

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2020 年 9 月 10 日 (10.09.2020)



(10) 国际公布号
WO 2020/177475 A1

(51) 国际专利分类号:
G06F 21/60 (2013.01)

(21) 国际申请号: PCT/CN2020/070303

(22) 国际申请日: 2020 年 1 月 3 日 (03.01.2020)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910161715.8 2019年3月4日 (04.03.2019) CN

(71) 申请人: 阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED) [—/CN]; 开曼群岛
大开曼资本大厦一座四层 847 号 邮箱,
Grand Cayman (KY)。

(72) 发明人: 林文珍 (LIN, Wenzhen); 中国浙江省杭州市
市余杭区文一西路 969 号 3 号楼 5 楼 阿里巴巴
集团法务部, Zhejiang 311121 (CN)。

(74) 代理人: 北京博思佳知识产权代理有限公司 (BEIJING BESTIPR INTELLECTUAL PROPERTY LAW CORPORATION); 中国北京市
海淀区上地三街 9 号 嘉华大厦 B 座 409,
Beijing 100085 (CN)。

(81) 指定国 (除另有指明, 要求每一种可提供的国家
保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG,
BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU,
CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB,
GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS,
JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK,
LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX,

(54) Title: SECURE FEATURE ENGINEERING METHOD AND APPARATUS

(54) 发明名称: 一种安全的特征工程方法和装置

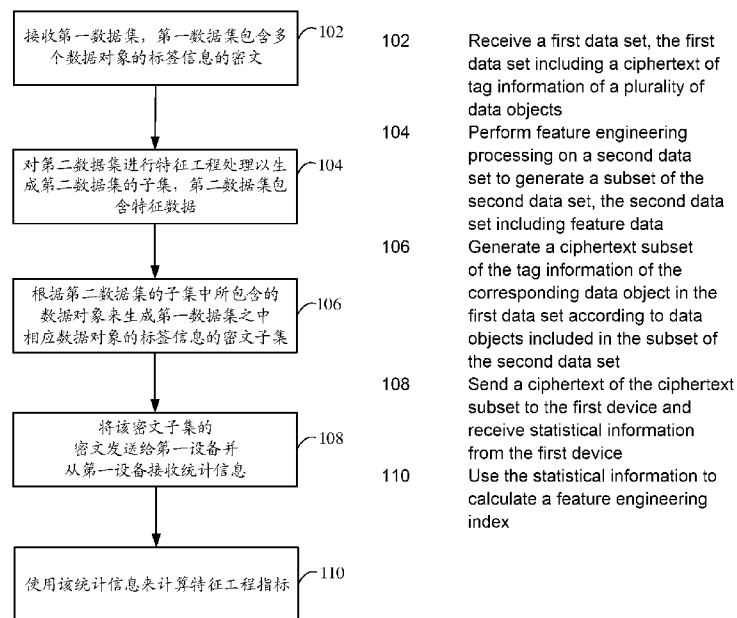


图 1

(57) Abstract: Provided are a secure feature engineering method and apparatus. A first device may transmit a first data set to a second device, the first data set including a ciphertext of tag information of a plurality of data objects. The second device may perform feature engineering processing on a second data set including feature data of the plurality of data objects so as to generate a subset of the second data set. The second device may generate a ciphertext subset of the tag information of the corresponding data object in the first data set according to the subset of the second data set, and send a ciphertext of the ciphertext subset to the first device. The first device

MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL,
PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,
ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US,
UZ, VC, VN, WS, ZA, ZM, ZW。

- (84) 指定国(除另有指明, 要求每一种可提供的地区
保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM,
AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG,
CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU,
IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT,
RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

- 包括国际检索报告(条约第21条(3))。

can decrypt the ciphertext subset and generate statistical information, and send the statistical information to the second device. The second device may use the statistical information to calculate a feature engineering index. Further provided is a corresponding feature engineering apparatus.

(57) 摘要: 本公开提供了一种安全的特征工程方法和装置。第一设备可向第二设备传送第一数据集, 第一数据集包含多个数据对象的标签信息的密文。第二设备可对包含该多个数据对象的特征数据的第二数据集进行特征工程处理以生成第二数据集的子集。第二设备可根据第二数据集的子集来生成第一数据集之中相应数据对象的标签信息的密文子集, 将密文子集的密文发送给第一设备。第一设备可解密该密文子集并生成统计信息, 将该统计信息发送给第二设备。第二设备可使用该统计信息来计算特征工程指标。本公开还提供了相应的特征工程装置。

一种安全的特征工程方法和装置

技术领域

[01]本公开涉及特征工程，尤其涉及一种安全的特征工程方法和装置。

背景技术

[02]建模过程需要对样本数据先进行基本的特征工程处理，然后用经过特征工程处理的数据集进行模型训练。特征工程是指从原始数据转换为特征向量的过程。特征工程是机器学习中最重要、最基础的起始步骤，会直接影响机器学习的效果，并通常需要大量的时间。典型的特征工程包括数据清理、特征离散化、特征提取、特征选择等过程。在大数据场景下，数据通常是分布式存储的，而特征工程处理需要把各方的建模样本数据融合在一起。例如，在进行数据分箱时，需要特征数据和标签信息配合才能完成。而在联合建模场景中，一般存在其中一方拥有特征数据，另一方拥有标签信息而没有或有少量不齐全的特征数据的情况，这就需要建模合作方的样本数据共同协作来完成特征工程处理，例如分箱及分箱指标统计。因此，在进行联合建模的场景中，数据合作双方需要把建模的样本数据融合在一起，才能进行特征工程处理。但这样做，数据合作双方的样本数据需要对外输出，存在隐私数据泄漏问题。

[03]现有技术中解决上述问题的一种方案是将数据集中到其中一方进行数据融合，并进一步做特征工程处理。这种方案的缺点是不安全，完全暴露了隐私数据明文，泄漏了数据隐私，而且还存在数据被恶意传播或售卖等风险。

[04]现有技术中解决上述问题的另一种方案是寻找一个安全可信的第三方，在安全环境中导入各方数据，在一个安全沙箱环境的公有云 / 第三方等环境中进行数据融合，集中进行特征工程处理。这种方案的缺点是数据明文仍然需要对外输出，不能完全杜绝隐私数据的泄漏问题，也不能解决本身的数据信任问题。此外，完全可信的第三方难求，数据 / 模型导入到可信第三方的传输过程中也有数据传输安全隐患。不仅如此，在可信第三方存在操作权限问题，对数据出入的安全检查粒度难把控，特征工程处理结果也可能被泄漏。

[05]因此，本领域需要一种安全的特征工程方法和装置。

发明内容

[06]本公开提供了安全的特征工程方法和装置，使得合作方均不必泄漏自己的隐私数据明文，但最终能完成特征工程处理。例如，第一设备可向第二设备传送第一数据集，第一数据集包含多个数据对象的标签信息的密文。第二设备可对包含该多个数据对象的特征数据的第二数据集进行特征工程处理以生成第二数据集的子集。第二设备可根据第二数据集的子集来生成第一数据集之中相应数据对象的标签信息的密文子集，将密文子集的密文发送给第一设备。第一设备可解密该密文子集并生成统计信息，将该统计信息发送给第二设备。第二设备可使用该统计信息来计算特征工程指标。

[07]根据本公开的一个实施例，提供了一种特征工程方法，其包括：接收第一数据集，第一数据集包含多个数据对象的标签信息的密文；对第二数据集进行特征工程处理以生成第二数据集的子集，第二数据集包含所述多个数据对象的特征数据；根据所述第二数据集的子集中所包含的数据对象来生成第一数据集之中相应数据对象的标签信息的密文子集；将所述密文子集的密文发送给第一设备；从第一设备接收在解密所述密文子集的密文后生成的统计信息；以及使用所述统计信息来计算对第二数据集进行的特征工程处理的特征工程指标。

[08]在一方面，该方法还包括：根据所述特征工程指标来生成特征工程结果；以及将所述特征工程结果发送给第一设备。

[09]在一方面，所述特征工程处理包括对第二数据集进行特征选择以生成第二数据集的特征数据子集，其中计算特征工程指标包括利用所述特征数据子集与所述统计信息来计算特征选择指标。

[10]在一方面，所述特征工程处理包括对第二数据集进行分箱以生成多个分箱段，并且所述密文子集包括与所述多个分箱段相对应的多个密文子集，其中每个密文子集包括第一数据集之中的落在相应分箱段中的数据对象的标签信息密文，其中计算特征工程指标包括利用所述第二数据集的分箱段与所述统计信息来计算分箱指标。

[11]在一方面，第一数据集的标签信息的密文是使用同态加密对所述标签信息进行加密来生成的，所述标签信息为二元值，并且将所述密文子集的密文发送给第一设备包括：对所述密文子集中的密文进行同态加法以得到密文总和并将所述密文总和发送给第一设备。

