**专利申请著录项目表**

接案号：TF230411114

(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 | 一种多方半监督学习的金融产品潜在用户推荐方法 | | | | | | |
| 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 |  | | | | | | |
| ④第一发明人国籍 | | 中国 | | 身份证号 |  | | |
| 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。

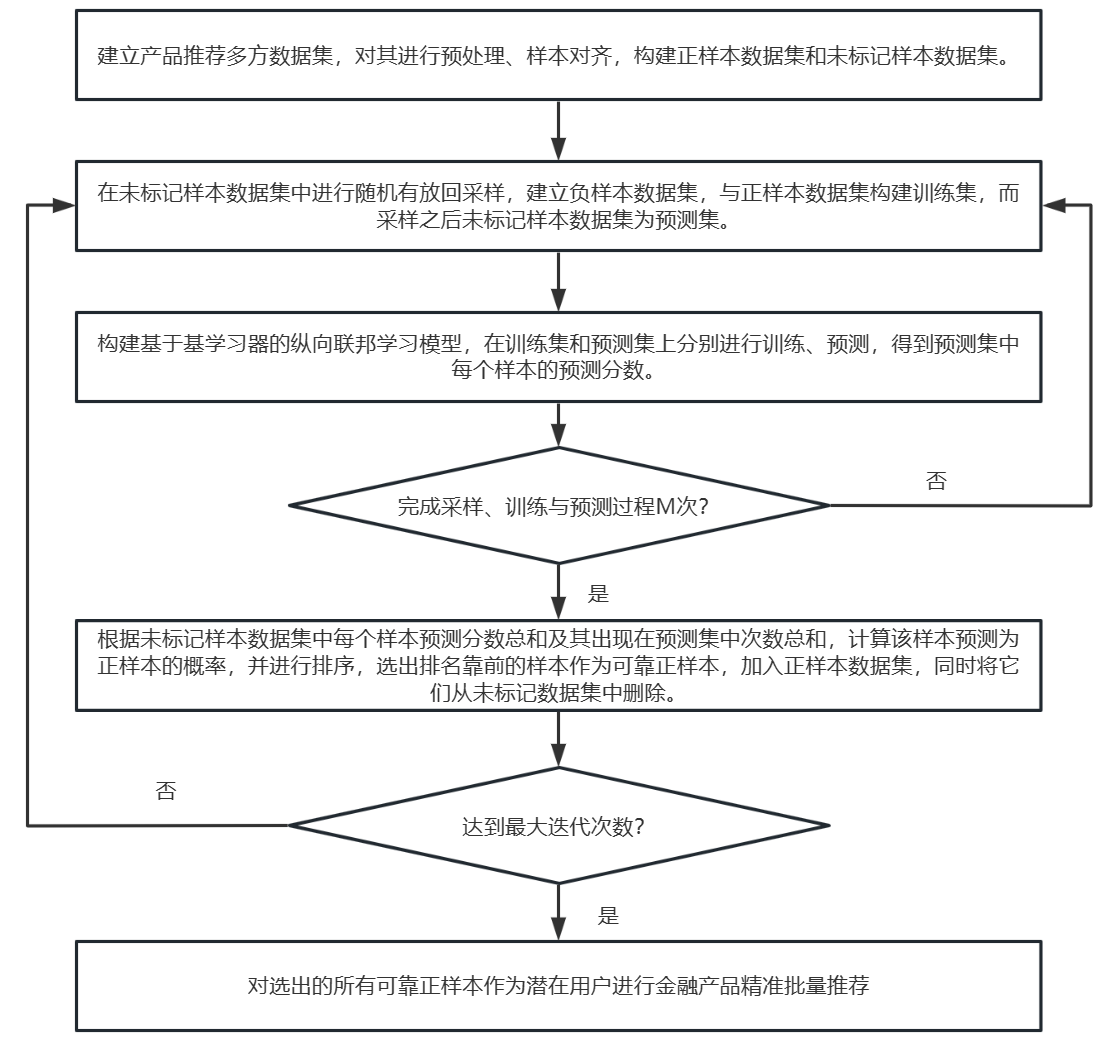
五、本表第⑦栏，实质审查是指审查员对发明专利申请是否符合授权条件（包括新颖性、创造性、实用性、公开充分、单一性问题等）进行审查。申请人可以在自申请日（有优先权的，指优先权日）起三年内提出实质审查请求来启动实质审查程序。如果在提交专利申请的同时提交实质审查请求，则该发明专利申请在公布后立即进入实质审查阶段。为了加快发明专利申请的审查进程，在申请人无特别要求的情况下，本公司默认勾选此栏。如申请人不想在提交专利申请的同时提交实质审查请求，请去除勾选并及时通知代理人。

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。

本发明涉及一种多方半监督学习的金融产品潜在用户推荐方法，属于大数据推荐领域，针对金融产品提供方只拥有正样本的己方数据而无法进行推荐的问题，在保护多方数据安全隐私的情况下，联合多方未标记数据进行多次随机采样，构建正负样本均衡的二分类数据集，训练基于基学习器的纵向联邦学习模型，根据其预测结果，从未标记样本数据中选出可靠正样本，多次迭代数据集重构采样、模型训练预测过程，选出批量可靠正样本。本专利方法有效地解决了只有少量正样本和大量未标记样本的批量推荐，提高了推荐的可靠性，实现了金融产品潜在用户的精准与批量推荐。



1. 一种多方半监督学习的金融产品潜在用户推荐方法，其特征在于：包括以下步骤：

S1：建立包含购买金融产品用户信息和其他多方用户信息的金融产品潜在用户推荐多方数据集，并对所述金融产品潜在用户推荐多方数据集进行预处理、样本对齐，构建正样本数据集和未标记样本数据集；

S2：在未标记样本数据集中进行随机有放回采样，建立负样本数据集；利用所述负样本数据集与正样本数据集构建训练集，利用未标记样本数据集中未被采样到的样本构建预测集；构建基于基学习器的纵向联邦学习模型，在训练集上进行训练，在预测集上进行预测，得到预测集中每个样本的预测分数；

S3：多次重复步骤S2的采样、训练与预测过程；根据所述未标记样本数据集中每个样本预测分数总和及其出现在预测集中的次数，计算该样本预测为正样本的概率；按照正样本的概率从大到小对所述未标记样本数据集中所有样本进行排序，根据先验知识选出排名靠前的样本作为可靠正样本，并将它们加入到所述正样本数据集中，同时将它们从所述未标记样本数据集中删除；

S4：重复步骤S2-S3，直到达到预设的最大迭代次数；由此，从所述未标记样本数据集中选出的所有可靠正样本即可作为潜在用户，向金融产品提供方进行精准批量推荐。

2. 根据权利要求1所述的多方半监督学习的金融产品潜在用户推荐方法，其特征在于：步骤S1所述的金融产品潜在用户推荐多方数据集包含拥有已购买金融产品用户信息的A方，以及除A方以外的其他多方数据源B方、C方；步骤S1具体包括以下步骤：

S11：对金融产品潜在用户推荐多方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、数据标准化处理以及标签数据处理，得到A方数据集，表示A方第i个样本特征向量，向量维度为，表示对应的标签，其中a；B方数据集，表示B方第i个样本特征向量，向量维度为，表示对应的标签，其中b；C方数据集，表示C方第i个样本特征向量，向量维度为，表示对应的标签，其中c；

S12：对B方和C方数据集和按照其样本ID进行样本加密对齐，保留B方和C方对齐的样本数据，丢弃未对齐样本数据，得到n个样本；对齐后的B方数据集为，C方数据集为，表示对应的标签，表示对应的标签，其中n；

