



(12)发明专利申请

(10)申请公布号 CN 111667022 A

(43)申请公布日 2020.09.15

(21)申请号 202010612623.X

(22)申请日 2020.06.30

(71)申请人 腾讯科技(深圳)有限公司

地址 518000 广东省深圳市南山区高新区
科技中一路腾讯大厦35层

(72)发明人 王星星 张悦 李救琦

(74)专利代理机构 广州华进联合专利商标代理
有限公司 44224

代理人 李文渊 曹瀚青

(51)Int.Cl.

G06K 9/62(2006.01)

G06N 20/00(2019.01)

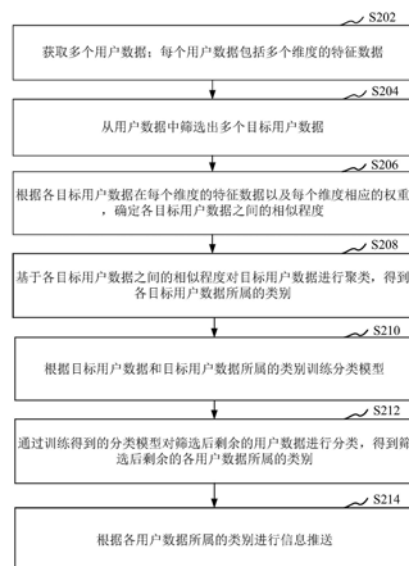
权利要求书3页 说明书16页 附图8页

(54)发明名称

用户数据处理方法、装置、计算机设备和存储介质

(57)摘要

本申请涉及一种用户数据处理方法、装置、计算机设备和存储介质。方法包括：获取多个用户数据；每个用户数据包括多个维度的特征数据；从用户数据中筛选出多个目标用户数据；根据各目标用户数据在每个维度的特征数据以及每个维度相应的权重，确定各目标用户数据之间的相似程度；基于各目标用户数据之间的相似程度对目标用户数据进行聚类，得到各目标用户数据所属的类别；根据目标用户数据和目标用户数据所属的类别训练分类模型；通过训练得到的分类模型对筛选后剩余的用户数据进行分类，得到筛选后剩余的各用户数据所属的类别；根据各用户数据所属的类别进行信息推送。采用本方法能够提高用户数据分类的准确性。



1. 一种用户数据处理方法,其特征在于,所述方法包括:
获取多个用户数据;每个所述用户数据包括多个维度的特征数据;
从所述用户数据中筛选出多个目标用户数据;
根据各所述目标用户数据在每个维度的特征数据以及每个维度相应的权重,确定各所述目标用户数据之间的相似程度;
基于各所述目标用户数据之间的相似程度对所述目标用户数据进行聚类,得到各所述目标用户数据所属的类别;
根据所述目标用户数据和所述目标用户数据所属的类别训练分类模型;
通过训练得到的所述分类模型对筛选后剩余的用户数据进行分类,得到筛选后剩余的各用户数据所属的类别;
根据各所述用户数据所属的类别进行信息推送。
2. 根据权利要求1所述的方法,其特征在于,所述根据各所述目标用户数据在每个维度的特征数据以及每个维度相应的权重,确定各所述目标用户数据之间的相似程度,包括:
向量化各所述目标用户数据,得到各所述目标用户数据的特征向量;所述特征向量包括所述多个维度的特征值;每个特征值对应一个维度的特征数据;
获取每个维度相应的权重;
对于任意两个所述目标用户数据的特征向量,将各维度的特征值按照各维度的权重计算距离,得到任意两个所述目标用户数据之间的相似程度。
3. 根据权利要求1所述的方法,其特征在于,所述根据各所述目标用户数据在每个维度的特征数据以及每个维度相应的权重,确定各所述目标用户数据之间的相似程度,包括:
获取每个维度相应的权重;
对每个所述目标用户数据,保留预设数量的、且权重最大的维度的特征数据,获得用于聚类的目标用户子数据;
对于任意两个所述目标用户子数据的特征向量,将各维度的特征值按照各维度的权重计算距离,得到任意两个所述目标用户子数据之间的相似程度。
4. 根据权利要求2或3所述的方法,其特征在于,所述获取每个维度相应的权重,包括:
获取训练完成的排序模型所输出的权重向量;
其中,所述排序模型通过存在训练标签的用户数据样本监督训练得到,并在监督训练的过程中对用户样本数据中多个维度的特征数据的影响程度排序;所述权重向量包括所述多个维度的特征数据的权重。
5. 根据权利要求4所述的方法,其特征在于,所述排序模型的训练步骤包括:
获取用户数据样本和所述用户数据样本对应的训练标签;所述用户数据样本包括多个维度的特征数据;
通过所述排序模型对所述用户数据样本进行预测,得到预测结果;
根据所述预测结果和所述训练标签之间的差异,优化所述排序模型;
通过所述排序模型对用户样本数据中多个维度的特征数据对预测的影响程度进行排序,并输出排序结果。
6. 根据权利要求4所述的方法,其特征在于,所述方法还包括:
获取所述权重向量中数值最大的向量元素和数值最小的向量元素;

确定所述数值最大的向量元素和所述数值最小的向量元素之间的差值；

根据所述数值最小的向量元素和所述差值，对所述权重向量中的各向量元素进行归一化处理。

7. 根据权利要求1所述的方法，其特征在于，所述根据所述目标用户数据和所述目标用户数据所属的类别训练分类模型，包括：

对每个所述目标用户数据，保留预设数量的、且权重最大的维度的特征数据，获得用于训练分类模型的目标用户子数据；

根据所述目标用户子数据和所述目标用户子数据所属类别训练分类模型；

所述通过训练得到的所述分类模型对筛选后剩余的用户数据进行分类，得到筛选后剩余的各用户数据所属的类别，包括：

对每个筛选后剩余的用户数据，保留所述预设数量的、且权重最大的维度的特征数据，获得用于使用分类模型的用户子数据；

通过训练得到的所述分类模型对所述用户子数据进行分类，得到各所述用户子数据所属的类别。

8. 根据权利要求1所述的方法，其特征在于，所述分类模型包括多个分类子结构；每个分类子结构包括注意力网络结构；

所述注意力网络结构的处理步骤包括：

对输入所述注意力网络结构的特征向量分别赋予权重，得到键向量、请求向量和值向量；

通过所述注意力网络结构对所述键向量、所述请求向量和所述值向量进行处理，得到处理结果。

9. 根据权利要求8所述的方法，其特征在于，所述通过所述注意力网络结构对所述键向量、所述请求向量和所述值向量进行处理，得到处理结果，包括：

通过所述注意力网络结构的非线性激活函数层，将所述键向量、所述请求向量和所述值向量映射为多于一组键向量、请求向量和值向量；

通过所述注意力网络结构的多个注意力机制层，分别对多于一组键向量、请求向量和值向量进行处理，得到中间结果；

通过所述注意力网络结构的拼接层和卷积层依次对中间结果进行处理，得到处理结果。

10. 根据权利要求8所述的方法，其特征在于，所述分类子结构还包括卷积网络结构和批标准化网络结构；

所述分类子结构的处理步骤包括：

通过所述卷积网络结构对输入所述分类子结构的数据进行卷积运算，并将卷积运算结果输出至所述批标准化网络结构；

通过所述批标准化网络结构对所述卷积运算结果进行分布调整，并将调整结果输出至所述注意力网络结构。

11. 根据权利要求1所述的方法，其特征在于，所述从所述用户数据中筛选出多个目标用户数据，包括：

对所述用户数据进行数据清洗；

将清洗后的用户数据进行多级划分,得到多个用户数据子集;

从各个最小数量的用户数据子集中随机选取用户数据作为目标用户数据。

12. 根据权利要求1所述的方法,其特征在于,所述根据各所述用户数据所属的类别进行信息推送,包括:

获取各个类别相应的推送信息;

向各所述用户数据所对应的用户终端,推送与各所述用户数据所属类别相应的推送信息。

13. 一种用户数据处理装置,其特征在于,所述装置包括:

获取模块,用于获取多个用户数据;每个所述用户数据包括多个维度的特征数据;

第一筛选模块,用于从所述用户数据中筛选出多个目标用户数据;

确定模块,用于根据各所述目标用户数据在每个维度的特征数据以及每个维度相应的权重,确定各所述目标用户数据之间的相似程度;

聚类模块,用于基于各所述目标用户数据之间的相似程度对所述目标用户数据进行聚类,得到各所述目标用户数据所属的类别;

训练模块,用于根据所述目标用户数据和所述目标用户数据所属的类别训练分类模型;

第二筛选模块,用于通过训练得到的所述分类模型对筛选后剩余的用户数据进行分类,得到筛选后剩余的各用户数据所属的类别;

应用模块,用于根据各所述用户数据所属的类别进行信息推送。

14. 一种计算机设备,包括存储器和处理器,所述存储器存储有计算机程序,其特征在于,所述处理器执行所述计算机程序时实现权利要求1至12中任一项所述的方法的步骤。

15. 一种计算机可读存储介质,存储有计算机程序,其特征在于,所述计算机程序被处理器执行时实现权利要求1至12中任一项所述的方法的步骤。

用户数据处理方法、装置、计算机设备和存储介质

技术领域

[0001] 本申请涉及计算机技术领域,特别是涉及一种用户数据处理方法、装置、计算机设备和存储介质。

背景技术

[0002] 随着计算机技术的发展,人们越来越多地依赖互联网来获取各方面的信息,为了达到向用户及时提供产品或者服务又尽量避免提供无用产品或者服务的目的,通常根据用户数据的分类标签确定接受产品或者服务的目标人群。

[0003] 然而,现有用户数据的分类主要是基于统计的用户画像进行,这种方式很难对用户稀疏的行为特征进行准确的评判,进而难以对用户数据真正所属的分类做出判断,从而导致得到的用户数据的分类结果准确率比较低。

发明内容

[0004] 基于此,有必要针对上述技术问题,提供一种能够提高用户数据分类准确性的用户数据处理方法、装置、计算机设备和存储介质。

[0005] 一种用户数据处理方法,所述方法包括:

[0006] 获取多个用户数据;每个所述用户数据包括多个维度的特征数据;

[0007] 从所述用户数据中筛选出多个目标用户数据;

[0008] 根据各所述目标用户数据在每个维度的特征数据以及每个维度相应的权重,确定各所述目标用户数据之间的相似程度;

[0009] 基于各所述目标用户数据之间的相似程度对所述目标用户数据进行聚类,得到各所述目标用户数据所属的类别;

