基于深度学习的用户画像研究

兰秋军，周妹璇

（湖南大学工商管理学院，湖南省、长沙市，410082）

摘要：用户画像在个性化推荐以及精准营销中起到了重要作用.如今在大数据背景下，传统的浅层学习方法由于不能深入挖掘特征之间的关系，尤其是在高维特征基础上预测用户画像标签时面临挑战。本文采用在原有特征基础上进行特征拓展与独热编码方法，并利用深度神经网络对用户画像标签进行预测，通过与决策树、逻辑回归算法的实验对比，本文方法获得较高的 AUC 值，达 0.792。本文进一步对深度神经网络的网络层数、神经元个数及正则化技术进行探索。

关键词：用户画像；高维特征；标签；深度神经网络

中图分类号： TP391　　文献标识码：A

0 引言

随着互联网技术的快速发展与人们物质生活水平的提高，移动终端设备迅速普及，各种移动应用也竞相进入人们的生活，导致互联网信息呈爆炸式增长。面对海量信息与当今不断加快的生活节奏，人们遭受信息筛选困难，信息处理效率低下的困扰，而企业由于对用户的营销不够细化，不合时宜与不合偏好的推送给用户造成骚扰，甚至流失重要客户。因此，如何利用各种与用户相关的信息精准地构建用户标签，在用户和信息需求之间实现自动化匹配成为当下各大电商企业重点研究的方向，用户画像及其相关技术就在这种背景下应运而生。

虽然目前已有多种用户画像的研究方法，但大多基于浅层学习方法，没有将数据特征之间的内在联系进一步挖掘，在某些情况下面临着挑战，比如广告点击率与转化率预测这类问题中，输入特征具有数量大、维度高以及稀疏性，浅层学习方法预测效果还有待进一步提升。如何准确高效地预测用户画像标签不仅有助于企业实现精准营销增加收益，还能提高用户对个性化服务的满意度，这也是本文研究目的与意义所在。

本文研究内容是对用户画像的定义和构建流程进行归纳总结，利用用户相关的基础特征进行拓展与独热编码，并使用深度神经网络方法来预测用户画像标签，从特征学习的角度挖掘出多维度特征之间隐藏的联系，从而提高用户画像的标签预测效果。本文在腾讯社交广告算法大赛的初赛数据上进行实验，并与逻辑回归（LR）和决策树进行对比，深度神经网络方法获得更高的 AUC 值，并进一步探索神经网络深度与神经元数量及 dropout 技术对模型性能的影响。

1 相关研究

用户画像首次被提出可以追溯到“交互设计之父”Alan Cooper 提出的 Personas are a concrete representation of target users.”[1]，他指出用户画像（Persona）是对目标用户的具体表示，又被称为用户角色，是基于大量目标用户群的真实信息构建的用户标签体系，是对产品或服务的目标人群做出的特征刻画。通过收集用户的人口统计信息、偏好信息及行为信息等，构建出用户画像，可以让产品经理更好地了解用户，设计出合适的产品原型，因此，用户画像是用户需求与产品设计的联系的桥梁。

通过查阅国内外用户画像相关文献不难发现，用户画像最早起步于国外，其理论和实践

相对于国内来说比较成熟。在早期，用户数据的来源渠道比较少，数据量也相对少的时期，用户画像的研究主要基于统计分析层面，通过用户调研来构建用户画像标签。加利福尼亚大学的 Syskill 和 Webert 就是通过显式地收集网站用户对页面的满意度，然后通过统计分析逐步学习构建出用户兴趣模型[2]。后来，随着互联网及信息采集技术的发展，CUM 大学开发的 Web Watcher 以及后来的 Personal Web Watcher[3]，可以通过数据采集器，记录互联网上用户产生的各种浏览行为及用户的兴趣偏好，实现对用户兴趣模型的构建，并随着数据的不断累积扩大而更新系统模型，用户画像标签也更加丰富。

近年来，随着互联网的海量数据呈爆炸式增长，统计方法的不断优化和硬件设备计算能力的不断提升，在大数据背景下，众多企业的用户画像研究开始面临新的机遇和极大的挑战，通过已有标签构建用户画像已经不能满足人们的个性化需求，如何使用算法模型实现用户画像中的用户行为标签的预测，已经成为产品经理及运营工作人员的关注重点。用户画像的含义也处于动态变化中，这种基于数据建模的用户画像模型被称之为 User Profile[4]。

目前的用户画像研究主要集中在三大方向上：用户属性，用户偏好，用户行为三个主要方面。其中用户属性的研究侧重于显式地搜集用户特征信息，主要体现在社会化标注系统领域[5] [6] [7] [8]，通过社会化标注系统搜集比较全面的用户信息，用于多方位的了解用户。用户偏好研究侧重于制定兴趣度度量方法[9]，评估用户的兴趣度，提高个性化推荐质量；用户行为的研究侧重于用户行为趋势的预测，比如用户流失行为的预测中[10]，有利于提前发现问题，找出对应策略，防止客户流失；在用户的欠费预测问题中[11]，有利于发现电力客户欠费特征，为电力公司提供决策支持。

不同研究领域的用户画像研究方法也会有所差异，常用的有决策树[12][13]、逻辑回归[14][15]、支持向量机[15]及神经网络[16][17]等模型。其中，在浅层学习方法中的逻辑回归算法模型凭借其简单性与易理解性得到了广泛的应用。但是作为一种浅层学习方法，在面临高维稀疏特时，其预测性能有待进一步提升。

