、Spark 等分布式计算工具即可完成。

表 3

用户通信行为特征指标

特征类别	呼叫号码特征	呼叫时长特征	呼叫频次特征	通联关系特征
指标	呼叫离散度	呼叫时长(天)	呼叫频次	拨打空号次数
	呼叫区域分布	最大呼叫时长(天)	呼叫号码数	出度
	主被叫占比	最小呼叫时长(天)	通话频次分布(天)	入度
	Top- k 呼叫占比	通话时长分布(天)	通话频次分布(时)	呼叫离散度
	—	通话时长分布(时)	主叫挂机次数	频繁主叫
	—	—	被叫挂机次数	频繁被叫

4.2 基于混合概率模型的呼叫行为建模

在构造通信用户特征向量的过程中,主被叫占比、呼叫频次、时长等数值特征可以直接应用。但对于用户通话频次分布、呼叫时长分布等非数值特征却无法直接应用。本节中引入高斯混合模型将具有不同特性的通话行为分布进行区分,并使用模型的参数值作为用户特征向量的一部分。考虑到用户的通话行为是一种不确定的随机过程,并且在不同时段内具有不同的分布情况,使用混合概率模型进行刻画可以达到更好的效果,文献[21]中也有一定应用尝试。

混合概率模型:使用混合概率模型对电信网用户呼叫行为进行建模,可以看作每个通信用户 U 以一定的概率属于 K 种通信模式中的一种,则用户所表现出的具体通信行为是第 $k(1 \leq k \leq K)$ 种通信模式的概率分布的某一具体表现量。将用户的历史行为数据(history behavior data)记作 $HD(u)$; α_k 是选取的第 k 种通信模式的概率分布的权重参数; $\phi(y|\theta_k)$ 是选择的概率模型的概率密度分布, θ 为相应的概率模型的参数,如针对高斯混合模型,则 $\theta_k = (\mu_k, \sigma_k^2)$, 其中 μ_k 为均值, σ_k^2 为方差。基于混合概率模型的电信网用户行为模型的公式化定义如下。

$$p(y|\theta) = \sum_{k=1}^K \alpha_k \phi(y|\theta_k) \quad (13)$$

$$\sum_{k=1}^K \alpha_k = 1$$

上述模型可以用作电信网用户的通用模型,当针对单个用户进行建模时,可对特定用户行为数据计算相应用户行为的分布参数,如一天中用户 u 的通话频次、通话时长分布,计算用户 u 的模型参数。

$$p(y_u|\theta) = \sum_{k=1}^K \alpha_{u,k} \phi(y|\theta_k)$$

$$\sum_{k=1}^K \alpha_{u,k} = 1 \quad (14)$$

若采用高斯(正态)分布、 $\phi(y|\theta_k)$ 正态分布的概率密度函数,则需要估计的参数为

$$\theta_k = (\mu_k, \sigma_k^2) \quad (15)$$

式(16)称为第 k 个分模型。

$$\phi(y|\theta_k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right) \quad (16)$$

即假设每个特征点由一个单高斯分布生成。

4.3 参数学习过程

当已知用户通话频次或通话时长的观测数据后,参数估计的过程就相当于找一组参数,这组参数确定的概率分布生成这些观测数据的概率最大。而这个概率就是观测数据整体的最大似然函数,即为个体的最大似然函数的乘积。

$$p(HD|\theta) = \prod_{i=1}^{N_U} p(HD_u|\theta) \quad (17)$$

$$p(HD_u|\theta) = \prod_{j=1}^{N_k} p(y_u|\theta)$$

需要进行估计的参数 θ 包括全局模型中的 α_k 及所选用的概率分布中对应的表征参数,选用最大似然估计的方法如式(18)。

$$p(HD|\theta) = \prod_{i=1}^{N_U} p(HD_u|\theta)$$

$$= \prod_{i=1}^{N_U} \prod_{j=1}^{N_k} p(y_u|\theta)$$

$$= \prod_{i=1}^{N_U} \prod_{j=1}^{N_k} \sum_{k=1}^K \alpha_{u,k} \phi(y|\theta_k)$$

$$\hat{\theta} = \arg \max \log p(HD|\theta) \quad (18)$$

由于 $p(HD|\theta)$ 中存在加和的形式,所以这个问题没有解析解,只有通过迭代的方式求解,通常采用 EM(expectation-maximization) 算法进行

参数估计计算。

EM 算法执行过程如下。

输入 观测数据 (即用户 CDR 数据) 经过特征提取后的用户通信频次或通信时长记录 $\{y_1, y_2, \dots, y_n\}$ (n 为通信记录的个数), 相应概率混合模型。