S13：对数据集和按照其样本ID进行样本加密对齐，将三方对齐的样本作为正样本，组成正样本数据集，其中，分别表示A方、B方、C方对齐后的第i个样本特征向量，为正样本标签，表示正样本个数，；三方未对齐的样本作为未标记样本，组成未标记样本数据集，其中，分别表示B方、C方对齐后的第i个样本特征向量，为未标记样本标签，表示未标记样本个数，。

3. 根据权利要求2所述的多方半监督学习的金融产品潜在用户推荐方法，其特征在于：在步骤S11中，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；

所述异常值处理具体包括：根据实际情况，确认异常值的阈值；对数据进行可视化分析，发现超出阈值的数据；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；

所述标签数据处理具体包括：A方数据集添加一列标签列，将其标签数据设为“1”，代表正样本数据；B方、C方数据集分别添加一列标签列，将其标签数据设为“0”，代表未标记数据。

4. 根据权利要求3所述的多方半监督学习的金融产品潜在用户推荐方法，其特征在于：所述步骤S2具体包括：建立预测未标记样本候选推荐过程，循环执行该过程M轮，第m轮预测过程如下：从U中随机有放回地抽取个样本，表示P的样本个数，将这个未标记样本组成负样本数据集；由P和组成训练集，U中未被抽取样本组成预测集；构建以梯度提升树（Gradient Boosting Decision Tree，GBDT）作为基学习器的纵向联邦模型，在上进行训练；将输入到训练好的模型中进行预测，得到中每个样本的分数，作为该样本的预测分数；具体包括以下步骤：

S21：使用bootstrap采样法从U中随机有放回地抽取个样本，将这个未标记样本组成第m轮负样本数据集，其中，分别是第m轮的第i个负样本中B方和C方的特征向量，表示第m轮第i个负样本的标签，，||表示第m轮负样本数据集的样本个数；由正样本数据集P和第m轮的负样本数据集组成对应的第m轮训练集，其中，分别表示第m轮的第i个样本中B方和C方的特征向量，表示第m轮第i个样本的标签，；未标记样本数据集U中未被抽取样本组成对应的第m轮预测集，其中表示第m轮的第j个样本的特征向量，；

S22：使用GBDT算法作为基学习器，构建一个纵向联邦模型，通过T颗决策树的集成对进行训练，对输入数据中每个样本来预测其第m轮输出 ，其中T表示第m轮的决策树的总数，表示第m轮第棵树的预测结果，表示第m轮第i个样本特征向量，；根据预测输出和真实标签之间的损失函数的梯度值，建立一阶导数和二阶导数梯度直方图，联合双方所有特征的梯度直方图找到当前节点的全局最佳分割从而构造出最优决策树；

S23：使用训练好的第m轮GBDT决策模型，对第m轮测试集进行预测，获得第m轮测试集中每个样本的预测分数，遍历U中所有样本，在U中找到与的ID相同的样本，并把的预测分数赋值给，作为样本的第m轮预测分数，用表示，其中为未标记数据集U中的样本编号；如果U中样本第m轮中未存在于中，则该样本对应的第m轮预测分数等于0。

5. 根据权利要求4所述的多方半监督学习的金融产品潜在用户推荐方法，其特征在于：在步骤S22中，首先初始化第m轮每个样本的预测结果，为一个随机值，接着第m轮第t颗树的联邦训练具体流程如下：

S221：从B方开始，首先对B方的每个样本计算第m轮模型损失函数的一阶梯度和二阶梯度，其中，表示第m轮由前棵树聚合的对样本的预测结果，表示第m轮第i个样本的真实标签，是损失函数；使用加法同态加密对和进行加密，得到和，B方将和发送给C方；

S222：对于C方，根据自己的特征数据建立第m轮梯度直方图，并且将加密后的梯度直方图发送给B方，

S223：B方对来自C方的第m轮加密梯度直方图进行解密，根据分裂增益计算公式，枚举每个特征梯度直方图进行最优解计算，找到全局最优分裂点，并将分裂信息返回给C方进行解析；

S224：C方根据从B方发送的特征编号和阈值编号确定特征的阈值，并对当前的样本空间进行划分；然后C方在本地建立一个查找表，记录选中特征的阈值，形成记录[记录编号，特征，阈值]，并将记录编号和划分后左侧的样本空间（）返回给B方；

S225：B方根据收到的[记录编号，]对当前节点进行划分，并将当前节点与[参与方，记录编号]关联；B方将当前节点的划分信息与C方同步，并进入对下一节点的分割；

S226：迭代步骤S222~S225，直至达到训练停止条件或树的最大深度。

6. 根据权利要求5所述的多方半监督学习的金融产品潜在用户推荐方法，其特征在于：步骤S222具体包括以下步骤：

S2221：对于当前第m轮样本中所有C方特征，根据每个特征的特征值对所有样本进行排序，然后将排序后的样本进行分桶划分到*q*个类别中，得到各个类别的相应特征阈值，其中表示特征编号，表示第m轮编号为的特征第q个类别的阈值；

S2222：根据从B方得到的第m轮的和，C方进行加密梯度信息聚合，构建第m轮加密梯度直方图即

其中，*v*=1,2,...,*q*；

S2223：C方将计算得到的第m轮的和发送给B方。

7. 根据权利要求5所述的多方半监督学习的金融产品潜在用户推荐方法，其特征在于：步骤S223具体包括以下步骤：

S2231：B方对当前节点空间的所有样本的一阶梯度和二阶梯度进行聚合，执行，其中表示当前节点的所有样本；

S2232：B方对从C方得到的第m轮和进行解密，得到第m轮解密值和，依次对C方的所有特征的所有类别进行如下计算得到、、和：

其中表示分割后左子节点的样本空间，表示分割后右子节点的样本空间，  
表示第m轮中左子节点样本空间的所有样本的损失函数一阶梯度之和，表示第m轮中左子节点样本空间的所有样本的损失函数二阶梯度之和，  
表示第m轮中右子节点样本空间的所有样本的损失函数一阶梯度之和，表示第m轮中右子节点样本空间的所有样本的损失函数二阶梯度之和；

S2233：计算第m轮当前节点的最佳分割值：

其中是超参数；

S2234：对于样本的每个特征的所有阈值均能得到一个值，选出其中最大的值，确定该特征阈值为第m轮的全局最佳分割，全局最佳分割用[参与方，特征编号()，阈值编号()]来表示，并将特征编号()和阈值编号()返回给C方。

8. 根据权利要求4所述的多方半监督学习的金融产品潜在用户推荐方法，其特征在于：步骤S23具体包括以下步骤：

S231：B方查询与当前节点相关联的[参与方，记录编号]记录；基于该记录，B方向C方发送待标注样本编号和记录编号，并且询问下一步的树搜索方向，即向左子节点或右子节点；

S232：C方接收到待标注样本编号和记录编号后，将待标注样本中相应特征的值与本地查找表中的记录[记录编号，特征，阈值]中的阈值进行比较，得出下一步的树搜索方向；然后，该C方将搜索决定发往B方；

S233：B方接收到C方传来的搜索决定，前往相应的子节点；

S234：迭代步骤S231~S233，直至到达一个叶节点，得到对应的分类标签以及该标签的权值，从而得到与样本对应的U中样本的第m轮预测分数：

其中表示叶节点的样本空间，是超参数。

9. 根据权利要求4所述的多方半监督学习的金融产品潜在用户推荐方法，其特征在于：所述步骤S3中，由U中每个样本的M轮预测分数总和及其在M轮预测集中出现次数总和，计算样本预测为正样本的概率；按照正样本的概率从大到小对U中所有样本进行排序，根据先验知识选出排名靠前的样本作为可靠正样本，并将它们加入到正样本数据集P中，同时将它们从U中删除，其具体包括以下步骤：

