(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 107330459 B
(45)授权公告日 2021.09.14

(21)申请号 201710509026.2
(22)申请日 2017.06.28
(65)同一申请的已公布的文献号
申请公布号 CN 107330459 A
(43)申请公布日 2017.11.07
(73)专利权人 联想(北京)有限公司
地址 100085 北京市海淀区上地信息产业基地创业路6号
(72)发明人 杨帆 王耀晖 金宝宝
(74)专利代理机构 北京集佳知识产权代理有限公司 11227
代理人 王宝筠
(51)Int.Cl.
G06K 9/62(2006.01)
G06Q 30/02(2012.01)

(56)对比文件
CN 105701498 A,2016.06.22
CN 104679835 A,2015.06.03

审查员 刘志军

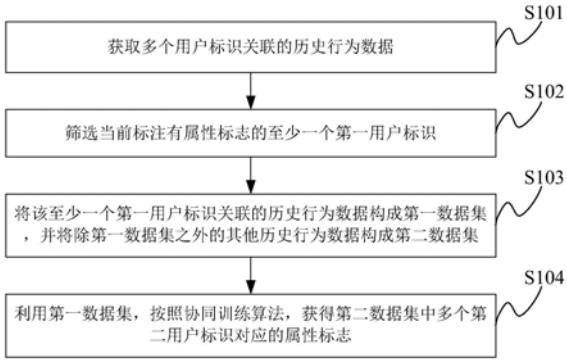
权利要求书4页 说明书17页 附图6页

(54)发明名称

一种数据处理方法、装置和电子设备

(57)摘要

本申请提供了一种数据处理方法、装置及电子设备，在获取到多个用户标识关联的历史行为数据，可以筛选当前标注有属性标志的至少一个第一用户标识，并将该至少一个用户标识关联的历史行为数据构成第一数据集，其他历史行为数据构成第二数据集，之后，本申请利用该第一数据集，按照协同训练算法，获得第二数据集中多个第二用户标识对应的属性标志。由此可见，本申请通过对少量标注有属性标志的历史行为数据进行训练扩展，自动且准确得到了大量用户标识对应的属性标志，无需人工一一标注各用户标识的属性标志，降低了人力成本，且大大提高了工作效率。



CN 107330459 B

1.一种数据处理方法,其特征在于,所述方法包括:

获取多个用户标识关联的历史行为数据;

对所述历史行为数据中部分行为数据对应的用户标识,人工标注该用户的属性标志,对这部分用户标识的属性标志进行验证以保证其真实性,以及保证后续依据人工标注的用户的属性标志训练扩展得到的其他用户标识的属性标志可靠性;

筛选当前标注有属性标志的至少一个第一用户标识;

将所述至少一个第一用户标识关联的历史行为数据构成第一数据集,并将除所述第一数据集之外的其他历史行为数据构成第二数据集;

利用所述第一数据集,按照协同训练算法,自动获得所述第二数据集中多个第二用户标识对应的属性标志;

其中,在所述筛选当前标注有属性标志的至少一个第一用户标识之前,所述方法还包括:获取所述历史行为数据的第一视图特征和第二视图特征,所述第一视图特征为增强型RFM特征,所述第二视图特征为行为序列特征;利用所述第一视图特征和所述第二视图特征,生成所述历史行为数据关联用户标识对应的特征数据;

其中,利用所述第一数据集,按照协同训练算法,自动获得所述第二数据集中多个第二用户标识对应的属性标志,包括:利用两种基于不同特征的算法,分别对第一数据集中标注有属性标志的特征数据中的增强型RFM特征和行为序列进行模型训练,得到相应的两个预测模型,利用两个预测模型对第二数据集中同一特征数据进行属性预测,得到的两个属性标志,分别计算这两个属性标志的置信度,选择置信度较高的属性标志作为该特征数据的属性标志。

2.根据权利要求1所述的方法,其特征在于,所述筛选当前标注有属性标志的至少一个第一用户标识,包括:

验证多个用户标识对应的特征数据中是否包含属性标志;

基于验证结果,确定具有属性标志的特征数据对应的第一用户标识。

3.根据权利要求2所述的方法,其特征在于,所述利用所述第一数据集,按照利用协同训练算法,自动获得所述第二数据集中多个第二用户标识对应的属性标志,包括:

分别对所述第一数据集和所述第二数据集进行备份,得到相应的第一备份数据集和第二备份数据集;

利用第一数据集,按照第一类算法,对所述第二数据集中多个第二用户标识关联的特征数据进行属性预测,确定至少一个第二用户标识对应的属性标志,更新所述第一备份数据集中与确定的属性标志对应的第二用户标识的特征数据;

利用第一备份数据集,按照第二类算法,对所述第二备份数据集中多个第二用户标识关联的特征数据进行属性预测,确定至少一个第二用户标识对应的属性标志,更新所述第一数据集中与确定的属性标志对应的第二用户标识的特征数据;

基于更新后的第一数据集以及更新后的第一备份数据集,确定所述多个第二用户标识对应的属性标志。

4.根据权利要求3所述的方法,其特征在于,所述利用第一数据集,按照第一类算法,对所述第二数据集中多个第二用户标识关联的特征数据进行属性预测,确定至少一个第二用户标识对应的属性标志,更新所述第一备份数据集中与确定的属性标志对应的第二用户标

识的特征数据，包括：

按照第一类算法，对所述第一数据集中的第一视图特征进行训练，生成第一预测模型；

利用所述第一预测模型，对所述第二数据集中的多个第二用户标识关联的特征数据进行计算，获得所述多个第二用户标识对应的属性标志；

判断所述属性标志的置信度是否大于第一阈值；

基于判断结果，利用大于所述第一阈值的置信度对应的属性标志，标注对应的用户标识关联的特征数据，并将标注后的特征数据更新到所述第一备份数据集。

5.根据权利要求3所述的方法，其特征在于，所述利用第一备份数据集，按照第二类算法，对所述第二备份数据集中多个第二用户标识关联的特征数据进行属性预测，确定至少一个第二用户标识对应的属性标志，更新所述第一数据集中与确定的属性标志对应的第二用户标识的特征数据，包括：

按照第二类算法，对所述第一备份数据集中的第二视图特征进行训练，生成第二预测模型；

利用所述第二预测模型，对所述第二备份数据集中的多个第二用户标识关联的特征数据进行计算，获得所述多个第二用户标识对应的属性标志；

判断所述属性标志的置信度是否大于第二阈值；

基于判断结果，利用大于所述第二阈值的置信度对应的属性标志，标注对应的用户标识关联的特征数据，并将标注后的特征数据更新到所述第一数据集。

6.根据权利要求3所述的方法，其特征在于，所述基于更新后的第一数据集以及更新后的第一备份数据集，确定所述多个第二用户标识对应的属性标志，包括：

验证更新后的第一数据集与更新后的第一备份数据集中，同一用户标识对应的属性标志是否相同；

如果相同，将所述属性标志确定为对应用户标识的目标属性标志；

如果不相同，将对应的用户标识关联的特征数据，以及当前所述第二数据集和所述第二备份数据集中未标注属性标志的特征数据，确定为待定数据集；

利用所述第一类算法和所述第二类算法，对所述待定数据集中的特征数据进行属性预测；

筛选符合预设置信度要求的属性标志，确定为相应特征数据的目标属性标志。

7.根据权利要求2所述的方法，其特征在于，所述获取所述历史行为数据的第一视图特征和第二视图特征，包括：

获取所述历史行为数据中多种行为类型对应的增强型RFM特征数据；

获取所述历史行为数据中预设时间段内至少一种行为类型对应的行为编码；

利用获取的行为编码，确定所述历史行为数据关联的用户标识的行为序列。

8.一种数据处理装置，其特征在于，所述装置包括：

第一获取模块，用于获取多个用户标识关联的历史行为数据；对所述历史行为数据中部分行为数据对应的用户标识，人工标注该用户的属性标志，对这部分用户标识的属性标志进行验证以保证其真实性，以及保证后续依据人工标注的用户的属性标志训练扩展得到的其他用户标识的属性标志可靠性；

筛选模块，用于筛选当前标注有属性标志的至少一个第一用户标识；

数据集构成模块，用于将所述至少一个第一用户标识关联的历史行为数据构成第一数据集，并将除所述第一数据集之外的其他历史行为数据构成第二数据集；

训练模块，用于利用所述第一数据集，按照协同训练算法，自动获得所述第二数据集中多个第二用户标识对应的属性标志；

第二获取模块，用于获取所述历史行为数据的第一视图特征和第二视图特征，所述第一视图特征为增强型RFM特征，所述第二视图特征为行为序列特征；

第一生成模块，用于利用所述第一视图特征和所述第二视图特征，生成所述历史行为数据关联用户标识对应的特征数据；

其中，所述训练模块，具体用于利用两种基于不同特征的算法，分别对第一数据集中标注有属性标志的特征数据中的增强型RFM特征和行为序列进行模型训练，得到相应的两个预测模型，利用两个预测模型对第二数据集中同一特征数据进行属性预测，得到的两个属性标志，分别计算这两个属性标志的置信度，选择置信度较高的属性标志作为该特征数据的属性标志。

9.根据权利要求8所述的装置，其特征在于，所述筛选模块包括：

第一验证单元，用于验证多个用户标识对应的特征数据中是否包含属性标志；

第一确定单元，用于基于验证结果，确定具有属性标志的特征数据对应的第一用户标识。

10.根据权利要求9所述的装置，其特征在于，所述训练模块包括：

备份单元，用于分别对所述第一数据集和所述第二数据集进行备份，得到相应的第一备份数据集和第二备份数据集；

第一预测更新单元，用于利用第一数据集，按照第一类算法，对所述第二数据集中多个第二用户标识关联的特征数据进行属性预测，确定至少一个第二用户标识对应的属性标志，更新所述第一备份数据集中与确定的属性标志对应的第二用户标识的特征数据；

第二预测更新单元，用于利用第一备份数据集，按照第二类算法，对所述第二备份数据集中多个第二用户标识关联的特征数据进行属性预测，确定至少一个第二用户标识对应的属性标志，更新所述第一数据集中与确定的属性标志对应的第二用户标识的特征数据；

