(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 109784407 A
(43)申请公布日 2019.05.21

(21)申请号 201910043827.3

(22)申请日 2019.01.17

(71)申请人 京东数字科技控股有限公司
地址 100176 北京市北京经济技术开发区科创十一街18号C座2层221室

(72)发明人 范叶亮 马云龙 卢周

(74)专利代理机构 中原信达知识产权代理有限责任公司 11219
代理人 张一军 张效荣

(51)Int.Cl.
G06K 9/62(2006.01)
G06F 16/22(2019.01)
G06F 21/62(2013.01)

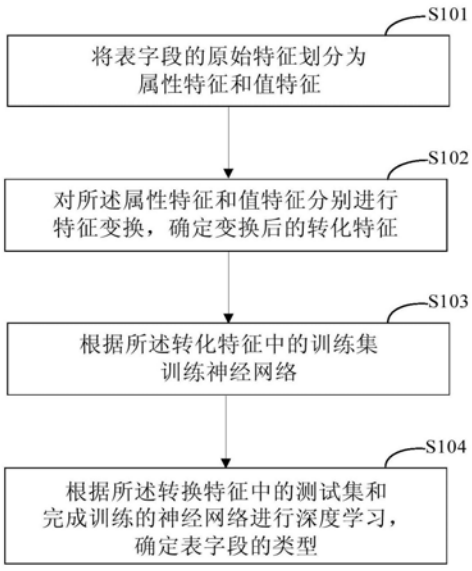
权利要求书2页 说明书13页 附图4页

(54)发明名称

确定表字段的类型的方法和装置

(57)摘要

本发明公开了确定表字段的类型的方法和装置，涉及计算机技术领域。该方法的一具体实施方式包括：将表字段的原始特征划分为属性特征和值特征；对所述属性特征和值特征分别进行特征变换，确定变换后的转化特征；根据所述转化特征中的训练集训练神经网络；根据所述转换特征中的测试集和完成训练的神经网络进行深度学习，确定表字段的类型。该实施方式解决了现有技术基于关键词匹配和传统机器学习方法的识别准确率和召回率都相对较低、人工成本过高的技术缺陷，进而达到深度学习更有针对性、充分利用表字段的原始特征使得确定的表字段的类型更准确的技术效果。



CN 109784407 A

1.一种确定表字段的类型的方法,其特征在于,包括:

将表字段的原始特征划分为属性特征和值特征;

对所述属性特征和值特征分别进行特征变换,确定变换后的转化特征;

根据所述转化特征中的训练集训练神经网络;

根据所述转换特征中的测试集和完成训练的神经网络进行深度学习,确定表字段的类型。

2.根据权利要求1所述的方法,其特征在于,所述属性特征包括:数值特征、分类特征、文本特征;

所述值特征包括文本特征。

3.根据权利要求2所述的方法,其特征在于,对所述属性特征和值特征分别进行特征变换,确定变换后的转化特征,包括:

将所述属性特征中的数值特征进行特征变换,得到的转化特征为宽度特征;

和/或,将所述属性特征中的文本特征和值特征进行特征变换,得到的转化特征为深文本特征;

和/或,将所述属性特征中的分类特征进行特征变换,得到的转化特征为深类别特征和/或宽度特征。

4.根据权利要求3所述的方法,其特征在于,将所述属性特征中的数值特征进行特征变换,得到的转化特征为宽度特征的变换公式为:

$$\text{wide\_feature} = \frac{min(\text{raw\_feature}, \text{max\_value})}{\text{max\_value}}$$

其中,wide_feature表示转换后的宽度特征,raw_feature表示原始特征,max_value表示该宽度特征的最大值,min表示所述数值特征与所述宽度特征的最大值中较小的数值。

5.根据权利要求4所述的方法,其特征在于,将所述属性特征中的分类特征进行特征变换,得到的转化特征为宽度特征,包括:

采用独热编码对所述分类特征进行编码,

将编码后的结果拼接为一个数值为0或1的向量;

拼接后的所述0或1的向量为宽度特征。

6.根据权利要求4所述的方法,其特征在于,将所述文本特征和值特征进行特征变换,得到的转化特征为深文本特征,包括:

将文本特征或值特征中的文本后加上终止字符;

针对文本特征设置预设长度;

当所述文本的加上终止字符的长度大于预设长度,则删除超出最大长度的部分,剩余部分为深文本特征;

当所述文本的加上终止字符的长度小于预设长度,则将不足预设长度的部分用补充字符补足得到深文本特征。

7.根据权利要求4所述的方法,其特征在于,将所述分类特征进行特征变换,得到的转化特征为深类别特征,包括:

将分类特征进行拼接;

将拼接的结果转化为一个向量;

所述向量为深类别特征。

8.根据权利要求3所述的方法，其特征在于，根据所述转化特征中的训练集训练神经网络，包括：

将所述训练集中的宽度特征作为训练神经网络中宽度网络的输入；

将所述训练集中的深文本特征和深类别特征作为训练神经网络中深度网络的输入；

根据所述宽度网络和深度网络，确定神经网络。

9.根据权利要求8所述的方法，其特征在于，所述宽度特征与所述深类别特征用于全连接神经网络进行深度学习；

所述深文本特征用于字符级的卷积神经网络进行深度学习。

10.根据权利要求3所述的方法，其特征在于，用于训练神经网络的函数模型为柔性最大值交叉熵损失函数。

11.根据权利要求5所述的方法，其特征在于，所述神经网络深度学习的训练集中的每个样本数据均由属性特征和值特征匹配确定。

12.根据权利要求1所述的方法，其特征在于，根据所述转化特征进行深度学习，确定表字段的类型，包括：

确定深度学习的预测结果；

确定所述预测结果的置信区间；

根据投票机制和最大置信区间，确定表字段的类型。

13.一种确定表字段的类型的装置，其特征在于，包括：

原始特征划分模块，用于将表字段的原始特征划分为属性特征和值特征；

转化特征模块，用于对所述属性特征和值特征分别进行特征变换，确定变换后的转化特征；

神经网络训练模块，用于根据所述转化特征中的训练集训练神经网络；

表字段的类型确定模块，用于根据所述转换特征中的测试集和完成训练的神经网络进行深度学习，确定表字段的类型。

14.一种确定表字段的类型的电子设备，其特征在于，包括：

一个或多个处理器；

存储装置，用于存储一个或多个程序，

当所述一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现如权利要求1-12中任一所述的方法。

15.一种计算机可读介质，其上存储有计算机程序，其特征在于，所述程序被处理器执行时实现如权利要求1-12中任一所述的方法。

确定表字段的类型的方法和装置

技术领域

[0001] 本发明涉及计算机技术领域，尤其涉及一种确定表字段的类型的方法和装置。

背景技术

[0002] 表字段的类型确定、以及表字段是否敏感的判断在存储到数据库之前非常重要，尤其是关于隐私保护或信息安全方面。故在将表字段加载进入数据库之前，需要对涉及敏感信息(包括但不限于:姓名，身份证号，手机号，银行卡号等)的表字段进行加密处理。

