(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111695042 B
(45)授权公告日 2023.04.18

(21)申请号 202010524285.4

(22)申请日 2020.06.10

(65)同一申请的已公布的文献号
申请公布号 CN 111695042 A

(43)申请公布日 2020.09.22

(73)专利权人 湖南湖大金科科技发展有限公司
地址 415000 湖南省常德市鼎城区灌溪镇(常德高新技术产业开发区樟窑路-常德科技创新创业孵化产业园第二层230号)

(72)发明人 陈佐 吴志良 杨胜刚 朱桑之 谷浩然 杨捷琳

(74)专利代理机构 成都行之专利代理事务所(普通合伙) 51220
专利代理师 林菲菲

(51)Int.Cl.
G06F 16/9536(2019.01)
G06F 16/955(2019.01)
G06F 40/30(2020.01)
G06F 18/2415(2023.01)
G06N 3/047(2023.01)
G06N 3/084(2023.01)
G06N 20/10(2019.01)
G06Q 30/0202(2023.01)

(56)对比文件
CN 107341571 A,2017.11.10
WO 2020083020 A1,2020.04.30
CN 109741112 A,2019.05.10
CN 111160483 A,2020.05.15
CN 109191240 A,2019.01.11
CN 109190030 A,2019.01.11
CN 108920641 A,2018.11.30
CN 110321494 A,2019.10.11
CN 110162690 A,2019.08.23
CN 109034960 A,2018.12.18
刘杨涛.基于嵌入式向量和循环神经网络的用户行为预测方法.现代电子技术.2016,第39卷(第23期),1-5.

审查员 张诗纬

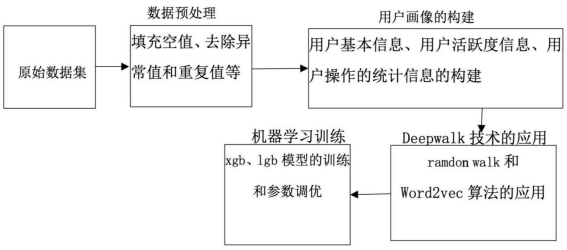
权利要求书2页 说明书16页 附图6页

(54)发明名称

基于深度游走和集成学习的用户行为预测方法及系统

(57)摘要

本发明公开了基于深度游走和集成学习的用户行为预测方法及系统,本发明对原始数据集中存在的重复、异常和冗余等问题进行了预处理工作,从预处理后的数据集中提取出能够反映消费者行为习惯和偏好程度的统计信息和活跃度信息,以此为用户构建用户画像,然后通过从用户购买商品的社交网络图结构进行随机漫步(Ramdon Walk)得到新的行为序列,然后在用Word2vec模型得到用户每个行为的上下信息加入到机器学习模型中去训练学习,提高了模型的预测可靠性和预测精度。



CN 111695042 B

1.基于深度游走和集成学习的用户行为预测方法,其特征在于,该方法包括以下步骤:

步骤S1,获取原始数据集并对其进行预处理;

步骤S2,基于预处理之后的数据集构建用户画像,形成商品社交网络图结构;

步骤S3,对商品社交网络图结构进行随机游走得到新的行为序列数据,然后利用Word2vec模型对新的行为序列数据进行训练生成embedding向量;

步骤S4,将embedding向量输入到机器学习模型中进行训练,得到单一用户行为预测模型。

2.根据权利要求1所述的基于深度游走和集成学习的用户行为预测方法,其特征在于,还包括:

步骤S5,对构造得到的多个单一用户行为预测模型中差异性最大的两个进行融合,得到用户行为预测模型。

3.根据权利要求2所述的基于深度游走和集成学习的用户行为预测方法,其特征在于,所述步骤S5具体包括:

步骤S51,通过调整随机游走的步长和embedding向量的维度,重复执行步骤S3和步骤S4,即可构造得到多个单一用户行为预测模型;

步骤S52,根据泛化能力从多个单一用户行为预测模型中选择n个模型;其中,n为大于等于3的正整数;

步骤S53,计算n个模型中每个模型与模型之间的最大信息系数MIC且构建混淆矩阵并进行可视化;

步骤S54,在得到的混淆矩阵上面找出相似度最小的两个单一模型进行融合,得到用户行为预测模型。

4.根据权利要求1或2或3所述的基于深度游走和集成学习的用户行为预测方法,其特征在于,所述步骤S2具体从三个角度进行用户画像的构建,分别是用户的基本信息、用户活跃度信息和用户操作行为统计信息。

5.根据权利要求1或2或3所述的基于深度游走和集成学习的用户行为预测方法,其特征在于,所述步骤S3中的随机游走过程具体为:从网络图结构的任意一个节点出发,游走的每一步都从与当前节点相连的多个点随机选择一个,不断重复这个过程,直到达到设定的游走长度后,停止游走,从而获得一条条新的用户行为序列数据。

6.基于深度游走和集成学习的用户行为预测系统,其特征在于,该系统包括数据获取模块、预处理模块、用户画像模块、随机游走模块和训练模块;

其中,所述数据获取模块用于获取用户的原始行为数据,构建原始数据集并将其发送给预处理模块;

所述预处理模块用于对原始数据集进行预处理并将预处理之后的数据发送给用户画像模块;

所述用户画像模块用于基于预处理之后的数据集构建用户画像,形成商品社交网络图结构并将其发送给游走模块;

所述游走模块用于对商品社交网络图结构进行随机游走得到新的行为序列数据,然后利用Word2vec模型对新的行为序列数据进行训练生成embedding向量并将其发送给训练模块;

所述训练模块用于将embedding向量输入到机器学习模型中进行训练,得到单一用户行为预测模型。

7.根据权利要求6所述的基于深度游走和集成学习的用户行为预测方法,其特征在于,还包括:融合模块;

所述融合模块用于接收由训练模块输出的多个单一用户行为预测模型,并将其中差异性最大的两个模型进行融合,得到用户行为预测模型。

8.根据权利要求7所述的基于深度游走和集成学习的用户行为预测方法,其特征在于,所述融合模块包括选择单元、计算单元和融合单元;

所述选择单元根据泛化能力从多个单一用户行为预测模型中选择n个模型;其中,n为大于等于3的正整数;

所述计算单元计算n个模型中每个模型与模型之间的最大信息系数MIC且构建混淆矩阵并进行可视化;

所述融合单元在得到的混淆矩阵上面找出相似度最小的两个单一模型进行融合,得到用户行为预测模型。

9.根据权利要求6或7或8所述的基于深度游走和集成学习的用户行为预测方法,其特征在于,所述用户画像模块具体从三个角度进行用户画像的构建,分别是用户的基本信息、用户活跃度信息和用户操作行为统计信息。

10.根据权利要求6或7或8所述的基于深度游走和集成学习的用户行为预测方法,其特征在于,所述随机游走模块被配置为执行如下过程:从网络图结构的任意一个节点出发,游走的每一步都从与当前节点相连的多个点随机选择一个,不断重复这个过程,直到达到设定的游走长度后,停止游走,从而获得一条条新的用户行为序列数据。

基于深度游走和集成学习的用户行为预测方法及系统

技术领域

[0001] 本发明涉及机器识别技术领域，具体涉及基于深度游走和集成学习的用户行为预测方法及系统。

背景技术

[0002] 伴随着互联网技术和电子商务的飞速发展，越来越多的人喜欢从互联网购物，解决日常物品需求问题。每天都会有成千上万的用户从电商网购平台上购买商品，利用人工智能算法对用户历史行为进行分析以此来判断该用户是否购买该商品具有重大的意义。例如，研究者发现，通过分析用户在某电商平台的历史购物数据，可以挖掘出用好的偏好及行为方面的特征，这对个性化推荐、用户关系管理以及广告的投放成本具有很大的作用。鉴于此，利用人工智能算法判断用户历史是否购买商品具有巨大的研究意义。

[0003] 机器学习算法一直是判断用户是否购买或收藏商品的常用方法。通过研究发现，一般从两个角度去建立和优化用户行为预测模型模型，一是从模型算法的角度去优化算法模型的泛化能力；另一种是通过去分析用户的行为序列从而建立算法模型来提高用户行为预测模型模型的泛化能力。然而随着集成学习的兴起，现有技术通过融合单一模型来提高算法的泛化能力。这两种方法都有各自的优点，但目前仍然存在着一些不足。包括：

[0004] (1)在研究用户行为序列的时候没有有效考虑用户每一个行为的上下文语义信息，导致训练得到的模型学习能力和预测准确性均较低；

[0005] (2)在进行集成学习的时候，大多数研究采用随机抽样的方法生成训练子集来构造若干个单一分类器，然而它们的多样性得不到保证，这可能会导致整体分类性能的下降。

发明内容

[0006] 为了至少能够解决上述问题之一，本发明提供了基于深度游走和集成学习的用户行为预测方法。

[0007] 本发明通过下述技术方案实现：

[0008] 基于深度游走和集成学习的用户行为预测方法，该方法包括以下步骤：

[0009] 步骤S1，获取原始数据集并对其进行预处理；

[0010] 步骤S2，基于预处理之后的数据集构建用户画像，形成商品社交网络图结构；

[0011] 步骤S3，对商品社交网络图结构进行随机游走得到新的行为序列数据，然后利用Word2vec模型对新的行为序列数据进行训练生成embedding向量；

[0012] 步骤S4，将embedding向量输入到机器学习模型中进行训练，得到单一用户行为预测模型。

[0013] 本发明通过对用户画像构建形成商品社交网络图结构，基于社交网络图结构，并利用深度游走技术能够提高用户行为预测的可靠性和精度。

[0014] 进一步的，为了进一步提高用户行为预测精度和可靠性，本发明还对单一用户行为预测模型进行集成(融合)，得到预测精度更高的融合模型。本发明的方法还包括步骤S5，

