



一种识别骚扰电话的组合算法研究

王彦青¹, 王瀚辰^{1,2}

(1. 东信北邮信息技术有限公司, 北京 100191; 2. 美国特尔菲学校, 美国 俄勒冈州 谢里登 97378)

摘要: 当前骚扰电话层出不穷, 严重影响了人们的日常生活。为了有效防范此类电话带来的不良社会影响, 采用了数据挖掘的分析手段, 深入研究了骚扰电话呼叫特点, 提出了一种基于用户反馈的分时段分析骚扰电话识别方法; 并对用户标识的疑似骚扰号码引入随机森林算法, 极大提高骚扰源识别率, 结合布控拦截机制, 整体实现对骚扰电话全方位的管控。通过实际数据验证, 效果明显。

关键词: 骚扰电话; 随机森林算法; 大数据

中图分类号: TP393

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2017127

Research on a combining algorithm for harassing calls to identify

WANG Yanqing¹, WANG Hanchen^{1,2}

1. Eastcom-BUPT Information Technology Co., Ltd., Beijing 100191, China

2. Delphian School, Sheridan, OR 97378, USA

Abstract: At present, people's daily life has been seriously affected by an endless stream of harassing calls. To prevent the adverse social influence, a harassing calls recognition method based on the analysis of users' feedbacks was proposed, which could make people look insight into the features of harassing calls by data mining. Also, the random forest algorithm was applied to identify the suspected harassment numbers. In this way, the recognition rate of harassment source has been enhanced greatly, and a comprehensive control in harassing calls can be achieved by integrating the portable interceptor. Simulation results also show its good performance.

Key words: harassing call, random forest algorithm, big data

1 引言

随着电话的普及, 各种恶意骚扰诈骗等非法语音电话日益猖獗, 向用户宣传广告、理财、售房等信息^[1], 严重影响了人们的日常生活, 目前虽有治理, 但效果仍不明显。通过抽样调查发现, 80.7%的用户被骚扰电话骚扰过, 其中每天被骚扰1~3次的用户占比约60%, 另外有83.8%的用户

认为日常生活已严重受到影响, 并希望运营商或政府采取措施解决骚扰电话的问题; 从运营商统计数据来看, 骚扰电话举报次数居高不下, 月均超过5万次, 侧面说明不仅给用户的财产造成了损失, 也给运营商的业务及品牌形象造成了不良影响。

针对这一现状, 本文在大数据分析^[2]基础上引入随机森林算法^[3], 分析骚扰电话的行为特征, 最

大限度找出骚扰源,为骚扰电话的治理提供依据。

2 骚扰电话特征及治理思路

2.1 骚扰电话特征

经过全面调查和近两年的研究,骚扰电话有明显的不同于普通电话的特征,总结出有网络环境中骚扰电话存在如下特点。

- 骚扰电话的骚扰目标离散度高,即被骚扰的号码归属地分散,分布于全国各地不同的物理位置。
- 骚扰电话的呼叫频次高,单位时间内骚扰号码发起的试呼次数高,远高于用户平均的呼叫频次。
- 骚扰号码呼出与呼入的比例高,骚扰号码主要用以发起外呼,而作为被叫的次数较少,因此单位时间内骚扰号码呼出次数/呼入次数异常高。
- 被骚扰号码的相关性较小,由于骚扰号码随机地选择被叫发起呼叫,所以被骚扰的被叫用户之间是通话圈的可能性较小,即相关性小。
- 骚扰电话的通话时长较小,而且被叫先挂机的概率大。
- 骚扰号码与所有的被骚扰号码间的通话次数相对较少。

2.2 骚扰电话治理思路

骚扰电话的治理重点是骚扰源的治理,通常包含3个环节,分步实施:首先,通过采集到的用户通信信令数据采用大数据分析挖掘手段^[4]对话单中的疑似骚扰电话进行识别;其次,通过录音、回拨、音频对比等审核确认方法对疑似号码进行确认;最后,疑似骚扰号码经确认后对其进行布控,实施拦截预警。在上述3个环节中,最为关键的是分析识别环节,本文重点研究并阐述了基于大数据分析判断骚扰电话的方法,包含基于用户反馈机制的骚扰电话判别方法与基于随机