输出 混合概率模型参数。

1) 取设定的参数初值进行迭代。

2) E 步: 根据当前的模型参数, 计算各个概率分布 k 对观测数据 y_j 的响应度 $\hat{\gamma}_{jk}$ 。对于高斯混合模型

$$\hat{\gamma}_{jk} = \frac{\alpha_k \phi(y_j | \theta_k)}{\sum_{k=1}^K \alpha_k \phi(y_j | \theta_k)} \quad (19)$$

其中, $j=1, 2, \dots, N; k=1, 2, \dots, K$ 。

3) M 步: 计算新一轮迭代参数

$$\begin{aligned} \hat{\mu}_k &= \frac{\sum_{j=1}^N \hat{\gamma}_{jk} y_j}{\sum_{j=1}^N \hat{\gamma}_{jk}}, k=1, 2, \dots, K \\ \hat{\mu}_k^2 &= \frac{\sum_{j=1}^N \hat{\gamma}_{jk} (y_j - \mu_k)^2}{\sum_{j=1}^N \hat{\gamma}_{jk}}, k=1, 2, \dots, K \\ \hat{\alpha}_k &= \frac{\sum_{j=1}^N \hat{\gamma}_{jk}}{N}, k=1, 2, \dots, K \end{aligned} \quad (20)$$

4) 重复步骤 2) 和步骤 3), 不断更新上面 3 个值, 直到参数值稳定 (收敛)。

$$|p(y|\theta) - \hat{p}(y|\theta)| < \varepsilon \quad (21)$$

其中, $\hat{p}(y|\theta)$ 为更新参数计算后的值, 式(21)表示前后两次迭代结果的变化小于阈值 ε , 通常取 $\varepsilon=10^{-5}$ 。计算所获得的混合模型参数作为用户通话频次或通话时长分布的特征用于用户号码分类实验。

5 实验与结果

本文在真实电信网数据集上测试验证所提出的用户呼叫行为特征及行为模式挖掘方法的有效性。以某地区约 1×10^7 电信网用户通信 CDR 数据作为数据集 (真实号码已脱敏处理), 数据情况如表 4 所示。

表 4 实验数据信息统计

用户数	记录数	时间	数据大小/GB
14 560 000	20 亿	2017.4.1~2017.4.30	200

5.1 评价标准

对于本文中电信网用户号码群体的分类问题, 常用的度量评价指标为查准率、查全率及 AUC 值。根据样例的真实类别与分类的预测类别划分为真正例 (TP, true positive)、假正例 (FP, false positive)、真反例 (TN, true negative)、假反例 (FN, false negative), 查准率与查全率分别定义为

$$precision = \frac{TP}{TP + FP} \times 100\% \quad (22)$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (23)$$

AUC 值为 ROC 曲线下区域的面积, 是研究学习器泛化性能的有力工具。ROC 曲线的纵轴是“真正例率” (TPR, true positive rate), 横轴是“假正例率” (FPR, false positive rate), 两者分别定义为

$$TPR = \frac{TP}{TP + FN} \quad (24)$$

$$FPR = \frac{FP}{TP + FP} \quad (25)$$

5.2 实验和结果

本文设计了 3 类实验来验证所提出的电信网用户行为特征的有效性及其可行性。实验 1 根据所提取的用户行为特征对 3 类服务号码和个人用户号码进行了分类, 证明了所提用户特征的有效性、计算可行性。实验 2 通过对用户号码聚类中的离群点检测, 分析了正常用户与异常用户的行为模式差异。实验 3 通过对用户分时段通信频次分布、呼损原因等特征的分析, 总结了快递类、列车/航班类、金融服务类电话的行为模式, 证明了所提特征的合理性。

实验 1 用户号码分类实验结果如表 5 所示, 选用了 C4.5 决策树、朴素贝叶斯、SVM 这 3 种常见的分类器进行用户号码分类实验。实验表明, 在用户特征提取良好的情况下, 决策树算法的整体效果 AUC 值略优于其他算法, 本文所提取的模型特征中多为连续型数值特征, 如呼叫频次、呼叫时长等; 并存在部分数值缺失的情况,

而该算法对连续型数据和不完整数据的效果较好。朴素贝叶斯算法需满足属性之间相互独立的假设，当属性个数增多且特征之间相关性较大时，分类效率比不上决策树，SVM 算法效果居中。在决策树算法中，通过计算比较各特征的信息增益率，可以找出对用户分类更有效的统计特征。用户号码的统计特征是准确进行用户行为识别的关键。