对构造得到的多个单一用户行为预测模型中差异性最大的两个进行融合，得到用户行为预测模型。

[0015] 优选的，本发明的模型融合步骤具体采用了MIC和混淆矩阵的模型差异性度量方法来实现模型融合，能够在提高模型的学习能力的同时得到更加卓越的泛化能力。本发明的步骤S5具体包括：

[0016] 步骤S51，通过调整随机游走的步长和embedding向量的维度，重复执行步骤S3和步骤S4，即可构造得到多个单一用户行为预测模型；

[0017] 步骤S52，根据泛化能力从多个单一用户行为预测模型中选择n个模型；其中，n为大于等于3的正整数；

[0018] 步骤S53，计算n个模型中每个模型与模型之间的最大信息系数MIC且构建混淆矩阵并进行可视化；

[0019] 步骤S54，在得到的混淆矩阵上面找出相似度最小的两个单一模型进行融合，得到用户行为预测模型。

[0020] 优选的，本发明步骤S2具体从三个角度进行用户画像的构建，分别是用户的基本信息、用户活跃度信息和用户操作行为统计信息。

[0021] 优选的，本发明的步骤S3中的随机游走过程具体为：从网络图结构的任意一个节点出发，游走的每一步都从与当前节点相连的多个点随机选择一个，不断重复这个过程，直到达到设定的游走长度后，停止游走，从而获得一条条新的用户行为序列数据。

[0022] 另一方面，本发明还提出了一种基于深度游走和集成学习的用户行为预测系统，本发明的系统包括数据获取模块、预处理模块、用户画像模块、随机游走模块和训练模块；

[0023] 其中，所述数据获取模块用于获取用户的原始行为数据，构建原始数据集并将其发送给预处理模块；

[0024] 所述预处理模块用于对原始数据集进行预处理并将预处理之后的数据发送给用户画像模块；

[0025] 所述用户画像模块用于基于预处理之后的数据集构建用户画像，形成商品社交网络图结构并将其发送给游走模块；

[0026] 所述游走模块用于对商品社交网络图结构进行随机游走得到新的行为序列数据，然后利用Word2vec模型对新的行为序列数据进行训练生成embedding向量并将其发送给训练模块；

[0027] 所述训练模块用于将embedding向量输入到机器学习模型中进行训练，得到单一用户行为预测模型。

[0028] 优选的，本发明的系统还包括：融合模块；

[0029] 所述融合模块用于接收由训练模块输出的多个单一用户行为预测模型，并将其中差异性最大的两个模型进行融合，得到用户行为预测模型。

[0030] 本发明的融合模块包括选择单元、计算单元和融合单元；

[0031] 所述选择单元根据泛化能力从多个单一用户行为预测模型中选择n个模型；其中，n为大于等于3的正整数；

[0032] 所述计算单元计算n个模型中每个模型与模型之间的最大信息系数MIC且构建混淆矩阵并进行可视化；

[0033] 所述融合单元在得到的混淆矩阵上面找出相似度最小的两个单一模型进行融合，得到用户行为预测模型。

[0034] 本发明的用户画像模块具体从三个角度进行用户画像的构建，分别是用户的基本信息、用户活跃度信息和用户操作行为统计信息。

[0035] 本发明的随机游走模块被配置为执行如下过程：从网络图结构的任意一个节点出发，游走的每一步都从与当前节点相连的多个点随机选择一个，不断重复这个过程，直到达到设定的游走长度后，停止游走，从而获得一条条新的用户行为序列数据。

[0036] 本发明具有如下的优点和有益效果：

[0037] 1、本发明对原始数据集中存在的重复、异常和冗余等问题进行了预处理工作，从预处理后的数据集中提取出能够反映消费者行为习惯和偏好程度的统计信息和活跃度信息，以此为用户构建用户画像，然后通过从用户购买商品的社交网络图结构进行随机漫步(Ramdon Walk)得到新的行为序列，然后在用Word2vec模型得到用户每个行为的上下信息加入到机器学习模型中去训练学习，提高了模型的预测可靠性和预测精度。

[0038] 2、本发明采用MIC和混淆矩阵法进一步对获得的单一模型进行选择性集成(融合)，进一步增强了模型的预测性能和可靠性。

附图说明

[0039] 此处所说明的附图用来提供对本发明实施例的进一步理解，构成本申请的一部分，并不构成对本发明实施例的限定。在附图中：

[0040] 图1为本发明第一实施方式的用户行为预测模型构建流程示意图。

[0041] 图2为本发明的随机搜索流程示意图。

[0042] 图3为本发明第二实施方式的用户行为预测模型构建流程示意图。

[0043] 图4为基于MIC和混淆矩阵的选择性模型融合流程图。

[0044] 图5为ROC曲线图。

[0045] 图6为本发明测试和验证时模型建立过程示意图。

[0046] 图7为本发明的用户画像和原始模型验证集AUC对比图。

[0047] 图8为本发明的用户画像和原始模型测试集AUC对比图。

[0048] 图9为本发明的混淆矩阵可视化。

[0049] 图10为本发明的模型融合验证集AUC对比图。

[0050] 图11为本发明的模型融合测试集AUC对比图。

[0051] 图12为本发明的AUC排序融合验证集AUC对比图。

[0052] 图13为本发明的AUC排序融合测试集AUC对比图。

具体实施方式

[0053] 为使本发明的目的、技术方案和优点更加清楚明白，下面结合实施例和附图，对本发明作进一步的详细说明，本发明的示意性实施方式及其说明仅用于解释本发明，并不作为对本发明的限定。

[0054] 实施例1

[0055] 本实施例提出了一种基于深度游走和集成学习的用户行为预测方法。

[0056] 本实施例把用户购买的行为序列的一个商品看作一个词,所有的商品看作一篇文档,这样就可以利用一些自然语言处理技术(NLP)去训练词向量。另一方面在用户购买行为序列场景下,数据与数据之间存在的大量的图结构信息,这些数据信息非常的重要,本实施例将深度游走(DeepWalk)技术很好的运用于购买行为网络结构中。深度游走(DeepWalk)技术利用随机游走(Random Walk)技术对图中的网络节点就行随机游走形成一个行为序列,在把用户的行为序列看成一个词,所有的行为序列文档,使用Word2vec算法模型去预训练词向量,并在原模型的基础上,加入DeepWalk技术,提出了基于深度游走的分类器算法。

[0057] 如图1所示,本实施例的方法主要包括以下几部分:

[0058] 1、获取原始数据集并进行预处理;

[0059] 2、基于预处理之后的数据集进行用户图像构建,形成用户行为序列的相关商品的图结构;

[0060] 3、在图结构中采用随机游走的方式随机选择起始点,重新产生商品的行为序列。具体为:

[0061] 从图结构中的某一个顶点访问其他剩余节点的过程叫做图的遍历,图的遍历方法通常有两种,一是广度优先搜索(BFS),而是深度优先搜索(DFS),图的遍历方法是求解图拓扑结构相关问题的前提。广度优先搜索(BFS)从起始点开始遍历其邻接的节点,由此向外不断扩散,优先考虑近端连接所带来的信息量。深度优先搜索(DFS)从一个顶点v出发,首先将v标记为已遍历的顶点,然后选择一个邻接于v的尚未遍历的顶点u,如果u不存在,本次搜素终止,如果u存在,那么从u又开始一次DFS,如此循环直到不存在这样的顶点,深度优先搜索(DFS)会利用到远端连接所隐含的信息量。RandomWalk是一种深度优先遍历算法,此算法可重复访问已访问节点。随机游走(Random Walk)就是在网络图结构中不断随机重复地随机游走路劲,从图结构中某个特定的顶点出发,游走的每一步都从与当前节点相连的几点随机选择一个,不断的重复这个过程,达到设定的游走长度后,停止游走,从而获得一条条序列数据。

[0062] 4、将新的行为序列输入到Word2vec模型中,训练生成该商品的embedding向量。具体为:

[0063] Word2vec算法就是通过学习文本然后通过词向量的方式来表征词的语义信息,即把原来词所在空间映射到一个新的空间中去,使得语义上相似的词在该空间内距离相近。Word2vec总共包含两个语言算法模型,CBOW模型和Skip-gram模型。CBOW模型和Skip-gram模型均包含输入层、隐含层和输出层,CBOW模型是以当前词w_t的上下文w_{t-1},w_{t-2},w_{t+1},w_{t+2}的前提下预测当前词w_t,而Skip-gram模型恰恰相反,是在已知当前词w_t的前提下,预测其上下文w_{t-1},w_{t-2},w_{t+1},w_{t+2}。Word2vec算法共提出了Hierarchical Softmax和Negative Sampling两种优化算法来减少词向量的训练时间。Hierarchical Softmax和Negative Sampling两种优化方法都使用了BP神经网络作为分类方法。每个单词最后都是由算法随机生成的一N维向量来表示,经过Woed2vec算法模型训练之后可以得到每个单词的最优词向量即embedding向量。

[0064] 学习某一个词向量的算法模型是在给定上下文单词的情况下去预测下一个词。该算法模型框架中,文档中的每个词语被投射到一个向量空间中,其中文档中的每一个词语对应着矩阵W里面唯一一个列向量,列向量的位置是以该词语在由该词语在文档中的位置

决定。然后上下文单词向量的级联或者相加作为特征向量来预测下一个词语。

[0065] 假设有一个句子W,该句子含有T个词,分别是$W_1,W_2,\cdots,W_i,\cdots,W_T$,我们的目标是最大化函数L,即

[0066] $$L=\prod_{t=1}^{V} p(w_t \mid w_{t-k}, w_{t-(k-1)}, \cdots, w_{t-1}, w_{t+1}, \cdots, w_{t+(k-1)}, w_{t+k})$$