森林算法的骚扰电话识别方法;其余两个环节目前已相对成熟,本文仅对取证审核及拦截相应办法做简要陈述。

3 基于大数据分析判断骚扰电话

针对目前治理骚扰电话面临的种种困难,尤其是目前现网所采用的传统业务分析识别骚扰电话方法的不足,具体来说,现网多采用单一维度或少数几个维度组合的方式,并人为设置阈值所生成的识别策略来对骚扰电话进行判断,这种方式受主观人为影响较大,导致刻画的骚扰电话特征存于表面,只要骚扰源稍作变化,就很难识别。所以一直以来现网存在识别号码覆盖较少、准确率低且审核工作量大的问题。

为了有效规避此类问题,本文采用骚扰电话识别的组合方法,从用户建立呼叫时刻入手分析,并全面深层次描述用户呼叫行为特征来分析挖掘。

(1) 从所有的呼叫号码中对疑似骚扰号码进行初步筛选,并尽可能地覆盖。

(2) 对第一步所圈定的疑似号码进行二次分析,进一步确认骚扰号码,以提高准确率:

- 首先,提出一种基于用户反馈的分时段分析确认骚扰电话识别方法,该方法利用被叫用户对主叫号码的反馈标签,初步圈定疑似骚扰号码源;
- 其次,利用随机森林算法进一步分析判断这些疑似号码在不同时间粒度下行为特征与骚扰号码特征的吻合度,确认疑似名单中的号码是否为骚扰源,如果是,加入黑名单。

该方法虽然在开始阶段,对部分用户有不同程度的骚扰,但随着标记号码数据的积累以及大数据分析的确认,最终将使得大部分用户受益不再被骚扰。

3.1 基于用户反馈的疑似骚扰电话识别

通常,正常的通话流程^[5]为:主叫拨打被叫



时, 呼叫信令会首先送到交换机, 再由交换机接续被叫号码, 实现主被叫通话。为了实现将疑似骚扰号码尽可能地标识出来, 本文提出一种在主被叫实现通话过程中增加标签投递系统的方法, 通过用户反馈机制实现骚扰电话的识别。该方法在实验过程中借助某省移动公司通信系统, 在充分验证不影响客户正常通话情况下增加标签投递装置, 从全网呼叫信令中实时反馈标注号码并沉淀。呼叫流程如图 1 所示。

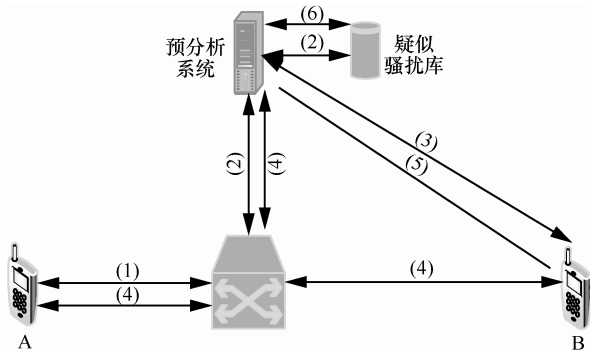


图 1 呼叫流程示意

具体识别方法如下。

(1) 当主叫 A 呼叫被叫 B 时, 信令首先到达交换机, 然后交换机把呼叫信令转送到标签投递系统。

(2) 标签投递系统将从信令中提取主叫号码和被叫号码, 判断主叫号码是否已存在于疑似骚扰库 DB 中, 其中, 疑似骚扰库 DB 中号码来源于第三方所标识号码并不断积累, 如 360、百度等互联网公司的标识号码。