表 5 用户号码分类实验

算法	准确率	召回率	AUC
决策树	0.82	0.85	0.862
朴素贝叶斯	0.84	0.76	0.848
SVM	0.85	0.79	0.853

在用户呼叫行为分布特征提取中，使用高斯混合概率模型，可以平滑地近似任意形状的密度分布。根据实际通信数据集情况，将用户在不同时段内的通话频次（通话时长）作为随机变量，观测数据通话频次序列的分布情况如图 5(a)所示，横坐标为 24 小时×30 天内小时数，纵坐标为对应小时内的通话频次数。图 5(b)为用户通话频次序列的统计直方图，横坐标为用户的通话频次，纵坐标为该频次在总体分布中所占的比例。图 5(c)为高斯混合模型中各高斯分量的曲线，图 5(d)为高斯混合模型与通话频次直方图的拟合图。由此，可以将用户一段时间内的通话频次序列数据转化为高斯混合概率模型的参数，实现

对用户通话频次特征的刻画，分析用户通话中隐含的通话行为模式。对于高斯混合模型中高斯分量个数的选取，根据采用混合模型对原始数据进行聚类的 ARI^[22]指标值确定，ARI 指标越大，表明拟合的效果越好，本文中采用的高斯分量个数为 4，如图 6 所示。同时表明典型的用户通信频次中呈现 4 种模式，可能与用户在一天中不同时间段内（凌晨、上午、下午、晚上）的通话习惯有关。

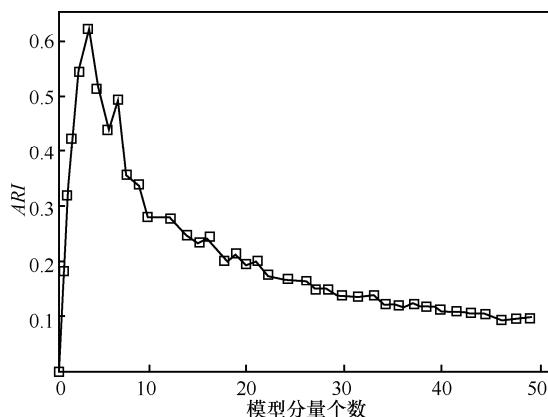


图 6 高斯混合模型分量个数影响

实验 2 通过对用户号码分类中的离群异常点进行检测，可以发现其中的行为异常用户。图 7(a)中，用户号码的呼叫行为主要集中于白天工作时间及晚上睡觉前，凌晨 0 点至早 9 点前用户呼叫几乎没有，且从呼叫频次上符合人工拨打的习惯，由此可大致推断该号码为用户推销电话或者业务工作繁忙的商务人士。图 7(b)中记录了另

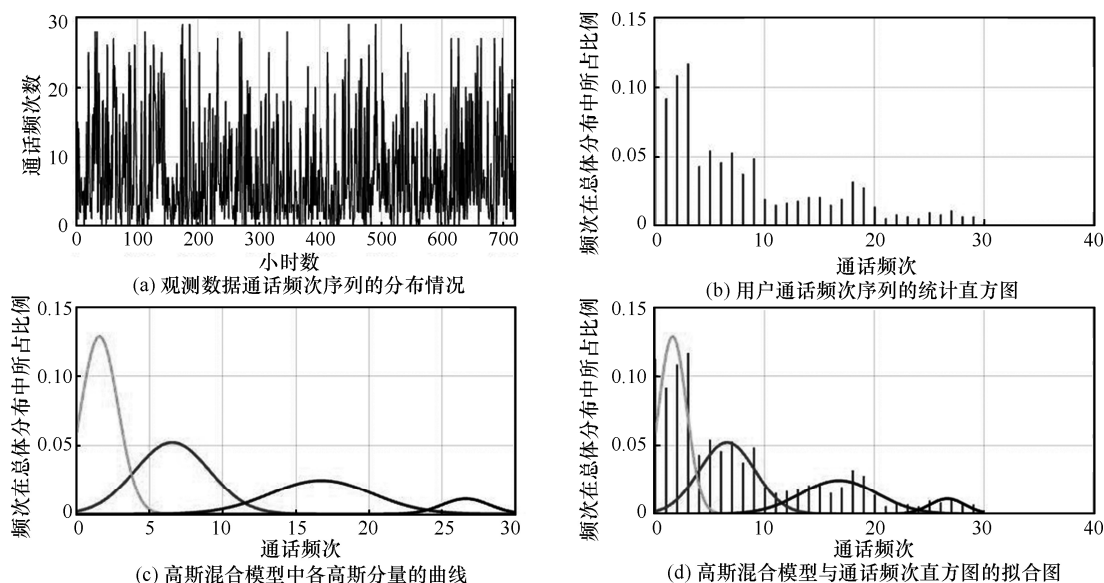


图 5 混合高斯模型拟合图

外一种类型的骚扰电话, 由图可知, 该号码每半小时内呼叫频次一直很高, 且全天呼叫频次虽有波动, 但一直相对稳定, 没有白天、晚上的周期性变化, 由此可推断该号码可能为一种由拨号软件发起的骚扰电话。

