数据挖掘大作业报告

社交广告高校算法大赛-相似人群拓展

郑安庆 3120181078　　白雪峰 3120180977
卞西墨 3120180978　　沙　九 3120181023

一、 问题描述：

基于社交关系的广告（即社交广告）已成为互联网广告行业中发展最为迅速的广告种类之一。而复杂的社交场景，多样的广告形态，以及庞大的用户数据，给实现为用户提供精准高效的广告这一目标带来了不小的挑战。本次选题的题目源于腾讯社交广告业务中的一个真实的广告产品——相似人群拓展（Lookalike）。该产品的目的是基于广告主提供的目标人群，从海量的人群中找出和目标人群相似的其他人群。在实际广告业务应用场景中，Lookalike 能基于广告主已有的消费者，找出和已有消费者相似的潜在消费者，以此有效帮助广告主挖掘新客、拓展业务。目前，Lookalike 相似人群拓展产品以广告主提供的第一方数据及广告投放效果数据（即后文提到的种子包人群）为基础，结合腾讯丰富的数据标签能力，透过深度神经网络挖掘，实现了可在线实时为多个广告主同时拓展具有相似特征的高质潜客的能力。

图 1 相似人群拓展问题

相似人群拓展（Lookalike）基于广告主提供的一个种子人群（又称为种子包），自动计算出与之相似的人群（称为扩展人群），如图 1 所示。本题目将为参赛选手提供几百个种子人群、海量候选人群对应的用户特征，以及种子人群对应的广告特征。出于业务数据安全保证的考虑，所有数据均为脱敏处理后的数据。整个数据集分为训练集和测试集。训练集中标定了人群中属于种子包的用户与不属于

种子包的用户（即正负样本）。测试集将检测参赛选手的算法能否准确标定测试集中的用户是否属于相应的种子包。训练集和测试集所对应的种子包完全一致。

二、 数据说明：

2.1 数据来源

本次实验数据来自 2018 年腾讯广告算法大赛-相似人群拓展任务官方提供的数据。数据提供了几百个种子人群、海量候选人群对应的用户特征，以及种子人群对应的广告特征。

2.2 数据说明

数据比赛数据抽取的时间范围是某连续 30 天的数据。总体而言，数据分为：训练集数据文件、测试集数据文件、用户特征文件以及种子包对应的广告特征文件四部分。

训练集数据文件 train.csv 每行代表一个训练样本，各字段之间由逗号分隔，格式为：“aid, uid, label”。其中，aid 唯一标识一个广告，uid 唯一标识一个用户。样本 label 的取值为+1 或-1，其中+1 表示种子用户，-1 表示非种子用户。为简化问题，一个种子包仅对应一个广告 aid，两者为一一对应的关系。

测试集数据文件 test.csv 每行代表一个训练样本，各字段之间由逗号分隔，格式为：“aid, uid”。字段含义同训练集。

用户特征文件 userFeature.data 每行代表一个用户的特征数据，格式为：“uid | features”，uid 和 features 用竖线“|”分隔。其中 feature 采用 vowpal wabbit 格式：“feature_group1 | feature_group2 | feature_group3 | ...”。每个 feature_group 代表一个特征组，多个特征组之间也以竖线“|”分隔。一个特征组若包括多个值则以空格分隔，格式为：“feature_group_name | fea_name1 | fea_name2 | ...”，其中 fea_name 采用数据编号的格式。

广告特征文件 adFeature.csv 格式为：“aid，advertiserId，campaignId，creativeId, creativeSize, adCategoryId, productId, productType”。其中，aid 唯一标识一个广告，其余字段为广告特征，各字段之间由逗号分隔。

