(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 108833720 B
(45)授权公告日 2021.11.30

(21)申请号 201810421601.8
(22)申请日 2018.05.04
(65)同一申请的已公布的文献号
申请公布号 CN 108833720 A
(43)申请公布日 2018.11.16
(73)专利权人 北京邮电大学
地址 100876 北京市海淀区西土城路10号北京邮电大学
(72)发明人 双锴 薛二彭 姚云腾 苏森
(74)专利代理机构 北京路浩知识产权代理有限公司 11002
代理人 王莹 吴欢燕
(51)Int.Cl.
H04W 12/128(2021.01)
H04M 3/22(2006.01)

审查员 廖薇

权利要求书3页 说明书8页 附图2页

(54)发明名称

诈骗电话号码识别方法与系统

(57)摘要

本发明实施例提供一种诈骗电话号码识别方法与系统，识别方法包括：对通话记录进行特征提取，获取提取特征后的通话记录，所述提取特征后的通话记录包括空号比例、好友通话比例和单位时间内电话号码的熵；将所述提取特征后的通话记录输入训练好的诈骗电话识别模型中，获取所述通话记录对应的诈骗电话识别结果，其中，所述诈骗电话识别模型通过梯度提升决策树算法训练。本发明实施例提供的诈骗电话号码识别方法与系统，能够对电话号码是否为诈骗电话号码进行更细致准确的识别；通过设置通过梯度提升决策树算法训练模型，能够对诈骗电话号码识别得更加精准。

CN 108833720 B

1.一种诈骗电话号码识别方法,其特征在于,包括:

对通话记录进行特征提取,获取提取特征后的通话记录,所述提取特征后的通话记录包括空号比例、好友通话比例和单位时间内电话号码的熵;

所述对通话记录进行特征提取,获取提取特征后的通话记录,具体包括:

获取所述通话记录对应的多个通话记录初始特征,其中,所述多个通话记录初始特征包括单位时间呼叫次数、单位时间通话为长话次数、单位时间通话为市话次数、单位时间内长话比例、单位时间被叫号码为空号次数、单位时间内被叫号码为空号比例、单位时间被叫用户直接挂断次数、单位时间被叫用户直接挂断比例、单位时间电话接通后被挂断次数、单位时间电话接通后被挂断比例、单位时间接通电话次数、单位时间呼叫的号码有回拨的次数、日最早通话时间、日最晚通话时间、单位时间内平均拨打电话时间间隔、单位时间被叫用户所属多少不同的地域、单位时间被叫用户号码熵值以及单位时间内被叫号码中好友比例中的一项或多项;

通过主成分分析法或者线型判别分析法,获取每一通话记录初始特征的重要性值;

按照重要性值由高到低的优先级,基于预设的特征个数值,获取所述提取特征后的通话记录,所述提取特征后的通话记录包括空号比例、好友通话比例和单位时间内电话号码的熵;

将所述提取特征后的通话记录输入训练好的诈骗电话识别模型中,获取所述通话记录对应的诈骗电话识别结果,其中,所述诈骗电话识别模型通过梯度提升决策树算法训练;

其中,所述好友为主叫用户与被叫用户双方互通电话大于预设次数的号码。

2.根据权利要求1所述的识别方法,其特征在于,对通话记录进行特征提取,获取提取特征后的通话记录,所述提取特征后的通话记录包括空号比例、好友通话比例和单位时间内电话号码的熵,之前还包括:

对所述通话记录进行数据处理,获取数据处理后的通话记录,所述数据处理包括归一化处理、去平均值处理、离散化处理、one-hot编码转化、词袋模型处理、n-gram处理、TF-IDF处理以及词向量转化处理中的至少一种;

对应的,对通话记录进行特征提取,获取提取特征后的通话记录,所述提取特征后的通话记录包括空号比例、好友通话比例和单位时间内电话号码的熵,具体包括:

对所述处理后的通话记录进行特征提取,获取提取特征后的通话记录,所述提取特征后的通话记录包括空号比例、好友通话比例和单位时间内电话号码的熵。

3.根据权利要求1所述的识别方法,其特征在于,所述诈骗电话识别模型通过下述步骤训练:

将电信网络中的原始训练通话记录进行预处理,获取训练通话记录,所述预处理包括数据清洗和数据采样;

对所述训练通话记录进行特征提取,获取提取特征后的训练通话记录,所述提取特征后的训练通话记录包括空号比例、好友通话比例和单位时间内电话号码的熵;

将所述提取特征后的训练通话记录输入未训练的诈骗电话识别模型中,基于梯度提升决策树算法,获取所述训练好的诈骗电话识别模型。

4.根据权利要求3所述的识别方法,其特征在于,所述将电信网络中的原始训练通话记录进行预处理,获取训练通话记录,所述预处理包括数据清洗和数据采样,以及,所述对所

述训练通话记录进行特征提取，获取提取特征后的训练通话记录，所述提取特征后的训练通话记录包括空号比例、好友通话比例和单位时间内电话号码的熵，之间还包括：

对所述训练通话记录进行数据处理，获取数据处理后的训练通话记录，所述数据处理包括归一化处理、去平均值处理、离散化处理、one-hot编码转化、词袋模型处理、n-gram处理、TF-IDF处理以及词向量转化处理中的至少一种；