随着目前大数据分析需求的越来越多，计算计硬件设备的不断更新，计算能力的不断加强，深度学习开始成为许多学者研究的重点。深度学习是模仿人脑结构，建立数据从输入到输出的映射，其中各层之间进行了特征的探索分析。深度学习的优势就在于能够发掘特征之间的隐含关系[18]，从原始特征中抽象出具有代表性的特征，在利用这些抽象性特征进行预测时能达到更好的效果。目前深度学习利用其特征提取优势，已经在语音识别[19]与图像识别[20]领域取得了巨大成果。本文将在深度神经网络的基础上对用户画像进行分析与探索。

2 用户画像的定义与构建流程

2.1 用户画像的定义

在当今的大数据背景下，伴随着机器学习及人工智能技术的不断发展与进步，众多企业重点关注的用户画像也越来越多地被媒体报道，进入人们的视野。但是通过浏览各种渠道关于用户画像的相关内容，发现用户画像的定义有些不清晰，通过阅读文献，本文对用户画像的定义做出归纳说明。

用户画像的概念最早是由交互设计领域的 Alan Cooper 提出的，他指出，用户画像（Persona）是现实中目标用户的可靠表示[1]，主要通过调研分析或者通过用户的信息填写及资料上传等方式的方式获取用户信息,然后根据用户的人口统计信息及行为信息，划分为不同类型的用户群，最后对每种类型中的用户打上典型的标签，包括姓名、性别、年龄、居住、

兴趣爱好及场景等，从而构建出用户画像。Persona 适用于产品开发的早期，在不确定用户具体需求与兴趣偏好的时候进行的用户研究，从而明确用户的产品需求、使用场景、以及预见产品使用过程中可能出现的问题等。但是 Persona 有一定的缺点，因为用户属性在被评估时难以量化与证伪，无法确定是否为真的目标用户，并且目标用户也会随时间推移而发生变化，所以需要对 Persona 进行不断的更新。

我们知道，一般在网络化产品上线前期，通过用户注册、填写信息、提交资料以及浏览网页等会不断有用户的信息积累，数量逐渐增多，信息的维度也逐渐变高。这些数据将会在产品运营的中后期发挥巨大作用。为了保持在激烈的市场竞争中持续占有强而有力的优势，企业需要有针对性地对用户进行个性化营销，留住老用户，吸引新用户。我们可以通过数据建模的方式来预测用户的行为标签，从而构建更完善的用户画像，实现用户的细分或者聚类，比如，使用用户数据预测其兴趣偏好及未来购买意向，从而为其推荐合适的广告或可能偏好的商品。这时候的用户画像建模过程称为 User Profile，也是本文用户画像研究的角度。

综上，目前用户画像主要有两种含义。第一种 Persona，优势又被称为用户角色，主要被产品及用户研究人员使用，适用于产品初期的定位与用户调研，第二种 User Profile，主要被数据分析人员使用，适用于数据及产品运营的中后期。本文是从 User Profile 角度对用户画像进行的研究。

2.2 用户画像的构建流程

结合上文提到的用户画像中的标签类型，本文认为，用户画像的构建主要分为以下四个阶段，如图 1 所示：

图 1　用户画像构建过程

各阶段具体说明如下：

（1）基础数据采集

基础数据的采集按照标签来源可以分为三种，用户信息、资源信息和上下文信息。对于用户个人信息采集，可以通过用户的注册信息以及上传的包含个人信息的文件来获得；用户的行为信息可以通过用户网上行为包括浏览记录、购物记录、视频观看记录等获得；用户的偏好信息可以从用户订阅内容的标签或者个性化定制标签中获得。对于资源信息，可以从资源本身的介绍信息来获得。对于上下文信息，通过业务场景的分析来获得。其实上下文信息

往往与用户行为相关联，因为用户的行为随着业务场景的变化而动态变化，在保存用户行为信息的时候往往将当时的业务场景一起存储，这样便于更好地进行分析。

（2）数据分析与处理

一般采集到的原始数据非常不规范，往往还有缺失值、异常值，以及一些格式不一致等问题，这时的数据不适合直接用来建模，需要经过数据清洗及预处理。该阶段主要对原始的数据进行探索性分析与处理，包括对数据分布的统计分析、相关性分析、数据清洗、归一化、缺失值处理等。当需要使用多个数据源文件时还需要数据的集成操作。

（3）用户标签建模

该阶段是对上阶段处理后的数据进行建模过程，抽象出用户的标签，进而预测出用户潜在的行为及偏好标签。在这个阶段，我们往往需要用到不同的算法模型来为用户贴标签。比如，电信用户的流失预测，根据用户话费套餐使用情况，利用决策树算法或者神经网络算法判断该用户是否会流失。不同的业务与数据，采用的算法模型也会有所不同，需要针对具体情况做好数据分析与特征处理工作，然后再选择比较合适的几种算法做实验，通过不同算法模型的实验结果与性能对比来选择最合适的模型，有时还可以通过模型融合的方式来提高预测精确度。

（4）用户画像基本成型与更新

该阶段是对上一阶段的深入，首选通过用户的基本信息、行为信息、兴趣偏好结合资源及上下文等信息对用户进行标签化，还可以根据业务需求将标签进行分层，使用户画像基本成型，再补充上一阶段预测的新标签，实现对用户画像的完善与更新，流程如图 2 所示。