[0067] 取log并平均化得

[0068] $$\frac{1}{T}\sum_{t=k}^{T-k} \log p(w_t \mid w_{t-k}, w_{t-(k-1)}, \cdots, w_{t-1}, w_{t+1}, \cdots, w_{t+(k-1)}, w_{t+k})$$

[0069] 预测任务中主要运用到了softmax多分类的思想,在上式中,后验概率为

[0070] $$p(w_t \mid w_{t-k}, w_{t-(k-1)}, \cdots, w_{t-1}, w_{t+1}, \cdots, w_{t+(k-1)}, w_{t+k})=\frac{e^{y_{w_i}}}{\sum e^{y_i}}$$

[0071] 上式中每个yi都是没有经过归一化处理概率的log值,计算方式如下式所示:

[0072] $y=b+Uh(w_{t-k},\cdots,w_{t+k};W)$

[0073] 通过Word2vec算法模型训练得到的词向量包含了上下文信息。Word2vec算法模型优点在于不仅能够获得上下文信息,相较于one-hotting还压缩了数据维度规模,大大降低了模型的时间效率。

[0074] 5、将embedding向量输入到机器学习模型中进行训练学习,从而得到单一的用户行为预测模型。

[0075] 本实施例中,经过数据预处理和用户画像的构建之后形成的商品图结构上进行Ramdon Walk对每个节点进行随机游走的得到新的行为序列,再把新的行为序列利用Word2vec算法模型进行训练和优化从而得到embedding向量,最后再利用机器学习模型进行训练和优化,得到最终的输出结果。在训练模型的时候,有两个超参数需要控制,分别是embedding向量的维度和随机游走的步长。embedding向量的维度不是越多越好,维度少的有可能影响模型的泛化能力,维度高有可能导致维度灾难,为此我们要进行网格搜索确定来embedding向量的最佳维度。网格搜索就是先指定词向量的维度,用于模型的训练,在这此基础上加大词向量的维度用于模型的训练和预测,如果该模型比上一步模型的效果要好,则在进一步扩大embedding向量的维度,反之,则达到了embedding向量的最佳维度,具体如图2所示。同理,随机游走的步长的的长度最优解也可以通过随机搜索可以得到。

[0076] 实施例2

[0077] 本实施例在上述实施例1的基础上,进一步对单一模型进行融合,如图3所示。本实施例的融合方法首先使用最大信息系数(MIC)分别度量每个单一学习器之间的差异性,然后在用混淆矩阵的形式表达出来,从而选择差异性最大的两个单一学习器进行模型融合以得到更加卓越的泛化能力。

[0078] 其中:

[0079] 1、最大信息系数(MIC)用来衡量两个变量之间的关联程度,是线性关系还是非线性关系。最大信息系数(MIC)的计算主要利用互信息(MI)和网格划分方法。互信息(MI)是用

来度量两个变量之间的关联程度，给定变量集合B＝{b_1,b_2,…,b_n}，n为样本的个数，互信息(MI)可以定义为

[0080] $$MI(A,B)=\sum_{a\in A}\sum_{b\in B}p(a,b)\log\frac{p(a,b)}{p(a)p(b)}$$

[0081] 其中p(a,b)是变量A和变量B之间的联合概率密度，p(a)和p(b)分别变量A和变量B的边缘概率密度，一般情况下联合概率计算相对来说比较复杂。最大信息系数(MIC)的思想是针对两个变量之间的关系，将其离散在二维空间中，并且使用散点图来表示，将当前二维空间在x,y方向分别划分为一定的区间数，然后查看当前的散点在各个方格中落入的情况，这就是联合概率的计算，这样就解决了在互信息中的联合概率难求的问题。假设有限的有序对的集D＝{(a1,b1),(a2,b2),…,(an,bn)}，定义划分G分别将变量A和变量B的值域分成x段和y段，G即为x×y坐标轴构成的网格。在得到每个网格划分片段分别计算内部计算互信息MI(A,B)，一种x×y的网格划分方式有许多方式，取不同划分法中的最大MI(A,B)值作为划分G的互信息值，于是定义划分G下D的最大互信息公式为：

[0082] MI*(D,x,y)＝maxMI(D|G)

[0083] 其中，D|G表示数据集合D在G上进行划分。将不同划分下得到的最大互信息进行归一化得到特征矩M(D)x,y，其计算公式为

[0084] $$M(D)_{x,y}=\frac{MI^{*}(D,x,y)}{\log\min\{x,\ y\}}$$

[0085] 则最大信息系数(MIC)定义为

[0086] $$MIC(D)=\max_{xy<B(n)}\{M(D)_{x,y}\}$$

[0087] 其中B(n)为变量，表示网格划分x×y的最大值，在文献中给出当B(n)＝$n^{0.6}$时效果最佳。

[0088] 2、混淆矩阵表达的是数据样本模型预测结果和真实属性值之间的关系，是评估分类器泛化能力一种采常用的方式。假设现在总共具有N个类别的分类任务，数据样本集合D总共包含T_0条数据记录，每个类别包含T_i条记录(1≤i≤N)。利用机器学习或深度学习模型构造一种分类器C，cmi_j表示第i个类别的数据记录被分类器C判断成第j类类别的数据记录第i类数据记录总数的百分率，于是可以得到混淆矩阵CM(C,D)，其维度为N×N：

[0089] $$CM(C,D)=\begin{bmatrix}cm_{11} & cm_{12} & \cdots & cm_{1i} & \cdots & cm_{1N}\\ cm_{21} & cm_{22} & \cdots & cm_{2i} & \cdots & cm_{2N}\\ \vdots & \vdots & & \vdots & & \vdots\\ cm_{i1} & cm_{i2} & \cdots & cm_{ii} & \cdots & cm_{iN}\\ \vdots & \vdots & & \vdots & & \vdots\\ cm_{N1} & cm_{N2} & \cdots & cm_{Ni} & \cdots & cm_{NN}\end{bmatrix}$$

[0090] 混淆矩阵中元素的列下标表示分类器C模型对数据样本的预测结果，行下标表示数据样本的真实标签值。混淆矩阵的对角线元素表示每种类别能被分类器C正确预测的概率，而非对角线部分表示分类器C判断错误的概率。

[0091] 在机器学习领域中，如果两种类别相似度比较大，它们的数据样本就有很大可能被分类器预测为对方类别。混淆矩阵行向量CM_i(1≤i≤N)表示类别i的数据样本在进行模型预测的时候对各个类别的倾向性。基于混淆矩阵，本实施例定义了一个单一分类器之间的相关性矩阵。假设一共有M个单一分类器，本实施例将每一个单一分类器对应的混淆矩阵转化成一个行向量，详细的做法是把混淆矩阵依次按行进行展开，如下所示：

[0092]
$$CM^{(i)} = (CM_1^{(i)}, CM_2^{(i)}, CM_3^{(i)}, \cdots, CM_N^{(i)})$$
$$= (cm_{11}^{(i)}, cm_{11}^{(i)}, \cdots, cm_{1N}^{(i)}, cm_{21}^{(i)}, cm_{22}^{(i)}, \cdots, cm_{2N}^{(i)}, \cdots, cm_{N1}^{(i)}, cm_{N2}^{(i)}, \cdots, cm_{NN}^{(i)})$$

[0093] 其中$CM^{(i)}$(1≤i≤M)。然后把所有的行向量$CM^{(i)}$(1≤i≤M)合并组成一个矩阵得到所有单一分类器的混淆矩阵，定义为CMS(C,D)，如下所示：

[0094]
$$CMS(C,D) = \begin{bmatrix} CM^{(1)} \\ CM^{(2)} \\ \vdots \\ CM^{(M)} \end{bmatrix}$$
$$= \begin{bmatrix} cm_{11}^{(1)} & \cdots & cm_{1w}^{(1)} & cm_{21}^{(1)} & & cm_{2w}^{(1)} & \cdots & cm_{w1}^{(1)} & \cdots & cm_{ww}^{(1)} \\ cm_{11}^{(2)} & \cdots & cm_{1w}^{(2)} & cm_{21}^{(2)} & \cdots & cm_{2w}^{(2)} & \cdots & cm_{w1}^{(2)} & \cdots & cm_{ww}^{(2)} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ cm_{11}^{(M)} & \cdots & cm_{1w}^{(M)} & cm_{21}^{(M)} & \cdots & cm_{2w}^{(M)} & \cdots & cm_{w1}^{(M)} & \cdots & cm_{ww}^{(2)} \end{bmatrix}$$

[0095] 基于上述得到的混淆矩阵和最大信息系数(MIC)可以得到集成学习中相似性度量矩阵Q。Q矩阵反应了每个单一分类器之间的相关性，Qij取值越小，说明两个单一分类器的相关性越小，反之越大。因此Q矩阵能很好的度量每个分类器之间的相似性，使用Q矩阵能够为集成学习中如何找到差异性很大的两个单一分类器提供一种方法。Q矩阵的计算公式如下式所示。

[0096]
$$Q_{ij} = MIC(D) = \max_{ij<B(n)} \{M(D)_{i,j}\}$$

[0097] 如图3所示，本实施例的方法在上述实施例1的基础上，还包括以下步骤：

[0098] 在上述实施例1中所构造的模型中选择n个泛化能力比较强的用户行为预测模型；第二步是计算每个模型与模型之间的最大信息系数MIC；第三步是构建混淆矩阵并进行可视化；第四步是在得的的混淆矩阵上面找出颜色比较浅即相似度比较小的两个单一模型就行Bagging融合。具体的流程图如图4所示。