对应的，对所述训练通话记录进行特征提取，获取提取特征后的训练通话记录，所述提取特征后的训练通话记录包括空号比例、好友通话比例和单位时间内电话号码的熵，具体包括：

对所述处理后的训练通话记录进行特征提取，获取提取特征后的训练通话记录，所述提取特征后的训练通话记录包括空号比例、好友通话比例和单位时间内电话号码的熵。

5.根据权利要求3所述的识别方法，其特征在于，所述对所述训练通话记录进行特征提取，获取提取特征后的训练通话记录，所述提取特征后的训练通话记录包括空号比例、好友通话比例和单位时间内电话号码的熵，具体包括：

获取所述训练通话记录对应的多个训练通话记录初始特征，其中，所述多个训练通话记录初始特征包括单位时间呼叫次数、单位时间通话为长话次数、单位时间通话为市话次数、单位时间内长话比例、单位时间被叫号码为空号次数、单位时间内被叫号码为空号比例、单位时间被叫用户直接挂断次数、单位时间被叫用户直接挂断比例、单位时间电话接通后被挂断次数、单位时间电话接通后被挂断比例、单位时间接通电话次数、单位时间呼叫的号码有回拨的次数、日最早通话时间、日最晚通话时间、单位时间内平均拨打电话时间间隔、单位时间被叫用户所属多少不同的地域、单位时间被叫用户号码熵值以及单位时间内被叫号码中好友比例中的一项或多项；

通过主成分分析法或者线型判别分析法，获取每一训练通话记录初始特征的重要性值；

按照重要性值由高到低的优先级，基于预设的特征个数值，获取所述提取特征后的训练通话记录，所述提取特征后的训练通话记录包括空号比例、好友通话比例和单位时间内电话号码的熵。

6.根据权利要求3所述的识别方法，其特征在于，所述将所述提取特征后的训练通话记录输入未训练的诈骗电话识别模型中，基于梯度提升决策树算法，获取所述训练好的诈骗电话识别模型，具体包括：

将所述提取特征后的训练通话记录输入未训练的诈骗电话识别模型中，生成决策树；

获取所述决策树的差值，并判断所述差值是否达到循环条件；

若所述差值达到循环条件，则所述训练好的诈骗电话识别模型获取完成，若所述差值未达到循环条件，则重复执行上述步骤，直至差值达到循环条件。

7.一种诈骗电话号码识别系统，其特征在于，包括：

特征获取模块，用于对通话记录进行特征提取，获取提取特征后的通话记录，所述提取特征后的通话记录包括空号比例、好友通话比例和单位时间内电话号码的熵；所述对通话记录进行特征提取，获取提取特征后的通话记录，具体包括：获取所述通话记录对应的多个通话记录初始特征，其中，所述多个通话记录初始特征包括单位时间呼叫次数、单位时间通话为长话次数、单位时间通话为市话次数、单位时间内长话比例、单位时间被叫号码为空号

次数、单位时间内被叫号码为空号比例、单位时间被叫用户直接挂断次数、单位时间被叫用户直接挂断比例、单位时间电话接通后被挂断次数、单位时间电话接通后被挂断比例、单位时间接通电话次数、单位时间呼叫的号码有回拨的次数、日最早通话时间、日最晚通话时间、单位时间内平均拨打电话时间间隔、单位时间被叫用户所属多少不同的地域、单位时间被叫用户号码熵值以及单位时间内被叫号码中好友比例中的一项或多项;通过主成分分析法或者线型判别分析法,获取每一通话记录初始特征的重要性值;按照重要性值由高到低的优先级,基于预设的特征个数值,获取所述提取特征后的通话记录,所述提取特征后的通话记录包括空号比例、好友通话比例和单位时间内电话号码的熵;

获取结果模块,用于将所述提取特征后的通话记录输入训练好的诈骗电话识别模型中,获取所述通话记录对应的诈骗电话识别结果,其中,所述诈骗电话识别模型通过梯度提升决策树算法训练;

其中,所述好友为主叫用户与被叫用户双方互通电话大于预设次数的号码。

8.一种诈骗电话号码识别设备,其特征在于,包括处理器、通信接口、存储器和总线,其中,处理器,通信接口,存储器通过总线完成相互间的通信,处理器可以调用存储器中的逻辑指令,以执行如权利要求1至6任一所述的识别方法。

