(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 108932434 A
(43)申请公布日 2018.12.04

(21)申请号 201810638537.9

(22)申请日 2018.06.20

(71)申请人 中国农业银行股份有限公司
地址 100005 北京市东城区建国门内大街69号

(72)发明人 赵维平 李现伟 李超 樊盛博 赵存超

(74)专利代理机构 北京集佳知识产权代理有限公司 11227
代理人 王宝筠

(51)Int.Cl.
G06F 21/60(2013.01)
G06K 9/62(2006.01)
H04L 29/06(2006.01)

权利要求书2页 说明书6页 附图3页

(54)发明名称

一种基于机器学习技术的数据加密方法及装置

(57)摘要

本申请公开了一种基于机器学习技术的数据加密方法，提取目标数据表中的每个数据列特征；将每个所述数据列特征输入到待加密字段识别模型中，使所述待加密字段识别模型对每个所述数据列特征进行识别，得到所述目标数据表中的待加密字段；采用预设映射加密算法对所述目标数据表中的待加密字段对应的数据进行加密。实现了自动识别待加密字段，不需要人工干预，提高了数据加密效率。

CN 108932434 A

1.一种基于机器学习技术的数据加密方法,其特征在于,包括:

提取目标数据表中的每个数据列特征;

将每个所述数据列特征输入到待加密字段识别模型中,使所述待加密字段识别模型对每个所述数据列特征进行识别,得到所述目标数据表中的待加密字段;

采用预设映射加密算法对所述目标数据表中的待加密字段对应的数据进行加密。

2.根据权利要求1所述的方法,其特征在于,所述方法还包括:

构建训练数据集和测试数据集,并根据所述训练数据集和测试数据集对分类决策模型进行训练,得到所述待加密字段识别模型。

3.根据权利要求2所述的方法,其特征在于,所述构建训练数据集和测试数据集,并根据所述训练数据集和测试数据集对分类决策模型进行训练,得到所述待加密字段识别模型,包括:

提取每条待处理数据中的数据列特征;

确定待处理数据中需要加密的数据和不需要加密的数据,每条待处理数据包括数据列特征和加密识别结果,所述加密识别结果为需要加密和不需要加密;

将待处理数据划分为训练数据集和测试数据集;

根据训练数据集对分类决策模型进行训练,并根据测试数据集对分类决策模型的训练结果进行验证,最终得到待加密字段识别模型。

4.根据权利要求1所述的方法,其特征在于,所述采用预设映射加密算法对所述目标数据表中的待加密字段对应的数据进行加密,包括:

对于所述目标数据表中待加密字段对应的每一条原始数据,根据取代加密法和预设密钥中的字母部分对原始数据进行映射加密,得到初始加密数据;

根据转位加密法和所述预设密钥中的数字部分对所述初始加密数据进行转位加密,得到最终加密数据;

在所述目标数据表中的所述原始数据替换为相应的将最终加密数据。

5.根据权利要求1所述的方法,其特征在于,在所述采用预设映射加密算法对所述目标数据表中的待加密字段对应的数据进行加密之后,所述方法还包括:

当需要对目标数据表中的已加密数据进行解密时,获取已加密数据加密时的预设秘钥;

根据转位加密法和所述预设秘钥中的数字部分对所述已加密数据进行反向转位解密,得到初始解密数据;

根据取代加密法和所述预设秘钥中的字母部分对所述初始解密数据进行反向映射解密,得到最终解密数据。

6.一种基于机器学习技术的数据加密装置,其特征在于,包括:

提取单元,用于提取目标数据表中的每个数据列特征;

识别单元,用于将每个所述数据列特征输入到待加密字段识别模型中,使所述待加密字段识别模型对每个所述数据列特征进行识别,得到所述目标数据表中的待加密字段;

