**专利申请著录项目表**

接案号：TF240811872

(与： 同日申报) (与： 一案两报) 投诉二维码

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 客户联系人 | 姓名：刘歆 电话：15923118908 电子邮箱：liuxin@cqupt.edu.cn | | | | | | |
| 专利代理人 | 姓名：安纹西 电话：023-86898822-807；13594600052 电子邮箱：wenxian@thycq.com  QQ：2355252419 | | | | | | |
| **特别提醒：**尊敬的客户，上述代理人的联系电话和邮箱为我司指定的工作电话和工作邮箱，为了保证您的技术方案的保密性和后续工作衔接的顺畅，请您务必采用上述联系方式与我方工作人员进行联系。 | | | | | | | |
| ①专利名称 | 一种基于纵向联邦学习的参与方对齐样本生成系统 | | | | | | |
| ②专利类型 | 发明 实用新型 外观设计 一案两报 | | | | | 是否费减： 是 否 | |
| ③所有发明人 | 刘歆，何杭轩，钱鹰，万邦睿，陈奉 | | | | | | |
| ④第一发明人国籍 | | 中国 | | 身份证号 | 51021919830324502X | | |
| ⑤申请人 | 申  请  人  (1) | 姓名或名称 | | 重庆邮电大学 | | | |
|  |  | 组织机构代码/身份证号 | | 125000004504018996 | | | |
|  |  | 邮政编码 | 400065 | 详细地址 | 重庆市南岸区黄桷垭崇文路2号 | | |
|  | 申  请  人  (2) | 姓名或名称 | |  | | | |
|  |  | 组织机构代码/身份证号 | |  | | | |
|  |  | 邮政编码 |  | 详细地址 |  | | |
|  | 申  请  人  (3) | 姓名或名称 | |  | | | |
|  |  | 组织机构代码/身份证号 | |  | | | |
|  |  | 邮政编码 |  | 详细地址 |  | | |
| ⑥提前公布 | 请求提前公布该专利申请（只适用于发明专利申请） | | | | | |
| ⑦实质审查 | 在提交专利申请的同时提交实质审查请求（只适用于发明专利申请） | | | | | |
| 特殊专利申请信息，涉及该项内容时填写 | | | | | | | |
| ⑧分案申请 | 原申请号： 针对的分案申请号： 原申请日： 年 月 日 | | | | | | |
| ⑨生物材料样品 | 保藏单位： 地址： | | | | | | |
|  | 保藏日期： 年 月 日 保藏编号： 分类命名： | | | | | | |
|  | 在提交专利申请的同时提交生物材料样品保藏及存活证明 | | | | | | |
| ⑩序列表 | 本专利申请涉及核苷酸或氨基酸序列表 | | | | | | |
| 要求优先权声明 | 原受理机构名称： 在先申请日： 年 月 日 在先申请号： | | | | | | |
|  | 原受理机构名称： 在先申请日： 年 月 日 在先申请号： | | | | | | |
| 备注 | 指定说明书附图中的图 1 为摘要附图 | | | | | | |

北京同恒源知识产权代理有限公司

一、本表由代理人预先填写，请联系人仔细核对信息是否正确，若有错误或缺漏请修改或补充。

二、本表第③栏，发明人是指对发明创造的实质性特点作出创造性贡献的人。发明人应当是个人（自然人）。发明人有两个发明人可以请求国家知识产权局不公布其姓名，若请求不公布姓名，应当在此栏所填写的相应发明人后面注明“（不公布姓名）”。

三、本表第④栏，第一发明人是中国港、澳、台地区的，其国籍应填写为“中国”。第一发明人为中国内地居民的，还应当同时填写居民身份证件号码（或军官证号码）。

四、本表第⑥栏，提前公布是指在发明专利申请初步审查合格后立即进入公布准备。如果不请求提前公布，则该发明专利申请将在自申请日起满十八个月时公布。由于发明专利申请必须在公布以后才能进入实质审查程序，为了加快申请的审查进程，在申请人无特别要求的情况下，本公司默认勾选此栏。如申请人不要求提前公布，请去除勾选并及时通知代理人。

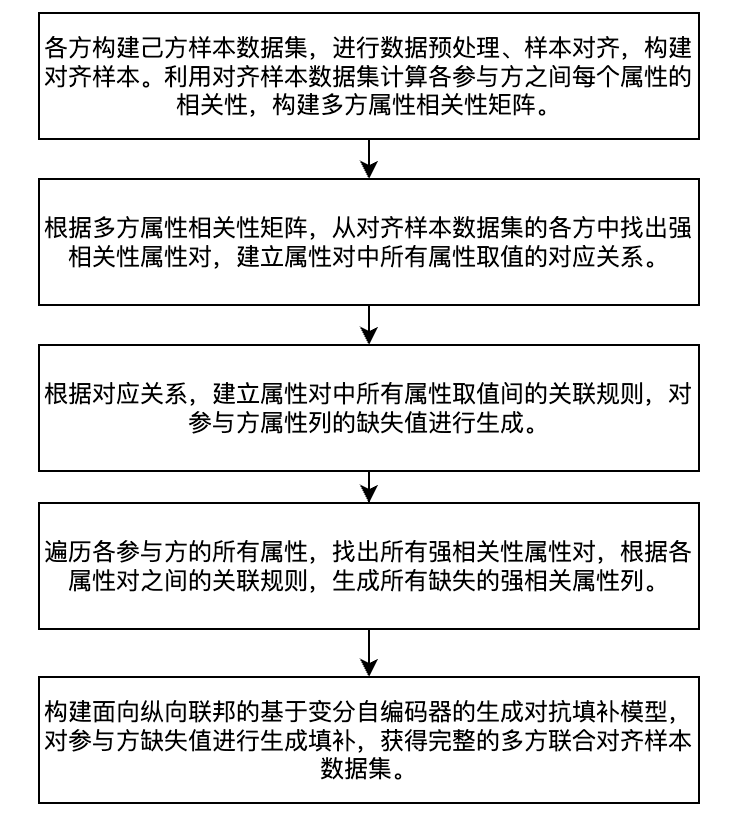
五、本表第⑦栏，实质审查是指审查员对发明专利申请是否符合授权条件（包括新颖性、创造性、实用性、公开充分、单一性问题等）进行审查。申请人可以在自申请日（有优先权的，指优先权日）起三年内提出实质审查请求来启动实质审查程序。如果在提交专利申请的同时提交实质审查请求，则该发明专利申请在公布后立即进入实质审查阶段。为了加快发明专利申请的审查进程，在申请人无特别要求的情况下，本公司默认勾选此栏。如申请人不想在提交专利申请的同时提交实质审查请求，请去除勾选并及时通知代理人。

六、本表第⑨栏，发明专利申请涉及公众不能得到的生物材料的，应当填写此栏，并自申请日起四个月内提交由保藏单位出具的该生物材料样品的保藏及存活证明。

七、本表第⑩栏，发明专利申请涉及核苷酸或氨基酸序列表的，应当填写此栏，并在提交专利申请文件的同时提交核苷酸或氨基酸序列表的计算机可读形式副本。

八、本表第栏，申请人要求外国或者本国优先权的，应当填写此栏。

本发明涉及一种基于纵向联邦学习的参与方对齐样本生成系统，属于数据生成填补技术领域，包括：多方属性相关性矩阵构建模块：用于将多个参与方的样本数据集进行对齐，并计算各参与方之间每个属性的相关性，构建多方属性相关性矩阵；属性对对应关系建立模块：用于根据多方属性相关性矩阵，分别从对齐样本数据集的各方中找出具有强相关性的属性对，建立该属性对中两个属性列所有取值间的对应关系；缺失值生成模块：用于根据对应关系，建立两个属性列取值的关联规则，对参与方属性列的缺失值进行生成；生成对抗填补模块：利用参与方数据样本和生成的属性列缺失值对参与方剩余属性列的缺失值进行生成填补，获得完整的多方联合对齐样本数据集。



