(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
2018 年 11 月 1 日 (01.11.2018)

WIPO | PCT

(10) 国际公布号
WO 2018/196798 A1

(51) 国际专利分类号:
G06F 17/30 (2006.01)

(21) 国际申请号: PCT/CN2018/084565

(22) 国际申请日: 2018 年 4 月 26 日 (26.04.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201710293870.6 2017年4月28日 (28.04.2017) CN

(71) 申请人: 腾 讯 科 技 (深 圳) 有 限 公 司 (TENCENT TECHNOLOGY (SHENZHEN) COMPANY LIMITED) [CN/CN]; 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。

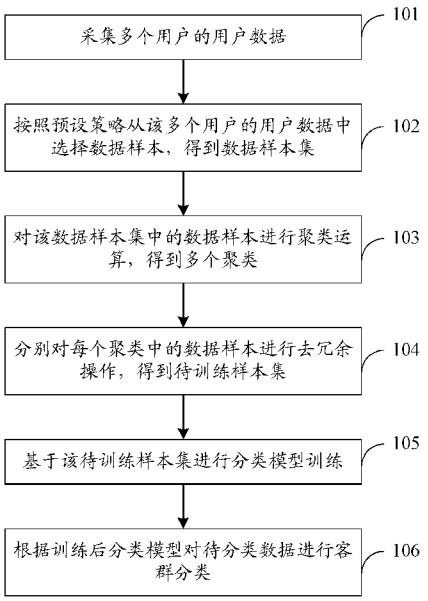
(72) 发明人: 陈玲(CHEN, Ling); 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。 户保田(HU, Baotian); 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。

(74) 代理人: 北 京 德 琦 知 识 产 权 代 理 有 限 公 司 (DEQI INTELLECTUAL PROPERTY LAW CORPORATION); 中国北京市海淀区知春路1号学院国际大厦7层, Beijing 100083 (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,

(54) **Title:** USER GROUP CLASSIFICATION METHOD AND DEVICE

(54) **发明名称**: 用户客群分类方法和装置



101 Collect user data of a plurality of users
102 Select data samples from the user data of the plurality of users according to a preset strategy to obtain a data sample set
103 Perform a clustering operation on data samples in the data sample set to obtain a plurality of clusters
104 Perform a de-redundant operation on the data samples in each cluster respectively to obtain a sample set to be trained
105 Perform classification model training on the sample set to be trained
106 Perform group classification on data to be classified according to a trained classification model

图 1B

(57) **Abstract:** Disclosed are a user group classification method and device. Embodiments of the present application can collect user data of a plurality of users, select data samples from the user data according to a preset strategy, then perform a clustering operation on these data samples, and perform a de-redundant operation on the data samples in each cluster respectively to obtain a sample set to be trained, perform classification model training on the sample set to be trained, and perform group classification on data to be classified according to a trained classification model.

WO 2018/196798 A1

[见续页]

LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, 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, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, 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))。

(57) 摘要：本申请实施例公开了一种用户客群分类方法和装置；本申请实施例可以采集多个用户的用户数据，按照预设策略从这些用户数据中选择数据样本，然后，对这些数据样本进行聚类运算，并分别对每个聚类中的数据样本进行去冗余操作，得到待训练样本集，再基于该待训练样本集进行分类模型训练，并根据训练后分类模型对待分类数据进行客群分类。

用户客群分类方法和装置

本申请要求于 2017 年 04 月 28 日提交中国专利局、申请号为201710293870.6、发明名称为“一种用户客群分类方法和装置”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及通信技术领域，具体涉及一种用户客群分类方法和装置。

背景技术

在许多数据分类应用中，如文本分类、图像分类、以及特殊客群的挖掘等，都需要大量的标注样本来进行分类模型训练，而标注样本通常难以自动获取，一般都需要人工进行标注。如何合理地从数据中选择样本进行标注为业界研究的一个问题。

发明内容

本申请实施例提供一种用户客群分类方法和装置，不仅可以大大加快分类模型的收敛，而且可以提高分类模型的准确率和泛化能力，有利于提高用户客群分类的效率。

本申请实施例提供一种用户客群分类方法，包括:

采集多个用户的用户数据，所述用户数据包括多种用户行为类别的数据;

按照预设策略从所述多个用户的用户数据中选择数据样本，得到数据样本集;

对所述数据样本集中的数据样本进行聚类运算，得到多个聚类;

分别对每个聚类中的数据样本进行去冗余操作，得到待训练样本集;

基于所述待训练样本集进行分类模型训练，并根据训练后分类模型对待分类数据进行客群分类。

相应地，本申请实施例还提供一种用户客群分类装置，包括：处理器和存储器，所述存储器上存储有计算机可读指令，所述计算机可读指令由所述处理器执行以完成以下操作:

采集多个用户的用户数据，所述用户数据包括多种用户行为类别的数据；

按照预设策略从所述多个用户的用户数据中选择数据样本，得到数据样本集；

对所述数据样本集中的数据样本进行聚类运算，得到多个聚类；

分别对每个聚类中的数据样本进行去冗余操作，得到待训练样本集；

基于所述待训练样本集进行分类模型训练，并根据训练后分类模型对待分类数据进行客群分类。

本申请实施例还提供一种用户客群分类方法，用于服务器，所述服务器包括：处理器和存储器，所述方法包括：

采集多个用户的用户数据，所述用户数据包括多种用户行为类别的数据；

按照预设策略从所述多个用户的用户数据中选择数据样本，得到数据样本集；

对所述数据样本集中的数据样本进行聚类运算，得到多个聚类；

分别对每个聚类中的数据样本进行去冗余操作，得到待训练样本集；

基于所述待训练样本集进行分类模型训练，并根据训练后分类模型对待分类数据进行客群分类。

本申请实施例还提供一种计算机可读存储介质，其中，所述计算机可读存储介质中存储有指令，当其在计算机上运行时，使得计算机执行以下操作：

采集多个用户的用户数据，所述用户数据包括多种用户行为类别的数据；

按照预设策略从所述多个用户的用户数据中选择数据样本，得到数据样本集；

对所述数据样本集中的数据样本进行聚类运算，得到多个聚类；

分别对每个聚类中的数据样本进行去冗余操作，得到待训练样本集；

基于所述待训练样本集进行分类模型训练，并根据训练后分类模型对待分类数据进行客群分类。

附图简要说明

为了更清楚地说明本申请实施例中的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请

的一些实施例，对于本领域技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

图1A是本申请实施例提供的用户客群分类方法的场景示意图；

图1B是本申请实施例提供的用户客群分类方法的流程图；

图2A是本申请实施例提供的用户客群分类方法的框架示例图；

图2B是本申请实施例提供的用户客群分类方法的另一流程图；

图3A是本申请实施例提供的用户客群分类装置的结构示意图；

图3B是本申请实施例提供的用户客群分类装置的另一结构示意图；

图4是本申请实施例提供的服务器的结构示意图。

实施本发明的方式

下面将结合本申请实施例中的附图，对本申请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申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

目前已有一些用于数据分类应用的样本标注方法，例如随机选择样本进行标注或者将信息熵或置信度较大的样本挑选出来进行标注。但随机选择样本进行标注可能会使样本分布与预期样本不一致，导致分类模型过拟合或者准确率不高。而使用将信息熵或者置信度较大的样本挑选出来进行标注的方法，所挑选的样本之间的相似性也可能较大，因此，往往会出现样本过于集中的情况，不利于分类模型的收敛，大大影响分类模型的准确率和泛化能力，导致数据分类的效率较低。

本申请实施例提供一种用户客群分类方法和装置。

其中，该用户客群分类装置具体可以集成在服务器，如业务服务器等设备中。例如，以集成在服务器中为例，参见图1A，该服务器可以采集多个用户的用户数据，按照预设策略从该用户数据中选择数据样本，得到数据样本集，然后，对该数据样本集中的数据样本进行聚类运算，得到多个聚类，并分别对每个聚类中的数据样本进行去冗余操作，得到待训练样本集，比如，以多个聚类分别为A类、B类、……N类等为例，则此时，可以分别对A类、B类、……N类等聚类中的数据样本进行去冗余操作，得到去冗余后的多个聚类，如A"类、B"

类、……N"类等，其中，为了描述方便，在本申请实施例中，将这去冗余后的多个聚类统称为待训练样本集；此后，服务器便可以基于该待训练样本集进行分类模型训练，比如，可以对该待训练样本集中的数据样本进行标注，然后，根据标注后的数据样本对初始分类模型进行训练，并根据训练后分类模型对待分类数据进行客群分类。

以下分别进行详细说明。

本申请实施例将从用户客群分类装置的角度进行描述，该用户客群分类装置具体可以集成在服务器，比如业务服务器等设备中。

一种用户客群分类方法，包括：采集多个用户的用户数据，按照预设策略从该多个用户的用户数据中选择数据样本，得到数据样本集，对该数据样本集中的数据样本进行聚类运算，得到多个聚类，分别对每个聚类中的数据样本进行去冗余操作，得到待训练样本集，基于该待训练样本集进行分类模型训练，并根据训练后分类模型对待分类数据进行客群分类。

如图1B所示，该用户客群分类方法的具体流程可以如下：

101、采集多个用户的用户数据。

例如，用户客群分类装置具体可以从网络中采集各种类型用户的用户数据，或者，也可以由其他的采集设备对各种类型用户的用户数据进行采集，然后在需要使用该用户数据时，用户客群分类装置从该采集设备中进行获取，等等。

其中，该用户数据可以包括多种用户行为类别的数据，其中，用户行为类别可以根据用户行为特性来进行划分，比如，与社交相关的数据可以划分为社交信息，与支付相关的数据可以划分为支付行为数据，与娱乐休闲相关的数据可以划分为娱乐休闲行为数据，等等；即该用户数据可以包括用户的帐号信息、社交信息、支付行为数据、娱乐休闲行为数据、教育行为数据、和/或其他网络行为数据，等等，例如，可以如下：

1）用户的帐号信息，比如性别、区域、年龄、和/或通用设置等。

2）用户的社交信息，比如聊天、电子邮件、语音通话、微博空间发表的信息、社区评论、和/或公众号文章阅读等。

3）用户的支付行为数据，比如网上购物、游戏点卡支付、和/或音乐流量包等。

4）娱乐休闲行为数据：如视频点播、音乐播放、K歌、和/或新闻阅读等。

5）教育行为数据：如线上读书、公开课学习、职业考试练习、技能培训、和/或翻译软件使用等.