9.一种计算机可读存储介质,其上存储有计算机程序,其特征在于,该计算机程序被处理器执行时实现如权利要求1至6任一所述的识别方法。

诈骗电话号码识别方法与系统

技术领域

[0001] 本发明实施例涉及通信网络安全领域，更具体地，涉及一种诈骗电话号码识别方法与系统。

背景技术

[0002] 近年来利用电话进行诈骗以及传播暴力、恐怖、邪教等信息的实践呈爆发趋势，不良呼叫已经成为广大群众的痛点和社会顽疾。2013年，因不良呼叫引起的相关案件被骗金额达到了百亿元；2014年，因不良呼叫引发的案件在前一年基础上再次大大增加；2015年，全国诈骗电话报案59.9余万起，被骗金额在上一年基础上增长一倍，各类由不良呼叫引起的案件触目惊心。这些损失的背后影射出了通讯信息诈骗黑产，整个黑产已经形成了分工明确、运作高效的产业链，从防范效果来看，仅仅靠电信基础企业，管局或公安的力量，无法达到有效的打击目的，需要联合各方资源，一起发力。

[0003] 电话沟通是连接亲人、朋友、同事、合作伙伴的重要手段，是网络空间的重要组成部分。不良呼叫不仅仅是造成了通讯网络的乌烟瘴气，生态恶化，还给人民群众造成了巨大财产损失乃至生命安全。

[0004] 目前对于诈骗电话的获取识别主要是通过语音识别，这种方式的学习性不强，获取识别不够准确。

发明内容

[0005] 针对现有技术存在的问题，本发明实施例提供一种诈骗电话号码识别方法与系统。

[0006] 本发明实施例提供一种诈骗电话号码识别方法，包括：对通话记录进行特征提取，获取提取特征后的通话记录，所述提取特征后的通话记录包括空号比例、好友通话比例和单位时间内电话号码的熵；将所述提取特征后的通话记录输入训练好的诈骗电话识别模型中，获取所述通话记录对应的诈骗电话识别结果，其中，所述诈骗电话识别模型通过梯度提升决策树算法训练。

[0007] 本发明实施例提供一种诈骗电话号码识别系统，包括：特征获取模块，用于对通话记录进行特征提取，获取提取特征后的通话记录，所述提取特征后的通话记录包括空号比例、好友通话比例和单位时间内电话号码的熵；获取结果模块，用于将所述提取特征后的通话记录输入训练好的诈骗电话识别模型中，获取所述通话记录对应的诈骗电话识别结果，其中，所述诈骗电话识别模型通过梯度提升决策树算法训练。

[0008] 本发明实施例提供一种诈骗电话号码识别设备，包括处理器、通信接口、存储器和总线，其中，处理器，通信接口，存储器通过总线完成相互间的通信，处理器可以调用存储器中的逻辑指令，以执行上述识别方法。

[0009] 本发明实施例提供一种计算机可读存储介质，该计算机程序被处理器执行时实现上述识别方法。

[0010] 本发明实施例提供的诈骗电话号码识别方法与系统，能够对电话号码是否为诈骗电话号码进行更细致准确的识别；通过设置通过梯度提升决策树算法训练模型，能够对诈骗电话号码识别得更加精准。

附图说明

[0011] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作一简单地介绍，显而易见地，下面描述中的附图是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0012] 图1为本发明诈骗电话号码识别方法实施例流程图；

[0013] 图2为本发明实施例中的诈骗电话识别模型训练流程图；

[0014] 图3为本发明诈骗电话号码识别系统实施例模块图；

[0015] 图4为本发明实施例的一种诈骗电话号码识别设备的结构示意图。

具体实施方式

[0016] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0017] 图1为本发明诈骗电话号码识别方法实施例流程图，如图1所示，包括：S1、对通话记录进行特征提取，获取提取特征后的通话记录，所述提取特征后的通话记录包括空号比例、好友通话比例和单位时间内电话号码的熵；S2、将所述提取特征后的通话记录输入训练好的诈骗电话识别模型中，获取所述通话记录对应的诈骗电话识别结果，其中，所述诈骗电话识别模型通过梯度提升决策树算法训练。

[0018] 具体地，步骤S1中的提取特征后的通话记录是诈骗电话识别模型的输入量。诈骗电话识别结果包括通话记录对应的电话号码是诈骗电话号码，或者，通话记录对应的电话号码不是诈骗电话号码。

[0019] 需要说明的是，提取特征后的通话记录优选包括空号比例、好友通话比例和单位时间内电话号码的熵，本发明还保护包括其它提取特征后的通话记录的方案。

[0020] 本发明实施例提供的诈骗电话号码识别方法，能够对电话号码是否为诈骗电话号码进行更细致准确的识别；通过设置通过梯度提升决策树算法训练模型，能够对诈骗电话号码识别得更加精准。