[12]在一方面，第一数据集和第二数据集之中的数据对象分别具有相关联的标识符，所

述方法还包括在生成所述第二数据集的子集之前对第一数据集和第二数据集进行标识符匹配,以使得第一数据集的标签信息与第二数据集的特征数据对应于相同的多个数据对象。

[13]在一方面,所述统计信息包括所述密文子集的正样本数、和/或负样本数。

[14]在一方面,对第二数据集进行特征工程处理是在接收第一数据集之前、之时、或之后发生的。

[15]在一方面,所述特征工程指标包括以下至少一者:所述第二数据集的子集的正样本率、所述第二数据集的子集的负样本率、所述第二数据集的子集与所述密文子集中的标签信息的相关度、所述第二数据集的子集的信息价值、所述第二数据集的子集的证据权重。

[16]根据本公开的另一个实施例,提供了一种特征工程方法,其包括:提供第一数据集,第一数据集包含多个数据对象的标签信息的密文;接收第一数据集的密文子集的密文,所述密文子集是根据对第二数据集进行的特征工程处理来生成的,第二数据集包含所述多个数据对象的特征数据,其中对第二数据集进行的特征工程处理生成第二数据集的子集,其中所述密文子集包含第一数据集之中的与所述第二数据集的子集相同的数据对象的标签信息密文;解密所述密文子集的密文并生成所述密文子集的统计信息;以及提供所述统计信息以用于计算对第二数据集的特征工程处理的特征工程指标。

[17]在一方面,所述方法还包括:接收根据所述特征工程指标生成的第一特征工程结果。

[18]在一方面,所述方法还包括:使用所述第一数据集的标签信息和第三数据集的特征数据来执行所述特征工程处理以生成第二特征工程结果;以及组合第一特征工程结果和第二特征工程结果以获得组合的特征工程结果。

[19]在一方面,所述特征工程处理包括对第二数据集进行特征选择以生成第二数据集的特征数据子集,其中计算特征工程指标包括利用所述特征数据子集与所述统计信息来计算特征选择指标。

[20]在一方面,所述特征工程处理包括对第二数据集进行分箱以生成多个分箱段,并且所述密文子集包括与所述多个分箱段相对应的多个密文子集,其中每个密文子集包括第一数据集之中的落在相应分箱段中的数据对象的标签信息密文,所述计算特征工程指标包括利用所述第二数据集的分箱段与所述统计信息来计算分箱指标。

[21]在一方面,第一数据集的标签信息的密文是使用同态加密对所述标签信息进行加密

来生成的，所述标签信息为二元值，所述密文子集的密文包括将所述密文子集中的密文进行同态加法得到的密文总和。

[22]在一方面，第一数据集和第二数据集之中的数据对象分别具有相关联的标识符，所述方法还包括在生成所述第二数据集的子集之前对第一数据集和第二数据集进行标识符匹配，以使得第一数据集的标签信息与第二数据集的特征数据对应于相同的多个数据对象。

[23]在一方面，所述统计信息包括所述密文子集的正样本数、和/或负样本数。

[24]在一方面，所述特征工程指标包括以下至少一者：所述第二数据集的子集的正样本率、所述第二数据集的子集的负样本率、所述第二数据集的子集与所述密文子集中的标签信息的相关度、所述第二数据集的子集的信息价值、所述第二数据集的子集的证据权重。

[25]根据本公开的另一个实施例，提供了一种特征工程装置，其包括：传输模块，其接收第一数据集，第一数据集包含多个数据对象的标签信息的密文；特征工程处理模块，其对第二数据集进行特征工程处理以生成第二数据集的子集，第二数据集包含所述多个数据对象的特征数据；以及密文子集生成模块，其根据所述第二数据集的子集中所包含的数据对象来生成第一数据集之中相应数据对象的标签信息的密文子集；其中所述传输模块将所述密文子集的密文发送给第一设备并从第一设备接收在解密所述密文子集的密文后生成的统计信息；并且所述特征工程处理模块使用所述统计信息来计算对第二数据集进行的特征工程处理的特征工程指标。

[26]在一方面，所述特征工程处理模块根据所述特征工程指标来生成特征工程结果；以及所述传输模块将所述特征工程结果发送给第一设备。

[27]在一方面，所述特征工程处理包括特征选择或分箱。

[28]在一方面，所述统计信息包括所述密文子集的正样本数、和/或负样本数。

[29]根据本公开的另一个实施例，提供了一种特征工程装置，其包括：传输模块，其向第二设备提供第一数据集，第一数据集包含多个数据对象的标签信息的密文，所述传输模块还从第二设备接收第一数据集的密文子集的密文，所述密文子集是根据对第二数据集进行的特征工程处理来生成的，第二数据集包含所述多个数据对象的特征数据，其中对第二数据集进行的特征工程处理生成第二数据集的子集，其中所述密文子集包含第一数据集之中的与所述第二数据集的子集相同的数据对象的标签信息密文；加解密模块，

其解密所述密文子集的密文；以及处理模块，其生成所述密文子集的统计信息，其中所述传输模块向第二设备提供所述统计信息以供第二设备用于计算对第二数据集的特征工程处理的特征工程指标。

[30]在一方面，其中所述传输模块从第二设备接收根据所述特征工程指标生成的第一特征工程结果，并且所述特征工程装置还包括：特征工程处理模块，其使用所述第一数据集的标签信息和第三数据集的特征数据来执行所述特征工程处理以生成第二特征工程结果，以及组合第一特征工程结果和第二特征工程结果以获得组合的特征工程结果。

[31]在一方面，所述特征工程处理包括特征选择或分箱。

[32]在一方面，所述统计信息包括所述密文子集的正样本数、和/或负样本数。

[33]根据本公开的另一个实施例，提供了一种特征工程系统，其包括：处理器；用于存储处理器可执行指令的存储器，其中所述处理器被配置成执行所述处理器可执行指令以实现如上所述的方法。

[34]根据本公开，合作方均不必泄漏自己的隐私数据明文，但最终能完成特征工程处理，从而提高了特征工程处理的安全性。

附图说明

[35]图 1 为根据本公开一个实施例的特征工程方法的流程图；

[36]图 2 为根据本公开一个实施例的特征工程方法的示意图；

[37]图 3 为根据本公开另一个实施例的特征工程方法的示意图；

[38]图 4 为根据本公开另一个实施例的特征工程方法的示意图；以及

[39]图 5 为根据本公开一个实施例的特征工程装置的框图。

具体实施方式

[40]下面结合具体实施例和附图对本公开作进一步说明，但不应以此限制本公开的保护范围。

[41]本公开提供了安全的特征工程方法和装置。在联合建模场景中，一般存在其中一方拥有特征数据，另一方拥有标签信息而没有或有少量不齐全的特征数据的情况，这就需要建模合作方的样本数据共同协作来完成特征工程处理。根据本公开的一个实施例，第

一设备可向第二设备传送第一数据集，第一数据集包含多个数据对象的标签信息的密文。第二设备可对包含该多个数据对象的特征数据的第二数据集进行特征工程处理以生成第二数据集的子集。第二设备可根据第二数据集的子集来生成第一数据集之中相应数据对象的标签信息的密文子集，将密文子集的密文发送给第一设备。第一设备可解密该密文子集并生成统计信息，将该统计信息发送给第二设备。第二设备可使用该统计信息来计算特征工程指标。

[42] 本公开的分布式特征工程和指标计算使得合作方均不必泄漏自己的隐私数据明文，但最终能完成特征工程处理，实现了安全的特征工程方法和装置。

[43] 图 1 为根据本公开一个实施例的特征工程方法的流程图。该方法可以是在例如第二设备处执行的。

[44] 步骤 102: 第二设备接收第一数据集，第一数据集包含多个数据对象的标签信息的密文。例如，第一设备可请求第二设备进行特征工程处理，并且第一设备可将第一数据集发送给第二设备，其中所发送的第一数据集包含多个数据对象的标签信息的密文。在另一示例中，第二设备需要进行特征工程处理并且可请求第一设备提供包含标签信息的第一数据集。在任一种情形中，第二设备可具有包含特征数据的第二数据集，并且可将第一数据集和第二数据集相结合地进行特征工程处理，例如特征选择、特征提取、特征离散化（如分箱）等。

[45] 特征数据可以表示对象的特性或行为等信息，例如用户的年龄、身高、消费记录等。标签信息可以标识对象的类别，例如用户是否为学生、用户是否为失信者等等。作为示例而非限定，标签信息可以为二元值，例如用 1、0 分别表示正样本和负样本，或者用 0、-1 分别表示正样本和负样本，或以其他方式区分正样本和负样本。

[46] 在一些情形中，虽然第二设备可将第一数据集的标签信息和第二数据集的特征数据相结合地进行特征工程处理，但拥有第一数据集的第一设备可能不希望将第一数据集的标签信息的具体值泄露给第二设备。由此，根据一个实施例，第一设备发送给第二设备的第一数据集包含多个数据对象的标签信息的密文，而不是这些标签信息的明文信息。