第二确定单元，用于基于更新后的第一数据集以及更新后的第一备份数据集，确定所述多个第二用户标识对应的属性标志。

11.根据权利要求10所述的装置，其特征在于，所述第一预测更新单元包括：

第一模型生成单元，用于按照第一类算法，对所述第一数据集中的第一视图特征进行训练，生成第一预测模型；

第一计算单元，用于利用所述第一预测模型，对所述第二数据集中的多个第二用户标识关联的特征数据进行计算，获得所述多个第二用户标识对应的属性标志；

第一判断单元，用于判断所述属性标志的置信度是否大于第一阈值；

第一更新单元，用于基于判断结果，利用大于所述第一阈值的置信度对应的属性标志，标注对应的用户标识关联的特征数据，并将标注后的特征数据更新到所述第一备份数据集。

12.根据权利要求10所述的装置，其特征在于，所述第二预测更新单元包括：

第二模型生成单元，用于按照第二类算法，对所述第一备份数据集中的第二视图特征

进行训练，生成第二预测模型；

第二计算单元，用于利用所述第二预测模型，对所述第二备份数据集中的多个第二用户标识关联的特征数据进行计算，获得所述多个第二用户标识对应的属性标志；

第二判断单元，用于判断所述属性标志的置信度是否大于第二阈值；

第二更新单元，用于基于判断结果，利用大于所述第二阈值的置信度对应的属性标志，标注对应的用户标识关联的特征数据，并将标注后的特征数据更新到所述第一数据集。

13.根据权利要求10所述的装置，其特征在于，所述第二确定单元包括：

第一验证单元，用于验证更新后的第一数据集与更新后的第一备份数据集中，同一用户标识对应的属性标志是否相同；

第三确定单元，用于所述第一验证单元的验证结果为是时，将所述属性标志确定为对应用户标识的目标属性标志；

第四确定单元，用于所述第一验证单元的验证结果为否时，将对应的用户标识关联的特征数据，以及当前所述第二数据集和所述第二备份数据集中未标注属性标志的特征数据，确定为待定数据集；

属性预测单元，用于利用所述第一类算法和所述第二类算法，对所述待定数据集中的特征数据进行属性预测；

筛选单元，用于筛选符合预设置信度要求的属性标志，确定为相应特征数据的目标属性标志。

14.一种电子设备，其特征在于，所述电子设备包括：

通信端口；

存储器，用于存储实现如权利要求1-7任意一项所述的数据处理方法的多个指令；

处理器，用于加载并执行所述多个指令，包括：

获取多个用户标识关联的历史行为数据；

对所述历史行为数据中部分行为数据对应的用户标识，人工标注该用户的属性标志，对这部分用户标识的属性标志进行验证以保证其真实性，以及保证后续依据人工标注的用户的属性标志训练扩展得到的其他用户标识的属性标志可靠性；

筛选当前标注有属性标志的至少一个第一用户标识；

将所述至少一个第一用户标识关联的历史行为数据构成第一数据集，并将除所述第一数据集之外的其他历史行为数据构成第二数据集；

利用所述第一数据集，按照协同训练算法，自动获得所述第二数据集中多个第二用户标识对应的属性标志。

一种数据处理方法、装置和电子设备

技术领域

[0001] 本申请主要涉及用户属性预测应用领域，更具体地说是涉及一种数据处理方法、装置和电子设备。

背景技术

[0002] 如今，随着网络技术的高速发展，在开发新产品或业务之前以及在使用过程中，通常会对用户的性别、年龄、收入、兴趣等属性信息进行研究，以便知晓并满足用户的潜在需求，并据此完成新产品或业务的功能完善，提高用户使用新产品或业务的体验感受。

[0003] 现有技术中，通常是通过注册用户填写的资料，得知用户属性信息，然而，由于用户避免个人信息泄露，经常会胡乱填写错误资料或不填写，将导致得到用户属性信息不准确。

[0004] 为了得到准确地用户属性信息，目前提出人工标注的方式来获得用户属性信息，但通常情况下，企业并不知道用户的性别、年龄、收入等属性信息，要想获得大批量的用户属性信息，需要付出大量的人力、物力，过程非常复杂，工作效率很低。

发明内容

[0005] 有鉴于此，本发明提供了一种数据处理方法、装置及电子设备，通过对标注有属性标志的少量用户标识关联的行为数据进行训练扩展，得到可靠且准确的大量标注属性标志的用户标识关联的行为数据，无需人工一一标注，大大节省了标注成本，且提高了属性标注可靠性以及准确性，进而提高了属性预测的效率以及准确性。

[0006] 为了实现上述发明目的，本申请提供了以下技术方案：

[0007] 一种数据处理方法，所述方法包括：

[0008] 获取多个用户标识关联的历史行为数据；

[0009] 筛选当前标注有属性标志的至少一个第一用户标识；

[0010] 将所述至少一个第一用户标识关联的历史行为数据构成第一数据集，并将除所述第一数据集之外的其他历史行为数据构成第二数据集；

[0011] 利用所述第一数据集，按照协同训练算法，获得所述第二数据集中多个第二用户标识对应的属性标志。

[0012] 优选的，在所述筛选当前标注有属性标志的至少一个第一用户标识之前，所述方法还包括：

[0013] 获取所述历史行为数据的第一视图特征和第二视图特征；

[0014] 利用所述第一视图特征和所述第二视图特征，生成所述历史行为数据关联用户标识对应的特征数据；

[0015] 所述筛选当前标注有属性标志的至少一个第一用户标识，包括：

[0016] 验证多个用户标识对应的特征数据中是否包含属性标志；

[0017] 基于验证结果，确定具有属性标志的特征数据对应的第一用户标识。

[0018] 优选的，所述利用所述第一数据集，按照利用协同训练算法，获得所述第二数据集中多个第二用户标识对应的属性标志，包括：

[0019] 分别对所述第一数据集和所述第二数据集进行备份，得到相应的第一备份数据集和第二备份数据集；

[0020] 利用第一数据集，按照第一类算法，对所述第二数据集中多个第二用户标识关联的特征数据进行属性预测，确定至少一个第二用户标识对应的属性标志，更新所述第一备份数据集中与确定的属性标志对应的第二用户标识的特征数据；

[0021] 利用第一备份数据集，按照第二类算法，对所述第二备份数据集中多个第二用户标识关联的特征数据进行属性预测，确定至少一个第二用户标识对应的属性标志，更新所述第一数据集中与确定的属性标志对应的第二用户标识的特征数据；

[0022] 基于更新后的第一数据集以及更新后的第一备份数据集，确定所述多个第二用户标识对应的属性标志。

[0023] 优选的，所述利用第一数据集，按照第一类算法，对所述第二数据集中多个第二用户标识关联的特征数据进行属性预测，确定至少一个第二用户标识对应的属性标志，更新所述第一备份数据集中与确定的属性标志对应的第二用户标识的特征数据，包括：

[0024] 按照第一类算法，对所述第一数据集中的第一视图特征进行训练，生成第一预测模型；

[0025] 利用所述第一预测模型，对所述第二数据集中的多个第二用户标识关联的特征数据进行计算，获得所述多个第二用户标识对应的属性标志；

[0026] 判断所述属性标志的置信度是否大于第一阈值；

[0027] 基于判断结果，利用大于所述第一阈值的置信度对应的属性标志，标注对应的用户标识关联的特征数据，并将标注后的特征数据更新到所述第一备份数据集。

[0028] 优选的，所述利用第一备份数据集，按照第二类算法，对所述第二备份数据集中多个第二用户标识关联的特征数据进行属性预测，确定至少一个第二用户标识对应的属性标志，更新所述第一数据集中与确定的属性标志对应的第二用户标识的特征数据，包括：

[0029] 按照第二类算法，对所述第一备份数据集中的第二视图特征进行训练，生成第二预测模型；

[0030] 利用所述第二预测模型，对所述第二备份数据集中的多个第二用户标识关联的特征数据进行计算，获得所述多个第二用户标识对应的属性标志；

[0031] 判断所述属性标志的置信度是否大于第二阈值；

[0032] 基于判断结果，利用大于所述第二阈值的置信度对应的属性标志，标注对应的用户标识关联的特征数据，并将标注后的特征数据更新到所述第一数据集。

[0033] 优选的，所述基于更新后的第一数据集以及更新后的第一备份数据集，确定所述多个第二用户标识对应的属性标志，包括：

[0034] 验证更新后的第一数据集与更新后的第一备份数据集中，同一用户标识对应的属性标志是否相同；

[0035] 如果相同，将所述属性标志确定为对应用户标识的目标属性标志；

[0036] 如果不相同，将对应的用户标识关联的特征数据，以及当前所述第二数据集和所述第二备份数据集中未标注属性标志的特征数据，确定为待定数据集；

[0037] 利用所述第一类算法和所述第二类算法，对所述待定数据集中的特征数据进行属性预测；

[0038] 筛选符合预设置信度要求的属性标志，确定为相应特征数据的目标属性标志。

[0039] 优选的，所述获取所述历史行为数据的第一视图特征和第二视图特征，包括：

[0040] 获取所述历史行为数据中多种行为类型对应的RFM特征数据；

[0041] 获取所述历史行为数据中预设时间段内至少一种行为类型对应的行为编码；

[0042] 利用获取的行为编码，确定所述历史行为数据关联的用户标识的行为序列。

[0043] 一种数据处理装置，所述装置包括：

[0044] 第一获取模块，用于获取多个用户标识关联的历史行为数据；

[0045] 筛选模块，用于筛选当前标注有属性标志的至少一个第一用户标识；

[0046] 数据集构成模块，用于将所述至少一个第一用户标识关联的历史行为数据构成第一数据集，并将除所述第一数据集之外的其他历史行为数据构成第二数据集；

[0047] 训练模块，用于利用所述第一数据集，按照协同训练算法，获得所述第二数据集中多个第二用户标识对应的属性标志。

[0048] 优选的，所述装置还包括：

[0049] 第二获取模块，用于获取所述历史行为数据的第一视图特征和第二视图特征；

[0050] 第一生成模块，用于利用所述第一视图特征和所述第二视图特征，生成所述历史行为数据关联用户标识对应的特征数据；

[0051] 相应的，所述筛选模块包括：

[0052] 第一验证单元，用于验证多个用户标识对应的特征数据中是否包含属性标志；

[0053] 第一确定单元，用于基于验证结果，确定具有属性标志的特征数据对应的第一用户标识。

[0054] 优选的，所述训练模块包括：

[0055] 备份单元，用于分别对所述第一数据集和所述第二数据集进行备份，得到相应的第一备份数据集和第二备份数据集；

[0056] 第一预测更新单元，用于利用第一数据集，按照第一类算法，对所述第二数据集中多个第二用户标识关联的特征数据进行属性预测，确定至少一个第二用户标识对应的属性标志，更新所述第一备份数据集中与确定的属性标志对应的第二用户标识的特征数据；

[0057] 第二预测更新单元，用于利用第一备份数据集，按照第二类算法，对所述第二备份数据集中多个第二用户标识关联的特征数据进行属性预测，确定至少一个第二用户标识对应的属性标志，更新所述第一数据集中与确定的属性标志对应的第二用户标识的特征数据；