[0021] 基于上述实施例，所述对通话记录进行特征提取，获取提取特征后的通话记录，之前还包括：对所述通话记录进行数据处理，获取数据处理后的通话记录，所述数据处理包括归一化处理、去平均值处理、离散化处理、one-hot编码转化、词袋模型处理、n-gram处理、TF-IDF处理以及词向量转化处理中的至少一种；对应的，对所述通话记录进行特征提取，获取提取特征后的通话记录，所述提取特征后的通话记录包括空号比例、好友通话比例和单位时间内电话号码的熵，具体包括：对所述处理后的通话记录进行特征提取，获取提取特征后的通话记录，所述提取特征后的通话记录包括空号比例、好友通话比例和单位时间内电

话号码的熵。

[0022] 具体地，归一化处理、去平均值处理和离散化处理针对的是数值型特征。one-hot编码转化针对的是类别型特征，类别型特征一般为文本信息，比如年龄阶段分为少年、中年、老年。词袋模型处理、n-gram处理、TF-IDF处理以及词向量转化处理针对的是文本型特征。

[0023] 本发明实施例提供的诈骗电话号码识别方法，通过设置数据处理，能够便于后续的特征提取处理。

[0024] 作为一个优选实施例，对所述通话记录进行数据处理，获取数据处理后的通话记录，所述数据处理包括归一化处理、去平均值处理、离散化处理、one-hot编码转化、词袋模型处理、n-gram处理、TF-IDF处理以及词向量转化处理中的任一种，之前还包括：将原始通话记录进行预处理，获取所述通话记录，所述预处理包括数据清洗和数据采样。

[0025] 具体地，预处理的目的是便于下面的特征提取处理。

[0026] 进一步地，数据采样是指，数据正负样本往往不够均衡，通过对数据进行采样，使正负样本基本一致。

[0027] 本发明实施例提供的诈骗电话号码识别方法，通过设置预处理，能够便于后续的特征提取处理。

[0028] 基于上述实施例，所述对通话记录进行特征提取，获取提取特征后的通话记录，具体包括：获取所述通话记录对应的多个通话记录初始特征，其中，所述多个通话记录初始特征包括单位时间呼叫次数、单位时间通话为长话次数、单位时间通话为市话次数、单位时间内长话比例、单位时间被叫号码为空号次数、单位时间内被叫号码为空号比例、单位时间被叫用户直接挂断次数、单位时间被叫用户直接挂断比例、单位时间电话接通后被挂断次数、单位时间电话接通后被挂断比例、单位时间接通电话次数、单位时间呼叫的号码有回拨的次数、日最早通话时间、日最晚通话时间、单位时间内平均拨打电话时间间隔、单位时间被叫用户所属多少不同的地域、单位时间被叫用户号码熵值以及单位时间内被叫号码中好友比例中的一项或多项；通过主成分分析法或者线型判别分析法，获取每一通话记录初始特征的重要性值；按照重要性值由高到低的优先级，基于预设的特征个数值，获取所述提取特征后的通话记录，所述提取特征后的通话记录包括空号比例、好友通话比例和单位时间内电话号码的熵。

[0029] 具体地，本步骤是选择对于诈骗电话号码识别结果影响较大的通话记录初始特征作为提取特征后的通话记录，其个数是预设的。

[0030] 进一步地，被叫用户号码熵值代表了某用户拨打电话号码的不确定性。不良呼叫具有拨打电话都为陌生号码，不确定性很大等特点。空号比例定义为拨打号码为空号的电话在所有拨打电话中的占比。被叫用户直接挂断比例定义为被叫用户没有接通电话直接将电话挂断的数量占总呼叫次数的比例。被叫号码好友比例，定义为被叫号码为主叫号码好友在通话中的比例。其中，好友定义为双方互通电话三次以上的号码，为用户的好友。

[0031] 具体地，主成分分析也称主分量分析，旨在利用降维的思想，把多指标转化为少数几个综合指标，其中每个主成分都能够反映原始变量的大部分信息，且所含信息互不重复。这种方法在引进多方面变量的同时将复杂因素归结为几个主成分，使问题简单化，同时得到的结果更加科学有效的数据信息。在实际问题研究中，为了全面、系统地分析问题，必须

考虑众多影响因素。这些涉及的因素一般称为指标，在多元统计分析中也称为变量。因为每个变量都在不同程度上反映了所研究问题的某些信息，并且指标之间彼此有一定的相关性，因而所得的统计数据反映的信息在一定程度上有重叠。

[0032] 进一步地，判别分析又称“分辨法”，是在分类确定的条件下，根据某一研究对象的各种特征值判别其类型归属问题的一种多变量统计分析方法。其基本原理是按照一定的判别准则，建立一个或多个判别函数，用研究对象的大量资料确定判别函数中的待定系数，并计算判别指标。据此即可确定某一样本属于何类。