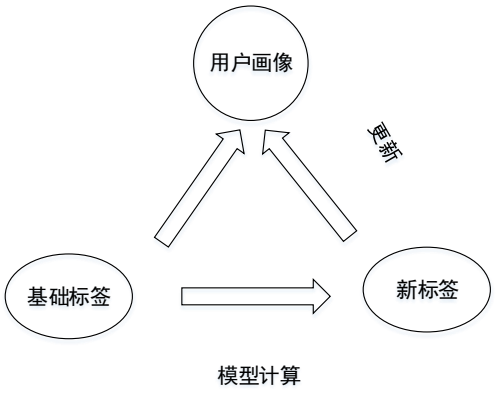

图 2　用户画像成型与更新

3　深度神经网络预测模型

3.1　深度神经网络概述

神经网络模型是模拟人脑原理，实现特征学习的一种模型，其基本组成单元是感知机，如图 3 所示，相当于人脑中的神经元结构。可以看出，感知器由两层神经元组成，输入层把接收到的外界信息传给输出层，感知器只对输出层进行激活函数处理，这种情况下学习能力是十分有限的，甚至不能解决非线性可分问题。但是把许多这样的感知器按一定的层次结构连接起来得到神经网络后，就能解决线性不可分问题。其中输入层与输出层之间的网络层统称为隐含层。深度神经网络中，输入层、隐含层及输出层都可以有多个，模型相对复杂。

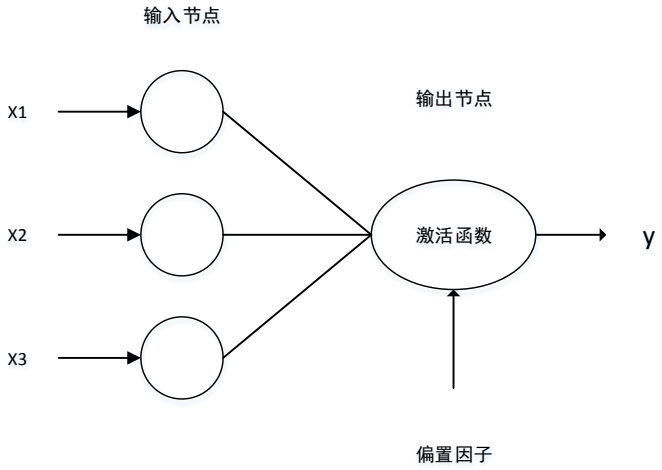

图 3　感知机组成

本文要预测的用户画像标签是用户转化或者未转化的概率值，输出神经元只有两个即可。关键在于输入特征的预处理，以及隐含层的网络设置。

3.2　激活函数的选择

深度神经网络中激活函数一般是非线性函数，定义了对神经元输出的映射，提供了网络的非线性建模能力。几乎所有的连续可导函数都可以作为激活函数。常用的激活函数有 Sigmoid、tanh。但是 Sigmoid 的软饱和性，即在定义域内任意处可导，但是两侧导数逐渐趋近于 0，这使得网络很容易陷入饱和区，发生梯度消失。这使得神经网络在很长一段时间难以得到有效的训练，也是神经网络发展的阻力之一。而 tanh 也具有软饱和性，而且收敛速度比 Sigmoid 更快，输出均值也更接近于 0，也同样容易出现梯度消失问题。后来 Relu 这一新激活函数的出现使得深度网络取得新的突破。与传统的激活函数相比，该函数的优点如下：

（1）由于 ReLu 的非饱和性，可以减轻梯度消失问题。

（2）能够强制某些神经元的输出数据为 0，使神经网络的稀疏表达能力提高。

（3）降低了参数之间的相互依赖关系，有效地减轻过拟合问题的产生。

（4）与传统的 Sigmoid 型激活函数相比较，Relu 不涉及除法和指数运算，具有运算速度快的优势。

（5）具有在计算中保持着分段线性的特点，不像 S 型激活函数容易将有用的特征在传播过程中丢弃。

由此可见，Relu 也同生物神经元的激活本质更相似，即与神经元信号激励原理更加符合。在进行反向传播求误差梯度时，ReLu 收敛速度很快，表现出了更突出的优点。

3.3　Dropout 技术

Dropout 是广泛应用在深度神经网络模型中的正则化技术[21]，与 L1 和 L2 正则化的区别在于它是通过改变神经网络的结构来防止过拟合，而后者是通过改变代价函数来实现的。Dropout，顾名思义，是在每次训练中通过随机删除一部分神经元来进行预测，但不是真的删除，只是不参与建模而已，并且每次随机删除的不一样，这相当于是训练了多种模型，最后相当于取各个模型的预测均值。

4 实验设计与分析

4.1 数据集

本次研究采用的数据是 2017 年腾讯社交广告高校算法大赛的初赛数据集，其中包含了用户的基本特征、广告特征及上下文特征，本次研究以广告的转化率为研究对象，这里的转化是指用户点击广告后安装并启用相应 APP，转化率属于用户画像中用户的行为标签。

该数据是从腾讯社交广告系统随机抽取的某连续半个月的日志数据，每一条训练样本代表的是一条广告点击日志，样本的转化标签对应的字段名为 label，有 0 和 1 两个取值，0 代表未发生转化，1 发生转化。因为出于对用户隐私及公司商业信息安全的考虑，数据经过加密处理，只给出各变量的说明，但是并未给出原始数据及具体数值的含义，从而无法从经验去分析各个特征，进而增加了分析的难度。本次比赛的完整数据文件的明细及规模如表 1 所示：