6）其网络行为数据：如终端应用（App，Application）下载、和/或搜索等。

102、按照预设策略从该多个用户的用户数据中选择数据样本，得到数据样本集。

其中，该预设策略可以根据实际应用的需求进行设置，例如，可以设置为“对用户数据进行清洗”。即步骤“按照预设策略从该用户数据中选择数据样本，得到数据样本集”具体可以包括:

将采集到的用户数据进行清洗，得到数据样本集。即此时可以将采集到的用户数据进行清洗，然后，将清洗后的用户数据作为数据样本，添加至数据样本集中。

其中，清洗的方式可以多种，比如“去除奇异点”、“补齐缺失值”、和/或“去掉重复采集的数据”，等等；例如，可以如下:

（1）去除奇异点;

将该用户数据中的奇异点剔除，并将剩余的用户数据作为数据样本，添加至数据样本集中。

其中，该奇异点指的是“由于数据采集过程中的错误或者其他行为造成的数据可能性取值外的数据”。

（2）补齐缺失值;

根据该用户数据确定本应采集但没有采集到数据，对没有采集到的数据赋予预设值（即赋默认值），然后，将采集到的用户数据和该赋值后的数据作为数据样本，添加至数据样本集中。

（3）去掉重复采集的数据;

将该用户数据中重复采集的数据进行剔除，并将剩余的用户数据作为数据样本，添加至数据样本集中。

在将用户数据进行清洗后，便可以将清洗后的数据，即数据样本集中的数据样本进行归类，并写入预设数据库中，以供后续步骤使用。

103、对该数据样本集中的数据样本进行聚类运算，得到多个聚类；例如，具体可以如下：

获取数据样本的用户行为类别，根据该用户行为类别对该数据样本集中的数据样本进行聚类运算，得到多个聚类。

其中，聚类算法也可以根据实际应用的需求而定，比如，可以采用邻近算法（kNN，k-NearestNeighbor）、或k-means算法（一种硬聚类算法）等。

比如，以用户行为类别包括社交信息和娱乐休闲信息为例，则经过聚类运算后，可以将属于社交信息的数据样本组成一个聚类，以及将属于娱乐休闲信息的数据样本组成一个聚类，以此类推，等等。

104、分别对每个聚类中的数据样本进行去冗余操作，得到待训练样本集。

例如，具体可以分别计算每个聚类中的数据样本的相似度，根据该相似度对每个聚类中的冗余样本进行剔除，得到待训练样本集，具体可以如下：

（1）分别计算每个聚类中的数据样本的相似度，比如，具体可以包括步骤S1、S2和S3，如下：

S1、利用最大化信息量算法分别计算每个聚类中的数据样本的最大信息量。

其中，最大信息量，指的是数据样本所能够提供的信息的最大含量，同一数据样本的最大信息量和信息量的和为1。比如，若数据样本的最大信息量为0.8，则其信息量为0.2，以此类推，等等。其中，信息量指的是从M个相等可能事件中选出一个事件所需要的信息度量或含量，也就是在辨识M个事件中特定的一个事件的过程中所需要提问“是或否”的最少次数。一个事件的可能性愈小（即出现的概率小），不确定性越多，确定其所需获取的信息越多，即信息量愈大；而事件的可能性愈大（即出现的概率大），不确定性越少，则确定其所需获取的信息愈少，即信息量越小。可见，该信息量的多少是与事件发生的概率大小有关，而最大信息量的多少同样与事件发生的概率大小有关。所以，可以通过计算事件发生的概率来计算该最大信息量，进而确定事件的不确定性的大小。其中，最大信息量越小（即信息量越大），不确定性越大，最大信息

量越大（即信息量越小），不确定性越小。比如，在本申请实施例中，事件发生的概率主要指的是“分类模型预测数据样本为某一类”的概率值，因此，在本申请实施例中，可以通过计算该概率值来得到该最大信息量。即，可选地，步骤“利用最大化信息量算法分别计算每个聚类中的数据样本的最大信息量”可以包括:

采用预设的初始分类模型对每个聚类中的数据样本的客群分类进行预测，根据预测结果计算相应数据样本的最大信息量。比如，可以直接根据预测结果（如该用户属于某一分类的概率）计算相应数据样本的最大信息量，或者，也可以计算预测结果的置信度，然后，根据该置信度来计算相应数据样本的最大信息量，等等。这里的置信度=最不确定性，最不确定性的计算可以由下文的公式1来计算（不仅限于此公式），即最不确定的样本置信度越低，反之越高。其中，该初始分类模型可以由维护人员进行预设，也可以由该用户客群分类装置通过训练来得到，比如，可以从每个聚类中随机选择一定数量的数据样本，对其所属的客群分类进行人工标注，构成初始训练样本（这些初始训练样本构成初始训练样本集），然后，利用这些初始训练样本对预设的原始模型进行训练，得到初始分类模型；即，可选地，在步骤“采用预设的初始分类模型对每个聚类中的数据样本的客群分类进行预测”之前，该用户客群分类方法还可以包括:

从每个聚类中随机选择预定数量的数据样本，对其所属的客群分类进行人工标注，根据人工标注后的数据样本对预设的原始模型进行训练，得到初始分类模型。

则此时，步骤“采用预设的初始分类模型对每个聚类中的数据样本的客群分类进行预测”可以包括：采用得到的初始分类模型，对每个聚类中的其他未进行人工标注的数据样本的客群分类进行预测。

其中，训练时所采用的原始模型可以根据实际应用的需求进行设置，而该预定数量也可以根据实际应用的需求进行设置，比如，以需要选择1000个初始训练样本为例，若总共有5个聚类，则每个聚类可以随机挑选200个数据样本作为相应聚类的初始训练样本，以此类推，等等。

例如，以二元分类（即存在两种分类：分类 1 和分类 2）为例，则其相应的最大信息量计算公式可以如下:

$$\chi_{LC}^{*} = \arg\min_{x \in U} \left| \Pr[x = | L] - \frac{1}{2} \right| \quad \text{（公式1）}$$

其中，x为数据样本，L为初始训练样本集（即人工标注后的数据样本），C为一个或一组分类模型（也称为分类器），$\Pr[x = 1 \mid L]$为在初始训练样本集L上训练出来的分类模型C对数据样本预测为“分类1”（其中的“1”表示“分类1”）的概率值，χ_{LC}^{*}为使得$\left| \Pr[x = 1 \mid L] - \frac{1}{2} \right|$取得最小值的数据样本，argmin为求使得$\left| \Pr[x = 1 \mid L] - \frac{1}{2} \right|$取得最小值的数据样本的函数，U为未标注样本集（即该聚类中未标注的数据样本）。

从公式$\chi_{LC}^{*} = \arg\min_{x \in U} \left| \Pr[x = 1 \mid L] - \frac{1}{2} \right|$可以看出，概率值越接近$\frac{1}{2}$的样本越容易被选中，因为越接近$\frac{1}{2}$表示分类模型对数据样本属于哪一类的预测越不能确定。例如聚类后某一聚类A中包括数据样本x_1、x_2和x_3，用分类模型C算出属于分类1的概率分别为0.49、0.1和0.8，则通过公式$\chi_{LC}^{*} = \arg\min_{x \in U} \left| \Pr[x = 1 \mid L] - \frac{1}{2} \right|$算出的最大信息量分别为0.01、0.4和0.3，即x_1的不确定性最大，x_3的不确定性次之，x_2的不确定性最小，按照最不确定性原则，三个样本入选待标注样本集的优先级是：x_1，x_3，x_2。依此类推，在此不再赘述。

S2、根据最大信息量从每个聚类中选取不确定性最大的前K个数据样本，得到待标注样本集。

其中，K为正整数，具体取值可以根据实际应用的需求进行设置，比如，以K为2为例，若在步骤S1中，三个样本入选待标注样本集的优先级是：x_1，x_3，x_2，则此时，可以将x_1和x_3添加至待标注样本集中。

S3、计算待标注样本集中两两数据样本之间的相似度，得到每个聚类中的数据样本的相似度。

例如，若该待标注样本集中包括数据样本x_1、x_3、x_5和x_6，则此时，需要计算x_1和x_3的相似度、x_1和x_5的相似度、x_1和x_6的相似度、x_3和x_5的相似度、x_3和x_6

的相似度、以及x_5和x_6的相似度，等等。

需说明的是，相似度的计算可以先将数据样本的用户信息进行量化，转换为向量空间，然后，用标准化欧氏距离、夹角余弦、以及相关系数等计算出相似度。

此外，还需说明的是，在计算出相似度之后，可以根据得到的相似度设置一个阈值，高于该阈值的则认为两个数据样本高度相似，为了描述方便，在申请实施例中，将该大于该预设阈值的相似度称为高度相似度。其中，该阈值可以根据实际应用的需求进行设置，在此不再赘述。

（2）根据该相似度对每个聚类中的冗余样本进行剔除，得到待训练样本集。

例如，具体可以获取相似度大于预设阈值的数据样本，得到候选集，从候选集中筛选最大信息量满足预设条件的数据样本，得到待训练样本集。

其中，该预设条件可以根据实际应用的需求而定，比如，可以通过统计高度相似计数，并将高度相似计数较多的样本进行剔除，等等。即可选地，步骤“从候选集中筛选最大信息量满足预设条件的数据样本，得到待训练样本集”可以包括:

统计该候选集中每个数据样本所具有的高度相似度（大于预设阈值的相似度）的数量，得到高度相似计数，从该候选集中剔除高度相似计数最高的前N个数据样本，得到待训练样本集。

其中，N为正整数，N的取值可以根据实际应用的需求而定。

105、基于该待训练样本集进行分类模型训练。

例如，具体可以获取预设的初始分类模型，对该待训练样本集中的数据样本进行标注，然后，根据标注后的数据样本对该初始分类模型进行训练，得到训练后分类模型。

106、根据训练后分类模型对待分类数据进行客群分类。

例如，可以获取待分类数据，将该待分类数据导入训练后分类模型中，得到输出结果，根据该输出结果确定该待分类数据的客群分类。

其中，客群指的是客户群体，比如可以划分为“学生”、“白领”和“金领”等，又比如，还可以划分为“儿童”、“青年”、“中年”和“老年”，等等，客群的具体划分需求可以根据实际应用的需求而定，在此不再赘述。

由上可知，本申请实施例在从采集到的用户数据中选择数据样本后，可以对这些数据样本进行聚类运算，并分别对每个聚类中的数据样本进行去冗余操作，得到待训练样本集，然后，才基于该待训练样本集对分类模型进行训练，并根据训练后分类模型对待分类数据进行客群分类；由于该方案在进行分类模型训练之前，先对数据样本进行了聚类，然后，再分别从每个聚类中选择数据样本进行训练，所以，可以保证每个聚类中都有数据样本被选中进入到分类模型的训练样本中，最大程度上使分类模型能从每个聚类中去学习，覆盖样本预期分布，大大提高分类模型的准确率和泛化能力；此外，由于该方案可以对每个聚类中的数据样本进行去冗余操作，以剔除相似度高的冗余数据样本，使得可以尽可能减少标注相似的数据样本，因此，可以大大加快分类模型的收敛；总而言之，该方案不仅可以大大加快分类模型的收敛，而且可以提高分类模型的准确率和泛化能力，有利于提高用户客群分类的效率。

根据前述实施例所描述的方法，以下将举例作进一步详细说明。

在本申请实施例中，将以该用户客群分类装置具体集成在服务器中为例进行说明。

如图2A和图2B所示，一种用户客群分类方法，具体流程可以如下：

201、服务器采集多个用户的用户数据。

例如，服务器具体可以从网络中采集各种类型用户的用户数据，或者，也可以由其他的采集设备对各种类型用户的用户数据进行采集，然后在需要使用该用户数据时，服务器从该采集设备中进行获取，等等。

其中，该用户数据可以包括多种用户行为类别的数据，比如，可以包括用户的帐号信息、社交信息、支付行为数据、娱乐休闲行为数据、教育行为数据、和/或其他网络行为数据等信息。

202、服务器将采集到的用户数据进行清洗，得到数据样本集。

其中，清洗的方式可以多种，比如“去除奇异点”、“补齐缺失值”、和/或“去掉重复采集的数据”，等等；即步骤“服务器将采集到的用户数据进行清洗，得到数据样本集”具体可以如下：

服务器具体可以将该用户数据中的奇异点剔除，并将剩余的用户数据作为数据样本，添加至数据样本集中；

或者，服务器也可以根据该用户数据确定本应采集但没有采集到数据，对没有采集到的数据赋予预设值（即赋默认值），然后，将采集到的用户数据和该赋值后的数据作为数据样本，添加至数据样本集中；

又或者，服务器还可以将该用户数据中重复采集的数据进行剔除，并将剩余的用户数据作为数据样本，添加至数据样本集中，等等。

需说明的是，该数据样本集的数据样本均为未标注所属的客群分类的数据样本。

203、服务器获取数据样本的用户行为类别，根据该用户行为类别对该数据样本集中的数据样本进行聚类运算，得到多个聚类。

例如，服务器具体可以根据该用户行为类别，采用kNN或k-means算法等聚类算法对该数据样本集中的数据样本进行聚类运算，得到多个聚类；为了描述方便，在本申请实施例中，将均以最终聚成5类：聚类A、聚类B、聚类C、聚类D、以及聚类E为例进行说明，比如，具体可参见图2A。

204、服务器从每个聚类中随机选择预定数量的数据样本，对其所属的客群分类进行人工标注，根据人工标注后的数据样本对预设的原始模型进行训练，得到初始分类模型。

其中，训练时所采用的原始模型可以根据实际应用的需求进行设置，而该预定数量也可以根据实际应用的需求进行设置。

例如，如图2A所示，如果总共需要挑选1000个数据样本，则此时，可以分别从聚类A、聚类B、聚类C、聚类D、以及聚类E等5个聚类中随机选择一定数量，如大约200个左右的未标注的数据样本，对其客群分类进行人工标注，构成初始训练样本，然后，利用这些初始训练样本对预设的原始模型进行训练，得到初始分类模型。

需说明的是，为了描述方便，可以将这些初始训练样本所在的集合称为初始训练样本集，该初始训练样本集中可以多个初始训练样本子集，每个初始训练样本子集对应一个聚类，比如，可以包括聚类A对应的初始训练样本子集、聚类B对应的初始训练样本子集、聚类C对应的初始训练样本子集、聚类D对应的初始训练样本子集、以及聚类E对应的初始训练样本子集，等等。其中，聚类A对应的初始训练样本子集可以包括聚类A对应的已标注的数据样本，聚类B对应的初始训练样本子集可以包括聚类B对应的已标注的数据样本，聚类C对应

的初始训练样本子集可以包括聚类C对应的已标注的数据样本，聚类D对应的初始训练样本子集可以包括聚类D对应的已标注的数据样本，而聚类E对应的初始训练样本子集可以包括聚类E对应的已标注的数据样本，等等，具体可参见图2A，在此不再赘述。

205、服务器采用步骤204中得到的初始分类模型对每个聚类中的数据样本的客群分类进行预测，根据预测结果计算相应数据样本的最大信息量。

例如，参见图2A，还是以最终聚成聚类A、聚类B、聚类C、聚类D、以及聚类E为例进行说明，则此时，服务器可以采用初始分类模型分别对聚类A、聚类B、聚类C、聚类D、以及聚类E中的数据样本的客群分类进行预测，然后，根据预测结果分别计算相应数据样本的最大信息量。

比如，以计算聚类A中的数据样本的最大信息量，且用户存在两种分类“学生”和“白领”为例，若聚类A中包括数据样本x_1、x_2和x_3，则此时，服务器可以采用初始分类模型对数据样本x_1、x_2和x_3的客群分类进行预测，若预测数据样本x_1、x_2和x_3均可能属于“学生”，且该属于“学生”的概率分别为0.49、0.1和0.8，则可以通过如下公式$\chi_{LC}^{*} = \arg\min_{x \in U} \left| \Pr[x = 1 \mid L] - \frac{1}{2} \right|$分别算出$x_1$、$x_2$和$x_3$的最大信息量：0.01、0.4和0.3。

其中，L为初始训练样本集（即人工标注后的数据样本），C为一个或一组分类模型，$\Pr[x = 1 \mid L]$为在初始训练样本集L上训练出来的分类模型C对数据样本预测为“学生”的概率值，χ_{LC}^{*}为使得$\left| \Pr[x = 1 \mid L] - \frac{1}{2} \right|$取得最小值的数据样本，U为未标注样本集（即该聚类中未标注的数据样本）。

同理，对于聚类B、C、D和E中的数据样本的最大信息量，也可以采用上述方式进行计算，在此不再赘述。

206、服务器对每个聚类进行筛选，得到待标注样本集。

其中，该待标注样本集可以包括多个待标注样本子集，每个待标注样本子集对应一个聚类，比如，可以包括聚类A对应的待标注样本子集、聚类B对应的待标注样本子集、聚类C对应的待标注样本子集、聚类D对应的待标注样本子集、以及聚类E对应的待标注样本子集，等等。其中，A对应的待标注样本子集

称为聚类A'，B对应的待标注样本子集称为聚类B'，C对应的待标注样本子集称为聚类C'，D对应的待标注样本子集称为聚类D'，E对应的待标注样本子集称为聚类E'，具体可参见图2A。

其中，筛选的方式可以有多种，例如，服务器具体可以根据最大信息量从每个聚类中选取不确定性最大的前K个数据样本，得到待标注样本集。

其中，K为正整数，具体取值可以根据实际应用的需求进行设置。

比如，以K为2为例，若在步骤205中，聚类A中有三个数据样本的最大信息量分别为：0.01、0.4和0.3，则按照不确定性从大到小进行排序，这三个数据样本的次序依次为：x_1，x_3，x_2，由于K为2，因此，此时可以将确定性最大的前两个数据样本，即x_1和x_3添加至待标注样本集中的聚类A'中（即聚类A对应的待标注样本子集中）。

同理，可以采用同样的方式从聚类B、C、D和E中选取最大信息量最大的前K个数据样本，分别添加至待标注样本集中的聚类B'、C'、D'和E'中。

207、服务器计算待标注样本集中两两数据样本之间的相似度，得到每个聚类中的数据样本的相似度。

例如，还是以聚类A为例，若该待标注样本集中（即该待标注样本集的聚类A对应的待标注样本子集中）包括数据样本x_1、x_3、x_5、x_6和x_7，则此时，需要计算x_1和x_3的相似度、x_1和x_5的相似度、x_1和x_6的相似度、x_1和x_7的相似度、x_3和x_5的相似度、x_3和x_6的相似度、x_3和x_7的相似度、x_5和x_6、x_5和x_7、以及x_6和x_7的相似度，即此时：