[0058] 第二确定单元，用于基于更新后的第一数据集以及更新后的第一备份数据集，确定所述多个第二用户标识对应的属性标志。

[0059] 优选的，所述第一预测更新单元包括：

[0060] 第一模型生成单元，用于按照第一类算法，对所述第一数据集中的第一视图特征进行训练，生成第一预测模型；

[0061] 第一计算单元，用于利用所述第一预测模型，对所述第二数据集中的多个第二用户标识关联的特征数据进行计算，获得所述多个第二用户标识对应的属性标志；

[0062] 第一判断单元，用于判断所述属性标志的置信度是否大于第一阈值；

[0063] 第一更新单元，用于基于判断结果，利用大于所述第一阈值的置信度对应的属性标志，标注对应的用户标识关联的特征数据，并将标注后的特征数据更新到所述第一备份数据集。

[0064] 优选的，所述第二预测更新单元包括：

[0065] 第二模型生成单元，用于按照第二类算法，对所述第一备份数据集中的第二视图特征进行训练，生成第二预测模型；

[0066] 第二计算单元，用于利用所述第二预测模型，对所述第二备份数据集中的多个第二用户标识关联的特征数据进行计算，获得所述多个第二用户标识对应的属性标志；

[0067] 第二判断单元，用于判断所述属性标志的置信度是否大于第二阈值；

[0068] 第二更新单元，用于基于判断结果，利用大于所述第二阈值的置信度对应的属性标志，标注对应的用户标识关联的特征数据，并将标注后的特征数据更新到所述第一数据集。

[0069] 优选的，所述第二确定单元包括：

[0070] 第一验证单元，用于验证更新后的第一数据集与更新后的第一备份数据集中，同一用户标识对应的属性标志是否相同；

[0071] 第三确定单元，用于所述第一验证单元的验证结果为是时，将所述属性标志确定为对应用户标识的目标属性标志；

[0072] 第四确定单元，用于所述第一验证单元的验证结果为否时，将对应的用户标识关联的特征数据，以及当前所述第二数据集和所述第二备份数据集中未标注属性标志的特征数据，确定为待定数据集；

[0073] 属性预测单元，用于利用所述第一类算法和所述第二类算法，对所述待定数据集中的特征数据进行属性预测；

[0074] 筛选单元，用于筛选符合预设置信度要求的属性标志，确定为相应特征数据的目标属性标志。

[0075] 一种电子设备，所述电子设备包括：

[0076] 通信端口；

[0077] 存储器，用于存储实现如上所述的数据处理方法的多个指令；

[0078] 处理器，用于加载并执行所述多个指令，包括：

[0079] 获取多个用户标识关联的历史行为数据；

[0080] 筛选当前标注有属性标志的至少一个第一用户标识；

[0081] 将所述至少一个第一用户标识关联的历史行为数据构成第一数据集，并将除所述第一数据集之外的其他历史行为数据构成第二数据集；

[0082] 利用所述第一数据集，按照协同训练算法，获得所述第二数据集中多个第二用户标识对应的属性标志。由此可见，与现有技术相比，本申请提供了一种数据处理方法、装置及电子设备，在获取到多个用户标识关联的历史行为数据，可以筛选当前标注有属性标志的至少一个第一用户标识，并将该至少一个用户标识关联的历史行为数据构成第一数据集，其他历史行为数据构成第二数据集，之后，本申请利用该第一数据集，按照协同训练算法，获得第二数据集中多个第二用户标识对应的属性标志。由此可见，本申请通过对少量标

注有属性标志的历史行为数据进行训练扩展，自动且准确得到了大量用户标识对应的属性标志，无需人工一一标注各用户标识的属性标志，降低了人力成本，且大大提高了工作效率。

附图说明

[0083] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术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发明的实施例，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，还可以根据提供的附图获得其他的附图。

[0084] 图1为本申请实施例提供的一种数据处理方法流程图；

[0085] 图2为本申请实施例提供的另一种数据处理方法流程图；

[0086] 图3本申请实施例提供的又一种数据处理方法流程图；

[0087] 图4为本申请实施例提供的一种数据处理装置的结构框图；

[0088] 图5为本申请实施例提供的另一种数据处理装置的结构框图；

[0089] 图6为本申请实施例提供的又一种数据处理装置的结构框图；

[0090] 图7为本申请实施例提供的又一种数据处理装置的结构框图；

[0091] 图8为本申请实施例提供的又一种数据处理装置的结构框图；

[0092] 图9为本申请实施例提供的一种电子设备的硬件结构图。

具体实施方式

[0093] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0094] 在实际应用中，为了提高用户属性预测的准确性，通常需要获取用户的大量的行为数据，作为其属性预测模型的训练样本，如通过产品日志、第三方关联或者人工标注等方式，获取用户的属性信息以及历史行为数据，之后，利用特征选择与提取算法，从中提取区别度较高的特征，作为属性预测模型的训练特征，采用SVM(support vector machine，支持向量机)、决策树、LR(Logistic Regression，逻辑回归)等算法，训练得到属性预测模型。基于此，可以利用该属性预测模型实现对用户的相关属性的预测。

[0095] 申请人经过分析得知，上述通过第三方数据关联获取的原始书记可信度不高，通过人工标注获取样本用户的属性信息的方式成本过高，且标注的属性信息通常也是不可靠的，导致训练得到的属性预测模型不可靠，这样，利用不可靠的属性预测模型预测得到的当前用户的属性必然也是不可靠的。

[0096] 为了解决上述问题，本申请提出通过少量的工作，获取一部分真实的标注数据，从而在此基础训练得到大量可靠地标准数据，再利用训练得到的大量的标注数据进行模型训练，得到可靠地属性预测模型，实现对用户难以标准的属性的准确预测。

[0097] 具体的，再获取到多个用户标识关联的历史行为数据后，本申请通过筛选当前标注有属性标志的至少一个第一用户标识，并将该至少一个用户标识关联的历史行为数据构

成第一数据集，其他历史行为数据构成第二数据集，之后，利用该第一数据集，按照协同训练算法，获得第二数据集中多个第二用户标识对应的属性标志。由此可见，本申请通过对少量标注有属性标志的历史行为数据进行训练扩展，自动且准确得到了大量用户标识对应的属性标志，无需人工一一标注各用户标识的属性标志，降低了人力成本，且大大提高了工作效率，且保证了得到的第二数据集对应的用户的属性标志的可靠性。

[0098] 为了使本申请的上述目的、特征和优点能够更加明显易懂，下面结合附图和具体实施方式对本发明作进一步详细的说明。

[0099] 参照图1，为本申请实施例提供的一种数据处理方法的流程图，该方法可以包括以下步骤：

[0100] 步骤S101，获取多个用户标识关联的历史行为数据；

[0101] 在本申请中，可以根据具体应用场景，通过不同途径获取多个用户的各种类型的用户行为数据，即多个用户标识关联的历史行为数据。

[0102] 例如，在网页浏览场景下，可以收集用户访问的网页类型、访问不同类型的网页的频率、在不同类型网页驻留的时长等等行为数据；在移动设备的使用场景中，可以收集用户通过该移动设备使用不同类型应用程序的频率、时长等行为数据；在购物应用场景下，可以收集用户购物次数和时长，以及购物类型等行为数据。可见，在不同应用场景下，收到的用户的行为数据通常是不同的，具体与具体的应用场景相关，本申请在此不再一一列举。

[0103] 其中，在实际应用中，为了区分收集到的不同的行为数据，可以为每一个用户设定对应的用户标识，如用户的账号、ID或编码等，在用户不同应用场景使用电子设备时，通过电子设备产生的行为数据通常携带有该用户的用户标识，以便通过行为数据与用户标识的关联关系，获取与每一个用户标识关联的所有历史行为数据。

[0104] 需要说明的是，本申请对获取每一个用户标识关联的历史行为数据的方式不作限定，可以包括本地电子设备收集到的与该用户标识关联的历史行为数据，还可以包括第三方应用平台(如支付平台、城市交通平台、生活服务平台等等)收集与该用户标识关联的历史行为数据，再从第三方应用平台获取该历史行为数据等等，本申请在此不再详述。

[0105] 可选的，本申请中，对于通过不同途径得到的历史行为数据，可以认为是用户行为流水数据，可以使形如“用户标识:用户某一时刻的行为”的格式的用户行为数据集A，此时，本申请可以按照用户标识进行分类存储，得到用户行为数据集B，如统计出“用户标识:用户所有的行为”的用户行为数据集B，但并不局限于这一种分类存储方式。

[0106] 步骤S102，筛选当前标注有属性标志的至少一个第一用户标识；

[0107] 在实际应用中，可以采用人工标注的方法，对得到的历史行为数据中部分行为数据对应的用户标识，标注该用户的基本属性，如对部分行为数据对应的用户标识标注该用户的性别、年龄、收入等基本属性，并将其作为相应的用户标识的属性标志进行保存。

[0108] 需要说明的是，由于本申请是基于部分用户标识的属性标志进行训练扩展，所以，本申请可以对这部分用户标识的属性标志进行验证，保证其真实性，从而保证后续据此训练扩展得到的其他用户标识的属性标志可靠。本申请对人工标注的属性标志的验证方式不做限定，可以发送至相应的用户客户端进行验证，也可以发送至相应的户籍平台等相关平台进行验证等等。

[0109] 可选的，对于获取的多个用户标识，可以按照存储顺序检测每一个用户标识是否

标注有对应的属性标志，若是，将其作为一个第一用户标识；若否，继续检测下一个用户标识，直至筛选出多个用户标识中所有的标注有属性标志的第一用户标识。但并不局限于本申请描述的这一种筛选方式。

[0110] 步骤S103，将该至少一个第一用户标识关联的历史行为数据构成第一数据集，并将除第一数据集之外的其他历史行为数据构成第二数据集；

[0111] 结合上述分析，本申请是基于标注有属性标志的用户标识关联的行为数据记性训练扩展，所以，本实施例可以将确定的至少一个第一用户标识关联的行为数据作为第一数据集，将获得的其他行为数据作为第二数据集。

[0112] 步骤S104，利用第一数据集，按照协同训练算法，获得第二数据集中多个第二用户标识对应的属性标志。

[0113] 在本申请中，可以利用协同训练算法实现对标注有属性标志的行为数据的训练扩展，具体可以使Co-Training方法，但并不局限于此，其实际上是一种半监督方法，利用少量已标记样本，通过两个或多个模型去学习，对未标记样本进行标记，从而选择出最优的样本加入已标记样本阵营。

[0114] 可见，本申请将利用人工标注了属性标志的第一数据集，通过协同训练算法给未人工标注属性标志的第二数据集中的行为数据添加标注，从而得到大量标准有属性标志的多个用户标识关联的行为数据，且保证了该标注过程可靠且准确。且不需要人工标注每一个用户标识关联的行为数据，大大降低了人工成本。