(3) 如果主叫号码为疑似骚扰号码, 则通过技术手段, 比如通过 USSD (unstructured supplementary service data) 方式向用户投递疑似骚扰提醒, 否则, 向被叫用户投递欢迎词之类的内容。

(4) 通过标签投递系统的快速比对确认后, 通知交换机接续被叫用户 B。

(5) 用户 B 接听电话, 如果用户 B 判断 A 为骚扰电话, 在挂机后会将 A 的信息反馈至疑似骚扰库, 则标记号码 A 为骚扰号码。

(6) 同时, 将 A 加入疑似骚扰库, 如 A 已存

在于疑似骚扰库, 则标记次数加 1; 如果否, 则标记为 1, 再次发现 A 后, 标记次数累加。

与传统的互联网公司提供的标注骚扰电话的方式不同, 本文提出的用户反馈骚扰号码的方式是基于通信网络侧实现的, 不需要用户手机客户端做任何改动或者安装专门的客户端。适用门槛低, 适用范围广, 用户受益面更大。

基于此反馈机制, 基本保证疑似骚扰库中的号码为疑似骚扰号码, 但不排除部分误标记, 或者被叫用户对部分主叫电话的内容反感而造成的激情标记行为。为了进一步把疑似骚扰号码确认为骚扰号码, 需根据骚扰电话的特点, 进一步采用大数据分析技术, 二次确认骚扰疑似号码。

3.2 基于随机森林算法的骚扰电话识别

通过第 3.1 节用户反馈机制识别出的疑似骚扰号码集, 可大量覆盖到全网骚扰号码, 并大幅压缩了数据范围, 进一步分析确认了骚扰号码以减少数据处理量, 但号码集内难免有部分正常号码被错误识别, 导致取证审核工作量加大, 增加了拦截管控风险引起用户投诉, 为避免上述情况, 本文采用随机森林挖掘算法针对用户反馈机制办法所识别的疑似骚扰号码集, 提取一段时间内其呼叫信令信息并衍生统计后的特征信息进行二次分析, 再次确认或强化骚扰情况, 在保证高覆盖的条件下提升准确性, 避免对正常号码形成误拦。

基于随机森林算法二次分析挖掘所需样本数据是以用户反馈机制法所识别出的所有疑似骚扰号码为唯一标识, 与该省全量通信信令数据中的主叫号码进行匹配, 并提取所有疑似骚扰号码的 7 天历史通话信令信息作为模型训练所需原始数据源, 主要包含: 主叫号码、被叫号码、通话开始时间、通话时长、振铃时延和释放原因等指标; 在训练样本中, 选择第 3.1 节疑似号码库中标识次数较高的号码和已确认为骚扰电话的号码作为正样本, 其余号码作为待识别样本或负样本, 并进行标识。另外, 为验证模型的准确性及衡量

模型运行稳定性, 按日提取用户反馈机制方法所识别出的最近 7 天疑似骚扰号码集, 分别以日为单位关联每日疑似号码的前 7 天历史通信信令数据作为模型测试所需原始数据源, 共形成 7 个连续的测试样本, 每个训练样本均需剔除已确认的骚扰电话号码和标记次数较高的疑似骚扰电话号码。

3.2.1 用户行为信息预处理

在进行训练样本数据处理时, 根据第 2.1 节所描述的骚扰电话共性, 首次提出时间分段法分析骚扰特性, 进一步分析骚扰特征在不同时间段的规律。具体做法是: 将全天划分多种时间粒度, 比如分为 9 种时间粒度, 分别为: 1 min、5 min、15 min、30 min、60 min、180 min、360 min、720 min、1 440 min。

对于样本数据不足 24 h, 仅保留话单数据起止时间内最长时间粒度以内的所有周期, 如每天只取 10:00—11:10 的数据, 则最长时间粒度只能取 60 min, 那么可用于计算的周期为 1 min、5 min、15 min、30 min、60 min。所有属性计算周期均从本周期的 00 s 开始, 以下周期的前 1 min 第 59 s 结束, 同时将呼叫频次过低而无法统计的其他属性的值置为 0。

