(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 110062114 A
(43)申请公布日 2019.07.26

(21)申请号 201910281812.0

(22)申请日 2019.04.09

(71)申请人 国家计算机网络与信息安全管理中心
地址 100029 北京市朝阳区裕民路甲三号
申请人 天津市国瑞数码安全系统股份有限公司

(72)发明人 王中华 夏光升 刘志会 李新

(74)专利代理机构 北京力量专利代理事务所（特殊普通合伙）11504
代理人 王鸿远

(51)Int.Cl.
H04M 3/22(2006.01)

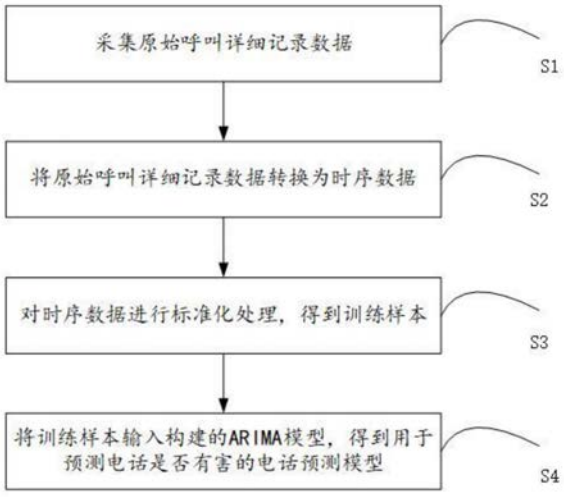
权利要求书1页 说明书5页 附图2页

(54)发明名称

基于ARIMA的诈骗电话预测方法及预测系统

(57)摘要

本发明涉及电信技术领域，尤其涉及一种基于ARIMA的诈骗电话预测方法及预测系统。该方法包括以下步骤：采集原始呼叫详细记录数据；将原始呼叫详细记录数据转换为时序数据；对时序数据进行标准化处理，得到训练样本；将训练样本输入构建的ARIMA模型，得到用于预测电话是否有害的电话预测模型。本发明根据原始呼叫详细记录数据对构建的ARIMA模型进行训练，得到预测有害电话的电话预测模型，该电话预测模型能够自动分析预测出主叫电话是否为有害电话，以及预测电信网的未来诈骗趋势，具有成本低、识别准确率高的优点。



CN 110062114 A

1.一种基于ARIMA的诈骗电话预测方法,其特征在于,包括以下步骤:

采集原始呼叫详细记录数据;

将原始呼叫详细记录数据转换为时序数据;

对时序数据进行标准化处理,得到训练样本;

将训练样本输入构建的ARIMA模型,得到用于预测电话是否有害的电话预测模型。

2.根据权利要求1所述的基于ARIMA的诈骗电话预测方法,其特征在于,对时序数据进行标准化处理,得到训练样本的步骤中包括以下步骤:

判断时序数据是否为平稳时间序列;

若判断结果为否,则对时序数据进行时间序列的差分,使时序数据成为平稳时间序列,作为训练样本。

3.根据权利要求2所述的基于ARIMA的诈骗电话预测方法,其特征在于,判断时序数据是否为平稳时间序列的步骤中,使用单位根检验方法检验时序数据是否为平稳时间序列。

4.根据权利要求3所述的基于ARIMA的诈骗电话预测方法,其特征在于,单位根检验方法为ADF检验。

5.根据权利要求4所述的基于ARIMA的诈骗电话预测方法,其特征在于,将训练样本输入构建的ARIMA模型的步骤之前还包括以下步骤:

采用基于贝叶斯信息准则训练模型进行网格搜索来确定参数p,d,q的值,其中,p为自回归项数,q为滑动平均项数,d为使时序数据成为平稳时间序列所做的差分次数;

根据p,d,q的值构建ARIMA模型。

6.根据权利要求5所述的基于ARIMA的诈骗电话预测方法,其特征在于,原始呼叫详细记录数据至少包括有害电话检出量和拦截量。

7.一种实现如权利要求1所述方法的基于ARIMA的诈骗电话预测系统,其特征在于,包括:

原始呼叫详细记录数据采集模块,用于采集原始呼叫详细记录数据;

时序数据转换模块,用于将原始呼叫详细记录数据转换为时序数据;

训练样本计算模块,用于对时序数据进行标准化处理,得到训练样本;

电话预测模型计算模块,用于将训练样本输入构建的ARIMA模型,得到用于预测电话是否有害的电话预测模型。

8.根据权利要求7所述的基于ARIMA的诈骗电话预测系统,其特征在于,训练样本计算模块包括:

判断单元,用于判断时序数据转换模块发送的时序数据是否为平稳时间序列,若判断结果为否,则将时序数据发送至时间序列差分单元,若判断结果为是,则将时序数据发送至电话预测模型计算模块;

时间序列差分单元,用于对时序数据进行时间序列的差分,使时序数据成为平稳时间序列,作为训练样本发送至电话预测模型计算模块。

9.根据权利要求8所述的基于ARIMA的诈骗电话预测系统,其特征在于,判断单元使用单位根检验方法检验时序数据是否为平稳时间序列。

10.根据权利要求9所述的基于ARIMA的诈骗电话预测系统,其特征在于,单位根检验方法为ADF检验。

基于ARIMA的诈骗电话预测方法及预测系统

技术领域

[0001] 本发明涉及电信技术领域，尤其涉及一种基于ARIMA的诈骗电话预测方法及预测系统。

背景技术

[0002] 近年来随着金融、通信业的快速发展，虚假信息诈骗犯罪迅速在我国发展蔓延。借助于手机、固定电话等通信工具和现代的网银技术实施的非接触式的电信诈骗犯罪可以说是迅速地发展蔓延，给人民群众造成了很大的损失。目前电信诈骗犯罪的手段如作案者冒充相关国家政府机关人员，例如电信局、公安局等单位工作人员，给受害者拨打电话，在通话中以受害人电话欠费、被他人盗用身份涉嫌经济犯罪，以没收受害人所有银行存款等进行恫吓威胁，骗取受害人像其汇转资金。

[0003] 现有技术中的诈骗电话识别方法，大都将智能终端与云服务器相结合，通过云服务器统计智能终端将某一电话号码的标记为诈骗电话的次数，当所得统计次数达到预设的限值时，认定该电话号码为诈骗电话，然后即对接到该电话号码呼叫的用户进行提醒，防止用户上当受骗。上述识别方法的实现，依赖于用户对电话号码的标记情况，只有对某一电话号码的标记次数达到预设的限值时，才会将该电话号码认定为诈骗电话，而这一过程往往需要经历较长的时间，导致诈骗电话的识别工作效率低下，滞后性比较严重。

[0004] ARIMA模型，Autoregressive Integrated Moving Average model，差分整合移动平均自回归模型，又称整合移动/滑动平均自回归模型，时间序列预测分析方法之一，ARIMA是可以适应时间序列数据的模型。

[0005] 目前电信网缺乏管理预见性的一种手段，不利于突发性的事件的处理，不能为业务开展提供指导。

[0006] 因此，急需一种基于ARIMA的诈骗电话预测方法及预测系统。

发明内容

[0007] 鉴于上述问题，提出了本发明以便提供一种克服上述问题或者至少部分地解决上述问题的一种基于ARIMA的诈骗电话预测方法及预测系统。

[0008] 本发明的一个方面，提供了一种基于ARIMA的诈骗电话预测方法，包括以下步骤：

[0009] 采集原始呼叫详细记录数据；

[0010] 将原始呼叫详细记录数据转换为时序数据；

[0011] 对时序数据进行标准化处理，得到训练样本；

[0012] 将训练样本输入构建的ARIMA模型，得到用于预测电话是否有害的电话预测模型。

[0013] 进一步地，对时序数据进行标准化处理，得到训练样本的步骤中包括以下步骤：

[0014] 判断时序数据是否为平稳时间序列；

[0015] 若判断结果为否，则对时序数据进行时间序列的差分，使时序数据成为平稳时间序列，作为训练样本。

[0016] 进一步地，判断时序数据是否为平稳时间序列的步骤中，使用单位根检验方法检验时序数据是否为平稳时间序列。

[0017] 进一步地，单位根检验方法为ADF检验。

[0018] 进一步地，将训练样本输入构建的ARIMA模型的步骤之前还包括以下步骤：

[0019] 采用基于贝叶斯信息准则训练模型进行网格搜索来确定参数p，d，q的值，其中，p为自回归项数，q为滑动平均项数，d为使时序数据成为平稳时间序列所做的差分次数；