[0033] 本发明实施例提供的诈骗电话号码识别方法，通过设置获取重要性较高的若干个特征作为提取特征后的通话记录，能够对诈骗电话号码识别得更加精准。

[0034] 需要说明的是，本发明实施例中的若干个指一个或者多个。

[0035] 作为一个优选实施例，所述数据清洗，具体包括数据清洗、数据空值处理和错误数据处理。

[0036] 具体地，下面来介绍特征工程的概念，特征工程就是利用相关知识与数据处理技巧，从海量数据中发现特征，使特征能够在模型建立以及结果预测中发挥作用的过程。好的特征工程，能够找出好的特征，发现隐藏在数据中的潜在规律，减小噪声的影响，从而使建立的模型达到很好的效果。

[0037] 特征工程的步骤一般分为数据清洗与采样，特征处理，特征选择三个过程。在本发明实施例中，具体可包括：

[0038] 第一，将原始通话记录进行预处理，获取通话记录，所述预处理包括数据清洗和数据采样。第二，对所述通话记录进行数据处理，获取数据处理后的通话记录，所述数据处理包括归一化处理、去平均值处理、离散化处理、one-hot编码转化、词袋模型处理、n-gram处理、TF-IDF处理以及词向量转化处理中的任一种。第三，对所述处理后的通话记录进行特征提取，获取提取特征后的通话记录。

[0039] 本发明实施例提供的诈骗电话号码识别方法，通过设置数据清洗，能够便于后续的特征提取处理。

[0040] 基于上述实施例，所述诈骗电话识别模型通过下述步骤训练：将电信网络中的原始训练通话记录进行预处理，获取训练通话记录，所述预处理包括数据清洗和数据采样；对所述训练通话记录进行特征提取，获取提取特征后的训练通话记录，所述提取特征后的训练通话记录包括空号比例、好友通话比例和单位时间内电话号码的熵；将所述提取特征后的训练通话记录输入未训练的诈骗电话识别模型中，基于梯度提升决策树算法，获取所述训练好的诈骗电话识别模型。

[0041] 具体地，提取特征后的训练通话记录是诈骗电话识别模型的输入量。

[0042] 进一步地，数据采样是指，数据正负样本往往不够均衡，通过对数据进行采样，使正负样本基本一致。数据清洗具体包括数据清洗、数据空值处理和错误数据处理。

[0043] 进一步地，梯度提升决策树算法是机器学习领域的一个分类模型，该模型由多棵决策树通过迭代生成。该模型具有较强的泛华能力。该模型在训练过程中，每一步都是利用已有模型与目标之间的差进行迭代，利用这个差值生成一棵新的决策树加入到模型中。重复上述过程，直到达到结束条件。

[0044] 本发明实施例提供的诈骗电话号码识别方法，通过设置基于梯度提升决策树算法

训练模型，能够使得训练好的诈骗电话识别模型对诈骗电话号码识别得更加精准。

[0045] 基于上述实施例，所述将电信网络中的原始训练通话记录进行预处理，获取训练通话记录，所述预处理包括数据清洗和数据采样，以及，所述对所述训练通话记录进行特征提取，获取提取特征后的训练通话记录，所述提取特征后的训练通话记录包括空号比例、好友通话比例和单位时间内电话号码的熵，之间还包括：对所述训练通话记录进行数据处理，获取数据处理后的训练通话记录，所述数据处理包括归一化处理、去平均值处理、离散化处理、one-hot编码转化、词袋模型处理、n-gram处理、TF-IDF处理以及词向量转化处理中的至少一种；对应的，对所述训练通话记录进行特征提取，获取提取特征后的训练通话记录，所述提取特征后的训练通话记录包括空号比例、好友通话比例和单位时间内电话号码的熵，具体包括：对所述处理后的训练通话记录进行特征提取，获取提取特征后的训练通话记录，所述提取特征后的训练通话记录包括空号比例、好友通话比例和单位时间内电话号码的熵。

[0046] 具体地，归一化处理、去平均值处理和离散化处理针对的是数值型特征。one-hot编码转化针对的是类别型特征，类别型特征一般为文本信息，比如年龄阶段分为少年、中年、老年。词袋模型处理、n-gram处理、TF-IDF处理以及词向量转化处理针对的是文本型特征。

[0047] 本发明实施例提供的诈骗电话号码识别方法，通过设置数据处理，能够便于后续的特征提取处理。

[0048] 基于上述实施例，所述对所述训练通话记录进行特征提取，获取提取特征后的训练通话记录，所述提取特征后的训练通话记录包括空号比例、好友通话比例和单位时间内电话号码的熵，具体包括：获取所述训练通话记录对应的多个训练通话记录初始特征，其中，所述多个训练通话记录初始特征包括单位时间呼叫次数、单位时间通话为长话次数、单位时间通话为市话次数、单位时间内长话比例、单位时间被叫号码为空号次数、单位时间内被叫号码为空号比例、单位时间被叫用户直接挂断次数、单位时间被叫用户直接挂断比例、单位时间电话接通后被挂断次数、单位时间电话接通后被挂断比例、单位时间接通电话次数、单位时间呼叫的号码有回拨的次数、日最早通话时间、日最晚通话时间、单位时间内平均拨打电话时间间隔、单位时间被叫用户所属多少不同的地域、单位时间被叫用户号码熵值以及单位时间内被叫号码中好友比例中的一项或多项；通过主成分分析法或者线型判别分析法，获取每一训练通话记录初始特征的重要性值；按照重要性值由高到低的优先级，基于预设的特征个数值，获取所述提取特征后的训练通话记录，所述提取特征后的训练通话记录包括空号比例、好友通话比例和单位时间内电话号码的熵。

