(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111104566 B
(45)授权公告日 2023.07.21

(21)申请号 201911362985.1
(22)申请日 2019.12.26
(65)同一申请的已公布的文献号
申请公布号 CN 111104566 A
(43)申请公布日 2020.05.05
(73)专利权人 腾讯科技(深圳)有限公司
地址 518057 广东省深圳市南山区高新区科技中一路腾讯大厦35层
(72)发明人 李伟
(74)专利代理机构 北京三高永信知识产权代理有限责任公司 11138
专利代理师 郭新禹
(51)Int.Cl.
G06F 16/901(2019.01)
G06N 20/00(2019.01)
G06N 20/20(2019.01)

(56)对比文件
CN 107633088 A,2018.01.26
CN 109656930 A,2019.04.19
CN 109919084 A,2019.06.21
CN 109933644 A,2019.06.25
CN 110134678 A,2019.08.16
US 2015220684 A1,2015.08.06
US 2019034882 A1,2019.01.31
US 8341417 B1,2012.12.25

审查员 郭坚

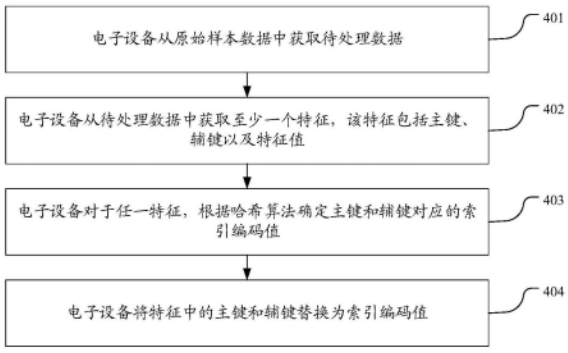
权利要求书2页 说明书13页 附图6页

(54)发明名称

特征索引编码方法、装置、电子设备及存储介质

(57)摘要

本申请提供了一种特征索引编码方法、装置、电子设备及存储介质，属于机器学习技术领域。所述方法包括：从待处理数据中获取至少一个特征，所述特征包括主键、辅键以及特征值；对于任一特征，根据哈希算法确定所述主键和所述辅键对应的索引编码值；将所述特征中的主键和辅键替换为所述索引编码值。通过哈希计算对特征的主键和辅键进行处理，确定对应的索引编码值，从而不需要对所有的待处理数据进行特征统计，一次计算即可为所有的待处理数据中特征的键值创建索引，降低了算法运行的时间复杂度，提高了数据的处理效率。



CN 111104566 B

1.一种特征索引编码方法,其特征在于,所述方法包括:

从待处理数据中获取至少一个特征,所述特征包括主键、辅键以及特征值,所述待处理数据由原始样本数据按照特征类别进行拆分得到;

对于任一特征,在所述特征为连续型特征的情况下,根据哈希算法确定所述主键对应的第一编码值;将目标占位符作为所述辅键对应的第二编码值;

在所述特征为离散特征的情况下,根据所述哈希算法确定所述主键对应的第一编码值和所述辅键对应的第二编码值;

将所述第一编码值和所述第二编码值进行拼接,得到索引编码值;

将所述特征中的主键和辅键替换为所述索引编码值。

2.根据权利要求1所述的方法,其特征在于,所述根据哈希算法确定所述主键对应的第一编码值和所述辅键对应的第二编码值,包括:

根据同一哈希算法,采用相同的参数分别确定所述主键对应的第一编码值和所述辅键对应的第二编码值。

3.根据权利要求1所述的方法,其特征在于,所述根据哈希算法确定所述主键对应的第一编码值和所述辅键对应的第二编码值,包括:

根据同一哈希算法,采用不同的参数分别确定所述主键对应的第一编码值和所述辅键对应的第二编码值。

4.根据权利要求1所述的方法,其特征在于,所述根据哈希算法确定所述主键对应的第一编码值和所述辅键对应的第二编码值,包括:

根据第一哈希算法确定所述主键对应的第一编码值;

根据第二哈希算法确定所述辅键对应的第二编码值,所述第一哈希算法和所述第二哈希算法为不同的哈希算法。

5.根据权利要求1所述的方法,其特征在于,所述哈希算法为MurmurHash3算法。

6.根据权利要求1-5任一项权利要求所述的方法,其特征在于,所述第一编码值位于所述索引编码值的尾部,所述第二编码值位于所述索引编码值的头部。

7.根据权利要求1所述的方法,其特征在于,所述从待处理数据中获取至少一个特征,包括:

对所述待处理数据中包括的字符串进行分割,得到多个字符串;

将包括至少一个目标字符的字符串作为所述特征,所述目标字符用于分隔所述主键、所述辅键和所述特征值。

8.根据权利要求1所述的方法,其特征在于,所述原始样本数据包括用户画像特征、用户行为特征、物品画像特征中的至少一种。

9.一种特征索引编码装置,其特征在于,所述装置包括:

获取模块,用于从待处理数据中获取至少一个特征,所述特征包括主键、辅键以及特征值,所述待处理数据由原始样本数据按照特征类别进行拆分得到;

确定模块,用于对于任一特征,在所述特征为连续型特征的情况下,根据哈希算法确定所述主键对应的第一编码值;将目标占位符作为所述辅键对应的第二编码值;在所述特征为离散特征的情况下,根据所述哈希算法确定所述主键对应的第一编码值和所述辅键对应的第二编码值;将所述第一编码值和所述第二编码值进行拼接,得到索引编码值;

替换模块，用于将所述特征中的主键和辅键替换为所述索引编码值。

10.根据权利要求9所述的装置，其特征在于，所述确定模块，用于根据同一哈希算法，采用相同的参数分别确定所述主键对应的第一编码值和所述辅键对应的第二编码值。

11.根据权利要求9所述的装置，其特征在于，所述确定模块，用于根据同一哈希算法，采用不同的参数分别确定所述主键对应的第一编码值和所述辅键对应的第二编码值；将所述第一编码值和所述第二编码值进行拼接，得到所述索引编码值。

12.根据权利要求9所述的装置，其特征在于，所述确定模块，用于根据第一哈希算法确定所述主键对应的第一编码值；根据第二哈希算法确定所述辅键对应的第二编码值，所述第一哈希算法和所述第二哈希算法为不同的哈希算法；将所述第一编码值和所述第二编码值进行拼接，得到所述索引编码值。

13.一种电子设备，其特征在于，所述电子设备包括处理器和存储器，所述存储器用于存储至少一段程序代码，所述至少一段程序代码由所述处理器加载并执行权利要求1至8任一权利要求所述的特征索引编码方法。

14.一种存储介质，其特征在于，所述存储介质用于存储至少一段程序代码，所述至少一段程序代码用于执行权利要求1至8任一权利要求所述的特征索引编码方法。

特征索引编码方法、装置、电子设备及存储介质

技术领域

[0001] 本申请涉及机器学习技术领域，特别涉及一种特征索引编码方法、装置、电子设备及存储介质。

背景技术

[0002] 在使用机器学习进行建模时，首先要做的一项工作是收集样本数据，使用收集到的样本数据来进行模型训练。通常情况下，收集样本数据大多是汇总用户行为、用户画像、物品画像及各类基于人类先验知识的统计类数据等。收集到的样本数据往往是如图1所示的明文数据。由于电子设备可以对数值进行计算，而无法对明文数据进行计算，因此需要对样本数据进行处理，即将明文数据转换为向量，再交由电子设备进行处理。将原始的明文数据转为向量的过程可以称为特征索引编码。

[0003] 相关技术中，通常使用统计类的方法来实现特征索引编码，即先对样本数据中特征的键值进行统计，为每个特征的键值分配全局唯一的索引标识。

[0004] 然而，在样本数据的数据量级非常大时，如果采用统计类的方法对所有样本数据中特征的键值进行统计，会花费大量的时间，甚至有可能建立特征索引编码的时间超过了模型训练的时间，导致样本数据处理的效率低。