[0115] 并且，本申请提供的这种数据处理方案，能够适应于不同应用场景的各种类型的行为数据的处理，从而得到该应用场景下大量标注有属性标志的多个用户标识关联的行为数据，以便据此训练得到可靠地属性预测模型。也就是说，本申请提供的数据处理方案的使用范围很广。

[0116] 参照图2，为本申请实施例提供的另一种数据处理方法的流程图，该方法可以包括：

[0117] 步骤S201，获取多个用户标识关联的历史行为数据；

[0118] 步骤S202，获取历史行为数据的第一视图特征和第二视图特征；

[0119] 在本申请中，由于协同训练算法实现对标注有属性标志的行为数据的训练扩展，这就要求待处理数据集至少存在两个独立的视图，本实施例在此仅以两个独立视图为例进行说明，记为第一视图和第二视图，并利用这两个视图的特性，对获取的多个用户标识关联的历史行为数据进行相应处理，从而得到相应的第一视图特征和第二视图特征。

[0120] 其中，视图是指看待数据的不同角度，如对于用户行为数据即可以从统计学的角度看而得到RFM特征，也可以从行为序列的角度看得到行为序列特征。所以说，本申请获得的两个视图特征是从不同角度看待数据提取的数据的特征。

[0121] 步骤S203，利用第一视图特征和第二视图特征，生成历史行为数据关联用户标识对应的特征数据；

[0122] 在本申请中，可以按照每一个用户标识，对从与该用户标识关联的行为特征中获取的第一视图特征和第二视图特征进行分类处理，若该用户标识已标注了属性标志，还可以对应添加属性标志。可见，本实施例可以生成“用户标识:第一视图特征，第二视图特征，[属性标志]”的形式的特征数据，每一个特征数据对应一个用户标识，通常情况下，生成的

特征数据的数量可以与获取的历史行为数据关联的用户标识的数量相同。

[0123] 步骤S204，验证多个用户标识对应的特征数据中是否包含属性标志，如果是，进入步骤S205；如果否，执行步骤S206；

[0124] 步骤S205，由包含有属性标志的至少一个第一用户标识对应的特征数据构成第一数据集，并对第一数据集进行备份，得到第一备份数据集；

[0125] 步骤S206，由未包含属性标志的多个第二用户标识对应的特征数据构成第二数据集，并对第二数据集进行备份，得到第二备份数据集；

[0126] 可见，本申请经过上述验证步骤后，将具有属性标志的特征数据对应的用户标识作为第一用户标识，并将获取的其他用户标识作为第二用户标识，即将未标注有属性标志的特征数据对应的用户标识作为第二用户标识。在实际应用中，第二用户标识的数量远大于第一用户标识的数量，或者说，未标注有属性标志的历史行为数据的数量，远远大于标注有属性标志的历史行为数据的数量。

[0127] 步骤S207，利用第一数据集，按照第一类算法，对第二数据集中多个第二用户标识关联的特征数据进行属性预测，确定至少一个第二用户标识对应的属性标志，更新第一备份数据集中与确定的属性标志对应的第二用户标识的特征数据；

[0128] 步骤S208，利用第一备份数据集，按照第二类算法，对第二备份数据集中多个第二用户标识关联的特征数据进行属性预测，确定至少一个第二用户标识对应的属性标志，更新第一数据集中与确定的属性标志对应的第二用户标识的特征数据；

[0129] 由此可见，本申请利用第一类算法进行属性预测，确定至少一个第二用户标识对应的属性标志后，会将新确定的第二用户标识的属性标志添加到第一备份数据集中，与相应的第二用户标识关联。同理，利用第二类算法进行属性预测后，也会将新确定的第二用户标识的属性标志添加到第一备份数据集中，与相应的第二用户标识关联，从而提高对第二用户标识标注属性标志的可靠性。

[0130] 步骤S209，基于更新后的第一数据集以及更新后的第一备份数据集，确定多个第二用户标识对应的属性标志；

[0131] 可选的，本申请可以通过验证更新后的第一数据集以及更新后的第一备份数据集中，同一用户标识对应的属性标志是否相同，如果相同，将该属性标志确定为对应第二用户标识的目标属性标志；如果不相同，可以利用第一类算法和第二类算法重新对特征数据进行属性预测，并选择置信度较高的属性标志确定为对应第二用户标识的目标属性标志。

[0132] 步骤S210，利用分类算法对得到的包含有属性标志的特征数据进行处理，得到属性预测模型；

[0133] 其中，分类算法可以包括GBDT(Gradient Boosting Decision Tree，迭代决策树)、逻辑回归、随机森林等机器学习算法，本申请对分类算法具体内容不作限定，且对于各分类算法的数据处理过程可以基于各自工作原理确定，本实施例在此不作详述。

[0134] 需要说明的是，利用不同分类算法训练得到的属性预测模型的格式可以不同，但都能够实现对待测数据集中用户属性的预测。

[0135] 步骤S211，获取目标用户标志关联的待测行为数据；

[0136] 步骤S212，利用待测行为数据对属性预测模型进行模型训练，得到目标用户标识对应的目标属性标志。

[0137] 综上，本实施例提出利用协同训练算法，为未标注属性标志的行为数据准确添加合适的属性标志，无需人工对用户的行为数据标注属性标志，即可得到大量标注有属性标志的样本数据，降低了标注成本，且在标注属性标志过程中，本实施例是通过两个角度分别训练模块，使得学习的特征数据多样，互补不足，且避免了相互之间的干扰，有效保证了得到的大量用户属性标志的准确性和可靠性，进而保证了训练得到的属性预测模型能够准确预测目标用户的属性标志，即提高了属性预测的准确性以及效率。

[0138] 参照图3，为本申请实施例提供的又一种数据处理方法实施例的流程图，该方法可以包括：

[0139] 步骤S301，获取多个用户标识关联的历史行为数据；

[0140] 步骤S302，获取该历史行为数据中多种行为类型对应的RFM(Recency Frequency Monetary)特征；

[0141] 在本实施例中，该RFM特征可以指上述实施例中的第一视图特征。

[0142] 其中，RFM是一种用户特征分析方法，通常情况下，R(Recency，最近一次消费)表示用户上次产生制定行为距今的时长，如上一次购物或使用APP距今的时长等；F(Frequency，消费频率)表示用户在限定时间范围内产生制定行为的次数，如近三个月内购物次数、使用APP次数等等；M(Monetary，消费金额)表示用户在限定时间内产生指定行为带来的收益，如近三个月内消费金额、近三个月使用APP时长等。

[0143] 基于此，在本申请中，为了增加模型训练的样本数据量，提高预测所得属性标志的可靠性，对上述传统的RFM算法进行增强扩展，得到的了增强型RFM算法，具体的，扩展后得到的增强型R表示不同业务类型用户产生制定行为距今的时长，以及无业务限制的情况下用户上次产生制定行为距今的时长。增强型F表示统计不同时长粒度、不同日期类型、不同时间段、不同业务类型下产生制定行为的次数；增强型M表示统计不同时长粒度、不同日期类型、不同时间段、不同业务类型下产生指定行为带来的收益。

[0144] 可见，本申请获取的RFM特征中的R特征、F特征以及M特征可以是通过多个维度确定的数据。需要说明的是，上述R特征、F特征以及M特征表示的多个维度并不局限于上文列举的几个维度。

[0145] 可选的，对于上述各维度，时长粒度可以包括：近一周、近一个月、近三个月、近半年、近一年或者全部等；日期类型可以包括：工作日、休息日；时间段可以包括：深夜、凌晨、早上、上午、中午、下午、傍晚、晚上以及夜晚等；业务类型可以包括：APP划分的工作学习、休闲娱乐、辅助工具、出行帮助等类型。

[0146] 举例说明，如下表1示出了一种时间粒度p、日期类型t以及时间段s的划分方式，但并不限于表1所示的划分方式。

[0147] 表1

[0148]

时长粒度（p）	日期类型(t)	时间段(s)
近一周 近一个月 近三个月 近半年 近一年 全部	工作日 休息日	深夜(0:00~2:00) 凌晨(2:00~5:30) 早上(5:30~9:30) 上午(9:30~11:30) 中午(11:30~14:30) 下午(14:30~17:30) 傍晚(17:30~19:30) 晚上(19:30~22:00) 夜晚(22:00~24:00)

[0149] 基于上述对增强型RFM的描述，本申请可以对历史行为数据进行处理，得到增强型R特征(如下表2所示)、增强型F特征(如下表3所示)以及增强型M特征(如下表4所示)。需要说明的是，关于增强型RFM特征的表示形式并不局限于下表示出的方式。

[0150] 表2

[0151]

Tag0	Tag1	Tag2	…	TagN
R-0	R-1	R-2	…	R-N

[0152] 表3

[0153]

Tag1/(p1,t1,s1)	Tag1/(p1,t1,s2)	...	Tag1/(pn, tm, sd)	…	TagN/(p1,t1, s1)	TagN/(p1, t1,s2)	...	TagN/(pn, tm, sd)
F1-111	F1-112	...	F1-nmd	…	FN-111	FN-112	...	FN-nmd

[0154] 表4

[0155]

Tag1/(p1,t1,s1)	Tag1/(p1,t1,s2)	...	Tag1/(pn, tm, sd)	…	TagN/(p1,t1, s1)	TagN/(p1, t1,s2)	...	TagN/(pn, tm, sd)
D1-111	D1-112	...	D1-nmd	…	DN-111	DN-112	...	DN-nmd

[0156] 上述表格中，Tag表示业务类型，N表示业务类型数，n表示时间段p类别数，m表示时长t粒度数，d表示日期s类型数。

[0157] 基于上述各表的特征，经整理可以得到如下形式的对应于每一个用户标识对应的RFM特征，本实施例中，用户标识可以指用户id，但并不局限于此。

[0158] [user_id,//用户标识；

[0159] R-0,R-1,R-2,...,R-N,//R特征；

[0160] F1-111,F1-112,...,FN-nmd,//F特征；

[0161] D1-111,D1-112,...,DN-nmd]//M特征；

[0162] 步骤S303，获取该历史行为数据中预设时间段内至少一种行为类型对应的行为编

码；

[0163] 在本申请中，除了采用基于RFM特征的算法外，本申请还可以采用基于行为序列的算法，实现对用户行为的分析。行为序列可以是基于时间序列的用户行为，具体可以是一段时间内，按照时间先后顺序记录的人从事某种活动的每一步行为。

[0164] 基于此，对于获取的多个用户标识对应的历史行为数据，本申请可以对预设时间段内的历史行为数据进行分类处理，即按照用户行为进行分类，并对用户行为的不同类型进行编码，以便通过编码确定相应类型的行为数据。