[0020] 根据p，d，q的值构建ARIMA模型。

[0021] 进一步地，原始呼叫详细记录数据至少包括有害电话检出量和拦截量。

[0022] 本发明的第二个方面，提供了一种实现如上述中所述方法的基于ARIMA的诈骗电话预测系统，包括：

[0023] 原始呼叫详细记录数据采集模块，用于采集原始呼叫详细记录数据；

[0024] 时序数据转换模块，用于将原始呼叫详细记录数据转换为时序数据；

[0025] 训练样本计算模块，用于对时序数据进行标准化处理，得到训练样本；

[0026] 电话预测模型计算模块，用于将训练样本输入构建的ARIMA模型，得到用于预测电话是否有害的电话预测模型。

[0027] 进一步地，训练样本计算模块包括：

[0028] 判断单元，用于判断时序数据转换模块发送的时序数据是否为平稳时间序列，若判断结果为否，则将时序数据发送至时间序列差分单元，若判断结果为是，则将时序数据发送至电话预测模型计算模块；

[0029] 时间序列差分单元，用于对时序数据进行时间序列的差分，使时序数据成为平稳时间序列，作为训练样本发送至电话预测模型计算模块。

[0030] 进一步地，判断单元，使用单位根检验方法检验时序数据是否为平稳时间序列。

[0031] 进一步地，单位根检验方法为ADF检验。

[0032] 本发明提供的基于ARIMA的诈骗电话预测方法及预测系统，与现有技术相比具有以下进步：

[0033] 本发明根据原始呼叫详细记录数据对构建的ARIMA模型进行训练，得到预测有害电话的电话预测模型，该电话预测模型能够自动分析预测出主叫电话是否为有害电话，以及预测电信网的未来诈骗趋势，具有成本低、识别准确率高的优点。

[0034] 上述说明仅是本发明技术方案的概述，为了能够更清楚了解本发明的技术手段，而可依照说明书的内容予以实施，并且为了让本发明的上述和其它目的、特征和优点能够更明显易懂，以下特举本发明的具体实施方式。

附图说明

[0035] 通过阅读下文优选实施方式的详细描述，各种其他的优点和益处对于本领域普通技术人员将变得清楚明了。附图仅用于示出优选实施方式的目的，而并不认为是对本发明的限制。而且在整个附图中，用相同的参考符号表示相同的部件。在附图中：

[0036] 图1为本发明实施例中基于ARIMA的诈骗电话预测方法的步骤图；

[0037] 图2为本发明实施例中基于ARIMA的诈骗电话预测系统的器件连接框图。

具体实施方式

[0038] 下面将参照附图更详细地描述本公开的示例性实施例。虽然附图中显示了本公开的示例性实施例，然而应当理解，可以以各种形式实现本公开而不应被这里阐述的实施例所限制。相反，提供这些实施例是为了能够更透彻地理解本公开，并且能够将本公开的范围完整的传达给本领域的技术人员。

[0039] 本技术领域技术人员可以理解，除非另外定义，这里使用的所有术语(包括技术术语和科学术语)，具有与本发明所属领域中的普通技术人员的一般理解相同的意义。还应该理解的是，诸如通用字典中定义的那些术语，应该被理解为具有与现有技术的上下文中的意义一致的意义，并且除非被特定定义，否则不会用理想化或过于正式的含义来解释。