分别统计 7 天内各个时间粒度下用户的通话频次:

$$Fre_{ijk}, i=1,2,\dots,7, j=(1,5,15,30,60,180,360,720,1440), k=1,2,\frac{1440}{j} \quad (1)$$

其中, i 表示样本数据的统计天数, j 表示按天划分的时间粒度 (单位: min)。

统计各类时间粒度下通话频次最高的时间 T , 其中:

$$\begin{aligned} T &= (I, j, K) \\ \text{if } Fre &= \max (Fre_{ijk}) \text{ then } I=i, K=k, \\ j &= (1,5,15,30,60,180,360,720,1440) \end{aligned} \quad (2)$$

并对这个时间 T 下主叫的通信行为进行统

计、衍生, 得到 a 个指标, 则 9 种时间粒度共计包含 $a \times 9$ 个指标, 另外, 完整的建模所需用户信息表还应包含用户号码、用户标识, 因此共计有 $(a \times 9 + 2)$ 个指标。

具体的每个时间粒度 T 时间下选取的指标如呼叫频次、被叫个数、通话时长、振铃时长、主动释放次数、被动释放次数、被叫离散度、同一主叫的被叫号码间相关系数、呼叫相同万号段最大频次、主叫占比 (呼出次数/呼出呼入总次数)、呼叫时间间隔标准差 (计算该指标被叫号码需 3 个或 3 个以上) 等。

同理, 采用上述方法对每个测试样本进行处理, 形成模型测试所需信息表。

3.2.2 随机森林算法识别方法

鉴于随机森林分类算法具有能够处理高维属性数据, 无需做特征选择, 训练速度快, 训练过程中可以检测到属性间的相互影响, 可实现并行化, 可输出属性重要程度以及分类概率和预测类别等优点, 选取随机森林分类算法用于进一步确认骚扰电话的识别模型。

随机森林的基本思想是: 首先, 利用有放回的随机抽样方法从原始训练样本 (N 个样本 M 维属性) 中抽取 K 组样本, 且每组样本的样本容量都与原始训练样本相同, 均为 N ; 其次, 随机选取 m 维属性, m 值 $\leq$ 总属性维数 M (其中 $m = \text{int}(2\sqrt{M})$); 然后, 每次对 N 个样本 m 维属性生成一棵决策树, 共建立 K 个决策树模型, 得到 K 种分类结果; 最后, 根据 K 种分类结果对每个记录进行投票表决, 决定其最终分类, 如图 2 所示。

由上述原理可知, 随机森林算法的构造主要有两部分: 一部分是决策树的构建, 形成决策树森林, 本文决策树采用 Gini 不纯度的方法, 不纯度越小, 属性越重要; 另一部分是决策过程, 采用投票表决方式输出最优分类结果。具体步骤如下。