这个聚类中数据样本x_1的相似度为：x_1和x_3的相似度、x_1和x_5的相似度、x_1和x_6的相似度、以及x_1和x_7的相似度；

聚类A中数据样本x_3的相似度为：x_1和x_3的相似度、x_3和x_5的相似度、x_3和x_6的相似度、以及x_3和x_7的相似度；

聚类A中数据样本x_5的相似度为：x_1和x_5的相似度、x_3和x_5的相似度、x_5和x_6的相似度、以及x_5和x_7的相似度；

聚类A中数据样本x_6的相似度为：x_1和x_6的相似度、x_3和x_6的相似度、x_5和x_6的相似度、以及x_5和x_7的相似度；

聚类A中数据样本x_7的相似度为：x_1和x_7的相似度、x_3和x_7的相似度、x_5和x_7的相似度、以及x_6和x_7的相似度。

需说明的是，在计算相似度时，可以先将数据样本的用户信息进行量化，转换为向量空间，然后，再用标准化欧氏距离、夹角余弦、以及相关系数等计算出相似度。

比如，参见表一，每个聚类中的数据样本的相似度可以如下：

表一：

数据样本	x_1	x_3	x_5	x_6	x_7
x_1	1	0.8	0.2	0.3	0.4
x_3	0.8	1	0.6	0.7	0.8
x_5	0.2	0.6	1	0.9	0.5
x_6	0.3	0.7	0.9	1	0.3
x_7	0.4	0.8	0.5	0.3	1

需说明的是，在计算出相似度之后，可以根据得到的相似度设置一个阈值，高于该阈值的则认为两个数据样本高度相似，称为高度相似度。其中，该阈值可以根据实际应用的需求进行设置，比如，该阈值可以设置为0.6，等等。

208、服务器根据该相似度对每个聚类中的冗余样本进行剔除（即进行去冗余操作），得到待训练样本集。

其中，该待训练样本集可以包括多个待训练样本子集，每个待训练样本子集对应一个聚类，比如，可以包括聚类A对应的待训练样本子集、聚类B对应的待训练样本子集、聚类C对应的待训练样本子集、聚类D对应的待训练样本子集、以及聚类E对应的待训练样本子集，等等。其中，聚类A对应的待训练样本子集称为聚类A"，聚类B对应的待训练样本子集称为聚类B"，聚类C对应的待训练样本子集称为聚类C"，聚类D对应的待训练样本子集称为聚类D", 聚类E对应的待训练样本子集称为聚类E"，具体可参见图2A。

其中，去冗余的方式可以由多种，例如，服务器具体可以获取相似度大于预设阈值的数据样本，得到候选集，然后，统计该候选集中每个数据样本所具有的高度相似度的数量，得到高度相似计数，并从该候选集中剔除高度相似计数最高的前N个数据样本，得到待训练样本集。

其中，N为正整数，N的取值可以根据实际应用的需求而定。

比如，以值大于0.6的相似度称为高度相似度为例，则根据表一可知，数据样本x_1的高度相似计数为1，数据样本x_3的高度相似计数为3，数据样本x_5的高度相似计数为1，数据样本x_6的高度相似计数为2，数据样本x_7的高度相似计数为1，可见，其中数据样本x_3的高度相似计数最高，若N为1，则此时，需要从该候选集中剔除数据样本x_3，以得到待训练样本集，即此时待训练样本集可以包括数据样本x_1、x_5、x_6和x_7。

209、服务器基于该待训练样本集进行分类模型训练。

例如，参见图2A，服务器具体可以获取预设的初始分类模型，对该待训练样本集中的数据样本进行标注，然后，根据标注后的数据样本对该初始分类模型进行训练，得到训练后分类模型。

其中，标注时可以采用人工标注，也可以由该服务器按照预设策略进行标注，在此不再赘述。

需说明的是，原始模型、初始分类模型、以及训练后分类模型分别为同一“分类模型”在不同训练期间的表现形式，因此，为了描述方便，在图2A中，将原始模型、初始分类模型、以及训练后分类模型均表示为“分类模型”，应当理解的是，在不同的时间，该分类模型的表现形式可以有所不同。

210、服务器根据训练后分类模型对待分类数据进行客群分类。

例如，服务器可以获取待分类数据，将该待分类数据导入训练后分类模型中，以进行客群分类预测，然后，根据训练后分类模型输出的结果确定该待分类数据的客群分类，得到分类后数据，具体可参见图2A。

比如，还是以用户存在两种分类“学生”和“白领”为例，则此时，可以将该待分类数据，比如待分类用户的帐号信息、社交信息、支付行为数据、娱乐休闲行为数据、教育行为数据、和/或其他网络行为数据等信息导入训练后分类模型中，然后，根据训练后分类模型输出的结果确定该待分类数据是“学生”对应的数据还是“白领”对应的数据，从而将相应的用户归类到相应的客群，比如“学生”或“白领”中。

由上可知，本申请实施例在从采集到的用户数据中选择数据样本后，可以对这些数据样本进行聚类运算，并分别对每个聚类中的数据样本进行去冗余操作，得到待训练样本集，然后，才基于该待训练样本集对分类模型进行训练，并根据训练后分类模型对待分类数据进行客群分类；由于该方案在进行分类模

型训练之前，先对数据样本进行了聚类，然后，再分别从每个聚类中选择数据样本进行训练，所以，可以保证每个聚类中都有数据样本被选中进入到分类模型的训练样本中，最大程度上使分类模型能从每个聚类中去学习，覆盖样本预期分布，大大提高分类模型的准确率和泛化能力；此外，由于该方案可以对每个聚类中的数据样本进行去冗余操作，以剔除相似度高的冗余数据样本，使得可以尽可能减少标注相似的数据样本，因此，可以大大加快分类模型的收敛；总而言之，该方案不仅可以大大加快分类模型的收敛，而且可以提高分类模型的准确率和泛化能力，有利于提高用户客群分类的效率。

为了更好地实施以上方法，本申请实施例还提供一种用户客群分类装置，该用户客群分类装置具体可以集成在服务器，比如业务服务器等设备中。

所述用户客群分类装置可以包括：处理器和存储器，所述存储器上存储有计算机可读指令，所述计算机可读指令由所述处理器执行以完成以下操作：采集多个用户的用户数据，所述用户数据包括多种用户行为类别的数据；按照预设策略从所述多个用户的用户数据中选择数据样本，得到数据样本集；对所述数据样本集中的数据样本进行聚类运算，得到多个聚类；分别对每个聚类中的数据样本进行去冗余操作，得到待训练样本集；基于所述待训练样本集进行分类模型训练，并根据训练后分类模型对待分类数据进行客群分类。

其中，所述对所述数据样本集中的数据样本进行聚类运算，得到多个聚类，包括：获取数据样本的用户行为类别；根据所述用户行为类别对所述数据样本集中的数据样本进行聚类运算，得到多个聚类。

其中，所述分别对每个聚类中的数据样本进行去冗余操作，得到待训练样本集，包括：分别计算每个聚类中的数据样本的相似度；根据所述相似度对每个聚类中的冗余样本进行剔除，得到待训练样本集。

其中，所述分别计算每个聚类中的数据样本的相似度，包括：利用最大化信息量算法分别计算每个聚类中的数据样本的最大信息量；根据最大信息量从每个聚类中选取不确定性最大的前K个数据样本，得到待标注样本集，其中，K为正整数；计算待标注样本集中两两数据样本之间的相似度，得到每个聚类中的数据样本的相似度。

其中，所述利用最大化信息量算法分别计算每个聚类中的数据样本的最大信息量，包括：采用预设的初始分类模型对每个聚类中的数据样本的客群分类进行预测；根据预测结果计算相应数据样本的最大信息量。

其中，所述采用预设的初始分类模型对每个聚类中的数据样本的客群分类进行预测之前，还包括：从每个聚类中随机选择预定数量的数据样本进行人工标注；根据人工标注后的数据样本对预设的原始模型进行训练，得到初始分类模型；所述采用预设的初始分类模型对每个聚类中的数据样本的客群分类进行预测，包括：采用得到的初始分类模型，对每个聚类中的其他未进行人工标注的数据样本的客群分类进行预测。

其中，所述根据所述相似度对每类数据样本中的冗余样本进行剔除，得到待训练样本集，包括：获取相似度大于预设阈值的数据样本，得到候选集；从候选集中筛选最大信息量满足预设条件的数据样本，得到待训练样本集。

所述指令例如为软件形式的指令模块，具体描述如下：

如图3A所示，该用户客群分类装置包括采集单元301、选择单元302、聚类单元303、去冗余单元304、以及分类单元305，如下：

（1）采集单元301；

采集单元301，用于采集多个用户的用户数据。

例如，该采集单元301可以从网络中采集各种类型用户的用户数据，或者，也可以由其他的采集设备对各种类型用户的用户数据进行采集，然后在需要使用该用户数据时，由采集单元301从该采集设备中进行获取，等等。

其中，该用户数据可以包括多种用户行为类别的数据，比如，可以包括用户的帐号信息、社交信息、支付行为数据、娱乐休闲行为数据、教育行为数据、和/或其他网络行为数据等信息。

（2）选择单元302；

选择单元302，用于按照预设策略从该用户数据中选择数据样本，得到数据样本集。

其中，该预设策略可以根据实际应用的需求进行设置，例如，可以设置为“对用户数据进行清洗”。其中，清洗的方式可以多种，比如“去除奇异点”、“补齐缺失值”、和/或“去掉重复采集的数据”，等等。

（3）聚类单元303；

聚类单元303，用于对该数据样本集中的数据样本进行聚类运算，得到多个聚类。

例如，该聚类单元303，具体可以用于获取数据样本的用户行为类别，根据该用户行为类别对该数据样本集中的数据样本进行聚类运算，得到多个聚类。

其中，该聚类算法也可以根据实际应用的需求而定，比如，可以采用kNN或k-means算法等。

（4）去冗余单元304;

去冗余单元304，用于分别对每个聚类中的数据样本进行去冗余操作，得到待训练样本集。

例如，该去冗余单元304可以包括计算子单元和筛选子单元，如下:

该计算子单元，用于分别计算每个聚类中的数据样本的相似度。

该筛选子单元，用于根据该相似度对每个聚类中的冗余样本进行剔除，得到待训练样本集。

比如，该计算子单元，具体可以用于利用最大化信息量算法分别计算每个聚类中的数据样本的最大信息量，根据该最大信息量从每个聚类中选取不确定性最大的前K个数据样本，得到待标注样本集，计算待标注样本集中两两数据样本之间的相似度，得到每个聚类中的数据样本的相似度；其中，K为正整数。

该筛选子单元，具体可以用于获取相似度大于预设阈值的数据样本，得到候选集，从候选集中筛选最大信息量满足预设条件的数据样本，得到待训练样本集；比如，可以统计该候选集中每个数据样本所具有的高度相似度的数量，得到高度相似计数，从该候选集中剔除高度相似计数最高的前N个数据样本，得到待训练样本集，等等。

其中，计算该最大信息量的方式可以有多种，比如，可以如下:

该计算子单元，具体可以用于采用预设的初始分类模型对每个聚类中的数据样本的客群分类进行预测，根据预测结果计算相应数据样本的最大信息量；比如，可以直接根据预测结果（如该用户属于某一分类的概率）计算相应数据样本的最大信息量，或者，也可以计算预测结果的置信度，然后，根据该置信度来计算相应数据样本的最大信息量，等等。

其中，该初始分类模型可以由维护人员进行预设，也可以由该用户客群分类装置通过训练来得到，比如，可以从每个聚类中随机选择一定数量的数据样本进行人工标注，构成初始训练样本，然后，利用这些初始训练样本对预设的原始模型进行训练，得到初始分类模型；即，可选地，如图3B所示，该用户客群分类装置还可以包括模型建立单元306，如下：

模型建立单元306，可以用于从每个聚类中随机选择预定数量的数据样本进行人工标注，根据人工标注后的数据样本对预设的原始模型进行训练，得到初始分类模型。

则此时，计算子单元，具体可以用于采用模型建立单元得到的初始分类模型，对每个聚类中的其他未进行人工标注的数据样本的客群分类进行预测。

其中，训练时所采用的原始模型可以根据实际应用的需求进行设置，而该预定数量也可以根据实际应用的需求进行设置，在此不再赘述。

（5）分类单元305;

分类单元305，用于基于该待训练样本集进行分类模型训练，并根据训练后分类模型对待分类数据进行客群分类。

例如，该分类单元305，具体用于获取预设的初始分类模型，对该待训练样本集中的数据样本进行标注，根据标注后的数据样本对该初始分类模型进行训练，得到训练后分类模型，获取待分类数据，将该待分类数据导入训练后分类模型中，得到输出结果，根据该输出结果确定该待分类数据的客群分类。

具体实施时，以上各个单元可以作为独立的实体来实现，也可以进行任意组合，作为同一或若干个实体来实现，以上各个单元的具体实施可参见前面的实施例，在此不再赘述。

由上可知，本申请实施例的用户客群分类装置在从采集到的用户数据中选择数据样本后，可以由聚类单元303对这些数据样本进行聚类运算，并由去冗余单元304分别对每个聚类中的数据样本进行去冗余操作，得到待训练样本集，然后，才由分类单元305基于该待训练样本集进行分类模型训练，并根据训练后分类模型对待分类数据进行客群分类；由于该方案在进行分类模型训练之前，先对数据样本进行了聚类，然后，再分别从每个聚类中选择数据样本进行训练，所以，可以保证每个聚类中都有数据样本被选中进入到分类模型的训练样本中，最大程度上使分类模型能从每个聚类中去学习，覆盖样本预期分

布，大大提高分类模型的准确率和泛化能力；此外，由于该方案可以对每个聚类中的数据样本进行去冗余操作，以剔除相似度高的冗余数据样本，使得可以尽可能减少标注相似的数据样本，因此，可以大大加快分类模型的收敛；总而言之，该方案不仅可以大大加快分类模型的收敛，而且可以提高分类模型的准确率和泛化能力，有利于提高用户客群分类的效率。

本申请实施例还提供一种服务器，如图 4 所示，其示出了本申请实施例所涉及的服务器的结构示意图，具体来讲:

该服务器可以包括一个或者一个以上处理核心的处理器 401、一个或一个以上存储器 402、电源 403 和输入单元 404 等部件。存储器 402 为计算机可读存储介质。图 4 中示出的服务器结构并不构成对服务器的限定，可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件布置。其中:

处理器 401 是该服务器的控制中心，利用各种接口和线路连接整个服务器的各个部分，通过运行或执行存储在存储器 402 内的软件程序和/或模块，以及调用存储在存储器 402 内的数据，执行服务器的各种功能和处理数据，从而对服务器进行整体监控。可选地，处理器 401 可包括一个或多个处理核心；优选的，处理器 401 可集成应用处理器和调制解调处理器，其中，应用处理器主要处理操作系统、用户界面和应用程序等，调制解调处理器主要处理无线通信。可以理解的是，上述调制解调处理器也可以不集成到处理器 401 中。

存储器 402 可用于存储软件程序以及模块，处理器 401 通过运行存储在存储器 402 的软件程序以及模块，从而执行各种功能应用以及数据处理。存储器 402 可主要包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需的应用程序（比如声音播放功能、图像播放功能等）等；存储数据区可存储根据服务器的使用所创建的数据等。此外，存储器 402 可以包括高速随机存取存储器，还可以包括非易失性存储器，例如至少一个磁盘存储器件、闪存器件、或其他易失性固态存储器件。相应地，存储器 402 还可以包括存储器控制器，以提供处理器 401 对存储器 402 的访问。

服务器还包括给各个部件供电的电源 403，优选的，电源 403 可以通过电源管理系统与处理器 401 逻辑相连，从而通过电源管理系统实现管理充电、放电、以及功耗管理等功能。电源 403 还可以包括一个或一个以上的直流或交流

电源、再充电系统、电源故障检测电路、电源转换器或者逆变器、电源状态指示器等任意组件。

该服务器还可包括输入单元 404，该输入单元 404 可用于接收输入的数字或字符信息，以及产生与用户设置以及功能控制有关的键盘、鼠标、操作杆、光学或者轨迹球信号输入。

尽管未示出，服务器还可以包括显示单元等，在此不再赘述。具体在本申请实施例中，服务器中的处理器401会按照如下的指令，将一个或一个以上的应用程序的进程对应的可执行文件加载到存储器402中，并由处理器401来运行存储在存储器402中的应用程序，从而实现各种功能，如下：

采集多个用户的用户数据，该用户数据包括多种用户行为类别的数据，按照预设策略从该用户数据中选择数据样本，得到数据样本集，对该数据样本集中的数据样本进行聚类运算，得到多个聚类，分别对每个聚类中的数据样本进行去冗余操作，得到待训练样本集，基于该待训练样本集进行分类模型训练，并根据训练后分类模型对待分类数据进行客群分类。

其中，去冗余的方式可以有多种，例如，可以分别计算每个聚类中的数据样本的相似度，根据该相似度对每个聚类中的冗余样本进行剔除，得到待训练样本集。

比如，可以利用最大化信息量算法分别计算每个聚类中的数据样本的最大信息量，根据最大信息量从每个聚类中选取不确定性最大的前K个数据样本，得到待标注样本集，计算待标注样本集中两两数据样本之间的相似度，得到每个聚类中的数据样本的相似度，获取相似度大于预设阈值的数据样本，得到候选集，统计该候选集中每个数据样本所具有的高度相似度的数量，得到高度相似计数，从该候选集中剔除高度相似计数最高的前N个数据样本，得到待训练样本集，等等。

以上各个操作的具体实施可参见前面的实施例，在此不再赘述。

由上可知，本申请实施例的服务器在从采集到的用户数据中选择数据样本后，可以对这些数据样本进行聚类运算，并分别对每个聚类中的数据样本进行去冗余操作，得到待训练样本集，然后，才基于该待训练样本集对分类模型进行训练，并根据训练后分类模型对待分类数据进行客群分类；由于该方案在进行分类模型训练之前，先对数据样本进行了聚类，然后，再分别从每个聚类中

选择数据样本进行训练，所以，可以保证每个聚类中都有数据样本被选中进入到分类模型的训练样本中，最大程度上使分类模型能从每个聚类中去学习，覆盖样本预期分布，大大提高分类模型的准确率和泛化能力；此外，由于该方案可以对每个聚类中的数据样本进行去冗余操作，以剔除相似度高的冗余数据样本，使得可以尽可能减少标注相似的数据样本，因此，可以大大加快分类模型的收敛；总而言之，该方案不仅可以大大加快分类模型的收敛，而且可以提高分类模型的准确率和泛化能力，有利于提高用户客群分类的效率。

上述实施例的各种方法中的全部或部分步骤是可以通过程序来指令相关的硬件来完成，该程序可以存储于一计算机可读存储介质中，存储介质可以包括：只读存储器（ROM，Read Only Memory）、随机存取记忆体（RAM，Random Access Memory）、磁盘或光盘等。

以上对本申请实施例所提供的一种用户客群分类方法和装置进行了详细介绍，本文中应用了具体个例对本申请的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本申请的方法及其核心思想；同时，对于本领域的技术人员，依据本申请的思想，在具体实施方式及应用范围上均会有改变之处，综上所述，本说明书内容不应理解为对本申请的限制。

权利要求书

1、一种用户客群分类方法，包括:

采集多个用户的用户数据，所述用户数据包括多种用户行为类别的数据;

按照预设策略从所述多个用户的用户数据中选择数据样本，得到数据样本集;

对所述数据样本集中的数据样本进行聚类运算，得到多个聚类;

分别对每个聚类中的数据样本进行去冗余操作，得到待训练样本集;