三、 模型方法：

3.1 数据预处理

3.1.1 长尾数据

这次比赛的任务是寻找相似人群，其实就可以看成是 CTR 问题。对于 CTR 的数据，都有一个特点就是长尾数据。对于这种类型的数据，进行的操作是对尾部的数据设置成一个新的类别，至于尾部数据的阈值，使用 plot 之后调整。通过这种方式我们构造了一个强特征，有一个 appInstall 字段，这个字段代表的是 60 几天的安装的 app，我们对这个字段进行了统计，统计了每个用户在这几天里 app 安装的数量，然后发现这是长尾数据，也就是大多数是这几天里没有安装 app 的，部分是安装了特别多 app 的，这部分结合实际的话，其实是安装 app 来刷单赚钱的人，对这部分人进行单独的分类处理效果不错，使用这个特征提升了 2 个千分点。

3.1.2 正负样本比例

还有就是正负样本的比例问题，对于此类问题，因模型而异，对于 lgb 模型的话，如果正负样本差的很多的话，可以采取少采一些样本的方式比较多的类型的样本，比如 ctr 中正样本极少，所以我们就抽了 90%的负样本和全量的正样本来训练我们的 lgb 模型，效果也是提升了 2 个千分点。但是对于 FFM 这样的流式训练的模型的话，不建议这么做，这样做会降低模型的效果。

3.1.3 特征处理

在广告点击率和转化率的特征中，特征可以分为三类：数值特征（numberical feature），有序特征 （ordinal feature），无序特征（categorical feature） 对于特征的处理：1.使用统计频率、转化次数特征、转化率特征代替 one-hot，由于数据量极大，one-hot 编码会出现一个很大维数的稀疏矩阵，有一定可能运行好长时间不出结果。2.对训练集和测试集中的重复样本构造是否第一次点击，是否中间点击，是否最后点击，第一次和最后一次间隔特征。3.大量使用组合特征，主要是用户特征和广告上下文特征。

3.2 模型介绍

3.2.1 GBDT

GBDT(梯度提升树)是一种基于迭代所构造的决策树算法，以 boost 为框架的加法模型的集成学习。GBDT 基于 GB 算法。GB 算法的主要思想是，每次建立模型是在之前建立模型损失函数的梯度下降方向。损失函数是评价模型性能(一般为拟合程度+正则项)，认为损失函数越小，性能越好。而让损失函数持续下降，就能使得模型不断调整提升性能，其最好的方法就是使损失函数沿着梯度方向下降。GBDT 再此基础上，基于负梯度(当损失函数为均方误差的时候，可以看作是残差)做学习。

3.2.2 LightGBM

LigthGBM 是 boosting 集合模型中的新进成员，它和 xgboost 一样是对 GBDT 的高效实现，很多方面会比 xgboost 表现的更为优秀。原理上它和 GBDT 及 xgboot 类似，都采用损失函数的负梯度作为当前决策树的残差近似值，去拟合新的决策树，并且支持高效率的并行训练。主要的特点如下：

- 使用直方图简化计算，计算 split 时只考虑直方图的 bin 做划分点，而不细化到每个 sample。
- 使用 leaf-wise 替代 level-wise，每次选择 delta-loss 最大的节点做分割。
- 计算直方图时，两个子节点只用计算其中一个，另一个通过 root 和前一

个做差可得。

- 基于 histogram 的算法，在寻找最佳 split 时，可以先顺序访问 data 的 gradient，填入对应 bin 中，提高 cache hit。
- 对于 category 类型的 feature，可以直接作为特征输入，不需要转化成 one-hot 之类的编码，据说在准确度差不多的情况下速度能快 8 倍以上。

3.2.3 FFM 集成

本题目属于商用推荐场景中的 CTR 预估类型问题，此类问题容易面临大规模稀疏数据的挑战。因此引入因子分解机（ Factorization Machine，简称 FM ）及 FFM（Field-aware Factorization Machine，场感知因子分解机）模型，通过对参数矩阵的低秩分解，来解决高维训练的低效问题。