[0040] 本发明实施例提供了一种基于ARIMA的诈骗电话预测方法及预测系统。

[0041] 如图1，本实施例的一种基于ARIMA的诈骗电话预测方法，包括以下步骤：

[0042] S1、采集原始呼叫详细记录数据；

[0043] S2、将原始呼叫详细记录数据转换为时序数据；

[0044] S3、对时序数据进行标准化处理，得到训练样本；

[0045] S4、将训练样本输入构建的ARIMA模型，得到用于预测电话是否有害的电话预测模型。

[0046] 本发明根据原始呼叫详细记录数据对构建的ARIMA模型进行训练，得到预测有害电话的电话预测模型，该电话预测模型能够自动分析预测出主叫电话是否为有害电话，以及预测电信网的未来诈骗趋势，具有成本低、识别准确率高的优点。

[0047] CDR话单数据(呼叫详细记录)，描述了呼叫接续的全过程。

[0048] 具体实施时，步骤S3对时序数据进行标准化处理，得到训练样本的步骤中包括以下步骤：

[0049] 判断时序数据是否为平稳时间序列；

[0050] 若判断结果为否，则对时序数据进行时间序列的差分，使时序数据成为平稳时间序列，作为训练样本。

[0051] 具体实施时，步骤S3对时序数据进行标准化处理的步骤中，还包括对时序数据进行数据缩放、数据标准化、数据归一化处理，这些处理方法都是将数据按比例缩放，使之落入一个小的特定区间，在数据训练过程中便于收敛。本发明中，对数据取了对数，缩小数据的绝对数值，方便计算，取对数后，不会改变数据的性质和相关关系，但缩小了变量的尺度，使得数据更加平稳，也消弱了模型的共线性、方差性等。

[0052] 具体实施时，判断时序数据是否为平稳时间序列的步骤中，使用单位根检验方法检验时序数据是否为平稳时间序列。具体实施时，单位根检验方法为ADF检验。

[0053] 具体实施时，将训练样本输入构建的ARIMA模型的步骤之前还包括以下步骤：

[0054] 采用基于贝叶斯信息准则(BIC)训练模型进行网格搜索来确定参数p，d，q的值，其中，p为自回归项数，q为滑动平均项数，d为使时序数据成为平稳时间序列所做的差分次数；根据BIC遍历p和q的值，取BIC最小值时对应的p和q的值。

[0055] 根据p，d，q的值构建ARIMA模型。

[0056] 参数p，d，q用于参数化ARIMA模型，从而利用模型对电信网时序数据进行预测。AR是自回归模型，p为自回归项数；MA为滑动平均模型，q为滑动平均项数；d为使时序数据成为

平稳时间序列所做的差分次数(阶数),ARIMA模型对时间序列的要求是平稳型,是在ARMA模型的基础上解决非平稳序列的模型。因此,当得到一个非平稳的时间序列时,首先要做的即是做时间序列的差分,直到得到一个平稳时间序列。如果对时间序列做d次差分才能得到一个平稳序列,那么才可以使用ARIMA(p,d,q)模型。

[0057] ARIMA模型的公式为

[0058] $$\left(1-\sum_{i=1}^{p}\phi_i L^i\right)(1-L)^d X_t=\left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t$$

[0059] 其中,L是滞后算子(Lag operator),$d\in\mathbb{Z},d>0$。

[0060] 对数据进行训练之前首先要看时序数据是否平稳的,如果不平稳,就要进行差分让数据平稳,即差分次数d。如何判断数据是不是平稳的,一般有三种方式:(1)看时序图,如果是平稳的数据会围绕一个常数上下波动,反之不平稳。(2)自相关系数(acf)和偏相关系数(pacf),观看自相关图和偏相关图,平稳的序列的自相关图和偏相关图不是拖尾就是截尾。(3)ADF单位根检验方法(本发明主要采用此种方法,上面两种为辅助方式),ADF检验,如果序列平稳,则不存在单位根,否则就会存在单位根。ADF原假设为,序列存在单位根,即非平稳,对于一个平稳的时序数据,就需要在给定的置信水平上显著,拒绝原假设。若得到的统计量显著小于3个置信度(1%,5%,10%)的临界统计值时,说明是拒绝原假设的。另外是看P-value是否非常接近如0.0001(4位小数基本即可)。

[0061] 具体实施时,原始呼叫详细记录数据至少包括有害电话检出量和拦截量。根据需要,也可以包括其他数据。

[0062] 如图2,本实施例的一种实现如上述实施例所述方法的基于ARIMA的诈骗电话预测系统,包括:

[0063] 原始呼叫详细记录数据采集模块,用于采集原始呼叫详细记录数据;

[0064] 时序数据转换模块,用于将原始呼叫详细记录数据转换为时序数据;

[0065] 训练样本计算模块,用于对时序数据进行标准化处理,得到训练样本;

[0066] 电话预测模型计算模块,用于将训练样本输入构建的ARIMA模型,得到用于预测电话是否有害的电话预测模型。

[0067] 本发明根据原始呼叫详细记录数据对构建的ARIMA模型进行训练,得到预测有害电话的电话预测模型,该电话预测模型能够自动分析预测出主叫电话是否为有害电话,以及预测电信网的未来诈骗趋势,具有成本低、识别准确率高的优点。

[0068] 如图2,具体实施时,训练样本计算模块包括:

[0069] 判断单元,用于判断时序数据转换模块发送的时序数据是否为平稳时间序列,若判断结果为否,则将时序数据发送至时间序列差分单元,若判断结果为是,则将时序数据发送至电话预测模型计算模块;

[0070] 时间序列差分单元,用于对时序数据进行时间序列的差分,使时序数据成为平稳时间序列,作为训练样本发送至电话预测模型计算模块。

[0071] 具体实施时,判断单元,使用单位根检验方法检验时序数据是否为平稳时间序列。

[0072] 具体实施时,单位根检验方法为ADF检验。

[0073] 以上方法实施例的改进和有益效果也属于系统实施例的改进和有益效果,系统实

施例中不再赘述。

[0074] 针对有害呼叫手段不断变化的特点，本发明采用大数据分析、系统自学习、自回溯验证、实时验证、趋势预测等技术，以现有电信网时间序列统计数据为基础进行下一节点或时间段上的数据预测，实现系统自动化分析预测。本发明最终得到的电话预测模型可通过系统模型自学习和趋势预测技术，有效发现集中电话骚扰情况，通过自动验证不断完善系统模型，实现系统自动化。

[0075] 对于方法实施例，为了简单描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本发明实施例并不受所描述的动作顺序的限制，因为依据本发明实施例，某些步骤可以采用其他顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于优选实施例，所涉及的动作并不一定是本发明实施例所必须的。