[0003] 现有技术采用基于关键字(例如:姓名，address，地址等)匹配和传统机器学习，加以人工判断辅助的方法对表字段的类型以及表字段是否为敏感字段进行识别。

[0004] 在实现本发明过程中，发明人发现现有技术中至少存在如下问题：

[0005] 1.基于关键词匹配和传统机器学习方法的识别准确率和召回率都相对较低。

[0006] 2.在识别准确率较低的情况下，加以人工判断辅助识别，人工成本过高。

发明内容

[0007] 有鉴于此，本发明实施例提供一种确定表字段的类型的方法和装置，能够达到深度学习更有针对性、充分利用表字段的原始特征使得确定的表字段的类型更准确的技术效果。

[0008] 为实现上述目的，根据本发明实施例的一个方面，提供了一种确定表字段的类型的方法，包括：

[0009] 将表字段的原始特征划分为属性特征和值特征；

[0010] 对所述属性特征和值特征分别进行特征变换，确定变换后的转化特征；

[0011] 根据所述转化特征中的训练集训练神经网络；

[0012] 根据所述转换特征中的测试集和完成训练的神经网络进行深度学习，确定表字段的类型。

[0013] 可选地，所述属性特征包括:数值特征、分类特征、文本特征；

[0014] 所述值特征包括文本特征。

[0015] 可选地，对所述属性特征和值特征分别进行特征变换，确定变换后的转化特征，包括：

[0016] 将所述属性特征中的数值特征进行特征变换，得到的转化特征为宽度特征；

[0017] 和/或，将所述属性特征中的文本特征和值特征进行特征变换，得到的转化特征为深文本特征；

[0018] 和/或，将所述属性特征中的分类特征进行特征变换，得到的转化特征为深类别特征和/或宽度特征。

[0019] 可选地，将所述属性特征中的数值特征进行特征变换，得到的转化特征为宽度特征的变换公式为：

[0020] $$\text{wide\_feature} = \frac{min(\text{raw\_feature}, \text{max\_value})}{\text{max\_value}}$$

[0021] 其中，wide_feature表示转换后的宽度特征，raw_feature表示原始特征，max_value表示该宽度特征的最大值，min表示所述数值特征与所述宽度特征的最大值中较小的数值。

[0022] 可选地，将所述属性特征中的分类特征进行特征变换，得到的转化特征为宽度特征，包括：

[0023] 采用独热编码对所述分类特征进行编码，

[0024] 将编码后的结果拼接为一个数值为0或1的向量；

[0025] 拼接后的所述0或1的向量为宽度特征。

[0026] 可选地，将所述文本特征和值特征进行特征变换，得到的转化特征为深文本特征，包括：

[0027] 将文本特征或值特征中的文本后加上终止字符；

[0028] 针对文本特征设置预设长度；

[0029] 当所述文本的加上终止字符的长度大于预设长度，则删除超出最大长度的部分，剩余部分为深文本特征；

[0030] 当所述文本的加上终止字符的长度小于预设长度，则将不足预设长度的部分用补充字符补足得到深文本特征。

[0031] 可选地，将所述分类特征进行特征变换，得到的转化特征为深类别特征，包括：

[0032] 将分类特征进行拼接；

[0033] 将拼接的结果转化为一个向量；

[0034] 所述向量为深类别特征。

[0035] 可选地，根据所述转化特征中的训练集训练神经网络，包括：

[0036] 将所述训练集中的宽度特征作为训练神经网络中宽度网络的输入；

[0037] 将所述训练集中的深文本特征和深类别特征作为训练神经网络中深度网络的输入；

[0038] 根据所述宽度网络和深度网络，确定神经网络。

[0039] 可选地，所述宽度特征与所述深类别特征用于全连接神经网络进行深度学习；

[0040] 所述深文本特征用于字符级的卷积神经网络进行深度学习。

[0041] 可选地，用于训练神经网络的函数模型为柔性最大值交叉熵损失函数。

[0042] 可选地，所述神经网络深度学习的训练集中的每个样本数据均由属性特征和值特征匹配确定。

[0043] 可选地，根据所述转化特征进行深度学习，确定表字段的类型，包括：

[0044] 确定深度学习的预测结果；

[0045] 确定所述预测结果的置信区间；

[0046] 根据投票机制和最大置信区间，确定表字段的类型。

[0047] 根据本发明实施例的再一个方面，提供了一种确定表字段的类型的装置，包括：

[0048] 原始特征划分模块，用于将表字段的原始特征划分为属性特征和值特征；

[0049] 转化特征模块，用于对所述属性特征和值特征分别进行特征变换，确定变换后的

转化特征；

[0050] 神经网络训练模块，用于根据所述转化特征中的训练集训练神经网络；

[0051] 表字段的类型确定模块，用于根据所述转换特征中的测试集和完成训练的神经网络进行深度学习，确定表字段的类型。

[0052] 根据本发明实施例的另一个方面，提供了一种确定表字段的类型电子设备，包括：

[0053] 一个或多个处理器；

[0054] 存储装置，用于存储一个或多个程序，

[0055] 当所述一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现本发明提供的确定表字段的类型方法。

[0056] 根据本发明实施例的还一个方面，提供了一种计算机可读介质，其上存储有计算机程序，所述程序被处理器执行时实现本发明提供的确定表字段的类型方法。

[0057] 上述发明中的一个实施例具有如下优点或有益效果：

[0058] 本发明通过采用表字段原始特征中的属性特征和值特征进行深度学习的技术手段，解决了现有技术基于关键词匹配和传统机器学习方法的识别准确率和召回率都相对较低、人工成本过高的技术缺陷，进而达到充分利用表字段的原始特征使得确定的表字段的类型更准确；

[0059] 通过对属性特征和值特征分别进行特征变换，使得变换后的转化数据可以应用于不同的神经网络做进一步训练，进而使得深度学习更有针对性、达到进一步提高准确率的技术效果，克服了现有技术确定表字段的类型不准确的缺陷。

[0060] 上述的非惯用的可选方式所具有的进一步效果将在下文中结合具体实施方式加以说明。

附图说明

[0061] 附图用于更好地理解本发明，不构成对本发明的不当限定。其中：

[0062] 图1是根据本发明实施例的确定表字段的类型的方法的主要流程的示意图；

[0063] 图2是根据本发明实施例的改进的宽度和深度网络结构；

[0064] 图3是根据本发明实施例的字符级别的卷积神经网络；

[0065] 图4是根据本发明实施例的确定表字段的类型的方法的训练与预测的具体流程图；

[0066] 图5是根据本发明实施例的确定表字段的类型的装置的主要模块的示意图；

[0067] 图6是本发明实施例可以应用于其中的示例性系统架构图；

[0068] 图7是适于用来实现本发明实施例的终端设备或服务器的计算机系统的结构示意图。

具体实施方式

[0069] 以下结合附图对本发明的示范性实施例做出说明，其中包括本发明实施例的各种细节以助于理解，应当将它们认为仅仅是示范性的。因此，本领域普通技术人员应当认识到，可以对这里描述的实施例做出各种改变和修改，而不会背离本发明的范围和精神。同样，为了清楚和简明，以下的描述中省略了对公知功能和结构的描述。