[0165] 可选的，不同行为类型的行为编码可以参照下表5，但并不局限于表5所示的表示形式。

[0166] 表5

[0167]

行为1	行为2	...	行为N
01	02	...	N

[0168] 举例说明，通过对获取的用户的历史行为数据的分析，确定用户A最近一段时间内产生了四个行为，玩游戏、玩游戏、看视频、玩游戏，则按照上述分析得到其行为序列可以为表6。可见，对于多个相同行为，在确定行为序列时，将自动化为一类行为，而且，当继续产生其他行为时，将在此基础上继续排序03、04、…、N。所以，本申请可以通过查询生成的行为编码，准确得知获取的历史行为数据中包含多少种行为，简单明了。

[0169] 表6

[0170]

玩游戏	看视频	...	买东西
01	02	...	N

[0171] 步骤S304，利用获取的行为编码，确定历史行为数据关联的多个用户标识的行为序列。

[0172] 在本实施例实际应用中，可以讲获取的历史行为数据包含的多个行为类型对应的行为编码，按照先后顺序排列，得到对应的行为序列。

[0173] 可选的，在本申请中，可以获取的历史行为数据中，按照用户行为的不同类型进行分类编码，从而获得用户最近一定数量(如100等)的行为序列，其中，可选的，本申请使用行为序列训练预测模型时，可以采用最长公共子序列、编辑距离等算法，计算两条记录之间的相似性。

[0174] 步骤S305，利用获取的RFM特征以及对应的行为序列，生成上述多个用户标识对应的特征数据；

[0175] 结合上述实施例步骤S203部分的描述，生成的特征数据中，每条记录格式可以是“用户标识:RFM特征矢量，行为序列，[属性标志]”的形式的数据，其中，属性标志只有标注了属性标志的记录有，也就是说，对于未标注属性标志的用户标识关联的历史行为数据得到的特征数据，每条记录格式可以是“用户标识:RFM特征矢量，行为序列”。

[0176] 步骤S306，验证多个用户标识对应的特征数据中是否包含属性标志，如果是，进入步骤S307；如果否，执行步骤S308；

[0177] 步骤S307，由包含有属性标志的至少一个第一用户标识对应的特征数据构成第一数据集，并对第一数据集进行备份，得到第一备份数据集；

[0178] 步骤S308，由未包含属性标志的多个第二用户标识对应的特征数据构成第二数据

集,并对第二数据集进行备份,得到第二备份数据集;

[0179] 在本申请中,对于第二数据集和第二备份数据集,可以从种随机选取部分(如100条)特征数据,分别发送至下文的第一预测模型和第二预测模型,进行属性预测,

[0180] 步骤S309,按照第一类算法,对第一数据集中的RFM特征进行训练,生成第一预测模型;

[0181] 在本申请中,第一类算法可以是SVM(Support Vector Machine,支持向量机)、逻辑回归、随机森林等分类算法,不同分类算法对第一数据集中的RFM特征训练过程可以不同,且使用RFM特征训练基于特征向量的分类模型即第一预测模型的表示形式也会有所差异,具体可以根据各分类算法的原理确定,本实施例在此不作详述。步骤S310,利用第一预测模型,对第二数据集中的多个第二用户标识关联的特征数据进行计算,获得多个第二用户标识对应的第一属性标志;

[0182] 如上述描述,本申请可以将第二数据集中随机选取的部分特征数据,发送至第一预测模型进行属性预测,如此循环,直至将第二数据集中的特征数据依次都发送至第一预测模型进行属性预测,或者循环次数达到预设次数。

[0183] 步骤S311,判断该第一属性标志的置信度是否大于第一阈值,如果是,进入步骤S312;如果否,执行步骤S313;

[0184] 其中,第一阈值是根据经验设定的,可以根据模型效果进行调整,本申请对其具体数值不作限定。

[0185] 步骤S312,利用该第一属性标志标注对应的第二用户标识关联的特征数据,并将标注后的特征数据添加到第一备份数据集;

[0186] 步骤S313,将该第一属性标志对应特征数据放回第二数据集;

[0187] 在本申请,对于第一预测模型预测得到的属性标志,并未直接添加到对应的特征数据中,而是对其置信度进行判断,从而选择置信度较高的属性标志添加到第一备份数据集中对应的特征数据,保证添加的属性标志可靠。

[0188] 而对于置信度不高的属性标志,将放回第二数据集,再重新从剩余第二数据集中选取一部分特征数据发送至第一预测模型。

[0189] 步骤S314,按照第二类算法,对第一备份数据集中的行为序列进行训练,生成第二预测模型;

[0190] 其中,第二类算法可以是knn算法(k-Nearest Neighbor algorithm,k最邻近结点算法)等。

[0191] 步骤S315,利用该第二预测模型,对第二备份数据集中的多个第二用户标识关联的特征数据进行计算,获得多个第二用户标识对应的第二属性标志;步骤S316,判断该第二属性标志的置信度是否大于第二阈值,如果是,进入步骤S317;如果否,执行步骤S308;

[0192] 其中,第二阈值也是人工设定的,可以根据第二预测模型的效果进行调整,其与第一阈值可以相同,也可以不同,本申请对两者数值不作限定。

[0193] 步骤S317,利用第二属性标志标注对应的第二用户标识关联的特征数据,并将标准后的特征数据添加到第一数据集;

[0194] 可见,本申请将第二备份数据集中未标注属性标志的特征数据,通过上述方式添加基本属性标志后,会将添加了属性标志的特征数据添加到第一备份数据集中,达到增大

标注属性标志的特征数据的规模的目的。

[0195] 步骤S318,将该第二属性标志对应的特征数据放回第二备份数据集;

[0196] 可选的,对于第二备份数据集的处理可以按照上述对第二数据集的处理,即选择部分第二备份数据集中的特征数据发送至第二预测模型,并将得到的置信度不高的行为序列对应的特征数据放回第二备份数据集,重新选择一部分特征数据继续发送至第二预测模块进行属性预测。

[0197] 步骤S319,验证更新后的第一数据集与更新后的第一备份数据集中,同一用户标识对应的第三属性标志是否相同,如果是,进入步骤S320;如果否,执行步骤S321;

[0198] 需要说明的是,关于本实施例中的第一属性标志、第二属性标志和第三属性标志,并不存在排序含义,其中的第一、第二和第三,是为了方便描述整个技术方案添加的。

[0199] 步骤S320,将该第三属性标志确定为对应第二用户标识的目标属性标志;

[0200] 这种情况下,第一属性标志与第二属性标志相同,均可作为第三属性标志。

[0201] 步骤S321,将对应的第二用户标识关联的特征数据,以及当前第二数据集和所述第二备份数据集中未标注属性标志的特征数据,确定为待定数据集;

[0202] 需要说明的是,若当前第二数据集和/或第二备份数据集为空集,可以不作处理,若都不为空集,可以将两者不重复的并集放入待定数据集中。

[0203] 步骤S322,利用第一类算法和第二类算法,对待定数据集中的特征数据进行属性预测;

[0204] 其中,关于对待定数据集中特征数据的模型预测过程,可以参照上述实施例相应部分的描述,本实施例在此不再详述。

[0205] 步骤S323,筛选符合预设置信度要求的属性标志,确定为相应特征数据的目标属性标志。

[0206] 可选的,对于两个预测模型对同一特征数据进行属性预测,得到的两个属性标志,可以分别计算这两个属性标志的置信度,之后,选择置信度较高的属性标志作为该特征数据的属性标志,但并不局限与此。

[0207] 需要说明的是,本申请对各预测模型得到的属性标志的置信度的计算方法不作限定。

[0208] 步骤S323,将具有目标属性标志的特征数据构成样本数据集。

[0209] 可选的,关于利用样本数据集训练得到属性预测模型,以及利用该属性预测模型实现对待测数据的属性预测过程,可以参照上述实施例相应部分的描述,本实施例在此不再详述。

[0210] 综上,本实施例利用两种基于不同特征的算法,分别对标注有属性标志的特征数据中的RFM特征和行为序列进行模型训练,得到相应的两个预测模型,从而实现对未标注属性标志的特征数据的属性预测,并选择置信度较高的属性标志添加到对应特征数据,进而添加到由标注有属性标志的特征数据构成的数据集中,而对于置信度较低的属性标志对应的特征数据,以及未标注属性标志的特征数据,还可以利用这两种算法再次进行属性预测,选择置信度较高的属性标志添加到特征数据中,进一步扩展由标注有属性标志的特征数据构成的数据集的规模。可见,本申请利用有限的标注有属性标志的特征数据,经过上述方式的训练扩展,得到了大量可靠的未标注属性标志的特征数据的属性标志,节省了人工标注

成本，且提高了属性标志的可靠性以及准确性，由于这使属性预测模型具有了大量可靠的样本数据，保证了属性预测模型的可靠性，进而提高了利用属性预测模型对待测数据进行属性预测所得结果的可靠性以及准确性，且提高了属性预测效率。

[0211] 如图4所示，为本申请实施例提供的一种数据处理装置的结构框图，该装置可以包括：

[0212] 第一获取模块41，用于获取多个用户标识关联的历史行为数据；

[0213] 筛选模块42，用于筛选当前标注有属性标志的至少一个第一用户标识；

[0214] 数据集构成模块43，用于将所述至少一个第一用户标识关联的历史行为数据构成第一数据集，并将除所述第一数据集之外的其他历史行为数据构成第二数据集；

[0215] 训练模块44，用于利用所述第一数据集，按照协同训练算法，获得所述第二数据集中多个第二用户标识对应的属性标志。

[0216] 其中，关于本实施例中，上述各模型的功能实现过程可以参照上述方法实施例相应部分的描述，本实施例在此不再详述。

[0217] 可选的，如图5所示，该装置还可以包括：

[0218] 第二获取模块45，用于获取所述历史行为数据的第一视图特征和第二视图特征；

[0219] 第一生成模块46，用于利用所述第一视图特征和所述第二视图特征，生成所述历史行为数据关联用户标识对应的特征数据；

[0220] 相应的，所述筛选模块42可以包括：

[0221] 第一验证单元421，用于验证多个用户标识对应的特征数据中是否包含属性标志；

[0222] 第一确定单元422，用于基于验证结果，确定具有属性标志的特征数据对应的第一用户标识。

[0223] 可选的，如图6所示，训练模块44可以包括：

[0224] 备份单元441，用于分别对所述第一数据集和所述第二数据集进行备份，得到相应的第一备份数据集和第二备份数据集；

[0225] 第一预测更新单元442，用于利用第一数据集，按照第一类算法，对所述第二数据集中多个第二用户标识关联的特征数据进行属性预测，确定至少一个第二用户标识对应的属性标志，更新所述第一备份数据集中与确定的属性标志对应的第二用户标识的特征数据；