[0099] 在进行模型融合的时候，因为进行的是简单的加权融合，如何确定每个模型的最佳系数通常是在进行加权融合时候的一个难点。进行加权融合时，最简单的是进行平均加权融合，因为从理论上来讲，如果模型融合能够带来模型效果提升的话，把每个单一模型平均化在进行融合一定可以带来模型效果方面的提升，但是将每个模型平均化有可能不是每个模型进行模型融合时候的最佳系数。本文不是简单的进行简单的平均加权平均融合，而是根据AUC排序的特性，提出了一种基于AUC排序的融合方法。AUC的计算公式如下

[0100] $$AUC=\frac{\sum_{insi\in postiveclass} rank_{insi}-\frac{M\times(M+1)}{2}}{M\times N}$$

[0101] 其中，M是正样本的个数，N是负样本的个数，$rank_{insi}$指的是正样本在数据集中的排列顺序。从上式可以看出，对于AUC其本质就是一个排序，在进行模型融合的时候我们可以利用此特性算出每个模型的系数，最后在进行模型融合。每个模型系数计算公式如下。具体的方式是将每个模型得到的结果按照AUC值进行降序排序，取每个样本进行排序后的倒数乘以模型预测得到的AUC相加得到最后的融合结果。

[0102] 模型融合最后得到的值为各个模型每个样本的AUC值乘以每个样本AUC值在单一模型相对排序的倒数的累加和。

[0103] 实施例3

[0104] 本实施例以某一银行购物APP后台日志为例对上述实施例提出的方法进行测试。

[0105] 1、原始数据集

[0106] 获取的银行购物APP后台日志的时间跨度为一个月，主要包括了4万多条用户消费行为数据，每一行对应着用户的一条操作记录，按照用户操作时间进行排序。该数据集包含的相关字段如表1所示。

[0107] 表1原始数据集基本信息表

[0108]

列名	类型	说明	示例
User-id	int	用户唯一标识	abc123
Product-id	int	产品 id	abc123
Seller-id	int	卖家 id	abc123
day	String	日期	1,2,3....30
type	int	用户操作相关变量	0
ProductInfo_X	String	产品相关变量	0
Webinfo_X	String	网络行为相关变量	1
Userinfo_X	String	用户相关变量	100
purcharse	int	购买	1:购买 0：未购买

[0109] 对原始数据集进行预处理。

[0110] 2、用户画像的构建

[0111] 未经过处理的数据集中的每一行记录是以用户的操作行为作为粒度的，是用户单个操作行为的信息记录，而本实施例是预测用户是否购买某一产品，需要的每一个用户更加细粒度的信息。所以对原始数据集进行需处理以及用户画像的构建工作非常关键，以此获得每一个用户的操作行为更加细粒度的特征。本实施例主要通过计算用户行为统计量来构建用户画像从而发现用户的行为习惯。

[0112] 根据用户User-id字段对未处理的数据集进行分组和排序，可以得到每个用户行为的统计信息。用户行为统计信息从多个方面描述了用户的行为习惯，主要包含以下几个

方面：

[0113] (1)用户基本信息

[0114] 这部分数据信息由原始数据中用户User-id字段和用户相关变量Userinfo_X构成。这部分数据字段不需要就行额外的重组和计算，只需要读到读到用户的每一条记录时，对相应的数值就行更新。

[0115] (2)用户活跃度信息

[0116] 用户的活动信息反映了用户对app的偏好，可以从多个方向来思考。本文主要使用的用户活动信息指标如表2所示。

[0117] 表2用户活跃度信息表

名称	描述
活跃天数	用户从第一次操作开始，一直到用户最后一次操作结束，在这段时间内用户操作天数的总数
用户是否当天有多次启动	用户当天是否有多次操作行为
用户是否有多次连续启动	用户当天是否有连续操作行为
用户连续启动次数的最大值、最小值、方差、总数	用户在30天内多次连续操作行为的最大值、最小值、方差、总数
用户的平均行为次数	用户在这30天内每天的操作行为平均次数
区间内最后一次活跃距离区间末端的天数	用户最后一次操作时间距离末端的时间间隔

[0118]

[0119] (2)用户操作行为统计信息

[0120] 用户操作行为信息反应了用户如何与APP交互的以及用户对APP某种功能的偏好程度。这部分数据字段主要包括用户每一种操作行为的次数、占比等，详细的特征字段如如表3所示。

[0121] 表3用户行为操作统计信息表

名称	描述
用户操作行为的统计量	用户每种操作行为的统计量
用户每种操作行为占比	用户每种操作行为占操作行为总次数的比例
用户最后一天的操作行为次数	用户最后一天每个操作行为次数
用户最后两天操作行为次数	用户最后两天每个操作行为次数
用户最后四天操作行为次数	用户最后四天天每个操作行为次数

[0122]

[0123] 3、评价指标

[0124] 模型的预测结果用混淆矩阵来表示，如表4所示。

[0125] 表4混淆矩阵

[0126]

预测类别 / 真实类别	预测为正（购买）	预测为负（为购买）
真实为正（购买）	TP(真正例)	FN（假负例）
真实为负（未购买）	FP（假正例）	TN（真负例）

[0127] 准确率(ACC)、精确率(Precision)、召回率(Recall)、假阳率(FRP)和F1-sorce可以有上面的混淆矩阵定义得到，计算公式分别如下所示：

[0128] ACC＝(TP+TN)/(TP+FN+FP+TN)

[0129] P＝TP/(TP+FP)

[0130] R＝TP/(TP+FN)

[0131] FRP＝FP/(TN+FP)

[0132] F1-sorce＝(2×P×R)/(P+R)

[0133] ROC曲线是以召回率(Recall)为纵坐标，假阳率(FRP)为横坐标而得的曲线，而ROC曲线下面的面积就是AUC值，如图5所示，显然这个面积的数值不会大于1。又由于ROC曲线一般都处于y＝x这条直线的上方，所以AUC的取值范围一般在0.5和1之间。原始数据集中用户购买的记录有17410条，未购买的记录有15590条，数据样本不平衡，这时候使用准确率作为模型评估指标并不合适，而对于此类用户行为预测问题，一般采取AUC作为评价方式，所以本实施例所选取的评价指标为AUC。AUC越大，模型效果越好。

[0134] 为了验证用户画像和基于深度游走用户行为预测模型的有效性，本实施例选取了xgboost以下简称xgb和lightgbm以下简称lgb两个基础机器学习模型为基础分类器，分别

与加上用户画像和Deepwalk技术的xgb和lgb扩展模型做对比实验。在Deepwalk技术里面随机游走步长walklength在[5,10)范围内以步长为1进行随机搜索，词向量维度size在[5,10)范围内以步长为1进行随机搜索。本实施例中数据预处理和用户画像建立的工作有pandas、numpy和sklearn实现。xgb和lgb分别使用python xgboost和lightgbm包实现。DeepWalk模型中Word2vec模型的实现由python gensim包实现。模型验证使用的五折交叉验证，由sklearn实现。xgb和lgb模型参数有网格搜索来确定最优参数值。本实施例所建立的模型结构图如图6所示。其中size表示训练Word2Vec embedded向量的维度，walklength表示随机游走的步长。

[0135] 原始数据集中的每一行记录是以某一用户的操作行为为粒度的，是单个操作的信息，而本实施例是预测用户是否购买某一产品，需要的单一用户粒度的信息。所以，有必要对原始数据集进行重新整合和计算建立用户画像，以获得每一个用户的操作行为更加细粒度的特征。本实施例主要是通过计算用户行为统计量来构建用户画像从而发现用户的行为习惯。

[0136] 如图7所示，是lgb、xgb模型和加上用户画像后验证集的AUC性能指标图。相比而言，xgb基础模型即模型2的AUC的值要比lgb基础模型即模型1的AUC值高0.0027，达到了0.5219，lgb和xgb模型加入用户画像之后形成的模型即模型3、模型4，模型3比模型4的AUC值高，达到了0.7131，而模型3比模型1的AUC值高了0.1939，模型4比模型2而的AUC值高了0.1905，实验证明了建立用户画像之后不管是xgb模型还是和lgb模型性能方面取得了较大的突破。

[0137] 图8是加上用户画像之后的模型和原始模型在测试集上即新的数据集上的实验效果。测试集中不包含训练集的数据，使得模型的预测会有很高的不可预见性，所以会使得模型结果的预测有适当的波动起伏。但是可以看出，加上用户画像之后的lgb和xgb模型的AUC值都比其他模型高，其中模型3的AUC值达到了0.7292，模型4的AUC值达到了0.7238。相比较验证集，lgb和xgb基础模型的AUC值低，但是加入用户画像之后，测试集的AUC值要比验证集高，这是由新数据上的不确定性导致的，模型3和模型4存在着欠拟合的风险。总体而言，无论是测试集还是验证集，加上用户画像之后的模型都要比基础模型的学习能力强。