[0070] 图1是根据本发明实施例的一种确定表字段的类型的方法的主要流程的示意图，如图1所示，包括：

[0071] 步骤S101、将表字段的原始特征划分为属性特征和值特征；

[0072] 步骤S102、对所述属性特征和值特征分别进行特征变换，确定变换后的转化特征；

[0073] 步骤S103、根据所述转化特征中的训练集训练神经网络；

[0074] 步骤S104、根据所述转换特征中的测试集和完成训练的神经网络进行深度学习，确定表字段的类型。

[0075] 所述表字段的类型包括：敏感字段和非敏感字段。特别是，当表字段为敏感字段时，在数据仓库中的用途非常重要。例如，将明文数据加载到数据仓库之前，需要对涉及敏感信息(包括但不限于：姓名，身份证号，手机号，银行卡号等)的字段进行加密处理。所述非敏感字段就是除敏感字段之外的字段，包括：年份信息、时刻信息等。

[0076] 本发明通过表字段原始特征中的属性特征和值特征进行深度学习的技术手段，解决了现有技术基于关键词匹配和传统机器学习方法的识别准确率和召回率都相对较低、人工成本过高的技术缺陷，进而达到充分利用表字段的原始特征使得确定的表字段的类型更准确的技术效果。

[0077] 对属性特征和值特征分别进行特征变换，使得变换后的转化数据可以应用于不同的神经网络做进一步训练，进而使得深度学习更有针对性、达到进一步提高准确率的技术效果，克服了现有技术确定的表字段的类型不准确的缺陷。

[0078] 可选地，所述属性特征(也就是表字段的属性信息：例如：表名，表注释，表类型，字段名，字段注释，字段类型等)包括：数值特征(例如：1000，1.0等)、分类特征(例如："是"和"否"等)、文本特征(例如："这是一个字段的注释信息")；

[0079] 所述值特征包括文本特征。

[0080] 具体地，如表1的原始特征示例所示。例如，数据库名称为属性特征，其特征值为对应的测试数据库名称，进一步属于文本特征；值特征对应的文本特征可能包含一个数组，如，"值1"，"值2"，…，"值n"。

[0081] 表1中的其他示例不再赘述，均遵循上述分类原则。

[0082] 表1原始特征示例

[0083]

特征名称	特征分类	特征值	特征值类型
数据库名称	属性特征	测试数据库名称	文本特征
数据库注释	属性特征	测试数据库注释	文本特征
数据库注释长度	属性特征	7	数值特征
数据库类型	属性特征	MYSQL	分类特征
…	…	…	…
字段名称	属性特征	测试字段名称	文本特征
字段注释	属性特征	测试字段注释	文本特征
字段值	值特征	["值1"，"值2"，…，"值n"]	文本特征

[0084] 图2是根据本发明实施例的改进的宽度和深度网络结构。如图2所示，该宽度和深度网络的输入可以包括：宽度(wide)特征、深文本(Deep-Text)特征、深类别(Deep-

Category)特征。

[0085] 可选地,对所述属性特征和值特征分别进行特征变换,确定变换后的转化特征,包括:将所述属性特征中的数值特征进行特征变换,得到的转化特征为宽度(wide)特征;

[0086] 和/或,将所述属性特征中的文本特征和值特征进行特征变换,得到的转化特征为深文本(Deep-Text)特征;

[0087] 和/或,将所述属性特征中的分类特征进行特征变换,得到的转化特征为深类别(Deep-Category)特征和/或宽度(wide)特征。

[0088] 其中,可选地,将所述属性特征中的数值特征进行特征变换,得到的转化特征为宽度特征的变换公式为:

[0089] $$\text{wide\_feature} = \frac{min(\text{raw\_feature}, \text{max\_value})}{\text{max\_value}}$$

[0090] 其中,wide_feature表示转换后的宽度特征,raw_feature表示原始特征,max_value表示该宽度特征的最大值,min表示所述数值特征与所述最大值中较小的数值。

[0091] 通过上述对数值特征的特征变变换,得到变换后的结果在0到1之间,进而达到特征归一化的技术效果。也就是变换后的wide特征点的所有维度值均在0和1之间。可选地,wide特征可以用于全连接神经网络。

[0092] 可选地,将所述文本特征和值特征进行特征变换,得到的转化特征为深文本特征(Deep-Text),包括:

[0093] 将文本特征或值特征中的文本后加上终止字符;

[0094] 针对每个文本设置一个预设长度;

[0095] 当所述文本的加上终止字符的长度大于预设长度,则删除超出最大长度的部分;

[0096] 当所述文本的加上终止字符的长度小于预设长度,则将不足预设长度的部分用补充字符补足。

[0097] 下面以具体实施方式的形式介绍深文本特征的确定过程:

[0098] 假设预设长度为6;待转换的原始特征可以包括属性特征和值特征。

[0099] 如下表5所示,在示例1中,首先在原始特征后加上终止字符<EOS>,得到四个字符。由于补充了终止字符后仍没有达到预设长度的6个字符,故还需要用补充字符<PAD>进行补足直至达到预设长度。

[0100] 在示例2中,在原始特征加上终止字符后刚好达到预设长度,故无需添加补充字符。

[0101] 在示例3中,在原始特征加上终止字符后已经超出预设长度,故需要删除超出部分的字符。

[0102] 表2单个文本特征转换后的特征

[0103]

索引	1	2	3	4	5	6	7	8	9
示例1	短	文	本	<EOS>	<PAD>	<PAD>			
示例2	中	等	长	文	本	<EOS>			
示例3	超	级	长	的	文	本	值	。	<EOS>

[0104] 因为该文本特征设置的最大长度为6,因此上表中超出第6个字符点的部分将被删除。

[0105] 可选地,当深文本特征对应多个文本特征或值特征时,可以采用拼接的方式生成对应的深文本,其长度(或维度)为$\dim_{Deep-Text} = \sum_{i=1}^{C_{Deep-Text}} \max_len_i^{Deep-Text}$,其中$\max_len_i^{Deep-Text}$为每个文本特征设置的预设长度。

[0106] 下面以一具体实施例详细说明采用拼接的方式生成对应的深文本:

[0107] 在本实施例中,共有2个文本特征为:表6和表7,每个文本特征且均包含3个示例。其中,表6的预设长度为4,表7的预设长度为5。故转换的过程中需要是删除表6中第5个字符点后的部分、删除表7中第6个字符点后的部分。再将处理后的表6和表7做拼接处理,得到表8转换后的深文本特征。其中,表8的1-4列为第一个转换后的文本特征,5-9列为第二个转换后的文本特征。

[0108] 表3转换前的Deep-Text特征1

[0109]