[47] 第一设备可以利用密钥对第一数据集之中的各个数据对象的标签信息逐个进行加密以生成标签信息的密文，或者第一设备可以接收包含由第三方加密后的标签信息密文的第一数据集。优选地，第一设备能够解密第一数据集的密文，而第二设备不能解密第一数据集的密文。因此，即使第一设备将第一数据集的标签信息密文发送给第二设备，第

二设备也不知晓每个数据对象的具体标签信息值。第一设备可以直接将包含密文的第一数据集发送给第二设备，也可以经由中间设备将第一数据集转发给第二设备。例如，第一设备可以将第一数据集上传到云端(比如阿里 oss 或者开源 sftp 等)供第二设备下载，也可以通过网络进行大文件的点对点传输提供给第二设备。第二设备可以从中间设备(例如，云端)接收第一数据集，或者接收来自匿名方的第一数据集，或者以其他方式获得包含密文的第一数据集。

[48] 步骤 104: 第二设备对第二数据集进行特征工程处理以生成第二数据集的子集。如上所述，第二数据集可包含多个数据对象的特征数据。第二设备可以对第二数据集之中所包含的多个数据对象的特征数据进行数据清理、数据筛选、特征离散化、特征提取、特征选择等特征工程处理，以生成第二数据集的子集，该子集可包括第二数据集之中的一些数据对象的特征数据。

[49] 对第二数据集的特征工程处理可以在步骤 102 之前或之后执行。例如，拥有第二数据集的第二设备可以在任何时间自主地对第二数据集进行特征工程处理，也可以在从第一设备接收到特征工程处理请求(和/或第一数据集)之后对第二数据集进行特征工程处理。

[50] 步骤 106: 第二设备根据第二数据集的子集中所包含的数据对象来生成第一数据集之中相应数据对象的标签信息的密文子集。

[51] 如上所述，第一数据集(例如，包含标签信息)可与第二数据集(例如，包含特征数据)联合地用于特征工程处理。然而，第二设备所接收到的第一数据集之中的标签信息是经加密的(即，为密文形式)，并且第二设备可能无法对第一数据集的密文进行解密，由此不知晓第一数据集的标签信息的具体值。以特征数据和经加密的标签信息为例，第二设备由于不知晓标签信息的具体值，因此不能直接用第二数据集的特征数据和第一数据集之中的经加密的标签信息来联合地进行特征工程处理。

[52] 根据本公开的一个实施例，第一数据集的标签信息与第二数据集的特征数据可对应于相同的多个数据对象。作为示例而非限定，第一设备拥有的标签信息和第二设备拥有的特征数据可各自具有标识符(id)，第一设备和第二设备可进行标识符匹配以分别在第一设备处生成第一数据集并在第二设备处生成第二数据集，使得第一数据集的标签信息与第二数据集的特征数据对应于相同的多个数据对象。此后，第一设备可在步骤 102 将第一数据集发送给第二数据集，第一数据集可包括该多个数据对象的标签信息的密文和相关联的标识符。第二设备可以根据第二数据集的子集中所包含的数据对象按照标识

符来生成第一数据集之中相应数据对象的标签信息的密文子集。

[53]作为进一步示例，如果在第一设备处生成的第一数据集和在第二设备处生成的第二数据集针对相同的多个数据对象，则还可以例如根据标识符按相同顺序来排列双方的数据，例如使得第一数据集和第二数据集的每条数据彼此对应，其中每条数据可对应于不同的对象。如果第一数据集和第二数据集的数据对象按相同的方式排序，则第一设备发送给第二设备的第一数据集可以不必包含标识符。第二设备可以按照两个数据集之间的对应关系，根据第二数据集的子集中所包含的数据对象来生成第一数据集之中相应数据对象的标签信息的密文子集。

[54]在另一实施例中，如果第二设备在接收第一数据集之前没有与第一设备进行标识符匹配，即，第一数据集可能包含与第二数据集不同的数据对象，则第二设备可在接收第一数据集之后对第一数据集和第二数据集进行标识符匹配，由此剔除一方独有的数据对象，以使得第一数据集的标签信息和第二数据集的特征数据对应于相同的多个数据对象。此后，第二设备可在步骤 106 根据第二数据集的子集中所包含的数据对象来生成第一数据集之中相应数据对象的标签信息的密文子集。

[55]以特征选择作为示例，对第二数据集的特征工程处理可包括选择第二数据集的特征数据子集，并且第二设备可生成与该特征数据子集中所包含的数据对象相对应的第一数据集之中的标签信息的密文子集。作为另一示例，对第二数据集的特征工程处理可包括对特征数据进行分箱以生成多个分箱段，由此第二设备可以生成第一数据集的与该多个分箱段相对应的多个密文子集，其中每个密文子集包括第一数据集之中的落在相应分箱段中的数据对象的标签信息密文。即，第二设备可以根据对第二数据集的特征数据的分箱来确定第一数据集之中相应的数据对象分别落在哪个分箱段中，并按分箱段来生成第一数据集之中的标签信息的密文子集。在步骤 106 中，第二设备可以生成第一数据集之中的标签信息的密文子集，并且可以知晓该密文子集中的标签信息的总数。然而如上所述，第一数据集的标签信息是加密的，第二设备不知晓这些标签信息的值，由此无法完成特征工程处理的指标计算。例如，以分箱为例，第二设备不知道该密文子集中哪些标签信息为正样本，哪些标签信息为负样本，由此不知晓每个分箱段中的标签信息的正样本数和负样本数，无法计算分箱指标，如 IV(Information Value, 信息价值)、Woe(Weight of Evidence, 证据权重)等。

[56]步骤 108: 第二设备将该密文子集的密文发送给第一设备，并从第一设备接收在解密该密文子集的密文后生成的统计信息。第二设备发送给第一设备的该密文子集的密文

可以是该密文子集中的各个标签信息的密文集合，或者对这些标签信息的密文进行某种操作或运算（例如，组合、级联、代数运算等）之后得到的密文。

[57]第一设备可以是第一数据集的拥有方并且能够对标签信息的密文进行解密。由此，第一设备可以对该密文子集的密文进行解密并生成统计信息（例如，该密文子集的正样本数和/或负样本数），并且可以将该统计信息发送给第二设备。应理解，第一设备发送的统计信息不同于第一数据集的标签信息密文的解密值，由此不会泄露第一数据集的标签信息值。

[58]以分箱为例，第二设备可以将与多个分箱段相对应的多个密文子集发送给第一设备，每个密文子集的密文可包括第一数据集之中的落在相应分箱段中的标签信息密文的集合。第一设备可以解密出每个分箱段对应的标签信息密文，并获得每个分箱段对应的标签信息的统计信息（例如，每个分箱段对应的标签信息正样本数和/或负样本数）作为上述统计信息发送给第二设备。对于其他特征工程处理（例如，特征选择等），第一设备可类似地解密每个密文子集的密文并生成特征工程处理所需的统计信息发送给第二设备。

[59]在一个示例中，第一数据集的标签信息的密文可以是使用同态加密（Homomorphic Encryption）对标签信息进行加密来生成的。同态加密允许对经同态加密后的密文进行特定的代数运算得到仍然是加密的运算结果，将其解密所得到的运算结果与对明文进行同样的运算得到的运算结果一样。换言之，这项技术令人们可以在经加密的数据中进行诸如检索、比较等操作，得出正确的结果，而在整个处理过程中无需对数据进行解密。如果标签信息为二元值并使用同态加密来加密，则第二设备可以对密文子集中的密文进行同态加法以得到密文总和，并将该密文总和发送给第一设备。第一设备解密该密文总和得到的值等同于第一设备解密该密文子集中的每个标签信息密文并求和得到的值。由此，第一设备可以解密该密文总和，并且解密结果可指示正样本数和/或负样本数。

[60]步骤 110：第二设备使用该统计信息来计算对第二数据集的特征工程处理的特征工程指标。特征工程指标可用于评价特征工程处理产生的结果。特征工程指标可根据具体的特征工程处理进行选择，例如第二数据集的子集的正样本率、第二数据集的子集的负样本率、第二数据集的子集与密文子集中的标签信息的相关度、第二数据集的子集的信息价值、第二数据集的子集的证据权重等。由此，第二设备可以根据第二数据集的子集和对应的统计信息来获得该子集的特征工程指标。以分箱为例，该统计信息可以是每个分箱段的关于标签信息的统计值，第二设备可以利用从第一设备接收的统计值并结合第

二设备在对特征数据进行分箱时获得的分箱段信息一起进行分箱指标计算。如果特征工程指标达到要求（例如，阈值），则第二设备可以生成特征工程处理结果并完成该特征工程处理。所生成的特征工程处理结果（例如，特征数据子集）可用于各种目的，例如模型训练、目标预测等。在可选的实施例 中，第二设备可以将特征工程结果发送给第一设备。在进一步的实施例中，如果特征工程指标不理想（例如，未达到阈值），第二设备可重复步骤 104-110，直到获得符合要求的特征工程指标或者直到达成终止特征工程处理的条件。