加密单元,用于采用预设映射加密算法对所述目标数据表中的待加密字段对应的数据进行加密。

7.根据权利要求6所述的装置,其特征在于,所述装置还包括训练单元,所述训练单元

用于构建训练数据集和测试数据集，并根据所述训练数据集和测试数据集对分类决策模型进行训练，得到所述待加密字段识别模型。

8.根据权利要求7所述的装置，其特征在于，所述训练单元具体用于：

提取每条待处理数据中的数据列特征；确定待处理数据中需要加密的数据和不需要加密的数据，每条待处理数据包括数据列特征和加密识别结果，所述加密识别结果为需要加密和不需要加密；将待处理数据划分为训练数据集和测试数据集；根据训练数据集对分类决策模型进行训练，并根据测试数据集对分类决策模型的训练结果进行验证，最终得到待加密字段识别模型。

9.根据权利要求6所述的装置，其特征在于，所述加密单元具体用于：

对于所述目标数据表中待加密字段对应的每一条原始数据，根据取代加密法和预设密钥中的字母部分对原始数据进行映射加密，得到初始加密数据；根据转位加密法和所述预设密钥中的数字部分对所述初始加密数据进行转位加密，得到最终加密数据；在所述目标数据表中的所述原始数据替换为相应的将最终加密数据。

10.根据权利要求6所述的装置，其特征在于，所述装置还包括：

解密单元，用于当需要对目标数据表中的已加密数据进行解密时，获取已加密数据加密时的预设秘钥；根据转位加密法和所述预设秘钥中的数字部分对所述已加密数据进行反向转位解密，得到初始解密数据；根据取代加密法和所述预设秘钥中的字母部分对所述初始解密数据进行反向映射解密，得到最终解密数据。

一种基于机器学习技术的数据加密方法及装置

技术领域

[0001] 本发明涉及数据处理技术领域，更具体的，涉及一种基于机器学习技术的数据加密方法及装置。

背景技术

[0002] 数据脱敏对于所有行业都十分重要，尤其是金融行业数据庞杂，隐私数据多，数据脱敏十分重要。

[0003] 传统的数据脱敏做法是筛选出需要脱敏的字段，使用替换，重排，加密，截断，掩码等方法对其进行脱敏，这种脱敏以后的数据丧失了隐私性，保证了数据安全。但是，对于数据质量比较差的机构，筛选出哪些数据库表，哪些数据库字段需要脱敏都是一件代价极高的工作。