[0049] 具体地，本步骤是选择对于诈骗电话号码识别结果影响较大的训练通话记录初始特征作为提取特征后的训练通话记录，其个数是预设的。

[0050] 具体地，主成分分析也称主分量分析，旨在利用降维的思想，把多指标转化为少数几个综合指标，其中每个主成分都能够反映原始变量的大部分信息，且所含信息互不重复。这种方法在引进多方面变量的同时将复杂因素归结为几个主成分，使问题简单化，同时得到的结果更加科学有效的数据信息。在实际问题研究中，为了全面、系统地分析问题，必须考虑众多影响因素。这些涉及的因素一般称为指标，在多元统计分析中也称为变量。因为每个变量都在不同程度上反映了所研究问题的某些信息，并且指标之间彼此有一定的相关

性,因而所得的统计数据反映的信息在一定程度上有重叠。

[0051] 进一步地,判别分析又称“分辨法”,是在分类确定的条件下,根据某一研究对象的各种特征值判别其类型归属问题的一种多变量统计分析方法。其基本原理是按照一定的判别准则,建立一个或多个判别函数,用研究对象的大量资料确定判别函数中的待定系数,并计算判别指标。据此即可确定某一样本属于何类。

[0052] 需要说明的是,被叫用户号码熵值代表了某用户拨打电话号码的不确定性。不良呼叫具有拨打电话都为陌生号码,不确定性很大等特点。空号比例定义为拨打号码为空号的电话在所有拨打电话中的占比。被叫用户直接挂断比例定义为被叫用户没有接通电话直接将电话挂断的数量占总呼叫次数的比例。被叫号码好友比例,定义为被叫号码为主叫号码好友在通话中的比例。其中,好友定义为双方互通电话三次以上的号码,为用户的好友。

[0053] 本发明实施例提供的诈骗电话号码识别方法,通过设置获取重要性较高的若干个特征作为提取特征后的训练通话记录,能够对诈骗电话号码识别得更加精准。

[0054] 基于上述实施例,图2为本发明实施例中的诈骗电话识别模型训练流程图,本实施例请参考图2。所述将所述提取特征后的训练通话记录输入未训练的诈骗电话识别模型中,基于梯度提升决策树算法,获取所述训练好的诈骗电话识别模型,具体包括:将所述提取特征后的训练通话记录输入未训练的诈骗电话识别模型中,生成决策树;获取所述决策树的差值,并判断所述差值是否达到循环条件;若所述差值达到循环条件,则所述训练好的诈骗电话识别模型获取完成,若所述差值未达到循环条件,则重复执行上述步骤,直至差值达到循环条件。

[0055] 下面来介绍特征工程的概念,特征工程就是利用相关知识与数据处理技巧,从海量数据中发现特征,使特征能够在模型建立以及结果预测中发挥作用的过程。好的特征工程,能够找出好的特征,发现隐藏在数据中的潜在规律,减小噪声的影响,从而使建立的模型达到很好的效果。

[0056] 特征工程的步骤一般分为数据清洗与采样,特征处理,特征选择三个过程。在本发明实施例中,具体可包括:

[0057] 第一,将电信网络中的原始训练通话记录进行预处理,获取训练通话记录,所述预处理包括数据清洗和数据采样。第二,对所述训练通话记录进行数据处理,获取数据处理后的训练通话记录,所述数据处理包括归一化处理、去平均值处理、离散化处理、one-hot编码转化、词袋模型处理、n-gram处理、TF-IDF处理以及词向量转化处理中的任一种。第三,对所述处理后的训练通话记录进行特征提取,获取提取特征后的训练通话记录。

[0058] 进一步地,诈骗电话识别模型建立过程主要包括特征工程与模型训练两个过程。

[0059] 本发明实施例通过采集电信网络中用户的通话信息,对其行为进行分析,采用大数据分析的手段,利用梯度提升决策树算法建立诈骗电话识别模型,实现对通信网中的电话进行监控,发现其中的诈骗电话号码,根据发现的结果,实现对诈骗电话的打击与治理,保护人民的生命财产安全。

[0060] 基于上述实施例,图3为本发明诈骗电话号码识别系统实施例模块图,如图3所示,包括:特征获取模块1,用于对通话记录进行特征提取,获取提取特征后的通话记录,所述提取特征后的通话记录包括空号比例、好友通话比例和单位时间内电话号码的熵;获取结果模块2,用于将所述提取特征后的通话记录输入训练好的诈骗电话识别模型中,获取所述通