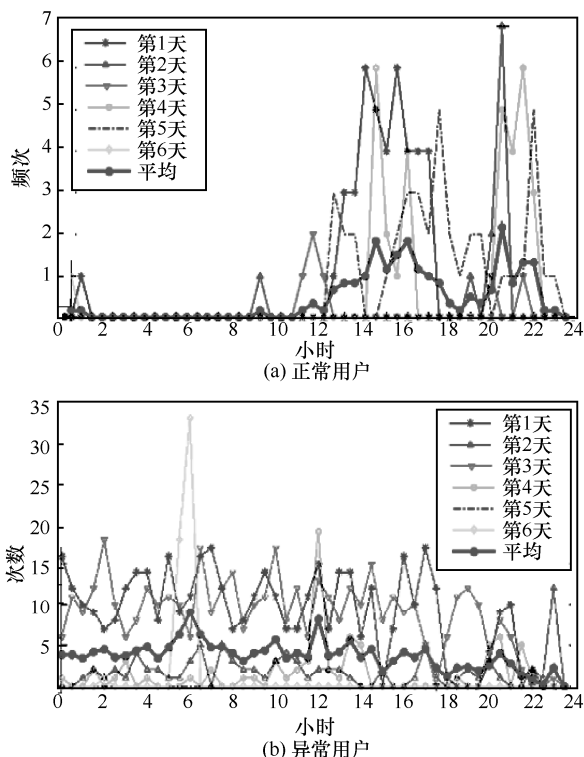


图7 正常用户与异常用户的行为比较

实验3 对不同类别的服务号码进行通话频次统计, 设定3种粒度, 即以周为粒度、以天为粒度、以小时为粒度。结果表明, 不同类别的号码在以小时为粒度时, 反映出较强的特点。将不同类型的号码每小时的通话频次数列化后可以看出较强的规律, 因此考虑以小时为粒度, 统计在1个月内所有号码在每天24个时段中的通话频次累加和作为号码分类的一个特征。下面分别对快递类、列车/航班类和银行类的1个月内以小时为粒度通话频次累加和进行可视化展示, 分析其各自的特点。

图8是快递类的通话频次分布, 横坐标从0时开始到24时, 第一个竖条表示2个月内从凌晨0时到1时之间的所有通话频次和, 以此类推。从图中可以明显看出, 快递类的通话频次分布有极强的相似性, 具体特点表现为: 从21时开始, 通话频次迅速提高; 到1时到4时, 通话

频次稳定在一个很高的水平; 从4时开始每小时频次有所下降, 但仍保持稳定; 到10时开始通话频次出现明显的几乎呈指数程度的递减, 在16时左右下降到几乎为0, 到21时通话数量又开始递增。出现这种情况的原因可能是快递公司与快递点之间的业务联系都是在凌晨和上午, 在这段时间快递公司会将快递送到各个快递点, 因而这段时间通话非常频繁。通过图8可以看出, 这种通话规律与普通电信网用户完全不同, 具有极强的时序特征。