[0010] 根据所述目标用户数据和所述目标用户数据所属的类别训练分类模型;

[0011] 通过训练得到的所述分类模型对筛选后剩余的用户数据进行分类,得到筛选后剩余的各用户数据所属的类别;

[0012] 根据各所述用户数据所属的类别进行信息推送。

[0013] 一种用户数据处理装置,所述装置包括:

[0014] 获取模块,用于获取多个用户数据;每个所述用户数据包括多个维度的特征数据;

[0015] 第一筛选模块,用于从所述用户数据中筛选出多个目标用户数据;

[0016] 确定模块,用于根据各所述目标用户数据在每个维度的特征数据以及每个维度相应的权重,确定各所述目标用户数据之间的相似程度;

[0017] 聚类模块,用于基于各所述目标用户数据之间的相似程度对所述目标用户数据进行聚类,得到各所述目标用户数据所属的类别;

[0018] 训练模块,用于根据所述目标用户数据和所述目标用户数据所属的类别训练分类模型;

[0019] 第二筛选模块,用于通过训练得到的所述分类模型对筛选后剩余的用户数据进行

分类,得到筛选后剩余的各用户数据所属的类别;

[0020] 应用模块,用于根据各所述用户数据所属的类别进行信息推送。

[0021] 一种计算机设备,包括存储器和处理器,所述存储器存储有计算机程序,所述处理器执行所述计算机程序时实现以下步骤:

[0022] 获取多个用户数据;每个所述用户数据包括多个维度的特征数据;

[0023] 从所述用户数据中筛选出多个目标用户数据;

[0024] 根据各所述目标用户数据在每个维度的特征数据以及每个维度相应的权重,确定各所述目标用户数据之间的相似程度;

[0025] 基于各所述目标用户数据之间的相似程度对所述目标用户数据进行聚类,得到各所述目标用户数据所属的类别;

[0026] 根据所述目标用户数据和所述目标用户数据所属的类别训练分类模型;

[0027] 通过训练得到的所述分类模型对筛选后剩余的用户数据进行分类,得到筛选后剩余的各用户数据所属的类别;

[0028] 根据各所述用户数据所属的类别进行信息推送。

[0029] 一种计算机可读存储介质,其上存储有计算机程序,所述计算机程序被处理器执行时实现以下步骤:

[0030] 获取多个用户数据;每个所述用户数据包括多个维度的特征数据;

[0031] 从所述用户数据中筛选出多个目标用户数据;

[0032] 根据各所述目标用户数据在每个维度的特征数据以及每个维度相应的权重,确定各所述目标用户数据之间的相似程度;

[0033] 基于各所述目标用户数据之间的相似程度对所述目标用户数据进行聚类,得到各所述目标用户数据所属的类别;

[0034] 根据所述目标用户数据和所述目标用户数据所属的类别训练分类模型;

[0035] 通过训练得到的所述分类模型对筛选后剩余的用户数据进行分类,得到筛选后剩余的各用户数据所属的类别;

[0036] 根据各所述用户数据所属的类别进行信息推送。

[0037] 上述用户数据处理方法、装置、计算机设备和存储介质,在获取到大量多维的用户数据后,筛选出少量用户数据作为目标来进行聚类 and 类别划分,得到带有类别的目标用户数据,然后利用带有类别的目标用户数据来训练分类模型,再利用训练得到的分类模型对筛选后剩余的用户数据进行分类。这样,一方面在聚类时,对不同维度的特征数据设置不同的权重参数,可以很好地对离散的用户数据进行有针对性处理,进而更加准确地根据不同维度的特征数据的重要程度对用户数据进行聚类,提高用户分类准确性;而且基于该部分带有类别的目标用户数据训练的分类模型,也能准确有效地对筛选后剩余的用户数据进行分类;另一方面,仅对部分用户数据进行聚类,还可以避免全量数据聚类带来的计算量和可能存在的计算资源不足的问题;另外在得到用户数据所属的类别后,可以更加准确且有针对性地进行信息推送。

附图说明

[0038] 图1为一个实施例中用户数据处理方法的应用环境图;

- [0039] 图2为一个实施例中用户数据处理方法的流程示意图；
- [0040] 图3为一个实施例中用户数据处理的流程框图；
- [0041] 图4为一个实施例中对目标用户数据进行聚类后的聚类效果示意图；
- [0042] 图5为一个实施例中对筛选后剩余的用户数据进行分类后的分类效果示意图；
- [0043] 图6为一个实施例中分类模型的结构示意图；
- [0044] 图7为一个实施例中注意力网络结构的结构示意图；
- [0045] 图8为一个实施例中用户数据处理装置的结构框图；
- [0046] 图9为一个实施例中计算机设备的内部结构图。

具体实施方式

[0047] 为了使本申请的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本申请进行进一步详细说明。应当理解，此处描述的具体实施例仅仅用以解释本申请，并不用于限定本申请。

[0048] 人工智能(Artificial Intelligence, AI)是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能,感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。换句话说,人工智能是计算机科学的一个综合技术,它企图了解智能的实质,并生产出一种新的能以人类智能相似的方式做出反应的智能机器。人工智能也就是研究各种智能机器的设计原理与实现方法,使机器具有感知、推理与决策的功能。

[0049] 人工智能技术是一门综合学科,涉及领域广泛,既有硬件层面的技术也有软件层面的技术。人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件技术包括自然语言处理技术和机器学习。

[0050] 机器学习(Machine Learning, ML)是一门多领域交叉学科,涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为,以获取新的知识或技能,重新组织已有的知识结构使之不断改善自身的性能。机器学习是人工智能的核心,是使计算机具有智能的根本途径,其应用遍及人工智能的各个领域。机器学习和深度学习通常包括人工神经网络、置信网络、强化学习、迁移学习、归纳学习、示教学习等技术。

[0051] 云技术(Cloud technology)是指在广域网或局域网内将硬件、软件、网络等系列资源统一起来,实现数据的计算、储存、处理和共享的一种托管技术。

[0052] 云技术(Cloud technology)基于云计算商业模式应用的网络技术、信息技术、整合技术、管理平台技术、应用技术等的总称,可以组成资源池,按需所用,灵活便利。云计算技术将变成重要支撑。技术网络系统的后台服务需要大量的计算、存储资源,如视频网站、图片类网站和更多的门户网站。伴随着互联网行业的高度发展和应用,将来每个物品都有可能存在自己的识别标志,都需要传输到后台系统进行逻辑处理,不同程度级别的数据将会分开处理,各类行业数据皆需要强大的系统后盾支撑,只能通过云计算来实现。

[0053] 云安全主要研究方向包括:1、云计算安全,主要研究如何保障云自身及云上各种应用的安全,包括云计算系统安全、用户数据的安全存储与隔离、用户接入认证、信息传输安全、网络攻击防护、合规审计等;2、安全基础设施的云化,主要研究如何采用云计算新

建与整合安全基础设施资源,优化安全防护机制,包括通过云计算技术构建超大规模安全事件、信息采集与处理平台,实现对海量信息的采集与关联分析,提升全网安全事件把控能力及风险控制能力;3、云安全服务,主要研究各种基于云计算平台为用户提供的安全服务,如防病毒服务等。

[0054] 以下将基于人工智能技术的机器学习和云技术,对本申请实施例提供的用户数据处理方法进行说明。

[0055] 本申请提供的用户数据处理方法,可以应用于如图1所示的应用环境中。其中,终端102通过网络与服务器104进行通信,终端102的数量多个。具体地,各终端102可向服务器104上传用户数据,每个用户数据包括多个维度的特征数据。服务器104即获取到多个用户数据;然后从这些用户数据中筛选出多个目标用户数据;根据各目标用户数据在每个维度的特征数据以及每个维度相应的权重,确定各目标用户数据之间的相似程度;然后基于各目标用户数据之间的相似程度对目标用户数据进行聚类,得到各目标用户数据所属的类别;再根据目标用户数据和目标用户数据所属的类别训练分类模型;再通过训练得到的分类模型对筛选后剩余的用户数据进行分类,得到筛选后剩余的各用户数据所属的类别;根据各用户数据所属的类别进行信息推送。在另外的实施例中,终端102也可通过其上运行的应用程序向服务器104上传用户数据。

[0056] 其中,终端102可以但不限于各种个人计算机、笔记本电脑、智能手机、平板电脑和便携式可穿戴设备,服务器104可以是独立的物理服务器,也可以是多个物理服务器构成的服务器集群或者分布式系统,还可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、CDN、以及大数据和人工智能平台等基础云计算服务的云服务器。终端可以是智能手机、平板电脑、笔记本电脑、台式计算机、智能音箱、智能手表等,但并不局限于此。终端以及服务器可以通过有线或无线通信方式进行直接或间接地连接,本申请在此不做限制。

[0057] 在一个实施例中,如图2所示,提供了一种用户数据处理方法,以该用户数据处理方法应用于计算机设备(例如图1中的服务器)为例进行说明,该用户数据处理方法包括以下步骤:

[0058] 步骤202,获取多个用户数据;每个用户数据包括多个维度的特征数据。

[0059] 其中,用户数据即用户特征数据是反映用户特性的数据。用户数据包括用户行为数据。用户行为数据是反映用户行为特征的数据。用户行为数据比如社交行为数据、浏览行为数据或者支付行为数据等。社交行为数据比如社交会话数据、社交信息发布数据或者社交消息评论信息等。浏览行为数据比如新闻浏览数据、音视频浏览数据或者商品浏览数据等。支付行为数据比如消费行为数据或者转账行为数据等。用户数据也可包括用户基础数据。用户基础数据是反映用户基础属性的数据。用户基础数据比如性别、年龄或者学历等。

[0060] 可以理解,本申请中多个包括两个和两个以上。在一个具体的实施例中,多个用户数据可以是百万级别、千万级别甚至上亿级别的数据集。

[0061] 维度可以是指特征数据所属的领域,也可以称为特征域。维度比如年龄维度、性别维度、视频维度或者图文维度等。

[0062] 在一个实施例中,维度的划分可根据实际需要自定义划分。在一种维度划分方式下划分得到的多个的维度,可以是另一种维度划分方式下划分得到的一个维度。比如,在A

方式下划分得到年龄维度、性别维度和地域维度,在B方式下划分得到基础信息维度。那么,可以认为年龄维度、性别维度和地域维度可以是基础信息维度的子维度。在一种维度划分方式下划分得到的一个维度,也可以是另一种维度划分方式下划分得到的多个的维度。比如,在A方式下划分得到视频维度,在B方式下划分得到体育视频维度和娱乐视频维度。那么,可以认为体育视频维度和娱乐视频维度可以是视频维度的子维度。