[0138] 其中类别1对应着模型5,6,7,8,9，类别2对应着模型10,11,12,13,14，类别3对应着模型15,16,17,18,19，类别4对应着模型20,21,22,23,24，类别5对应着模型25,26,27,28,29，类别6对应着模型30,31,32,33,34，类别7对应着模35,36,37,38,39，类别8对应着40,41,42,43,44，类别9对应着45,46,47,48,49，类别10对应着模型50,51,52,53,54。实验表明，在加入用户画像模型3和模型4后加入DeepWalk技术之后得到的众多模型中，不管Word2vec embedding向量的维度和随机游走的步长的的取值是多少，模型的泛化能力都有所提升，其中，就模型3利用DeepWalk而言，当size＝7，walklength＝9时所的到的用户行为预测模型即模型29，AUC达到了0.7431，要比模型3的auc高了0.03，当size＝7，walklength＝8时所得到的模型28，在由模型3演变得到的众多模型中AUC值最低，尽管如此，模型28的AUC值比模型的AUC值高了0.026。由模型4演变得到的众多模型的AUC值都比有模型3演变得到的模型的AUC值低，当size＝9，walklength＝6时所得到的模型51，AUC值最高，达到了0.7374，AUC值比模型4高了0.025，效果最低的模型为模型53，AUC为0.7342，比模型4的AUC高了0.021。总的来说，加入DeepWalk技术的用户行为预测模型比其他基础模型的性能更

优。各个模型在验证集上的具体实验数据如表5所示。

[0139] 表5验证集各模型AUC值

size \ walklength	5	6	7	8	9
5	0.7412	0.7415	0.7408	0.7415	0.7419
	（模型 5）	（模型 6）	（模型 7）	（模型 8）	（模型 9）
	0.7363	0.7360	0.7361	0.7350	0.7351
	（模型 10）	（模型 11）	（模型 12）	（模型 13）	（模型 14）
6	0.7411	0.7423	0.7410	0.7413	0.7421
	（模型 15）	（模型 16）	（模型 17）	（模型 18）	（模型 19）
	0.7362	0.7360	0.7351	0.7349	0.7358
	（模型 20）	（模型 21）	（模型 22）	（模型 23）	（模型 24）
7	0.7420	0.7406	0.7417	0.7393	0.7431
	（模型 25）	（模型 26）	（模型 27）	（模型 28）	（模型 29）
	0.7369	0.7359	0.7355	0.7338	0.7350
	（模型 30）	（模型 31）	（模型 32）	（模型 33）	（模型 34）
8	0.7410	0.7402	0.7416	0.7420	0.7413
	（模型 35）	（模型 36）	（模型 37）	（模型 38）	（模型 39）
	0.7366	0.7365	0.7357	0.7358	0.7363
	（模型 40）	（模型 41）	（模型 42）	（模型 43）	（模型 44）
9	0.7416	0.7411	0.7427	0.7425	0.7413
	（模型 45）	（模型 46）	（模型 47）	（模型 48）	（模型 49）
	0.7366	0.7374	0.7360	0.7343	0.7364
	（模型 50）	（模型 51）	（模型 52）	（模型 53）	（模型 54）

[0140]

[0141] 在验证集上，有lgb和xgb加入DeepWalk急速扩展得到的模型的学习能力差不多，AUC值都达到了0.74。由lgb扩展的模型中，当size＝5，walklength＝8时所的到的用户行为预测模型性能最佳即模型8，AUC值达到了0.7479，比模型3在验证集上的AUC值高了0.0187，当size＝6，walklength＝7时所的到的用户行为预测模型的AUC值最低即模型17，AUC值达到了0.7451，只比模型8低了0.0028。有xgb扩展的模型中，当size＝7，walklength＝6时所的到的用户行为预测模型性能最佳即模型31，AUC值达到了0.7466，比模型4在验证集上的AUC值高了0.0228，当size＝6，walklength＝7时所的到的用户行为预测模型的AUC值最低即模型23，AUC值达到了0.7437，只比模型31低了0.0029。各个模型在验证集上的具体实验数据如表6所示。综上所述，无论是在验证集还是在测试集上，基于深度游走的用户行为预

测模型要比其他模型性能更优。

[0142] 表6测试集各模型AUC值

size \ walklength	5	6	7	8	9
5	0.7468（模型5）	0.7457（模型6）	0.7466（模型7）	0.7479（模型8）	0.7472（模型9）
	0.7450（模型10）	0.7447（模型11）	0.7449（模型12）	0.7456（模型13）	0.7455（模型14）
6	0.7465（模型15）	0.7478（模型16）	0.7451（模型17）	0.7446（模型18）	0.7471（模型19）
	0.7441（模型20）	0.7466（模型21）	0.7444（模型22）	0.7437（模型23）	0.7455（模型24）
7	0.7460（模型25）	0.7462（模型26）	0.7465（模型27）	0.7458（模型28）	0.7461（模型29）
	0.7447（模型30）	0.7445（模型31）	0.7456（模型32）	0.7442（模型33）	0.7446（模型34）
8	0.7470（模型35）	0.7469（模型36）	0.7468（模型37）	0.7462（模型38）	0.7453（模型39）
	0.7448（模型40）	0.7438（模型41）	0.7444（模型42）	0.7444（模型43）	0.7443（模型44）
9	0.7470（模型45）	0.7455（模型46）	0.7468（模型47）	0.7463（模型48）	0.7465（模型49）
	0.7451（模型50）	0.7449（模型51）	0.7441（模型52）	0.7443（模型53）	0.7440（模型54）

[0143]

[0144] 本实施例使用到的最大互信息和混淆矩阵的构建的计算由pandas和numpy共同实现，混淆矩阵的可视化由matplotlib实现。进行模型融合时，第一步所要求的是单一学习器的学习能力比较强。本实施例选取了上述模型中泛化能力较强的6个单一模型，分别是模型29、模型47、模型16、模型51、模型30、模型54、模型38和模型20进行集成学习。对这6个模型与模型之间进行最大互信息(MIC)计算并构建混淆矩阵可视化如图9所示。每个模型之间最大信息(MIC)值如表7所示。

[0145] 表7模型与模型之间的MIC值

[0146]

	模型 16	模型 20	模型 29	模型 30	模型 47	模型 51
模型 16	1	0.86	0.64	0.72	0.86	0.76
模型 20	0.86	1	0.58	0.46	0.44	0.88
模型 29	0.64	0.58	1	0.42	0.59	0.72
模型 30	0.72	0.46	0.42	1	0.84	0.68
模型 47	0.86	0.44	0.59	0.84	1	0.82
模型 51	0.76	0.88	0.72	0.68	0.82	1

[0147] 图9所示，颜色越浅表示两个模型之间的差异性越大。本实施例可以选择模型20与模型30进行融合、模型30与模型29进行融合、模型47与模型20这三对颜色较浅且最大互信息MIC低于0.5的进行模型融合，其中模型30与模型29两个模型在混淆矩阵中颜色最浅，最大信息系数MIC最低。可以将颜色较深且最大信息MIC高于于0.5将模型16和模型51融合、模型16与模型20这两对做模型融合形成对比实验。模型融合的方式为Bagging融合，每个单一模型的权重设置为0.5。

[0148] 如图10所示是模型融合之后验证集的效果对比图。从表上可以看出，当把混淆矩阵颜色较浅即差异性比较大的两个单一模型融合时，集成学习才能有效果。把模型30和模型29进行融合时，效果最佳，AUC值达到了0.7561，因为模型30和模型29两个单一模型之间的差异性最大，AUC值比模型30和模型29分别提高了0.0192、0.013，相同的把模型20和模型30进行模型融合和将模型47和模型20进行融合，其AUC值都比单一模型要高，模型20和模型30融合后的AUC值比型20和模型30分别提高了0.008、0.0073，模型47和模型20进行融合后其AUC值比模型47和模型20分别提高了0.0072、0.00139，相反的把两个差异性比较小的模型融合的效果并不好，模型16和模型51或者模型16和模型20相融合后的AUC值都比其单一模型的表达能力更弱。实验证明了在验证集上，模型融合只有找到相似性比较小的单一学习器进行融合才能获得更好的学习能力，差异性最大的两个模型进行融合时所带来的增益最大，而将两个相似的模型进行融合的时候不仅不能带来模型表达能力的提升，反而使得模型的表达能力变弱。具体实验结果如表8所示。

[0149] 表8模型融合验证集AUC对比表

[0150]

模型	AUC 值	模型	AUC 值	融合后 AUC 值
模型 20	0.7362	模型 30	0.7369	0.7442
模型 30	0.7369	模型 29	0.7431	0.7561
模型 47	0.7427	模型 20	0.7362	0.7501
模型 16	0.7423	模型 51	0.7366	0.7303
模型 16	0.7423	模型 20	0.7362	0.7226

[0151] 如图11所示是模型融合后模型在测试集上即新的数据集上的表现结果。从该图上可以看出，模型融合依然可以适用于新的数据集。和训练集保持一致，当把模型30和模型29

融合之后，AUC值最高，达到了0.7612，比单一模型模型30和模型29的AUC值分别高出了0.0265、0.0151，把模型20和模型30融合后AUC值达到了0.7586，比单一模型模型20和模型30的AUC值分别提高了0.0145、0.0139，而把模型47和模型20融合后AUC值达到了0.7566，比单一模型47和模型20的AUC值分别高出了0.0098、0.0209。当把模型16和模型51或者模型16和模型20相融合后的AUC值都比其单一模型学习能力弱。验证集和测试集上的实验结果表明，基于MIC和混淆矩阵的选择性模型融合方法能够找出差异项比较大的单一学习器进行融合，获得了比较卓越的泛化能力，而把差异性比较小的两个单一模型进行融合时，模型的表达能力不增反降。具体实验结果如表9所示。

[0152] 表9模型融合验证集AUC对比表

[0153]

模型	AUC 值	模型	AUC 值	融合后 AUC 值
模型 20	0.7441	模型 30	0.7447	0.7586
模型 30	0.7347	模型 29	0.7461	0.7612
模型 47	0.7468	模型 20	0.7362	0.7566
模型 16	0.7478	模型 51	0.7449	0.7400
模型 16	0.7478	模型 20	0.7441	0.7211