S31：由U中每个样本的预测分数总和及其在M轮预测集中出现次数总和，计算样本预测为正样本的概率,计算公式如下：

其中为指示函数，表示如果U中的样本存在于第m轮的测试集中，则*I*=1，否则*I*=0；

S32：对U中所有样本使用其概率按降序排序，选出排名前的样本作为可靠正样本，并将它们加入到正样本数据集P中，同时将它们从U中删除，其中的值是根据先验知识设定的。

S4：重复步骤S2-S3，直到达到预设的最大迭代次数。在每次迭代中，为了推荐的可靠性和精准性，每次迭代只能选出少量可靠正样本，所以为了选出一定量的可靠正样本进行批量推荐，需要进行多次迭代。在每次迭代中，步骤S3每次会从未标记数据集U中选出部分可靠正样本加入到正样本数据集P中，同时将他们从未标记数据集U中删除。这种情况下，|P|的样本大小会随着迭代次数的增加而增大，|U|的样本大小会随着迭代次数的增加而减少，直至达到预设的最大迭代次数。由此，在多次迭代过程中从U中选出的所有可靠正样本即可作为金融产品潜在购买用户，向金融产品拥有方进行精准批量推荐。

**一种多方半监督学习的金融产品潜在用户推荐方法**

### 技术领域

本发明属于大数据推荐领域，涉及一种多方半监督学习的金融产品潜在用户推荐方法。

### 背景技术

随着社会经济的不断发展，越来越多的金融机构开始不断向具有不同特点的用户群体推出不同的金融产品。金融产品是指银行、证券、保险、公积金等金融机构提供的各种金融服务和产品，如贷款、信用卡、基金、股票、公积金缴存等。但由于不同用户对不同金融产品的需求不同，当一款特定的新金融产品发布后，如何将其快速精准地推荐给潜在用户存在巨大挑战。利用机器学习和大数据技术可以对已购买某金融机构的某金融产品的用户特点进行分析，并从尚未购买该金融产品人员中找到具有相同特点的潜在用户，向金融机构进行精准推荐，满足用户和金融机构的需求。但当金融机构只拥有少量已购买该金融产品的人员信息，未拥有尚未购买该金融产品人员信息时，金融机构需要从其他机构或渠道获取更多数据以辅助推荐。但由于各方数据隐私安全保护的要求，数据彼此孤立，形成数据孤岛，使得不同参与方想要将数据聚合在一起构建一个更高性能的机器学习模型变得十分困难。

### 发明内容

有鉴于此，本发明的目的在于提供一种联合多方半监督学习的金融产品潜在用户推荐方法。

为达到上述目的，本发明提供如下技术方案：

一种多方半监督学习的金融产品潜在用户推荐方法，包括以下步骤：

S1：建立包含购买金融产品用户信息和其他多方用户信息的金融产品潜在用户推荐多方数据集，并对所述金融产品潜在用户推荐多方数据集进行预处理、样本对齐，构建正样本数据集和未标记样本数据集；

S2：在未标记样本数据集中进行随机有放回采样，建立负样本数据集；利用所述负样本数据集与正样本数据集构建训练集，利用未标记样本数据集中未被采样到的样本构建预测集；构建基于基学习器的纵向联邦学习模型，在训练集上进行训练，在预测集上进行预测，得到预测集中每个样本的预测分数；

S3：多次重复步骤S2的采样、训练与预测过程；根据所述未标记样本数据集中每个样本预测分数总和及其出现在预测集中的次数，计算该样本预测为正样本的概率；按照正样本的概率从大到小对所述未标记样本数据集中所有样本进行排序，根据先验知识选出排名靠前的样本作为可靠正样本，并将它们加入到所述正样本数据集中，同时将它们从所述未标记样本数据集中删除；

S4：重复步骤S2-S3，直到达到预设的最大迭代次数；由此，从所述未标记样本数据集中选出的所有可靠正样本即可作为潜在用户，向金融产品提供方进行精准批量推荐。

进一步，步骤S1所述的金融产品潜在用户推荐多方数据集包含拥有已购买金融产品用户信息的A方，以及除A方以外的其他多方数据源B方、C方；步骤S1具体包括以下步骤：

S11：对金融产品潜在用户推荐多方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、数据标准化处理以及标签数据处理，得到A方数据集，表示A方第i个样本特征向量，向量维度为，表示对应的标签，其中a；B方数据集，表示B方第i个样本特征向量，向量维度为，表示对应的标签，其中b；C方数据集，表示C方第i个样本特征向量，向量维度为，表示对应的标签，其中c；

S12：对B方和C方数据集和按照其样本ID进行样本加密对齐，保留B方和C方对齐的样本数据，丢弃未对齐样本数据，得到n个样本；对齐后的B方数据集为，C方数据集为，表示对应的标签，表示对应的标签，其中n；

S13：对数据集和按照其样本ID进行样本加密对齐，将三方对齐的样本作为正样本，组成正样本数据集，其中，分别表示A方、B方、C方对齐后的第i个样本特征向量，为正样本标签，表示正样本个数，；三方未对齐的样本作为未标记样本，组成未标记样本数据集，其中，分别表示B方、C方对齐后的第i个样本特征向量，为未标记样本标签，表示未标记样本个数，。

进一步，在步骤S11中，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；

所述异常值处理具体包括：根据实际情况，确认异常值的阈值；对数据进行可视化分析，发现超出阈值的数据；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；

所述标签数据处理具体包括：A方数据集添加一列标签列，将其标签数据设为“1”，代表正样本数据；B方、C方数据集分别添加一列标签列，将其标签数据设为“0”，代表未标记数据。

进一步，所述步骤S2具体包括：建立预测未标记样本候选推荐过程，循环执行该过程M轮，第m轮预测过程如下：从U中随机有放回地抽取个样本，表示P的样本个数，将这个未标记样本组成负样本数据集；由P和组成训练集，U中未被抽取样本组成预测集；构建以梯度提升树（Gradient Boosting Decision Tree，GBDT）作为基学习器的纵向联邦模型，在上进行训练；将输入到训练好的模型中进行预测，得到中每个样本的分数，作为该样本的预测分数；具体包括以下步骤：

S21：使用bootstrap采样法从U中随机有放回地抽取个样本，将这个未标记样本组成第m轮负样本数据集，其中，分别是第m轮的第i个负样本中B方和C方的特征向量，表示第m轮第i个负样本的标签，，||表示第m轮负样本数据集的样本个数；由正样本数据集P和第m轮的负样本数据集组成对应的第m轮训练集，其中，分别表示第m轮的第i个样本中B方和C方的特征向量，表示第m轮第i个样本的标签，；未标记样本数据集U中未被抽取样本组成对应的第m轮预测集，其中表示第m轮的第j个样本的特征向量，；

S22：使用GBDT算法作为基学习器，构建一个纵向联邦模型，通过T颗决策树的集成对进行训练，对输入数据中每个样本来预测其第m轮输出 ，其中T表示第m轮的决策树的总数，表示第m轮第棵树的预测结果，表示第m轮第i个样本特征向量，；根据预测输出和真实标签之间的损失函数的梯度值，建立一阶导数和二阶导数梯度直方图，联合双方所有特征的梯度直方图找到当前节点的全局最佳分割从而构造出最优决策树；

S23：使用训练好的第m轮GBDT决策模型，对第m轮测试集进行预测，获得第m轮测试集中每个样本的预测分数，遍历U中所有样本，在U中找到与的ID相同的样本，并把的预测分数赋值给，作为样本的第m轮预测分数，用表示，其中为未标记数据集U中的样本编号；如果U中样本第m轮中未存在于中，则该样本对应的第m轮预测分数等于0。