[0063] 可以理解,通常情况下每个用户一般只会对自己感兴趣的零散内容产生行为。那么,基于这些零散内容产生的行为相应地也会产生一些零散的用户数据,那么计算机设备得到的也就是各用户的离散的用户数据。现有技术中通常无法很好的基于这些高维度、且离散的用户数据来进行用户分类。

[0064] 在一个实施例中,计算机设备在步骤202获取的用户数据可以是原始的用户数据,此时需要对这些用户数据进行筛选后再进行后续的处理。比如,计算机设备获取M维用户数据,从这M维中筛选出N维,得到N维用户数据,再进行后续的处理;其中, $M > N$ 。

[0065] 在一个实施例中,计算机设备在步骤202获取的用户数据可以是预处理后的用户数据,预处理后的用户数据的包括特征数据的维度是特定的维度。比如权重较高的维度。比如,计算机设备获取的就是筛选出的N个维度的用户数据。

[0066] 步骤204,从用户数据中筛选出多个目标用户数据。

[0067] 其中,目标用户数据是指从大量用户数据中挑选出的作为目标进行处理的用户数据。目标用户数据也可以称为种子用户数据。

[0068] 具体地,种子用户数据可以是种子用户的用户数据。计算机设备可从用户数据中筛选出多个种子用户的用户数据作为目标用户数据。

[0069] 需要说明的是,种子用户数据不一定是种子用户的用户数据。

[0070] 在一个实施例中,步骤204包括:对用户数据进行数据清洗;将清洗后的用户数据进行多级划分,得到多个用户数据子集;从各个最小数量的用户数据子集中随机选取用户数据作为目标用户数据。

[0071] 具体地,计算机设备可对用户数据进行清洗,比如删除重复信息以及纠正存在的错误信息等将脏数据转化为满足数据质量要求的数据。由于用户数据的数量较大,计算机设备可以将这些用户数据按照简单的划分方式进行初次划分,比如按照年龄段进行划分,再对划分后的用户数据继续划分,实现对用户数据的多级划分,得到多个用户数据子集,每个用户数据子集的数量较小,再从各个最小数量的用户数据子集中随机选取用户数据作为目标用户数据。

[0072] 举例说明,参考图3,该图示出了一个实施例中用户数据处理的流程框图。可以看到计算机设备可先收集原始用户数据,然后对收集到的原始用户数据进行数据清洗,然后从数据清洗完成的用户数据中筛选出一部分作为种子用户数据,将筛选后剩余的用户数据作为非种子用户数据。其中,种子用户数据的数量远小于非种子用户数据的数量,这样可以减少对种子用户数据聚类的运算量。

[0073] 在本实施例中,先将用户数据进行多级划分,拆分成数量较小的子集,再从子集中随机选取用户数据,相较于直接从大量数据中随机选取可以比较全面地挑选出各种具有代表性的用户数据,避免筛选出用户数据之间过于相似集中。

[0074] 步骤206,根据各目标用户数据在每个维度的特征数据以及每个维度相应的权重,

确定各目标用户数据之间的相似程度。

[0075] 可以理解,用户数据包括多个维度的特征数据。目前无论是在何种具体场景下对用户数据进行聚类时,通常都是对各维度的特征数据进行无差别处理。这样可能会使得聚类结果并不适合当前的应用场景,进而不能很好的区别具体场景下的用户的价值。比如,在视频推广场景下,用户的视频浏览行为数据对用户数据聚类的重要程度要高于用户的社交会话数据的重要程度。在本实施例中,对不同的维度的特征数据设置不同的权重参数,可以使得聚类时能够更多地关注当前应用场景下更重要的特征,得到的聚类结果更适合当前的应用场景。

[0076] 具体地,计算机设备可对于这些目标用户数据中的任意两个目标用户数据,计算这两个目标用户数据在每个维度的特征数据之间的第一相似程度,然后根据每个维度的权重对每个维度下的第一相似程度融合,得到第二相似程度,作为这两个目标用户数据之间的相似程度。其中,根据每个维度的权重对每个维度下的第一相似程度融合,得到第二相似程度,具体可以是将每个维度下的第一相似程度按照相应维度的权重加权求和,得到第二相似程度。

[0077] 举例说明,假设每个目标用户数据包括3个维度(维度1、维度2和维度3)的特征数据,那么对于目标用户数据A和目标用户数据B,分别计算A和B在维度1的第一相似程度1,维度2的第一相似程度2,以及维度3的第一相似程度3,然后根据每个维度的权重对每个维度下的第一相似程度融合,得到第二相似程度。假设维度1的权重1,维度2的权重2,以及维度3的权重3,将第一相似程度1、第一相似程度2和第一相似程度3按照对应维度的权重加权求和,得到第一相似程度1*权重1+第一相似程度2*权重2+第一相似程度3*权重3,即第二相似程度,作为A和B之间的相似程度。

[0078] 其中,在目标用户数据之间的相似程度越高时,可以认为目标用户数据所对应的用户之间的相似程度越高,即用户越相似。用户数据可以向量化,特征数据可以数值化,特征数据之间的相似程度则可以通过数值之间的距离量化,用户数据之间的相似程度则可以通过向量之间的距离量化。

[0079] 在一个实施例中,步骤206包括:向量化各目标用户数据,得到各目标用户数据的特征向量;特征向量包括多个维度的特征值;每个特征值对应一个维度的特征数据;获取每个维度相应的权重;对于任意两个目标用户数据的特征向量,将各维度的特征值按照各维度的权重计算距离,得到任意两个目标用户数据之间的相似程度。

[0080] 其中,向量化是指将其他形式的数据以数学形式进行表达。比如,文本形式的“XXX”表示为数学形式“[0 0 0 1 0 0 0 0 0 0 0...]”,此时,“[0 0 0 1 0 0 0 0 0 0 0...]”即为对“XXX”进行向量化的结果,即为“XXX”的向量。可以理解,这里不限定将其他形式的数据转换为何种表示的向量,只要能够将其他形式的数据进行数学化表示即可。

[0081] 特征向量的每一个向量元素表示一个维度的特征数据所对应的特征值。举例说明,假设目标用户A的用户数据包括1、2、3和4这四个维度的特征数据,向量化用户A的用户数据得到的特征向量为(X1,X2,X3,X4)。那么,X1为维度1的特征数据对应的特征值,X2为维度2的特征数据对应的特征值,X3为维度3的特征数据对应的特征值,X4为维度4的特征数据对应的特征值。

[0082] 具体地,计算机设备可以通过对用户数据经过Embedding操作,将离散的用户数据

转化为连续的特征向量进行表示。

[0083] 在一个实施例中,每个维度相应的权重,可以是根据历史经验数据设定,也可以通过机器学习模型对现有分类的用户数据学习得到,还可以通过其他方式得到,本申请实施例在此不做限定。

[0084] 在一个具体的实施例中,计算机设备可以通过加权欧式距离衡量用户数据之间的相似度。加权欧式距离计算公式如下所示:

$$[0085] \quad \text{EuclideanDistance}(X,Y) = \sum_{k=1}^n \sqrt{\delta_k (x_k - y_k)^2} \quad (1)$$

[0086] 其中, $X = \{x_1, x_2, x_3, \dots, x_k, \dots, x_n\}$ 为一个目标用户数据对应的特征向量, $Y = \{y_1, y_2, y_3, \dots, y_k, \dots, y_n\}$ 为另一个目标用户数据对应的特征向量, n 为特征数量的维度, $\text{EuclideanDistance}(X,Y)$ 为 X 和 Y 之间的加权欧式距离,反映了 X 和 Y 之间的相似程度。 δ_k 为第 k 个维度的特征数据的权重。

[0087] 可以理解,在另外的实施例中,计算机设备还可以采用其他距离衡量用户数据之间的相似度,比如曼哈顿距离、切比雪夫距离以及搬土距离等。在采用其他距离衡量用户数据之间的相似度时,同样通过加权的方式引入权重参数。

[0088] 上述实施例中,在对用户数据进行聚类时,对不同维度的特征数据分别设置不同的权重参数,可以很好地对离散的用户数据进行有针对性地处理,进而更加准确地根据不同维度的特征数据的重要程度对用户数据进行聚类,提高用户数据的分类准确性,以提高用户的价值划分准确性。

[0089] 在一个实施例中,步骤206包括:获取每个维度相应的权重;对每个目标用户数据,保留预设数量的、且权重最大的维度的特征数据,获得用于聚类的目标用户子数据;对于任意两个目标用户子数据的特征向量,将各维度的特征值按照各维度的权重计算距离,得到任意两个目标用户子数据之间的相似程度。

[0090] 可以理解,在步骤202获取的用户数据是原始的用户数据时,在计算用户数据之间的相似程度时,可以将所有维度的特征数据均考虑进去,即按照上一个实施例中的步骤计算得到任意两个目标用户数据之间的相似程度。在步骤202获取的用户数据是特定维度的用户数据时,按照上一个实施例中的步骤计算得到任意两个目标用户数据之间的相似程度,则考虑的是特定维度的特征数据,可以在减少计算量的同时,满足当前场景下对于用户数据相似程度计算的需求。

[0091] 在本实施例中,在步骤202获取的用户数据是原始的用户数据时,还可继续对原始的用户数据进行维度筛选,保留预设数量的、且权重最大的维度的特征数据,获得用于聚类的目标用户子数据,然后对这些目标用户子数据进行聚类。此时,可以认为任意两个目标用户子数据之间的相似程度,即可以代表相应的两个目标用户数据之间的相似程度,进而代表相应的用户之间的相似程度。

[0092] 步骤208,基于各目标用户数据之间的相似程度对目标用户数据进行聚类,得到各目标用户数据所属的类别。

[0093] 具体地,计算机设备在计算得到这些目标用户数据中两两之间的相似程度后,可根据这些相似程度对目标用户数据进行聚类,得到多个聚类簇。每个聚类簇对应一个类别。由聚类所生成的聚类簇是一组目标用户数据的集合,同一个聚类簇中的目标用户数据之间

的相似程度高,不同聚类簇中的目标用户数据之间的相似程度低。

[0094] 继续参考图3,计算机设备在筛选出种子用户数据之后,即可对种子用户数据进行聚类,得到多个聚类簇。每个聚类簇对应一个类别。可以理解,对用户数据的分类可以用于对用户进行分组,从而得到用户的分组结果。