基于所述待训练样本集进行分类模型训练，并根据训练后分类模型对待分类数据进行客群分类。

2、根据权利要求1所述的方法，其中，所述对所述数据样本集中的数据样本进行聚类运算，得到多个聚类，包括:

获取数据样本的用户行为类别;

根据所述用户行为类别对所述数据样本集中的数据样本进行聚类运算，得到多个聚类。

3、根据权利要求1所述的方法，其中，所述分别对每个聚类中的数据样本进行去冗余操作，得到待训练样本集，包括:

分别计算每个聚类中的数据样本的相似度;

根据所述相似度对每个聚类中的冗余样本进行剔除，得到待训练样本集。

4、根据权利要求3所述的方法，其中，所述分别计算每个聚类中的数据样本的相似度，包括:

利用最大化信息量算法分别计算每个聚类中的数据样本的最大信息量;

根据最大信息量从每个聚类中选取不确定性最大的前K个数据样本，得到待标注样本集，其中，K为正整数;

计算待标注样本集中两两数据样本之间的相似度，得到每个聚类中的数据样本的相似度。

5、根据权利要求4所述的方法，其中，所述利用最大化信息量算法分别计算每个聚类中的数据样本的最大信息量，包括:

采用预设的初始分类模型对每个聚类中的数据样本的客群分类进行预测;

根据预测结果计算相应数据样本的最大信息量。

6、根据权利要求5所述的方法，其中，所述采用预设的初始分类模型对每个聚类中的数据样本的客群分类进行预测之前，还包括:

从每个聚类中随机选择预定数量的数据样本进行人工标注;

根据人工标注后的数据样本对预设的原始模型进行训练，得到初始分类模型;

所述采用预设的初始分类模型对每个聚类中的数据样本的客群分类进行预测，包括：采用得到的初始分类模型，对每个聚类中的其他未进行人工标注的数据样本的客群分类进行预测。

7、根据权利要求3所述的方法，其中，所述根据所述相似度对每类数据样本中的冗余样本进行剔除，得到待训练样本集，包括:

获取相似度大于预设阈值的数据样本，得到候选集;

从候选集中筛选最大信息量满足预设条件的数据样本，得到待训练样本集。

8、根据权利要求7所述的方法，其中，所述从候选集中筛选最大信息量满足预设条件的数据样本，得到待训练样本集，包括:

统计所述候选集中每个数据样本所具有的高度相似度的数量，得到高度相似计数，所述高度相似度为大于预设阈值的相似度;

从所述候选集中剔除高度相似计数最高的前N个数据样本，得到待训练样本集，其中，N为正整数。

9、根据权利要求1至8任一项所述的方法，其中，所述基于所述待训练样本集进行分类模型训练，包括:

获取预设的初始分类模型;

对所述待训练样本集中的数据样本进行标注;

根据标注后的数据样本对所述初始分类模型进行训练，得到训练后分类模型。

10、根据权利要求1至8任一项所述的方法，其中，所述根据训练后分类模型对待分类数据进行客群分类，包括:

获取待分类数据;

将所述待分类数据导入训练后分类模型中，得到输出结果;

根据所述输出结果确定所述待分类数据的客群分类。

11、一种用户客群分类装置，包括：处理器和存储器，所述存储器上存储有计算机可读指令，所述计算机可读指令由所述处理器执行以完成以下操作：

采集多个用户的用户数据，所述用户数据包括多种用户行为类别的数据；

按照预设策略从所述多个用户的用户数据中选择数据样本，得到数据样本集；

对所述数据样本集中的数据样本进行聚类运算，得到多个聚类；

分别对每个聚类中的数据样本进行去冗余操作，得到待训练样本集；

基于所述待训练样本集进行分类模型训练，并根据训练后分类模型对待分类数据进行客群分类。

12、根据权利要求11所述的装置，其中，所述对所述数据样本集中的数据样本进行聚类运算，得到多个聚类，包括：

获取数据样本的用户行为类别；

根据所述用户行为类别对所述数据样本集中的数据样本进行聚类运算，得到多个聚类。

13、根据权利要求11所述的装置，其中，所述分别对每个聚类中的数据样本进行去冗余操作，得到待训练样本集，包括：

分别计算每个聚类中的数据样本的相似度；

根据所述相似度对每个聚类中的冗余样本进行剔除，得到待训练样本集。

14、根据权利要求13所述的装置，其中，所述分别计算每个聚类中的数据样本的相似度，包括：

利用最大化信息量算法分别计算每个聚类中的数据样本的最大信息量；

根据最大信息量从每个聚类中选取不确定性最大的前K个数据样本，得到待标注样本集，其中，K为正整数；

计算待标注样本集中两两数据样本之间的相似度，得到每个聚类中的数据样本的相似度。

15、根据权利要求13所述的装置，其中，所述根据所述相似度对每类数据样本中的冗余样本进行剔除，得到待训练样本集，包括：

获取相似度大于预设阈值的数据样本，得到候选集；

从候选集中筛选最大信息量满足预设条件的数据样本，得到待训练样本集。

16、一种用户客群分类方法，用于服务器，所述服务器包括：处理器和存储器，所述方法包括：

采集多个用户的用户数据，所述用户数据包括多种用户行为类别的数据；

按照预设策略从所述多个用户的用户数据中选择数据样本，得到数据样本集；

对所述数据样本集中的数据样本进行聚类运算，得到多个聚类；

分别对每个聚类中的数据样本进行去冗余操作，得到待训练样本集；

基于所述待训练样本集进行分类模型训练，并根据训练后分类模型对待分类数据进行客群分类。

17、根据权利要求16所述的方法，其中，所述对所述数据样本集中的数据样本进行聚类运算，得到多个聚类，包括：

获取数据样本的用户行为类别；

根据所述用户行为类别对所述数据样本集中的数据样本进行聚类运算，得到多个聚类。

18、根据权利要求16所述的方法，其中，所述分别对每个聚类中的数据样本进行去冗余操作，得到待训练样本集，包括：

分别计算每个聚类中的数据样本的相似度；

根据所述相似度对每个聚类中的冗余样本进行剔除，得到待训练样本集。

19、根据权利要求18所述的方法，其中，所述分别计算每个聚类中的数据样本的相似度，包括：

利用最大化信息量算法分别计算每个聚类中的数据样本的最大信息量；

根据最大信息量从每个聚类中选取不确定性最大的前K个数据样本，得到待标注样本集，其中，K为正整数；

计算待标注样本集中两两数据样本之间的相似度，得到每个聚类中的数据样本的相似度。

20、根据权利要求19所述的方法，其中，所述利用最大化信息量算法分别计算每个聚类中的数据样本的最大信息量，包括：

采用预设的初始分类模型对每个聚类中的数据样本的客群分类进行预测；

根据预测结果计算相应数据样本的最大信息量。

21、根据权利要求20所述的方法，其中，所述采用预设的初始分类模型对每个聚类中的数据样本的客群分类进行预测之前，还包括:

从每个聚类中随机选择预定数量的数据样本进行人工标注;

根据人工标注后的数据样本对预设的原始模型进行训练，得到初始分类模型;

所述采用预设的初始分类模型对每个聚类中的数据样本的客群分类进行预测，包括：采用得到的初始分类模型，对每个聚类中的其他未进行人工标注的数据样本的客群分类进行预测。

22、根据权利要求18所述的方法，其中，所述根据所述相似度对每类数据样本中的冗余样本进行剔除，得到待训练样本集，包括:

获取相似度大于预设阈值的数据样本，得到候选集;

从候选集中筛选最大信息量满足预设条件的数据样本，得到待训练样本集。

23、根据权利要求22所述的方法，其中，所述从候选集中筛选最大信息量满足预设条件的数据样本，得到待训练样本集，包括:

统计所述候选集中每个数据样本所具有的高度相似度的数量，得到高度相似计数，所述高度相似度为大于预设阈值的相似度;

从所述候选集中剔除高度相似计数最高的前N个数据样本，得到待训练样本集，其中，N为正整数。

24、根据权利要求16至23任一项所述的方法，其中，所述基于所述待训练样本集进行分类模型训练，包括:

获取预设的初始分类模型;

对所述待训练样本集中的数据样本进行标注;

根据标注后的数据样本对所述初始分类模型进行训练，得到训练后分类模型。

25、根据权利要求16至23任一项所述的方法，其中，所述根据训练后分类模型对待分类数据进行客群分类，包括:

获取待分类数据;

将所述待分类数据导入训练后分类模型中，得到输出结果;