1. 一种基于纵向联邦学习的参与方对齐样本生成系统，其特征在于：包括：

多方属性相关性矩阵构建模块：用于将多个参与方的样本数据集进行对齐，并计算各参与方之间每个属性的相关性，构建多方属性相关性矩阵；

属性对对应关系建立模块：用于根据多方属性相关性矩阵，分别从对齐样本数据集的各方中找出具有强相关性的属性对，建立该属性对中两个属性列所有取值间的对应关系；

缺失值生成模块：用于根据对应关系，建立两个属性列取值的关联规则，对参与方属性列的缺失值进行生成；

生成对抗填补模块：将参与方数据样本和生成的属性列缺失值输入基于变分自编码器的生成对抗填补模型，对参与方剩余属性列的缺失值进行生成填补，获得完整的多方联合对齐样本数据集。

2. 根据权利要求1所述的基于纵向联邦学习的参与方对齐样本生成系统，其特征在于：对于所述参与方包括拥有大量样本数据的A方和少量样本数据的B方，以及A方和B方均信任的协调方C方的情况下，所述多方属性相关性矩阵构建模块构建多方属性相关性矩阵的步骤如下：

S11：A方和B方各自构建己方样本数据集进行数据预处理，得到A方数据集，共有个样本数据；B方数据集，共有个样本数据；

S12：对A方和B方数据集和按照其样本ID进行样本加密对齐，得到个对齐样本，剩余的-个样本为未对齐样本；A方的对齐数据集为，未对齐数据集为；和共同构成了A方数据集，表示A方数据集第*i*个样本，，，表示A方第*i*个样本的第个属性，其中a；A方参与对齐数据集样本个数，未对齐数据集样本个数-；B方的对齐数据集为，表示B方数据集第*i*个样本，，，表示B方第个样本的第个属性，其中，B方参与对齐数据集样本个数；

S13：利用对齐数据集和，从中获取A方所有样本第*m*个属性的集合，表示A方第*i*个样本的第个属性，；从中获取B方所有样本第*n*个属性的集合，表示B方第*i*个样本的第*n*个属性，；根据斯皮尔曼相关性系数计算A方第*m*个属性和B方第*n*个属性之间的属性相关性；

S14：迭代步骤S13，计算A方所有属性与B方所有属性之间的相关性系数，其中，得到多方属性相关性矩阵M：

。

3. 根据权利要求2所述的基于纵向联邦学习的参与方对齐样本生成系统，其特征在于：步骤S11所述数据预处理包括冗余数据处理、缺失值处理、异常值处理以及数据标准化处理；

所述冗余数据处理具体包括：通过数据记录中的某些字段判断数据是否重复，对于重复数据，只保留其中一个，将其他重复数据从数据集中删除，同时保留原始数据的备份以便需要时进行回溯和对比分析；

所述缺失值处理具体包括：对数据中的每个特征进行缺失率统计，对于缺失特征数目大于总体样本特征规模一半的数据视为无效数据，进行剔除；其余数据根据数据分布，采用特定方法进行填充，其中包括众数填充、均值填充、中位数填充和插值法填充；

所述异常值处理具体包括：根据实际情况，使用标准差、四分位数范围方法确认异常值的阈值；利用箱线图、散点图、直方图或密度图工具对数据进行可视化分析，发现超出阈值的数据；选择用中位数或均值代替异常值的方法对异常值进行处理，处理完之后再次检查数据集，确保异常值已被处理且数据集的基本特征没有发生重大变化；

所述数据标准化处理具体包括：将已有特征按照数据类型分为连续型特征和离散型特征，对连续型特征采用最大-最小标准化进行处理，对离散型特征采用独热编码进行处理。

4. 根据权利要求2所述的基于纵向联邦学习的参与方对齐样本生成系统，其特征在于：在步骤S13中，A方第*m*个属性和B方第*n*个属性相关性系数的具体计算流程如下：

S131：协调方C方生成密钥对，所述密钥对包括公钥和私钥，利用非对称加密对公钥进行加密，并将加密后的公钥发送至A方和B方；

S132：A方获得从C方发送过来的加密公钥，进行解密操作，获得解密后的公钥；针对A方第*m*个属性集合，循环遍历A方第*m*个属性集合中的所有样本的第*m*个属性值，其中，利用解密后的公钥对每个属性值进行加密，并且将加密后得每个属性值发送给C方；

S133：B方获得从C方发送过来的加密公钥，进行解密操作，获得解密后的公钥；针对B方第*n*个属性集合，循环遍历的每个属性值，其中，利用解密后的公钥对每个属性值进行加密，并且将加密后得每个属性值发送给C方；

S134：协调方C方采用私钥对加密后的A方第*m*个属性集合所对应的每个属性值和B方第*n*个属性集合所对应的每个属性值进行解密，得到解密后的A方第*m*个属性集合所对应的每个属性值和B方第*n*个属性集合所对应的每个属性值；

S135：对A方第*m*个属性集合所对应的每个属性值进行从小到大排序，对B方第*n*个属性集合所对应的每个属性值进行从小到大排序，依次计算与之间的差值，其中；

S136：根据差值进行斯皮尔曼相关性系数计算，得到相关性系数，计算公式如下：

其中为总的样本数，即。

5. 根据权利要求1所述的基于纵向联邦学习的参与方对齐样本生成系统，其特征在于：所述属性对对应关系建立模块根据多方属性相关性矩阵M，找出其中最大的相关性系数，假设该相关性系数分别由A方第*m*个属性和B方第*n*个属性计算得到，则认为A方属性和B方属性具有强相关性，可组成一组强相关性属性对)，同时将重新设为0；根据该属性对)，循环遍历和中所有取值和，其中，建立属性中各取值与属性的取值之间的对应关系R，具体包括以下步骤：

S21：循环遍历相关性矩阵M中的每一个值，找出当前矩阵中最大的相关性系数，根据找出具有强相关性的一组属性对)，同时重新将的值设为0；

S22：根据属性对)，循环遍历和中所有属性列取值和，建立属性中各取值与属性的取值之间的对应关系表R。

6. 根据权利要求5所述的基于纵向联邦学习的参与方对齐样本生成系统，其特征在于：步骤S22中，设属性总共包含个属性取值，其中第*k*个取值为*mk*，，属性总共包含t个属性取值，其中第q个取值为，建立与属性取值之间的对应关系表R的具体流程如下：

S221：协调方C方初始化一个空的对应关系表，用来存储属性中的每个取值与属性的取值之间的对应关系，同时C方生成密钥对，所述密钥对包括公钥和私钥，利用非对称加密对公钥进行加密，并将加密后的公钥发送至A方和B方；

S222：A方和B方各自循环遍历属性和，并且同时将对应的属性值和通过加密的公钥进行加密发送给C方；

S223：C方记录当分别为时，对应的属性取值；同时C方记录和各取值之间的多对多的关系表，记为R。

7. 根据权利要求1所述的基于纵向联邦学习的参与方对齐样本生成系统，其特征在于：所述缺失值生成模块根据对应关系R，建立和两个属性列中各取值的关联规则；循环遍历A方未对齐数据集中第*m*个属性中的每个样本值，其中，利用关联规则，为A方未对齐数据集中第*m*个属性中的每个样本值生成对应的B方缺失数据集中第*n*个属性中的样本值；具体包括以下步骤：