[0095] 参考图4,该图示出了一个实施例中目标用户数据进行聚类后的聚类效果示意图。对用户数据的分类可以用于对用户进行价值分组,如图4所示,在该实施例中,计算机设备采用权重最大的三个维度的特征数据对用户数据进行聚类,该图的三维坐标分别代表不同的特征数据的维度。计算机设备通过聚类将用户分为八组:重要价值用户、重要保持用户、重要发展用户、重要挽留用户、一般价值用户、一般保持用户、一般发展用户以及一般挽留用户。

[0096] 步骤210,根据目标用户数据和目标用户数据所属的类别训练分类模型。

[0097] 具体地,计算机设备在得到各目标用户数据所属的类别后,可以将各目标用户数据所属的类别作为各目标用户数据相应的训练标签,然后根据目标用户数据及其相应的训练标签,有监督地训练分类模型。

[0098] 其中,分类模型训练完成后,可以用于对用户数据进行分类。分类模型对用户数据进行分类得到的类别包括于目标用户数据所属的这些类别。

[0099] 在一个实施例中,在采用目标用户数据训练分类模型时,可采用所有维度的特征数据作为训练样本,也可以采用部分维度的特征数据作为训练样本。但无论采用怎样的特征数据,用户数据所属的类别不发生变化。

[0100] 在一个实施例中,步骤210,包括:对每个目标用户数据,保留预设数量的、且权重最大的维度的特征数据,获得用于训练分类模型的目标用户子数据;根据目标用户子数据和目标用户子数据所属类别训练分类模型。

[0101] 假设,权重为 $\delta = \{\delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \dots, \delta_n\}$,预设数量为3且权重大小排序前三为 δ_3 、 δ_5 和 δ_1 ,那么则包括这三个权重对应的维度的特征数据,将这些维度的特征数据向量化后作为分类模型的输入。具体地,计算机设备可以通过对用户数据经过Embedding操作,将离散的用户数据转为连续的特征向量进行表示,这样特征向量可以直接输入给分类模型进行处理。

[0102] 这样,采用部分重要程度较高的维度的用户数据进行模型训练,既可以减少模型训练的数据量和运算量,又能保障模型训练的效果。

[0103] 步骤212,通过训练得到的分类模型对筛选后剩余的用户数据进行分类,得到筛选后剩余的各用户数据所属的类别。

[0104] 具体地,计算机设备可根据分类模型训练时的训练样本的数据形式,对筛选后剩余的用户数据进行相应的处理后,通过训练得到的分类模型对筛选后剩余的用户数据进行分类,得到筛选后剩余的各用户数据所属的类别。这里的数据形式包括特征数据的维度,特征数据的表示方式。

[0105] 在一个实施例中,步骤212,包括:对每个筛选后剩余的用户数据,保留预设数量的、且权重最大的维度的特征数据,获得用于使用分类模型的用户子数据;通过训练得到的分类模型对用户子数据进行分类,得到各用户子数据所属的类别。

[0106] 参考图5,该图示出了一个实施例中对筛选后剩余的用户数据进行分类后的分类

效果示意图。参考图5,分类模型在训练时采用的是种子用户数据权重最大的三个维度的特征数据作为训练样本,且种子用户数据采用权重最大的三个维度的特征数据进行聚类,那么对于非种子用户数据在使用分类模型时,则同样需要采用权重最大的三个维度的特征数据,该图的三维坐标分别代表不同的特征数据的维度。分类模型输出的分类结果也包括于对种子用户数据聚类得到的八个分组:重要价值用户、重要保持用户、重要发展用户、重要挽留用户、一般价值用户、一般保持用户、一般发展用户以及一般挽留用户。

[0107] 继续参考图3,计算机设备在对种子用户数据聚类,得到带有类别的种子用户数据后,可采用带有类别的种子用户数据训练分类模型,在分类模型训练完成后,利用分类模型对剩余的非种子用户数据进行分类,得到各非种子用户数据所述的类别。可以理解,对用户数据的分类可以用于对用户进行分组,从而得到用户的分组结果。

[0108] 步骤214,根据各用户数据所属的类别进行信息推送。

[0109] 具体地,计算机设备在得到各用户数据所属的类别后,可根据各用户数据所属的类别进行有针对性地信息推送。根据各用户数据所属的类别进行信息推送,具体可以是根据各用户数据所属的类别,对各用户数据所对应的用户进行信息推送。对于属于不同类别的用户可以推送不同的数据。

[0110] 其中,推送的信息可以是商品、新闻、视频、资源或者音频等等。

[0111] 在一个实施例中,步骤214,包括:获取各个类别相应的推送信息;向各用户数据所对应的用户终端,推送与各用户数据所属类别相应的推送信息。

[0112] 其中,推送信息是具体应用场景下向用户推送的信息。比如视频推广场景下的视频消息,或者商品推广场景下的商品消息等等。通过对用户进行分类,利用用户不同价值分组进行不同的业务处理,比如采用不同的红包奖励等。

[0113] 可以理解,高维数据计算距离都会出现内存爆炸,资源不够的情况,同时耗时严重,无论是欧式距离、曼哈顿距离,切比雪夫距离等都存在类似的情况。在面对数量巨大的高维用户数据集时,无法像传统聚类算法K-Means、DBSCAN以及层次聚类一样,将全部数据载入内存进行两两计算。在本申请中,巧妙地提出了大数据分而治之的处理方式,通过筛选少量的种子用户数据,对种子用户数据进行聚类,得到带有类别的种子用户数据,再用带有类别的种子用户数据,对分类模型进行有监督训练,得到更加精确的分类模型。再利用训练得到的分类模型对大量剩余的非种子用户数据进行分类。

[0114] 而且本申请一方面对不同特征设置不同的权重参数,通过加权的方式量化用户数据之间的相似程度,即用户距离,可以很好地进行具体场景下用户数据的分类,进而更好的区分具体场景下用户的价值;而且权重参数,根据机器学习模型学习大量用户数据得到的,客观且准确。另一方面对分类模型进行了设计,使得分类模型能够更加精确的学习用户的特征数据,学习之后的分类模型可以准确的对用户数据进行分类,从而可以很好地对用户价值进行识别。

[0115] 在得到了用户数据的类别之后,可以得到用户的价值分组,进而根据用户的价值分组,进行不同的处理。比如根据用户的价值分组,进行不同的营销活动,对不同的用户采用不同的红包奖励,高价值用户会得到更高的奖励,低价值用户会推送不同的新品,进行更大的折扣激活用户的购买能力,等产品方案。

[0116] 上述用户数据处理方法,在获取到大量多维的用户数据后,筛选出少量用户数据

作为目标来进行聚类 and 类别划分,得到带有类别的目标用户数据,然后利用带有类别的目标用户数据来训练分类模型,再利用训练得到的分类模型对筛选后剩余的用户数据进行分类。这样,一方面在聚类时,对不同维度的特征数据设置不同的权重参数,可以很好地对离散的用户数据进行有针对性处理,进而更加准确地根据不同维度的特征数据的重要程度对用户数据进行聚类,提高用户分类准确性;而且基于该部分带有类别的目标用户数据训练的分类模型,也能准确有效地对筛选后剩余的用户数据进行分类;另一方面,仅对部分用户数据进行聚类,还可以避免全量数据聚类带来的计算量和可能存在的计算资源不足的问题;另外在得到用户数据所属的类别后,可以更加准确且有针对性地进行信息推送。

[0117] 在一个实施例中,获取每个维度相应的权重,包括:获取训练完成的排序模型所输出的权重向量;其中,排序模型通过存在训练标签的用户数据样本监督训练得到,并在监督训练的过程中对用户样本数据中多个维度的特征数据的影响程度排序;权重向量包括多个维度的特征数据的权重。

[0118] 其中,排序模型可以为树模型、SVM(Support Vector Machine,支持向量机)模型、LR(Logistic Regression,逻辑回归)模型、神经网络模型等多种类型的模型,或者排序模型也可以为树模型、SVM(Support Vector Machine,支持向量机)模型、LR(Logistic Regression,逻辑回归)模型、神经网络模型等多种类型算法结合的模型,或者排序模型还可以采用SVM模型、LR模型或者神经网络模型结合SHAP(Shapley AdditiveexPlanations,模型可解释性)、Permutation(置换)等可解释性算法来得到,本申请实施例对此不做限定。

[0119] 具体地,当排序模型为树模型时,该树模型可以包括GBDT(Gradient Boosting Decision Tree,梯度下降树)模型、LightGBM(Light Gradient Boosting Machine,轻型梯度提升树)模型或者LightGBM与GBDT的结合等。LightGBM模型采用histogram(直方图)算法进行特征选择,将连续的多个数值转换为直方图中的预设数量个离散值,计算效率高。LightGBM模型抛除了按层生长的策略,采用按叶子生长的策略,在分裂次数相同的情况下,可以减少不必要的搜索和分裂,提高了模型的准确率。

[0120] 排序模型所输出的权重向量具体可以为 $\delta = \{\delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \dots, \delta_n\}$ 。其中, δ_1 可以表示维度1的特征数据对应的权重, δ_2 可以表示维度2的特征数据对应的权重, δ_3 可以表示维度3的特征数据对应的权重,以此类推,n为维度数量。

[0121] 在一个实施例中,排序模型的训练步骤包括:获取用户数据样本和用户数据样本对应的训练标签;用户数据样本包括多个维度的特征数据;通过排序模型对用户数据样本进行预测,得到预测结果;根据预测结果和训练标签之间的差异,优化排序模型;通过排序模型对用户样本数据中多个维度的特征数据对预测的影响程度进行排序,并输出排序结果。

[0122] 其中,训练排序模型所通过的用户数据样本的训练标签可以与用户数据样本的重要程度相关。可以理解,对各个不同维度的特征数据设置权重,实质上是对各个不同维度的特征数据对用户数据分类的影响程度进行排序,哪个维度的特征数据对用户数据的重要程度划分的影响程度更大,该维度的特征数据所对应的权重即更高。

[0123] 具体地,计算机设备可将用户数据样本划分为两类,一类认为是重要的用户数据样本作为正样本,另一类认为是不重要的用户数据样本作为负样本。通过正样本和负样本对排序模型进行有监督地训练。排序模型的输出包括两种数据,一种是对用户数据样本的