索引	1	2	3	4	5	6
示例1	特	征	1	<EOS>		
示例2	还	是	特	征	<EOS>	
示例3	长	的	特	征	值	<EOS>

[0110] 表4转换前的Deep-Text特征2

[0111]

索引	1	2	3	4	5	6
示例1	特	征	2	<EOS>	<PAD>	
示例2	特	征	<EOS>	<PAD>	<PAD>	
示例3	也	是	长	特	征	<EOS>

[0112] 表5转换后的Deep-Text特征

[0113]

索引	1	2	3	4	5	6	7	8	9
示例1	特	征	1	<EOS>	特	征	2	<EOS>	<PAD>
示例2	还	是	特	征	特	征	<EOS>	<PAD>	<PAD>
示例3	长	的	特	征	也	是	长	特	征

[0114] 对于Deep-Text特征,其维度已经确定为$\dim_{Deep-Text}$,维护一个包含所有可能字符的字典表,该字典表将任意一个字符映射成为$\dim_{Text-Embedding}$维的向量,例如:

[0115] 表6一个字在字典表中的查询结果

[0116]

索引	1	2	3	…	…	i	…	…	$dim_{embedding}$
值	0.012	0.231	0.986	…	…	0.123	…	…	0.689

[0117] 该字典表中每个字对应的向量随机初始化，其具体的值在训练过程中不断学习更新。Deep-Text为文本型特征，对应着字符级的卷积神经网络，用于提取文本的内在信息。其中，图3是根据本发明实施例的字符级别的卷积神经网络。

[0118] 如图3所示，先将文本特征转化为特征向量，并且将多个文本他正向量拼接在一起进行多尺度卷积，再后对卷积的结果进行池化，最后通过神经网络的输出层输出训练的结果。

[0119] 可选地，采用独热编码对所述分类特征进行编码，

[0120] 将编码后的结果拼接为一个数值为0或1的向量；

[0121] 拼接后的所述0或1的向量为宽度特征。

[0122] 由于类别特征采用独热编码的方式能够更好地体现出分类特征对应的类别。可选地，将特征变换后的最大长度(或维度)设置为dim_{Wide}。

[0123] 如下所示，假设该特征的类型共有m种可能，针对第i种特征类型，将其转化为如下仅包含0和1的向量，如下表2中的值所示：

[0124] 表7包含m种可能类型的第i种类型的独热编码

[0125]

索引	1	2	…	i	…	m
值	0	0	…	1	…	0

[0126] 下面以一具体的实施例介绍包含数值特征与分类特征进行转换的具体操作：

[0127] 具体地，原始特征中有2个数值类型特征(f_1和f_2)，2个分类类型特征(f_3和f_4)，每个特征的统计信息如下：

[0128] 1)f_1的最大值为20，最小值为0

[0129] 2)f_2的最大值为100，最小值为0

[0130] 3)f_3的所有可能性为:A,B,共2种可能

[0131] 4)f_4的所有可能性为:A,B,C,共3种可能

[0132] 转换前的Wide特征示例如下表3所示：

[0133] 表8转换前Wide特征示例

[0134]

特征	f_1	f_2	f_3	f_4
值	3	78	B	A

[0135] 由于f_1和f_2为wide特征中的数值类型特征，则依照归一化的方式进行特征变换；f_3和f_4为wide特征中的分类类型特征，则依照上述独热编码的方式进行特征变换。故转换后的Wide特征示例如下表4所示：

[0136] 表9转换后Wide特征示例

[0137]

特征	f'_1	f'_2	f'^{1}_{3}	f'^{2}_{3}	f'^{1}_{4}	f'^{2}_{4}	f'^{3}_{4}
值	0.15	0.78	0	1	1	0	0

[0138] 其中，f'_1为f_1转换后的特征，f'_2为f_2转换后的特征，$\mathbf{f}'^{1}_{3}$和$\mathbf{f}'^{2}_{3}$为f_3转换后的特征，$\mathbf{f}'^{1}_{4}$，$\mathbf{f}'^{2}_{4}$和$\mathbf{f}'^{3}_{4}$为f_4转换后的特征。

[0139] 可选地，将所述分类特征进行特征变换，得到的转化特征为深类别(Deep-Category)特征，包括：

[0140] 将分类特征进行拼接；

[0141] 将拼接的结果转化为一个向量；

[0142] 所述向量为深类别特征。

[0143] 所述Deep-Category特征来源为属性特征中的分类类型。针对属性特征中分类类型的原始特征不做其他处理，仅将其拼接为一个长向量，其长度(或维度)为$dim_{Deep-Category}$。

[0144] 可选地，根据所述转化特征中的训练集训练神经网络，包括：

[0145] 将所述训练集中的宽度特征作为训练神经网络中宽度网络的输入；

[0146] 将所述训练集中的深文本特征和深类别特征作为训练神经网络中深度网络的输入；

[0147] 根据所述宽度网络和深度网络，确定神经网络。

[0148] 通过宽度网络和深度网络进行深度学习，使得整个模型的参数由宽度网络和深度网络共同影响。宽度网络通过全连接网络进行深度学习，深度网络通过嵌入式的卷积神经网络进行深度学习，使得学习的准确度更高、模型大小和复杂度均能够得到控制，进而整体学习的效果也能显著提升。

[0149] 可选地，所述宽度特征与所述深类别特征用于全连接神经网络进行深度学习；

[0150] 所述深文本特征用于字符级的卷积神经网络进行深度学习。

[0151] 由于转换后的Wide特征所有维度的值均在0和1之间，可以将Wide特征用于全连接的神经网络进行训练。

[0152] 文本特征对应Deep-Text特征，Deep-Text特征的维度已经确定为$dim_{Deep-Text}$。将文本特征转化为深文本特征可以依赖包含所有可能字符的字典表，该字典表将任意一个字符映射成为$dim_{Text-Embedding}$维的向量，例如：

[0153] 表10一个字在字典表中的查询结果

[0154]

索引	1	2	3	…	…	i	…	$dim_{embedding}$
值	0.012	0.231	0.986	…	…	0.123	…	0.689

[0155] 字典表中每个字对应的向量随机初始化，其具体的值在训练过程中不断学习更新。Deep-Text为文本型特征，可以用于字符级的卷积神经网络，进而方便提取文本的内在信息。

[0156] 如图3所示，Deep-Category特征对应一个包含所有类型的字典表，该字典表将一种类型映射成为$dim_{Category-Embedding}$维的向量。与Deep-Text特征不同的是，Deep-Category特

征用于全连接的神经网络。再后，将Wide特征，Deep-Text特征和Deep-Category特征的输出拼接为一个向量。最后再通过一个全连接层得到最终的输出。

[0157] 可选地，用于训练神经网络的函数模型为柔性最大值交叉熵损失函数。其中，采用柔性最大值交叉熵损失(softmax cross entropy)函数作为损失函数，利用随机梯度下降算法对模型进行训练函数模型。