S31：利用Apriori算法通过对应关系表R，建立和两个属性列中的关联规则；

S32：循环遍历A方未对齐数据集中每个样本在第*m*个属性集合中的值，利用关联规则，为B方缺失数据集生成与A方对应的样本值在B方第*n*个属性集合的样本值。

8. 根据权利要求7所述的基于纵向联邦学习的参与方对齐样本生成系统，其特征在于：步骤S31建立和两个属性列中各取值的关联规则具体流程如下：

S311：A方遍历属性，记录属性取值为的样本个数，并将加密发送给C方；B方遍历属性集合，根据的关系表R，记录对应中的各属性取值的样本个数，并加密发送给C方；

S312：C方计算各属性取值的样本个数占总样本个数的比率，计算公式如下：

其中。

S313：同理，按照所述步骤S311与S312，计算所有属性取值与在对应关系表R中各属性取值的样本占比；建立属性与的关联规则rule，表示如下：

其中表示对应的的样本个数占总样本个数的比率，表示对应的的样本个数占总样本个数的比率。

9. 根据权利要求1所述的基于纵向联邦学习的参与方对齐样本生成系统，其特征在于：循环执行所述属性对对应关系建立模块和缺失值生成模块H轮，每轮循环找出当前相关性最强一组属性对)建立该组属性对各属性取值之间的对应关系，利用每轮的对应关系建立当前属性对之间的关联规则；根据属性列之间的关联规则，对B方缺失数据集中与A方属性具有相关性大于的属性列进行生成，其中为人为给定的阈值；循环迭代轮数H等于A、B两方相关性大于的属性列对数；

所述生成对抗填补模块中构建有面向纵向联邦的基于变分自编码器的生成对抗填补模型，将和由所述属性对对应关系建立模块和缺失值生成模块循环执行H轮后生成的属性列放入填补模型中进行训练，对参与方剩余属性列的缺失值进行生成填补，获得完整的多方联合对齐样本数据集，具体包括以下步骤：

S41：A方和B方各自构建己方的底层子模型，其中包括各方的生成器模型和判别器模型；C方构建顶层判别器模型，用于对A、B两方的数样本进行联合判定；

S42：A、B双方的底层生成器模型由一个变分自编码器组成，包括编码器、编码向量和解码器，其双方底层输入包括各方数据样本、各方掩码向量以及噪声样本，双方对输入数据做如下处理：

(1)A方掩码向量，，B方掩码向量，，A方和B方时，代表该位置的数据缺失；当时，代表数据完整；A方特征数据的编码向量为，A方剩余缺失样本数据向量：

即，其中表示向量间的元素级乘法；

B方特征数据的编码向量为，A方剩余缺失样本数据向量：

即，其中表示向量间的元素级乘法；

(2)A方和B方分别对噪声样本和噪声样本进行编码，并用编码后的结果对缺失样本进行填充，得到和，，；

S43：A、B、C三方的判别模型由三个全连接网络构成，A、B两方各自对己方生成模型的生成数据进行处理以满足判别器的输入；

S44：A方将生成数据、缺失数据与掩码向量做元素级乘法得到判别模型的输入数据，用表示，即；B方将生成数据、缺失数据与掩码向量做元素级乘法得到判别模型的输入数据，用表示，即；

S45：A、B两方底层损失函数的包含三部分：生成样本中观测变量与生成变量间的误差损失，生成样本输入判别器中得到的判别损失以及变分自编码器的KL散度损失；其中A、B双方重建损失表示为：

双方判别损失表示为：

其中为A方对掩码矩阵进行随机采样得到的提示矩阵，为B方对掩码矩阵进行随机采样得到的提示矩阵；

KL散度损失表示为：

其中和表示为A、B双方方编码器产生的潜在向量的以为条件的数据分布；为标准正态分布；是编码器输出的潜在向量的均值；是编码器输出的潜在向量的方差；

分别得到A方、B方各自的总损失如下：

其中为超参数，为C方判别器损失，计算如下：

S46：将数据集和生成的属性列缺失值放入填补模型中进行训练，对参与方剩余属性列的缺失值进行生成填补，获得完整的多方联合对齐样本数据集。

10. 根据权利要求9所述的基于纵向联邦学习的参与方对齐样本生成系统，其特征在于：所述步骤S46具体包括以下步骤：

S461：C方初始化顶部判别器网络，A、B双方初始化各自的底部生成器网络和底部判别器网络；

S462：A、B方固定生成器训练判别器，具体包括：A、B双方采用同一批次的掩码向量、真实数据和随机噪声；经过数据编码、噪声填充以及向量拼接操作后输入至双方生成器得到生成数据样本与；A、B双方分别对掩码矩阵进行随机采样得到各自的提示矩阵与；进行向量拼接操作后输入至双方判别器网络得到双方数据的特征表示为与，将与发送至C方后，C方进行向量拼接并输入至C方判别器网络得到生成数据的判别结果；A、B双方将真实数据输入至双方各自的判别器网络后得到与，将与发送至C方后，C方进行向量拼接并输入至C方判别器网络得到真实数据的判别结果；经过计算，通过Adam优化器更新C方判别器网络参数以及A、B双方判别器参数，计算公式如下：

其中代表期望值，和分别代表集合和集合服从、概率分布；

S463：A、B双方固定判别器训练生成器，具体包括：双方采样同一批次的真实数据、掩码向量和随机噪声；经过数据编码、噪声填充以及向量拼接等操作后输入至双方生成器得到生成数据样本与；A、B双方分别对掩码矩阵进行随机采样得到各自的提示矩阵与；进行向量拼接操作后输入至双方判别器网络得到生成数据的判别结果；A、B双方计算各自的均方误差损失通过同态加密发送至服务端进行损失聚合，得到总损失，其中为超参数，通过Adam优化器更新A、B双方生成器网络。

**一种基于纵向联邦学习的参与方对齐样本生成系统**

### 技术领域

本发明属于数据生成填补技术领域，涉及一种基于纵向联邦学习的参与方对齐样本生成系统。

### 背景技术

在纵向联邦学习场景下，某些参与方样本较多，而另一些参与方样本较少，样本对齐后，基于多方的联合对齐样本数量将会参照样本较少的参与方，即可视为联合对齐样本在这些参与方中出现了数据缺失。这将导致对齐后的样本无法支持纵向联邦场景下机器学习模型基于大量样本训练的需要。现有的数据生成填补方法大部分都是针对集中式的数据，即单方数据样本进行生成填补，而对于纵向联邦学习场景下的多方数据集的缺失，利用现有数据生成填补方法结合纵向联邦学习只能解决多方数据集缺失的其中一种情况，即各参与方对齐数据样本的缺失，无法解决上述联合对齐样本在样本较少的参与方出现的数据缺失问题。

### 发明内容

有鉴于此，本发明的目的在于提供基于纵向联邦学习的参与方对齐样本生成系统。

为达到上述目的，本发明提供如下技术方案：

一种基于纵向联邦学习的参与方对齐样本生成系统，包括：

多方属性相关性矩阵构建模块：用于将多个参与方的样本数据集进行对齐，并计算各参与方之间每个属性的相关性，构建多方属性相关性矩阵；

属性对对应关系建立模块：用于根据多方属性相关性矩阵，分别从对齐样本数据集的各方中找出具有强相关性的属性对，建立该属性对中两个属性列所有取值间的对应关系；

缺失值生成模块：用于根据对应关系，建立两个属性列取值的关联规则，对参与方属性列的缺失值进行生成；

生成对抗填补模块：将参与方数据样本和生成的属性列缺失值输入基于变分自编码器的生成对抗填补模型，对参与方剩余属性列的缺失值进行生成填补，获得完整的多方联合对齐样本数据集。

进一步，对于所述参与方包括拥有大量样本数据的A方和少量样本数据的B方，以及A方和B方均信任的协调方C方的情况下，所述多方属性相关性矩阵构建模块构建多方属性相关性矩阵的步骤如下：

S11：A方和B方各自构建己方样本数据集进行数据预处理，得到A方数据集，共有个样本数据；B方数据集，共有个样本数据；

S12：对A方和B方数据集和按照其样本ID进行样本加密对齐，得到个对齐样本，剩余的-个样本为未对齐样本；A方的对齐数据集为，未对齐数据集为；和共同构成了A方数据集，表示A方数据集第*i*个样本，，，表示A方第*i*个样本的第个属性，其中a；A方参与对齐数据集样本个数，未对齐数据集样本个数-；B方的对齐数据集为，表示B方数据集第*i*个样本，，，表示B方第个样本的第个属性，其中，B方参与对齐数据集样本个数；