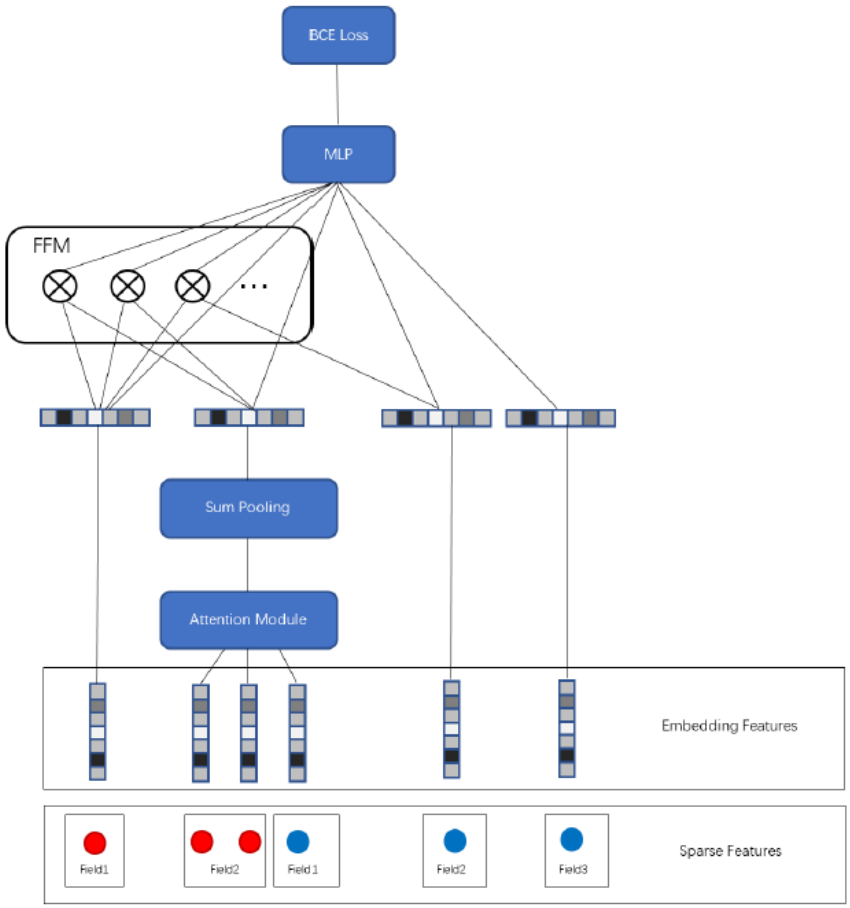

图 2 FFM 模型整体框架

如图 2 所示，首先我们将稀疏的特征作为输入，然后将其映射到一个六维的

空间，将特征转化成一个向量表示。对于一些重要的特征，例如用户的兴趣、关键字和关注的主题等，我们引入了注意力机制（如图 3 所示），使用一个简单的网络对其进行加权求和，然后通过 sigmoid 函数生成一个新的向量表示。之后通过 FFM 将输入的特征的信息加入到学习之中，实现场感知（field-aware），进一步提升了特征组合的表达能力。

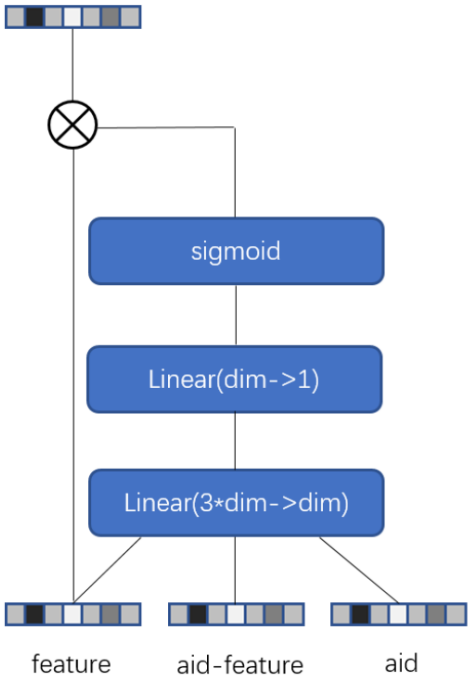

图 3 注意力机制模型

3.3 项目评估

对于扩展后的相似用户，如果在广告投放上有相关的效果行为（点击或者转化），则认为是正例；如果不产生效果行为，则认为是负例。每个待评估的种子包会提供如下信息：种子包对应的广告 aid 及其特征，以及对应的候选用户集合（uid 及其特征）。选手需要为每个种子包计算测试集中用户的得分，比赛会据此计算每个种子包的 AUC 指标，AUC_i 表示第 i 个包的 AUC 值，并以所有待评估的 m 个种子包的平均 AUC 作为最终的评估指标。