根据所述输出结果确定所述待分类数据的客群分类。

26、一种计算机可读存储介质，其中，所述计算机可读存储介质中存储有指令，当其在计算机上运行时，使得计算机执行上述权利要求 1-10 中任一项用户客群分类方法。

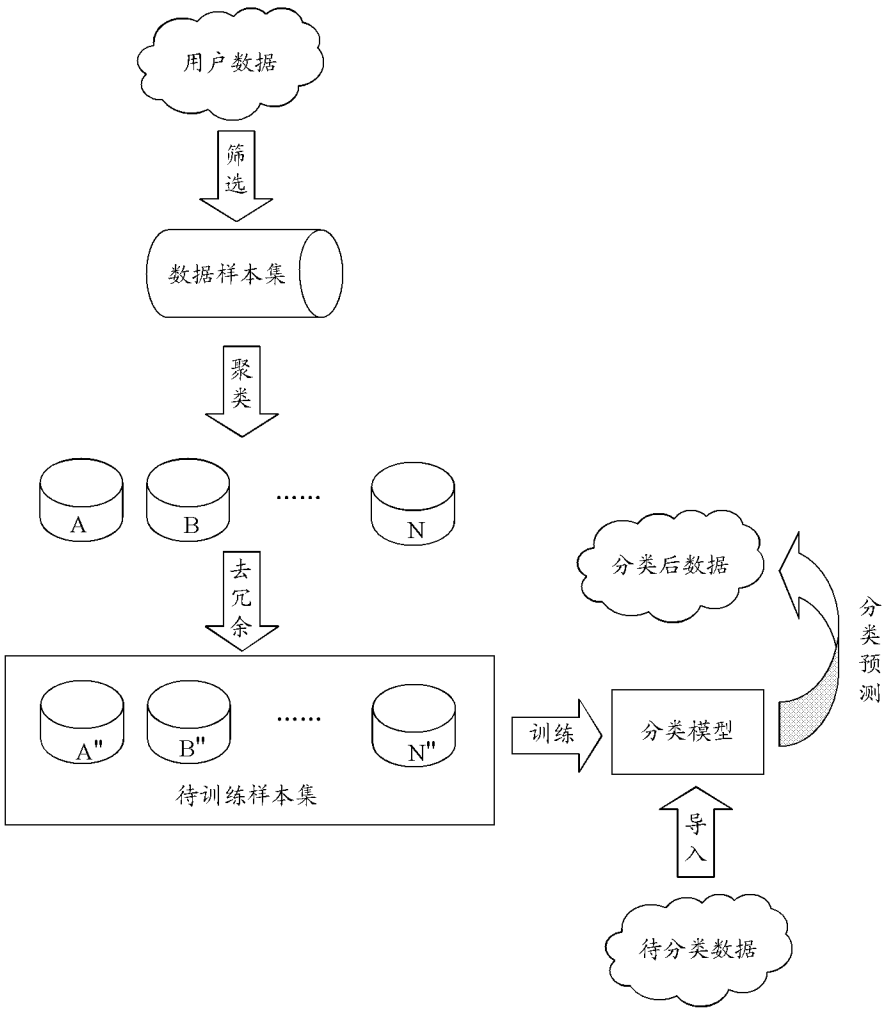

用户数据
筛选
数据样本集
聚类
A
B
.....
N
去冗余
A"
B"
.....
N"
待训练样本集
训练
分类模型
分类后数据
分类预测
导入
待分类数据

图 1A

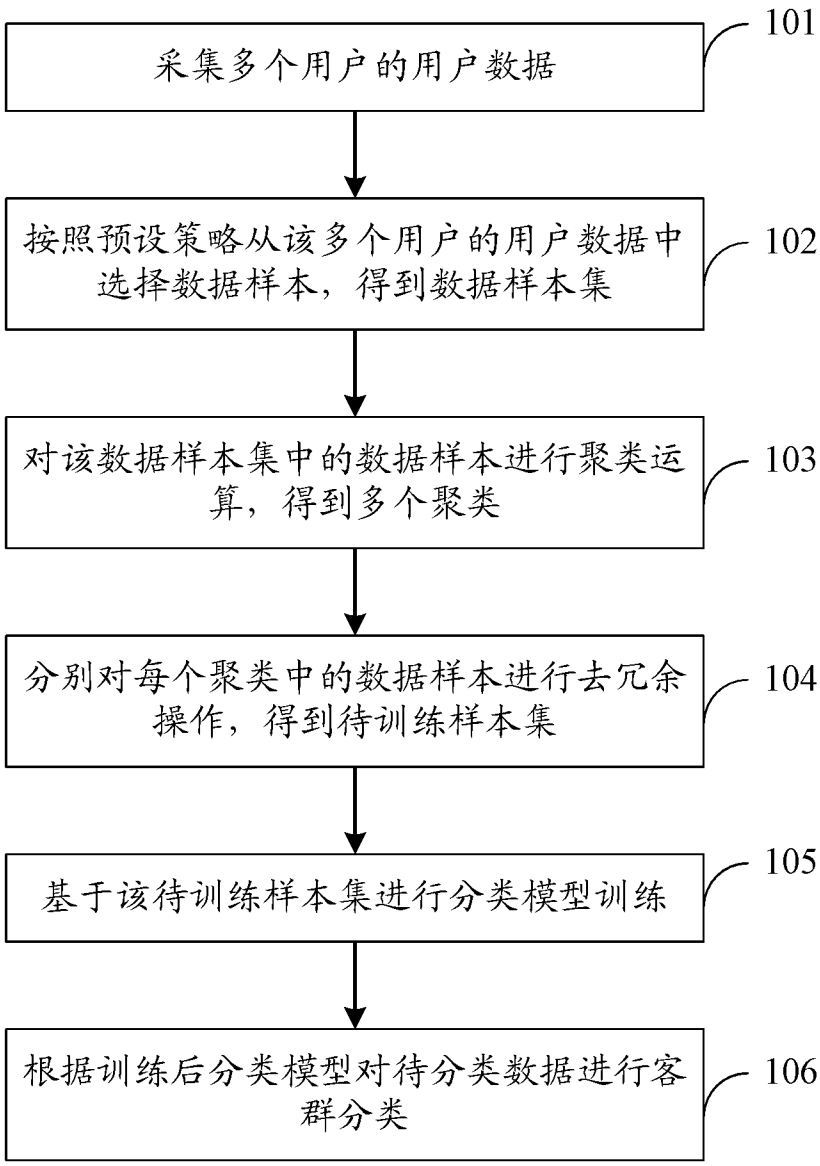

采集多个用户的用户数据
101
按照预设策略从该多个用户的用户数据中选择数据样本，得到数据样本集
102
对该数据样本集中的数据样本进行聚类运算，得到多个聚类
103
分别对每个聚类中的数据样本进行去冗余操作，得到待训练样本集
104
基于该待训练样本集进行分类模型训练
105
根据训练后分类模型对待分类数据进行客群分类
106

图 1B

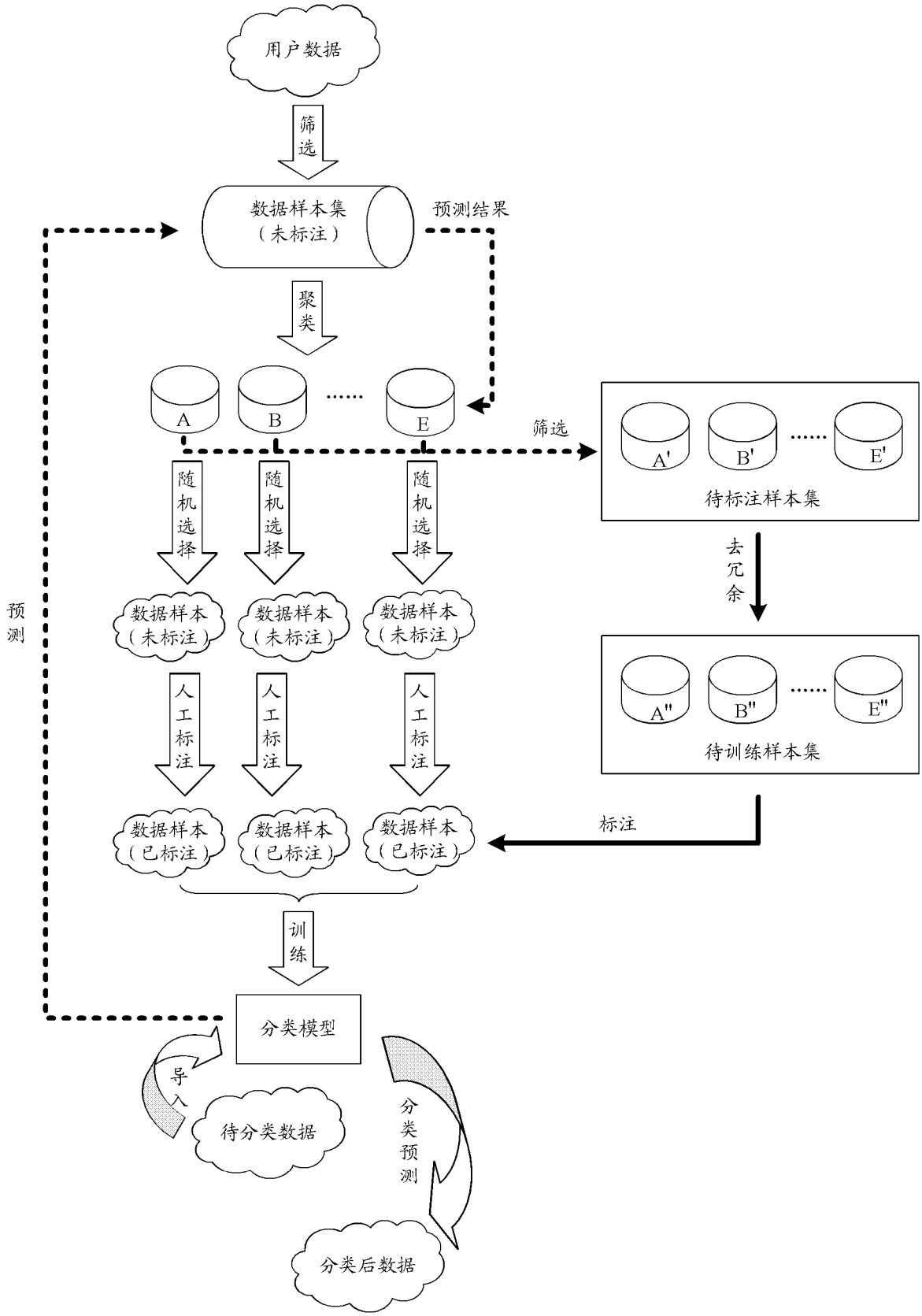

用户数据
筛选
数据样本集
（未标注）
预测结果
聚类
A
B
.....
E
筛选
A'
B'
.....
E'
待标注样本集
随机选择
随机选择
随机选择
去冗余
数据样本
（未标注）
数据样本
（未标注）
数据样本
（未标注）
预测
人工标注
人工标注
人工标注
A''
B''
.....
E''
待训练样本集
数据样本
（已标注）
数据样本
（已标注）
数据样本
（已标注）
标注
训练
分类模型
导入
待分类数据
分类预测
分类后数据