S13：利用对齐数据集和，从中获取A方所有样本的第*m*个属性的集合，表示A方第*i*个样本的第个属性，；从中获取B方所有样本第*n*个属性的集合，表示B方第*i*个样本的第*n*个属性，；根据斯皮尔曼相关性系数计算A方第*m*个属性和B方第*n*个属性之间的属性相关性；

S14：迭代步骤S13，计算A方所有属性与B方所有属性之间的相关性系数，其中，得到多方属性相关性矩阵M：

。

进一步，步骤S11所述数据预处理包括冗余数据处理、缺失值处理、异常值处理以及数据标准化处理；

所述冗余数据处理具体包括：通过数据记录中的某些字段判断数据是否重复，对于重复数据，只保留其中一个，将其他重复数据从数据集中删除，同时保留原始数据的备份以便需要时进行回溯和对比分析；

所述缺失值处理具体包括：对数据中的每个特征进行缺失率统计，对于缺失特征数目大于总体样本特征规模一半的数据视为无效数据，进行剔除；其余数据根据数据分布，采用特定方法进行填充，其中包括众数填充、均值填充、中位数填充和插值法填充；

所述异常值处理具体包括：根据实际情况，使用标准差、四分位数范围方法确认异常值的阈值；利用箱线图、散点图、直方图或密度图工具对数据进行可视化分析，发现超出阈值的数据；选择用中位数或均值代替异常值的方法对异常值进行处理，处理完之后再次检查数据集，确保异常值已被处理且数据集的基本特征没有发生重大变化；

所述数据标准化处理具体包括：将已有特征按照数据类型分为连续型特征和离散型特征，对连续型特征采用最大-最小标准化进行处理，对离散型特征采用独热编码进行处理。

进一步，在步骤S13中，A方第*m*个属性和B方第*n*个属性相关性系数的具体计算流程如下：

S131：协调方C方生成密钥对，所述密钥对包括公钥和私钥，利用非对称加密对公钥进行加密，并将加密后的公钥发送至A方和B方；

S132：A方获得从C方发送过来的加密公钥，进行解密操作，获得解密后的公钥；针对A方第*m*个属性集合，循环遍历A方第*m*个属性集合中的所有样本的第*m*个属性值，其中，利用解密后的公钥对每个属性值进行加密，并且将加密后得每个属性值发送给C方；

S133：B方获得从C方发送过来的加密公钥，进行解密操作，获得解密后的公钥；针对B方第*n*个属性集合，循环遍历的每个属性值，其中，利用解密后的公钥对每个属性值进行加密，并且将加密后得每个属性值发送给C方；

S134：协调方C方采用私钥对加密后的A方第*m*个属性所对应的每个属性值和B方第*n*个属性所对应的每个属性值进行解密，得到解密后的A方第*m*个属性所对应的每个属性值和B方第*n*个属性所对应的每个属性值；

S135：对A方第*m*个属性集合所对应的每个属性值进行从小到大排序，对B方第*n*个属性集合所对应的每个属性值进行从小到大排序，依次计算与之间的差值，其中；

S136：根据差值进行斯皮尔曼相关性系数计算，得到相关性系数，计算公式如下：

其中为总的样本数，即。

进一步，所述属性对对应关系建立模块包括根据多方属性相关性矩阵M，找出其中最大的相关性系数，假设该相关性系数分别由A方第*m*个属性和B方第*n*个属性计算得到，则认为A方属性和B方属性具有强相关性，可组成一组强相关性属性对)，同时将重新设为0；根据该属性对)，循环遍历和中所有属性列取值和，其中，建立属性中各取值与属性的取值之间的对应关系R，具体包括以下步骤：

S21：循环遍历相关性矩阵M中的每一个值，找出当前矩阵中最大的相关性系数，根据找出具有强相关性的一组属性对)，同时重新将的值设为0；

S22：根据属性对)，循环遍历和中所有属性列取值和，建立属性中各取值与属性的取值之间的对应关系表R。

进一步，步骤S22中，设属性总共包含k个属性取值：，属性总共包含q个属性取值：，建立与属性取值之间的对应关系表R的具体流程如下：

S221：协调方C方初始化一个空的对应关系表，用来存储属性中的每个取值与属性的取值之间的对应关系，同时C方生成密钥对，所述密钥对包括公钥和私钥，利用非对称加密对公钥进行加密，并将加密后的公钥发送至A方和B方；

S222：A方和B方各自循环遍历属性和，并且同时将对应的属性值和通过加密的公钥进行加密发送给C方；

S223：C方记录当分别为时，对应的属性取值；同时C方记录和各取值之间的多对多的关系表，记为R。

进一步，所述缺失值生成模块根据对应关系R，建立和两个属性列中各取值的关联规则；循环遍历A方未对齐数据集中第*m*个属性中的每个样本值，其中，利用关联规则，为A方未对齐数据集中第*m*个属性中的每个样本值生成对应的B方缺失数据集中第*n*个属性中的样本值；具体包括以下步骤：

S31：利用Apriori算法通过对应关系表R，建立和两个属性列中的关联规则；

S32：循环遍历A方未对齐数据集中每个样本在第*m*个属性集合中的值，利用关联规则，为B方缺失数据集生成与A方对应的样本值在B方第*n*个属性集合的样本值；

进一步，步骤S31建立和两个属性列中各取值的关联规则具体流程如下：

S311：A方遍历属性，记录属性取值为的样本个数，并将加密发送给C方；B方遍历属性，根据的关系表R，记录对应中的各属性取值的样本个数，并加密发送给C方；

S312：C方计算各属性取值的样本个数占总样本个数的比率，计算公式如下：

其中。

S313：同理，按照所述步骤S311与S312，计算所有属性取值与在对应关系表R中各属性取值的样本占比；建立属性与的关联规则rule，表示如下：

其中表示对应的的样本个数占总样本个数的比率，表示对应的的样本个数占总样本个数的比率。

进一步，循环执行所述属性对对应关系建立模块和缺失值生成模块H轮，每轮循环找出当前相关性最强一组属性对)建立该组属性对各属性取值之间的对应关系，利用每轮的对应关系建立当前属性对之间的关联规则；根据属性列之间的关联规则，对B方缺失数据集中与A方属性具有相关性大于的属性列进行生成，其中为人为给定的阈值；循环迭代轮数H等于A、B两方相关性大于的属性列对数。

进一步，所述生成对抗填补模块包括构建面向纵向联邦的基于变分自编码器(VAE)的生成对抗填补模型，将和由所述属性对对应关系建立模块和缺失值生成模块循环执行H轮后生成的属性列放入填补模型中进行训练，对参与方剩余属性列的缺失值进行生成填补，获得完整的多方联合对齐样本数据集，具体包括以下步骤：

S41：A方和B方各自构建己方的底层子模型，其中包括各方的生成器模型和判别器模型；C方构建顶层判别器模型，用于对A、B两方的数样本进行联合判定；

S42：A、B双方的底层生成器模型由一个变分自编码器组成，包括编码器、编码向量和解码器，其双方底层输入包括各方数据样本、各方掩码向量以及噪声样本，双方对输入数据做如下处理：

(1)A方掩码向量，，B方掩码向量，，A方和B方时，代表该位置的数据缺失；当时，代表数据完整；A方特征数据的编码向量为，A方剩余缺失样本数据向量：

即，其中表示向量间的元素级乘法；

B方特征数据的编码向量为，A方剩余缺失样本数据向量：

即，其中表示向量间的元素级乘法；

(2)A方和B方分别对噪声样本和噪声样本进行编码，并用编码后的结果对缺失样本进行填充，得到和，，；

S43：A、B、C三方的判别模型由三个全连接网络构成，A、B两方各自对己方生成模型的生成数据进行处理以满足判别器的输入；

S44：A方将生成数据、缺失数据与掩码向量做元素级乘法得到判别模型的输入数据，用表示，即；B方将生成数据、缺失数据与掩码向量做元素级乘法得到判别模型的输入数据，用表示，即；