[0154] 如图12所示是在验证集基于AUC排序的融合方法和简单融合方法的对比图。从图上可以看出，基于AUC排序的融合方法要比普通的平均加权融合方法AUC值高，模型的表达能力更强。其中基于AUC排序后，模型20与模型30融合后的AUC值比其他普通的加权融合提升最高，AUC达到了0.7611，提高了0.0025，而模型30和模型29融合后AUC是三个模型融合后AUC中最高的，达到了0.7622，要比普通融合方法的AUC值提高了0.001，模型47和模型20融合后AUC值提高了0.002，模型的表达能力比前两个模型融合后的模型表达能力差。通过实验证明基于AUC排序的融合方法要比普通的加权融合方法在验证集上学习能力更强。具体实验结果如表10所示。

[0155] 表10 AUC排序融合验证集对比表

[0156]

模型	AUC 值	模型	AUC 值	融合后 AUC 值	AUC 排序融合后的 AUC 值
模型 20	0.7362	模型 30	0.7369	0.7586	0.7611
模型 30	0.7369	模型 29	0.7431	0.7612	0.7622
模型 47	0.7427	模型 20	0.7362	0.7566	0.7586

[0157] 如图13所示是在测试集上即新的数据集上基于AUC排序的融合方法和简单融合方法的对比图。从图上可以看出，基于AUC排序的融合方法要比普通的平均加权融合方法在验证集上AUC值高，模型的表达能力更强。其中基于AUC排序后，模型20与模型30融合后的AUC值比其他普通的加权融合方法AUC值提升最高，AUC达到了0.7662，提高了0.0076，而模型30和模型29融合后AUC是三个模型融合后AUC值提高了0.0031，达到了0.7622，要比普通融合

方法的AUC值提高了0.001，模型47和模型20融合后AUC值提高了0.006，模型的表达能力比前两个模型融合后的模型表达能力差。通过实验证明基于AUC排序的融合方法要比普通的加权融合方法不管是在验证集还是在测试集上，模型的表达能力更强。具体实验结果如表11所示。

[0158] 表11 AUC排序融合测试集对比表

模型	AUC 值	模型	AUC 值	融合后 AUC 值	AUC 排序融合后的 AUC 值
模型 20	0.7441	模型 30	0.7447	0.7586	0.7662
模型 30	0.7347	模型 29	0.7461	0.7612	0.7643
模型 47	0.7468	模型 20	0.7362	0.7566	0.7626

[0159]

[0160] 通过上述测试结果和分析，本实施例采用相似性比较小的单一学习器进行模型融合要比单一学习器性能更优，而差异性最大的两个单一学习器进行集成学习性能最佳。而基于AUC排序的模型融合方法所得到的用户行为预测模型要比简单的加权融合方法性能更优。

[0161] 本领域内的技术人员应明白，本申请的实施例可提供为方法、系统、或计算机程序产品。因此，本申请可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0162] 本申请是参照根据本申请实施例的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0163] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0164] 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0165] 以上所述的具体实施方式，对本发明的目的、技术方案和有益效果进行了进一步详细说明，所应理解的是，以上所述仅为本发明的具体实施方式而已，并不用于限定本发明的保护范围，凡在本发明的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

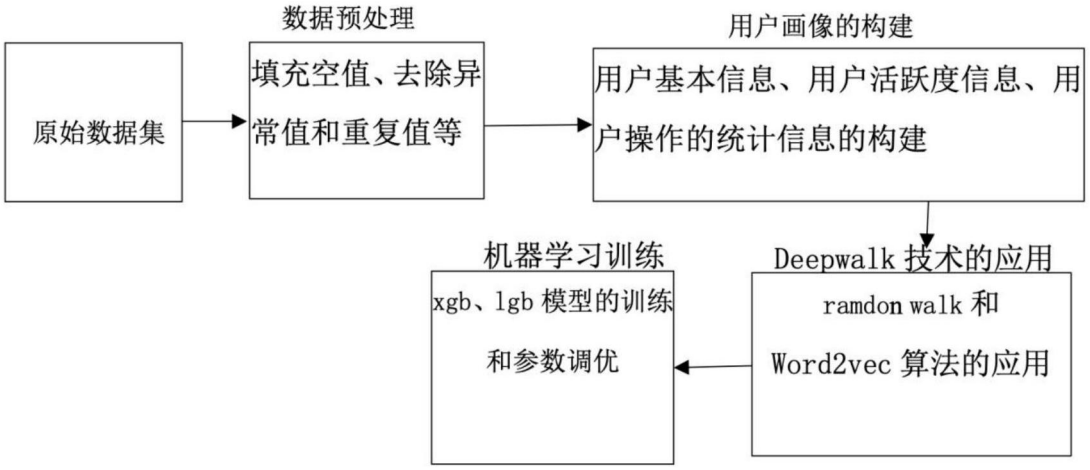

数据预处理
原始数据集
填充空值、去除异常值和重复值等
用户画像的构建
用户基本信息、用户活跃度信息、用户操作的统计信息的构建
机器学习训练
xgb、lgb 模型的训练和参数调优
Deepwalk 技术的应用
ramdon walk 和 Word2vec 算法的应用

图1

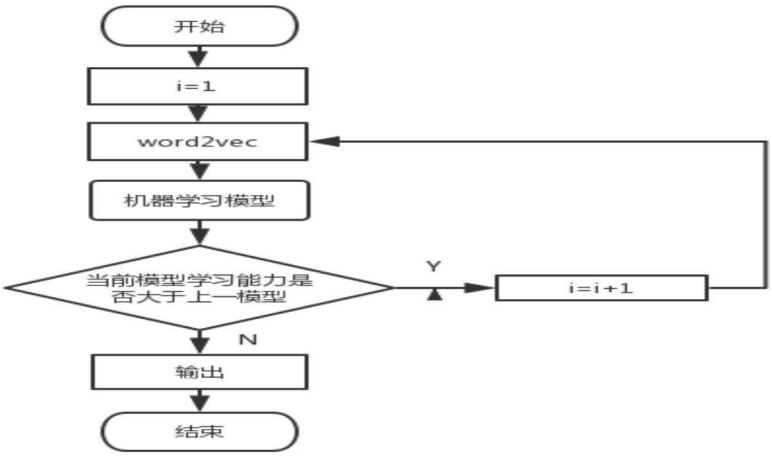

开始
i=1
word2vec
机器学习模型
当前模型学习能力是否大于上一模型
Y
i=i+1
N
输出
结束

图2

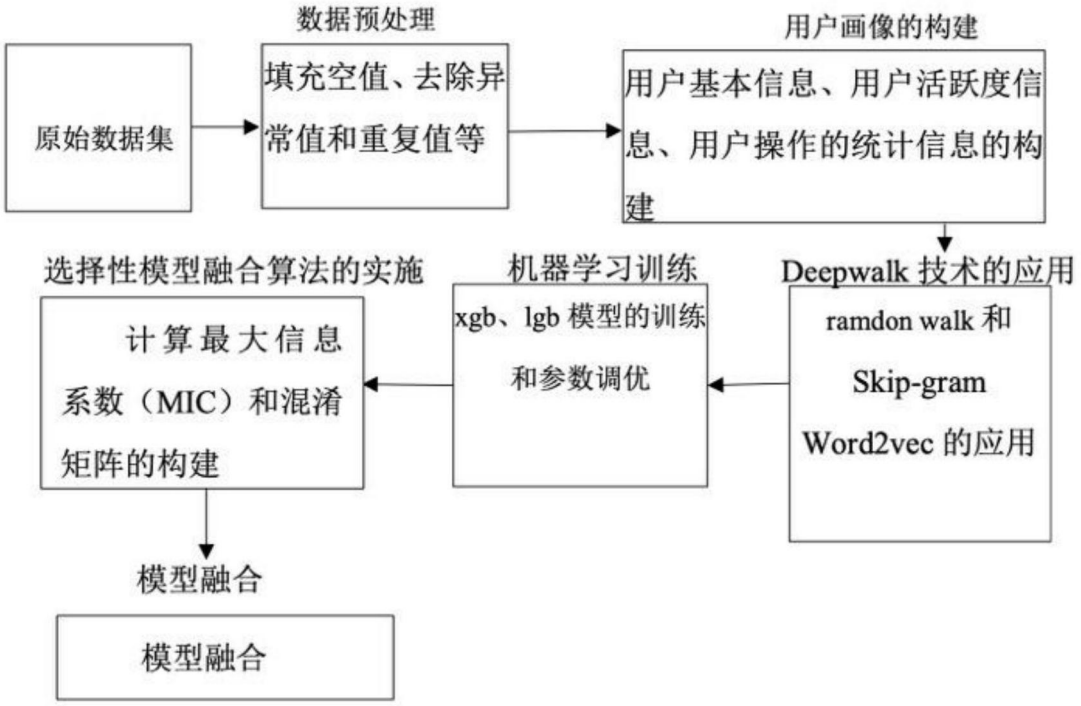

数据预处理
用户画像的构建
原始数据集
填充空值、去除异常值和重复值等
用户基本信息、用户活跃度信息、用户操作的统计信息的构建
选择性模型融合算法的实施
机器学习训练
Deepwalk 技术的应用
计算最大信息系数（MIC）和混淆矩阵的构建
xgb、lgb 模型的训练和参数调优
ramdon walk 和 Skip-gram Word2vec 的应用
模型融合
模型融合