[61]根据上述方法，合作方均不必泄漏自己的隐私数据明文，但最终能完成特征工程处理。以下结合图 2-4 进一步详细描述本公开的应用场景。

[62]图 2 为根据本公开一个实施例的特征工程方法的示意图。作为示例而非限定，以第一设备拥有标签信息 Y（第一数据集），第二设备拥有特征数据 X（第二数据集）的特征选择场景为例进行说明。为简化描述，假设标签信息 Y（第一数据集）和特征数据 X（第二数据集）对应于相同的多个数据对象（例如，通过进行标识符匹配）。特征选择的目的是寻找最优特征子集。特征选择能剔除不相关（irrelevant）或冗余（redundant）的特征，从而达到减少特征个数、降低数据集维度、提高模型精确度、减少运行时间的目的。另一方面，特征选择可用于选取出真正相关的特征简化模型，协助理解数据产生的过程。

[63]在步骤 202，第一设备可加密第一数据集之中的 Y 样本并生成经加密的 Y' 样本。在步骤 204，第一设备可将经加密的 Y' 样本发送给第二设备。作为示例而非限定，第一设备可在发送之前进一步压缩经加密的 Y' 样本，并将压缩后的加密 Y' 样本发送给第二设备。若 Y' 样本数据量大，可考虑将密文文件上传到双方可访问的安全云端（例如 oss 等）。

[64]在一个实施例中，第一设备可对第一数据集之中的 Y 样本进行逐个加密以生成 Y 样本的密文 Y'。第一设备可以利用密钥以任何合适的加密算法对第一数据集之中的 Y 样本逐个进行加密，并将 Y' 样本（例如，不压缩或压缩后）发送给第二设备。作为示例而非限定，可以采用同态加密（Homomorphic Encryption）来对第一数据集之中的 Y 样本进行逐个全量同态加密。

[65]在步骤 206，第二设备可接收经加密的 Y' 样本。例如，第二设备可从第一设备或其他中间方（例如，oss 等云端）获取经加密的 Y' 样本。如果 Y' 样本是经压缩的，则第二设备可解压 Y' 样本。

[66]在步骤 208, 第二设备可选择本地拥有的 X 样本 (例如, 特征数据) 的子集。应理解, 第二设备可以按各种准则或组合来选择 X 样本子集。此外, 第二设备可以选择一个或多个 X 样本子集并依次或并行地评价这一个或多个 X 样本子集。

[67]虽然图 2 中示出了步骤 208 在步骤 206 之后执行, 但是应理解, 步骤 208 可以在步骤 202-206 中的任一个步骤之前或之后执行, 也可以与步骤 202-206 中的任一个步骤并发地执行。

[68]在步骤 210, 第二设备可根据所选择的 X 样本子集来读取相对应的 Y' 样本子集。类似于以上所述, X 样本的 id 和 Y 样本的 id 可以相匹配, 以使得 X 样本和 Y 样本具有对应关系。由此, 第二设备可根据所选择的 X 样本子集来读取相对应 id 的 Y' 样本子集。如上所述, 由于 Y' 样本是加密的, 因此第二设备不知晓 Y' 样本的值, 从而该 Y' 样本子集可包括经加密的 Y' 样本 (包含密文)。

[69]在步骤 212, 第二设备可以将 Y' 样本子集的密文发送给第一设备进行解密。在一个示例中, 在步骤 212 发送的 Y' 样本子集的密文可包括 Y' 样本子集所包含的各个 Y' 样本的集合。

[70]在步骤 214, 第一设备解密 Y' 样本子集的密文。例如, 第一设备可解密该子集中包含的每个 Y' 样本以获得经解密的 Y 样本。

[71]在步骤 216, 第一设备在解密 Y' 样本子集的密文之后可以生成关于特征选择的统计信息。例如, 第一设备可解密出各个 Y' 样本的值, 并由此获得用于评价特征选择子集的统计信息 (例如正样本数、负样本数等)。

[72]在另一个示例中, 如果 Y 样本为二元值并使用同态加密来生成 Y' 样本, 则第二设备可以对 Y' 样本子集的各个 Y' 样本进行同态加法以得到密文总和, 并在步骤 212 将该密文总和发送给第一设备, 而不必将该 Y' 样本子集完整地发送给第一设备。第一设备可在步骤 214 解密该密文总和并直接生成统计信息。例如, 如果用 1、0 分别表示正样本和负样本, 则 Y' 样本子集的同态密文总和即为所有 1 的密文总和, 该密文总和被解密后即所有正样本的数量。反之, 如果用 1、0 分别表示负样本和正样本, 则解密 Y' 样本子集的同态密文总和可以得到负样本数量。

[73]在步骤 218, 第一设备可将 Y' 样本子集的统计信息发送给第二设备。应理解, 第一设备发送 Y' 样本子集的统计信息不会泄露样本 Y 的值。

[74]在步骤 220, 第二设备可根据该统计信息来计算特征选择指标, 如 X 样本子集的正

样本率、X 样本子集的负样本率、X 样本子集与预测目标（例如，Y 样本为正或负）的相关度、X 样本子集的信息价值、X 样本子集的证据权重等。作为示例而非限定，第二设备可从步骤 220 返回到步骤 208 以选择下一组的一个或多个 X 样本子集进行评价。

[75]在步骤 222，在获得了符合要求的（例如，最优的）特征子集的情况下，第二设备可以生成特征选择结果并完成特征选择。

[76]在可选步骤 224，第二设备可将特征选择结果发送给第一设备。相应地，在可选步骤 226，第一设备可从第二设备接收特征选择结果。

[77]图 3 为根据本公开另一个实施例的特征工程方法的示意图。作为示例而非限定，以第一设备（例如，需求方）拥有标签信息 Y（第一数据集），第二设备（例如，数据方）拥有特征数据 X（第二数据集）的分箱场景为例进行说明。为简化描述，假设标签信息 Y（第一数据集）和特征数据 X（第二数据集）对应于相同的多个数据对象。

[78]分箱是一种常用的数据预处理方法，其按照属性值划分的子区间，如果一个属性值处于某个子区间范围内，就称把该把待处理的数据（某列属性值）按照一定的规则放进一些箱子中，考察每一个箱子中的数据，采用某种方法分别对各个箱子中的数据进行处理。分箱可包括等深分箱法、等宽分箱法、最小熵法和用户自定义区间法等。

[79]图 3 中的步骤 302-306 与图 2 中的步骤 202-206 相似并且不再赘述。

[80]在步骤 308，第二设备可对本地拥有的 X 样本（例如，特征数据）进行分箱。虽然图 3 中示出了步骤 308 在步骤 306 之后执行，但是应理解，步骤 308 的分箱可以在步骤 302-306 中的任一个步骤之前或之后执行，也可以与步骤 302-306 中的任一个步骤并发地执行。例如，步骤 308 处对 X 样本的分箱可以是预先（例如，在步骤 302 之前）执行的。在其他实施例中，对 X 样本的分箱可以在步骤 306 之后（或者，在接收到来自第一设备的分箱请求之后）执行。

[81]在步骤 310，第二设备以 X 样本的分箱段为单位读取相应的 Y' 样本子集。

[82]根据本公开的一个实施例中，可在步骤 308 之前的任何时间对 X 样本的 id 和 Y 样本的 id 进行匹配，以使得 X 样本和 Y 样本具有对应关系。作为示例而非限定，第一设备发送给第二设备的加密 Y' 样本具有相关联的标识符（id），第二设备本地的 X 样本具有相关联的标识符，其中 X 样本和 Y 样本的标识符是匹配的（例如，对应于相同的一组对象）。由此，第二设备可以根据 X 样本的分箱来确定相应的 id 分别落在哪个分箱段中，并按分箱段来读取各 id 对应的 Y' 样本。在一个实施例中，第一设备发送给第

二设备的加密 Y'样本的标识符可以被加密（例如，进行哈希散列）或不被加密。在另一实施例中，如果第一设备和第二设备先进行了匹配，以使得 X 样本和 Y'样本按特定方式排序（例如使得每项数据一一对应），则第一设备发送给第二设备的加密 Y'样本可以不必包括 id 信息，第二设备可以直接根据 X 样本的分箱来读取相对应的 Y'样本。

[83]在步骤 312, 第二设备可以将各个分箱段对应的 Y'样本子集的密文发送给第一设备。在一个实施例中，可在步骤 312 发送多个 Y'样本子集的密文，其中每个 Y'样本子集的密文可包括相应分箱段对应的 Y'样本的集合。在另一个实施例中，如果 Y'样本是采用同态加密来加密的，则在步骤 312 发送的每个 Y'样本子集的密文可包括将该分箱段对应的 Y'样本进行同态加法得到的密文总和。