话记录对应的诈骗电话识别结果，其中，所述诈骗电话识别模型通过梯度提升决策树算法训练。

[0061] 需要说明的是，上述特征获取模块1和获取结果模块2配合以执行上述实施例中的一种诈骗电话号码识别方法，该系统的具体功能参见上述的识别方法的实施例，此处不再赘述。

[0062] 图4为本发明实施例的一种诈骗电话号码识别设备的结构示意图，如图4所示，该设备包括：处理器(processor)401、通信接口(Communications Interface)402、存储器(memory)403和总线404，其中，处理器401，通信接口402，存储器403通过总线404完成相互间的通信。处理器401可以调用存储器403中的逻辑指令，以执行如下方法：对通话记录进行特征提取，获取提取特征后的通话记录，所述提取特征后的通话记录包括空号比例、好友通话比例和单位时间内电话号码的熵；将所述提取特征后的通话记录输入训练好的诈骗电话识别模型中，获取所述通话记录对应的诈骗电话识别结果，其中，所述诈骗电话识别模型通过梯度提升决策树算法训练。

[0063] 本发明实施例公开一种计算机程序产品，所述计算机程序产品包括存储在非暂态计算机可读存储介质上的计算机程序，所述计算机程序包括程序指令，当所述程序指令被计算机执行时，计算机能够执行上述各方法实施例所提供的方法，例如包括：对通话记录进行特征提取，获取提取特征后的通话记录，所述提取特征后的通话记录包括空号比例、好友通话比例和单位时间内电话号码的熵；将所述提取特征后的通话记录输入训练好的诈骗电话识别模型中，获取所述通话记录对应的诈骗电话识别结果，其中，所述诈骗电话识别模型通过梯度提升决策树算法训练。

[0064] 本实施例提供一种非暂态计算机可读存储介质，所述非暂态计算机可读存储介质存储计算机指令，所述计算机指令使所述计算机执行上述各方法实施例所提供的方法，例如包括：对通话记录进行特征提取，获取提取特征后的通话记录，所述提取特征后的通话记录包括空号比例、好友通话比例和单位时间内电话号码的熵；将所述提取特征后的通话记录输入训练好的诈骗电话识别模型中，获取所述通话记录对应的诈骗电话识别结果，其中，所述诈骗电话识别模型通过梯度提升决策树算法训练。

[0065] 本领域普通技术人员可以理解：实现上述方法实施例的全部或部分步骤可以通过程序指令相关的硬件来完成，前述的程序可以存储于一计算机可读取存储介质中，该程序在执行时，执行包括上述方法实施例的步骤；而前述的存储介质包括：ROM、RAM、磁碟或者光盘等各种可以存储程序代码的介质。

[0066] 以上所描述的显示装置的测试设备等实施例仅仅是示意性的，其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。本领域普通技术人员在不付出创造性的劳动的情况下，即可以理解并实施。

[0067] 通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到各实施方式可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件。基于这样的理解，上述技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品可以存储在计算机可读存储介质中，如ROM/RAM、磁碟、光盘等，包括若干指

令用以使得一台计算机设备(可以是个人计算机,服务器,或者网络设备等)执行各个实施例或者实施例的某些部分所述的方法。

[0068] 最后应说明的是:以上实施例仅用以说明本发明的技术方案,而非对其限制;尽管参照前述实施例对本发明进行了详细的说明,本领域的普通技术人员应当理解:其依然可以对前述各实施例所记载的技术方案进行修改,或者对其中部分技术特征进行等同替换;而这些修改或者替换,并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

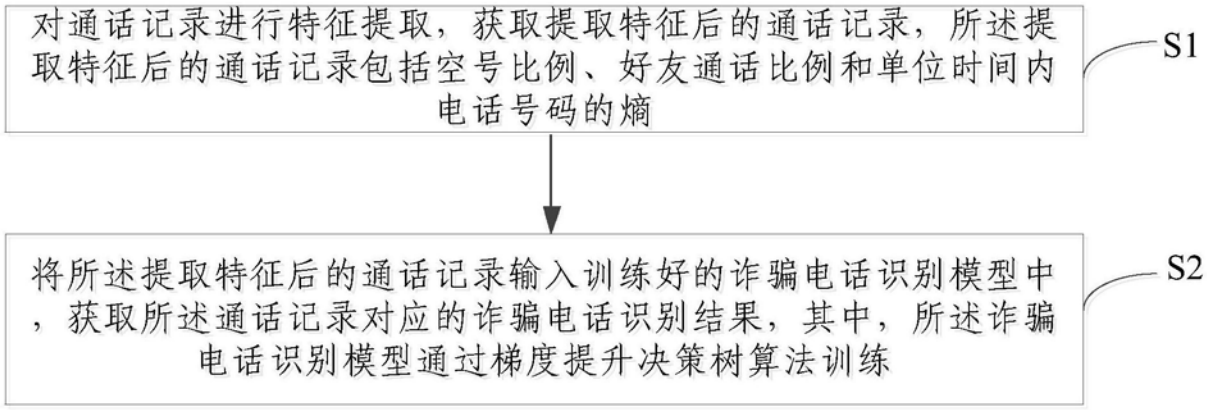
对通话记录进行特征提取，获取提取特征后的通话记录，所述提取特征后的通话记录包括空号比例、好友通话比例和单位时间内电话号码的熵
S1
将所述提取特征后的通话记录输入训练好的诈骗电话识别模型中，获取所述通话记录对应的诈骗电话识别结果，其中，所述诈骗电话识别模型通过梯度提升决策树算法训练
S2