分类预测,另一种是对各个不同维度的特征数据的重要性排序。排序模型在不断地迭代训练中,学会关注对分类结果准确率提高更有用的维度的特征数据,进而可以实现对各个不同维度的特征数据对用户数据分类的影响程度进行排序。排序模型在训练完后会输出一个稳定的排序结果。该排序结果中包括各个维度的特征数据对应的权重,即为权重向量。

[0124] 在一个实施例中,该用户数据处理方法还包括:获取权重向量中数值最大的向量元素和数值最小的向量元素;确定数值最大的向量元素和数值最小的向量元素之间的差值;根据数值最小的向量元素和差值,对权重向量中的各向量元素进行归一化处理。

[0125] 具体地,排序模型输出权重向量为 $\delta = \{\delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \dots, \delta_n\}$,获取该权重向量中数值最大的向量元素 $\delta_{\max}$ 和数值最小的向量元素 $\delta_{\min}$,确定数值最大的向量元素和数值最小的向量元素之间的差值 $\delta_{\max} - \delta_{\min}$,根据数值最小的向量元素和差值,对权重向量中的各向量元素进行归一化处理:

$$[0126] \quad \text{Normal_}\delta = (\delta - \delta_{\min}) / (\delta_{\max} - \delta_{\min}) \quad (2)$$

[0127] 进一步地,计算机设备可以采用归一化处理后的权重,对用户数据计算相似程度。归一化是一种无量纲处理手段,使物理系统数值的绝对值变成某种相对值关系,可以简化计算,缩小量值。

[0128] 上述实施例中,根据机器学习模型学习大量用户数据得到各个维度的特征数据所对应的权重参数,客观、可靠且准确程度高。

[0129] 在一个实施例中,分类模型包括多个分类子结构;每个分类子结构包括注意力网络结构;注意力网络结构的处理步骤包括:对输入注意力网络结构的特征向量分别赋予权重,得到键向量、请求向量和值向量;通过注意力网络结构对键向量、请求向量和值向量进行处理,得到处理结果。

[0130] 具体地,本申请创造性地设计了分类模型的模型结构,该分类模型包括多个分类子结构,每个分类子结构包括注意力网络结构。其中,注意力网络结构是基于注意力机制(Attention Mechanism)的网络结构。注意力机制是基于编码器与解码器的隐藏状态之间的依赖关系建立模型的方式。多个注意力网络结构可以分别用来捕获不同表示空间的特征数据。每个注意力网络结构处理数据后均会得到一个表示空间下的特征数据。这样,注意力网络结构可以使得分类模型更加关注重要的特征,学习到对分类模型的分类目的更加有用的信息,侧重关注多维特征中某些维度的特征信息。

[0131] 分类模型的输入是将离散的用户数据经过Embedding操作向量化得到的特征向量(User Feature Embedding)。其中,在对离散的用户数据进行处理时,可以仅对权重较高的预设数量个维度的特征数据进行Embedding操作。比如,仅对权重最高的3个维度的特征数据进行Embedding操作,得到三维特征向量。

[0132] 离散的用户数据向量化得到的特征向量(User Feature Embedding)输入分类模型后,依次通过分类模型的每一个分类子结构进行处理,得到新的特征向量输入下一个分类子结构进行处理,直至分类模型的输出层输出分类结果。

[0133] 分类模型的输出是对用户数据进行分类的分类结果。比如,分类的类别有8个,那么分类模型的输出可以是(0,1,2,3,4,5,6,7)中各自对应的概率,概率最大的类别即为用户数据所属的类别。

[0134] 举例说明,参考图6,该图示出了一个实施例中分类模型的结构示意图。由该图可

以看出,分类模型(SA-NET)包括多个分类子结构,每个分类子结构中包括一个注意力网络结构(SA,SP-Attention)。相邻的分类子结构之间可直接相连,也可通过池化层过渡连接。分类模型最后通过回归层输出分类结果。其中,池化层的池化操作比如最大池化(Max Pooling)或者全局最大池化(Global Max Pooling)等。回归层比如Softmax层等。

[0135] 参考图7,该图示出了一个实施例中注意力网络结构(SA,SP-Attention)的结构示意图。由该图可以看出,输入到注意力网络结构的特征向量 z 被分别赋予权重,得到键向量 $K(z) = w_2 z_2$,请求向量 $Q(z) = w_1 z_1$ 和值向量 $V(z) = w_3 z_3$, $z = z_1 = z_2 = z_3$ 。其中, w_1 、 w_2 和 w_3 是分类模型在训练时需要学习的权重参数。

[0136] 注意力网络结构输出如下式所示:

$$[0137] \quad SA(Q, K, V) = softmax(\frac{QK^T}{\sqrt{|Q-K|}} + \frac{KQ^T}{\sqrt{|Q-K|}})V \quad (3)$$

[0138] 其中, K^T 是 K 的转置, $\sqrt{|Q-K|}$ 表示 Q 与 K 的哈夫曼距离。

[0139] 在一个实施例中,输入到注意力网络结构的特征向量 z 被分别赋予权重,得到键向量 $K(z) = w_2 z_2$,请求向量 $Q(z) = w_1 z_1$ 和值向量 $V(z) = w_3 z_3$ 时, z_1 、 z_2 和 z_3 也可以是 z 的不同维度的特征数据。

[0140] 在一个实施例中,通过注意力网络结构对键向量、请求向量和值向量进行处理,得到处理结果,包括:通过注意力网络结构的非线性激活函数层,将键向量、请求向量和值向量映射为多于一组键向量、请求向量和值向量;通过注意力网络结构的多个注意力机制层,分别对多于一组键向量、请求向量和值向量进行处理,得到中间结果;通过注意力网络结构的拼接层和卷积层依次对中间结果进行处理,得到处理结果。

[0141] 继续参考图7,在得到键向量 $K(z) = w_2 z_2$ 、请求向量 $Q(z) = w_1 z_1$ 和值向量 $V(z) = w_3 z_3$ 后,通过注意力网络结构的非线性激活函数层,如prelu激活函数层,将键向量、请求向量和值向量映射为多于一组键向量、请求向量和值向量。比如,映射为 h 组,每组键向量、请求向量和值向量均相同,都是, $K(z) = w_2 z_2$ 、 $Q(z) = w_1 z_1$ 和 $V(z) = w_3 z_3$ 。进而通过注意力网络结构的多个注意力机制层,如Scaled Dot-Product Attention,分别对多于一组键向量、请求向量和值向量进行处理,得到中间结果;再通过注意力网络结构的拼接层(Concatenate)和卷积层(Convolution)依次对中间结果进行处理,得到处理结果 $SA(Q, K, V)$ 。

[0142] 在一个实施例中,分类子结构还包括卷积网络结构和批标准化网络结构;分类子结构的处理步骤包括:通过卷积网络结构对输入分类子结构的数据进行卷积运算,并将卷积运算结果输出至批标准化网络结构;通过批标准化网络结构对卷积运算结果进行分布调整,并将调整结果输出至注意力网络结构。

[0143] 其中,卷积网络结构是用于对数据进行卷积操作的网络结构。不同的分类子结构所包括的卷积网络结构中卷积层的数量可以相同也可以不相同。

[0144] 批标准化网络结构(BN, Batch Normalization)用于在模型的训练过程中使得每一层网络结构的输入保持相同的分布。批标准化网络结构可以加快模型的训练速度,提高模型的泛化能力。

[0145] 继续参考图6,可以看到分类网络包括13个分类子结构,第5个分类子结构与第6个分类子结构以及第10个分类子结构与第11个分类子结构之间通过Max Pooling过渡连接,

其他分类子结构之间直接连接,第13个分类子结构之后通过全局最大池化操作(Global Max Pooling)、内积操作(Inner Product)以及归一化操作(Softmax)之后输出分类结果。

[0146] 在一个具体场景中,第1个分类子结构的卷积层(Convolution)包括66个3x3的卷积核;第2、3、4个分类子结构的卷积层(Convolution)包括128个3x3的卷积核;第5、6、7、8、9个分类子结构的层卷积层(Convolution)包括192个3x3的卷积核;第10、11个分类子结构的卷积层(Convolution)包括288个3x3的卷积核;第12个分类子结构的卷积层(Convolution)包括355个3x3的卷积核;第13个分类子结构的层卷积层(Convolution)包括432个3x3的卷积核。当然,分类子结构的卷积层的卷积核的数量和尺寸还可以是其他情形。

[0147] 应该理解的是,虽然上述实施例的流程图中的各个步骤按照箭头的指示依次显示,但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本文中有明确的说明,这些步骤的执行并没有严格的顺序限制,这些步骤可以以其它的顺序执行。而且,上述实施例的流程图中的至少一部分步骤可以包括多个步骤或者多个阶段,这些步骤或者阶段并不必然是在同一时刻执行完成,而是可以在不同的时刻执行,这些步骤或者阶段的执行顺序也不必然是依次进行,而是可以与其它步骤或者其它步骤中的步骤或者阶段的至少一部分轮流或者交替地执行。

[0148] 在一个实施例中,如图8所示,提供了一种用户数据处理装置,该用户数据处理装置可以采用软件模块或硬件模块,或者是二者的结合成为计算机设备的一部分,该用户数据处理装置具体包括:获取模块801、第一筛选模块802、确定模块803、聚类模块804、训练模块805、第二筛选模块806和应用模块807。

[0149] 其中,

[0150] 获取模块801,用于获取多个用户数据;每个用户数据包括多个维度的特征数据;

[0151] 第一筛选模块802,用于从用户数据中筛选出多个目标用户数据;

[0152] 确定模块803,用于根据各目标用户数据在每个维度的特征数据以及每个维度相应的权重,确定各目标用户数据之间的相似程度;

[0153] 聚类模块804,用于基于各目标用户数据之间的相似程度对目标用户数据进行聚类,得到各目标用户数据所属的类别;

[0154] 训练模块805,用于根据目标用户数据和目标用户数据所属的类别训练分类模型;

[0155] 第二筛选模块806,用于通过训练得到的分类模型对筛选后剩余的用户数据进行分类,得到筛选后剩余的各用户数据所属的类别;

[0156] 应用模块807,用于根据各用户数据所属的类别进行信息推送。

[0157] 在一个实施例中,确定模块803还用于向量化各目标用户数据,得到各目标用户数据的特征向量;特征向量包括多个维度的特征值;获取每个维度相应的权重;对于任意两个目标用户数据的特征向量,将各维度的特征值按照各维度的权重计算距离,得到任意两个目标用户数据之间的相似程度。