进一步，在步骤S22中，首先初始化第m轮每个样本的预测结果，为一个随机值，接着第m轮第t颗树的联邦训练具体流程如下：

S221：从B方开始，首先对B方的每个样本计算第m轮模型损失函数的一阶梯度和二阶梯度，其中，表示第m轮由前棵树聚合的对样本的预测结果，表示第m轮第i个样本的真实标签，是损失函数；使用加法同态加密对和进行加密，得到和，B方将和发送给C方；

S222：对于C方，根据自己的特征数据建立第m轮梯度直方图，并且将加密后的梯度直方图发送给B方；

S223：B方对来自C方的第m轮加密梯度直方图进行解密，根据分裂增益计算公式，枚举每个特征梯度直方图进行最优解计算，找到全局最优分裂点，并将分裂信息返回给C方进行解析；

S224：C方根据从B方发送的特征编号和阈值编号确定特征的阈值，并对当前的样本空间进行划分；然后C方在本地建立一个查找表，记录选中特征的阈值，形成记录[记录编号，特征，阈值]，并将记录编号和划分后左侧的样本空间（）返回给B方；

S225：B方根据收到的[记录编号，]对当前节点进行划分，并将当前节点与[参与方，记录编号]关联；B方将当前节点的划分信息与C方同步，并进入对下一节点的分割；

S226：迭代步骤S222~S225，直至达到训练停止条件或树的最大深度。

进一步，步骤S222具体包括以下步骤：

S2221：对于当前第m轮样本中所有C方特征，根据每个特征的特征值对所有样本进行排序，然后将排序后的样本进行分桶划分到*q*个类别中，得到各个类别的相应特征阈值，其中表示特征编号，表示第m轮编号为的特征第q个类别的阈值；

S2222：根据从B方得到的第m轮的和，C方进行加密梯度信息聚合，构建第m轮加密梯度直方图即

其中，*v*=1,2,...,*q*；

S2223：C方将计算得到的第m轮的和发送给B方。

进一步，步骤S223具体包括以下步骤：

S2231：B方对当前节点空间的所有样本的一阶梯度和二阶梯度进行聚合，执行，其中表示当前节点的所有样本；

S2232：B方对从C方得到的第m轮和进行解密，得到第m轮解密值和，依次对C方的所有特征的所有类别进行如下计算得到、、和：

其中表示分割后左子节点的样本空间，表示分割后右子节点的样本空间，  
表示第m轮中左子节点样本空间的所有样本的损失函数一阶梯度之和，表示第m轮中左子节点样本空间的所有样本的损失函数二阶梯度之和，  
表示第m轮中右子节点样本空间的所有样本的损失函数一阶梯度之和，表示第m轮中右子节点样本空间的所有样本的损失函数二阶梯度之和；

S2233：计算第m轮当前节点的最佳分割值：

其中是超参数；

S2234：对于样本的每个特征的所有阈值均能得到一个值，选出其中最大的值，确定该特征阈值为第m轮的全局最佳分割，全局最佳分割用[参与方，特征编号()，阈值编号()]来表示，并将特征编号()和阈值编号()返回给C方。

进一步，步骤S23具体包括以下步骤：

S231：B方查询与当前节点相关联的[参与方，记录编号]记录；基于该记录，B方向C方发送待标注样本编号和记录编号，并且询问下一步的树搜索方向，即向左子节点或右子节点；

S232：C方接收到待标注样本编号和记录编号后，将待标注样本中相应特征的值与本地查找表中的记录[记录编号，特征，阈值]中的阈值进行比较，得出下一步的树搜索方向；然后，该C方将搜索决定发往B方；

S233：B方接收到C方传来的搜索决定，前往相应的子节点；

S234：迭代步骤S231~S233，直至到达一个叶节点，得到对应的分类标签以及该标签的权值，从而得到与样本对应的U中样本的第m轮预测分数：

其中表示叶节点的样本空间，是超参数。

进一步，所述步骤S3中，由U中每个样本的M轮预测分数总和及其在M轮预测集中出现次数总和，计算样本预测为正样本的概率；按照正样本的概率从大到小对U中所有样本进行排序，根据先验知识选出排名靠前的样本作为可靠正样本，并将它们加入到正样本数据集P中，同时将它们从U中删除，其具体包括以下步骤：

S31：由U中每个样本的预测分数总和及其在M轮预测集中出现次数总和，计算样本预测为正样本的概率,计算公式如下：

其中为指示函数，表示如果U中的样本存在于第m轮的测试集中，则*I*=1，否则*I*=0；

S32：对U中所有样本使用其概率按降序排序，选出排名前的样本作为可靠正样本，并将它们加入到正样本数据集P中，同时将它们从U中删除，其中的值是根据先验知识设定的。

进一步的，所述S4步骤具体为：重复步骤S2-S3，直到达到预设的最大迭代次数。在每次迭代中，为了推荐的可靠性和精准性，每次迭代只能选出少量可靠正样本，所以为了选出一定量的可靠正样本进行批量推荐，需要进行多次迭代。在每次迭代中，步骤S3每次会从未标记数据集U中选出部分可靠正样本加入到正样本数据集P中，同时将他们从未标记数据集U中删除。这种情况下，|P|的样本大小会随着迭代次数的增加而增大，|U|的样本大小会随着迭代次数的增加而减少，直至达到预设的最大迭代次数。由此，在多次迭代过程中从U中选出的所有可靠正样本即可作为金融产品潜在购买用户，向金融产品拥有方进行精准批量推荐。

本发明的有益效果在于：本发明针对金融产品提供方只拥有正样本数据而无法进行推荐的问题，结合纵向联邦学习和半监督学习方法，在保护多方数据安全隐私的情况下，联合多方进行金融产品潜在用户推荐模型训练与预测，有效地解决了只有少量正样本和大量未标记样本的批量推荐，提高了推荐的可靠性，实现了金融产品潜在用户的精准与批量推荐。

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为多方半监督学习的金融产品潜在用户推荐方法的流程示意图；

图2为纵向联邦学习多方数据样本对齐示意图；

图3为随机采样组成训练集和测试集示意图；

图4为构建梯度直方图示意图；

图5为纵向联邦GBDT模型的训练过程示意图；

图6为纵向联邦GBDT模型的预测过程示意图。

### 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。

其中，附图仅用于示例性说明，表示的仅是示意图，而非实物图，不能理解为对本发明的限制；为了更好地说明本发明的实施例，附图某些部件会有省略、放大或缩小，并不代表实际产品的尺寸；对本领域技术人员来说，附图中某些公知结构及其说明可能省略是可以理解的。

本发明实施例的附图中相同或相似的标号对应相同或相似的部件；在本发明的描述中，需要理解的是，若有术语“上”、“下”、“左”、“右”、“前”、“后”等指示的方位或位置关系为基于附图所示的方位或位置关系，仅是为了便于描述本发明和简化描述，而不是指示或暗示所指的装置或元件必须具有特定的方位、以特定的方位构造和操作，因此附图中描述位置关系的用语仅用于示例性说明，不能理解为对本发明的限制，对于本领域的普通技术人员而言，可以根据具体情况理解上述术语的具体含义。

以灵活就业人员公积金缴存推荐应用场景为例，尚未缴存灵活就业人员公积金的潜在用户相关信息在多方，且各方都有数据安全隐私保护的要求。当多方数据进行纵向联邦学习，进行样本对齐后，A方拥有的是已缴存公积金的灵活就业人员信息，B方和C方拥有已缴存公积金的灵活就业人员信息和未缴存公积金的潜在灵活就业人员信息。