[0158] 可选地，所述神经网络的样本由属性特征和值特征匹配确定。

[0159] 由于原始特征中必然包含值特征，可以将值特征中的一个数值与属性特征结合训练神经网络。

[0160] 可选地，将所述文本特征和值特征进行特征变换，得到的转化特征为深文本特征，包括：

[0161] 将文本特征或值特征中的文本后加上终止字符；

[0162] 针对文本特征设置预设长度；

[0163] 当所述文本的加上终止字符的长度大于预设长度，则删除超出最大长度的部分，剩余部分为深文本特征；

[0164] 当所述文本的加上终止字符的长度小于预设长度，则将不足预设长度的部分用补充字符补足得到深文本特征。

[0165] 可选地，根据所述转化特征进行深度学习，确定表字段的类型，包括：

[0166] 确定深度学习的预测结果；

[0167] 确定所述预测结果的置信区间；

[0168] 根据投票机制和最大置信区间，确定表字段的类型。

[0169] 具体地，利用表字段的属性特征和n个值特征(即取不为空的n个值)可以构造n个样本。对于判断一个表字段是否为敏感字段，利用训练好的模型进行预测分别对这n个样本进行预测。最后利用投票机制确定该表字段的敏感类型(即预测类型的概率最大的类型为该表字段的敏感类型)。

[0170] 具体地，假设预测结果如下表所示：

[0171] 表11一组示例预测结果

[0172]

组	1	2	3	…	i	…	n
预测	手机号	手机号	固话	…	手机号	…	手机号

[0173] 假设预测结果中仅第3组特征预测结果为“固话”，其他组别均预测为“手机号”，则利用投票机制确定该表字段最终的预测结果为“手机号”。所述投票的机制，即预测类型数目最多的分类标签即为最终的预测结果。

[0174] 可选地，计算最终预测结果的置信度，所述置信度的公式为：

[0175] $$conf = \frac{counter(max\_freq\_label)}{n}$$

[0176] 其中，max_freq_label为出现频次最高的分类标签；counter为计数算子，即计算出现的频次；n为输入特征的组数，conf值越大，表示模型结果的可信度越高；conf为该预测结果的置信度。在使用过程中可以通过调节conf值的大小在准确率和覆盖率之间进行权衡。

[0177] 下面以一具体实施方式说明本发明的具体流程。

[0178] 图4是根据本发明实施例的确定表字段的类型的方法的训练与预测的具体流程图。如图4所示：

[0179] 将原始特征进行特征变换。其中需要将原始特征进行分类，具体地分为属性特征和值特征。得到的转换后的特征包括：wide特征、Deep-Text特征、Deep-Category特征。根据所述wide特征、Deep-Text特征、Deep-Category特征进行深度学习，得到Wide&Deep模型。

[0180] 当用户发出预测请求时，也需要对待预测的原始特征进行特征变换，并将转化后的特征用于Wide&Deep模型确定模型输出。最后根据投票机制确定最终的预测响应。

[0181] 图5是根据本发明实施例的确定表字段的类型的装置的主要模块的示意图。如图所示，一种确定表字段的类型的装置500，包括：

[0182] 模块501、原始特征划分模块，用于将表字段的原始特征划分为属性特征和值特征；

[0183] 模块502、转化特征模块，用于对所述属性特征和值特征分别进行特征变换，确定变换后的转化特征；

[0184] 模块503、神经网络训练模块，用于根据所述转化特征中的训练集训练神经网络；

[0185] 模块504、表字段的类型确定模块，用于根据所述转换特征中的测试集和完成训练的神经网络进行深度学习，确定表字段的类型。

[0186] 可选地，所述属性特征包括：数值特征、分类特征、文本特征；

[0187] 所述值特征包括文本特征。

[0188] 可选地，对所述属性特征和值特征分别进行特征变换，确定变换后的转化特征，包括：

[0189] 将所述属性特征中的数值特征进行特征变换，得到的转化特征为宽度特征；

[0190] 和/或，将所述属性特征中的文本特征和值特征进行特征变换，得到的转化特征为深文本特征；

[0191] 和/或，将所述属性特征中的分类特征进行特征变换，得到的转化特征为深类别特征和/或宽度特征。

[0192] 可选地，将所述属性特征中的数值特征进行特征变换，得到的转化特征为宽度特征的变换公式为：

[0193] $$\text{wide\_feature} = \frac{min(\text{raw\_feature}, \text{max\_value})}{\text{max\_value}}$$

[0194] 其中，wide_feature表示转换后的宽度特征，raw_feature表示原始特征，max_value表示该宽度特征的最大值，min表示所述数值特征与所述宽度特征的最大值中较小的数值。

[0195] 可选地，将所述属性特征中的分类特征进行特征变换，得到的转化特征为宽度特征，包括：

[0196] 采用独热编码对所述分类特征进行编码，

[0197] 将编码后的结果拼接为一个数值为0或1的向量；

[0198] 拼接后的所述0或1的向量为宽度特征。

[0199] 可选地，将所述文本特征和值特征进行特征变换，得到的转化特征为深文本特征，

包括：

[0200] 将文本特征或值特征中的文本后加上终止字符；

[0201] 针对文本特征设置预设长度；

[0202] 当所述文本的加上终止字符的长度大于预设长度，则删除超出最大长度的部分，剩余部分为深文本特征；

[0203] 当所述文本的加上终止字符的长度小于预设长度，则将不足预设长度的部分用补充字符补足得到深文本特征。

[0204] 可选地，将所述分类特征进行特征变换，得到的转化特征为深类别特征，包括：

[0205] 将分类特征进行拼接；

[0206] 将拼接的结果转化为一个向量；

[0207] 所述向量为深类别特征。

[0208] 可选地，根据所述转化特征中的训练集训练神经网络，包括：

[0209] 将所述训练集中的宽度特征作为训练神经网络中宽度网络的输入；

[0210] 将所述训练集中的深文本特征和深类别特征作为训练神经网络中深度网络的输入；

[0211] 根据所述宽度网络和深度网络，确定神经网络。

[0212] 可选地，所述宽度特征与所述深类别特征用于全连接神经网络进行深度学习；

[0213] 所述深文本特征用于字符级的卷积神经网络进行深度学习。

[0214] 可选地，用于训练神经网络的函数模型为柔性最大值交叉熵损失函数。

[0215] 可选地，所述神经网络深度学习的训练集中的每个样本数据均由属性特征和值特征匹配确定。

[0216] 可选地，根据所述转化特征进行深度学习，确定表字段的类型，包括：

[0217] 确定深度学习的预测结果；

[0218] 确定所述预测结果的置信区间；

[0219] 根据投票机制和最大置信区间，确定表字段的类型。

[0220] 图6示出了可以应用本发明实施例的确定表字段的类型方法或确定表字段的类型装置的示例性系统架构600。

[0221] 如图6所示，系统架构600可以包括终端设备601、602、603，网络604和服务器605。网络604用以在终端设备601、602、603和服务器605之间提供通信链路的介质。网络604可以包括各种连接类型，例如有线、无线通信链路或者光纤电缆等等。