步骤 1 根据训练样本, 采用随机森林算法进行模型训练。假设给定训练样本集为 S , 其中样

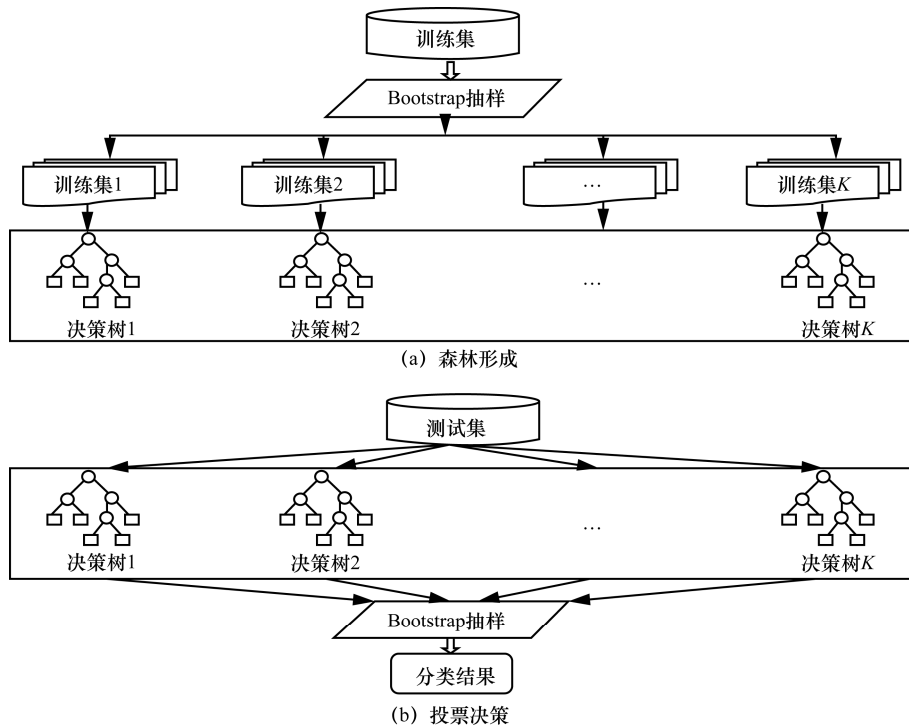


图2 随机森林算法流程

本量为 N (指一个电话号码对应一条记录), 且样本中包含特征维数为 M , 构建 K 棵决策树, 单棵树记为 k 。那么, 每棵树的输入样本采用 Bootstrap 抽样方法^[6]从训练样本集 S 中有放回的抽取大小和 N 一样的 K 个样本, 同时每棵树的输入维度是从 M 维特征中随机提取 $m = \text{int}(2\sqrt{M})$ 个特征为第 k 棵树的特征输入。另外, 需事先设定每棵决策树 k 采用 Gini 不纯度判断标准进行分裂, 并设置分支覆盖的样本的最小个数为 d 。

针对第 k 棵树, $k=1,2,\dots,K$, 设样本集 $s(k)$, m 为第 k 棵树的特征集, 根据 Gini 不纯度最小的原则分别计算 m 维特征的 Gini 不纯度:

$$I_G(j_k) = \frac{|N(Z_1)|}{N} \times \text{Gini}(Z_1) + \frac{|N(Z_2)|}{N} \times \text{Gini}(Z_2) \quad (3)$$

其中, $I_G(j_k)$ 是第 k 棵决策树的第 j 维特征的 Gini 不纯度, Z_1 是根节点的样本集 $s(k)$ 在 j 的特征条件下经最优二叉分隔后所产生的左节点的样本集, Z_2 是根节点的样本集 $s(k)$ 在 j 的特征

条件下经最优二叉分隔后所产生的右节点的样本集, $N(Z_1)$ 、 $N(Z_2)$ 分别是 Z_1 、 Z_2 的样本量, $\text{Gini}(Z_1)$ 、 $\text{Gini}(Z_2)$ 分别是 Z_1 、 Z_2 的 Gini 不纯度, 且 $\text{Gini}(Z_l)$ ($l=1$ 或 $l=2$) 可以进一步表示为:

$$\text{Gini}(Z_1) = 1 - \sum_{i=0}^l f_{kj}^{Z_1}(i), i=0 \text{ 或 } 1 \quad (4)$$

$i=0$ 则表示非骚扰号码, $i=1$ 则表示骚扰号码, $f_{kj}^{Z_1}(i)$ 是第 k 棵决策树中第 j 个特征条件下的 Z_l 对应分支中, 所识别的骚扰号码或非骚扰号码的概率。

经计算, 从 m 维特征中选出 Gini 不纯度最小的一个特征 $\min(I_G(j \in m))$ 作为根节点, 将该节点分裂为左节点和右节点, 然后以根节点作为限制条件, 再次计算 m 个指标的 Gini 不纯度, 按其值的大小决定子节点指标, 继续向下生成, 以此类推, 形成决策树; 如果某一分支覆盖的样本的个数 d 小于一个设定阈值, 则设置当前节点为叶子节点, 即该节点停止生长, 转而继续训练其他节点, 直到所有节点都训练过了或者被标记为叶子节点, 则该棵树训练完毕。