发明内容

[0005] 本申请实施例提供了一种特征索引编码方法、装置、电子设备及存储介质，可以降低算法运行的时间复杂度，提高数据的处理效率。所述技术方案如下：

[0006] 一方面，提供了一种特征索引编码方法，其特征在于，所述方法包括：

[0007] 从待处理数据中获取至少一个特征，所述特征包括主键、辅键以及特征值；

[0008] 对于任一特征，根据哈希算法确定所述主键和所述辅键对应的索引编码值；

[0009] 将所述特征中的主键和辅键替换为所述索引编码值。

[0010] 另一方面，提供了一种特征索引编码装置，其特征在于，所述装置包括：

[0011] 获取模块，用于从待处理数据中获取至少一个特征，所述特征包括主键、辅键以及特征值；

[0012] 确定模块，用于对于任一特征，根据哈希算法确定所述主键和所述辅键对应的索引编码值；

[0013] 替换模块，用于将所述特征中的主键和辅键替换为所述索引编码值。

[0014] 在一种可选的实现方式中，所述特征为离散型特征；

[0015] 所述确定模块，还用于根据同一哈希算法，采用相同的参数分别确定所述主键对应的第一编码值和所述辅键对应的第二编码值；将所述第一编码值和所述第二编码值进行拼接，得到所述索引编码值。

[0016] 在一种可选的实现方式中，所述特征为离散型特征；

[0017] 所述确定模块，还用于根据同一哈希算法，采用不同的参数分别确定所述主键对

应的第一编码值和所述辅键对应的第二编码值;将所述第一编码值和所述第二编码值进行拼接,得到所述索引编码值。

[0018] 在一种可选的实现方式中,所述特征为离散型特征;

[0019] 所述确定模块,还用于根据第一哈希算法确定所述主键对应的第一编码值;根据第二哈希算法确定所述辅键对应的第二编码值,所述第一哈希算法和所述第二哈希算法为不同的哈希算法;将所述第一编码值和所述第二编码值进行拼接,得到所述索引编码值。

[0020] 在一种可选的实现方式中,所述特征为连续型特征;

[0021] 所述确定模块,还用于根据所述哈希算法确定所述主键对应的第一编码值;

[0022] 将目标占位符作为所述辅键对应的第二编码值;将所述第一编码值和所述第二编码值进行拼接,得到所述索引编码值。

[0023] 在一种可选的实现方式中,所述哈希算法为MurmurHash3算法。

[0024] 在一种可选的实现方式中,所述第一编码值位于所述索引编码值的尾部,所述第二编码值位于所述索引编码值的头部。

[0025] 在一种可选的实现方式中,所述获取模块,还用于对所述待处理数据中包括的字符串进行分割,得到多个字符串;将包括至少一个目标字符的字符串作为所述特征,所述目标字符用于分隔所述主键、所述辅键和所述特征值。

[0026] 在一种可选的实现方式中,所述装置还包括:

[0027] 所述获取模块,还用于获取原始样本数据,所述原始样本数据包括用户画像特征、用户行为特征、物品画像特征中的至少一种;

[0028] 拆分模块,用于对原始样本数据按照特征类别进行拆分,得到至少一种所述待处理数据。

[0029] 另一方面,提供了一种电子设备,所述电子设备包括处理器和存储器,所述存储器用于存储至少一段程序代码,所述至少一段程序代码由所述处理器加载并执行以实现本申请实施例中的特征索引编码方法中所执行的操作。

[0030] 另一方面,提供了一种存储介质,所述存储介质中存储有至少一段程序代码,所述至少一段程序代码用于执行本申请实施例中的特征索引编码方法。

[0031] 本申请实施例提供的技术方案带来的有益效果是:

[0032] 在本申请实施例中,通过哈希计算对特征的主键和辅键进行处理,确定对应的索引编码值,从而不需要对所有的待处理数据进行特征统计,一次计算即可为所有的待处理数据中特征的键值创建索引,降低了算法运行的时间复杂度,提高了数据的处理效率。

附图说明

[0033] 为了更清楚地说明本申请实施例中的技术方案,下面将对实施例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申请的一些实施例,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。

[0034] 图1是一种明文类型的样本数据的示意图;

[0035] 图2是一种向量索引的样本数据的示意图;

[0036] 图3是根据本申请实施例提供的一种编码系统的结构框图;

[0037] 图4是根据本申请实施例提供的一种特征索引编码方法的流程图；

[0038] 图5是根据本申请实施例提供的一种索引编码拼接生成示意图；

[0039] 图6是根据本申请实施例提供的一种哈希算法将字符串转换为编码值的流程；

[0040] 图7是根据本申请实施例提供的一种特征索引编码装置的框图；

[0041] 图8是根据本申请实施例提供的一种终端的结构示意图；

[0042] 图9是根据本申请实施例提供的一种服务器的结构示意图。

具体实施方式

[0043] 为使本申请的目的、技术方案和优点更加清楚，下面将结合附图对本申请实施方式作进一步地详细描述。

[0044] 这里将详细地对示例性实施例进行说明，其示例表示在附图中。下面的描述涉及附图时，除非另有表示，不同附图中的相同数字表示相同或相似的要素。以下示例性实施例中所描述的实施方式并不代表与本申请相一致的所有实施方式。相反，它们仅是与如所附权利要求书中所详述的、本申请的一些方面相一致的装置和方法的例子。

[0045] 本申请实施例提供了一种特征索引编码方法，可以用于机器学习过程中对样本数据进行处理的场景。在使用机器学习进行建模时，需要对模型进行训练，而模型训练离不开样本数据。通常情况下，收集到的样本数据往往是如图1所示的明文数据。参见图1，图1示例性示出了两条样本数据101和102，以样本数据101为例，样本1->因变量:1.0表示样本1的因变量的值为1.0，而相应的自变量包括“点击行为->餐饮类次数:23.0，性别->男性:1.0，年龄->20-30岁之间:1.0，物品->类别为餐饮:1.0，物品1000085->点击率:0.02，…”。样本数据102与样本数据101相类似，不再赘述。这些收集到的数据为原始样本数据，通常来自于用户行为、用户画像、物品画像以及各类基于人类先验知识的统计类数据，最终以图1所示的方式进行呈现。由于电子设备不能对明文数据进行直接计算，因此在获取到原始样本数据后，需要对原始样本数据进行处理，将原始样本数据转化为电子设备可以计算的形式，如向量等。例如，参见图2所示，图2示例性示出了两条向量索引的样本数据201和202，以样本数据201为例，样本数据201对应于图1中的样本数据101。其中，1.0对应于样本1的因变量，1:23.0对应于点击行为->餐饮类次数:23.0，3:1.0对应于性别->男性:1.0，5:1.0对应于年龄->20-30岁之间:1.0，8:10.对应于物品->类别为餐饮:1.0，12:0.02对应于物品1000085->点击率:0.02。也即是点击行为->餐饮类次数，性别->男性，年龄->20-30岁之间，物品->类别为餐饮，物品1000085->点击率分别用索引值1，3，5，8，12来表示。这样电子设备即可对处理后的样本数据进行计算。本申请实施例提供的特征索引编码方法用于实现将明文数据转化为索引值。

[0046] 下面简单介绍一下现有技术存在的缺陷。目前现有技术在将明文数据转化为索引值时，通常是采用统计类方法，对所有原始样本数据中的特征的键值进行统计，即统计上述性别->男性、点击行为->餐饮类次数等，统计完毕后为每个键值分配唯一的索引值。这种方式适合于原始样本数据的数量级较小时的场景，如几千条样本数据，需要统计几千个特征的键值，分配几千个索引值。而当原始样本数据的数量级很大时，如几千万条样本数据，甚至几亿条样本数据，需要统计的特征的键值数也是几千万个甚至上亿个，很显然统计的成本非常的高，甚至明文数据转化为索引值所花费的时间比模型训练花费的时间还长。另外，

对于在线学习的场景，由于在线学习通常是采用增量学习的方式来进行训练，以捕捉用户的兴趣变化，使模型更加契合当前的数据分布，因此不能预先对数据进行统计，除非是预先划定特征的范围，将范围外的特征舍弃。例如，在电商大促时，商品的更新频率非常高，对于新的商品，会实时产生大量的相关特征，显然无法通过统计类方法将明文数据转化为索引值。