[0158] 在一个实施例中,确定模块803还用于获取每个维度相应的权重;对每个目标用户数据,保留预设数量的、且权重最大的维度的特征数据,获得用于聚类的目标用户子数据;对于任意两个目标用户子数据的特征向量,将各维度的特征值按照各维度的权重计算距离,得到任意两个目标用户子数据之间的相似程度。

[0159] 在一个实施例中,确定模块803还用于获取训练完成的排序模型所输出的权重向

量;其中,排序模型通过存在训练标签的用户数据样本监督训练得到,并在监督训练的过程中对用户样本数据中多个维度的特征数据的影响程度排序;权重向量包括多个维度的特征数据的权重。

[0160] 在一个实施例中,训练模块805还用于获取用户数据样本和用户数据样本对应的训练标签;用户数据样本包括多个维度的特征数据;通过排序模型对用户数据样本进行预测,得到预测结果;根据预测结果和训练标签之间的差异,优化排序模型;通过排序模型对用户样本数据中多个维度的特征数据对预测的影响程度进行排序,并输出排序结果。

[0161] 在一个实施例中,确定模块803还用于获取权重向量中数值最大的向量元素和数值最小的向量元素;确定数值最大的向量元素和数值最小的向量元素之间的差值;根据数值最小的向量元素和差值,对权重向量中的各向量元素进行归一化处理。

[0162] 在一个实施例中,训练模块805还用于对每个目标用户数据,保留预设数量的、且权重最大的维度的特征数据,获得用于训练分类模型的目标用户子数据;根据目标用户子数据和目标用户子数据所属类别训练分类模型。第二筛选模块806还用于对每个筛选后剩余的用户数据,保留预设数量的、且权重最大的维度的特征数据,获得用于使用分类模型的用户子数据;通过训练得到的分类模型对用户子数据进行分类,得到各用户子数据所属的类别。

[0163] 在一个实施例中,分类模型包括多个分类子结构;每个分类子结构包括注意力网络结构。第二筛选模块806还用于对输入注意力网络结构的特征向量分别赋予权重,得到键向量、请求向量和值向量;通过注意力网络结构对键向量、请求向量和值向量进行处理,得到处理结果。

[0164] 在一个实施例中,第二筛选模块806还用于通过注意力网络结构的非线性激活函数层,将键向量、请求向量和值向量映射为多于一组键向量、请求向量和值向量;通过注意力网络结构的多个注意力机制层,分别对多于一组键向量、请求向量和值向量进行处理,得到中间结果;通过注意力网络结构的拼接层和卷积层依次对中间结果进行处理,得到处理结果。

[0165] 在一个实施例中,分类子结构还包括卷积网络结构和批标准化网络结构。第二筛选模块806还用于通过卷积网络结构对输入分类子结构的数据进行卷积运算,并将卷积运算结果输出至批标准化网络结构;通过批标准化网络结构对卷积运算结果进行分布调整,并将调整结果输出至注意力网络结构。

[0166] 在一个实施例中,第一筛选模块802还用于对用户数据进行数据清洗;将清洗后的用户数据进行多级划分,得到多个用户数据子集;从各个最小数量的用户数据子集中随机选取用户数据作为目标用户数据。

[0167] 在一个实施例中,应用模块807还用于获取各个类别相应的推送信息;向各用户数据所对应的用户终端,推送与各用户数据所属类别相应的推送信息。

[0168] 上述用户数据处理装置,在获取到大量多维的用户数据后,筛选出少量用户数据作为目标来进行聚类 and 类别划分,得到带有类别的目标用户数据,然后利用带有类别的目标用户数据来训练分类模型,再利用训练得到的分类模型对筛选后剩余的用户数据进行分类。这样,一方面在聚类时,对不同维度的特征数据设置不同的权重参数,可以很好地对离散的用户数据进行有针对性处理,进而更加准确地根据不同维度的特征数据的重要程度对

用户数据进行聚类,提高用户分类准确性;而且基于该部分带有类别的目标用户数据训练的分类模型,也能准确有效地对筛选后剩余的用户数据进行分类;另一方面,仅对部分用户数据进行聚类,还可以避免全量数据聚类带来的计算量和可能存在的计算资源不足的问题;另外在得到用户数据所属的类别后,可以更加准确且有针对性地进行信息推送。

[0169] 关于用户数据处理装置的具体限定可以参见上文中对于用户数据处理方法的限定,在此不再赘述。上述用户数据处理装置中的各个模块可全部或部分通过软件、硬件及其组合来实现。上述各模块可以硬件形式内嵌于或独立于计算机设备中的处理器中,也可以以软件形式存储于计算机设备中的存储器中,以便于处理器调用执行以上各个模块对应的操作。

[0170] 在一个实施例中,提供了一种计算机设备,该计算机设备可以是服务器,其内部结构图可以如图9所示。该计算机设备包括通过系统总线连接的处理器、存储器和网络接口。其中,该计算机设备的处理器用于提供计算和控制能力。该计算机设备的存储器包括非易失性存储介质和内存。该非易失性存储介质存储有操作系统、计算机程序和数据库。该内存为易失性存储介质中的操作系统和计算机程序的运行提供环境。该计算机设备的数据库用于存储用户数据等。该计算机设备的网络接口用于与外部的终端通过网络连接通信。该计算机程序被处理器执行时以实现一种由服务器执行的用户数据处理方法。

[0171] 本领域技术人员可以理解,图9中示出的结构,仅仅是与本申请方案相关的部分结构的框图,并不构成对本申请方案所应用于其上的计算机设备的限定,具体的计算机设备可以包括比图中所示更多或更少的部件,或者组合某些部件,或者具有不同的部件布置。

[0172] 在一个实施例中,还提供了一种计算机设备,包括存储器和处理器,存储器中存储有计算机程序,该处理器执行计算机程序时实现上述各方法实施例中的步骤。

[0173] 在一个实施例中,提供了一种计算机可读存储介质,存储有计算机程序,该计算机程序被处理器执行时实现上述各方法实施例中的步骤。

[0174] 在一个实施例中,提供了一种计算机程序产品或计算机程序,该计算机程序产品或计算机程序包括计算机指令,该计算机指令存储在计算机可读存储介质中。计算机设备的处理器从计算机可读存储介质读取该计算机指令,处理器执行该计算机指令,使得该计算机设备执行上述各方法实施例中的步骤。

[0175] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程,是可以通过计算机程序来指令相关的硬件来完成,所述的计算机程序可存储于一非易失性计算机可读存储介质中,该计算机程序在执行时,可包括如上述各方法的实施例的流程。其中,本申请所提供的各实施例中所使用的对存储器、存储、数据库或其它介质的任何引用,均可包括非易失性和易失性存储器中的至少一种。非易失性存储器可包括只读存储器(Read-Only Memory,ROM)、磁带、软盘、闪存或光存储器等。易失性存储器可包括随机存取存储器(Random Access Memory,RAM)或外部高速缓冲存储器。作为说明而非局限,RAM可以是多种形式,比如静态随机存取存储器(Static Random Access Memory,SRAM)或动态随机存取存储器(Dynamic Random Access Memory,DRAM)等。

[0176] 以上实施例的各技术特征可以进行任意的组合,为使描述简洁,未对上述实施例中的各个技术特征所有可能的组合都进行描述,然而,只要这些技术特征的组合不存在矛盾,都应当认为是本说明书记载的范围。

[0177] 以上所述实施例仅表达了本申请的几种实施方式,其描述较为具体和详细,但不能因此而理解为对发明专利范围的限制。应当指出的是,对于本领域的普通技术人员来说,在不脱离本申请构思的前提下,还可以做出若干变形和改进,这些都属于本申请的保护范围。因此,本申请专利的保护范围应以所附权利要求为准。