图3

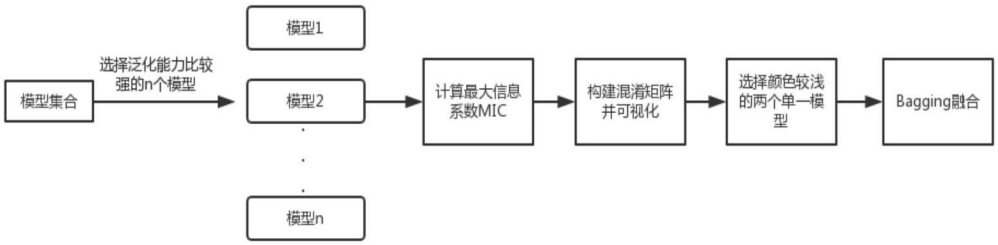

模型集合
选择泛化能力比较强的n个模型
模型1
模型2
模型n
计算最大信息系数MIC
构建混淆矩阵并可视化
选择颜色较浅的两个单一模型
Bagging融合

图4

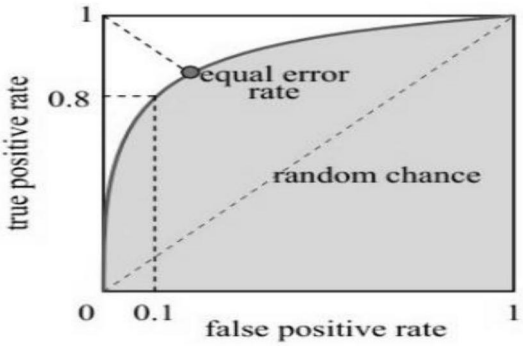

1
0.8
true positive rate
equal error rate
random chance
0
0.1
1
false positive rate

图5

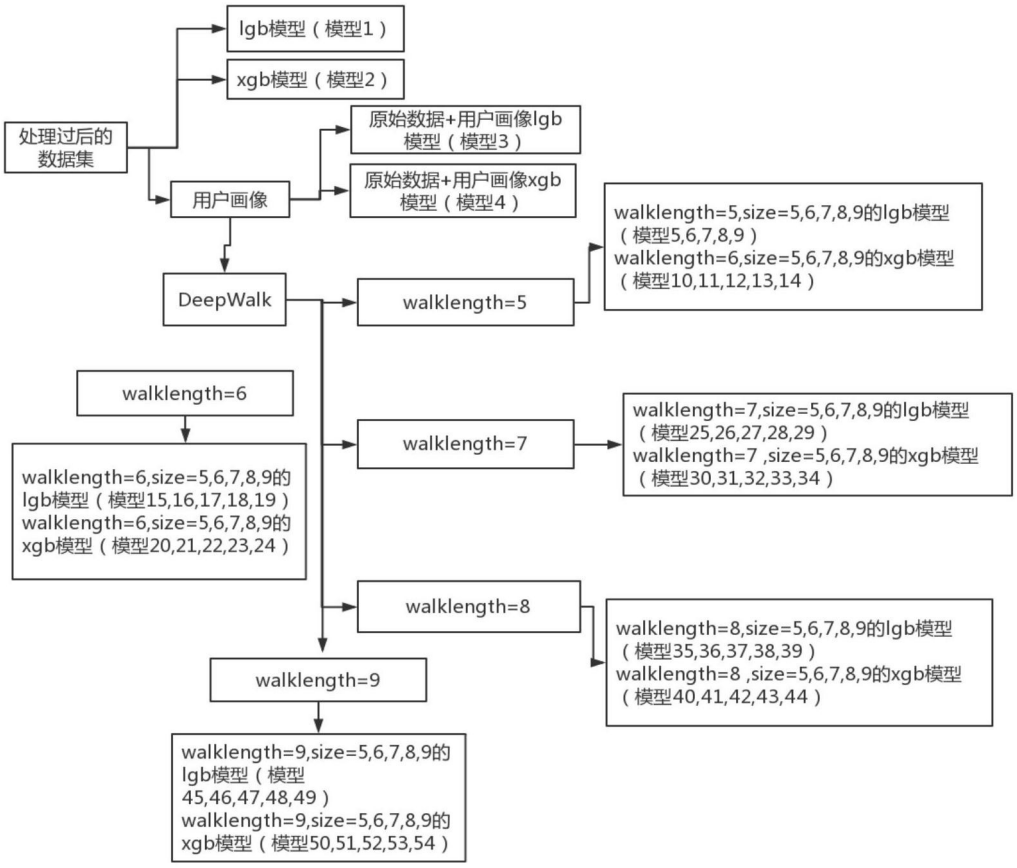

lgb模型（模型1）
xgb模型（模型2）
处理过后的数据集
原始数据+用户画像lgb模型（模型3）
原始数据+用户画像xgb模型（模型4）
用户画像
DeepWalk
walklength=5
walklength=5,size=5,6,7,8,9的lgb模型（模型5,6,7,8,9）
walklength=6,size=5,6,7,8,9的xgb模型（模型10,11,12,13,14）
walklength=6
walklength=6,size=5,6,7,8,9的lgb模型（模型15,16,17,18,19）
walklength=6,size=5,6,7,8,9的xgb模型（模型20,21,22,23,24）
walklength=7
walklength=7,size=5,6,7,8,9的lgb模型（模型25,26,27,28,29）
walklength=7 ,size=5,6,7,8,9的xgb模型（模型30,31,32,33,34）
walklength=8
walklength=8,size=5,6,7,8,9的lgb模型（模型35,36,37,38,39）
walklength=8 ,size=5,6,7,8,9的xgb模型（模型40,41,42,43,44）
walklength=9
walklength=9,size=5,6,7,8,9的lgb模型（模型45,46,47,48,49）
walklength=9,size=5,6,7,8,9的xgb模型（模型50,51,52,53,54）

图6

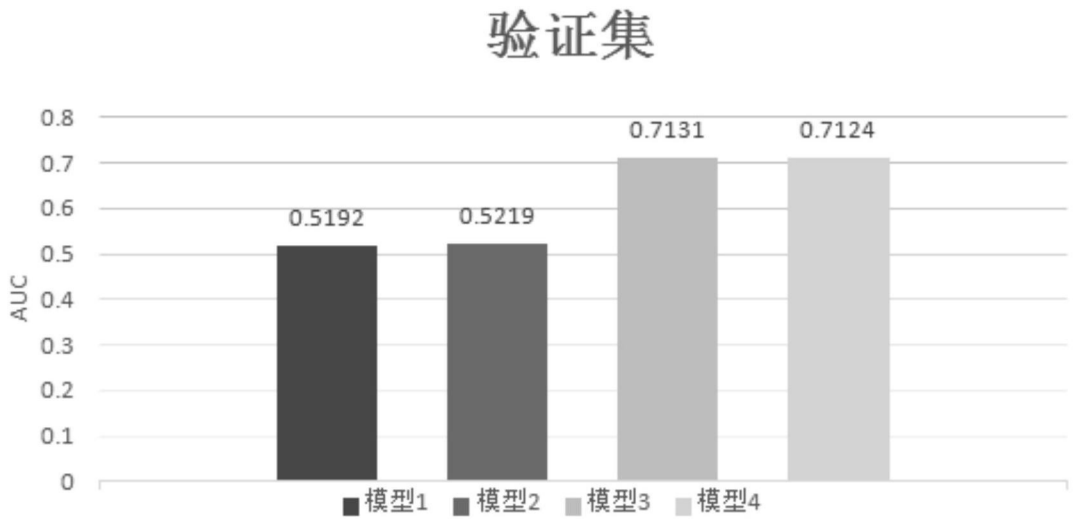

验证集
AUC
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.5192
0.5219
0.7131
0.7124
模型1
模型2
模型3
模型4