表 1　数据文件

文件名	数据规模（行×列）	大小	描述
train.csv	3749528×8	112M	训练集
test.csv	338489×8	12.2M	测试集
app_categories.csv	217041×2	2.04M	APP 特征文件
user.csv	2805118×8	70.5M	用户基础特征文件
user_app_actions.csv	6003471×3	109M	用户 APP 安装流水
user_installedapps.csv	84039009×2	941M	用户 APP 列表
app_categories.csv	217041×2	2.04M	APP 特征文件
ad.csv	6582×6	144K	广告特征文件
position.csv	7645×3	66K	广告位特征文件

4.2 数据分析与处理

4.2.1 数据集成

在训练集中，每一条样本包含 8 个属性，我们可以将这些属性归为三大类：分别是广告属性、用户属性、上下文属性，属性明细如表 2 所示。

表 2　实验数据基础属性说明

属性名	属性描述	属性类别
label	是否转化	用户属性
clickTime	点击时间	上下文属性
conversionTime	转化时间	上下文属性
creativeID	广告素材	广告属性
userID	用户	用户属性
positionID	广告位	上下文属性
connectionType	联网方式	上下文属性
telecomsOperator	手机运营商	上下文属性
label	是否转化	用户属性

训练集与测试集以外的文件作为特征扩展数据，将这些文件进行集成使用户特征得以拓展。除了基本的文件集成外，还可以通过用户安装列表与安装流水文件，统计每个用户历史安装 APP 数量，可以理解为用户的活跃度，作为新的特征列，字段名设置为 userapp，可供

建模使用。用户特征扩展后的具体信息如表 3 所示，可供建模使用。

表 3　用户特征说明

属性名	说明	属性类别
userID	用户 ID	离散型
age	年龄	连续型
gender	性别	离散型
education	学历	离散型
marriageStatus	婚恋状态	离散型
haveBaby	育儿状态	离散型
hometown	家乡/籍贯	离散型
residence	常住地	离散型
userapp	用户 app 的安装数量	连续型

广告特征扩展后的具体信息如表 4 所示，还可以统计每个 app 被安装的次数，可以理解为 app 的热度，作为新的特征列，字段名设置为 appuser，可供建模使用。

表 4　广告特征说明

属性名	说明	属性类别
advertiserID	广告主	离散型
campaignID	推广计划	离散型
adID	广告创意与展示设置	离散型
creativeID	广告素材 ID	离散型
appID	移动应用 ID	离散型
appCategory	移动应用的类型	离散型
appPlatform	APP 操作系统平台	离散型
appuser	App 的用户群数量	连续型
advertiserID	广告主	离散型

上下文特征扩展后的具体信息如表 5 所示，可供建模使用。

表 5　上下文特征说明

属性名	说明	属性类别
positionID	广告曝光的具体位置	离散型
sitesetID	多个广告位的聚合	离散型
positionType	广告位规格的类别	离散型
connectionType	移动设备当前使用的联网方式	离散型
connectionType	移动设备当前使用的联网方式	离散型

4.2.2　缺失值处理

因为本次实验数据是涉及到用户注册使用 APP 的信息数据，所以难免会有缺失值，缺失的原因主要有两方面：一方面是 APP 在用户注册使用时，往往不需要填写很全面的信息；另一方面，用户出于对个人隐私的保护，会对年龄、婚育等信息有所隐瞒，而 APP 更无法检测出数据的真实性。在数据分析中，缺失值的统计是十分必要的，此次研究统计各个变量中的缺失值占比，可以发现主要是用户的信息缺失，其中育儿状态（haveBaby）缺失率达 80%，婚恋状态（marriageStatus）缺失率达 40%，籍贯（hometown）缺失率达 36.84%，构建的新特征 userapp, appuser 缺失率高达 60% 、50%。对于缺失值的处理，本文采用简单的删除方式，不考虑将以上五个缺失率比较大的特征参与建模。除了对以上缺失率比较高的特

征进行剔除外，剩余缺失特征依然按照官方文件，统一取值为 0，隐含了用户的某种特征，可以参与建模。

4.2.3 特征探索与处理

关于特征变量的分布分析，根据特征变量的取值类型可以分为两种情况：

（1）特征类型为连续型

可以自定义区间进行类别的重新划分，作为新的属性。比如 age，可以按照 0-12，13-30，31-50，51-80 划分，其中 0 表示未知，但是根据我们的经验判断，12 岁以下的用户很有可能是虚假年龄，所以也归为未知一类。13-30 属于年轻人，31-50 中年人，50 以上为老年人。

（2）特征类型为离散型

可以统计每个离散值的频数分布及转化率，分别代表了该特征对应的点击热度与转化率强度，然后再对频数及转化率进行离散化，得到新特征。在此次的数据中，基础特征主要是离散型特征，各种 ID 特征取值以及个别多值类特征均按照这种方式处理，具体统计如图 4 所示。