[0047] 图3是根据本申请实施例提供的一种编码系统300的结构框图。该编码系统300包括：终端310和编码平台320。

[0048] 终端310通过无线网络或有线网络与编码平台310相连。终端310可以是智能手机、游戏主机、台式计算机、平板电脑、电子书阅读器、MP3播放器、MP4播放器和膝上型便携计算机中的至少一种。终端310安装和运行有用于数据采集的应用程序。该应用程序可以是购物类应用程序、社交通讯类应用程序或者资讯类应用程序等。示意性的，终端310是用户使用的终端，终端310中运行的应用程序内登录有用户账号。其中，采集的数据均为用户已授权的信息。

[0049] 编码平台320包括一台服务器、多台服务器、云计算平台和虚拟化中心中的至少一种。编码平台320用于从至少一个终端中获取样本数据以及对样本数据进行处理。可选地，编码平台320承担主要编码工作，终端310承担次要编码工作；或者，编码平台320承担次要编码工作，终端310承担主要编码工作；或者，编码平台320或终端310分别可以单独承担编码工作。

[0050] 可选地，编码平台320包括：接入服务器、编码服务器和数据库。接入服务器用于为终端310提供接入服务。编码服务器用于提供样本数据的处理服务。编码服务器可以是一台或多台。当编码服务器是多台时，存在至少两台编码服务器用于提供不同的服务，和/或，存在至少两台编码服务器用于提供相同的服务，比如以负载均衡方式提供同一种服务，本申请实施例对此不加以限定。

[0051] 终端310可以泛指多个终端中的一个，本实施例仅以终端310来举例说明。

[0052] 本领域技术人员可以知晓，上述终端310的数量可以更多或更少。比如上述终端可以仅为一个，或者上述终端为几十个或几百个，或者更多数量，此时上述编码系统还包括其他终端。本申请实施例对终端的数量和设备类型不加以限定。

[0053] 图4是本申请实施例提供的一种特征索引编码方法的流程图，如图4所示，在本申请实施例中以电子设备为例进行说明。该特征索引编码方法包括以下步骤：

[0054] 401、电子设备从原始样本数据中获取待处理数据。

[0055] 在本申请实施例中，电子设备可以从至少一个终端中获取样本数据，这些获取到的样本数据为未经处理的数据，可以称为原始样本数据，例如参见图1所示。电子设备可以将任一条原始样本数据作为待处理数据，通过本申请实施例提供的特征索引方法对该待处理数据进行处理。

[0056] 在一种可选的实现方式中，原始样本数据包括用户画像特征、用户行为特征、物品画像特征中的至少一种，电子设备在得到原始样本数据后，可以对原始样本数据进行特征分段，即对原始样本数据按照特征类别进行拆分，得到至少一种待处理数据。例如包括至少一个用户画像特征的待处理数据，或者包括至少一个物品画像特征的待处理数据等，本申请实施例对此不进行限制。

[0057] 402、电子设备从待处理数据中获取至少一个特征，该特征包括主键、辅键以及特征值。

[0058] 在本申请实施例中，该待处理数据可以为字符串形式的数据，该字符串形式的数据具有固定的数据格式。对于任一条待处理数据，该待处理数据包括至少一个特征，各特征之间可以通过固定的字符进行间隔。可选的，电子设备可以对待处理数据中包括的字符串进行分割，得到多个字符串。对于任一字符串，电子设备可以通过是否包括目标字符来判断该字符串是否为特征，若该字符串中包括至少一个目标字符，则可以将该字符串作为特征；若该字符串不包括任一目标字符，则该字符串不为特征。即电子设备可以将包括至少一个目标字符的字符串作为一个特征。其中，目标字符用于分隔特征中包括的主键、辅键和特征值。

[0059] 例如，以图1中示出的样本数据101为待处理数据进行说明，样本数据101为一个字符串，该字符串包括自变量部分和因变量部分，其因变量部分和自变量部分通过分号间隔，而自变量部分包括的特征都在大括号中，因此，电子设备可以通过大括号分割出自变量部分包括的特征，通过分号分割出因变量部分。对于自变量部分，由图1可知表示自变量部分的字符串中，各特征之间通过分号进行间隔。电子设备可以基于该分号对自变量部分包括的特征进行分割，得到多个字符串。如果字符串中包括目标字符“->”和“:”中的至少一种，如“物品->类别为餐饮:1.0”、“性别->男性:1.0”，则可以将该字符串作为特征。由于目标字符“->”用于分隔主键和辅键，目标字符“:”用于分隔辅键和特征值，则特征“物品->类别为餐饮:1.0”中的主键为物品，辅键为类别为餐饮，特征值为1.0；特征“性别->男性:1.0”中的主键为性别，辅键为男性，特征值为1.0。

[0060] 403、电子设备对于任一特征，根据哈希算法确定主键和辅键对应的索引编码值。

[0061] 在本申请实施例中，特征可以分为离散型特征和连续型特征，离散型特征可以表示为【维度，等级，特征值】的形式，如【性别，男性，1.0】，【类别，零售，1.0】等；而连续型特征可以表示为【维度，特征值】的形式，如【点击率，0.334】，【曝光次数，1234】等。在本申请实施例中，将上述两种结构形式通过【主键，辅键，特征值】的形式来表示，也即主键对应维度、辅键对应等级，特征值对应特征值。对于连续型特征没有等级的情况，可以通过占位符来表示辅键。在确定特征的表示方式后，电子设备可以通过哈希算法对主键和辅键进行计算，得到对应的索引编码值。

[0062] 在一种可选的实现方式中，对于离散型特征，电子设备可以根据同一哈希算法，采用相同的参数分别确定主键对应的第一编码值和辅键对应的第二编码值，然后将上述第一编码值和上述第二编码值进行拼接，得到上述索引编码值。由于采用的是同一哈希算法，且参数相同，因此不需要对哈希算法进行过多的调整，得到的第一编码值和第二编码值的取值范围相同。

[0063] 在一种可选的实现方式中，对于离散型特征，电子设备还可以根据同一哈希算法采用不同的参数分别确定主键对应的第一编码值和辅键对应的第二编码值，然后将上述第一编码值和上述第二编码值进行拼接，得到上述索引编码值。上述参数可以为主键的最大个数、辅键的最大个数、主键映射区间的扩大倍数以及辅键映射区间的扩展倍数等参数。其中，映射区间影响索引编码的取值范围，映射区间越大，越不容易冲突。由于采用的是同一哈希算法，而参数不同，如由于辅键的数量相对于主键的数量较少，因此可以为辅键设置较

小的辅键映射区间，使得辅键的取值范围小，从而辅键经过哈希计算后得到第二编码值更紧凑。

[0064] 在一种可选的实现方式中，对于离散型特征，电子设备还可以采用不同的哈希算法确定主键和辅键对应的索引编码值。相应的，本步骤可以为：电子设备根据第一哈希算法确定主键对应的第一编码值，根据第二哈希算法确定辅键对应的第二编码值，将第一编码值和第二编码值进行拼接，得到索引编码值。

[0065] 在一种可能的实现方式中，对于连续性特征，电子设备可以根据哈希算法确定主键对应的第一编码值，将目标占位符作为辅键对应的第二编码值，将第一编码值和第二编码值进行拼接，得到索引编码值。其中，目标占位符可以根据需求进行定义，本申请实施例对目标占位符不进行限制。如000、111或者222等。

[0066] 在一种可能的实现方式中，由于辅键通常比主键的个数少，因此辅键的取值范围比主键相对较小。电子设备在对第一编码值和第二编码值进行拼接时，可以将第一编码值作为索引编码值的尾部，将第二编码值作为索引编码值的头部，使得索引编码值的取值范围也较小。当然，电子设备也可以在将第一编码值作为索引编码值的头部，将第二编码值作为索引编码值的尾部。本申请实施例对此不进行限制。