[0222] 用户可以使用终端设备601、602、603通过网络604与服务器605交互，以接收或发送消息等。终端设备601、602、603上可以安装有各种通讯客户端应用，例如购物类应用、网页浏览器应用、搜索类应用、即时通信工具、邮箱客户端、社交平台软件等(仅为示例)。

[0223] 终端设备601、602、603可以是具有显示屏并且支持网页浏览的各种电子设备，包括但不限于智能手机、平板电脑、膝上型便携计算机和台式计算机等等。

[0224] 服务器605可以是提供各种服务的服务器，例如对用户利用终端设备601、602、603所浏览的购物类网站提供支持的后台管理服务器(仅为示例)。后台管理服务器可以对接收到的产品信息查询请求等数据进行分析等处理，并将处理结果(例如目标推送信息、产品信息--仅为示例)反馈给终端设备。

[0225] 需要说明的是，本发明实施例所提供的确定表字段的类型方法一般由服务器605执行，相应地，确定表字段的类型装置一般设置于服务器605中。

[0226] 应该理解，图6中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要，可以具有任意数目的终端设备、网络和服务器。

[0227] 下面参考图7，其示出了适于用来实现本发明实施例的终端设备的计算机系统700的结构示意图。图7示出的终端设备仅仅是一个示例，不应对本发明实施例的功能和使用范围带来任何限制。

[0228] 如图7所示，计算机系统700包括中央处理模块(CPU)701，其可以根据存储在只读存储器(ROM)702中的程序或者从存储部分708加载到随机访问存储器(RAM)703中的程序而执行各种适当的动作和处理。在RAM 703中，还存储有系统700操作所需的各种程序和数据。CPU 701、ROM 702以及RAM 703通过总线704彼此相连。输入/输出(I/O)接口705也连接至总线704。

[0229] 以下部件连接至I/O接口705：包括键盘、鼠标等的输入部分706；包括诸如阴极射线管(CRT)、液晶显示器(LCD)等以及扬声器等的输出部分707；包括硬盘等的存储部分708；以及包括诸如LAN卡、调制解调器等的网络接口卡的通信部分709。通信部分709经由诸如因特网的网络执行通信处理。驱动器710也根据需要连接至I/O接口705。可拆卸介质711，诸如磁盘、光盘、磁光盘、半导体存储器等等，根据需要安装在驱动器710上，以便于从其上读出的计算机程序根据需要被安装入存储部分708。

[0230] 特别地，根据本发明公开的实施例，上文参考流程图描述的过程可以被实现为计算机软件程序。例如，本发明公开的实施例包括一种计算机程序产品，其包括承载在计算机可读介质上的计算机程序，该计算机程序包含用于执行流程图所示的方法的程序代码。在这样的实施例中，该计算机程序可以通过通信部分709从网络上被下载和安装，和/或从可拆卸介质711被安装。在该计算机程序被中央处理模块(CPU)701执行时，执行本发明的系统中限定的上述功能。

[0231] 需要说明的是，本发明所示的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。计算机可读存储介质的更具体的例子可以包括但不限于：具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑磁盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本发明中，计算机可读存储介质可以是任何包含或存储程序的有形介质，该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本发明中，计算机可读的信号介质可以包括在基带中或者作为载波一部分传播的数据信号，其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式，包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质，该计算机可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含的程序代码可以用任何适当的介质传输，包括但不限于：无线、电线、光缆、RF等等，或者上述的任意合适的组合。

[0232] 附图中的流程图和框图，图示了按照本发明各种实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分，上述模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意，在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个接连地表示的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的是，框图或流程图中的每个方框、以及框图或流程图中的方框的组合，可以用执行规定的功能或操作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

[0233] 描述于本发明实施例中所涉及到的模块可以通过软件的方式实现，也可以通过硬件的方式来实现。所描述的模块也可以设置在处理器中，例如，可以描述为：一种处理器包括发送模块、获取模块、确定模块和第一处理模块。其中，这些模块的名称在某种情况下并不构成对该模块本身的限定，例如，发送模块还可以被描述为“向所连接的服务端发送图片获取请求的模块”。

[0234] 作为另一方面，本发明还提供了一种计算机可读介质，该计算机可读介质可以是上述实施例中描述的设备中所包含的；也可以是单独存在，而未装配入该设备中。上述计算机可读介质承载有一个或者多个程序，当上述一个或者多个程序被一个该设备执行时，使得该设备包括：

[0235] 将表字段的原始特征划分为属性特征和值特征；

[0236] 对所述属性特征和值特征分别进行特征变换，确定变换后的转化特征；

[0237] 根据所述转化特征中的训练集训练神经网络；

[0238] 根据所述转换特征中的测试集和完成训练的神经网络进行深度学习，确定表字段的类型。

[0239] 根据本发明实施例的技术方案，可以达到如下有益效果：

[0240] 本发明通过采用表字段的原始特征中的属性特征和值特征进行深度学习确定训练模型的技术手段，解决了现有技术基于关键词匹配和传统机器学习方法的识别准确率和召回率都相对较低、人工成本过高的技术缺陷，进而达到充分利用表字段的原始特征使得确定的表字段的类型更准确；

[0241] 通过对属性特征和值特征分别进行特征变换，使得变换后的转化数据可以应用于不同的神经网络做进一步训练，进而使得深度学习更有针对性、进一步提高准确率的技术效果，克服了现有技术确定的表字段的类型不准确的缺陷。

[0242] 上述具体实施方式，并不构成对本发明保护范围的限制。本领域技术人员应该明白的是，取决于设计要求和其他因素，可以发生各种各样的修改、组合、子组合和替代。任何在本发明的精神和原则之内所作的修改、等同替换和改进等，均应包含在本发明保护范围之内。

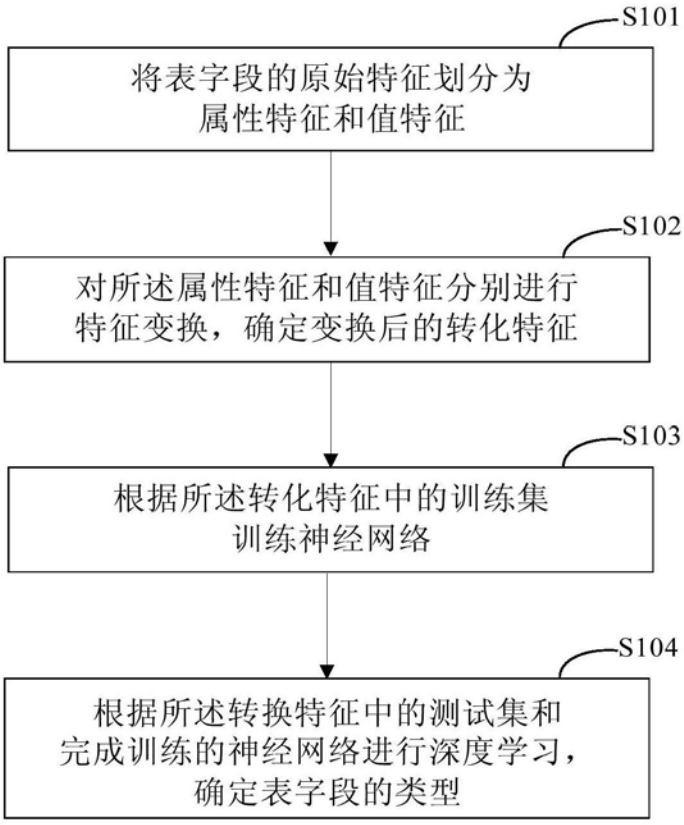

S101
将表字段的原始特征划分为
属性特征和值特征
S102
对所述属性特征和值特征分别进行
特征变换，确定变换后的转化特征
S103
根据所述转化特征中的训练集
训练神经网络
S104
根据所述转换特征中的测试集和
完成训练的神经网络进行深度学习，
确定表字段的类型