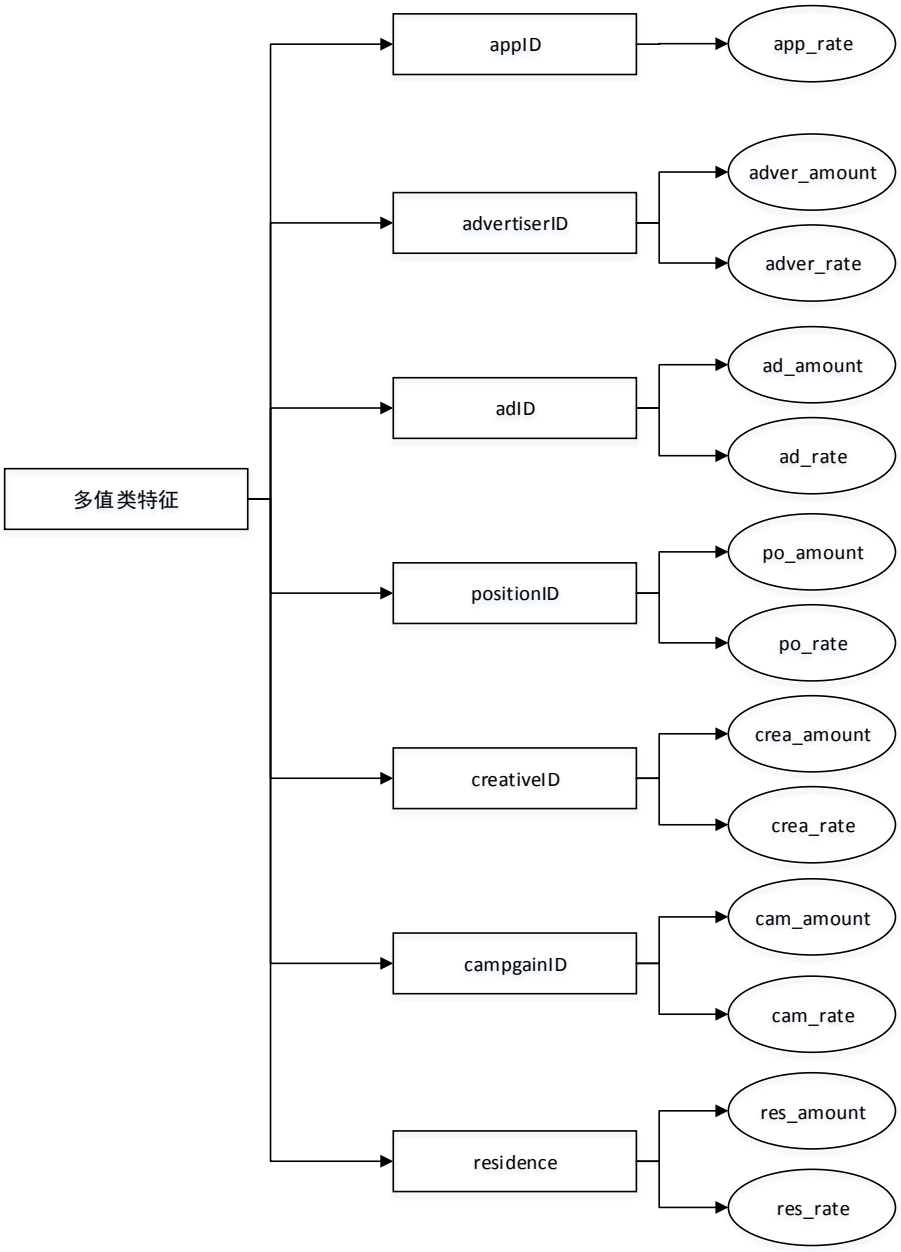

图 4　多值类特征的拓展特征

但是一般不会将 ID 类特征作为模型的输入，而 ID 类特征其实隐含着其他信息，比如转化率高低、点击量等。可以统计其对应的转化率等信息，再进行离散化处理。本文采用分位数分组的方式对以上特征根据频数统计进行离散化，使各组变量值总和近似，分组数量设置为 30，作为 ID 类特征的拓展特征。

4.2.4　时间相关性分析

由于广告的时效性与用户移动端上网时间段的分布规律，我们需要从时间的角度考虑其与转化率的相关性。一方面以天为单位统计历史的转化率情况，另一方面以小时为单位进行转化率的统计分析。

（1）历史日转化率分布

将 clicktime 以天为单位进行划分，分析统计周期内每天的转化率，得到如图 5 所示分

布结果：

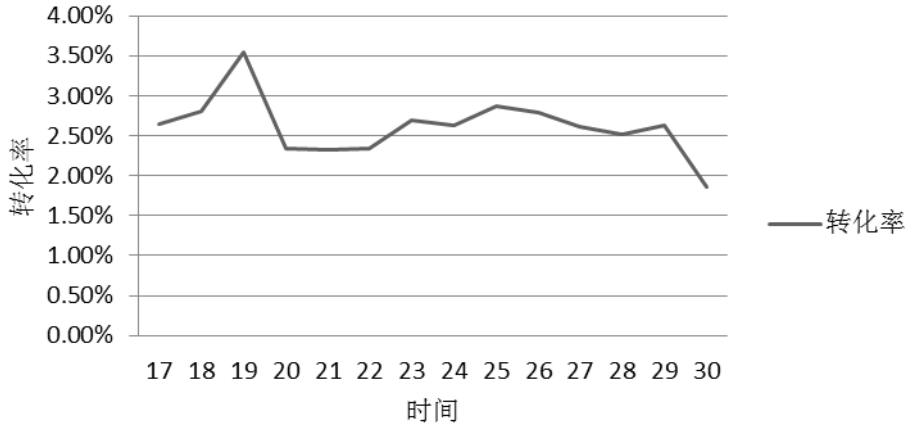

图 5　日转化率分布

由图 4 可以看出，整体的日转化率在 2.5%上下波动，而 19 日的转化率偏高，30 日的偏低。关于 30 日转化率偏低问题，我们通过腾讯官方的数据说明可以理解，官方数据说明中指出，从用户点击广告到广告系统得知用户转化即激活了 App（如果有），通常会有一段时间间隔，主要原因由以下两个方面导致：①用户可能在下载之后并没有立刻激活 app，可能直接卸载或者过一段时间才开始启用,这也符合我们生活场景,有时一时兴起下载了 APP，但是又不会立刻使用，有时因为无意间点击了下载，又觉得不感兴趣直接卸载；②用户启动 App 的行为不能及时同步到广告系统，而是需要等到广告主上报回传才得以知晓，因此，通常会有一定的延时。这里回流时间表示了广告主把 App 激活数据上报给广告系统的时间，回流时间超过 5 天的数据会被系统忽略。