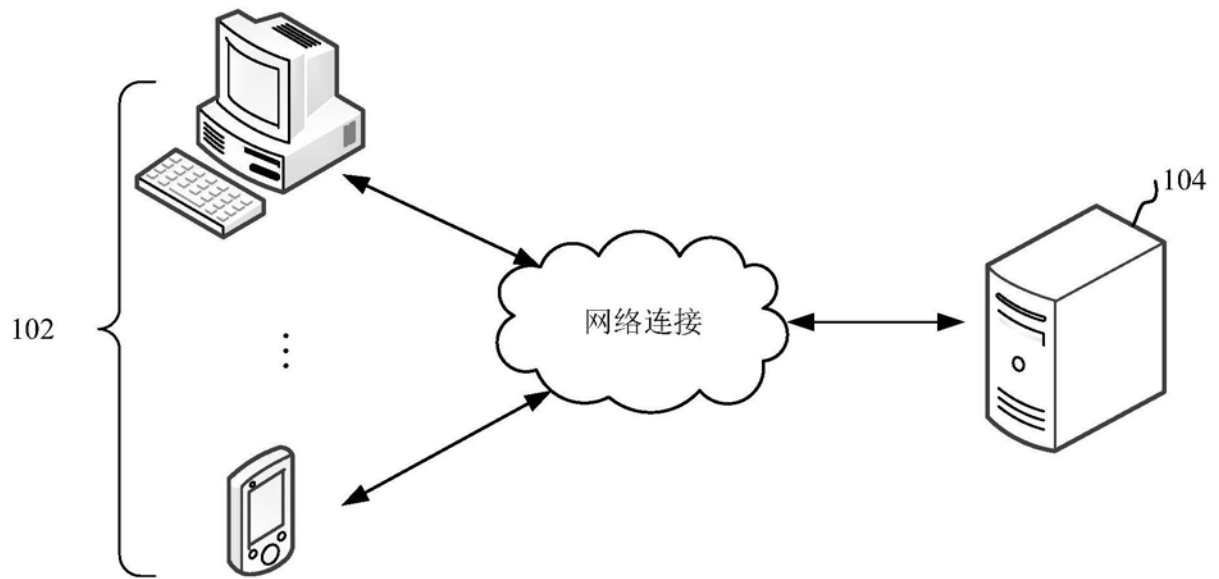


图1

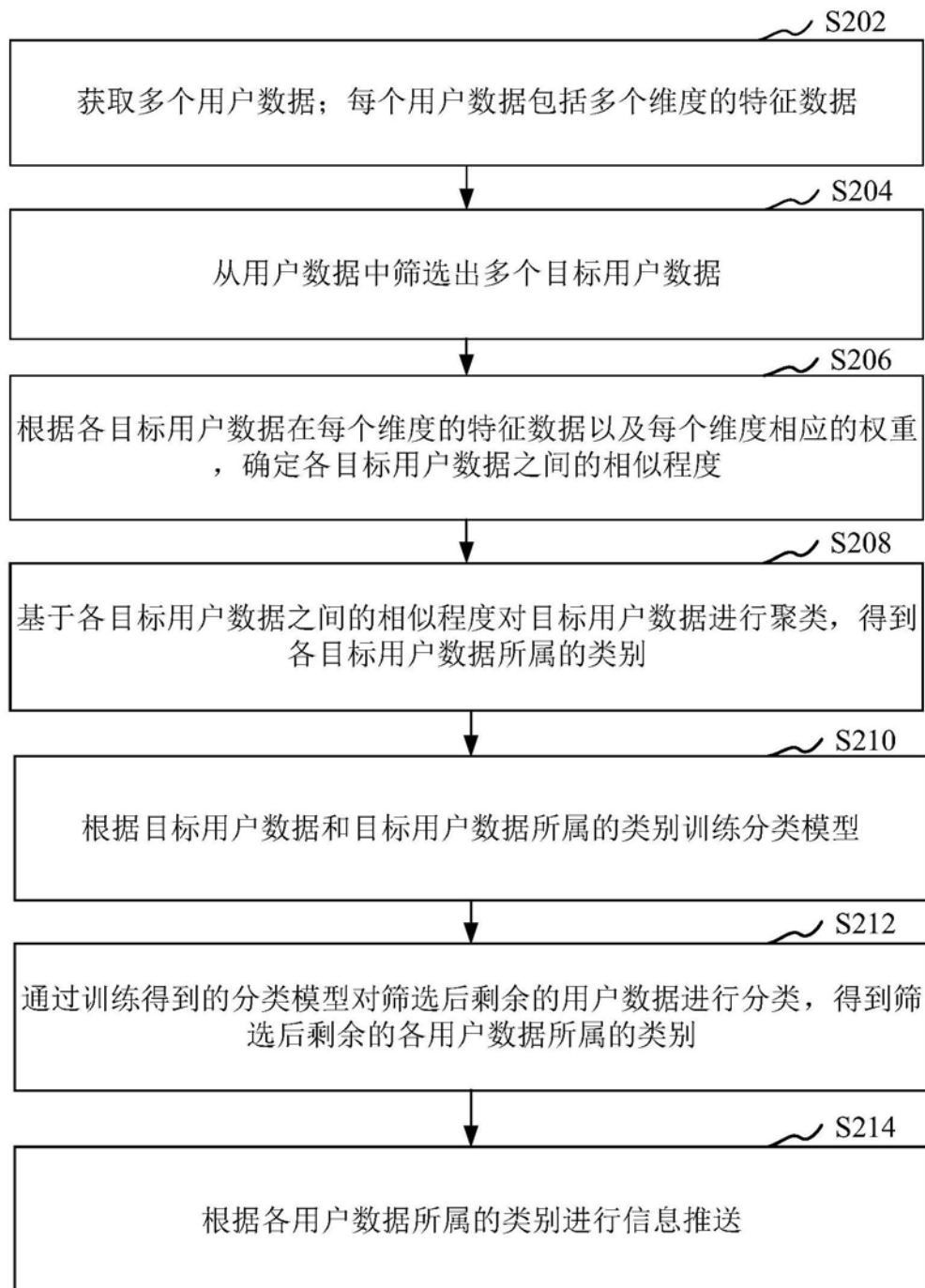


图2

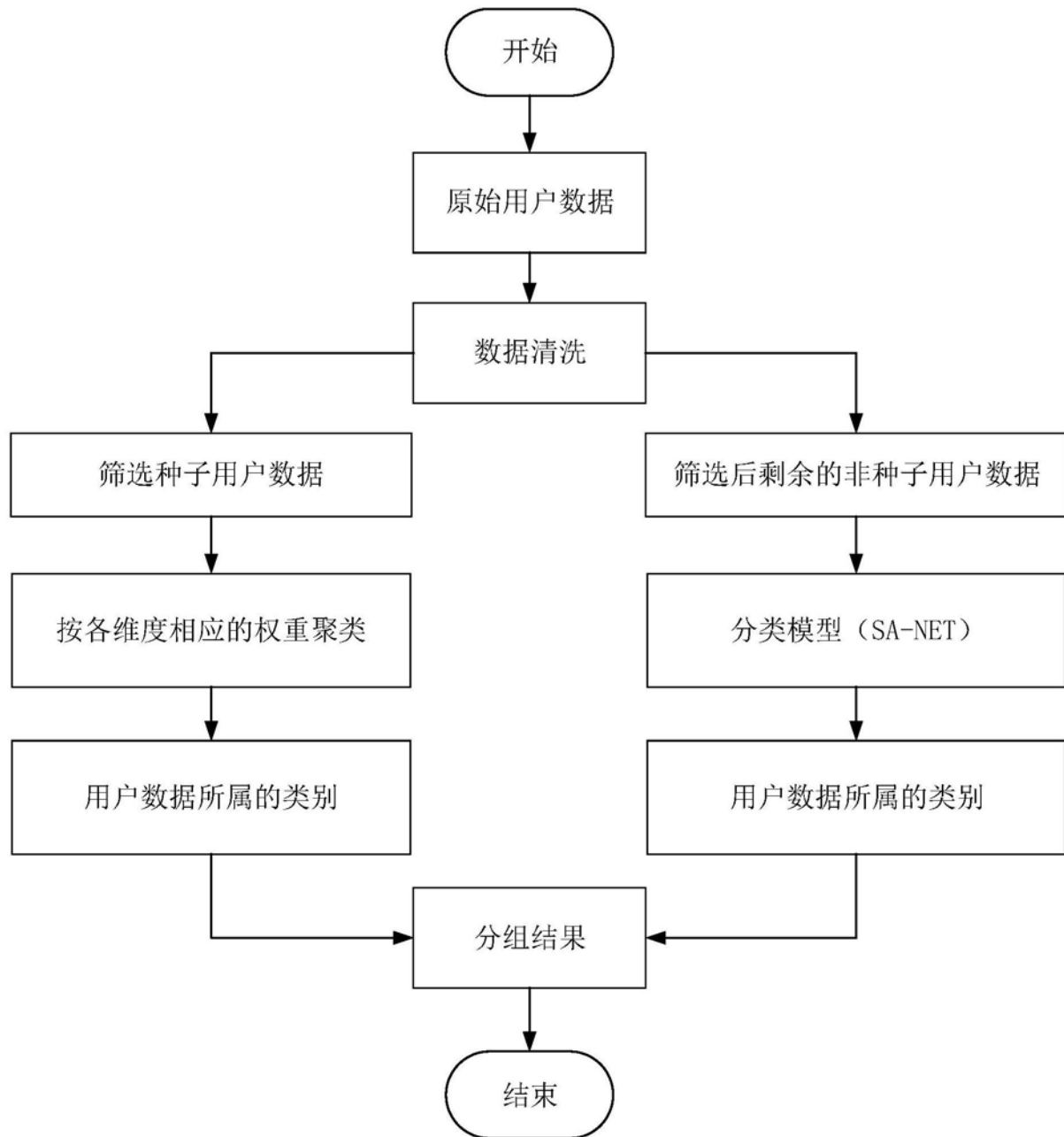


图3

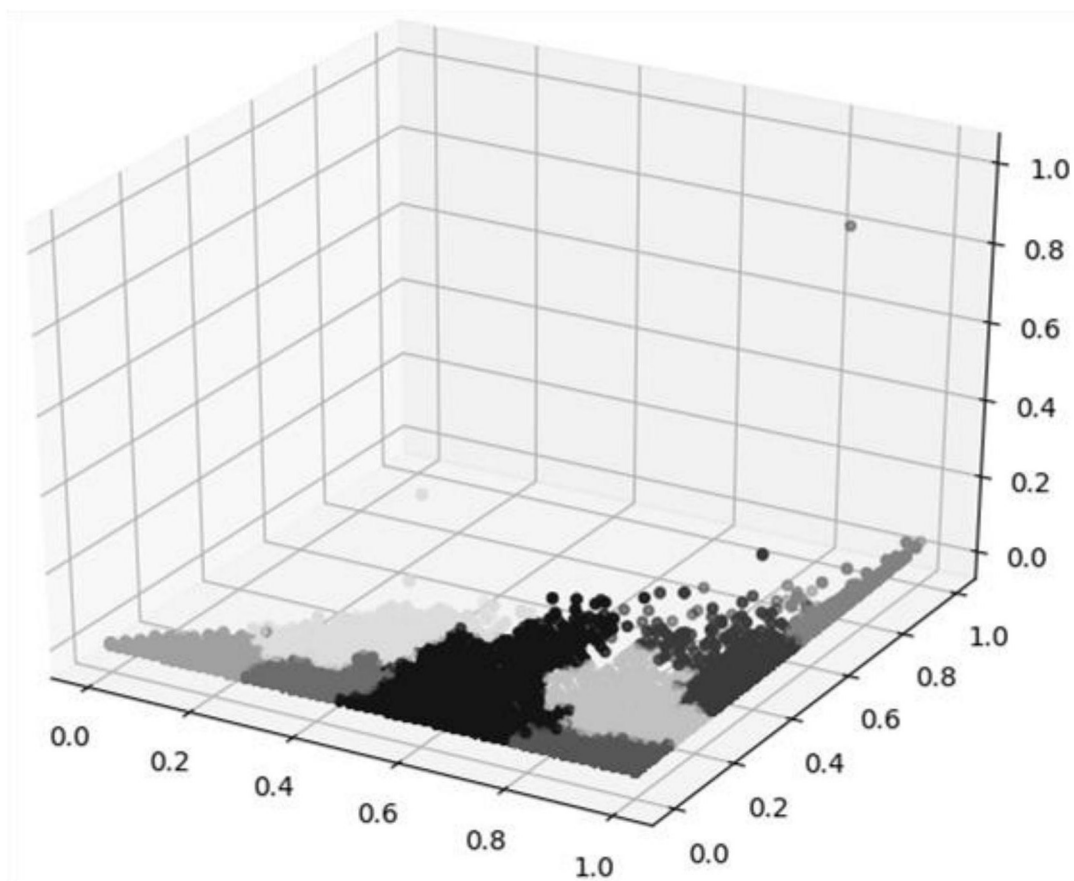


图4

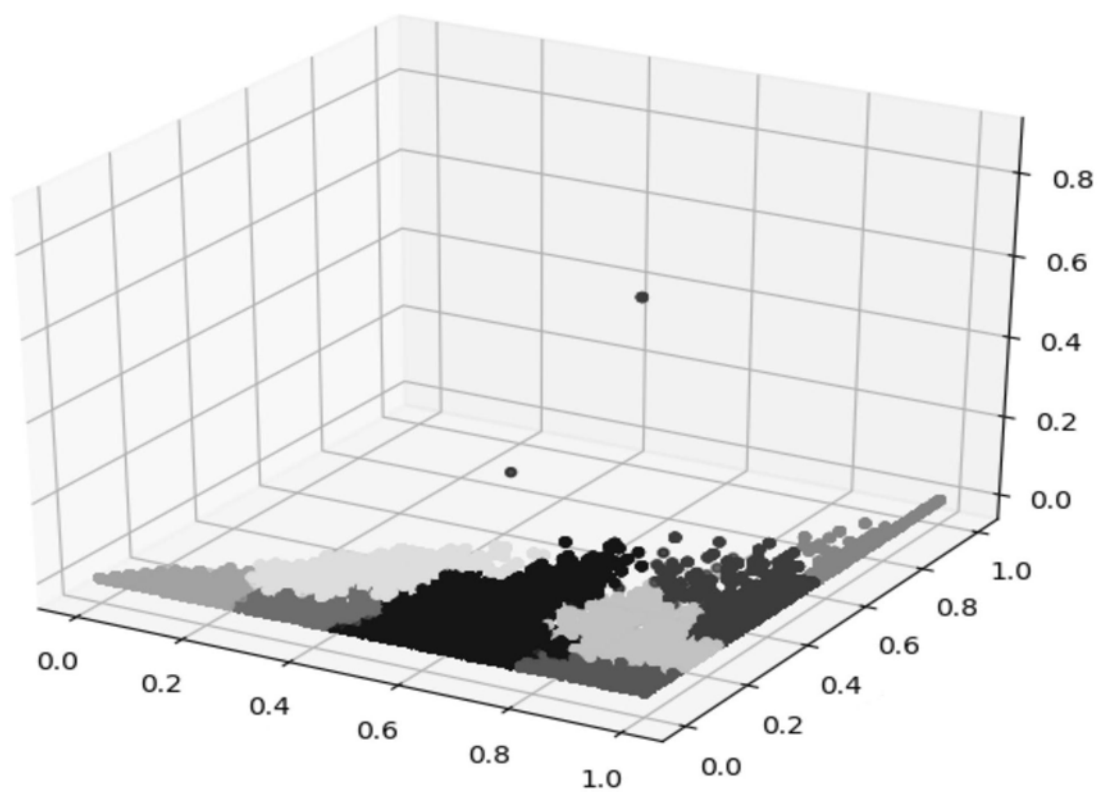