[0004] 因此，目前迫切需要一种能够自动识别需要脱敏的数据库表，以及相关数据库字段的技术。

发明内容

[0005] 有鉴于此，本发明提供了一种基于机器学习技术的数据加密方法及装置，实现了自动识别待加密字段，不需要人工干预，提高了数据加密效率。

[0006] 为了实现上述发明目的，本发明提供的具体技术方案如下：

[0007] 一种基于机器学习技术的数据加密方法，包括：

[0008] 提取目标数据表中的每个数据列特征；

[0009] 将每个所述数据列特征输入到待加密字段识别模型中，使所述待加密字段识别模型对每个所述数据列特征进行识别，得到所述目标数据表中的待加密字段；

[0010] 采用预设映射加密算法对所述目标数据表中的待加密字段对应的数据进行加密。

[0011] 可选的，所述方法还包括：

[0012] 构建训练数据集和测试数据集，并根据所述训练数据集和测试数据集对分类决策模型进行训练，得到所述待加密字段识别模型。

[0013] 可选的，所述构建训练数据集和测试数据集，并根据所述训练数据集和测试数据集对分类决策模型进行训练，得到所述待加密字段识别模型，包括：

[0014] 提取每条待处理数据中的数据列特征；

[0015] 确定待处理数据中需要加密的数据和不需要加密的数据，每条待处理数据包括数据列特征和加密识别结果，所述加密识别结果为需要加密和不需要加密；

[0016] 将待处理数据划分为训练数据集和测试数据集；

[0017] 根据训练数据集对分类决策模型进行训练，并根据测试数据集对分类决策模型的训练结果进行验证，最终得到待加密字段识别模型。

[0018] 可选的，所述采用预设映射加密算法对所述目标数据表中的待加密字段对应的数据进行加密，包括：

[0019] 对于所述目标数据表中待加密字段对应的每一条原始数据，根据取代加密法和预设密钥中的字母部分对原始数据进行映射加密，得到初始加密数据；

[0020] 根据转位加密法和所述预设密钥中的数字部分对所述初始加密数据进行转位加密，得到最终加密数据；

[0021] 在所述目标数据表中的所述原始数据替换为相应的将最终加密数据。

[0022] 可选的，在所述采用预设映射加密算法对所述目标数据表中的待加密字段对应的数据进行加密之后，所述方法还包括：

[0023] 当需要对目标数据表中的已加密数据进行解密时，获取已加密数据加密时的预设秘钥；

[0024] 根据转位加密法和所述预设秘钥中的数字部分对所述已加密数据进行反向转位解密，得到初始解密数据；

[0025] 根据取代加密法和所述预设秘钥中的字母部分对所述初始解密数据进行反向映射解密，得到最终解密数据。

[0026] 一种基于机器学习技术的数据加密装置，包括：

[0027] 提取单元，用于提取目标数据表中的每个数据列特征；

[0028] 识别单元，用于将每个所述数据列特征输入到待加密字段识别模型中，使所述待加密字段识别模型对每个所述数据列特征进行识别，得到所述目标数据表中的待加密字段；

[0029] 加密单元，用于采用预设映射加密算法对所述目标数据表中的待加密字段对应的数据进行加密。

[0030] 可选的，所述装置还包括训练单元，所述训练单元用于构建训练数据集和测试数据集，并根据所述训练数据集和测试数据集对分类决策模型进行训练，得到所述待加密字段识别模型。

[0031] 可选的，所述训练单元具体用于：

[0032] 提取每条待处理数据中的数据列特征；确定待处理数据中需要加密的数据和不需要加密的数据，每条待处理数据包括数据列特征和加密识别结果，所述加密识别结果为需要加密和不需要加密；将待处理数据划分为训练数据集和测试数据集；根据训练数据集对分类决策模型进行训练，并根据测试数据集对分类决策模型的训练结果进行验证，最终得到待加密字段识别模型。

[0033] 可选的，所述加密单元具体用于：

[0034] 对于所述目标数据表中待加密字段对应的每一条原始数据，根据取代加密法和预设密钥中的字母部分对原始数据进行映射加密，得到初始加密数据；根据转位加密法和所述预设密钥中的数字部分对所述初始加密数据进行转位加密，得到最终加密数据；在所述目标数据表中的所述原始数据替换为相应的将最终加密数据。

[0035] 可选的，所述装置还包括：

[0036] 解密单元，用于当需要对目标数据表中的已加密数据进行解密时，获取已加密数据加密时的预设秘钥；根据转位加密法和所述预设秘钥中的数字部分对所述已加密数据进行反向转位解密，得到初始解密数据；根据取代加密法和所述预设秘钥中的字母部分对所述初始解密数据进行反向映射解密，得到最终解密数据。

[0037] 相对于现有技术，本发明的有益效果如下：

[0038] 本发明公开的基于机器学习技术的数据加密方法，构建待加密字段识别模式，将每个所述数据列特征输入到待加密字段识别模型中，使所述待加密字段识别模型对每个所述数据列特征进行识别，得到所述目标数据表中的待加密字段，实现了自动识别待加密字段，不需要大量人工识别或干预，提高了数据加密效率。

附图说明

[0039] 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简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的实施例，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，还可以根据提供的附图获得其他的附图。

[0040] 图1为本发明实施例公开的一种基于机器学习技术的数据加密方法流程图；

[0041] 图2为本发明实施例公开的一种待加密字段识别模型的训练方法流程图；

[0042] 图3为本发明实施例公开的另一种基于机器学习技术的数据加密方法流程图；

[0043] 图4为本发明实施例公开的一种基于机器学习技术的数据加密装置结构示意图。

具体实施方式

[0044] 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，所描述的实施例仅仅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做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。