值得注意的是，本次竞赛的训练数据提供的截止第 31 天 0 点的广告日志，因此，对于最后几天的训练数据，某些 label=0 并不够准确，可能广告系统会在第 31 天之后得知 label 实际上为 1。而关于 19 日转化率偏高的问题，由于数据的保密性较强，我们无法分析得知其中的原因，也进行了删除处理，使模型更具泛化能力。

（2）历史每日各时间段频数分布

从 clicktime 中提取出 hour，得到新的特征 clickTime_hour，以小时为单位进行划分，统计各个时间段的频数分布，如图 6 所示：

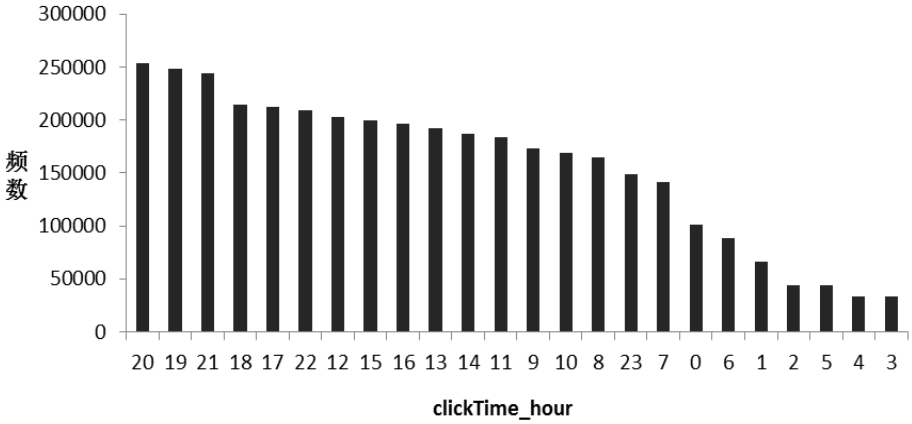

图 6　clickTime_hour 频数分布

从图中数据的分析，可以将每日的时间划分为 3 个阶段，其中 17-22 为晚上，8-16 为白

天，0-7 和 23 为闲时。频数的高低可以理解为用户在这个时间段点击广告的活跃度。

4.3 实验环境与评价指标

本文的实验运行环境为 Windows7 操作系统，编程语言为 python，主要利用 pandas 库对数据进行预处理，用 Scikit-Learn 模块实现了逻辑回归及决策树算法模型，使用基于 TensorFlow 后端的 Keras 编写深度神经网络模型。

转化率预估数据是不平衡的，正负样本比例为 1:39，类似于点击率预估，常采用 AUC（Area Under Curve）作为转化率标签预测的评价标准，代表的是 ROC 曲线下的面积。混淆矩阵是 ROC 曲线的重要知识，表 6 所示，Positive 和 Negative 表示实际所属类别，True 和 False 表示分类是否正确，ROC 曲线的纵坐标代表真正率(True Positive Rate，TPR)，TPR 表示正样本分为正的概率，TPR=TP/(TP+FN)，横坐标代表假正率(False Positive Rate，FPR)，FPR=FP/(FP+TN)，表示负样本错误的分为正的概率。面积值 AUC 的范围是[0.5,1]。

表 6 混淆矩阵

	Positive	Negative
True	TP	TN
False	FP	FN

广告转化率的预估是以广告是否发生转化作为分类标准的二分问题。混淆矩阵对应广告点击率预估问题，(Positive，Negative)表示转化率预估实例的样本类别，(True，False)表示广告转化率预估是否正确。AUC 的值越大，转化率预测越准确。

4.4 实验结果与分析

根据本章以上各小节的数据探索分析，具有多值的 ID 类特征及现居住城市 residence 按照频数以及转化率进行分组，然后将这些新的分类特征进行 one-hot 编码。考虑到硬件设备的能力，采用随机抽样的方式，从中抽取 30 万条样本进行实验，其中 75%作为训练集，25%作为测试集，采用五折交叉验证。

使用逻辑回归，决策树作为深度神经网络模型的对比进行实验，实验结果如下：

表 7 不同模型的 AUC 值对比

模型	AUC
LR	0.787
DT	0.768
DNN	0.792

可以得出，高维稀疏特征作为模型的输入进行预测时，深度神经网络的预测能力相对较强，这是由于其能深入挖掘特征之间的隐含关系，通过层层处理，在最后一层得到的是特征的抽象化表示，虽然不易解释，但的确是特征的一种表示，在这种特征下进行预测，效果较好。

4.5 深度神经网络的进一步探索

4.5.1 隐含层个数的分析

本文进一步对深度神经网络的层数及神经元个数进行探索分析，图 7 为各层神经元个数为 50 的情况下，随着网络层数的逐渐加深预测效果的变化。可以发现，一开始网络层数增

多，预测效果明显提高，但是增加到一定程度，实验结果不再有明显提高，而是有波动变化，考虑到神经网络参数越多训练时间越长的问题，我们选择效果较好的少数层数作为进一步实验的基础，这里最大层数取值为 4 层，作为进一步实验的基数。