在本实施例中：A方为公积金方，拥有灵活就业人员公积金缴存数据，包括个人基本信息、公积金账户号、公积金余额、公积金缴存期数、公积金缴存比例、公积金缴存额等信息。可选的，B方可以为税务数据方，包括个人基本信息、个人所得税，房管税等信息。可选的，C方可以为社保数据方，包括个人基本信息、医疗保险、养老保险、失业保险等信息。

请参阅图1～图6，为一种多方半监督学习的金融产品潜在用户推荐方法，该方法包含以下步骤：

S1：建立灵活就业人员缴存推荐多方数据集，包含有灵活就业人员公积金缴存信息的A方以及B方、C方。对多方数据集进行预处理，并进行样本对齐，组成正样本数据集P和未标记样本数据集U。具体步骤如下：

S11：对灵活就业人员缴存推荐多方数据集进行数据预处理操作包括冗余数据处理、缺失值处理、异常值处理、数据标准化处理以及标签数据处理，得到A方数据集，表示A方第i个样本特征向量，向量维度为，表示对应的标签，其中a；B方数据集，表示B方第i个样本特征向量，向量维度为，表示对应的标签，其中b；C方数据集，表示C方第i个样本特征向量，向量维度为，表示对应的标签，其中c。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。可选的，采用中位数填充的方法进行缺失值填充。

所述异常值处理具体包括：根据实际情况，可选的，使用标准差方法确认异常值的阈值。可选的，利用箱线图对数据进行可视化分析，发现超出阈值的数据。可选的，选择用中位数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，可选的，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。

所述标签数据处理具体包括：A方数据集添加一列标签列，将其标签数据设为“1”，代表正样本数据；B方、C方数据集分别添加一列标签列，将其标签数据设为“0”，代表未标记数据。

S12：对B方和C方数据集和按照其样本ID进行样本加密对齐，保留B方和C方对齐的样本数据，丢弃未对齐样本数据，得到n个样本。对齐后的B方数据集为，C方数据集为，表示对应的标签，表示对应的标签，其中n。

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：对数据集和按照其样本ID进行加密对齐，将三方对齐的样本作为正样本，组成正样本数据集，其中，分别表示A方、B方、C方对齐后的第i个样本特征向量，为正样本标签，表示正样本个数，；三方未对齐的样本作为未标记样本，组成未标记样本数据集，其中，分别表示B方、C方对齐后的第i个样本特征向量，为未标记样本标签，表示未标记样本个数，。

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。

S2：建立预测未标记样本候选推荐过程，循环执行该过程M轮。可选的，将循环总轮数M设为10。第m轮预测过程如下：从U中随机有放回地抽取个样本，表示P的样本个数，将这个未标记样本组成负样本数据集。由P和组成训练集，U中未被抽取样本组成预测集。构建以梯度提升树（Gradient Boosting Decision Tree，GBDT）作为基学习器的纵向联邦模型，在上进行训练。将输入到训练好的模型中进行预测，得到中每个样本的分数，作为该样本的预测分数。

S21：使用bootstrap采样法从U中随机有放回地抽取个样本，将这个未标记样本组成第m轮负样本数据集，其中，分别是第m轮的第i个负样本中B方和C方的特征向量，表示第m轮第i个负样本的标签，，||表示第m轮负样本数据集的样本个数。由正样本数据集P和第m轮的负样本数据集组成对应的第m轮训练集，其中，分别表示第m轮的第i个样本中B方和C方的特征向量，表示第m轮第i个样本的标签，；未标记样本数据集U中未被抽取样本组成对应的第m轮预测集，其中表示第m轮的第j个样本的特征向量，。

S22：使用GBDT算法作为基学习器，构建一个纵向联邦模型，通过T颗决策树的集成对进行训练，对输入数据中每个样本来预测其第m轮输出 ，其中T表示第m轮的决策树的总数，可选的，将决策树总数T的值设为40，表示第m轮第棵树的预测结果，表示第m轮第i个样本特征向量，。根据预测输出和真实标签之间的损失函数的梯度值，建立一阶导数和二阶导数梯度直方图，联合双方所有特征的梯度直方图找到当前节点的全局最佳分割从而构造出最优决策树。

首先，初始化第m轮每个样本的预测结果，为一个随机值，可选的，的随机值为-1或1，接着第m轮第t颗树的联邦训练具体流程如下：

S221：从B方开始，首先对B方的每个样本计算第m轮模型损失函数的一阶梯度和二阶梯度，其中，表示第m轮由前棵树聚合的对样本的预测结果，表示第m轮第i个样本的真实标签，是损失函数。使用加法同态加密对和进行加密，得到和，B方将和发送给C方。

可选的，这里损失函数使用均方误差损失函数，采用的同态加密方案为Paillier同态加密。使用表示对数据a进行了同态加密。

S222：对于C方，根据自己的特征数据建立第m轮梯度直方图，并且将加密后的梯度直方图发送给B方。具体流程如下：

S2221：对于当前第m轮样本中所有C方特征，根据每个特征的特征值对所有样本进行排序，然后将排序后的样本进行分桶划分到*q*个类别中，得到各个类别的相应特征阈值，其中表示特征编号，表示第m轮编号为的特征第q个类别的阈值。可选的，这里类别q设为10。

S2222：根据从B方得到的第m轮的和，C方进行加密梯度信息聚合，构建第m轮加密梯度直方图即

其中，*v*=1,2,...,*q*。

S2223：C方将计算得到的第m轮的和发送给B方。

S223：B方对来自C方的第m轮加密梯度直方图进行解密，根据分裂增益计算公式，枚举每个特征梯度直方图进行最优解计算，找到全局最优分裂点，并将分裂信息返回给C方进行解析。具体流程如下：

S2231：B方对当前节点空间的所有样本的一阶梯度和二阶梯度进行聚合，执行，其中表示当前节点的所有样本。

S2232：B方对从C方得到的第m轮和进行解密，得到第m轮解密值和，依次对C方的所有特征的所有类别进行如下计算得到、、和：

其中表示分割后左子节点的样本空间，表示分割后右子节点的样本空间，

表示第m轮中左子节点样本空间的所有样本的损失函数一阶梯度之和，

表示第m轮中左子节点样本空间的所有样本的损失函数二阶梯度之和，

表示第m轮中右子节点样本空间的所有样本的损失函数一阶梯度之和，

表示第m轮中右子节点样本空间的所有样本的损失函数二阶梯度之和。

S2233：计算第m轮当前节点的最佳分割值：

其中是超参数。可选的，超参数设为0.5。

S2234：对于样本的每个特征的所有阈值均可以得到一个值，选出其中最大的值，确定该特征阈值为第m轮的全局最佳分割，全局最佳分割可以用[参与方，特征编号()，阈值编号()]来表示，并将特征编号()和阈值编号()返回给C方。

S224：C方根据从B方发送的特征编号和阈值编号确定特征的阈值，并对当前的样本空间进行划分。然后，C方在本地建立一个查找表，记录选中特征的阈值，形成记录[记录编号，特征，阈值]，并将记录编号和划分后左侧的样本空间（）返回给B方。

S225：B方将会根据收到的[记录编号，]对当前节点进行划分，并将当前节点与[参与方，记录编号]关联。B方将当前节点的划分信息与C方同步，并进入对下一节点的分割。

S226：迭代步骤S222~S225，直至达到训练停止条件或树的最大深度。

S23：使用训练好的第m轮GBDT决策模型，对第m轮测试集进行预测，获得第m轮测试集中每个样本的预测分数，遍历U中所有样本，在U中找到与的ID相同的样本，并把的预测分数赋值给，作为样本的第m轮预测分数，用表示，其中为未标记数据集U中的样本编号。如果U中样本第m轮中未存在于中，则该样本对应的第m轮预测分数等于0。具体流程如下：