[84]在步骤 314, 第一设备解密每个分箱段对应的 Y'样本子集的密文。第一设备可以是曾对 Y 进行加密的设备并且能够解密 Y'，或者即使加密不是由第一设备执行的，但第一设备具有密钥权限并由此能够解密 Y'。例如，第一设备可解密 Y'样本子集中包含的每个 Y'样本，或者在该子集包含同态密文总和的情况下解密该同态密文总和。如果 Y'样本是二元值且采用同态加密来加密，则解密该密文总和可以直接得出正样本数或负样本数。

[85]在步骤 316, 第一设备在解密每个分箱段对应的 Y'样本子集的密文之后可以生成关于每个分箱段的统计信息。例如，第一设备可解密出每个分箱段的各个 Y'样本的值，并由此获得每个分箱段的正样本数、负样本数、和/或总样本数等参数。在另一个实施例中，在步骤 312 处发送的 Y'样本子集的密文包含同态密文总和的情况下解密同态密文总和提供了各个分箱段的正样本数或负样本数。例如，如果用 1、0 分别表示正样本和负样本，则同态密文总和即为所有正样本的数量。

[86]在步骤 318, 第一设备可将各个分箱段的统计信息（例如 Y 样本统计信息，如正样本数、负样本数、和/或总样本数等）发送给第二设备。应理解，第一设备发送各个分箱段的统计信息不会泄露样本 Y 的值。

[87]在步骤 320, 第二设备可根据该统计信息来计算分箱指标。例如，第二设备可以利用从第一设备接收的关于每个分箱段的统计值并结合第二设备在对特征数据 X 进行分箱时获得的其他信息一起进行分箱指标计算，如各 X 样本分箱段的正样本率、各 X 样本分箱段的负样本率、各 X 样本分箱段与预测目标（例如，Y 样本为正或负）的相关度、各 X 样本分箱段的 IV（Information Value, 信息价值）、各 X 样本分箱段的 Woe（Weight of Evidence, 证据权重）等。

[88]在步骤 322, 在分箱指标有效(例如, 达到阈值)的情况下, 第二设备可以生成分箱结果并完成分箱。在分箱指标不理想的情况下, 第二设备可重复步骤 308-320 以重新执行分箱并进行分箱指标计算。

[89]在可选步骤 324, 第二设备可将分箱结果发送给第一设备。相应地, 在可选步骤 326, 第一设备可以从第二设备接收分箱结果。

[90]图 4 为根据本公开另一个实施例的特征工程方法的示意图。在一个实施例中, 一方(例如, 第一设备)可以具有特征数据 X2 和标签信息 Y, 另一方(例如, 第二设备)具有特征数据 X1。特征数据 X1 和特征数据 X2 可以彼此相同、部分相同、或者彼此不同。为简化描述, 假设标签信息 Y(第一数据集)、特征数据 X1(第二数据集)、和特征数据 X2(第三数据集)对应于相同的多个数据对象。第一设备可以如图 3 所述地将经加密的标签信息 Y' 发送给第二设备以结合特征数据 X1 进行分箱, 还可以在本地针对特征数据 X2 和标签信息 Y 进行本地分箱。

[91]在步骤 402-422 中, 第一设备和第二设备可以如参考图 3 中的步骤 302-322 所述地根据特征数据 X1 和标签信息 Y' 进行分箱并生成第一分箱结果。

[92]在步骤 424, 第二设备可以将第一分箱结果发送给第一设备。

[93]在步骤 426, 第一设备可在本地针对特征数据 X2 和未加密的标签信息 Y 进行分箱以获得第二分箱结果。应注意, 第一设备在步骤 426 进行的本地分箱可以在步骤 402-424 中的任一个步骤之前或之后执行, 或与其中任一个步骤并行地执行。

[94]在步骤 428, 第一设备可以组合第一分箱结果和第二分箱结果, 以得到组合的分箱结果。

[95]上文以特征工程处理中的特征选择和分箱为例介绍了分布式特征工程和指标计算, 使得合作方均不必泄漏自己的隐私数据明文, 但最终能完成特征工程处理。本领域技术人员可以将本公开的分式特征工程方法和装置广泛地应用于需要把各方的建模样本数据融合在一起进行处理而不想泄露样本数据明文信息的场合, 比如数据清理、特征离散化、特征提取、特征选择等。

[96]图 5 示出了根据本公开一个实施例的特征工程装置的框图。例如, 第一设备 510 可执行以上所描述的关于第一设备的操作, 第二设备 520 可执行以上所描述的关于第二设备的操作。

[97]如图 5 所示, 第一设备 510 可包括加解密模块 512、传输模块 514、处理模块 516。

加解密模块 512 可对第一数据集的标签信息进行加密以生成密文。传输模块 514 可将包含标签信息密文的第一数据集传送给第二设备 520 或其他设备或云端。

[98] 第二设备 520 可包括传输模块 522、特征工程处理模块 524 和密文子集生成模块 526。传输模块 522 可接收包含标签信息密文的第一数据集。此外，第二设备 520 可具有第二数据集（例如，存储在存储器中），第二数据集包含多个数据对象的特征数据。特征工程处理模块 524 可对第二数据集进行特征工程处理以生成第二数据集的子集。第一数据集的标签信息与第二数据集的特征数据可对应于相同的多个数据对象。密文子集生成模块 526 可根据第二数据集的子集中所包含的数据对象来生成第一数据集之中相应数据对象的标签信息的密文子集。传输模块 522 可进一步将该密文子集的密文发送给第一设备。

[99] 在第一设备 510 中，传输模块 514 可接收该密文子集的密文并将其传递给加解密模块 512，加解密模块 512 可解密所述密文子集的密文。处理模块 516 可根据该密文子集的密文被解密后的数据生成用于特征工程处理的统计信息，例如正样本数、负样本数、总样本数等。传输模块 514 可将所述统计信息提供给第二设备 520。

[100] 在第二设备 520 中，传输模块 522 接收所述统计信息并将其提供给特征工程处理模块 524。特征工程处理模块 524 可使用所述统计信息来计算对第二数据集进行的特征工程处理的特征工程指标并完成所述特征工程处理。可选地，在完成特征工程处理后，传输模块 522 还可以将特征工程处理结果发送给第一设备 510。

[101] 在进一步的实施例中，第一设备 510 可任选地包括特征工程处理模块 518，其可使用第一数据集的未加密的标签信息和第三数据集（例如，包含特征数据 X2）来执行特征工程处理以生成第二特征工程结果。第一数据集的标签信息与第三数据集的特征数据可对应于相同的多个数据对象。第一设备 510 还可以组合由第二设备 520 生成的上述特征工程结果和第二特征工程结果以获得组合的特征工程结果。在一个实施例中，处理模块 516 和特征工程处理模块 518 可以一起实现或分开实现。

[102] 根据本公开，联合建模的数据合作各方（例如，第一设备 510 和第二设备 520）不对外输出隐私数据明文，但最终得到特征工程处理结果，防止了私有数据泄漏，解决了数据合作过程中的隐私数据泄漏和数据信任问题。

[103] 以上描述的特征工程方法和装置的各个步骤和模块可以用硬件、软件、或其组合来实现。如果在硬件中实现，结合本公开描述的各种说明性步骤、模块、以及电路可用通用处理器、数字信号处理器（DSP）、专用集成电路（ASIC）、现场可编程门阵列

(FPGA)、或其他可编程逻辑组件、硬件组件、或其任何组合来实现或执行。通用处理器可以是处理器、微处理器、控制器、微控制器、或状态机等。如果在软件中实现,则结合本公开描述的各种说明性步骤、模块可以作为一条或多条指令或代码存储在计算机可读介质上或进行传送。实现本公开的各种操作的软件模块可驻留在存储介质中,如RAM、闪存、ROM、EPROM、EEPROM、寄存器、硬盘、可移动盘、CD-ROM、云存储等。存储介质可耦合到处理器以使得该处理器能从/向该存储介质读写信息,并执行相应的程序模块以实现本公开的各个步骤。而且,基于软件的实施例可以通过适当的通信手段被上载、下载或远程地访问。这种适当的通信手段包括例如互联网、万维网、内联网、软件应用、电缆(包括光纤电缆)、磁通信、电磁通信(包括RF、微波和红外通信)、电子通信或者其他这样的通信手段。

[104] 还应注意,这些实施例可能是作为被描绘为流程图、流图、结构图、或框图的过程来描述的。尽管流程图可能会把诸操作描述为顺序过程,但是这些操作中有许多操作能够并行或并发地执行。另外,这些操作的次序可被重新安排。

[105] 所公开的方法、装置和系统不应以任何方式被限制。相反,本公开涵盖各种所公开的实施例(单独和彼此的各种组合和子组合)的所有新颖和非显而易见的特征和方面。所公开的方法、装置和系统不限于任何具体方面或特征或它们的组合,所公开的任何实施例也不要求存在任一个或多个具体优点或者解决特定或所有技术问题。