[0067] 例如，参见图5所示，图5是本申请实施例提供的一种索引编码拼接生成示意图。在图5中，对于待处理的特征501，该特征为“性别->男性:1.0”，该特征501可以分为键值502和特征值503两部分，其中键值501包括主键和辅键，特征值503包括特征值，主键为性别、辅键为男性、特征值为1.0，通过哈希算法分别对主键和辅键进行处理，得到性别对应的第一编码值为121，男性对应的第二编码值为234。电子设备将第二编码值234作为索引编码值的头部，将第一编码值121作为索引编码值的尾部，从而得到索引编码值234121，该索引编码值与“性别->男性”具有映射关系。则特征501经过处理后被转换为特征504。

[0068] 需要说明的是，现有的哈希算法有很多，如Checksum(总和校验码)(8,16,32,or 64bit)，CRC16(16bit)(Cyclic Redundancy Check 16，循环冗余校验16比特版)，CRC32(32bit)，MD5(128bit)(Message Digest Algorithm 5，消息摘要算法5)，SHA-1(160bit)(Secure Hash Algorithm 1，安全散列算法1)，SHA-256(256bit)(哈希值长度为256位的安全散列算法)，RipeMD-128(128bit)(原始完整性校验讯息摘要)，RipeMD-160(160bit)，MD4(128bit)(Message Digest Algorithm 4，消息摘要算法4)，Ed2k(128bit)(eDonkey2000 network，一种文件共享网络)，Adler32(用于计算数据流的校验和的类)，MurmurHash3(32,128bit)(murmur哈希3，一种非加密散列函数)。其中，哈希算法将字符串转换为编码值的流程可以参见图6所示，该流程包括以下步骤：601、输入字符串s，602、对s进行Bytes(比特)编码，将s转换为二进制编码的形式b，603、对h进行初始化，即将种子值赋予h，该种子值可以为素数等，604、判断b的所有比特位的二进制值是否处理完毕，605、如没有则对当前比特位值k进行转换，如位移或者与素数相乘等，606、将h与k合并，如通过指数、位移、相乘或相加素数的方式，然后继续处理b的下一个比特位，直到b的所有比特位均已处理完毕，607、得到s对应的哈希值h，608、将h对映射区间取余，得到该字符串对应的编码值。

[0069] 还需要说明的是，由于MurmurHash3对于规律性较强的特征的键值(特征的键值一般为各种英文单词及数字的组合)，随机分布特征表现更加良好，所以上述哈希算法可以为Murmur3算法。目前已实现MurmurHash算法的编程语言包括C++、Python、C、C#、Perl、Ruby、

PHP、Scala、Java、JavaScript,Spark(UC Berkeley AMP lab(加州大学伯克利分校的AMP实验室)所开源的类Hadoop MapReduce的通用并行框架)中也实现了32bit的Murmur3_x86_32,对于使用Spark处理大数据的应用程序可以直接调用该API(Application Programming Interface,应用程序接口)来完成确定索引编码值的功能。即电子设备将主键和辅键作为该API的输入参数,结合其他参数即可得到上述索引编码值。

[0070] 示例代码如下:

[0071] import org.apach.spark.mllib.feature.HashingTF

[0072] import scala.math

[0073] /**

[0074] *特征的键值映射,策略为将最终映射区间分为两部分,一部分为主键映射区,一部分为辅键映射区,这样设计的目的是尽可能使用主键、辅键不同则映射索引肯定不同的先验知识来表征低冲突率

[0075] *

[0076] *@param pk特征主键

[0077] *@param sk特征辅键

[0078] *@param pkMax模型所用特征集合中,包含的主键个数

[0079] *@param skMax模型所用特征集合中,包含的主键所属辅键最大个数

[0080] *@param pkScaling使用哈希算法映射时,主键映射区域间扩展倍数,映射区间越大,越不容易冲突

[0081] *@param skScaling使用哈希算法映射时,辅键映射区域间扩展倍数,映射区间越大,越不容易冲突

[0082] *@return索引编码值

[0083] */

[0084] Def hashIndex(pk:String,sk:String,pkMax:Int,skMax:Int,pkScaling:Int,skScaling:Int):Int＝{

[0085] val pkTF＝new HashingTF(pkMax*pkScaling)

[0086] val skTF＝new HashingTF(skMax*skScaling)

[0087] val pkLength＝math.log10(pkMax*pkScaling.toDouble).ceiltoInt//pk最长用多少位表示

[0088] val pkIx＝s“%0${pkLength}d”.format(pkTF.indexof(pk))

[0089] val skIx＝skTF.indexof(sk)

[0090] s“$skIx$pkIx”.toInt

[0091] }

[0092] 需要说明的是,由于Spark原生Murmur3_x86_32对索引有最大值限制INT.MAX_VALUE(2^32-1),经过验证,千万维特征规模使用原生Murmur3_x86_32是可靠的。更高维度特征的需求,需要复写该算法,扩大特征映射区间。

[0093] 404、电子设备将特征中的主键和辅键替换为索引编码值。

[0094] 在本申请实施例中,电子设备在确定主键和辅键对应的索引编码值后,可以将特征中的主键和辅键替换为对应的索引编码值,从而电子设备可以对该特征进行计算。

[0095] 例如，对于特征“性别->男性:1.0”，电子设备将该特征中的性别->男性替换为234121，替换后的特征为“234121:1.0”。

[0096] 本申请实施例提供的特征索引编码方法，相对于优化前的方案，即统计+分配索引的方案，具有较多的优点，当然也存在一定的缺点，如存在误差等。相应的，统计+分配索引的方案(优化前方案)与本申请提供的方案(优化后方案)的优缺点对比可以参见表1所示。

[0097] 表1

[0098]

方案	统计+分配索引的方案（优化前方案）	本申请提供的方案（优化后方案）
优点	准确、无误差	● 效率高（计算转换速度快） ● 避免数据处理阶段成为系统瓶颈（理论上能够处理亿维度的特征） ● 维护简单（不需要键值和唯一标识映射数据，特别是在线服务是，减少了数据出库，状态管理等工作） ● 例行化算法效果上升，方便定位问题
缺点	● 效率低（统计+网络输入输出+分配索引） ● 系统瓶颈（几万维特征不明显，百万维成为瓶颈） ● 维护麻烦（训练、离线预测、在线预测，且需要维护一份键值和唯一标识映射数据） ● 例行化算法效果下降，定位问题较难	可能存在一定误差，但是从数学/实验上给出了可靠结论

[0099] 另外，为了验证本申请实施例提供的特征索引编码方法的可靠性，还通过不同模型进行了性能实验和精度实验。参见表2所示，性能实验采用XGBoost算法(eXtreme Gradient Boosting，是GradientBoosting Machine的一个c++实现)进行，对比统计类特征索引编码方法(统计+分配索引的方案)和哈希特征索引编码方法(本申请提供的方案)的特征索引编码的生成时间，样本采用的是亿维样本、十万维特征。哈希特征索引编码方法的生成时间为4.7分钟，相对于统计类特征索引编码方法，效率提高了6倍左右，且随着特征规模的扩大，该值会越来越大。精度实验采用XGBoost算法、FM_LBFGS算法(一种在牛顿法基础上提出的求解函数根的算法)以及LR_LiBLinear算法(是针对线性场景而专门实现和优化的工具包，同时支持线性svm和线性Logistic Regression模型)，对比统计类特征索引编码方法和哈希特征索引编码方法的训练集AUC(Area Under Curve，受试者工作特征曲线下与坐标轴围成的面积)和测试集AUC，样本采用亿维样本、十万维特征和百万维样本、万维特征。

对比结果参见表2所示。

[0100] 表2

[0101]

实验类型	算法类型	统计类特征索引编码方法	哈希特征索引编码方法	备注
性能实验	XGBoost	30.3 分钟	4.7 分钟	亿维样本、十万维特征
精度实验	XGBoost	训练集 AUC: 0.7443↑ 测试集 AUC: 0.7432	训练集 AUC: 0.7442 测试集 AUC: 0.7436↑	亿维样本、十万维特征
	FM_LBFGS	训练集 AUC: 0.7665 测试集 AUC: 0.7651	训练集 AUC: 0.7682↑ 测试集 AUC: 0.7658	百万维样本、万维特征
	LR_LiBLinear	训练集 AUC: 0.6238 测试集 AUC: 0.6247	训练集 AUC: 0.6263↑ 测试集 AUC: 0.6270↑	百万维样本、万维特征