图1

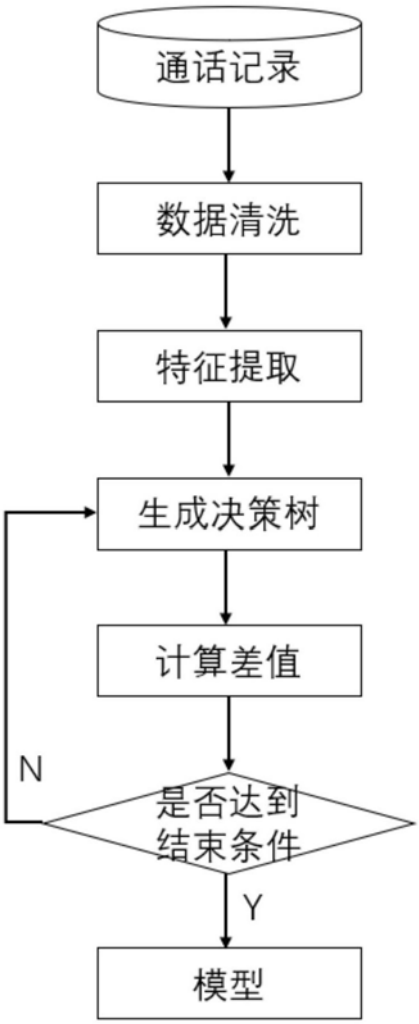
通话记录
数据清洗
特征提取
生成决策树
计算差值
是否达到结束条件
N
Y
模型

图2


诈骗电话号码识别系统
1
特征获取模块
2
获取结果模块

图3

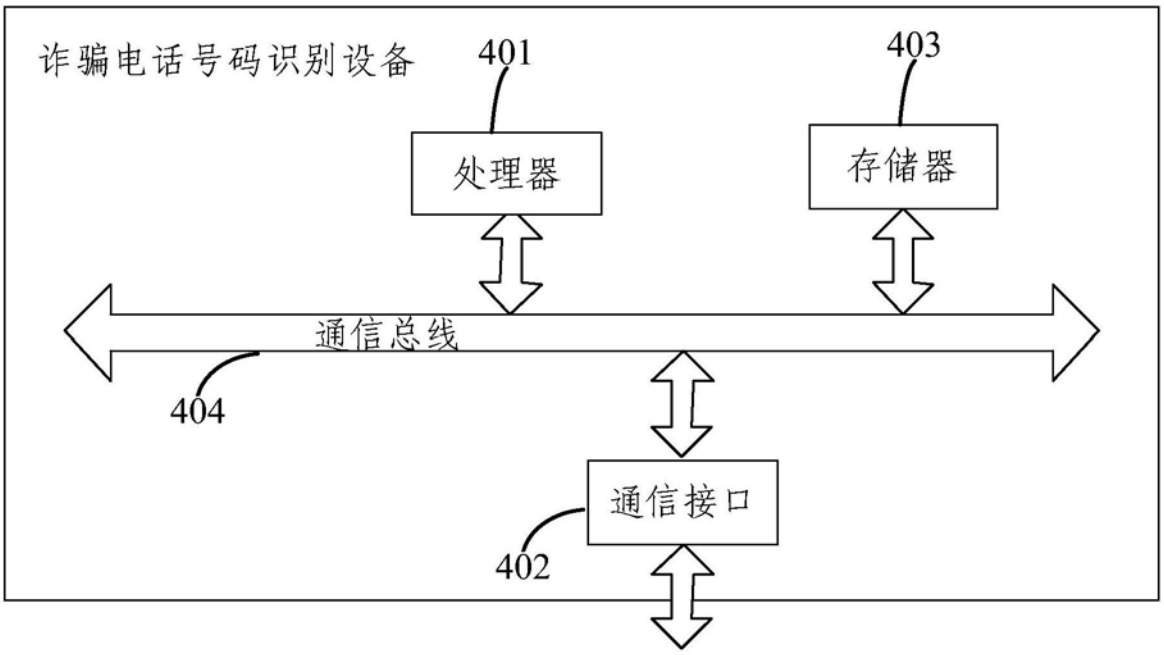
诈骗电话号码识别设备
401
处理器
403
存储器
通信总线
404
通信接口
402

图4