S45：A、B两方底层损失函数的包含三部分：生成样本中观测变量与生成变量间的误差损失，生成样本输入判别器中得到的判别损失以及变分自编码器的KL散度损失；其中A、B双方重建损失表示为：

双方判别损失表示为：

其中为A方对掩码矩阵进行随机采样得到的提示矩阵，为B方对掩码矩阵进行随机采样得到的提示矩阵；

KL散度损失表示为：

其中和表示为A、B双方方编码器产生的潜在向量的以为条件的数据分布；为标准正态分布；是编码器输出的潜在向量的均值；是编码器输出的潜在向量的方差；

分别得到A方、B方各自的总损失如下：

其中为超参数，为C方判别器损失，计算如下：

S46：将数据集和生成的属性列缺失值放入填补模型中进行训练，对参与方剩余属性列的缺失值进行生成填补，获得完整的多方联合对齐样本数据集。

进一步，所述步骤S46具体包括以下步骤：

S461：C方初始化顶部判别器网络，A、B双方初始化各自的底部生成器网络和底部判别器网络；

S462：A、B方固定生成器训练判别器，具体包括：A、B双方采用同一批次的掩码向量、真实数据和随机噪声；经过数据编码、噪声填充以及向量拼接操作后输入至双方生成器得到生成数据样本与；A、B双方分别对掩码矩阵进行随机采样得到各自的提示矩阵与；进行向量拼接操作后输入至双方判别器网络得到双方数据的特征表示为与，将与发送至C方后，C方进行向量拼接并输入至C方判别器网络得到生成数据的判别结果；A、B双方将真实数据输入至双方各自的判别器网络后得到与，将与发送至C方后，C方进行向量拼接并输入至C方判别器网络得到真实数据的判别结果；经过计算，通过Adam优化器更新C方判别器网络参数以及A、B双方判别器参数，计算公式如下：

其中代表期望值，和分别代表集合和集合服从、概率分布；

S463：A、B双方固定判别器训练生成器，具体包括：双方采样同一批次的真实数据、掩码向量和随机噪声；经过数据编码、噪声填充以及向量拼接等操作后输入至双方生成器得到生成数据样本与；A、B双方分别对掩码矩阵进行随机采样得到各自的提示矩阵与；进行向量拼接操作后输入至双方判别器网络得到生成数据的判别结果；A、B双方计算各自的均方误差损失通过同态加密发送至服务端进行损失聚合，得到总损失，其中为超参数。通过Adam优化器更新A、B双方生成器网络。

本发明的有益效果在于：本发明提出了一种基于纵向联邦学习的参与方对齐样本生成系统，针对纵向联邦学习场景下，某些参与方样本较多，而另一些参与方样本较少，从而导致对齐后的样本无法支持纵向联邦场景下机器学习模型基于大量样本训练的需要的问题，结合纵向联邦学习、关联规则分析以及生成填补网络，对样本较少的参与方进行数据生成，以构建高质量的多方联合样本数据集，满足和提升纵向联邦场景下机器学习模型的训练效果。

本发明的其他优点、目标和特征在某种程度上将在随后的说明书中进行阐述，并且在某种程度上，基于对下文的考察研究对本领域技术人员而言将是显而易见的，或者可以从本发明的实践中得到教导。本发明的目标和其他优点可以通过下面的说明书来实现和获得。

### 附图说明

为了使本发明的目的、技术方案和优点更加清楚，下面将结合附图对本发明作优选的详细描述，其中：

图1为基于纵向联邦学习的参与方对齐样本生成方法的流程示意图。

图2为纵向联邦学习多方数据样本对齐示意图。

图3为A方第m个属性与B方第n个属性之间的对应关系示例图。

图4为B方基于变分自编码器的生成填补网络示意图。

### 具体实施方式

以下通过特定的具体实例说明本发明的实施方式，本领域技术人员可由本说明书所揭露的内容轻易地了解本发明的其他优点与功效。本发明还可以通过另外不同的具体实施方式加以实施或应用，本说明书中的各项细节也可以基于不同观点与应用，在没有背离本发明的精神下进行各种修饰或改变。需要说明的是，以下实施例中所提供的图示仅以示意方式说明本发明的基本构想，在不冲突的情况下，以下实施例及实施例中的特征可以相互组合。

需要说明的是，以下实施例中所提供的图示仅以示意方式说明本发明的基本构想，遂图式中仅显示与本发明中有关的组件而非按照实际实施时的组件数目、形状及尺寸绘制，其实际实施时各组件的型态、数量及比例可为一种随意的改变，且其组件布局型态也可能更为复杂。

在下文描述中，探讨了大量细节，以提供对本发明实施例的更透彻的解释，然而，对本领域技术人员来说，可以在没有这些具体细节的情况下实施本发明的实施例是显而易见的，在其他实施例中，以方框图的形式而不是以细节的形式来示出公知的结构和设备，以避免使本发明的实施例难以理解。

在本实施例中，以两个参与方A和B为例，各参与方包括拥有大量样本数据的A方和少量样本数据的B方，以及A方和B方均信任的协调方C方；请参阅图1～图4，为一种基于纵向联邦学习的参与方对齐样本生成系统，包括：

多方属性相关性矩阵构建模块：A、B两方构建己方样本数据集，并进行数据预处理。联合多个参与方进行样本对齐，构建对齐样本数据集。利用对齐样本数据集计算各参与方之间每个属性的相关性，构建多方属性相关性矩阵。具体包括以下步骤：

S11：A方和B方各自构建己方样本数据集进行数据预处理，包括：冗余数据处理、缺失值处理、异常值处理以及数据标准化处理，得到A方数据集，共有个样本数据；B方数据集，共有个样本数据；数据预处理具体操作如下：

所述冗余数据处理具体包括：通过数据记录中的某些字段（如账户号、时间等）判断数据是否重复，对于重复数据，只保留其中一个，将其他重复数据从数据集中删除，同时保留原始数据的备份以便需要时进行回溯和对比分析；

所述缺失值处理具体包括：对数据中的每个特征进行缺失率统计，对于缺失特征数目大于总体样本特征规模一半的数据视为无效数据，进行剔除；其余数据根据数据分布，采用特定方法进行填充，其中包括众数填充、均值填充、中位数填充和插值法填充等；

所述异常值处理具体包括：根据实际情况，使用标准差、四分位数范围等方法确认异常值的阈值。利用箱线图、散点图、直方图或密度图等工具对数据进行可视化分析，发现超出阈值的数据。选择用中位数或均值代替异常值的方法对异常值进行处理，处理完之后再次检查数据集，确保异常值已被处理且数据集的基本特征没有发生重大变化；

所述数据标准化处理具体包括：将已有特征按照数据类型分为连续型特征和离散型特征，对连续型特征采用最大-最小标准化进行处理，对离散型特征采用独热编码进行处理。

S12：对A方和B方数据集和按照其样本ID进行样本加密对齐，得到个对齐样本，剩余的-个样本为未对齐样本；A方的对齐数据集为，未对齐数据集为；和共同构成了A方数据集，表示A方数据集第*i*个样本，，，表示A方第*i*个样本的第个属性，其中a；A方参与对齐数据集样本个数，未对齐数据集样本个数-；B方的对齐数据集为，表示B方数据集第*i*个样本，，，表示B方第个样本的第个属性，其中，B方参与对齐数据集样本个数；

可选的，可以使用基于RSA算法和散列函数来进行样本的加密对齐。

S13：利用对齐数据集和，从中获取A方所有样本的第*m*个属性的集合，表示A方第*i*个样本的第个属性，；从中获取B方所有样本第*n*个属性的集合，表示B方第*i*个样本的第*n*个属性，；根据斯皮尔曼相关性系数计算A方第*m*个属性集合和B方第*n*个属性集合之间的属性相关性；A方第*m*个属性和B方第*n*个属性相关性系数的具体计算流程如下：