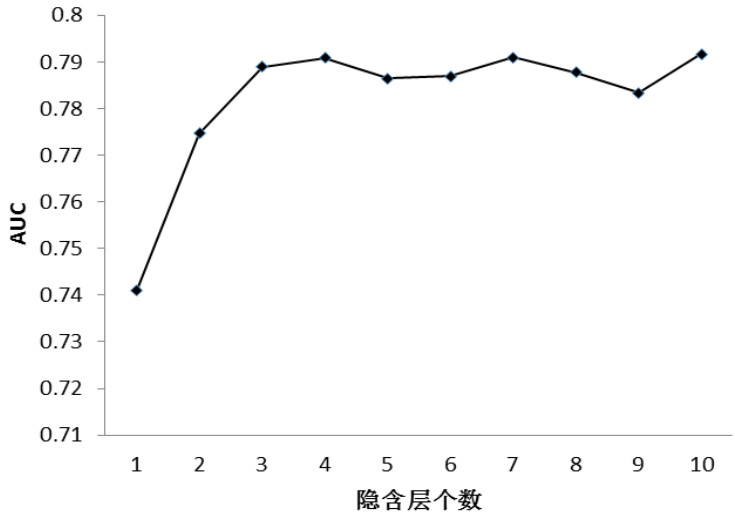

图 7　不同隐含层个数下的实验 AUC 值

4.5.2　神经网络个数的分析

接下来在隐含层个数 4 层以内做实验，对每层的神经元个数进行分析，得到如图 8 实验结果，可以发现，每层的神经元个数并不是越多越好。在网络层数较低时，神经元个数越多，预测结果越好，但是随着层数的增多，神经元个数的影响程度逐渐降低，同样，考虑到神经网络的模型复杂度，这里我们取神经元个数为 50 作为进一步实验的基础。

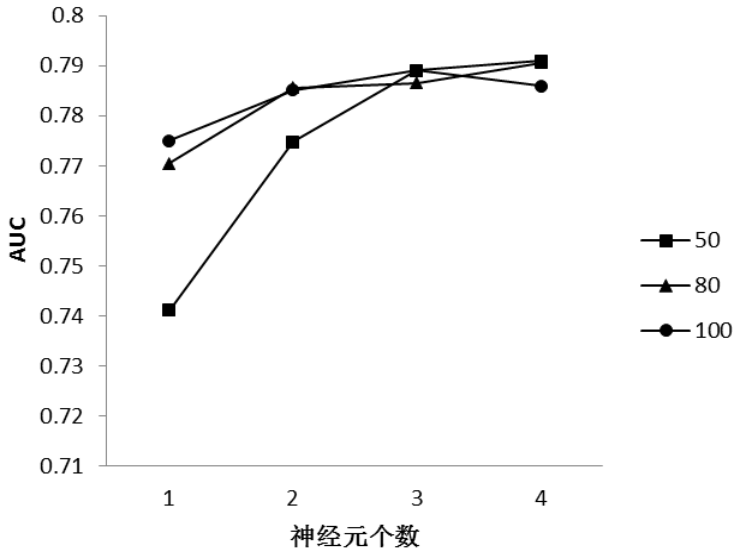

图 8　不同神经元个数下的实验 AUC 值

4.5.3　Dropout 正则化分析

接下来，通过使用 dropout 正则化对模型进行优化，防止过拟合。如图 9 所示，在两种不同神经网络结构的情况下对第一隐含层的 dropout 值进行分析，发现，神经元个数越多，

需要增大 dropout 值来避免过拟合，但是 dropout 也不宜过大，否则会丢弃太多有用信息，模型效果反而降低。本次实验第一层 dropout 最终取值为 0.2。

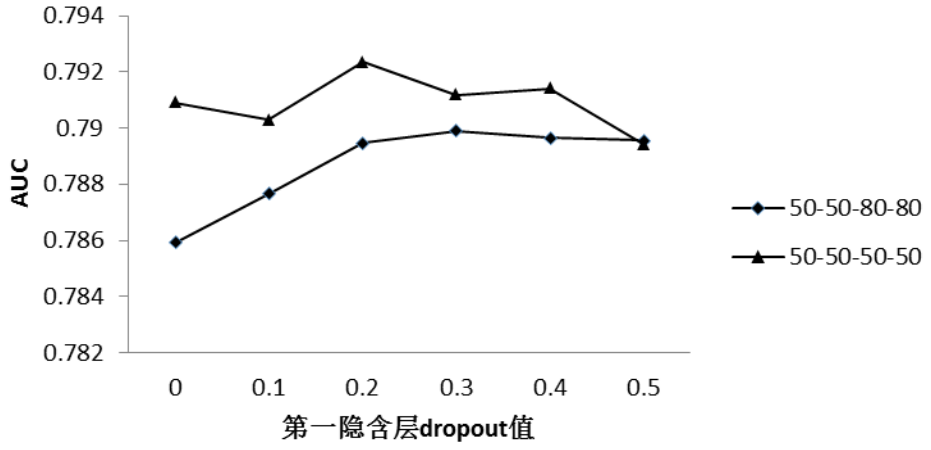

图 9　第一隐含层 dropout 值变化下的实验 AUC 值

然后对第二隐含层的 dropout 正则化的影响进行研究，此时神经网络固定为每层 50 个神经元，隐含层数量为 4 层，通过改变一二层的 dropout 比例，得到如表 8 所示实验结果，可以发现，当第一层的 dropout 值调到比较好的时候，第二层的 dropout 有时可以省去，从而使模型也更简洁。