重复上述方法, 直至 $k=K$, 所有分类树训练完毕, 即最终生成 K 棵决策树, 每棵决策树会包含自身的所有骚扰电话识别规则以及每条规则所对应的骚扰电话识别概率。具体每棵树中, 计算其每个叶子节点对待识别号码所属类别的预测概率:

$$p_k^r(i) = \frac{N_k^r(i)}{N_k^r}, \text{ 其中, } i=0 \text{ 或 } 1, i=0 \text{ 则表示为}$$

非骚扰号码, $i=1$ 则表示为骚扰号码, $p_k^r(i)$ 是第 k 棵决策树的第 r 个叶子节点对待识别号码属于第 i 个类别的预测概率, $N_k^r(i)$ 是第 k 棵决策树的第 r 个叶子节点中属于第 i 个类别的号码数, N_k^r 是第 k 棵决策树的第 r 个叶子节点包含的号码总数。

步骤 2 根据测试样本, 从测试样本库中无放回地提取任意一个主叫号码的记录 w 。

步骤 3 将步骤 2 中提取出的任意一个主叫号码 w 的所有特征作为输入, 利用随机森林算法生成的 K 棵树中所包含的规则对该主叫号码 w 进行判定, 如果 K 棵树中判定为骚扰电话的 K 个概率之和大于判定为非骚扰电话的 K 个概率之和, 则判定结果为骚扰电话, 则进行步骤 4; 如果否, 则转向步骤 2, 直至提取完所有主叫号码; 详细判别方式表示为:

$$\begin{cases} \sum_{k=1}^K p(i=0) > \sum_{k=1}^K p(i=1), w \in \{0\} \\ \sum_{k=1}^K p(i=1) > \sum_{k=1}^K p(i=0), w \in \{1\} \end{cases} \quad (5)$$

其中, $w \in \{0\}$ 表示步骤 2 中任取电话号码 w 归属类别为非骚扰; $w \in \{1\}$ 表示步骤 2 中任取电

话号码 w 归属类别为骚扰; $\sum_{k=1}^K p(i=1)$ 代表所有分

类树中类别为 1 的预测概率总和; $\sum_{k=1}^K p(i=0)$ 代表

所有分类树中类别为 0 的预测概率总和。

输出的电话号码归属类别为所有分类树中预测概率总和最大的那一个类。

步骤 4 通过步骤 3 将判定为骚扰号码的主叫号码 w 写入骚扰号码表中, 并判断是否已从话单记录中提取完所有主叫号码, 如果是, 则模型识别流程结束; 如果否, 则转向步骤 2。

3.3 模型验证

通过将随机森林算法模型得到的规则集分别作用于 7 个连续的测试样本后, 对输出结果进行了分析评估, 日均约可识别出 28 万个高度疑似骚扰号码, 整体模型稳定性较强; 在模型准确性评估时, 将识别出的高度疑似骚扰号码采用与第三方标识号码匹配、回拨确认的方式初步进行核验, 待该类号码再次进行通话时采用监听录音的办法来最终确认, 同时与白名单号码进行匹配, 查验模型错误识别率, 如图 3 所示。对比验证结果见表 1。

由表 1 中实验结果可知, 基于用户反馈机制方法和随机森林算法结合的二次骚扰识别防范办法, 与现网传统的业务分析办法相比, 无论从号码覆盖量还是准确率上均有大幅度的提高。本文模型所识别各类号码的查准率均达到 80% 以上, 较现网传统业务分析识别方法提高了近 30%, 同