S231：B方查询与当前节点相关联的[参与方，记录编号]记录。基于该记录，B方向C方发送待标注样本编号和记录编号，并且询问下一步的树搜索方向，即向左子节点或右子节点。

S232：C方接收到待标注样本编号和记录编号后，将待标注样本中相应特征的值与本地查找表中的记录[记录编号，特征，阈值]中的阈值进行比较，得出下一步的树搜索方向。然后，该C方将搜索决定发往B方。

S233：B方接收到C方传来的搜索决定，前往相应的子节点。

S234：迭代步骤S231~S233，直至到达一个叶节点，得到对应的分类标签以及该标签的权值，即可得到与样本对应的U中样本的第m轮预测分数。

其中表示叶节点的样本空间，是超参数。可选的，超参数设为0.5。

S3：由U中每个样本的M轮预测分数总和及其在M轮预测集中出现次数总和，计算样本预测为正样本的概率。按照正样本的概率从大到小对U中所有样本进行排序，根据先验知识选出排名靠前的样本作为可靠正样本，并将它们加入到正样本数据集P中，同时将它们从U中删除。可选的，总轮数M设为10。

S31：由U中每个样本的预测分数总和及其在M轮预测集中出现次数总和，计算样本预测为正样本的概率,计算公式如下：

其中为指示函数，表示如果U中的样本存在于第m轮的测试集中，则*I*=1，否则*I*=0。

S32：对U中所有样本使用其概率按降序排序，选出排名前的样本作为可靠正样本，并将它们加入到正样本数据集P中，同时将它们从U中删除，其中的值是根据先验知识设定的。可选的，设为0.1。

S4：重复步骤S2-S3，直到达到预设的最大迭代次数。可选的，最大迭代次数设为5。在每次迭代中，为了推荐的可靠性和精准性，每次迭代只能选出少量可靠正样本，所以为了选出一定量的可靠正样本进行批量推荐，需要进行多次迭代。在每次迭代中，步骤S3每次会从未标记数据集U中选出部分可靠正样本加入到正样本数据集P中，同时将他们从未标记数据集U中删除。这种情况下，|P|的样本大小会随着迭代次数的增加而增大，|U|的样本大小会随着迭代次数的增加而减少，直至达到预设的最大迭代次数。由此，在多次迭代过程中从U中选出的所有可靠正样本即可作为公积金潜在缴存用户，向公积金方进行批量推荐。

例如，一批客户在B方税务数据中的个人基本信息、个人所得税，房管税等与应用场景相关信息；该批客户在C方社保数据中的个人基本信息、医疗保险、养老保险、失业保险等与应用场景信息，经预处理，分别放入训练好的纵向联邦GBDT模型中，各方根据该批客户特征特点判断该批客户中是否存在灵活就业人员公积金缴存潜在用户，然后将判断为潜在用户的客户们向公积金方进行批量推荐。