S131：协调方C方生成密钥对，所述密钥对包括公钥和私钥，利用非对称加密对公钥进行加密，并将加密后的公钥发送至A方和B方；

S132：A方获得从C方发送过来的加密公钥，进行解密操作，获得解密后的公钥。针对A方第*m*个属性，循环遍历A方第*m*个属性集合中所有样本的第*m*个属性值，其中，利用解密后的公钥对每个属性值进行加密，并且将加密后得每个属性值发送给C方；

S133：B方获得从C方发送过来的加密公钥，进行解密操作，获得解密后的公钥。针对B方第*n*个属性集合，循环遍历的每个属性值，其中，利用解密后的公钥对每个属性值进行加密，并且将加密后得每个属性值发送给C方；

S134：协调方C方采用私钥对加密后的A方第*m*个属性集合所对应的每个属性值和B方第*n*个属性所对应的每个属性值进行解密，得到解密后的A方第*m*个属性集合所对应的每个属性值和B方第*n*个属性集合所对应的每个属性值；

S135：对A方第*m*个属性所对应的每个属性值进行从小到大排序，对B方第*n*个属性集合所对应的每个属性值进行从小到大排序，依次计算与之间的差值，其中；

可选的，排序算法采用冒泡排序。

S136：根据差值进行斯皮尔曼相关性系数计算，得到相关性系数，计算公式如下：

其中为总的样本数，即。

S14：迭代步骤S13，计算A方所有属性与B方所有属性之间的相关性系数，其中，得到多方属性相关性矩阵M：

属性对对应关系建立模块：根据多方属性相关性矩阵M，找出其中最大的相关性系数，假设该相关性系数分别由A方第*m*个属性和B方第*n*个属性计算得到，则认为A方属性和B方属性具有强相关性，可以组成一组强相关性属性对)，同时将重新设为0；根据该属性对)，循环遍历和中所有属性列取值和，其中，建立属性中各取值与属性的取值之间的对应关系R；具体包括以下步骤：

S21：循环遍历相关性矩阵M中的每一个值，找出当前矩阵中最大的相关性系数，根据找出具有强相关性的一组属性对)，同时重新将的值设为0；

S22：根据属性对)，循环遍历和中所有属性列取值和，建立属性中各取值与属性的取值之间的对应关系R。假设属性总共包含3个属性取值：*m*1、*m*2、*m*3，属性总共包含4个属性取值：*n*1*、n*2*、n*3*、n*4，建立与属性取值之间的对应关系表R的具体流程如下：

S221：协调方C方初始化一个空的对应关系表，用来存储属性中的每个取值与属性的取值之间的对应关系，同时C方生成密钥对，所述密钥对包括公钥和私钥，利用非对称加密对公钥进行加密，并将加密后的公钥发送至A方和B方；

S222：A方和B方各自循环遍历属性集合和，并且同时将对应的属性值和通过加密的公钥进行加密发送给C方；

S223：C方记录当分别为*m*1、*m*2、*m*3时，对应的属性取值。假设*m*1对应的的属性取值有，*m*2对应的的属性取值有，*m*3对应的的属性取值有*n*2*、n*3，则各属性取值的对应关系记为；则C方将和各取值之间的对应关系表R表示为：

缺失值生成模块：根据对应关系R，建立和两个属性列中各取值的关联规则；循环遍历A方未对齐数据集中第*m*个属性中的每个样本值，其中，利用关联规则，为A方未对齐数据集中第*m*个属性中的每个样本值生成对应的B方缺失数据集中第*n*个属性中的样本值；具体包括以下步骤：

S31：利用Apriori算法通过对应关系R，建立和两个属性列中的关联规则。具体流程如下：

S311：A方遍历属性，记录属性取值为的样本个数，并将加密发送给C方；B方遍历属性，记录对应的属性取值分别为与的样本个数和，并将和加密发送给C方；

S312：C方计算的样本个数占总样本个数的比率，的样本个数占总样本个数的比率，计算公式如下：

S313：同理，对于剩余属性取值*m*2、*m*3，按照所述步骤S311与S312计算其对应关系中各对应属性取值所占比率。建立属性与的关联规则rule，表示如下：

，

其中表示对应的的样本个数占总样本个数的比率，

表示对应的的样本个数占总样本个数的比率，

表示对应的的样本个数占总样本个数的比率，

表示对应的的样本个数占总样本个数的比率，

表示对应的的样本个数占总样本个数的比率，

表示对应的的样本个数占总样本个数的比率，

表示对应的的样本个数占总样本个数的比率。

S32：循环遍历A方未对齐数据集中第*m*个属性中的每个样本值，利用关联规则，为A方未对齐数据集中第*m*个属性中的每个样本值生成对应的B方缺失数据集中第*n*个属性中的样本值。

循环迭代步骤S21~S32 H轮，每轮循环找出当前相关性最强一组属性对)建立该组属性对各属性取值之间的对应关系，利用每轮的对应关系建立当前属性对之间的关联规则；根据属性列之间的关联规则，对B方缺失数据集中与A方属性具有相关性大于的属性列进行生成，其中为人为给定的阈值，可选的，；循环迭代轮数H等于A、B两方相关性大于的属性列对数；

生成对抗填补模块：构建面向纵向联邦的基于变分自编码器(VAE)的生成对抗填补模型。将和由循环迭代步骤S21~S32 H轮生成的属性列放入填补模型中进行训练，对参与方剩余属性列的缺失值进行生成填补，获得完整的多方联合对齐样本数据集；具体包括以下步骤：

S41：A方和B方各自构建己方的底层子模型，其中包括了各方的生成器模型和判别器模型；C方构建顶层判别器模型，用于对A、B两方的数样本进行联合判定；

S42：A、B双方的底层生成器模型由一个变分自编码器组成，包括了编码器、编码向量和解码器，其双方底层输入包括各方数据样本、各方掩码向量以及噪声样本，双方对输入数据做如下处理：

(1)A方掩码向量，，B方掩码向量，，A方和B方时，代表该位置的数据缺失；当时，代表数据完整；A方特征数据的编码向量为，A方剩余缺失样本数据向量：

即，其中表示向量间的元素级乘法；

B方特征数据的编码向量为，A方剩余缺失样本数据向量：

即，其中表示向量间的元素级乘法；

(2)A方和B方分别对噪声样本和噪声样本进行编码，并用编码后的结果对缺失样本进行填充，得到和，，；

S43：A、B、C三方的判别模型由三个全连接网络构成，A、B两方各自对己方生成模型的生成数据进行处理以满足判别器的输入；

S44：A方将生成数据、缺失数据与掩码向量做元素级乘法得到判别模型的输入数据，用表示，即；B方将生成数据、缺失数据与掩码向量做元素级乘法得到判别模型的输入数据，用表示，即；

S45：A、B两方底层损失函数的包含三部分：生成样本中观测变量与生成变量间的误差损失，生成样本输入判别器中得到的判别损失以及变分自编码器的KL散度损失。其中A、B双方重建损失表示为：

双方判别损失表示为：

其中为A方对掩码矩阵进行随机采样得到的提示矩阵，为B方对掩码矩阵进行随机采样得到的提示矩阵；

KL散度损失表示为：

其中和表示为A、B双方方编码器产生的潜在向量的以为条件的数据分布；为标准正态分布；是编码器输出的潜在向量的均值；是编码器输出的潜在向量的方差。

此时分别得到A方、B方各自的总损失如下：

其中为超参数，取值为10，为C方判别器损失，计算如下：

S46：将数据集和由循环迭代步骤S21~S32 H轮生成的属性列放入填补模型中进行训练，对参与方剩余属性列的缺失值进行生成填补，获得完整的多方联合对齐样本数据集。具体包括：