[0045] 请参阅图1，本实施例公开了一种基于机器学习技术的数据加密方法，具体包括以下步骤：

[0046] S101：提取目标数据表中的每个数据列特征；

[0047] 目标数据表是需要进行脱敏处理的数据表，目标数据表中包括很多列数据，有些字段的数据是需要加密的，有些字段的数据是不需要加密的。

[0048] 数据列特征包括：数据长度的众数、数据列是否以数字和X占大多数、数据列英文名是否含有关键词、数据列中文名是否含有关键词、数据长度的众数是否为特定位数、数据列所在表英文名是否含有关键词、数据列所在表中文名是否含有关键词。

[0049] 其中，数据长度的众数为在对应列存储数据的长度中取众数；数据是否以数字和X占大多数具体为对应列存储数据中超过70％(可调)由数字和X组成；数据列英文名是否含有关键词具体为对应列列名中是否含有telephone cert等及其缩写的关键词；数据列中文名是否含有关键词具体为对应列列名中是否含有电话、身份证等关键词；数据长度的众数是否为特定位数表示众数是否为身份证号18位、手机号11位等特定位数；数据列所在表英文名是否含有关键词具体为对应表英文名是否含有info address telephone等及其缩写的关键词；数据列所在表中文名是否含有关键词具体为对应表中文名是否含有信息、电话、身份证等关键词。

[0050] S102：将每个所述数据列特征输入到待加密字段识别模型中，使所述待加密字段识别模型对每个所述数据列特征进行识别，得到所述目标数据表中的待加密字段；

[0051] 待加密字段识别模型的输出结果为是或否，即对于一组数据列特征，输出结果为是，则该组数据列特征对应的字段是待加密字段，输出结果为否，则该组数据列特征对应的字段不是待加密字段。

[0052] S103：采用预设映射加密算法对所述目标数据表中的待加密字段对应的数据进行加密。

[0053] 需要说明的是，所述数据加密方法还包括：

[0054] 构建训练数据集和测试数据集，并根据所述训练数据集和测试数据集对分类决策模型进行训练，得到所述待加密字段识别模型。

[0055] 请参阅图2，待加密字段识别模型的训练过程如下：

[0056] S201：提取每条待处理数据中的数据列特征；

[0057] 每条待处理数据为作为训练及测试样本的多个数据表中的多个列数据；

[0058] S202：确定待处理数据中需要加密的数据和不需要加密的数据，每条待处理数据包括数据列特征和加密识别结果，所述加密识别结果为需要加密和不需要加密；

[0059] 确定待处理数据中需要加密的数据和不需要加密的数据的方法可以为初级自动识别结合人工标注的方法，初级自动识别算法如下：

[0060] a.提取需识别字段的全部数据；

[0061] b.去除两侧空格，计算每个数据对应的数据长度；

[0062] c.判断最多的数据长度，取出众数；

[0063] d.如果众数不为指定位数，则识别数据列不为需加密字段；

[0064] e.将数据进行拼合，判断数据中是否超过一定比例不由数字和X组成。超过一定比例的则识别数据列不为需加密字段；

[0065] f.其它情况，则判断数据列为需要加密的字段。

[0066] S203：将待处理数据划分为训练数据集和测试数据集；

[0067] S204：根据训练数据集对分类决策模型进行训练，并根据测试数据集对分类决策模型的训练结果进行验证，最终得到待加密字段识别模型。

[0068] 分类决策模型可以逻辑回归模型、决策树模型、支持向量机等模型，在此不做具体限定。

[0069] 针对提取得到的每个数据列特征首先进行一定的数据预处理。这一步骤的目的是优化学习结果，将模型不能接受的数据类型转化成可处理的类型，同时将有扰乱性的数据分布转换成对机器学习有利的分布。由于目前所使用的特征多为离散二类特征值，所需要的预处理加工量较小，包括对部分连续属性进行离散化处理、对类别类属性进行二元化处理和对部分连续属性进行属性变换。

[0070] 在实际训练过程中，为了避免过拟合问题，需要通过测试数据集对分类决策模型训练后的结果进行验证，并通过验证结果对模型的训练进行调整，得到识别精度较理想的待加密字段识别模型。

[0071] 本实施例公开的基于机器学习技术的数据加密方法，构建待加密字段识别模式，将每个所述数据列特征输入到待加密字段识别模型中，使所述待加密字段识别模型对每个所述数据列特征进行识别，得到所述目标数据表中的待加密字段，实现了自动识别待加密字段，不需要大量人工识别或干预，提高了数据加密效率。