[0102] 在本申请实施例中，通过哈希计算对特征的主键和辅键进行处理，确定对应的索引编码值，从而不需要对所有的待处理数据进行特征统计，一次计算即可为所有的待处理数据中特征的键值创建索引，降低了算法运行的时间复杂度，提高了数据的处理效率。

[0103] 图7是据一示例性实施例提供的一种特征索引编码装置的框图。该装置用于执行上述特征索引编码方法执行时的步骤，参见图7，装置包括：获取模块701、确定模块702以及替换模块703。

[0104] 获取模块，用于从待处理数据中获取至少一个特征，特征包括主键、辅键以及特征值；

[0105] 确定模块，用于对于任一特征，根据哈希算法确定主键和辅键对应的索引编码值；

[0106] 替换模块，用于将特征中的主键和辅键替换为索引编码值。

[0107] 在一种可选的实现方式中，特征为离散型特征；

[0108] 确定模块，还用于根据同一哈希算法，采用相同的参数分别确定主键对应的第一编码值和辅键对应的第二编码值；将第一编码值和第二编码值进行拼接，得到索引编码值。

[0109] 在一种可选的实现方式中，特征为离散型特征；

[0110] 确定模块，还用于根据同一哈希算法，采用不同的参数分别确定主键对应的第一编码值和辅键对应的第二编码值；将第一编码值和第二编码值进行拼接，得到索引编码值。

[0111] 在一种可选的实现方式中，特征为离散型特征；

[0112] 确定模块，还用于根据第一哈希算法确定主键对应的第一编码值；根据第二哈希算法确定辅键对应的第二编码值，第一哈希算法和第二哈希算法为不同的哈希算法；将第一编码值和第二编码值进行拼接，得到索引编码值。

[0113] 在一种可选的实现方式中，特征为连续型特征；

[0114] 确定模块,还用于根据哈希算法确定主键对应的第一编码值;

[0115] 将目标占位符作为辅键对应的第二编码值;将第一编码值和第二编码值进行拼接,得到索引编码值。

[0116] 在一种可选的实现方式中,哈希算法为MurmurHash3算法。

[0117] 在一种可选的实现方式中,第一编码值位于索引编码值的尾部,第二编码值位于索引编码值的头部。

[0118] 在一种可选的实现方式中,获取模块,还用于对待处理数据中包括的字符串进行分割,得到多个字符串;将包括至少一个目标字符的字符串作为特征,目标字符用于分隔主键、辅键和特征值。

[0119] 在一种可选的实现方式中,装置还包括:

[0120] 获取模块,还用于获取原始样本数据,原始样本数据包括用户画像特征、用户行为特征、物品画像特征中的至少一种;

[0121] 拆分模块,用于对原始样本数据按照特征类别进行拆分,得到至少一种待处理数据。

[0122] 在本申请实施例中,通过哈希计算对特征的主键和辅键进行处理,确定对应的索引编码值,从而不需要对所有的待处理数据进行特征统计,一次计算即可为所有的待处理数据中特征的键值创建索引,降低了算法运行的时间复杂度,提高了数据的处理效率。

[0123] 需要说明的是:上述实施例提供的特征索引编码装置在运行应用程序时,仅以上述各功能模块的划分进行举例说明,实际应用中,可以根据需要而将上述功能分配由不同的功能模块完成,即将装置的内部结构划分成不同的功能模块,以完成以上描述的全部或者部分功能。另外,上述实施例提供的特征索引编码装置与特征索引编码方法实施例属于同一构思,其具体实现过程详见方法实施例,这里不再赘述。

[0124] 在本申请实施例中,电子设备可以提供为终端或者服务器,当提供为终端时,可以由该终端实现上述的特征索引编码方法所执行的操作,当提供为服务器时,可以通过该服务器和终端的交互来实现上述的特征索引编码方法所执行的操作,也可以由服务器单独实现上述的特征索引编码方法所执行的操作。

[0125] 图8示出了本申请一个示例性实施例提供的终端800的结构框图。该终端图8示出了本发明一个示例性实施例提供的终端800的结构框图。该终端800可以是:智能手机、平板电脑、MP3播放器(Moving Picture Experts Group Audio Layer III,动态影像专家压缩标准音频层面3)、MP4(Moving Picture Experts Group Audio Layer IV,动态影像专家压缩标准音频层面4)播放器、笔记本电脑或台式电脑。终端800还可能被称为用户设备、便携式终端、膝上型终端、台式终端等其他名称。

[0126] 通常,终端800包括有:处理器801和存储器802。

[0127] 处理器801可以包括一个或多个处理核心,比如4核心处理器、8核心处理器等。处理器801可以采用DSP(Digital Signal Processing,数字信号处理)、FPGA(Field-Programmable Gate Array,现场可编程门阵列)、PLA(Programmable Logic Array,可编程逻辑阵列)中的至少一种硬件形式来实现。处理器801也可以包括主处理器和协处理器,主处理器是用于对在唤醒状态下的数据进行处理的处理器,也称CPU(Central Processing Unit,中央处理器);协处理器是用于对在待机状态下的数据进行处理的低功耗处理器。在

一些实施例中，处理器801可以在集成有GPU(Graphics Processing Unit，图像处理器)，GPU用于负责显示屏所需要显示的内容的渲染和绘制。一些实施例中，处理器801还可以包括AI(Artificial Intelligence，人工智能)处理器，该AI处理器用于处理有关机器学习的计算操作。

[0128] 存储器802可以包括一个或多个计算机可读存储介质，该计算机可读存储介质可以是非暂态的。存储器802还可包括高速随机存取存储器，以及非易失性存储器，比如一个或多个磁盘存储设备、闪存存储设备。在一些实施例中，存储器802中的非暂态的计算机可读存储介质用于存储至少一个指令，该至少一个指令用于被处理器801所执行以实现本申请中方法实施例提供的特征索引编码方法。

[0129] 在一些实施例中，终端800还可选包括有：外围设备接口803和至少一个外围设备。处理器801、存储器802和外围设备接口803之间可以通过总线或信号线相连。各个外围设备可以通过总线、信号线或电路板与外围设备接口803相连。具体地，外围设备包括：射频电路804、显示屏805、摄像头组件806、音频电路807和电源809中的至少一种。

[0130] 外围设备接口803可被用于将I/O(Input/Output，输入/输出)相关的至少一个外围设备连接到处理器801和存储器802。在一些实施例中，处理器801、存储器802和外围设备接口803被集成在同一芯片或电路板上；在一些其他实施例中，处理器801、存储器802和外围设备接口803中的任意一个或两个可以在单独的芯片或电路板上实现，本实施例对此不加以限定。

[0131] 射频电路804用于接收和发射RF(Radio Frequency，射频)信号，也称电磁信号。射频电路804通过电磁信号与通信网络以及其他通信设备进行通信。射频电路804将电信号转换为电磁信号进行发送，或者，将接收到的电磁信号转换为电信号。可选地，射频电路804包括：天线系统、RF收发器、一个或多个放大器、调谐器、振荡器、数字信号处理器、编解码芯片组、用户身份模块卡等等。射频电路804可以通过至少一种无线通信协议来与其它终端进行通信。该无线通信协议包括但不限于：城域网、各代移动通信网络(2G、3G、4G及5G)、无线局域网和/或WiFi(Wireless Fidelity，无线保真)网络。在一些实施例中，射频电路804还可以包括NFC(Near Field Communication，近距离无线通信)有关的电路，本申请对此不加以限定。