AUC（Area under the Curve of ROC）是 ROC 曲线下方的面积，是判断二分类预测模型优劣的标准。ROC（receiver operating characteristic curve）接收者操作特征曲线，是由二战中的电子工程师和雷达工程师发明用来侦测战场上敌军载

具（飞机、船舰）的指标，属于信号检测理论。ROC 曲线的横坐标是伪阳性率（也叫假正类率，False Positive Rate），纵坐标是真阳性率（真正类率，True Positive Rate），相应的还有真阴性率（真负类率，True Negative Rate）和伪阴性率（假负类率，False Negative Rate）。这四类的计算方法如下：

- 伪阳性率（FPR）

  判定为正例却不是真正例的概率。

- 真阳性率（TPR）

  判定为正例也是真正例的概率。

- 伪阴性率（FNR）

  判定为负例却不是真负例的概率。

- 真阴性率（TNR）

  判定为负例也是真负例的概率。

3.4 实验结果及分析

图 3 为 LightGBM 模型特征重要性的输出结果：

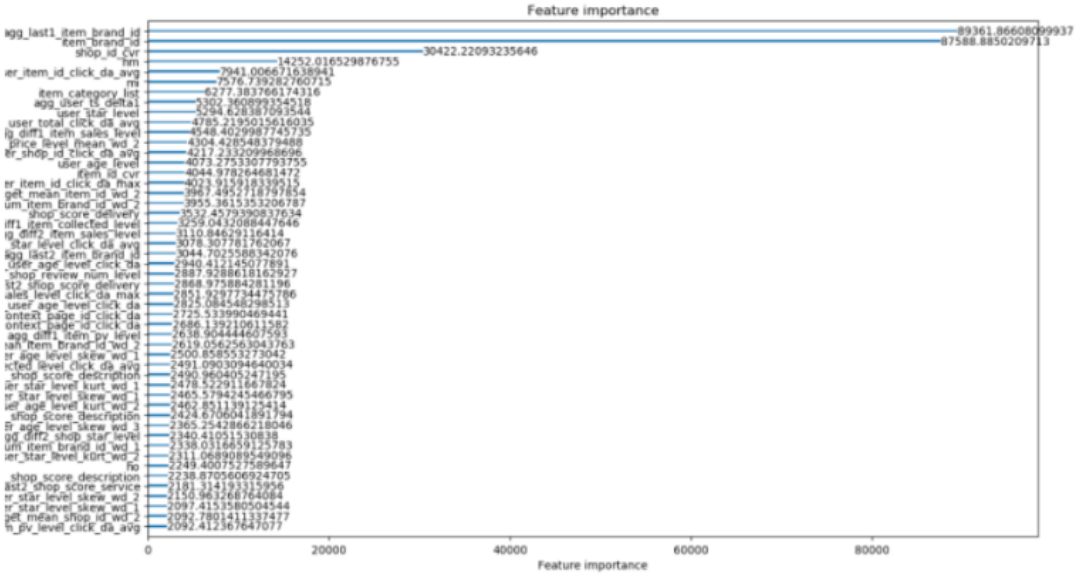

图 3 LightGBM 模型结果

从 LightGBM 输出结果来看，procurtType、productId 和 creativeId 这三个特征显得较为重要，这与我们在阶段报告中数据分析结果较为相似，如图 4 所示。