[0072] 请参阅图3,本实施例公开的基于机器学习技术的数据加密方法,包括以下步骤:

[0073] S301:提取目标数据表中的每个数据列特征;

[0074] S302:将每个所述数据列特征输入到待加密字段识别模型中,使所述待加密字段识别模型对每个所述数据列特征进行识别,得到所述目标数据表中的待加密字段;

[0075] S303:采用预设映射加密算法对所述目标数据表中的待加密字段对应的数据进行加密;

[0076] 具体的,预设映射加密算法符合使用了取代加密法和转位加密法,以对电话号码信息进行加密为例,首先需要设计密钥。本算法的预设密钥由两部分组成,第一部分为字母,如取古诗词拼音首字母,举例如“商女不知亡国恨”,“SNBZWGH”;第二部分为数字,如取一事先约定好的数字,如2。

[0077] 具体加密过程如下:

[0078] 对于所述目标数据表中待加密字段对应的每一条原始数据,根据取代加密法和预设密钥中的字母部分对原始数据进行映射加密,得到初始加密数据;

[0079] 如将“123”加密为“SNB”。

[0080] 根据转位加密法和所述预设密钥中的数字部分对所述初始加密数据进行转位加密,得到最终加密数据;

[0081] 如将“SNB”加密为“NBS”,即从第2位开始转位。

[0082] 在所述目标数据表中的所述原始数据替换为相应的将最终加密数据。

[0083] 如将“123”最终加密为“NBS”。

[0084] 本实施例公开的加密方法,时间复杂度为O(n),加密速度快。

[0085] S304:当需要对目标数据表中的已加密数据进行解密时,获取已加密数据加密时的预设秘钥;

[0086] 需要说明的是,加密的密钥和解密的密钥相同。

[0087] S305:根据转位加密法和所述预设秘钥中的数字部分对所述已加密数据进行反向转位解密,得到初始解密数据;

[0088] S306:根据取代加密法和所述预设秘钥中的字母部分对所述初始解密数据进行反向映射解密,得到最终解密数据。

[0089] 加密后对SQL的select、insert、update、join等操作无影响;join关联的时候,相同内容加密后结果一致,所以不影响join操作;对于select、insert、update等操作,通过关联MAP表,相当于对加密前的数据内容进行操作,所以对select、insert、update等SQL操作也无影响;在解密方面,通过简单的SQL关联操作就可以实现,因此效率比传统数据库加密方法速度要快很多。

[0090] 采用本实施例公开的加密方法,不改变加密数据的长度,加密后密文的长度与原数据的长度相同,不影响数据库结构,加密后不影响数据库的SQL操作。

[0091] 基于上述实施例公开的一种基于机器学习技术的数据加密方法,请参阅图,本实施例对应公开了一种基于机器学习技术的数据加密装置,包括:

[0092] 提取单元401,用于提取目标数据表中的每个数据列特征;

[0093] 识别单元402,用于将每个所述数据列特征输入到待加密字段识别模型中,使所述待加密字段识别模型对每个所述数据列特征进行识别,得到所述目标数据表中的待加密字

段；

[0094] 加密单元403，用于采用预设映射加密算法对所述目标数据表中的待加密字段对应的数据进行加密。

[0095] 可选的，所述装置还包括训练单元，所述训练单元用于构建训练数据集和测试数据集，并根据所述训练数据集和测试数据集对分类决策模型进行训练，得到所述待加密字段识别模型。

[0096] 可选的，所述训练单元具体用于：

[0097] 提取每条待处理数据中的数据列特征；确定待处理数据中需要加密的数据和不需要加密的数据，每条待处理数据包括数据列特征和加密识别结果，所述加密识别结果为需要加密和不需要加密；将待处理数据划分为训练数据集和测试数据集；根据训练数据集对分类决策模型进行训练，并根据测试数据集对分类决策模型的训练结果进行验证，最终得到待加密字段识别模型。