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。

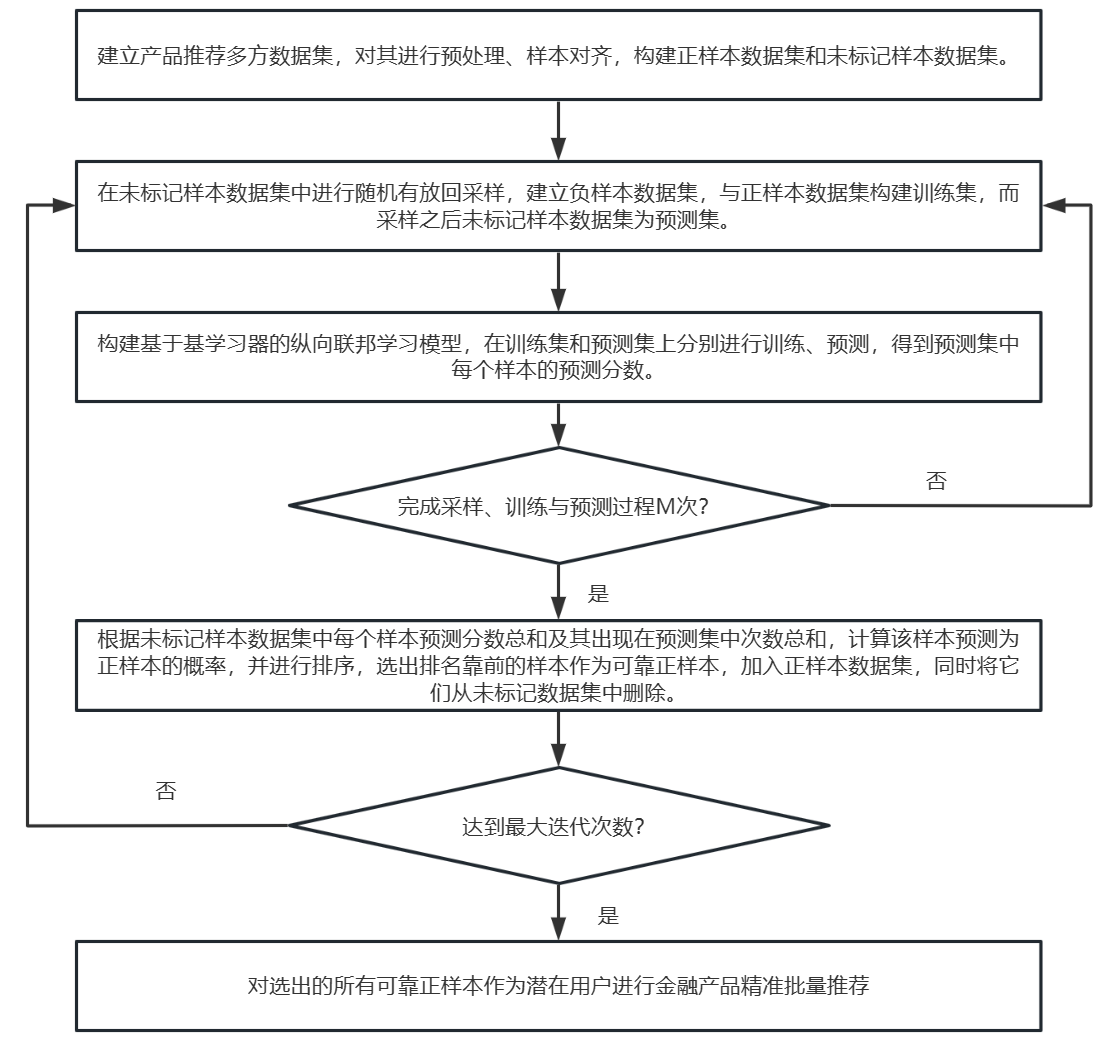


图1

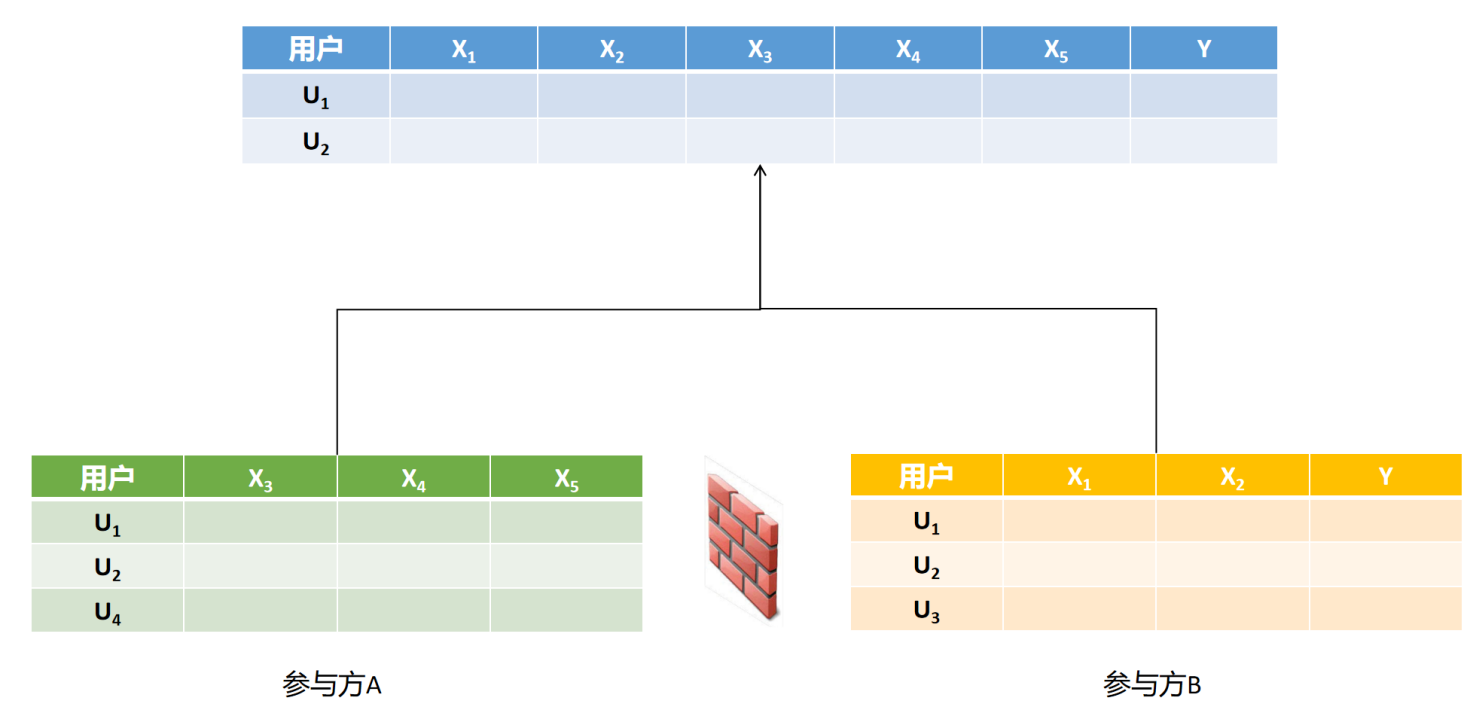


图2