图7

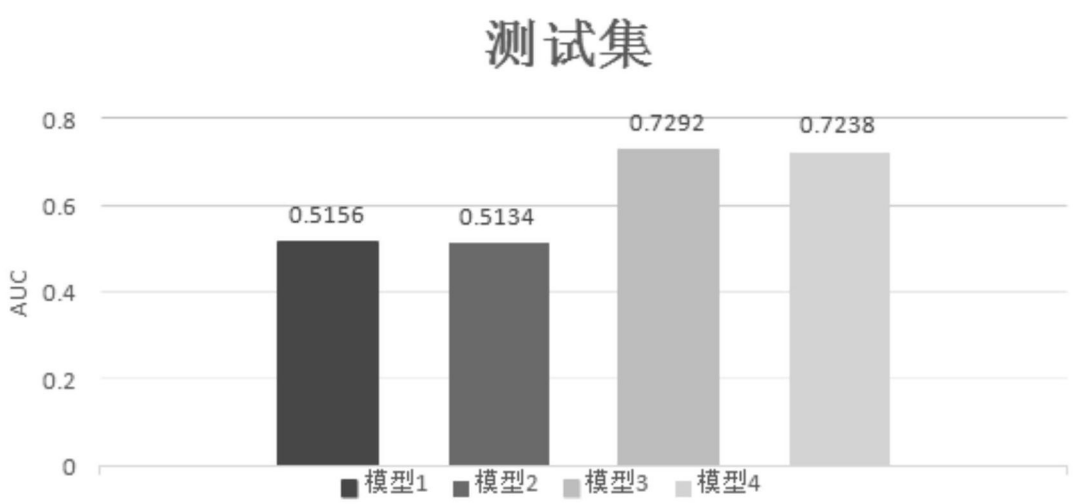

测试集
AUC
0.8
0.6
0.4
0.2
0
0.5156
0.5134
0.7292
0.7238
模型1
模型2
模型3
模型4

图8

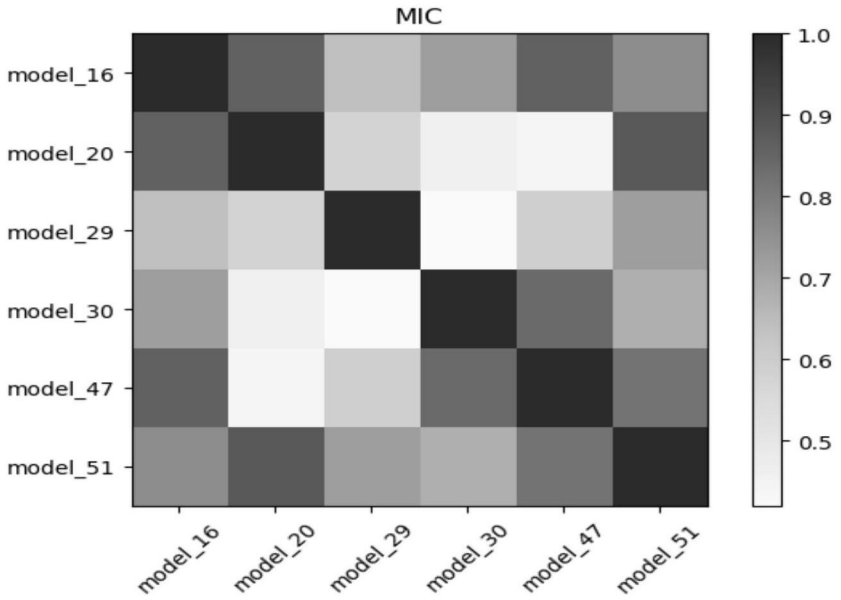

MIC
model_16
model_20
model_29
model_30
model_47
model_51
1.0
0.9
0.8
0.7
0.6
0.5

图9

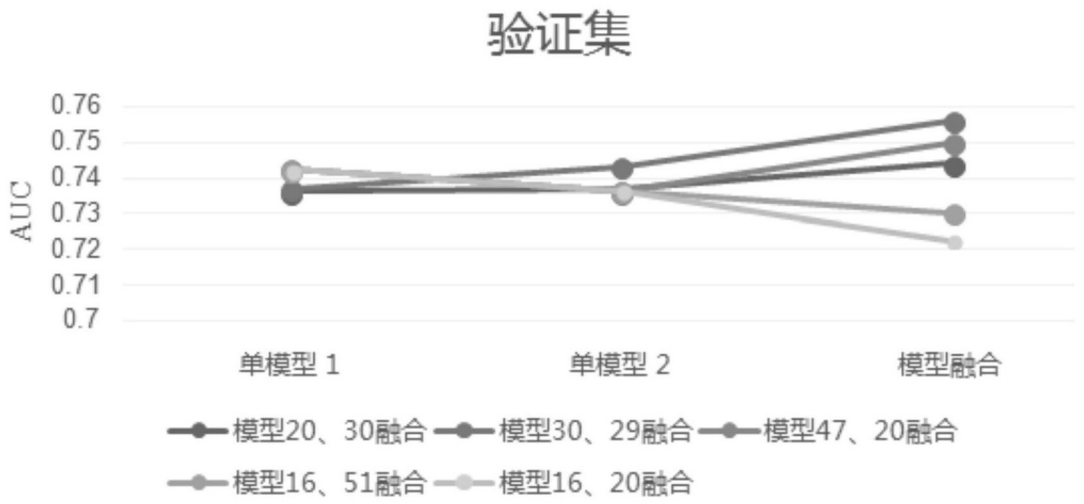

验证集
AUC
0.76
0.75
0.74
0.73
0.72
0.71
0.7
单模型 1
单模型 2
模型融合
模型20、30融合
模型30、29融合
模型47、20融合
模型16、51融合
模型16、20融合

图10

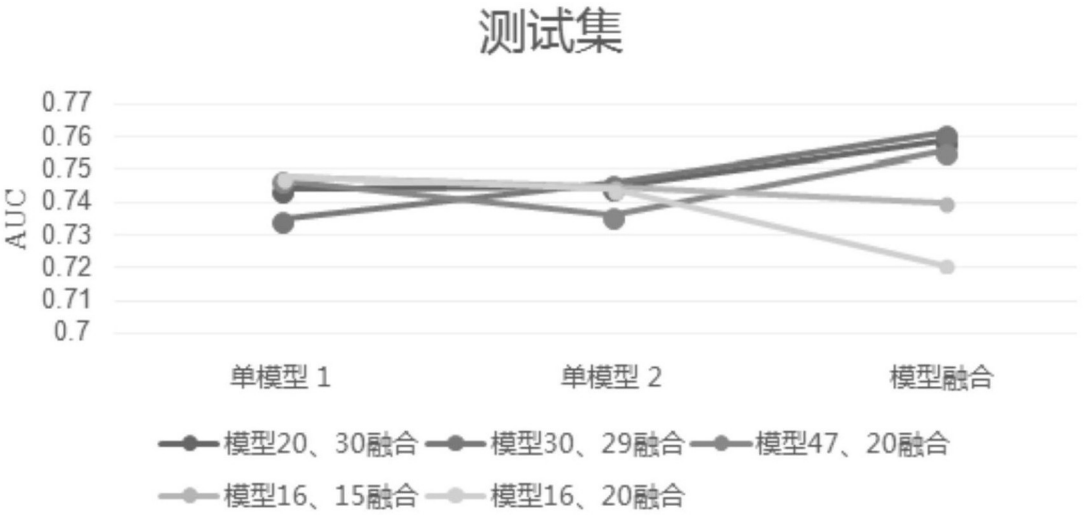

测试集
AUC
0.77
0.76
0.75
0.74
0.73
0.72
0.71
0.7
单模型 1
单模型 2
模型融合
模型20、30融合
模型30、29融合
模型47、20融合
模型16、15融合
模型16、20融合

图11

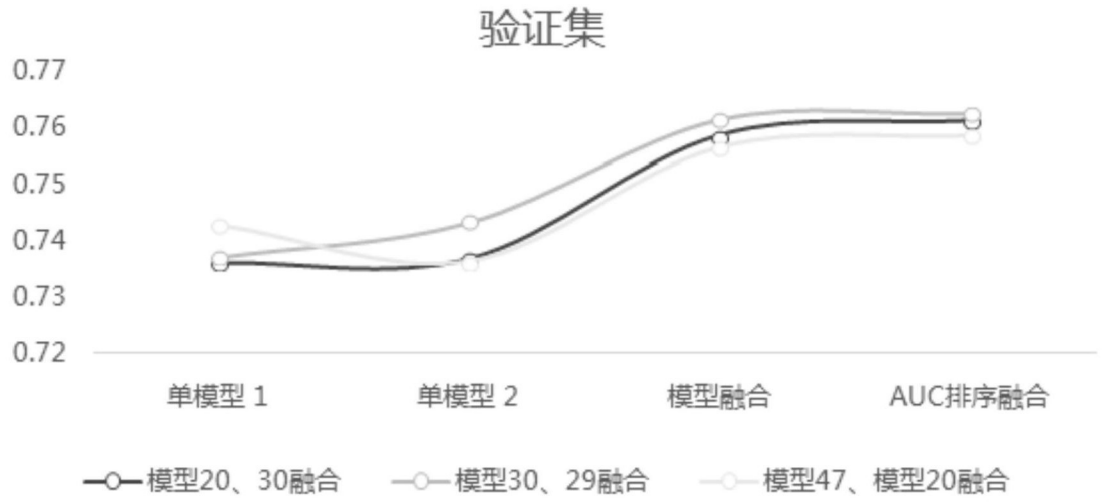

验证集
0.77
0.76
0.75
0.74
0.73
0.72
单模型 1
单模型 2
模型融合
AUC排序融合
模型20、30融合
模型30、29融合
模型47、模型20融合

图12

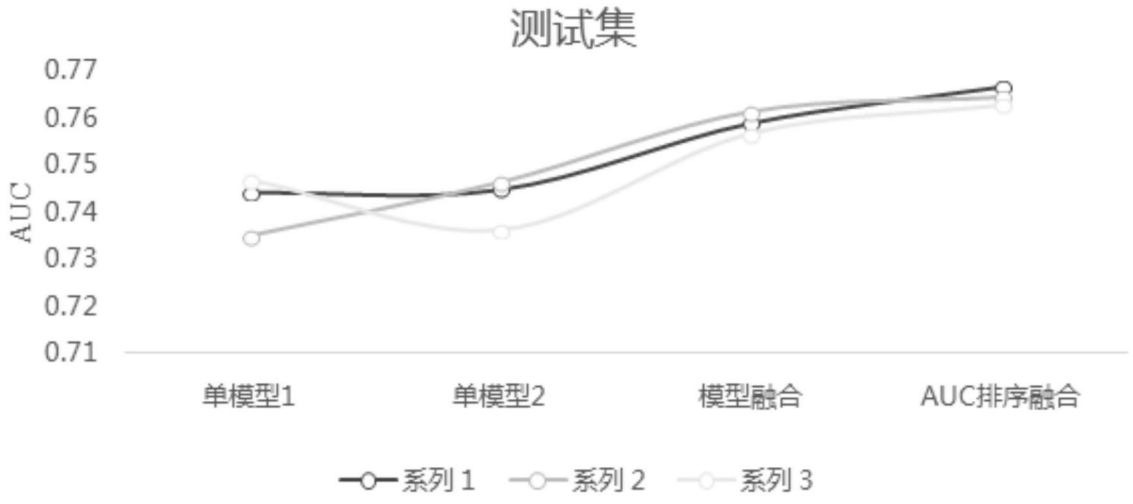

测试集
AUC
0.77
0.76
0.75
0.74
0.73
0.72
0.71
单模型1
单模型2
模型融合
AUC排序融合
系列 1
系列 2
系列 3

图13