表 8　第二隐含层 dropout 值变化下的实验 AUC 值

Dropout1	Dropout2	AUC
0.1	0	0.7902
0.1	0.1	0.7903
0.1	0.2	0.7919
0.1	0.3	0.7895
0.2	0	0.7923
0.2	0.1	0.7918
0.2	0.2	0.7910

5 结论

本文研究了用户画像标签在深度神经网络基础上的预测效果，并与传统的浅层学习模型决策树、逻辑回归做实验对比，证明了深度神经网络利用其特征提取优势，能够深入挖掘特征之间的隐含关系，在预测广告用户转化率标签问题中，展现出了其在高维稀疏特征上的预测优势。本文进一步对深度神经网络的层数、神经元数量及 dropout 正则化进行探索，得出以下结论：

（1）在高维稀疏特征基础上进行用户标签预测时，深度神经网络比浅层的逻辑回归模型及决策树更具有优势。

（2）深度神经网络不是神经元个数越多，网络层数越多，越复杂效果越好， dropout 正则化能有效地避免过拟合。

（3）深度神经网络首个隐含层的 dropout 值的确定很重要，能够简化后隐含层数的 dropout 的调整。

未来的研究方向为将深度神经网络作为特征提取器，验证其对浅层学习模型的提升效果。

参考文献

【1】Alan Cooper.交互设计之[M].北京：电子工业出版社，2006

【2】Mladenic, D. Machine learning for better Web browsing. In: Rogers, S., Iba, W., eds. AAAI 2000 Spring Symposium Technical Reports on Adaptive User Interfaces. Menlo Park, CA: AAAI Press, 2000. 82-84.

【3】Chen P M, Kuo F C. An information retrieval system based on a user profile[J]. Journal of Systems & Software, 2000, 54(1):3-8.

【4】Zhou D, Lawless S, Wu X, et al. A study of user profile representation for personalized cross-language information retrieval[J]. Aslib Journal of Information Management, 2016, 68(4):448-477.

【5】曾鸿，吴苏倪. 基于微博的大数据用户画像与精准营销[J]. 现代经济信息, 2016(16):306-308.

【6】熊回香，杨雪萍. 社会化标注系统中的个性化信息推荐研究[J]. 情报学报, 2016, 35(5):549-560.

【7】武慧娟，秦雯，窦平安,等. 社会化标注系统中个性化信息推荐动态模型研究[J]. 情报科学，2016, V34(6):43-46.

【8】Kumar S. A Hybrid Personalized Tag Recommendationsfor Social E-Learning System[J]. International Journal of Control Theory & Applications, 2016, 9(8):1187-1199.

【9】邢玲，宋章浩，马强. 基于混合行为兴趣度的用户兴趣模型[J]. 计算机应用研究, 2016, 33(3):661-664.

【10】廖俊峰，陈天歌，陈旭. 基于客户流失理论的社交媒体用户流向研究[J]. 情报科学，2018, V36(1):45-48.

【11】杨一帆，傅军，朱天博,等. 电力客户用电行为特征挖掘与预测[J]. 电测与仪表, 2016, 53(s1):111-114.

【12】杨晓峰，严建峰，刘晓升,等. 深度随机森林在离网预测中的应用[J]. 计算机科学, 2016, 43(6):208-213.

【13】王靖，王兴伟，赵悦. 基于变精度粗糙集决策树垃圾邮件过滤[J]. 系统仿真学报, 2016, 28(3):705-710.

【14】徐建民，粟武林，吴树芳,等. 基于逻辑回归的微博用户可信度建模[J]. 计算机工程与设计，2015(3):772-777.

【15】Dong L, Wesseloo J, Potvin Y, et al. Discrimination of Mine Seismic Events and Blasts Using the Fisher Classifier, Naive Bayesian Classifier and Logistic Regression[J]. Rock Mechanics & Rock Engineering, 2016, 49(1):183-211.

【16】李翠停. 基于神经网络模型的苹果手机预约数据的分析[D].南京：东南大学, 2016.

【17】陈巧红，孙超红，余仕敏,等. 基于递归神经网络的广告点击率预估研究[J]. 浙江理工大学学报, 2016, 35(6):880-885.

【18】Lecun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553):436.

【19】Dahl G E, Acero A. Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition[J]. IEEE Transactions on Audio Speech & Language Processing, 2011, 20(1):30-42.

【20】胡二雷，冯瑞. 基于深度学习的图像检索系统[J]. 计算机系统应用, 2017, 26(3):8-19.

【21】Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. Computer Science, 2012, 3(4):pages. 212-223.

User Profile Research based on Deep Neural Network

LAN Qiujun, ZHOU Meixuan

(Business School of Hunan University, Changsha / Hunan Province, 410082)

Abstract： User profile plays an important role in personalized recommendation and precision marketing. In the era of big data, due to the lack of mining the relation of the features, the traditional shallow learning methods are faced with difficulties in predicting the tags of user profile especially based on high-dimensional features. In this paper, we expand the primeval basic features and convert them with one-hot encoding, and we predict the tag of user profile by Deep Neural Network (DNN). By the contrast experiments with Decision Tree (DT) and Logistic Regression (LR), our method achieves a higher score at 0.792 in AUC. Further, this paper explores the numbers of layers and neurons of DNN and regularization technology.

Keywords: User Profile; high-dimensional features; tags; Deep Neural Network