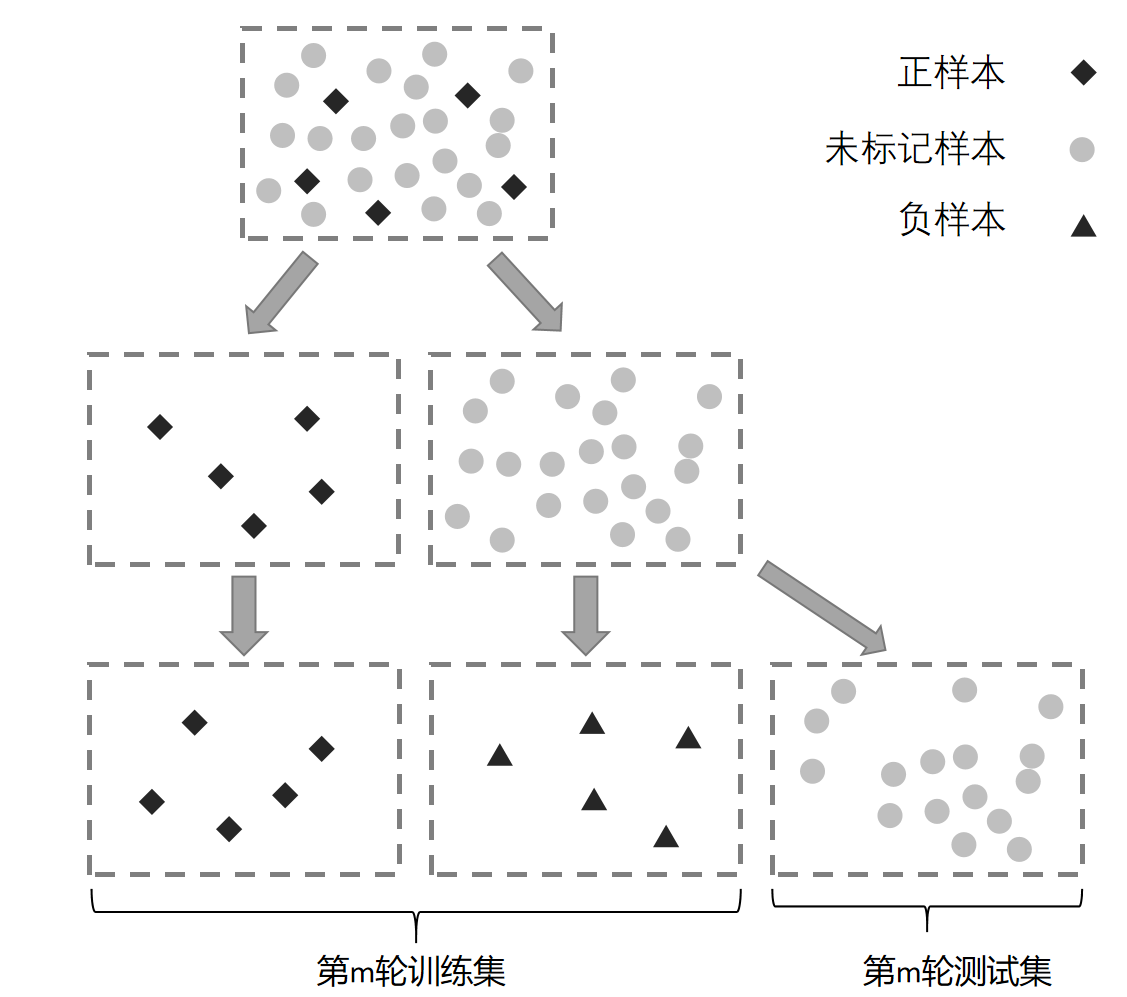


图3

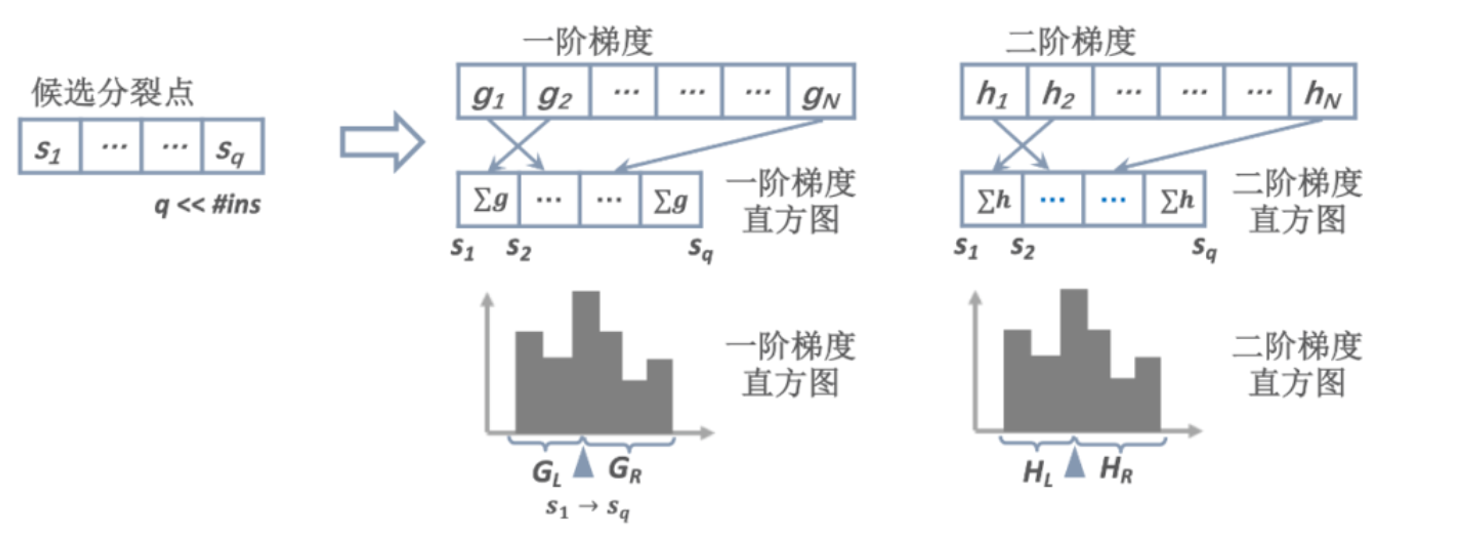


图4

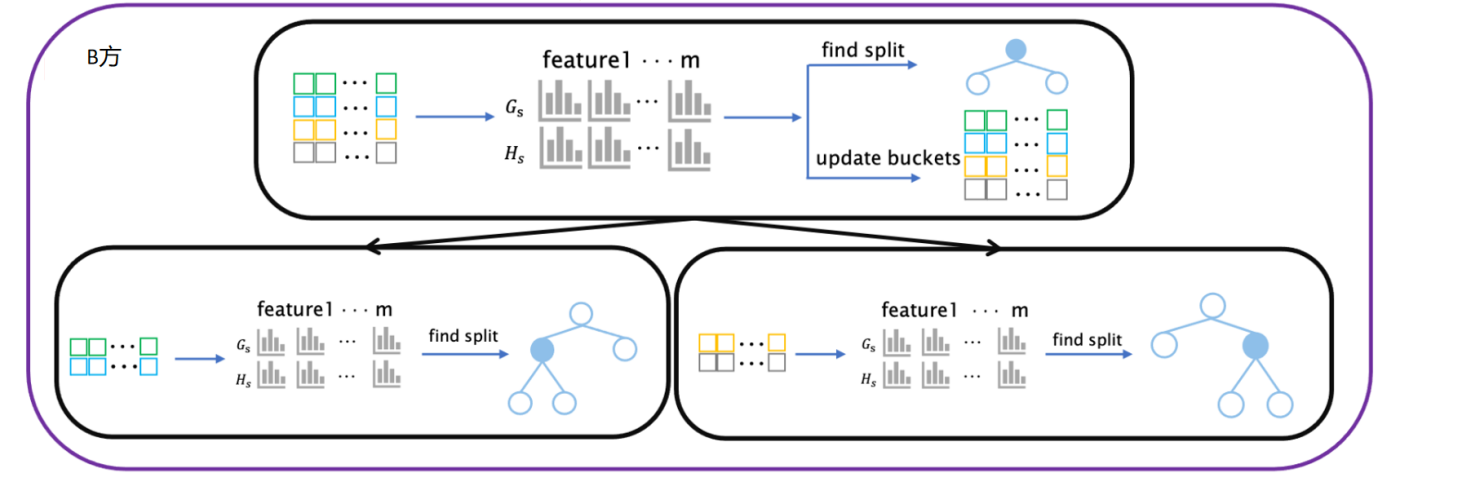


图5

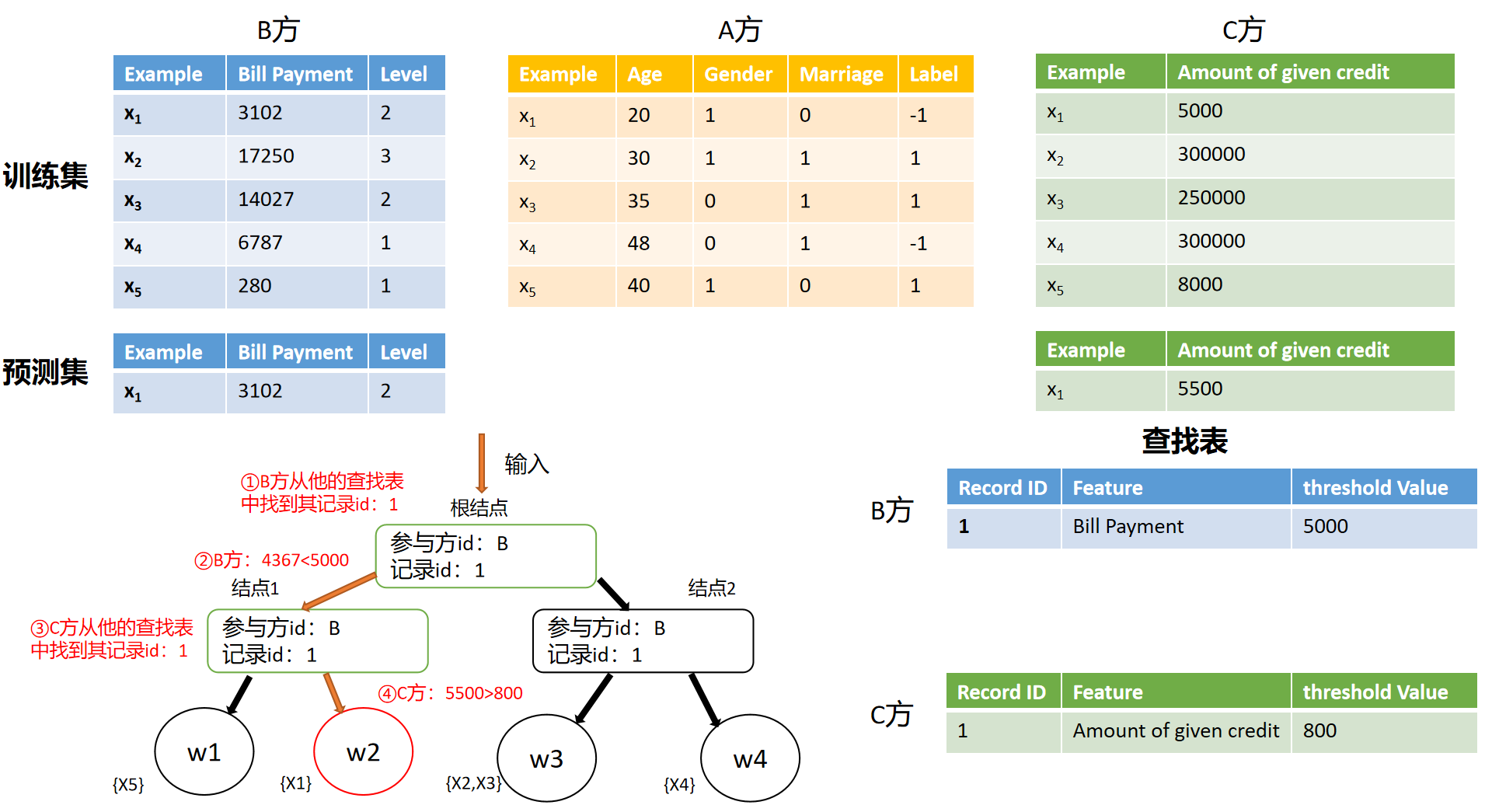


图6