图1

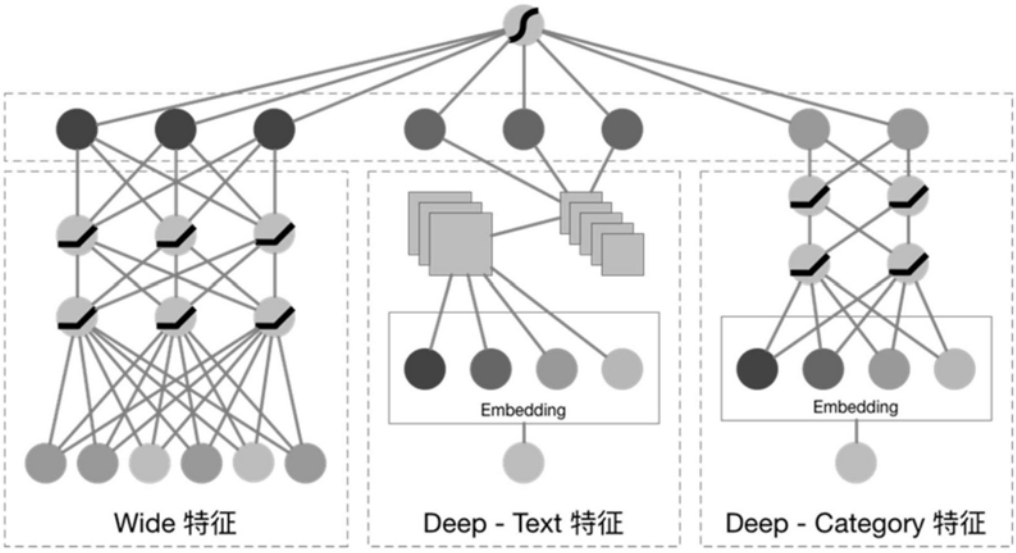

Embedding
Embedding
Wide 特征
Deep - Text 特征
Deep - Category 特征

图2

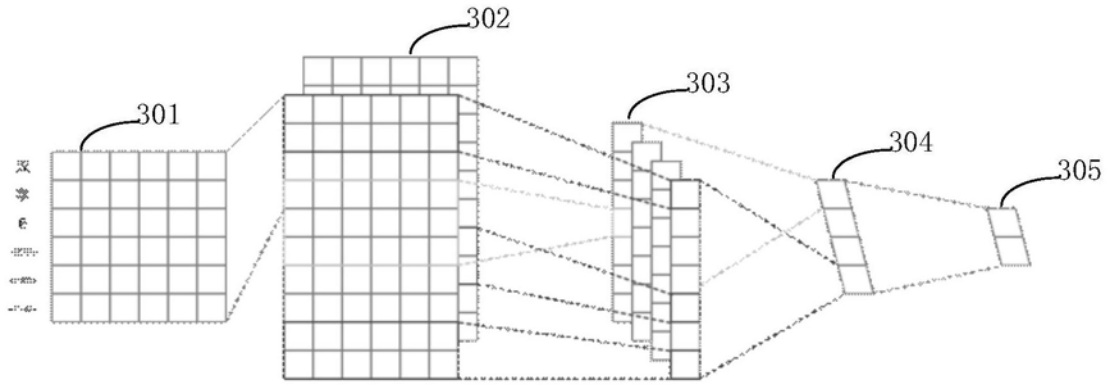

302
301
303
304
305

图3

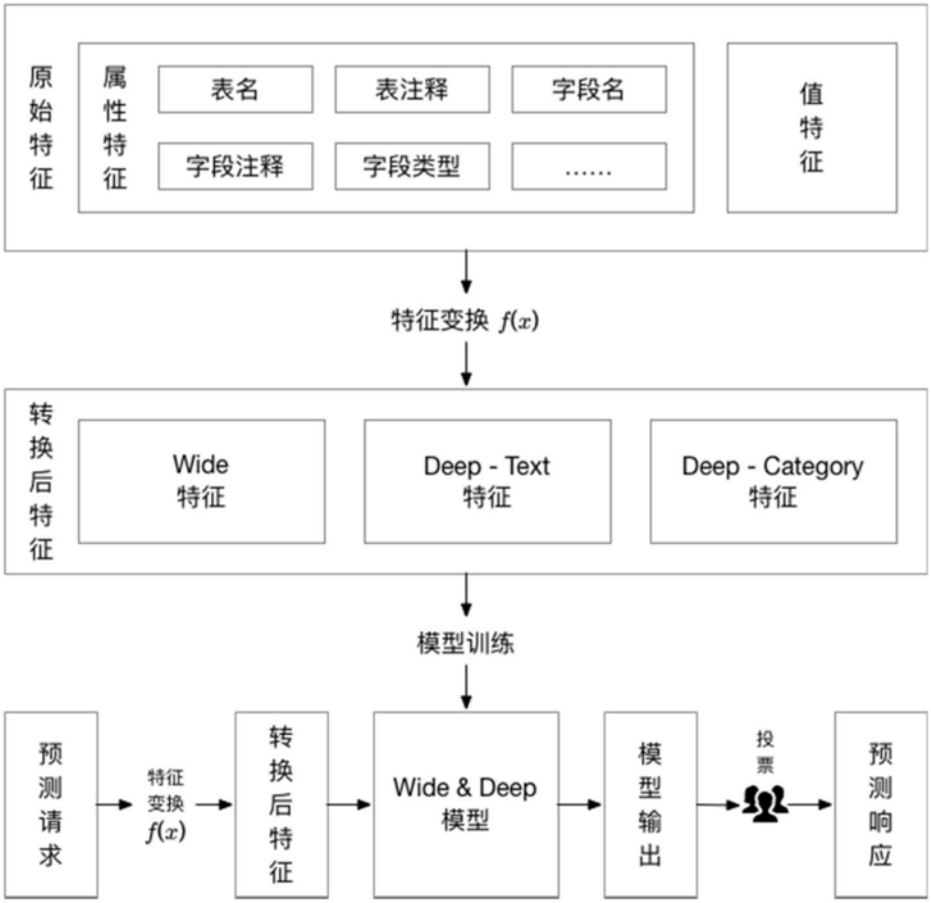

原始特征
属性特征
表名
表注释
字段名
字段注释
字段类型
.....
值特征
特征变换 $f(x)$
转换后特征
Wide 特征
Deep - Text 特征
Deep - Category 特征
模型训练
预测请求
特征变换 $f(x)$
转换后特征
Wide & Deep 模型
模型输出
投票
预测响应