[0226] 第二预测更新单元443，用于利用第一备份数据集，按照第二类算法，对所述第二备份数据集中多个第二用户标识关联的特征数据进行属性预测，确定至少一个第二用户标识对应的属性标志，更新所述第一数据集中与确定的属性标志对应的第二用户标识的特征数据；

[0227] 第二确定单元444，用于基于更新后的第一数据集以及更新后的第一备份数据集，确定所述多个第二用户标识对应的属性标志。

[0228] 进一步，如图7所示，上述第一预测更新单元442可以包括：

[0229] 第一模型生成单元4421，用于按照第一类算法，对所述第一数据集中的第一视图特征进行训练，生成第一预测模型；

[0230] 第一计算单元4422，用于利用所述第一预测模型，对所述第二数据集中的多个第二用户标识关联的特征数据进行计算，获得所述多个第二用户标识对应的属性标志；

[0231] 第一判断单元4423，用于判断所述属性标志的置信度是否大于第一阈值；

[0232] 第一更新单元4424，用于基于判断结果，利用大于所述第一阈值的置信度对应的属性标志，标注对应的用户标识关联的特征数据，并将标注后的特征数据更新到所述第一备份数据集。

[0233] 并且，参照图7，该第二预测更新单元443可以包括：

[0234] 第二模型生成单元4431，用于按照第二类算法，对所述第一备份数据集中的第二视图特征进行训练，生成第二预测模型；

[0235] 第二计算单元4432，用于利用所述第二预测模型，对所述第二备份数据集中的多个第二用户标识关联的特征数据进行计算，获得所述多个第二用户标识对应的属性标志；

[0236] 第二判断单元4433，用于判断所述属性标志的置信度是否大于第二阈值；

[0237] 第二更新单元4434，用于基于判断结果，利用大于所述第二阈值的置信度对应的属性标志，标注对应的用户标识关联的特征数据，并将标注后的特征数据更新到所述第一数据集。

[0238] 另外，如图7所示，第二确定单元444可以包括：

[0239] 第一验证单元4441，用于验证更新后的第一数据集与更新后的第一备份数据集中，同一用户标识对应的属性标志是否相同；

[0240] 第三确定单元4442，用于所述第一验证单元的验证结果为是时，将所述属性标志确定为对应用户标识的目标属性标志；

[0241] 第四确定单元4443，用于所述第一验证单元的验证结果为否时，将对应的用户标识关联的特征数据，以及当前所述第二数据集和所述第二备份数据集中未标注属性标志的特征数据，确定为待定数据集；

[0242] 属性预测单元4444，用于利用所述第一类算法和所述第二类算法，对所述待定数据集中的特征数据进行属性预测；

[0243] 筛选单元4445，用于筛选符合预设置信度要求的属性标志，确定为相应特征数据的目标属性标志。

[0244] 可选的，如图8所示，第二获取模块45可以包括：第一获取单元451，用于获取所述历史行为数据中多种行为类型对应的RFM特征数据；

[0245] 第二获取单元452，用于获取所述历史行为数据中预设时间段内至少一种行为类型对应的行为编码；

[0246] 第五确定单元453，用于利用获取的行为编码，确定所述历史行为数据关联的用户标识的行为序列。

[0247] 综上所述，本实施例提出利用协同训练算法，为未标注属性标志的行为数据准确添加合适的属性标志，无需人工对用户的行为数据标注属性标志，即可得到大量标注有属性标志的样本数据，降低了标注成本，且在标注属性标志过程中，本实施例是通过两个角度分别训练模块，使得学习的特征数据多样，互补不足，且避免了相互之间的干扰，有效保证了得到的大量用户属性标志的准确性和可靠性，进而保证了训练得到的属性预测模型能够准确预测目标用户的属性标志，即提高了属性预测的准确性以及效率。

[0248] 上面主要是通过功能结构来描述数据处理装置的结构，下面将从硬件组成上描述电子设备的结构。

[0249] 如图9所示，为本申请实施例提供的一种电子设备的硬件结构图，该电子设备可以包括：通信端口91、存储器92以及处理器93，其中：

[0250] 通信端口91可以为通信模块的接口，如GSM模块或WIFI模块的接口等，用于获取本地存储的历史行为数据，或者接收第三方应用平台发送的多个用户标识关联的历史行为数据。

[0251] 存储器92，用于存储实现上述数据处理方法的多个指令，还可以存储电子设备通信期间产生的数据信息，因此，存储器92可以分为存储程序区和存储数据区。其中，存储程序区可以存储操作系统、至少一个功能所需的应用程序(如上述多个指令构成的应用程序等)等；存储数据区可以存储电子设备使用过程中产生的各种数据，以及接收到其他设备传输的数据信息等等。

[0252] 可选的，存储器92可能包含高速RAM存储器，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、闪存器件或其他易失性固态存储器件等。

[0253] 处理器93是电子设备的控制中心，利用各种接口和线路连接整个电子设备部的各个部分，通过运行或执行存储器92存储的软件程序和/或模块，调用存储器92内的数据，还可以对接收或发送的数据信息进行处理，实现电子设备的各种功能。

[0254] 可选的，处理器93可以是一个中央处理器CPU，或者是特定集成电路ASIC(Application Specific Integrated Circuit)，或者是被配置成实施本发明实施例的一个或多个集成电路。

[0255] 在本申请中，处理器93可以用于加载并执行存储器92存储的多个指令，包括：

[0256] 获取多个用户标识关联的历史行为数据；

[0257] 筛选当前标注有属性标志的至少一个第一用户标识；

[0258] 将所述至少一个第一用户标识关联的历史行为数据构成第一数据集，并将除所述第一数据集之外的其他历史行为数据构成第二数据集；

[0259] 利用所述第一数据集，按照协同训练算法，获得所述第二数据集中多个第二用户标识对应的属性标志。

[0260] 需要说明的是，关于处理器83的数据处理过程可以参照上述方法实施例相应部分的描述，本实施例在此不再详述。

[0261] 另外，对于电子设备中的通信端口81、存储器82以及处理器83可以通过通信总线进行通信，且电子设备除了上述列举的硬件结构外，还可以包括显示器、各种传感器等硬件结构，本申请不再一一详述。

[0262] 可见，本申请提供的电子设备，利用协同训练算法对少量标注属性标志的行为数据进行训练扩展，对大量未标注属性标志的多个用户标识关联的历史行为数据，准确添加可靠的属性标志，从而得到大量用户训练属性预测模型的可靠样本数据，保证所得属性预测模型的可靠性，进而提高了对待测数据的用户的属性预测准确性以及效率。

[0263] 最后，需要说明的是，关于上述各实施例中，诸如第一、第二等之类的关系术语仅仅用来将一个操作、单元或模块与另一个操作、单元或模块区分开来，而不一定要求或者暗示这些单元、操作或模块之间存在任何这种实际的关系或者顺序。而且，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法或者系统不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种

过程、方法或者系统所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法或者系统中还存在另外的相同要素。

[0264] 本说明书中各个实施例采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似部分互相参见即可。对于实施例公开的装置和电子设备而言，由于其与实施例公开的方法对应，所以描述的比较简单，相关之处参见方法部分说明即可。

[0265] 专业人员还可以进一步意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、计算机软件或者二者的结合来实现，为了清楚地说明硬件和软件的可互换性，在上述说明中已经按照功能一般性地描述了各示例的组成及步骤。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

[0266] 结合本文中所公开的实施例描述的方法或算法的步骤可以直接用硬件、处理器执行的软件模块，或者二者的结合来实施。软件模块可以置于随机存储器(RAM)、内存、只读存储器(ROM)、电可编程ROM、电可擦除可编程ROM、寄存器、硬盘、可移动磁盘、CD-ROM、或技术领域内所公知的任意其它形式的存储介质中。

[0267] 对所公开的实施例的上述说明，使本领域专业技术人员能够实现或使用本发明。对这些实施例的多种修改对本领域的专业技术人员来说将是显而易见的，本文中所定义的一般原理可以在不脱离本发明的精神或范围的情况下，在其它实施例中实现。因此，本发明将不会被限制于本文所示的这些实施例，而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

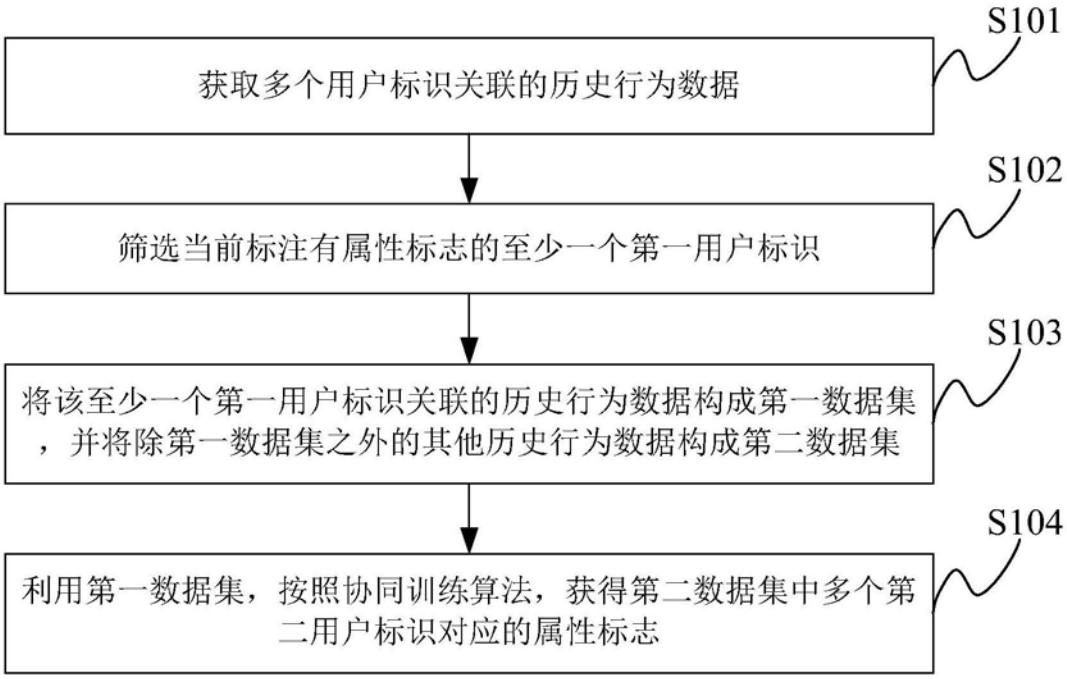

S101
获取多个用户标识关联的历史行为数据
S102
筛选当前标注有属性标志的至少一个第一用户标识
S103
将该至少一个第一用户标识关联的历史行为数据构成第一数据集，并将除第一数据集之外的其他历史行为数据构成第二数据集
S104
利用第一数据集，按照协同训练算法，获得第二数据集中多个第二用户标识对应的属性标志