[0098] 可选的，所述加密单元具体用于：

[0099] 对于所述目标数据表中待加密字段对应的每一条原始数据，根据取代加密法和预设密钥中的字母部分对原始数据进行映射加密，得到初始加密数据；根据转位加密法和所述预设密钥中的数字部分对所述初始加密数据进行转位加密，得到最终加密数据；在所述目标数据表中的所述原始数据替换为相应的将最终加密数据。

[0100] 可选的，所述装置还包括：

[0101] 解密单元，用于当需要对目标数据表中的已加密数据进行解密时，获取已加密数据加密时的预设秘钥；根据转位加密法和所述预设秘钥中的数字部分对所述已加密数据进行反向转位解密，得到初始解密数据；根据取代加密法和所述预设秘钥中的字母部分对所述初始解密数据进行反向映射解密，得到最终解密数据。

[0102] 本实施例公开的基于机器学习技术的数据加密装置，构建待加密字段识别模式，将每个所述数据列特征输入到待加密字段识别模型中，使所述待加密字段识别模型对每个所述数据列特征进行识别，得到所述目标数据表中的待加密字段，实现了自动识别待加密字段，不需要大量人工识别或干预，提高了数据加密效率。

[0103] 对所公开的实施例的上述说明，使本领域专业技术人员能够实现或使用本发明。对这些实施例的多种修改对本领域的专业技术人员来说将是显而易见的，本文中所定义的一般原理可以在不脱离本发明的精神或范围的情况下，在其它实施例中实现。因此，本发明将不会被限制于本文所示的这些实施例，而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

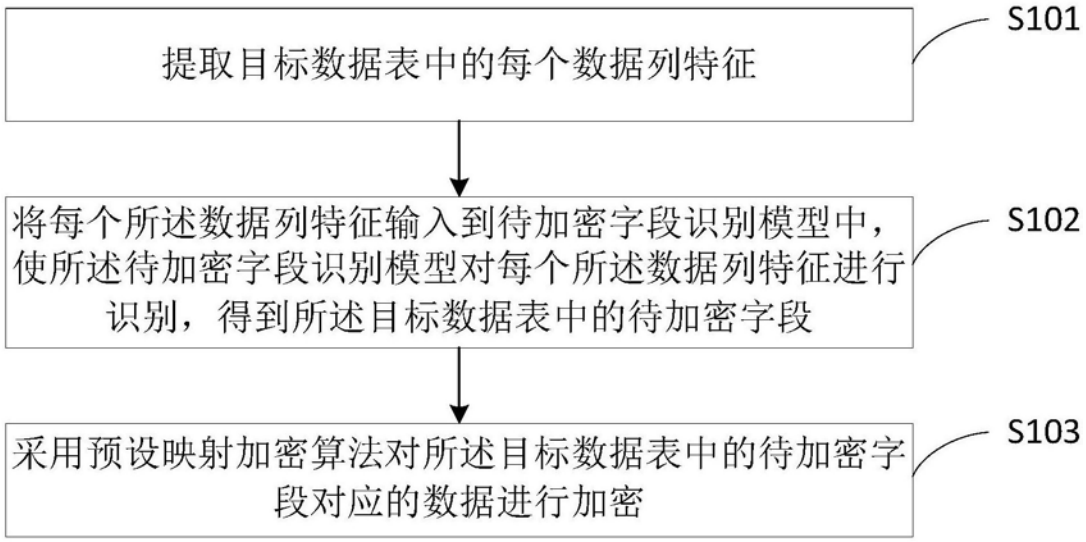
提取目标数据表中的每个数据列特征
S101
将每个所述数据列特征输入到待加密字段识别模型中，使所述待加密字段识别模型对每个所述数据列特征进行识别，得到所述目标数据表中的待加密字段
S102
采用预设映射加密算法对所述目标数据表中的待加密字段对应的数据进行加密
S103