[106] 上面结合附图对本公开的实施例进行了描述,但是本公开并不局限于上述的具体实施方式,上述的具体实施方式仅仅是示意性的,而不是限制性的,本领域的普通技术人员在本公开的启示下,在不脱离本公开宗旨和权利要求所保护的范围情况下,还可做出很多更改,这些均落在本公开的保护范围之内。

权利要求书

1、一种特征工程方法，其特征在于，包括：

接收第一数据集，第一数据集包含多个数据对象的标签信息的密文；

对第二数据集进行特征工程处理以生成第二数据集的子集，第二数据集包含所述多个数据对象的特征数据；

根据所述第二数据集的子集中所包含的数据对象来生成第一数据集之中相应数据对象的标签信息的密文子集；

将所述密文子集的密文发送给第一设备；

从第一设备接收在解密所述密文子集的密文后生成的统计信息；以及

使用所述统计信息来计算对第二数据集进行的特征工程处理的特征工程指标。

2、如权利要求 1 所述的特征工程方法，其特征在于，还包括：

根据所述特征工程指标来生成特征工程结果；以及

将所述特征工程结果发送给第一设备。

3、如权利要求 1 所述的特征工程方法，其特征在于，所述特征工程处理包括对第二数据集进行特征选择以生成第二数据集的特征数据子集，其中计算特征工程指标包括利用所述特征数据子集与所述统计信息来计算特征选择指标。

4、如权利要求 1 所述的特征工程方法，其特征在于，所述特征工程处理包括对第二数据集进行分箱以生成多个分箱段，并且所述密文子集包括与所述多个分箱段相对应的多个密文子集，其中每个密文子集包括第一数据集之中的落在相应分箱段中的数据对象的标签信息密文，其中计算特征工程指标包括利用所述第二数据集的分箱段与所述统计信息来计算分箱指标。

5、如权利要求 1 所述的特征工程方法，其特征在于，第一数据集的标签信息的密文是使用同态加密对所述标签信息进行加密来生成的，所述标签信息为二元值，并且将所述密文子集的密文发送给第一设备包括：对所述密文子集中的密文进行同态加法以得到密文总和并将所述密文总和发送给第一设备。

6、如权利要求 1 所述的特征工程方法，其特征在于，第一数据集和第二数据集之中的数据对象分别具有相关联的标识符，所述方法还包括在生成所述第二数据集的子集之前对第一数据集和第二数据集进行标识符匹配，以使得第一数据集的标签信息与第二数据集的特征数据对应于相同的多个数据对象。

7、如权利要求 1 所述的特征工程方法，其特征在于，所述统计信息包括所述密文子集的正样本数、和/或负样本数。

8、如权利要求 1 所述的特征工程方法，其特征在于，对第二数据集进行特征工程处理是在接收第一数据集之前、之时、或之后发生的。

9、如权利要求 1 所述的特征工程方法，其特征在于，所述特征工程指标包括以下至少一者：

所述第二数据集的子集的正样本率、所述第二数据集的子集的负样本率、所述第二数据集的子集与所述密文子集中的标签信息的相关度、所述第二数据集的子集的信息价值、所述第二数据集的子集的证据权重。

10、一种特征工程方法，其特征在于，包括：

提供第一数据集，第一数据集包含多个数据对象的标签信息的密文；

接收第一数据集的密文子集的密文，所述密文子集是根据对第二数据集进行的特征工程处理来生成的，第二数据集包含所述多个数据对象的特征数据，其中对第二数据集进行的特征工程处理生成第二数据集的子集，其中所述密文子集包含第一数据集之中的与所述第二数据集的子集相同的数据对象的标签信息密文；

解密所述密文子集的密文并生成所述密文子集的统计信息；以及

提供所述统计信息以用于计算对第二数据集的特征工程处理的特征工程指标。

11、如权利要求 10 所述的特征工程方法，其特征在于，所述方法还包括：

接收根据所述特征工程指标生成的第一特征工程结果。

12、如权利要求 11 所述的特征工程方法，其特征在于，所述方法还包括：

使用所述第一数据集的标签信息和第三数据集的特征数据来执行所述特征工程处理以生成第二特征工程结果；以及

组合第一特征工程结果和第二特征工程结果以获得组合的特征工程结果。

13、如权利要求 10 所述的特征工程方法，其特征在于，所述特征工程处理包括对第二数据集进行特征选择以生成第二数据集的特征数据子集，其中计算特征工程指标包括利用所述特征数据子集与所述统计信息来计算特征选择指标。

14、如权利要求 10 所述的特征工程方法，其特征在于，所述特征工程处理包括对第二数据集进行分箱以生成多个分箱段，并且所述密文子集包括与所述多个分箱段相对应的多个密文子集，其中每个密文子集包括第一数据集之中的落在相应分箱段中的数据对象的标签信息密文，所述计算特征工程指标包括利用所述第二数据集的分箱段与所述统计信息来计算分箱指标。

15、如权利要求 10 所述的特征工程方法，其特征在于，第一数据集的标签信息的密文是使用同态加密对所述标签信息进行加密来生成的，所述标签信息为二元值，所述

密文子集的密文包括将所述密文子集中的密文进行同态加法得到的密文总和。

16、如权利要求 10 所述的特征工程方法，其特征在于，第一数据集和第二数据集之中的数据对象分别具有相关联的标识符，所述方法还包括在生成所述第二数据集的子集之前对第一数据集和第二数据集进行标识符匹配，以使得第一数据集的标签信息与第二数据集的特征数据对应于相同的多个数据对象。

17、如权利要求 10 所述的特征工程方法，其特征在于，所述统计信息包括所述密文子集的正样本数、和/或负样本数。

18、如权利要求 10 所述的特征工程方法，其特征在于，所述特征工程指标包括以下至少一者：

所述第二数据集的子集的正样本率、所述第二数据集的子集的负样本率、所述第二数据集的子集与所述密文子集中的标签信息的相关度、所述第二数据集的子集的信息价值、所述第二数据集的子集的证据权重。

19、一种特征工程装置，其特征在于，包括：

传输模块，其接收第一数据集，第一数据集包含多个数据对象的标签信息的密文；

特征工程处理模块，其对第二数据集进行特征工程处理以生成第二数据集的子集，第二数据集包含所述多个数据对象的特征数据；以及

密文子集生成模块，其根据所述第二数据集的子集中所包含的数据对象来生成第一数据集之中相应数据对象的标签信息的密文子集；

其中所述传输模块将所述密文子集的密文发送给第一设备并从第一设备接收在解密所述密文子集的密文后生成的统计信息；并且

所述特征工程处理模块使用所述统计信息来计算对第二数据集进行的特征工程处理的特征工程指标。

20、如权利要求 19 所述的特征工程装置，其特征在于，

所述特征工程处理模块根据所述特征工程指标来生成特征工程结果；以及

所述传输模块将所述特征工程结果发送给第一设备。

21、如权利要求 19 所述的特征工程装置，其特征在于，所述特征工程处理包括特征选择或分箱。

22、如权利要求 19 所述的特征工程装置，其特征在于，所述统计信息包括所述密文子集的正样本数、和/或负样本数。

23、一种特征工程装置，其特征在于，包括：

传输模块，其向第二设备提供第一数据集，第一数据集包含多个数据对象的标签信

息的密文，所述传输模块还从第二设备接收第一数据集的密文子集的密文，所述密文子集是根据对第二数据集进行的特征工程处理来生成的，第二数据集包含所述多个数据对象的特征数据，其中对第二数据集进行的特征工程处理生成第二数据集的子集，其中所述密文子集包含第一数据集之中的与所述第二数据集的子集相同的数据对象的标签信息密文；

加解密模块，其解密所述密文子集的密文；以及

处理模块，其生成所述密文子集的统计信息，

其中所述传输模块向第二设备提供所述统计信息以供第二设备用于计算对第二数据集的特征工程处理的特征工程指标。

24、如权利要求 23 所述的特征工程装置，其特征在于，其中所述传输模块从第二设备接收根据所述特征工程指标生成的第一特征工程结果，并且所述特征工程装置还包括：

特征工程处理模块，其使用所述第一数据集的标签信息和第三数据集的特征数据来执行所述特征工程处理以生成第二特征工程结果，以及组合第一特征工程结果和第二特征工程结果以获得组合的特征工程结果。