图1

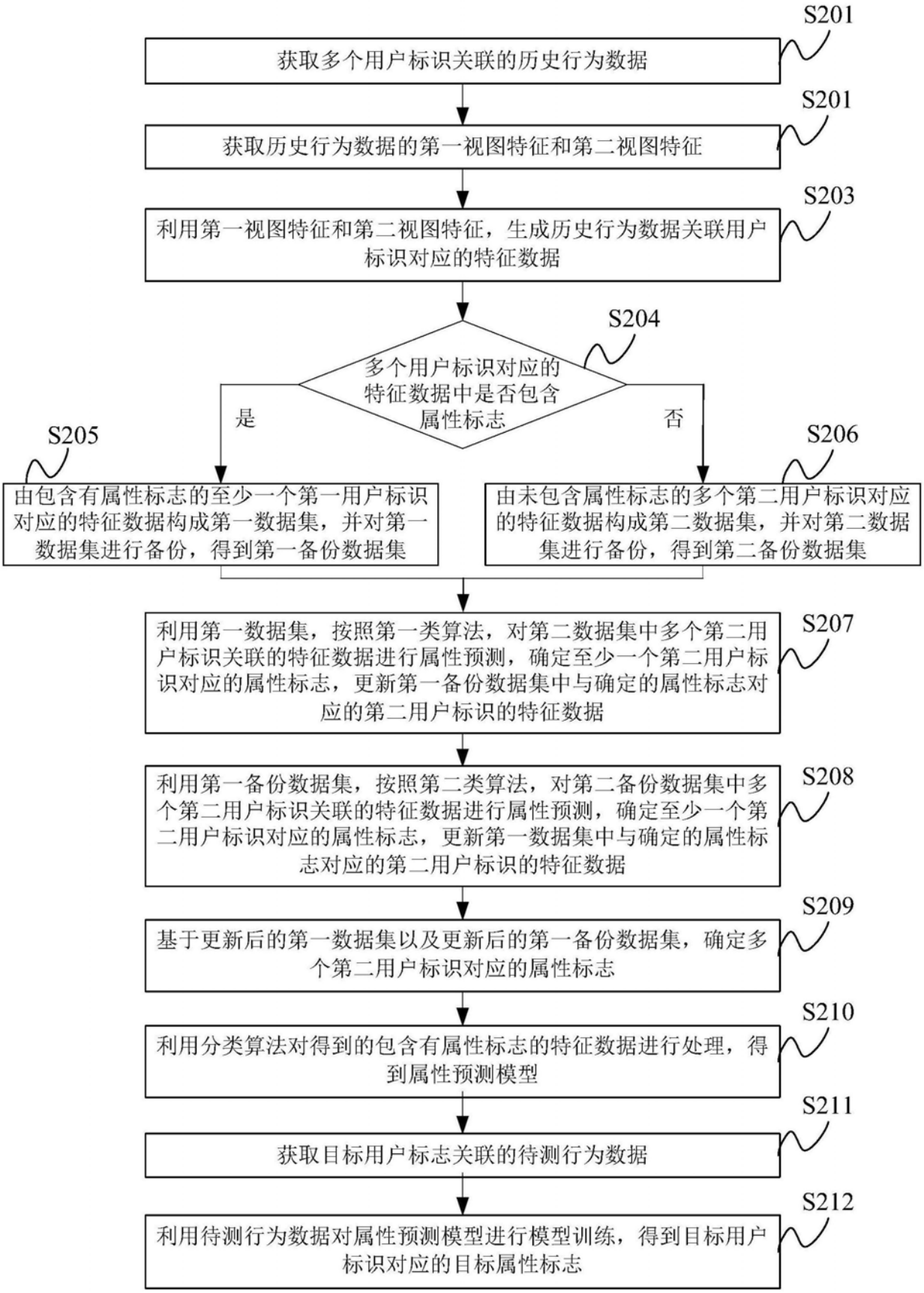

S201
获取多个用户标识关联的历史行为数据
S201
获取历史行为数据的第一视图特征和第二视图特征
S203
利用第一视图特征和第二视图特征，生成历史行为数据关联用户标识对应的特征数据
S204
多个用户标识对应的特征数据中是否包含属性标志
是
否
S205
由包含有属性标志的至少一个第一用户标识对应的特征数据构成第一数据集，并对第一数据集进行备份，得到第一备份数据集
S206
由未包含属性标志的多个第二用户标识对应的特征数据构成第二数据集，并对第二数据集进行备份，得到第二备份数据集
S207
利用第一数据集，按照第一类算法，对第二数据集中多个第二用户标识关联的特征数据进行属性预测，确定至少一个第二用户标识对应的属性标志，更新第一备份数据集中与确定的属性标志对应的第二用户标识的特征数据
S208
利用第一备份数据集，按照第二类算法，对第二备份数据集中多个第二用户标识关联的特征数据进行属性预测，确定至少一个第二用户标识对应的属性标志，更新第一数据集中与确定的属性标志对应的第二用户标识的特征数据
S209
基于更新后的第一数据集以及更新后的第一备份数据集，确定多个第二用户标识对应的属性标志
S210
利用分类算法对得到的包含有属性标志的特征数据进行处理，得到属性预测模型
S211
获取目标用户标志关联的待测行为数据
S212
利用待测行为数据对属性预测模型进行模型训练，得到目标用户标识对应的目标属性标志

图2

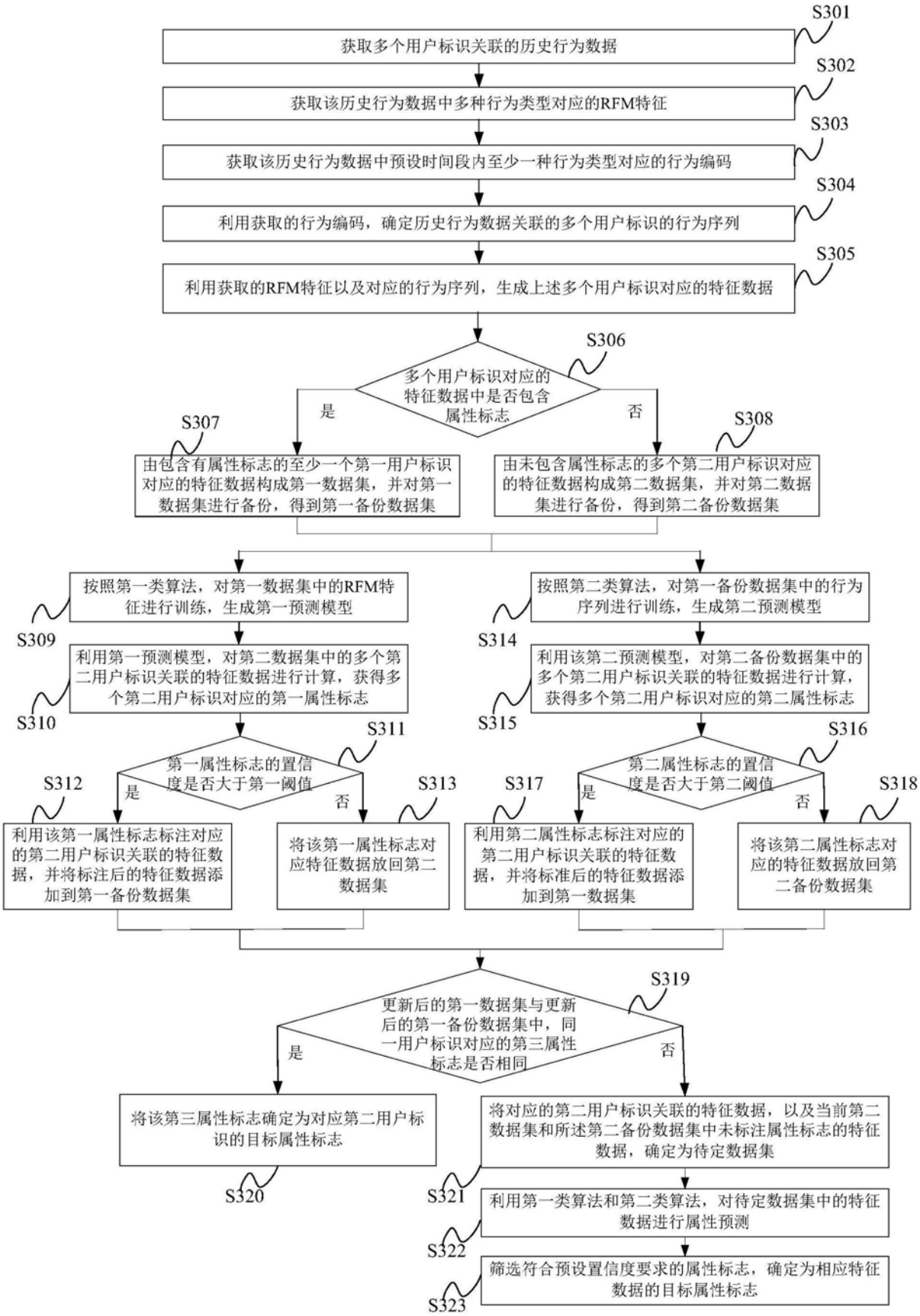

S301
获取多个用户标识关联的历史行为数据
S302
获取该历史行为数据中多种行为类型对应的RFM特征
S303
获取该历史行为数据中预设时间段内至少一种行为类型对应的行为编码
S304
利用获取的行为编码，确定历史行为数据关联的多个用户标识的行为序列
S305
利用获取的RFM特征以及对应的行为序列，生成上述多个用户标识对应的特征数据
S306
多个用户标识对应的特征数据中是否包含属性标志
是
否
S307
由包含有属性标志的至少一个第一用户标识对应的特征数据构成第一数据集，并对第一数据集进行备份，得到第一备份数据集
S308
由未包含属性标志的多个第二用户标识对应的特征数据构成第二数据集，并对第二数据集进行备份，得到第二备份数据集
S309
按照第一类算法，对第一数据集中的RFM特征进行训练，生成第一预测模型
S310
利用第一预测模型，对第二数据集中的多个第二用户标识关联的特征数据进行计算，获得多个第二用户标识对应的第一属性标志
S311
第一属性标志的置信度是否大于第一阈值
是
否
S312
利用该第一属性标志标注对应的第二用户标识关联的特征数据，并将标注后的特征数据添加到第一备份数据集
S313
将该第一属性标志对应特征数据放回第二数据集
S314
按照第二类算法，对第一备份数据集中的行为序列进行训练，生成第二预测模型
S315
利用该第二预测模型，对第二备份数据集中的多个第二用户标识关联的特征数据进行计算，获得多个第二用户标识对应的第二属性标志
S316
第二属性标志的置信度是否大于第二阈值
是
否
S317
利用第二属性标志标注对应的第二用户标识关联的特征数据，并将标准后的特征数据添加到第一数据集
S318
将该第二属性标志对应的特征数据放回第二备份数据集
S319
更新后的第一数据集与更新后的第一备份数据集中，同一用户标识对应的第三属性标志是否相同
是
否
S320
将该第三属性标志确定为对应第二用户标识的目标属性标志
S321
将对应的第二用户标识关联的特征数据，以及当前第二数据集和所述第二备份数据集中未标注属性标志的特征数据，确定为待定数据集
S322
利用第一类算法和第二类算法，对待定数据集中的特征数据进行属性预测
S323
筛选符合预设置信度要求的属性标志，确定为相应特征数据的目标属性标志