图1

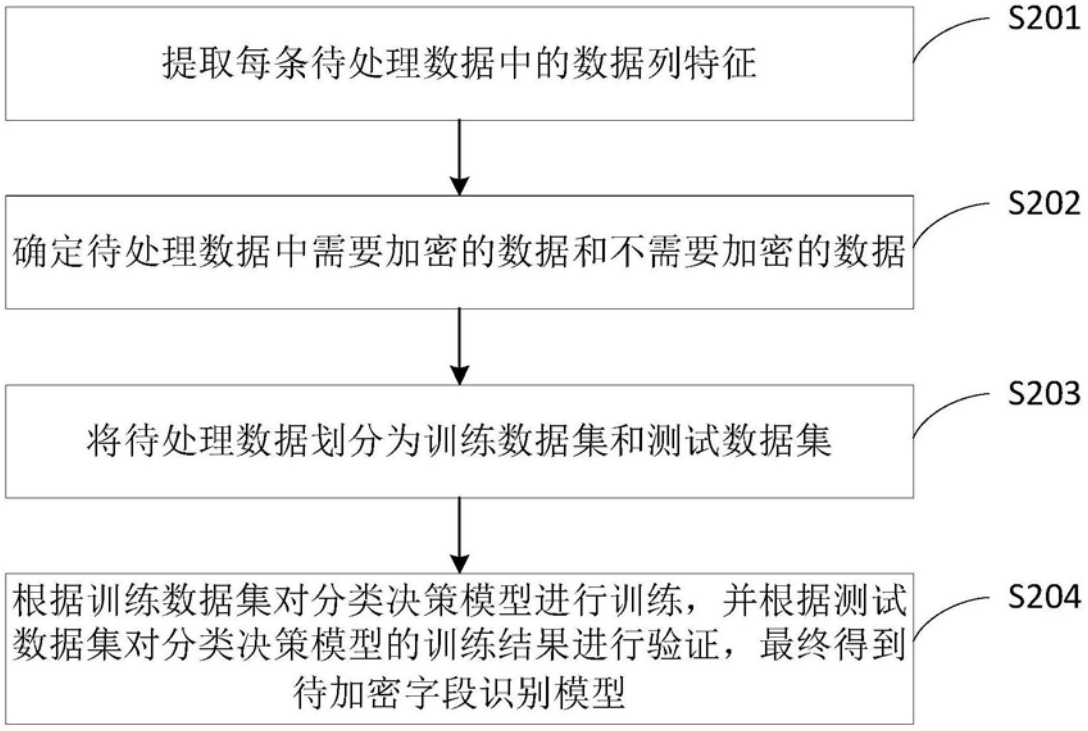
提取每条待处理数据中的数据列特征
S201
确定待处理数据中需要加密的数据和不需要加密的数据
S202
将待处理数据划分为训练数据集和测试数据集
S203
根据训练数据集对分类决策模型进行训练，并根据测试数据集对分类决策模型的训练结果进行验证，最终得到待加密字段识别模型
S204