[0076] 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。

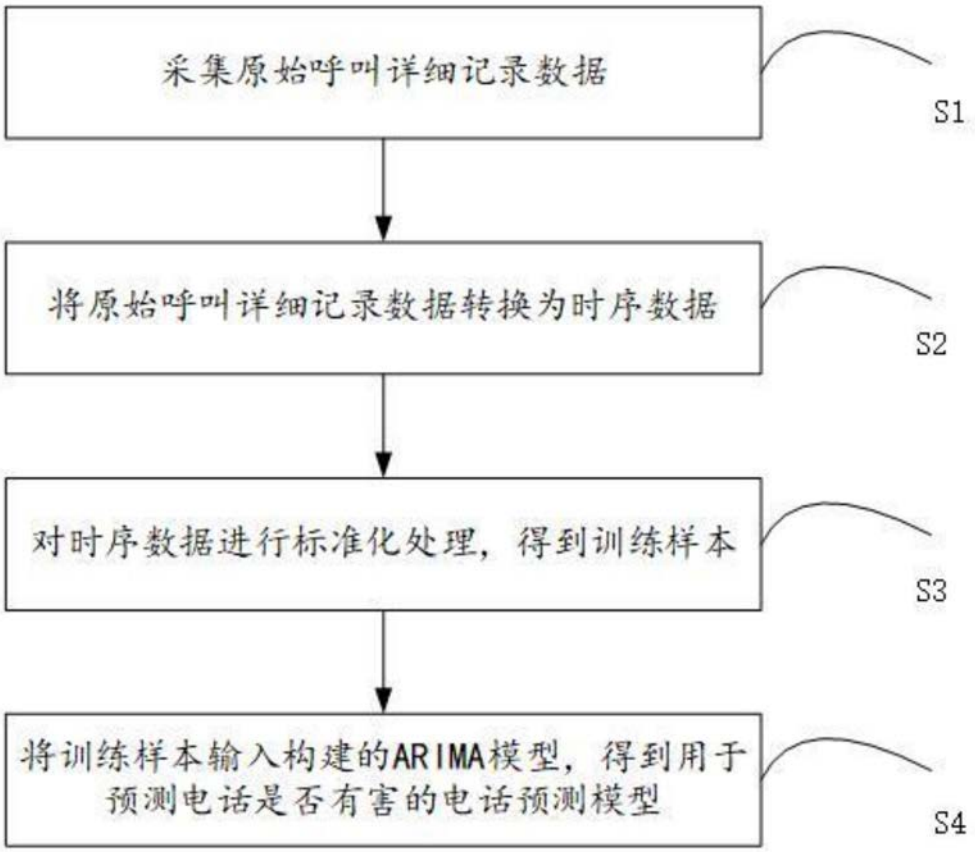

采集原始呼叫详细记录数据
S1
将原始呼叫详细记录数据转换为时序数据
S2
对时序数据进行标准化处理，得到训练样本
S3
将训练样本输入构建的ARIMA模型，得到用于预测电话是否有害的电话预测模型
S4

图1

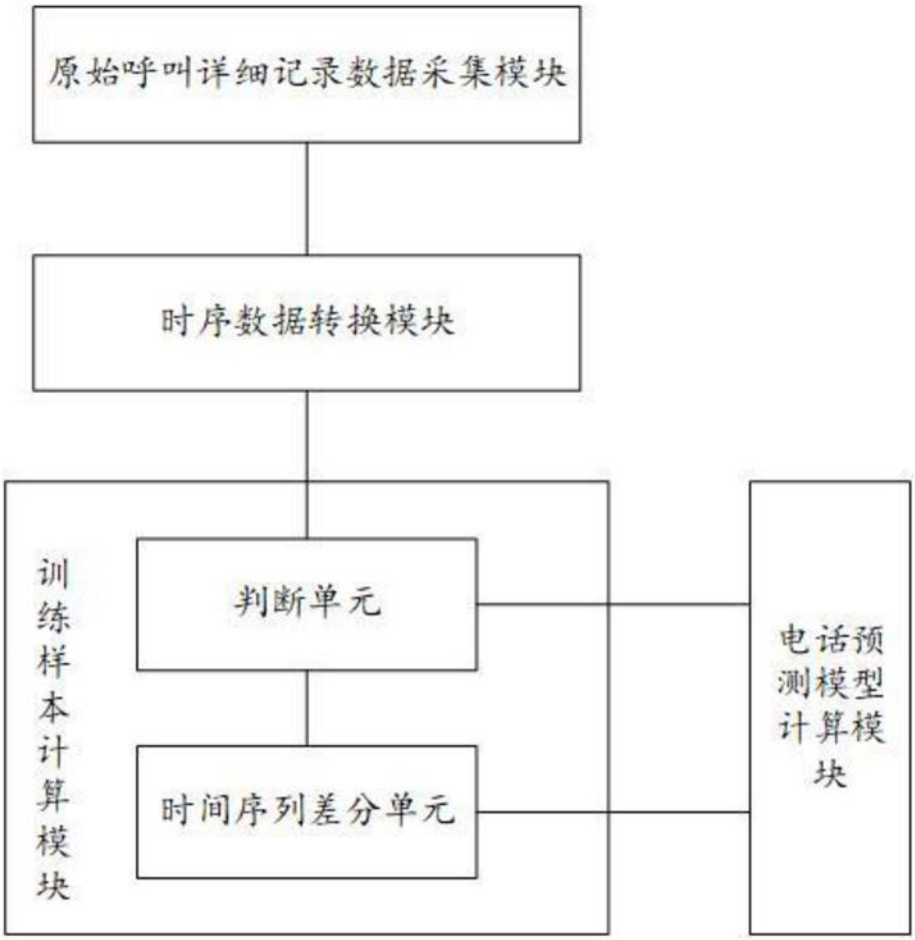

原始呼叫详细记录数据采集模块
时序数据转换模块
训练样本计算模块
判断单元
时间序列差分单元
电话预测计算模型模块

图2