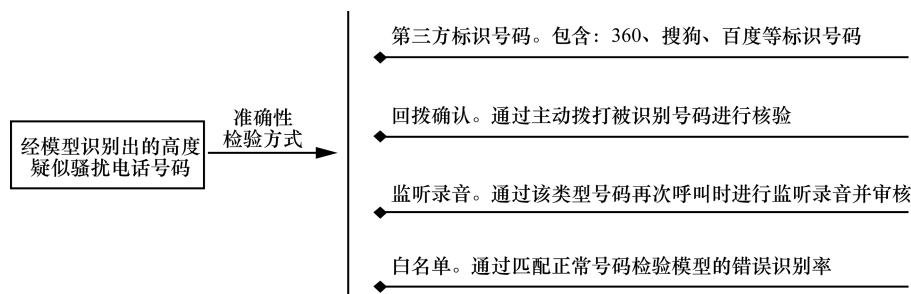


图 3 验证对比号码构成



表 1 对比验证结果

验证指标	省内群呼	省内响一声	网间群呼		网间响一声	
	未剔除短号	剔除短号	未剔除短号	剔除短号	未剔除短号	剔除短号
高度疑似号码量	10 471	5 713	250 001	151 660	134 100	116 529
总计确认号码个数(第三方标识匹配量+回拨确认+监听录音)	8 682	4 773	172 921	127 837	111 338	98 247
模型准确率	82.90%	83.50%	69.20%	84.30%	83.00%	84.30%
现网已有规则当时段确认量(当时段指与测试样本同一时间周期)	920	750	1 158	1 158	2 162	2 162
提升倍数	9.4	6.4	149.3	110.4	51.5	45.4
与白名单匹配后识别错误量	79	50	176	176	53	53
错误识别率	0.75%	0.88%	0.07%	0.12%	0.04%	0.05%

说明: 短号是指主叫号码位数在 6 位以下的号码; 因部分短号是由骚扰源改号后进行呼叫, 具有多变的特点, 无法进行回拨验证且即使确认后也很难进行拦截处置, 故结合实际拦截考虑在验证中剔除短号。现网已有规则是指通过传统的业务分析办法所得。

时误报率低于 1%, 从覆盖号码量来看, 本文模型所覆盖号码均达上万个, 而现网已有的方法所覆盖号码仅为 2 000 个左右, 通过本文方法较现网已有方法提升倍数达 5 倍以上; 总之, 本文模型很大程度上可对话单中骚扰号码进行识别, 在保证一定准确率的前提下, 覆盖骚扰号码更广, 并能有效避免误拦导致的用户投诉, 利于综合治理。

4 取证审核与拦截管控

4.1 取证审核

取证审核是将通过大数据分析判断得出的号码进行取证和审核确认, 审核分成人工审核和自动审核两种。

人工审核是业务开展前期的一种审核手段, 骚扰电话通过大数据分析得出的高疑似号码, 由客服人员对内容进行审核确认, 确认为骚扰号码的进入拦截管控环节。人工审核准确率高, 但是效率相对较低。

自动审核功能采用频谱分析比对技术, 目前市场上也已经出现基于语音识别的骚扰电话呼叫检测方式^[7], 速度快、效率高, 是人工审核的必要补充。频谱分析比对技术是使用基于能量和时间双维度语音对比的核心技术, 将疑似号码的频谱与样本库进行比对, 当某号码的全部频谱特征都与样本库成功匹配时, 将该号码视为确定的骚扰号码, 后续

该号码的所有来话由拦截模块进行拦截管控。利用频谱分析比对技术进行自动审核, 大大提高了审核的周期和效率, 放大了人工审核的价值。

4.2 拦截管控

拦截管控^[8]是骚扰电话治理的关键环节, 采用直接治理和间接治理两种。

直接治理的思路是将骚扰电话直接拦截, 比如可按网元拦截、号首+号长拦截和 DTMF (dual tone multi frequency, 双音多频信号, 即被叫的按键行为) 拦截等多种拦截手段。