图2

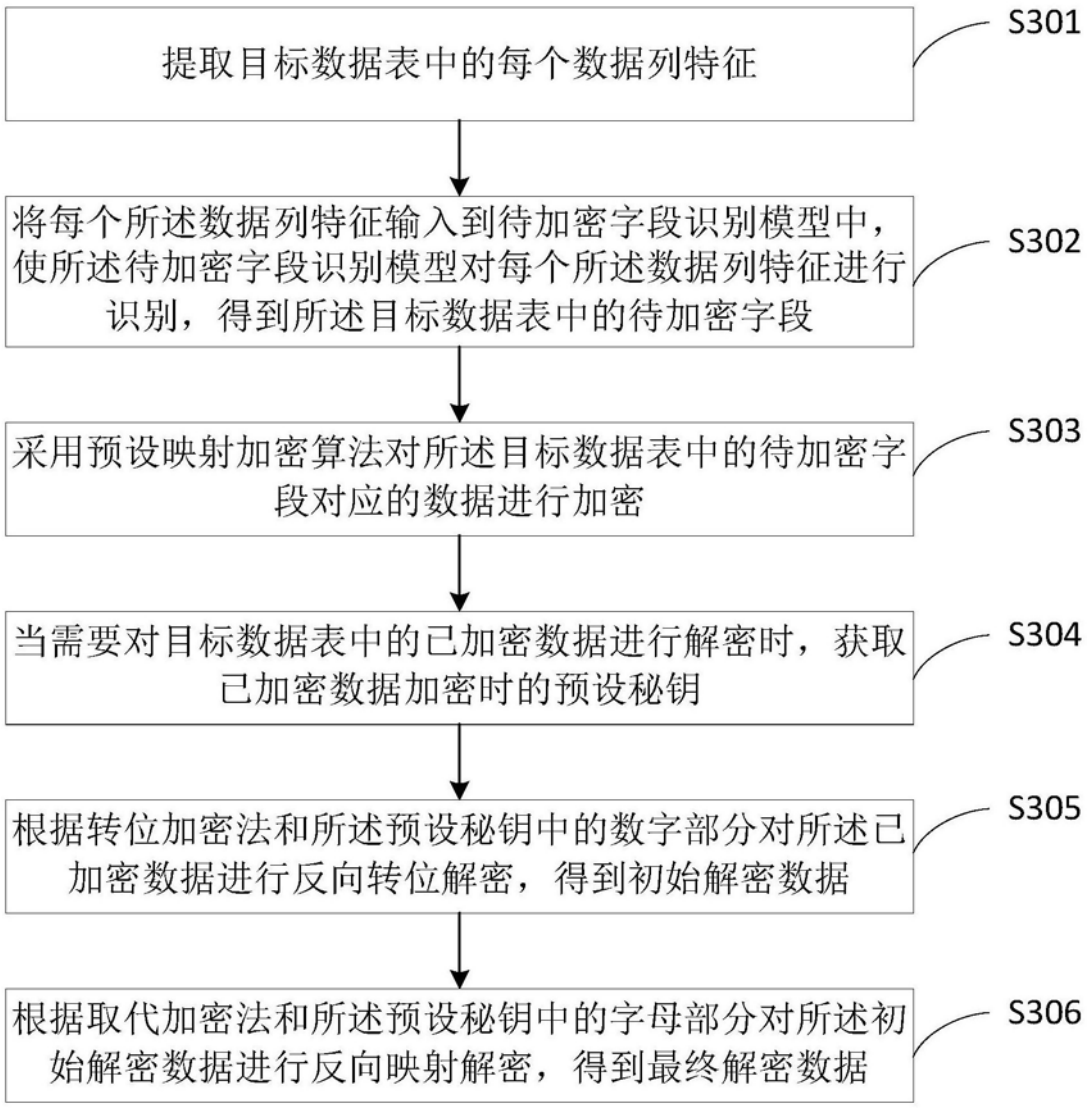
提取目标数据表中的每个数据列特征
S301
将每个所述数据列特征输入到待加密字段识别模型中，使所述待加密字段识别模型对每个所述数据列特征进行识别，得到所述目标数据表中的待加密字段
S302
采用预设映射加密算法对所述目标数据表中的待加密字段对应的数据进行加密
S303
当需要对目标数据表中的已加密数据进行解密时，获取已加密数据加密时的预设秘钥
S304
根据转位加密法和所述预设秘钥中的数字部分对所述已加密数据进行反向转位解密，得到初始解密数据
S305
根据取代加密法和所述预设秘钥中的字母部分对所述初始解密数据进行反向映射解密，得到最终解密数据
S306

图3

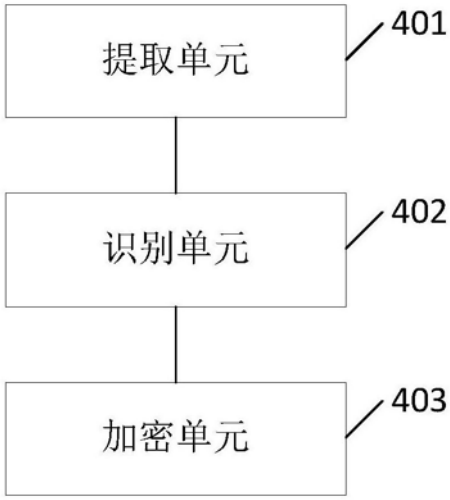

图4