图3

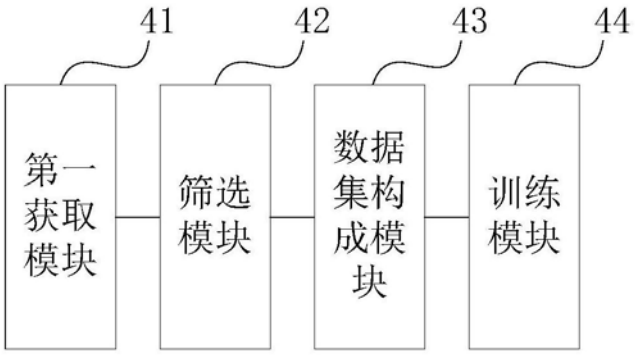

41
42
43
44
第一获取模块
筛选模块
数据集构成模块
训练模块

图4

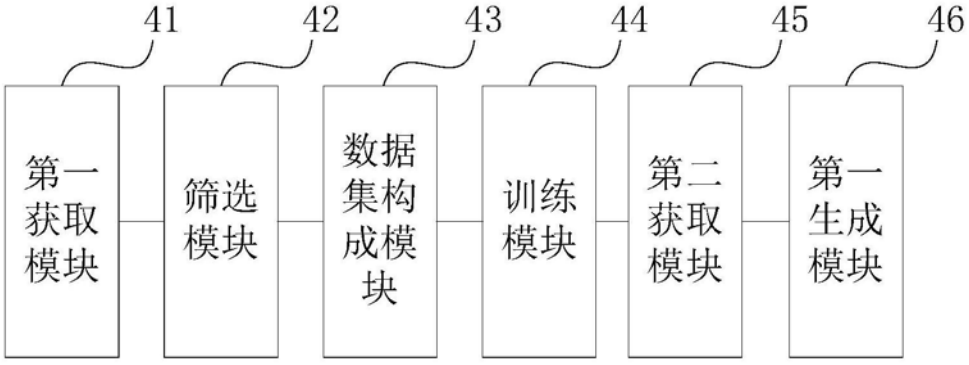

41
42
43
44
45
46
第一获取模块
筛选模块
数据集构成模块
训练模块
第二获取模块
第一生成模块

图5

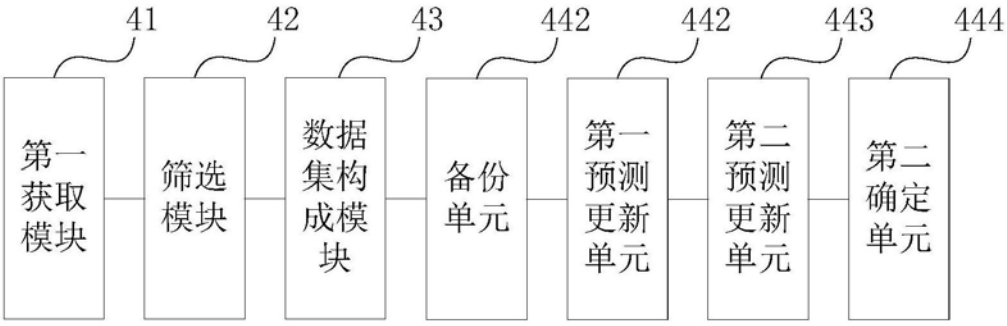

41
42
43
442
442
443
444
第一获取模块
筛选模块
数据集构成模块
备份单元
第一预测更新单元
第二预测更新单元
第二确定单元

图6

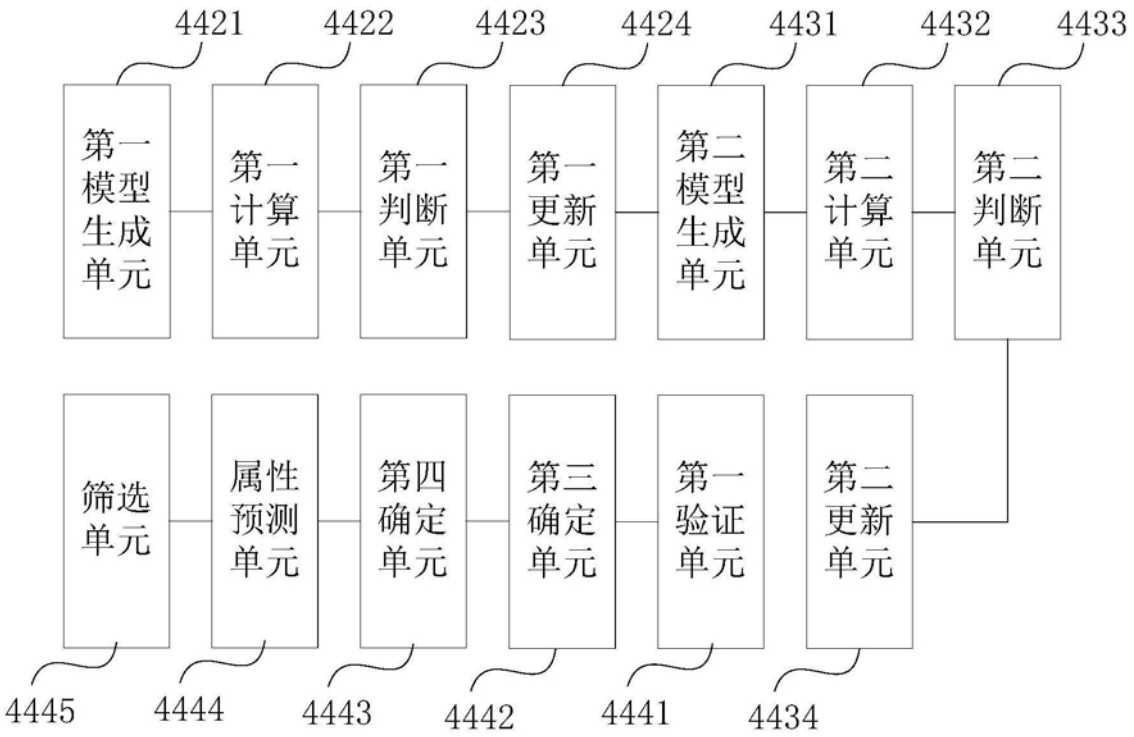

4421
4422
4423
4424
4431
4432
4433
第一模型生成单元
第一计算单元
第一判断单元
第一更新单元
第二模型生成单元
第二计算单元
第二判断单元
筛选单元
属性预测单元
第四确定单元
第三确定单元
第一验证单元
第二更新单元
4445
4444
4443
4442
4441
4434

图7

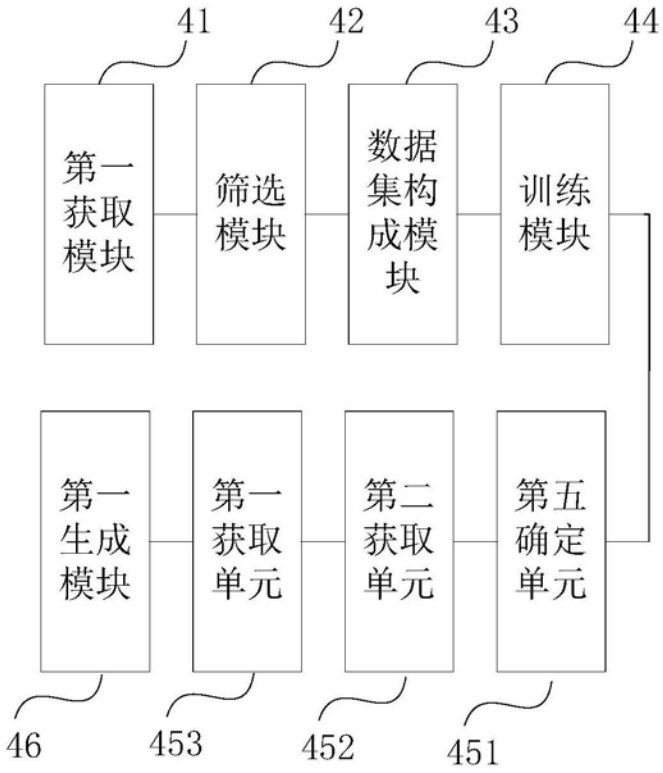

41
42
43
44
第一获取模块
筛选模块
数据集构成模块
训练模块
第一生成模块
第一获取单元
第二获取单元
第五确定单元
46
453
452
451

图8

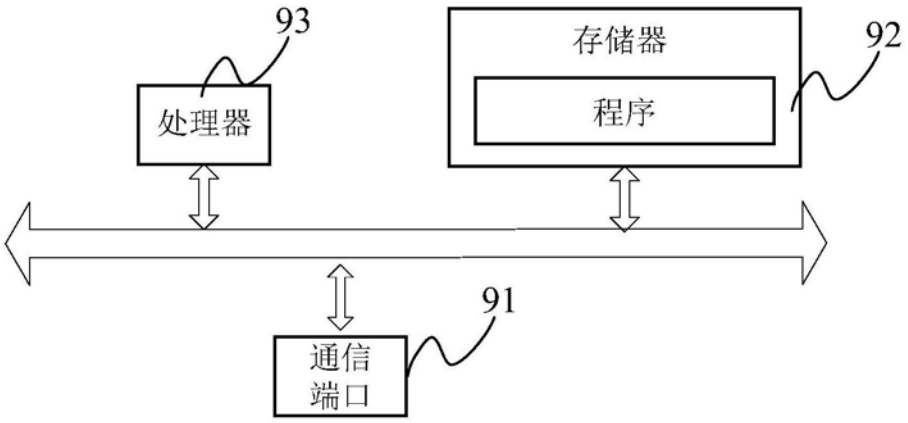


图9