图 2A

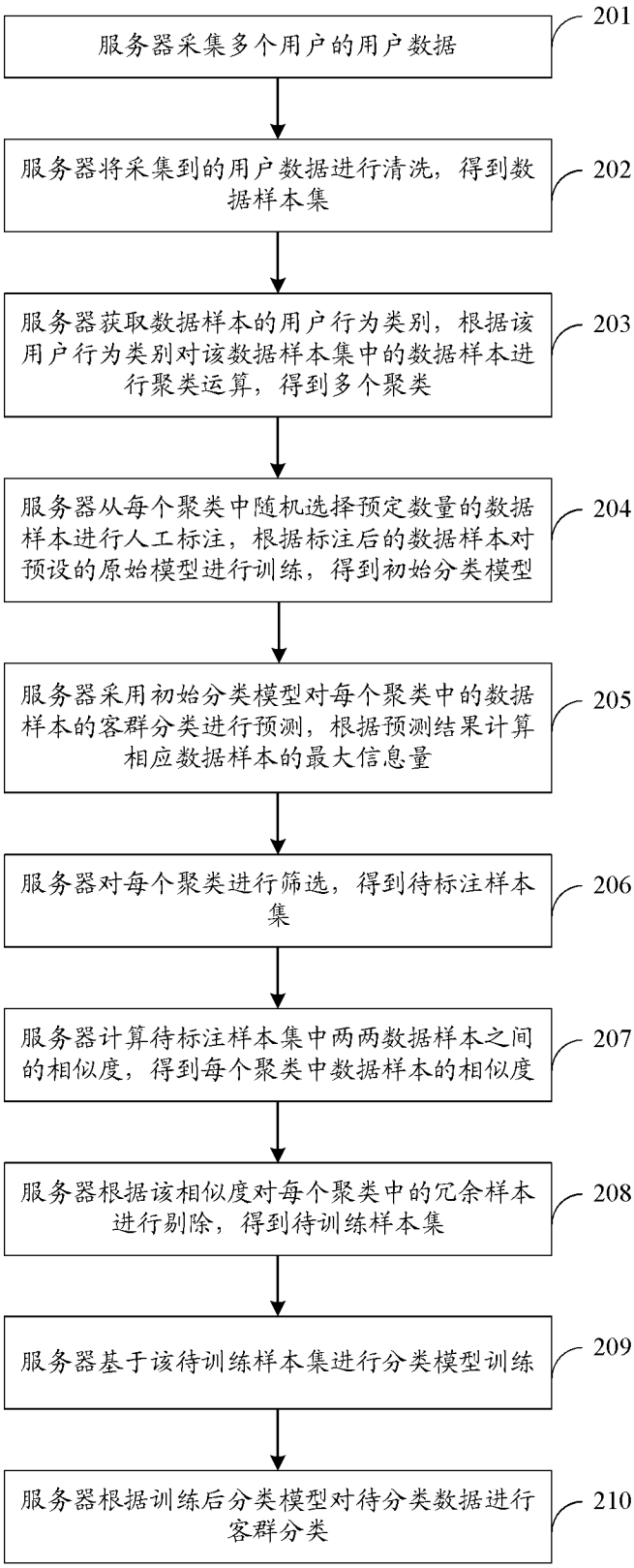

服务器采集多个用户的用户数据
201
服务器将采集到的用户数据进行清洗，得到数据样本集
202
服务器获取数据样本的用户行为类别，根据该用户行为类别对该数据样本集中的数据样本进行聚类运算，得到多个聚类
203
服务器从每个聚类中随机选择预定数量的数据样本进行人工标注，根据标注后的数据样本对预设的原始模型进行训练，得到初始分类模型
204
服务器采用初始分类模型对每个聚类中的数据样本的客群分类进行预测，根据预测结果计算相应数据样本的最大信息量
205
服务器对每个聚类进行筛选，得到待标注样本集
206
服务器计算待标注样本集中两两数据样本之间的相似度，得到每个聚类中数据样本的相似度
207
服务器根据该相似度对每个聚类中的冗余样本进行剔除，得到待训练样本集
208
服务器基于该待训练样本集进行分类模型训练
209
服务器根据训练后分类模型对待分类数据进行客群分类
210

图 2B

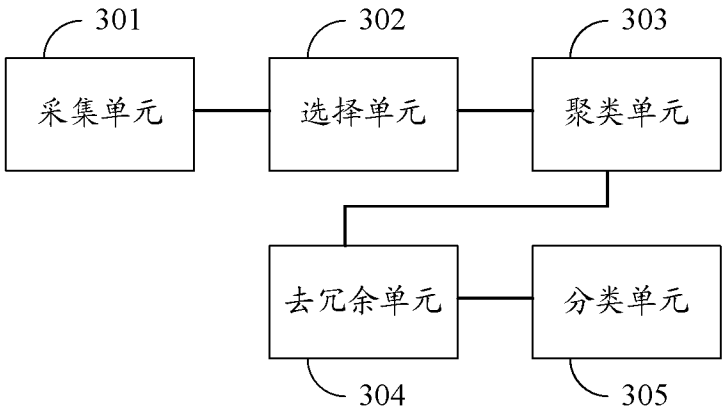

301
302
303
采集单元
选择单元
聚类单元
去冗余单元
分类单元
304
305

图 3A

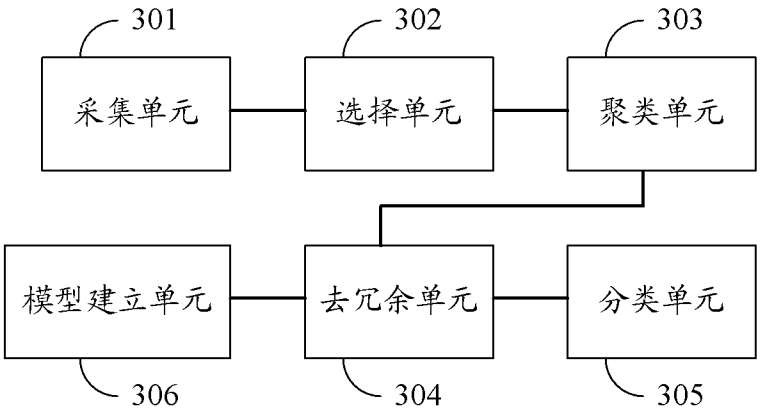

301
302
303
采集单元
选择单元
聚类单元
模型建立单元
去冗余单元
分类单元
306
304
305

图 3B

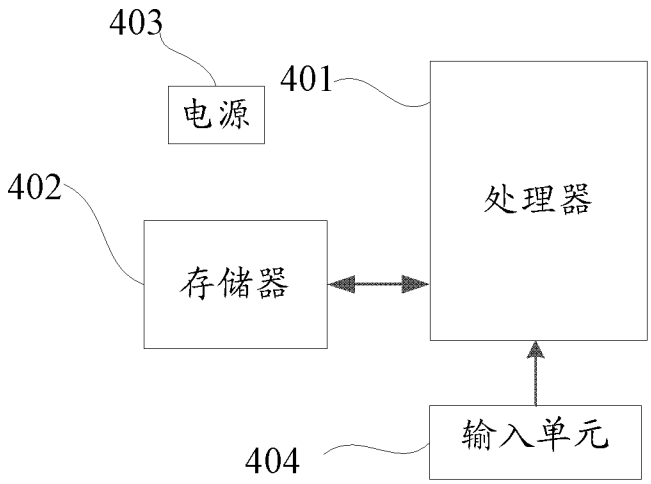


图 4

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2018/084565

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30 (2006.01) 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)

CPRSABS, CNTXT, CNKI, DWPI, SIPOABS, USTXT, IEEE: 分类, 聚类, 冗余, 去重, 训练, 样本, 用户, classify, cluster, redundancy, dereplication, train, sample, user

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 103617429 A (SOOCHOW UNIVERSITY) 05 March 2014 (05.03.2014), description, paragraphs [0005]-[0116]	1-26
A	CN 101211339 A (SHANGHAI XINSHENG ELECTRONIC TECHNOLOGY CO., LTD.) 02 July 2008 (02.07.2008), entire document	1-26
A	CN 103885977 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 25 June 2014 (25.06.2014), entire document	1-26
A	CN 101980202 A (XIDIAN UNIVERSITY) 23 February 2011 (23.02.2011), entire document	1-26

☐ 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	Date of mailing of the international search report
20 July 2018	01 August 2018
Name and mailing address of the ISA State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	Authorized officer WANG, Xuelian Telephone No. (86-10) 62411747

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2018/084565

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 103617429 A	05 March 2014	None	
CN 101211339 A	02 July 2008	None	
CN 103885977 A	25 June 2014	None	
CN 101980202 A	23 February 2011	None	

国际检索报告

国际申请号
PCT/CN2018/084565

A. 主题的分类

G06F 17/30(2006.01)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））

CPRSABS，CNTXT，CNKI，DWPI，SIPOABS，USTXT，IEEE：分类，聚类，冗余，去重，训练，样本，用户，classify，cluster，redundancy，dereplication，train，sample，user

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
X	CN 103617429 A (苏州大学) 2014年 3月 5日 (2014-03-05) 说明书第5-116段	1-26
A	CN 101211339 A (上海芯盛电子科技有限公司) 2008年 7月 2日 (2008-07-02) 全文	1-26
A	CN 103885977 A (腾讯科技深圳有限公司) 2014年 6月 25日 (2014-06-25) 全文	1-26
A	CN 101980202 A (西安电子科技大学) 2011年 2月 23日 (2011-02-23) 全文	1-26

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 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
2018年 7月 20日	2018年 8月 1日
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局(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	受权官员 王雪莲 电话号码 62411747

表 PCT/ISA/210 (第2页) (2015年1月)

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2018/084565

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 103617429 A	2014年 3月 5日	无	
CN 101211339 A	2008年 7月 2日	无	
CN 103885977 A	2014年 6月 25日	无	
CN 101980202 A	2011年 2月 23日	无	