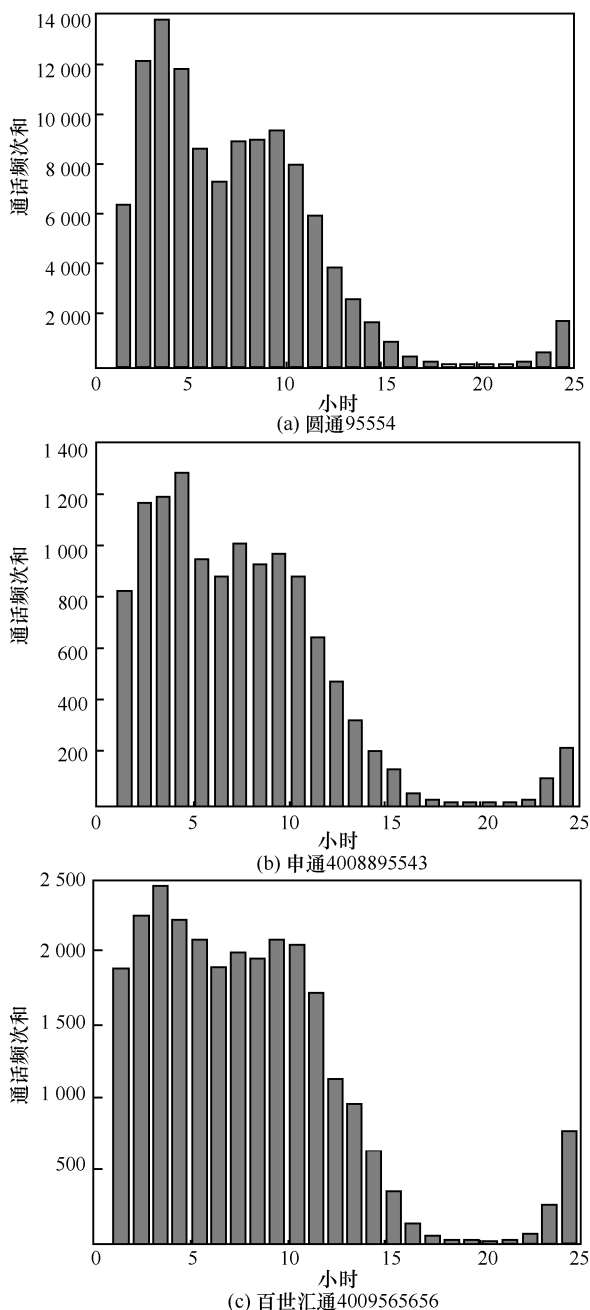


图8 快递类按时刻通话频次累加和分布

图 9 给出列车/航班类的通话频次分布。从图中可以看出，通话频次的分布仍然具有相似性，具体表现为：17 时到 22 时通话频次处在一个较低的水平；22 时到 3 时通话频次迅速上升，4 时通话频次略微下降但仍然较高；11 时到 15 时通话频次显著下降；15 时到 17 时迅速下降到极低的水平。通过图 10 中银行类服务电话的分布可以看出，这种规律也显示出了与众不同的特性，并且类内相似度极高。