间接治理是放行转接功能, 放行转接是对部分骚扰号码来电, 路由到独立放音设备, 对主叫放音。一方面麻痹主叫侧的犯罪分子, 以减少主叫频繁换号给系统增加的追踪难度, 另一方面通过接通骚扰来电, 对骚扰通话进一步分析, 发现更多可疑特征。

信息提醒及反馈包括省内移动用户在接听电话过程中进行事中提醒及挂机后给被叫用户下发短信, 鼓励用户进行众包举报。

5 结束语

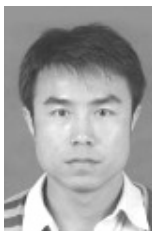
本文利用大数据分析技术, 先后完成了用户反馈机制方法的实践以及基于随机森林算法挖掘模型的构建, 并采用嵌套组合方式对骚扰电话进行识别, 经评估验证取得了良好效果, 已具备现

网实施部署可行性。然而,随着骚扰电话治理的不断深入,骚扰电话的剧本和技术手段也可能不断更新,治理的难度也将逐渐增加,所以仍需要在今后研究工作中持续改进和优化,继续深入探索研究以“业务+大数据”为方向的骚扰电话联动分析识别能力,“由点及面、逐步完善”,逐步扩大覆盖范围,持续优化应用效果。同时,骚扰电话的全面治理还需集合各方力量,多部门协同作战,以政府为主导落实治理责任,从呼叫源头拒绝骚扰业务接入,全方位保护人民群众利益。

参考文献:

- [1] 杜磊. 骚扰电话何时休[J]. 计算机与网络, 2015, 41(3): 1.
DU L. Harassing phone calls[J]. Computer and Network, 2015, 41(3): 1.
- [2] 漆晨曦. 立足小数据基础的电信企业大数据分析应用发展策略[J]. 电信科学, 2014, 30(10): 15-20.
QI C X. Analysis and application development strategy of telecom enterprise big data based on small data [J]. Telecommunications Science, 2014, 30(10): 15-20.
- [3] 董师师, 黄哲学. 随机森林理论浅析[J]. 集成技术, 2013(1): 1-7.
DONG S S, HUANG Z X. A brief theoretical overview of random forests[J]. Journal of Integration Technology, 2013(1): 1-7.
- [4] 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. Discussion on the development strategy of telecom operators[J]. Telecommunications Science, 2013, 29(3): 6-11.
- [5] 卢卫, 陆希玉. 4G时代移动互联网的发展趋势[J]. 电信科学, 2014, 30(5): 51-54.
LU W, LU X Y. Mobile internet trends in 4G era [J]. Telecommunications Science, 2014, 30(5): 51-54.
- [6] 张艳芳, 黄群. 基于 Bootstrap 方法的样本相等性检验[J]. 防灾科技学院学报, 2008, 10(1): 95-98.
ZHANG Y F, HUANG Q. Tests based on bootstrap method for the equality of distribution[J]. Journal of Institute of Disaster- prevention Science and Technology, 2008, 10(1): 95-98.
- [7] 杜海涛, 张峰, 高曼颖, 等. 基于语音识别的骚扰电话呼叫检测技术分析[J]. 电信工程技术与标准化, 2014(12): 5-9.
DU H T, ZHANG F, GAO M Y, et al. Analysis and experiment of disturbing call detection technology based on speech recognition[J]. Telecom Engineering Technics and Standardization, 2014(12): 5-9.
- [8] 胡坤, 刘镒, 刘明辉, 等. 大数据的安全理解及应对策略研究[J]. 电信科学, 2014, 30(2): 112-117, 122.
HU K, LIU D, LIU M H, et al. Research on security connotation and response strategies for big data [J]. Telecommunications Science, 2014, 30(2): 112-117, 122.

作者简介



王彦青, 男, 东信北邮信息技术有限公司数据分析师, 主要从事算法研究方面的工作。



王瀚辰, 男, 现就职于东信北邮信息技术有限公司。