[0132] 显示屏805用于显示UI(User Interface，用户界面)。该UI可以包括图形、文本、图标、视频及其它们的任意组合。当显示屏805是触摸显示屏时，显示屏805还具有采集在显示屏805的表面或表面上方的触摸信号的能力。该触摸信号可以作为控制信号输入至处理器801进行处理。此时，显示屏805还可以用于提供虚拟按钮和/或虚拟键盘，也称软按钮和/或软键盘。在一些实施例中，显示屏805可以为一个，设置终端800的前面板；在另一些实施例中，显示屏805可以为至少两个，分别设置在终端800的不同表面或呈折叠设计；在再一些实施例中，显示屏805可以是柔性显示屏，设置在终端800的弯曲表面上或折叠面上。甚至，显示屏805还可以设置成非矩形的不规则图形，也即异形屏。显示屏805可以采用LCD(Liquid Crystal Display，液晶显示屏)、OLED(Organic Light-Emitting Diode，有机发光二极管)等材质制备。

[0133] 摄像头组件806用于采集图像或视频。可选地，摄像头组件806包括前置摄像头和后置摄像头。通常，前置摄像头设置在终端的前面板，后置摄像头设置在终端的背面。在一

些实施例中，后置摄像头为至少两个，分别为主摄像头、景深摄像头、广角摄像头、长焦摄像头中的任意一种，以实现主摄像头和景深摄像头融合实现背景虚化功能、主摄像头和广角摄像头融合实现全景拍摄以及VR(Virtual Reality，虚拟现实)拍摄功能或者其它融合拍摄功能。在一些实施例中，摄像头组件806还可以包括闪光灯。闪光灯可以是单色温闪光灯，也可以是双色温闪光灯。双色温闪光灯是指暖光闪光灯和冷光闪光灯的组合，可以用于不同色温下的光线补偿。

[0134] 音频电路807可以包括麦克风和扬声器。麦克风用于采集用户及环境的声波，并将声波转换为电信号输入至处理器801进行处理，或者输入至射频电路804以实现语音通信。出于立体声采集或降噪的目的，麦克风可以为多个，分别设置在终端800的不同部位。麦克风还可以是阵列麦克风或全向采集型麦克风。扬声器则用于将来自处理器801或射频电路804的电信号转换为声波。扬声器可以是传统的薄膜扬声器，也可以是压电陶瓷扬声器。当扬声器是压电陶瓷扬声器时，不仅可以将电信号转换为人类可听见的声波，也可以将电信号转换为人类听不见的声波以进行测距等用途。在一些实施例中，音频电路807还可以包括耳机插孔。

[0135] 电源809用于为终端800中的各个组件进行供电。电源809可以是交流电、直流电、一次性电池或可充电电池。当电源809包括可充电电池时，该可充电电池可以支持有线充电或无线充电。该可充电电池还可以用于支持快充技术。

[0136] 在一些实施例中，终端800还包括有一个或多个传感器810。该一个或多个传感器810包括但不限于：加速度传感器811、陀螺仪传感器812、压力传感器813、光学传感器815以及接近传感器816。

[0137] 加速度传感器811可以检测以终端800建立的坐标系的三个坐标轴上的加速度大小。比如，加速度传感器811可以用于检测重力加速度在三个坐标轴上的分量。处理器801可以根据加速度传感器811采集的重力加速度信号，控制显示屏805以横向视图或纵向视图进行用户界面的显示。加速度传感器811还可以用于游戏或者用户的运动数据的采集。

[0138] 陀螺仪传感器812可以检测终端800的机体方向及转动角度，陀螺仪传感器812可以与加速度传感器811协同采集用户对终端800的3D动作。处理器801根据陀螺仪传感器812采集的数据，可以实现如下功能：动作感应(比如根据用户的倾斜操作来改变UI)、拍摄时的图像稳定、游戏控制以及惯性导航。

[0139] 压力传感器813可以设置在终端800的侧边框和/或显示屏805的下层。当压力传感器813设置在终端800的侧边框时，可以检测用户对终端800的握持信号，由处理器801根据压力传感器813采集的握持信号进行左右手识别或快捷操作。当压力传感器813设置在显示屏805的下层时，由处理器801根据用户对显示屏805的压力操作，实现对UI界面上的可操作性控件进行控制。可操作性控件包括按钮控件、滚动条控件、图标控件、菜单控件中的至少一种。

[0140] 光学传感器815用于采集环境光强度。在一个实施例中，处理器801可以根据光学传感器815采集的环境光强度，控制显示屏805的显示亮度。具体地，当环境光强度较高时，调高显示屏805的显示亮度；当环境光强度较低时，调低显示屏805的显示亮度。在另一个实施例中，处理器801还可以根据光学传感器815采集的环境光强度，动态调整摄像头组件806的拍摄参数。

[0141] 接近传感器816，也称距离传感器，通常设置在终端800的前面板。接近传感器816用于采集用户与终端800的正面之间的距离。在一个实施例中，当接近传感器816检测到用户与终端800的正面之间的距离逐渐变小时，由处理器801控制显示屏805从亮屏状态切换为息屏状态；当接近传感器816检测到用户与终端800的正面之间的距离逐渐变大时，由处理器801控制显示屏805从息屏状态切换为亮屏状态。

[0142] 本领域技术人员可以理解，图8中示出的结构并不构成对终端800的限定，可以包括比图示更多或更少的组件，或者组合某些组件，或者采用不同的组件布置。

[0143] 图9是本申请实施例提供的一种服务器900的结构示意图。该服务器900可因配置或性能不同而产生比较大的差异，可以包括一个或一个以上处理器(Central Processing Units，CPU)901和一个或一个以上的存储器902，其中，所述存储器902中存储有至少一条指令，所述至少一条指令由所述处理器901加载并执行以实现上述各个方法实施例提供的方法。当然，该服务器还可以具有有线或无线网络接口、键盘以及输入输出接口等部件，以便进行输入输出，该服务器还可以包括其他用于实现设备功能的部件，在此不做赘述。

[0144] 本申请实施例还提供了一种计算机可读存储介质，该计算机可读存储介质应用于电子设备，该计算机可读存储介质中存储有至少一条程序代码，该至少一条程序代码用于被处理器执行并实现本申请实施例中的特征索引编码方法中电子设备所执行的操作。

[0145] 本领域普通技术人员可以理解实现上述实施例的全部或部分步骤可以通过硬件来完成，也可以通过程序来指令相关的硬件完成，所述的程序可以存储于一种计算机可读存储介质中，上述提到的存储介质可以是只读存储器，磁盘或光盘等。

[0146] 以上所述仅为本申请的可选实施例，并不用以限制本申请，凡在本申请的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本申请的保护范围之内。

101 样本1->因变量：1.0；自变量：{点击行为->餐饮类次数：23.0，性别->男性：1.0，年龄->20-30岁之间：1.0，物品->类别为餐饮：1.0，物品1000085->点击率：0.02，…}

102 样本2->因变量：0.0；自变量：{点击行为->零售类次数：12.0，性别->女性：1.0，年龄->10-20岁之间：1.0，物品->类别为零售：1.0，物品1000080->点击率：0.015，…}