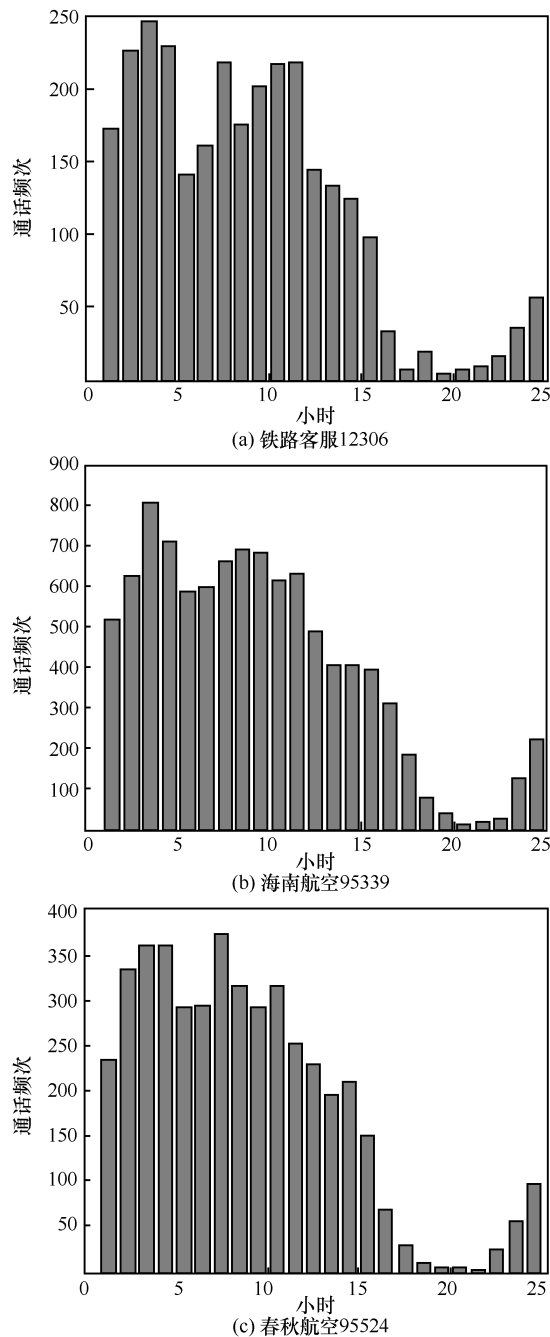


图 9 列车航班类按时刻通话频次累加和分布

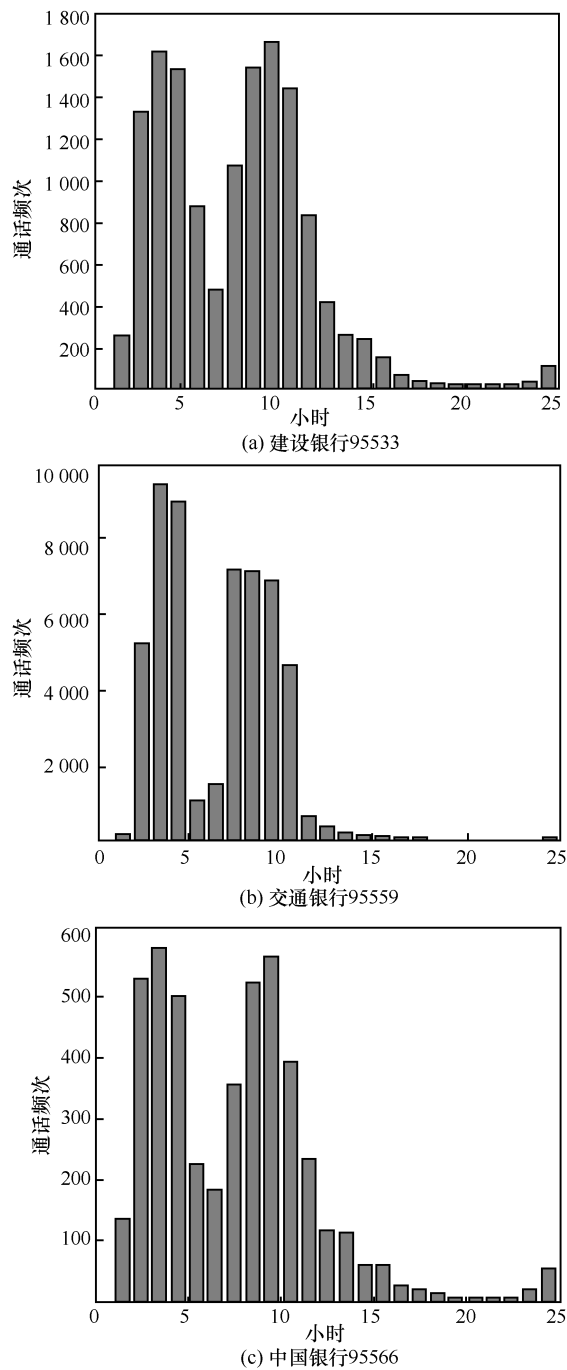
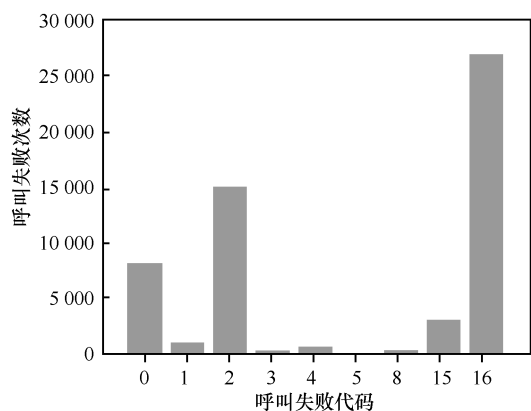
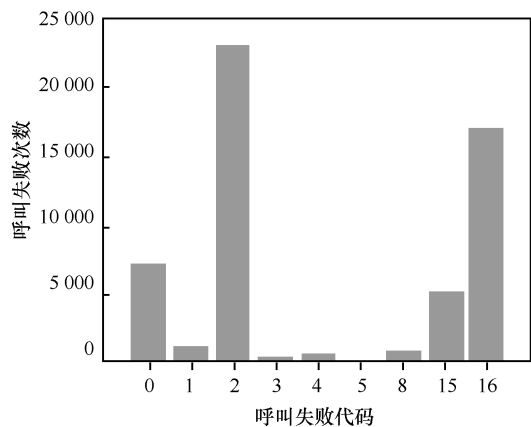


图 10 银行类按时刻通话频次累加和分布

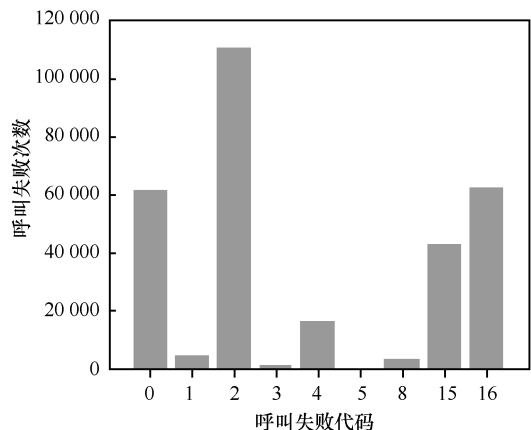
图 11 给出银行类号码作为主叫时的失败原因统计，从图中可以看出，该类号码的呼叫失败原因分布非常相似，呼叫失败代码 2（振铃中途挂机）占了较大的比重，说明这些号码在作为主叫呼叫用户时，有很大一部分振铃时就被用户挂断；另外呼叫失败代码 16 占比较高，即接通后被叫（用户）先挂机，这也符合其作为服务类号码不主动挂机的特点。