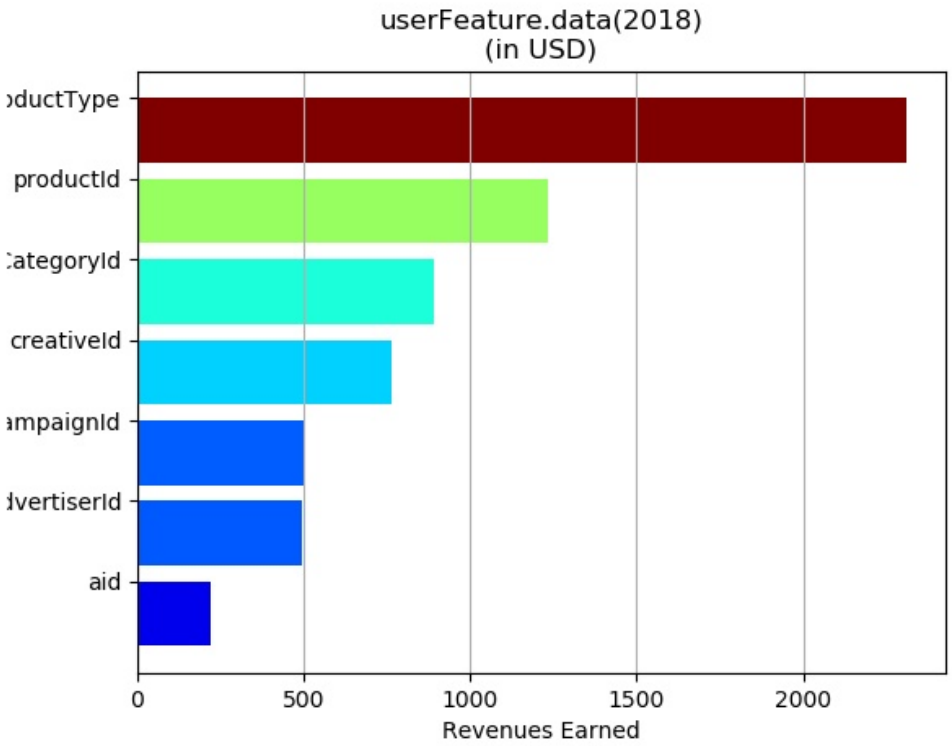

图 4 阶段报告中的数据分析结果

对于 FFM 模型，我们将预处理时构造的统计特征，比例特征和转化率特征作为输入。和以往不同的是，构造这样特征时不仅考虑单个特征的统计度量，还考虑了所有可能的组合特征。也因此发现了很多不易想到的强特，如 uid 相关特征，uid 点击次数，uid 转化率，将这些特征作为注意力机制的输入。之后将数据分块，一块作为验证集进行调参，其余分块做 dropout 交叉统计，测试集则用全部训练集数据进行统计，最终，在测试集上的准确率稳定在了 0.5 左右，准确率变化如图 4 所示。

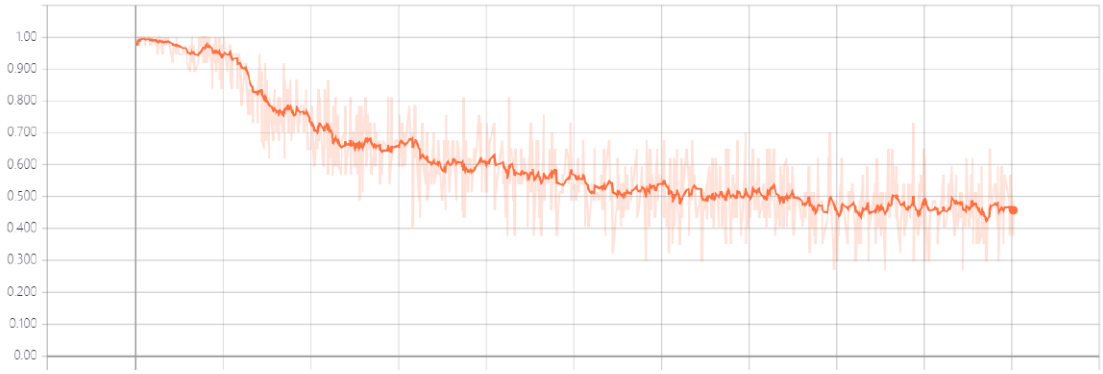

图 4 在测试集上的准确率