…

图1

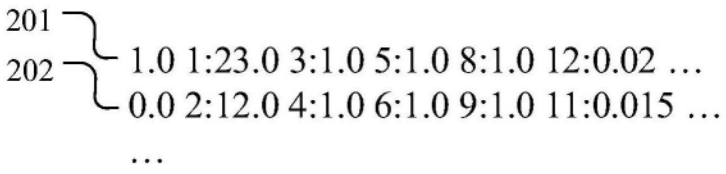


图2

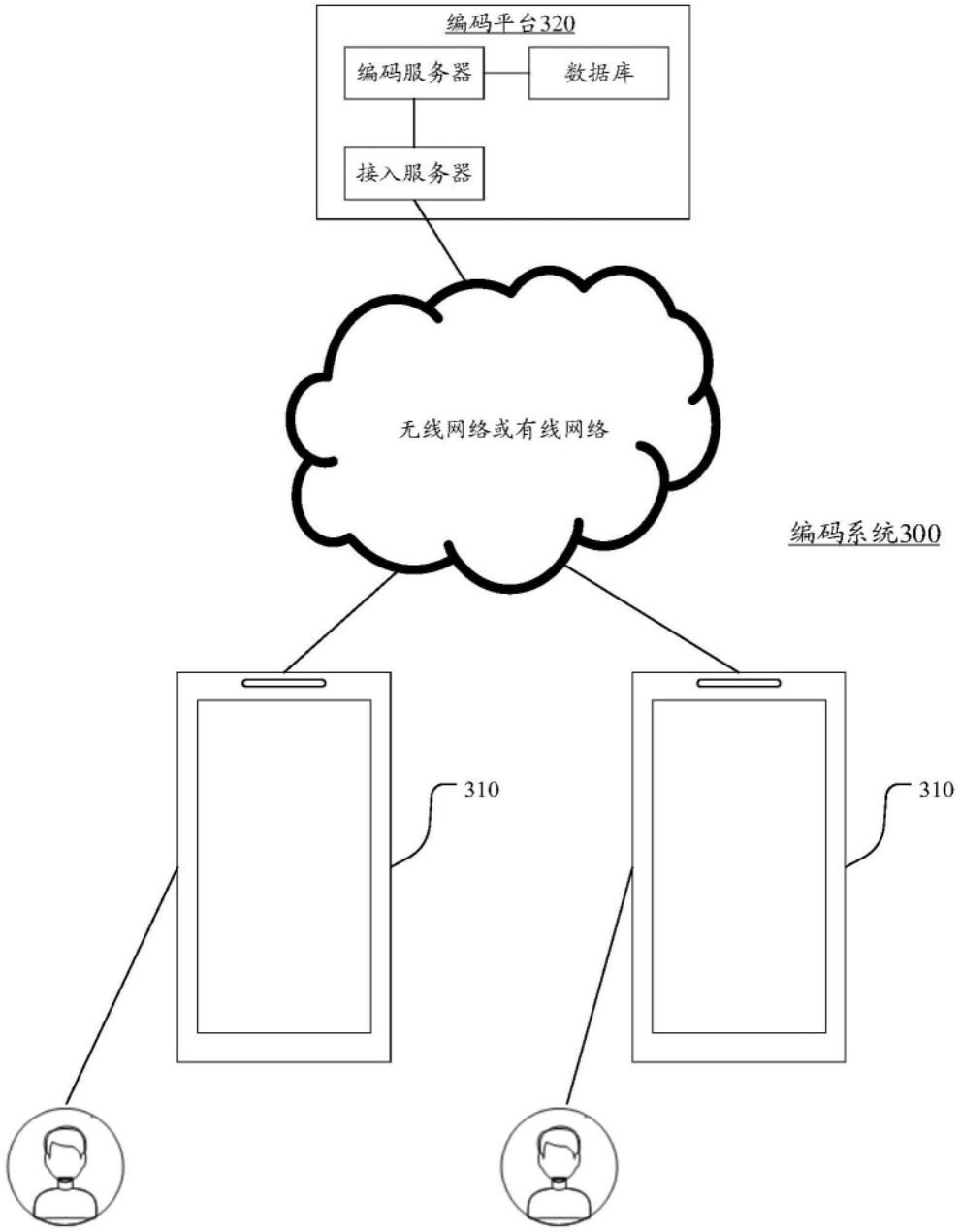


图3

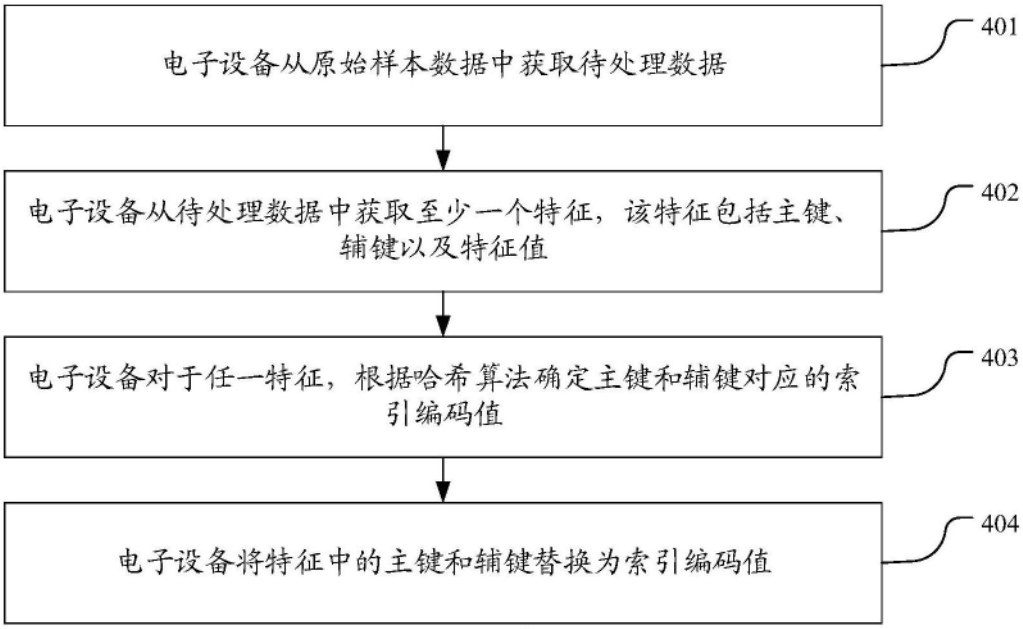

电子设备从原始样本数据中获取待处理数据
401
电子设备从待处理数据中获取至少一个特征，该特征包括主键、辅键以及特征值
402
电子设备对于任一特征，根据哈希算法确定主键和辅键对应的索引编码值
403
电子设备将特征中的主键和辅键替换为索引编码值
404

图4

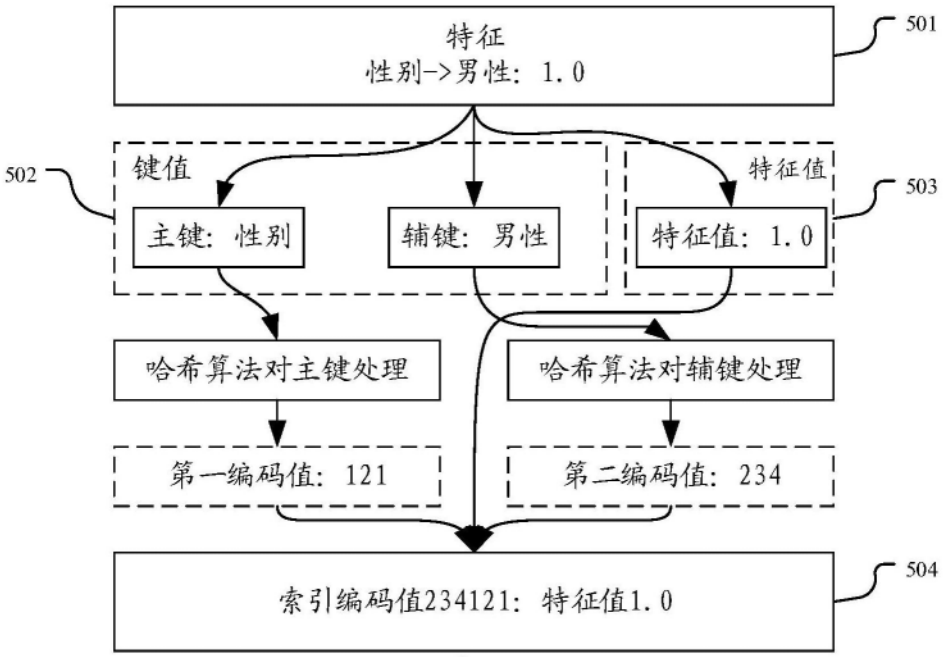

特征
性别->男性：1.0
501
502
键值
主键：性别
辅键：男性
特征值
特征值：1.0
503
哈希算法对主键处理
哈希算法对辅键处理
第一编码值：121
第二编码值：234
索引编码值234121：特征值1.0
504