(a) 中信银行95558



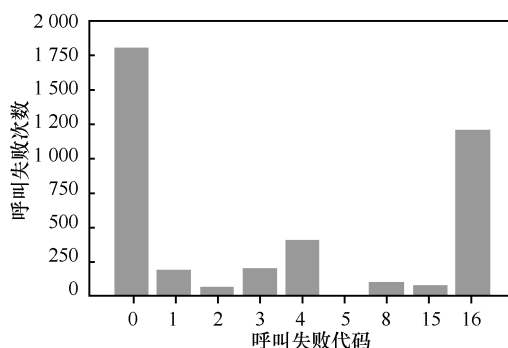
(b) 交通银行95559



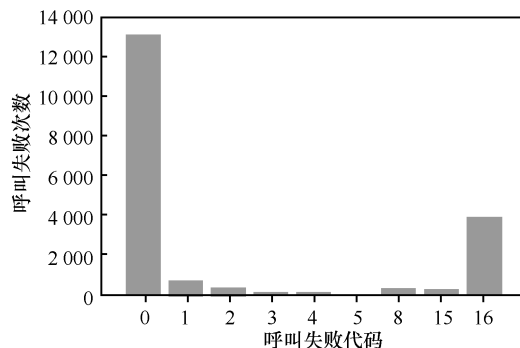
(c) 工商银行95588

图 11 银行类作为主叫时失败原因统计

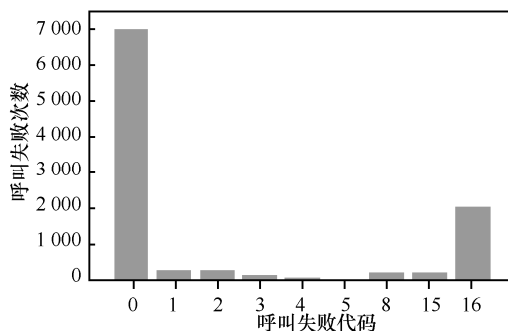
图 12 给出航班类号码作为被叫时的失败原因统计。从图中可以看出, 当列车航班类作为被叫时, 振铃中途挂机 (代码 2) 现象明显减少, 说明服务类号码通常不会挂断用户的电话, 个别中途挂机现象可能是主叫用户误拨导致。同时, 大部分呼叫结束的代码为 0 和 16, 说明绝大多数的通话都是成功的, 并且多数情况是主叫用户先挂机, 同样符合服务类号码的特点。



(a) 春秋航空95524



(b) 东方航空95530



(b) 海南航空95339

图 12 航班类作为被叫时失败原因统计

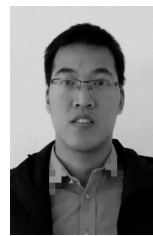
6 结束语

在大数据时代, 针对海量通信数据的挖掘与分析尤为重要, 通信数据的复杂多样性以及数据挖掘技术在各行业应用的特殊性也为数据挖掘提出了新的理论与技术挑战。本文针对海量通信数据中用户行为模式挖掘困难这一挑战, 提出一种基于多维粒度的电信网用户行为模式挖掘方法, 有效解决了电信网用户行为数据中的特征提取问题。在电信网用户行为特征分析过程中, 目前的数据源仅采用了电信网的 CDR 数据, 来源相对比较单一, 后续可以考虑加入用户语音及上网数据信息, 辅助进行用户行为特征的分析挖掘。