25、如权利要求 23 所述的特征工程装置，其特征在于，所述特征工程处理包括特征选择或分箱。

26、如权利要求 23 所述的特征工程装置，其特征在于，所述统计信息包括所述密文子集的正样本数、和/或负样本数。

27、一种特征工程系统，其特征在于，包括：

处理器；

用于存储处理器可执行指令的存储器，

其中所述处理器被配置成执行所述处理器可执行指令以实现如权利要求 1-18 中任一项所述的方法。

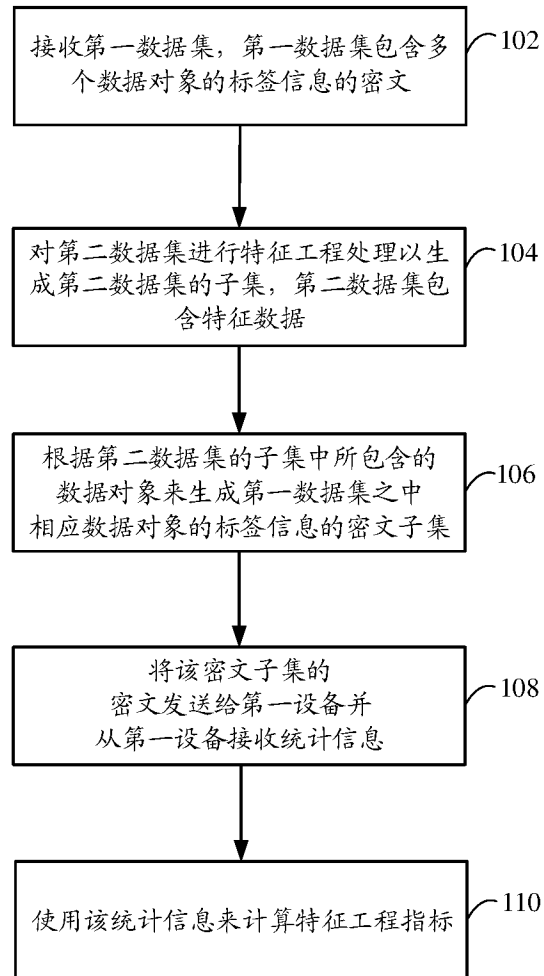


图 1

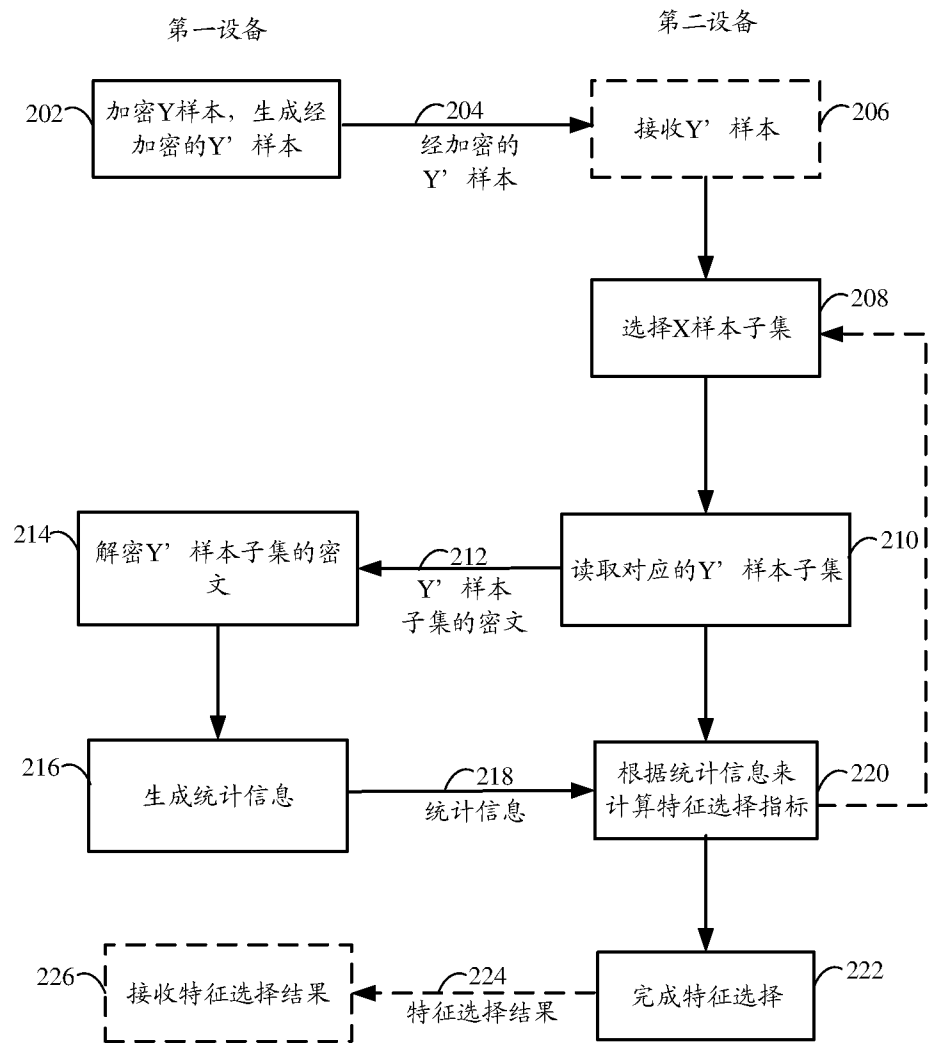


图 2

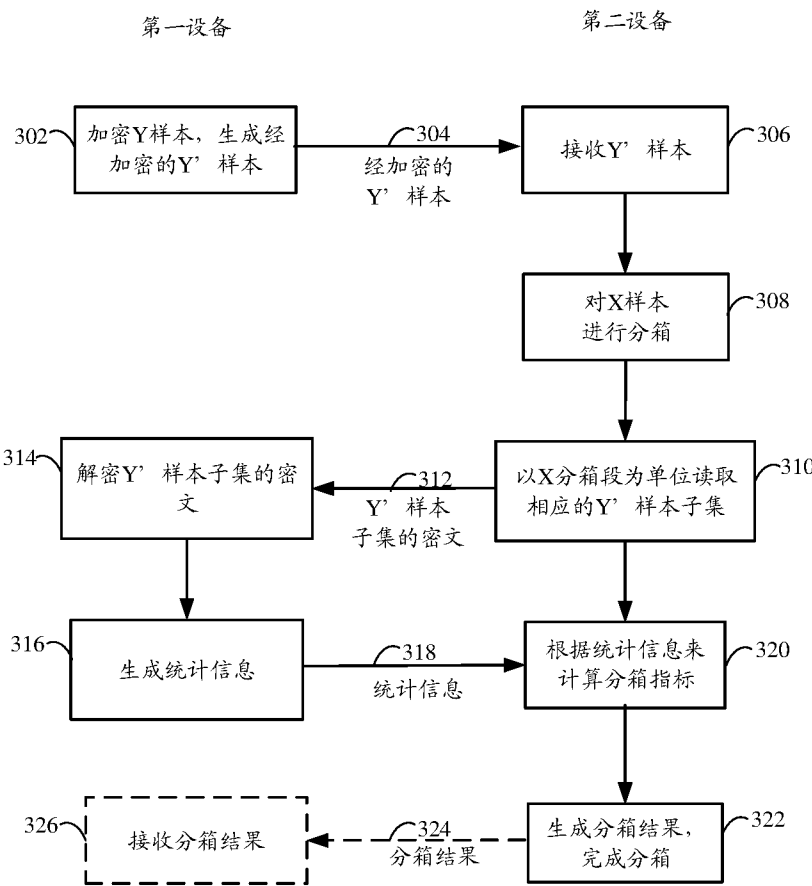


图 3

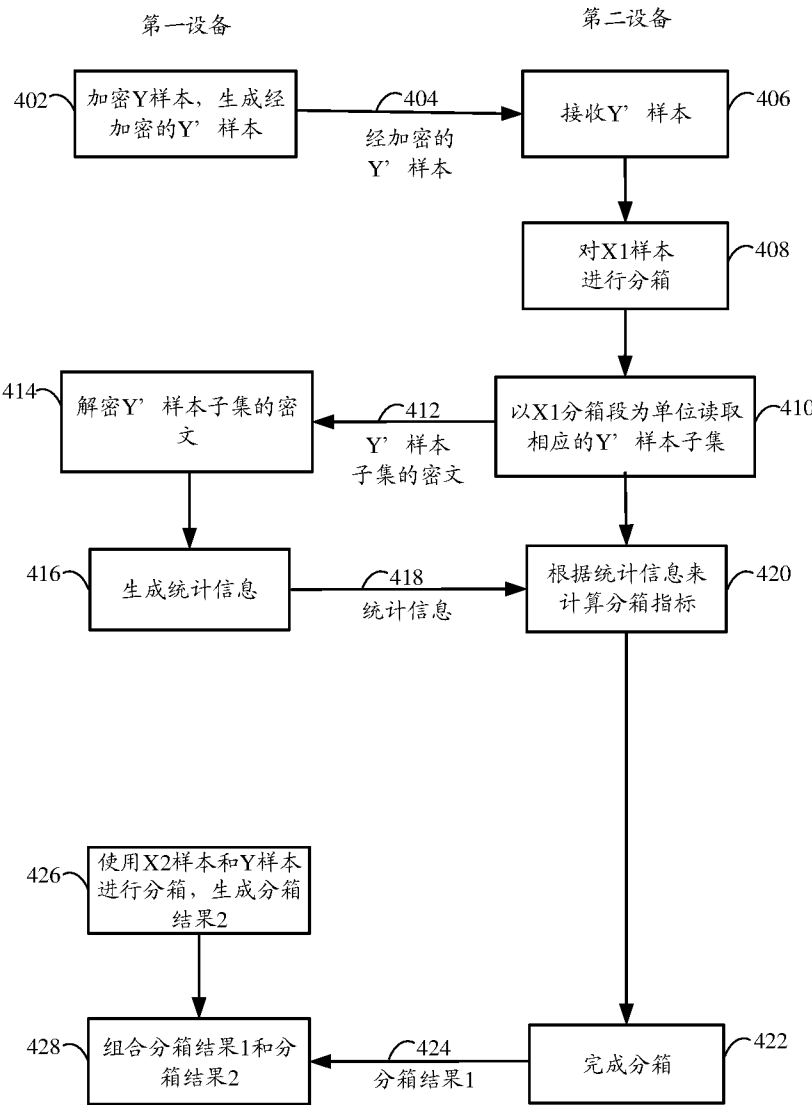


图 4

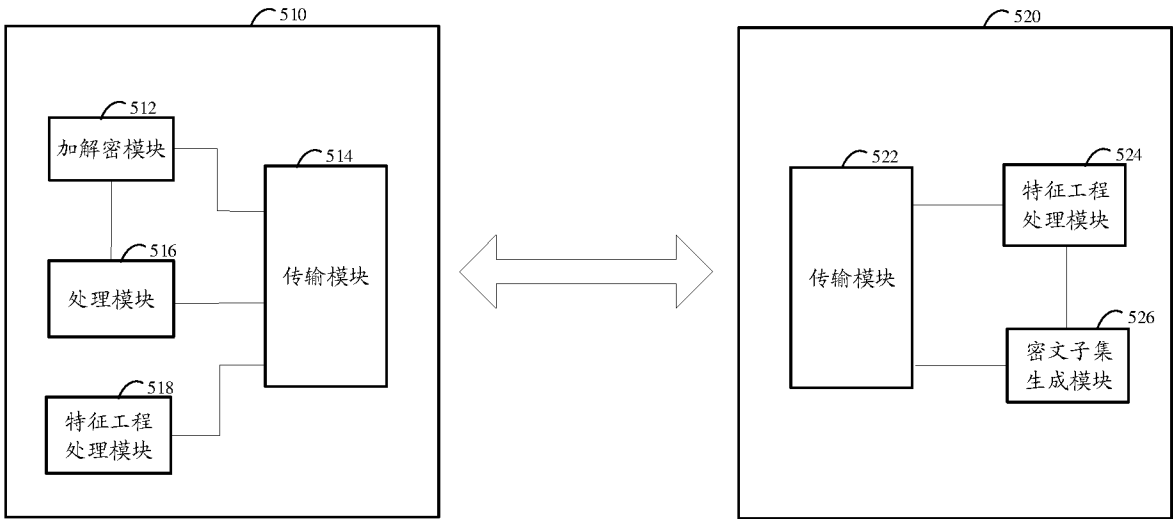


图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2020/070303

A. CLASSIFICATION OF SUBJECT MATTER

G06F 21/60(2013.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS; CNTXT; CNKI; SIPOABS; DWPI; USTXT; WOTXT; EPTXT: 加密, 同态, 标签, 正样本, 特征, 建模, 机器学习, encrypt, homomorphic, label, positive, sample, feature, modeling, machine, learning

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 110032878 A (ALIBABA GROUP HOLDING LIMITED) 19 July 2019 (2019-07-19) claims 1-27, and description, paragraphs [0041]-[0105]	1-27
X	CN 109241770 A (SHENZHEN QIANHAI WEIZHONG BANK CO., LTD.) 18 January 2019 (2019-01-18) description, paragraphs [0045]-[0155]	1-27
A	CN 105760932 A (BEIJING WUSI CHUANGXIANG TECHNOLOGY CO., LTD.) 13 July 2016 (2016-07-13) entire document	1-27
A	US 2017310643 A1 (NAT ICT AUSTRALIA LTD.) 26 October 2017 (2017-10-26) entire document	1-27



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

10 March 2020

Date of mailing of the international search report

25 March 2020

Name and mailing address of the ISA/CN

China National Intellectual Property Administration
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2020/070303

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	110032878	A	19 July 2019	None			
CN	109241770	A	18 January 2019	None			
CN	105760932	A	13 July 2016	EP	3418950	A4	23 October 2019
				EP	3418950	A1	26 December 2018
				US	2019258927	A1	22 August 2019
				CN	105760932	B	06 April 2018
				WO	2017140248	A1	24 August 2017
US	2017310643	A1	26 October 2017	WO	2016061635	A1	28 April 2016
				US	10320752	B2	11 June 2019
				AU	2015336949	A1	18 May 2017

国际检索报告

国际申请号

PCT/CN2020/070303

A. 主题的分类 G06F 21/60 (2013.01) i 按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类																	
B. 检索领域 检索的最低限度文献 (标明分类系统和分类号) G06F 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用)) CNABS; CNTXT; CNKI; SIPOABS; DWPI; USTXT; WOTXT; EPTXT: 加密, 同态, 标签, 正样本, 特征, 建模, 机器学习, encrypt, homomorphic, label, positive, sample, feature, modeling, machine, learning																	