S461：C方初始化顶部判别器网络，A、B双方初始化各自的底部生成器网络和底部判别器网络；

S462：A、B方固定生成器训练判别器。双方采用同一批次的掩码向量、真实数据和随机噪声。经过数据编码、噪声填充以及向量拼接等操作后输入至双方生成器得到生成数据样本与。A、B双方分别对掩码矩阵进行随机采样得到各自的提示矩阵与。进行向量拼接操作后输入至双方判别器网络得到双方数据的特征表示为与，将与发送至C方后，C方进行向量拼接并输入至C方判别器网络得到生成数据的判别结果。A、B双方将真实数据输入至双方各自的判别器网络后得到与，将与发送至C方后，C方进行向量拼接并输入至C方判别器网络得到真实数据的判别结果。经过计算，通过Adam优化器更新C方判别器网络参数以及A、B双方判别器参数，计算公式如下：

其中代表期望值，和分别代表集合和集合服从、概率分布；

S463：A、B双方固定判别器训练生成器。双方采样同一批次的真实数据、掩码向量和随机噪声。经过数据编码、噪声填充以及向量拼接等操作后输入至双方生成器得到生成数据样本与。A、B双方分别对掩码矩阵进行随机采样得到各自的提示矩阵与。进行向量拼接操作后输入至双方判别器网络得到生成数据的判别结果。A、B双方计算各自的均方误差损失通过同态加密发送至服务端进行损失聚合，得到总损失，其中为超参数，可选的10。通过Adam优化器更新A、B双方生成器网络。

在上述实施例中，说明书对“本实施例”、的提及表示结合实施例说明的特定特征、结构或特性包括在至少一些实施例中，但不必是全部实施例。“本实施例”的多次出现不一定全部都指代相同的实施例。

在上述实施例中，尽管已经结合了本发明的具体实施例对本发明进行了描述，但是根据前面的描述，这些实施例的很多替换、修改和变形对本领域普通技术人员来说将是显而易见的。例如，其他存储结构（例如，动态RAM（DRAM））可以使用所讨论的实施例。本发明的实施例旨在涵盖落入所附权利要求的宽泛范围之内的所有这样的替换、修改和变型。

本实施例还提供一种计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现本实施例中的任一项方法。

本实施例还提供一种电子终端，包括：处理器及存储器；

所述存储器用于存储计算机程序，所述处理器用于执行所述存储器存储的计算机程序，以使所述终端执行本实施例中任一项方法。

本实施例中的计算机可读存储介质，本领域普通技术人员可以理解：实现上述各方法实施例的全部或部分步骤可以通过计算机程序相关的硬件来完成。前述的计算机程序可以存储于一计算机可读存储介质中。该程序在执行时，执行包括上述各方法实施例的步骤；而前述的存储介质包括：ROM、RAM、磁碟或者光盘等各种可以存储程序代码的介质。

本实施例提供的电子终端，包括处理器、存储器、收发器和通信接口，存储器和通信接口与处理器和收发器连接并完成相互间的通信，存储器用于存储计算机程序，通信接口用于进行通信，处理器和收发器用于运行计算机程序，使电子终端执行如上方法的各个步骤。

在本实施例中，存储器可能包含随机存取存储器(Random Access Memory，简称RAM)，也可能还包括非易失性存储器(non-volatile memory)，例如至少一个磁盘存储器。

上述的处理器可以是通用处理器，包括中央处理器(Central Processing Unit，简称CPU)、网络处理器(Network Processor，简称NP)等；还可以是数字信号处理器(Digital Signal Processing，简称DSP)、专用集成电路(Application Specific Integrated Circuit，简称ASIC)、现场可编程门阵列(Field－Programmable Gate Array，简称FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件。

本发明可用于众多通用或专用的计算系统环境或配置中。例如：个人计算机、服务器计算机、手持设备或便携式设备、平板型设备、多处理器系统、基于微处理器的系统、置顶盒、可编程的消费电子设备、网络PC、小型计算机、大型计算机、包括以上任何系统或设备的分布式计算环境等等。

本发明可以在由计算机执行的计算机可执行指令的一般上下文中描述，例如程序模块。一般地，程序模块包括执行特定任务或实现特定抽象数据类型的例程、程序、对象、组件、数据结构等等。也可以在分布式计算环境中实践本发明，在这些分布式计算环境中，由通过通信网络而被连接的远程处理设备来执行任务。在分布式计算环境中，程序模块可以位于包括存储设备在内的本地和远程计算机存储介质中。

最后说明的是，以上实施例仅用以说明本发明的技术方案而非限制，尽管参照较佳实施例对本发明进行了详细说明，本领域的普通技术人员应当理解，可以对本发明的技术方案进行修改或者等同替换，而不脱离本技术方案的宗旨和范围，其均应涵盖在本发明的权利要求范围当中。