参考文献：

- [1] 刘衍珩, 李飞鹏, 孙鑫, 朱建启. 基于信息传播的社交网络拓扑模型[J]. 通信学报, 2013, 34(4): 1-9.
LIU Y Y, LI F P, SUN X, ZHU J Q. Social network model based on the transmission of information[J]. Journal on Communications, 2013, 34(4): 1-9.
- [2] 曹玖新, 吴江林, 石伟, 等. 新浪微博网信息传播分析与预测[J]. 计算机学报, 2014, 37(4): 779-790.
CAO J X, WU J L, SHI W, et al. Sina microblog information diffusion analysis and prediction[J]. Journal of Computers, 2014, 37(4): 779-790.
- [3] 杨杰, 李小平, 陈焱. 基于增量时空轨迹大数据的群体挖掘方法[J]. 计算机研究与发展, 2014(s2): 76-85.
YANG J, LI X P, CHEN T. A group mining method for incremental spatio-temporal trajectory bigdata[J]. Journal of Computer Research and Development, 2014(s2): 76-85.
- [4] 李艳丽, 刘阳, 谢文波, 等. 大数据发现非法传销网络[J]. 大数据, 2017, 3(5): 106-112.
LI Y L, LIU Y, XIE W P, et al. Uncovering the illegal pyramid networks by big data[J]. Big Data Research, 2017, 3(5): 106-112.
- [5] 史殿习, 李寒, 杨若松, 等. 用户日常频繁行为模式挖掘[J]. 国防科技大学学报, 2017, 39(1): 74-80.
SHI D X, LI H, YANG R S, et al. Mining user frequent behavior patterns in daily life[J]. Journal of National University of Defense Technology, 2017, 39(1): 74-80.
- [6] 文雨, 王伟平, 孟丹. 面向内部威胁检测的用户跨域行为模式挖掘[J]. 计算机学报, 2016, 39(8): 1555-1569.
WEN Y, WANG W P, MENG D. Mining user cross-domain behavior patterns for insider threat detection[J]. Journal of Computer, 2016, 39(8): 1555-1569.
- [7] 李全刚, 时金桥, 秦志光, 等. 面向邮件网络事件检测的用户行为模式挖掘[J]. 计算机学报, 2014, 37(5): 1135-1146.
LI Q G, SHI J Q, QIN Z G, et al. Mining user behavior patterns for event detection in email networks[J]. Journal of Computer, 2014, 37(5): 1135-1146.
- [8] 郝秀兰, 胡运发, 申倩. 中文论坛内容监测的方法研究[J]. 中文信息学报, 2012, 26(3): 129-137.
HAO X L, HU Y F, SHEN Q. Research on content monitoring on Chinese web forums[J]. Journal of Chinese Information Processing, 2012, 26(3): 129-137.
- [9] AIELLO W, CHUNG F, LU L. A random graph model for power law graphs[J]. Experimental Mathematics, 2001, 10(1): 53-66.
- [10] ONNELA J P, SARAMÄKI J, HYVÖNEN J, et al. Structure and tie strengths in mobile communication networks[J]. Proceedings of the national academy of sciences, 2007, 104(18): 7332-7336.
- [11] JIANG Z Q, XIE W J, LI M X, et al. Calling patterns in human communication dynamics[J]. Proceedings of the National Academy of Sciences, 2013, 110(5): 1600-1605.
- [12] 余晓平, 裴韬. 手机通话网络度特征分析[J]. 物理学报, 2013, 62(2): 1-9.
YU X P, PEI T. Analysis on degree characteristics of mobile call network[J]. Acta Phys Sin, 2013, 62(2): 1-9.
- [13] CAPDEVILA J, CERQUIDES J, TORRES J. Recognizing warblers: a probabilistic model for event detection in Twitter[C]. ICML2016 Anomaly Detection Workshop, New York, USA, 2016.
- [14] MANAVOGLU E, PAVLOV D, GILES C L. Probabilistic user behavior models[C]//Proceedings of the Third IEEE International Conference on Data Mining, Melbourne, Florida. 2003: 203-210.
- [15] BEUTEL A. User behavior modeling with large-scale graph analysis[D]. Computer Science Department School of Computer Science Carnegie Mellon University Pittsburgh, PA. 2016
- [16] XU Z H, ZHANG Y, WU, et al. Modeling user posting behavior on social media[C]. SIGIR'12, Portland, Oregon, USA. 2012.
- [17] 张啸剑, 王淼, 孟小峰. 差分隐私保护下一种精确挖掘 top-k 频繁模式方法[J]. 计算机研究与发展, 2014, 51(1): 104-114.
ZHANG X J, WANG M, MENG X F. An accurate method for mining top-k frequent pattern under differential privacy[J]. Journal of Computer Research and Development, 2014, 51(1): 104-114.
- [18] CUI P, LIU H, AGGARWAL C, et al. Uncovering and predicting human behaviors[J]. IEEE Intelligent Systems. 2016, 03.
- [19] 张宇翔, 孙菀, 杨家海, 等. 新浪微博反垃圾中特征选择的重要性分析[J]. 通信学报, 2016, 37(8): 2016152-10.
ZHANG Y X, SUN Y, YANG J H, et al. Feature importance analysis for spammer detection in Sina Weibo[J]. Journal of Communications, 2016, 37(8): 2016152-10.
- [20] 曹菁华, 胡访宇, 华焯. 基于电信数据的通话行为模式发现[J]. 计算机仿真, 2014, 11(31): 163-167.
CAO J H, HU F Y, HUA Y. Call behavior pattern discovery based on telecom data[J]. Computer Simulation, 2014, 11(31): 163-167.
- [21] OLSZEWSKI D. A probabilistic approach to fraud detection in telecommunications[J]. Knowledge-Based Systems, 2012, 26(3): 246-258.
- [22] STEINLEY D. Properties of the hubert-arable adjusted rand index[J]. Psychological methods, 2004, 9(3): 386.

[作者简介]



程晓涛(1990-), 男, 河北邢台人, 国家数字交换系统工程技术研究中心博士生, 主要研究方向为电信网安全、网络大数据处理与分析。



吉立新(1969-), 男, 江苏淮安人, 国家数字交换系统工程技术研究中心研究员, 主要研究方向为电信网信息安全。

黄瑞阳(1986-), 男, 福建漳州人, 国家数字交换系统工程技术研究中心助理研究员, 主要研究方向为文本挖掘和图挖掘。

于洪涛(1970-), 男, 辽宁丹东人, 国家数字交换系统工程技术研究中心研究员, 主要研究方向为网络大数据分析处理。

杨奕卓(1994-), 男, 吉林省吉林市人, 国家数字交换系统工程技术研究中心硕士生, 主要研究方向为社会网络分析。