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>PX</td> <td>CN 110032878 A (阿里巴巴集团控股有限公司) 2019年 7月 19日 (2019 - 07 - 19) 权利要求第1-27项, 说明书第[0041]-[0105]段</td> <td>1-27</td> </tr> <tr> <td>X</td> <td>CN 109241770 A (深圳前海微众银行股份有限公司) 2019年 1月 18日 (2019 - 01 - 18) 说明书第[0045]-[0155]段</td> <td>1-27</td> </tr> <tr> <td>A</td> <td>CN 105760932 A (北京物思创想科技有限公司) 2016年 7月 13日 (2016 - 07 - 13) 全文</td> <td>1-27</td> </tr> <tr> <td>A</td> <td>US 2017310643 A1 (NAT ICT AUSTRALIA LTD) 2017年 10月 26日 (2017 - 10 - 26) 全文</td> <td>1-27</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	PX	CN 110032878 A (阿里巴巴集团控股有限公司) 2019年 7月 19日 (2019 - 07 - 19) 权利要求第1-27项, 说明书第[0041]-[0105]段	1-27	X	CN 109241770 A (深圳前海微众银行股份有限公司) 2019年 1月 18日 (2019 - 01 - 18) 说明书第[0045]-[0155]段	1-27	A	CN 105760932 A (北京物思创想科技有限公司) 2016年 7月 13日 (2016 - 07 - 13) 全文	1-27	A	US 2017310643 A1 (NAT ICT AUSTRALIA LTD) 2017年 10月 26日 (2017 - 10 - 26) 全文	1-27
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求															
PX	CN 110032878 A (阿里巴巴集团控股有限公司) 2019年 7月 19日 (2019 - 07 - 19) 权利要求第1-27项, 说明书第[0041]-[0105]段	1-27															
X	CN 109241770 A (深圳前海微众银行股份有限公司) 2019年 1月 18日 (2019 - 01 - 18) 说明书第[0045]-[0155]段	1-27															
A	CN 105760932 A (北京物思创想科技有限公司) 2016年 7月 13日 (2016 - 07 - 13) 全文	1-27															
A	US 2017310643 A1 (NAT ICT AUSTRALIA LTD) 2017年 10月 26日 (2017 - 10 - 26) 全文	1-27															
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																	
<table border="0"> <tr> <td> * 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 </td> <td> “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件 </td> </tr> </table>			* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件	“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件													
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件	“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件																
国际检索实际完成的日期		国际检索报告邮寄日期															
2020年 3月 10日		2020年 3月 25日															
ISA/CN的名称和邮寄地址		受权官员															
中国国家知识产权局 (ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451		穆小川 电话号码 (86-512) 88995837															

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2020/070303

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	110032878	A	2019年 7月 19日	无			
CN	109241770	A	2019年 1月 18日	无			
CN	105760932	A	2016年 7月 13日	EP	3418950	A4	2019年 10月 23日
				EP	3418950	A1	2018年 12月 26日
				US	2019258927	A1	2019年 8月 22日
				CN	105760932	B	2018年 4月 6日
				WO	2017140248	A1	2017年 8月 24日
US	2017310643	A1	2017年 10月 26日	WO	2016061635	A1	2016年 4月 28日
				US	10320752	B2	2019年 6月 11日
				AU	2015336949	A1	2017年 5月 18日