图5

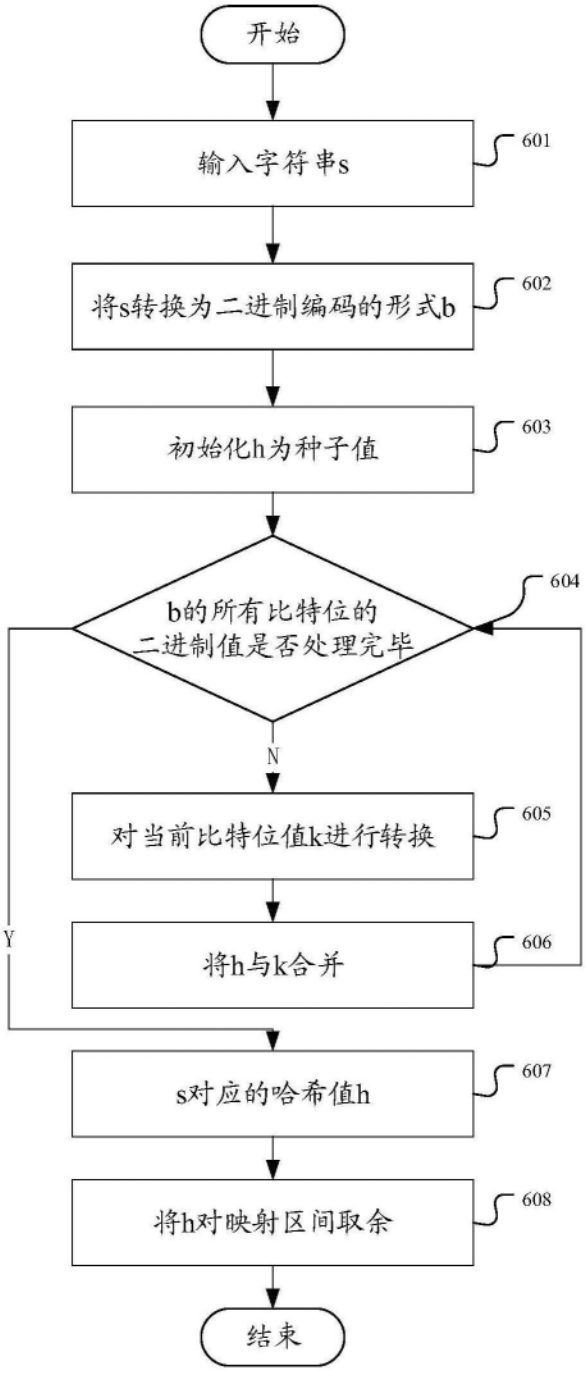

开始
输入字符串s
601
将s转换为二进制编码的形式b
602
初始化h为种子值
603
b的所有比特位的
二进制值是否处理完毕
604
N
对当前比特位值k进行转换
605
Y
将h与k合并
606
s对应的哈希值h
607
将h对映射区间取余
608
结束

图6

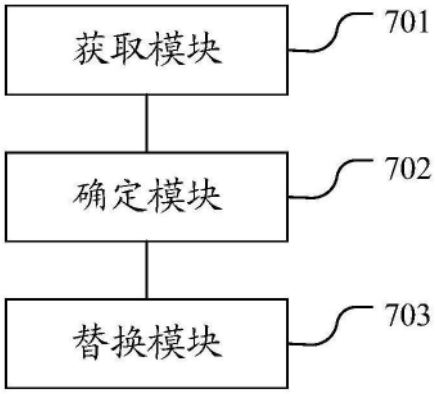


图7

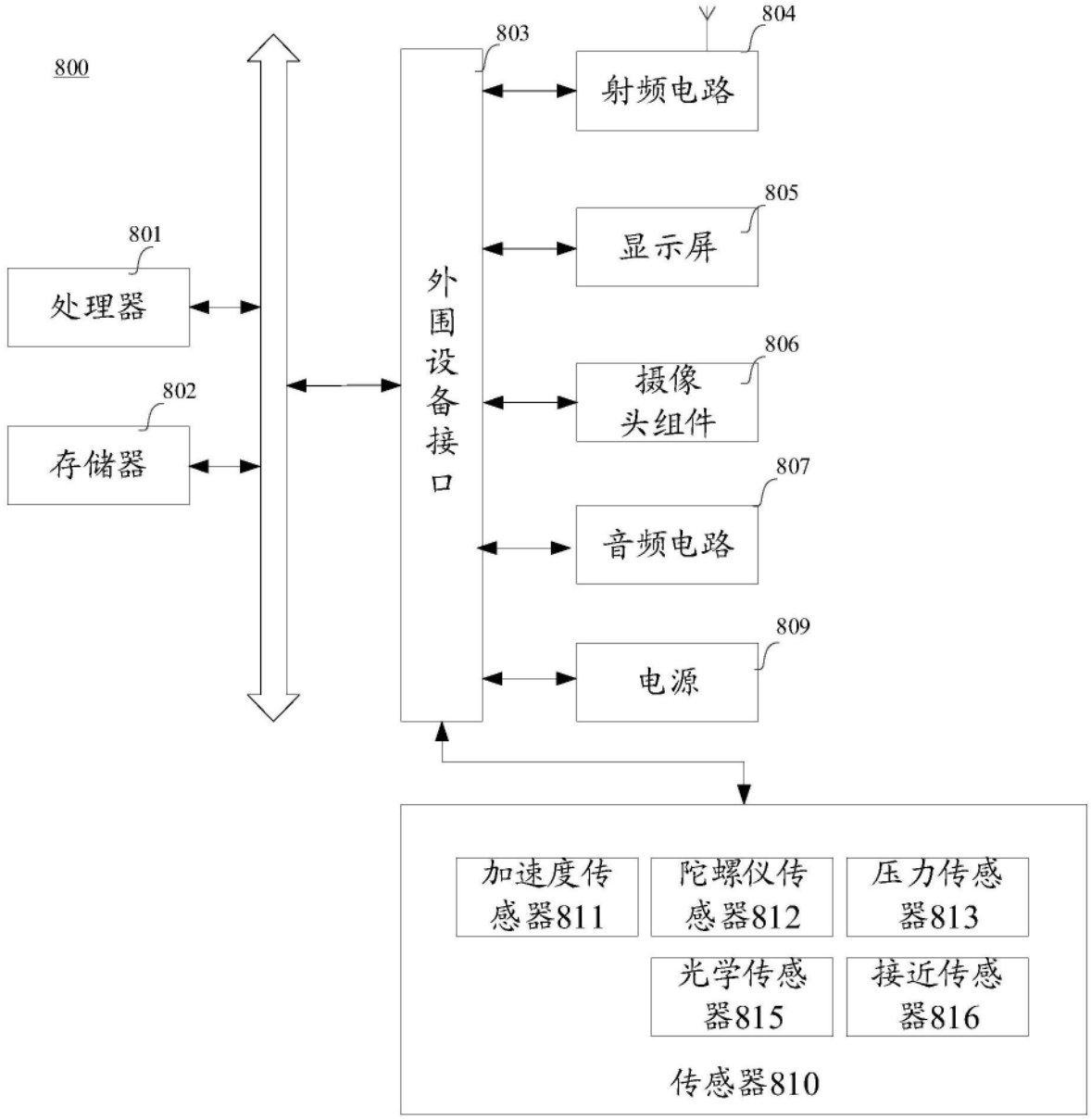

800
801
处理器
802
存储器
803
外围设备接口
804
射频电路
805
显示屏
806
摄像头组件
807
音频电路
809
电源
加速度传感器811
陀螺仪传感器812
压力传感器813
光学传感器815
接近传感器816
传感器810

图8

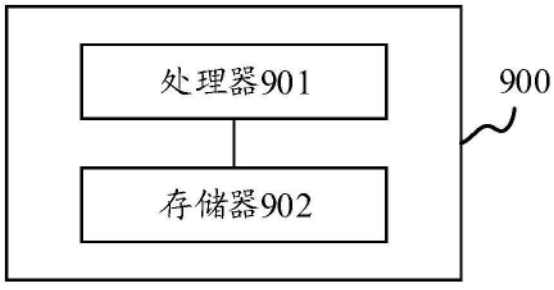

处理器901
存储器902
900

图9