图5

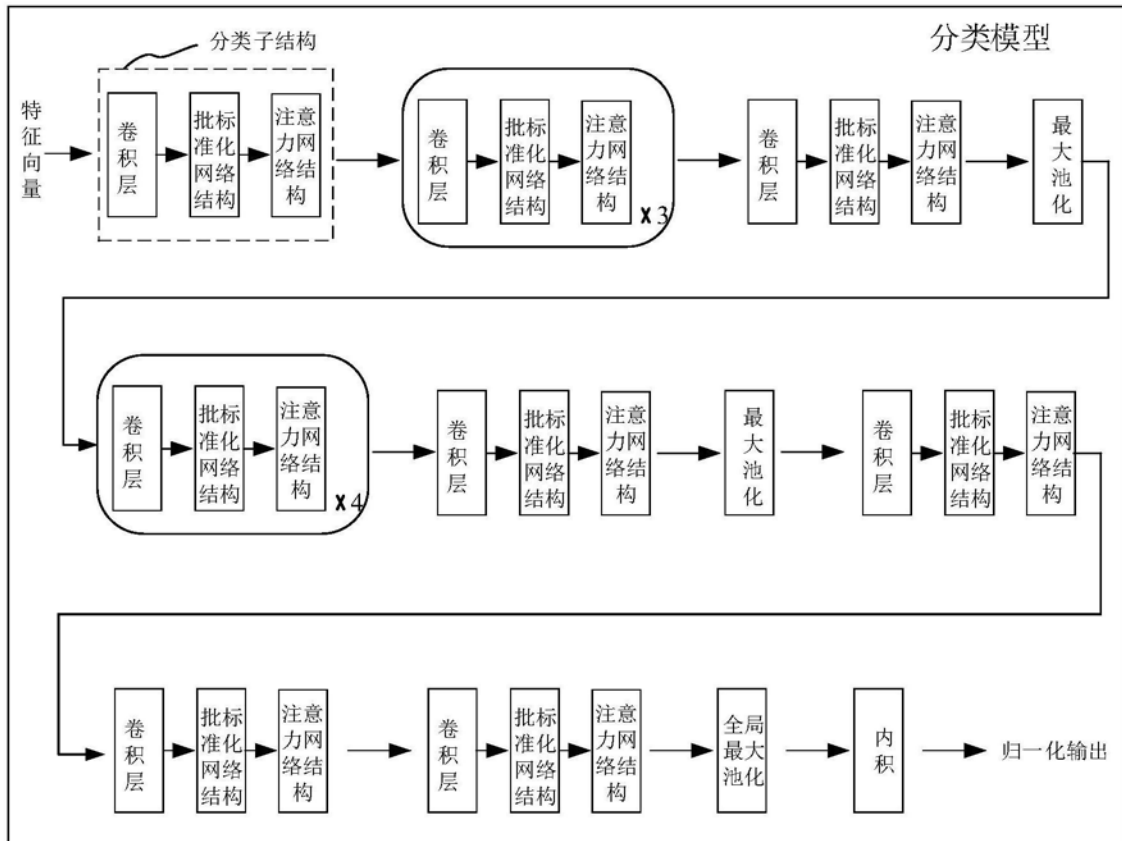


图6

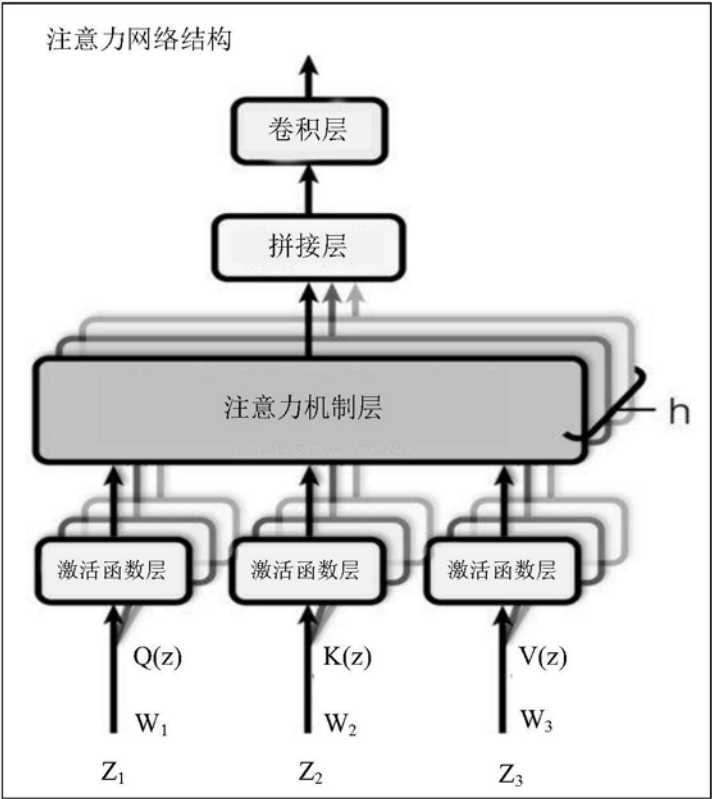


图7

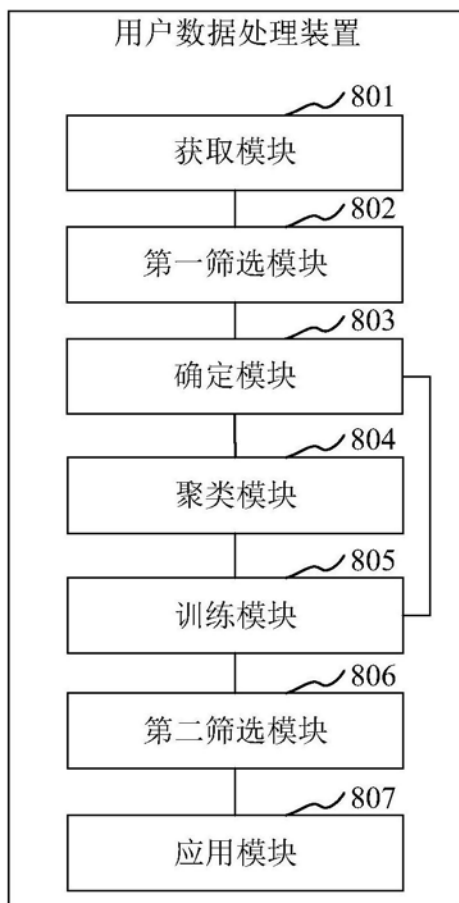


图8

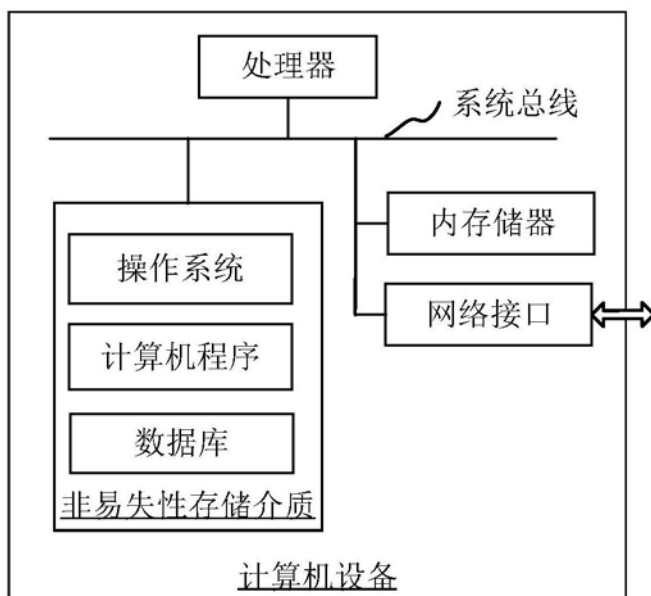


图9