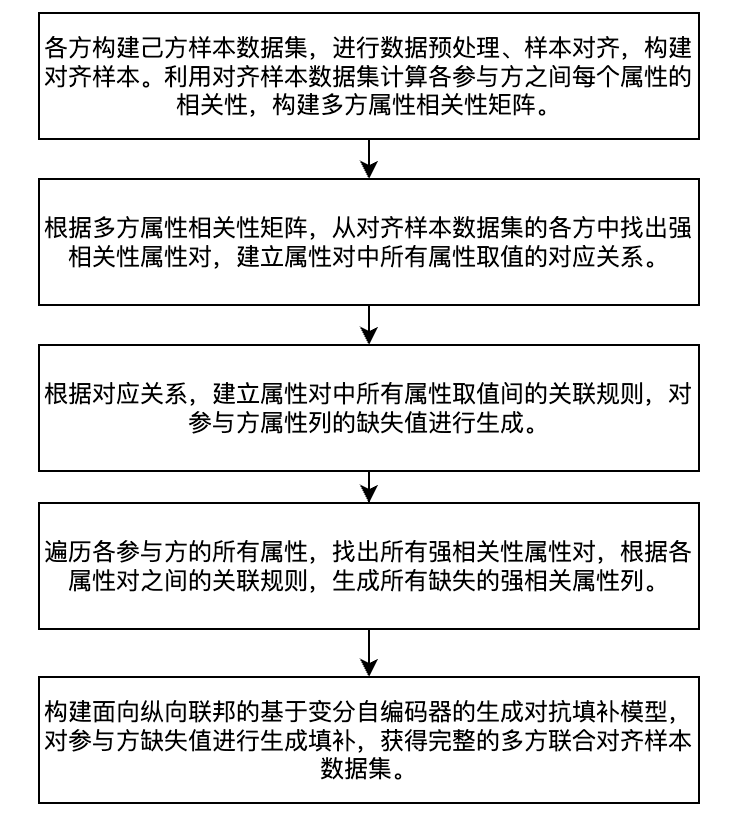


图1

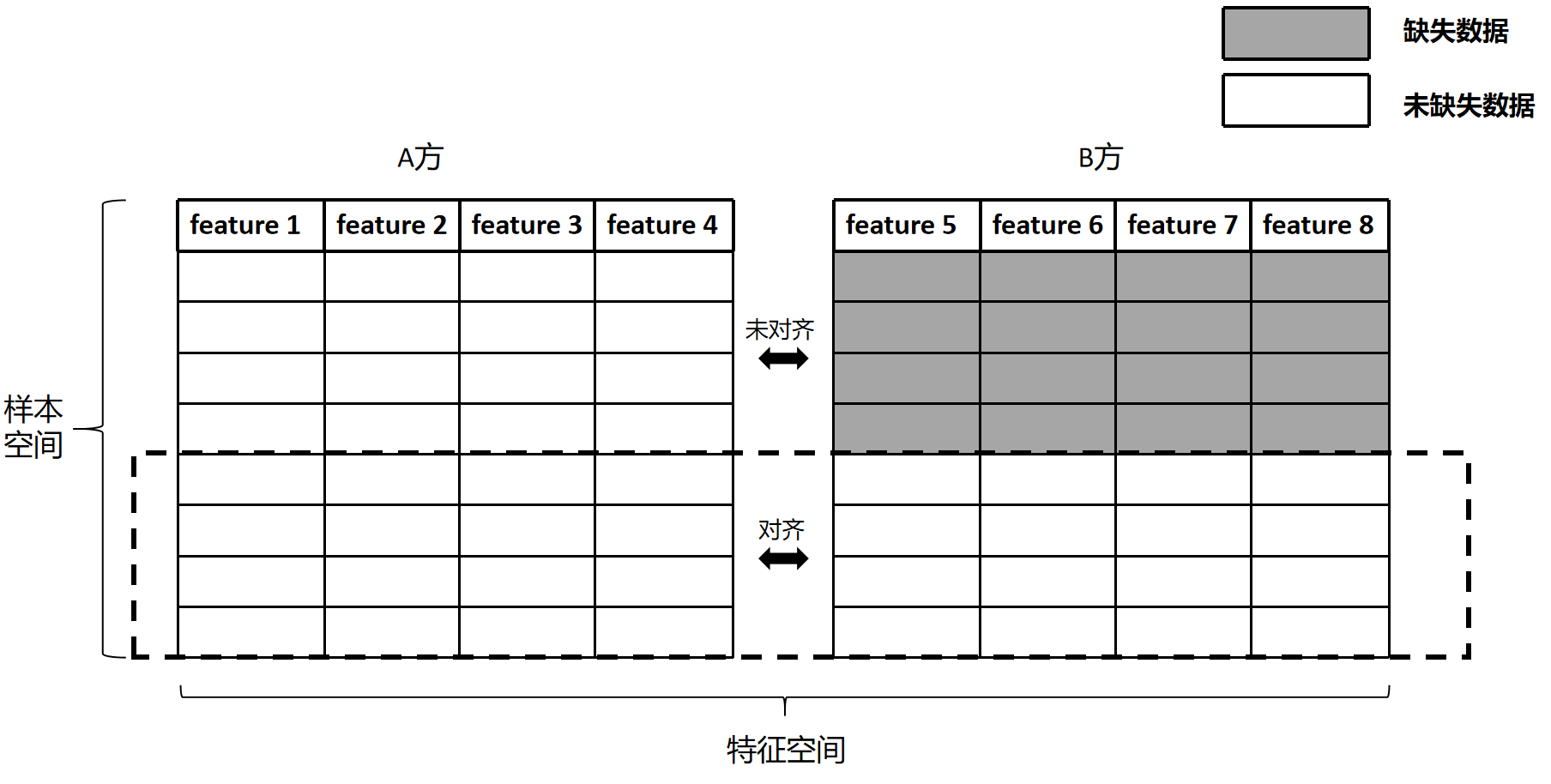


图2

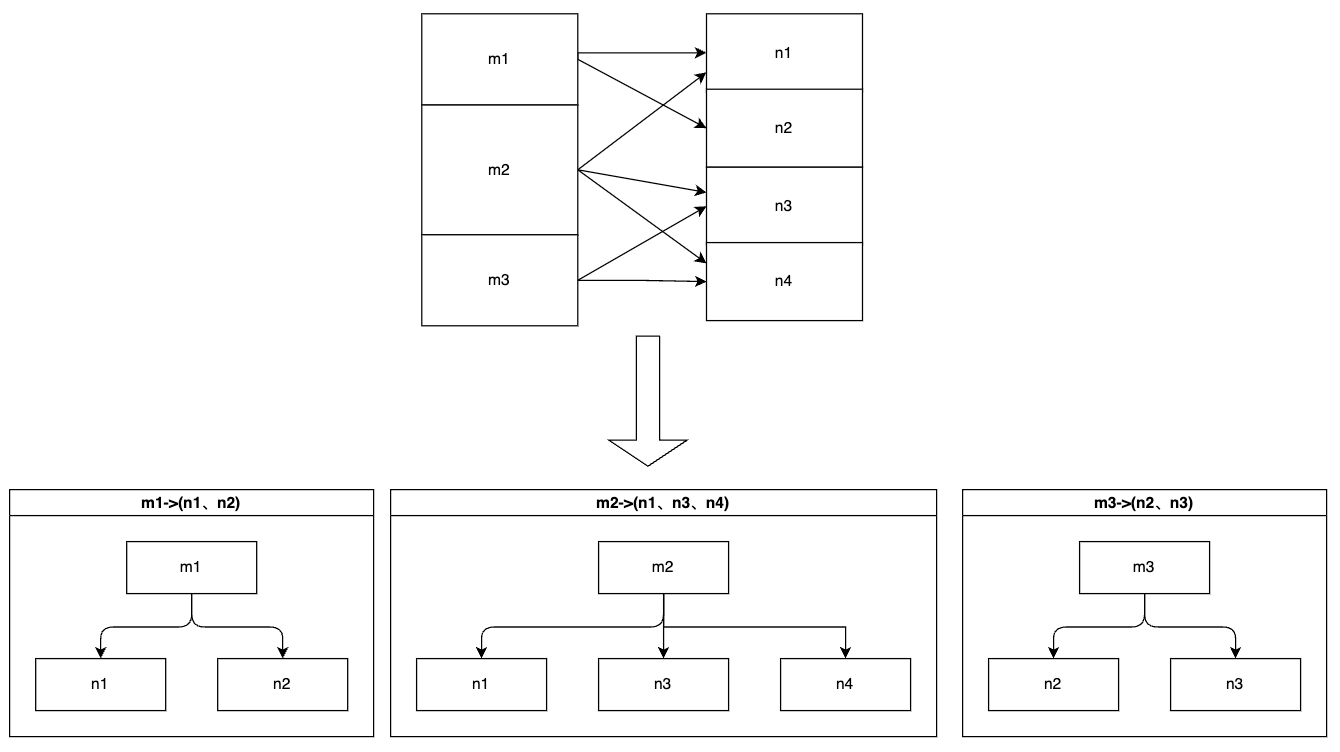


图3

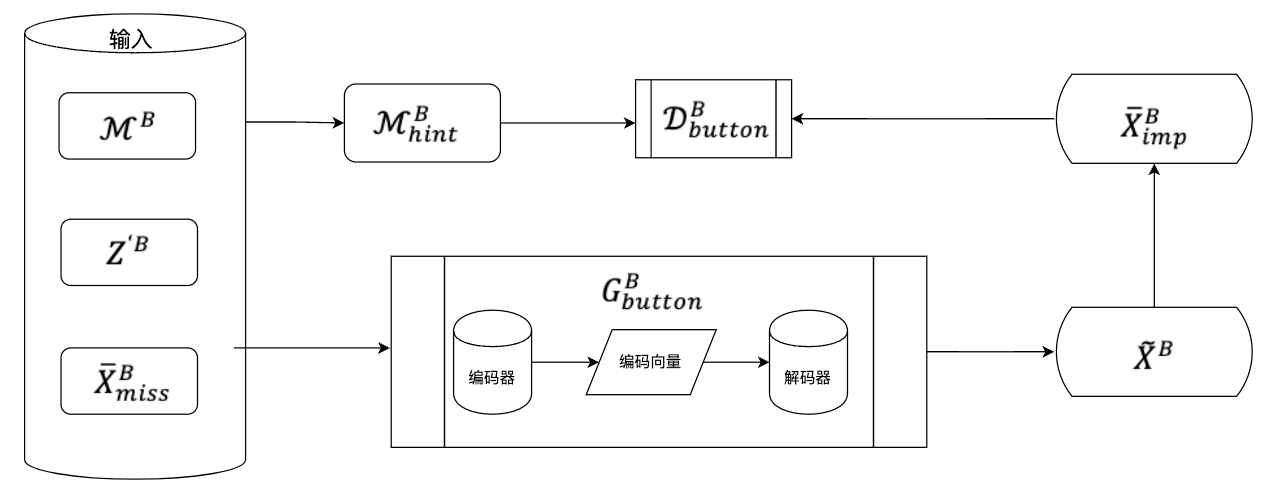


图4