图4

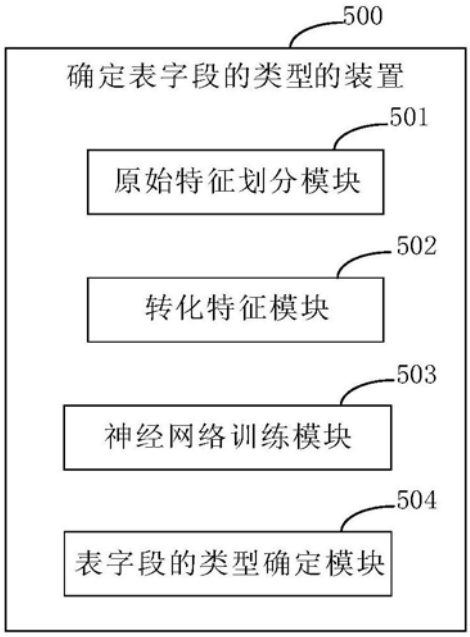

500
确定表字段的类型的装置
501
原始特征划分模块
502
转化特征模块
503
神经网络训练模块
504
表字段的类型确定模块

图5

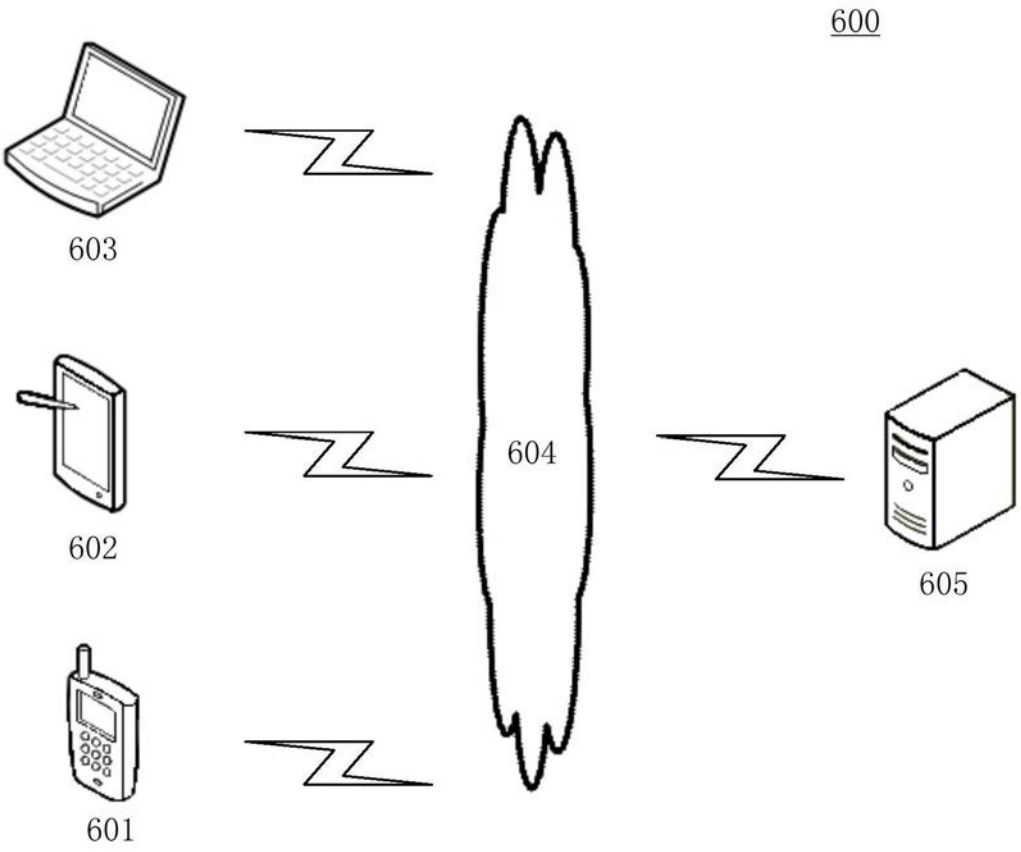

600
603
602
601
604
605

图6

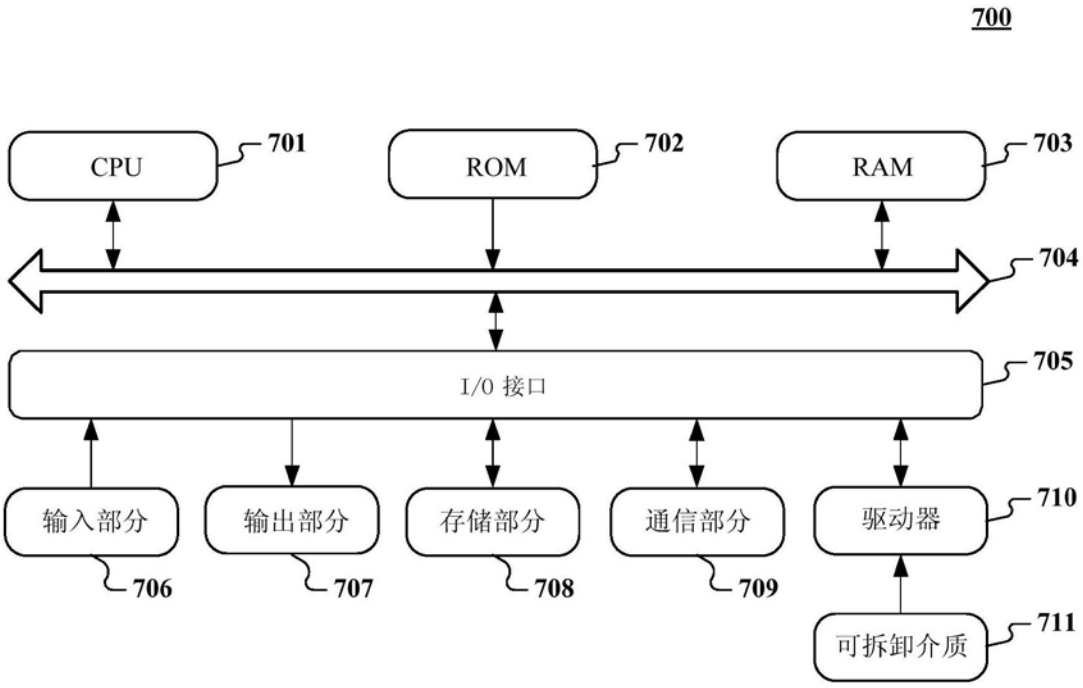

700
CPU
701
ROM
702
RAM
703
704
I/O 接口
705
输入部分
706
输出部分
707
存储部分
708
通信部分
709
驱动器
710
可拆卸介质
711

图7