2019 智源-看山杯 专家发现算法大赛

王雪
学号：201944768
wangxue574@126.com

摘要

知识分享服务已经成为目前全球互联网最重要、最受欢迎的应用类型之一。在知识分享或问答社区中，问题数远远超过有质量的回复数。因此，如何连接知识、专家和用户，增加专家的回答意愿，成为了此类服务的中心课题。知乎自 2011 年创办至今，已经成为一个拥有 2.2 亿用户，每天有数以十万计的新问题以及用户原创内容产生的网站。其中，如何高效的将这些用户新提出的问题邀请其他用户进行解答，以及挖掘用户有能力且感兴趣的问题进行邀请下发，优化邀请回答的准确率，提高问题解答率以及回答生产数，成为知乎最重要的课题之一。在本次比赛中针对这一问题主要利用 LightGBM 算法和 K 折交叉验证进行处理，来计算将问题 Q 推荐给用户 U 之后，用户 U 回答问题 Q 的概率。

关键字

数据挖掘　推荐系统　LightGBM　K 折交叉验证

简介

知乎目前已拥有超过 2.2 亿用户，每天产生海量的提问，传统的手动邀请他人回答问题功能已经不能满足用户的需求。故知乎专家推荐系统即问题路由应运而生。问题路由推荐系统每日可以对 10 万+的问题进行分发，并保证问题提问后 3 日内的解答率达到 70% 以上；系统对千万级的创作群体进行精准推荐，经由系统智能分发推荐下每日产生的回答数超过 20 万。问题路由同时也是本次看山杯的题目来源，比赛旨在从选手中征集高效精准的推荐算法，挖掘有能力且感兴趣的用户进行问题地精准推荐。

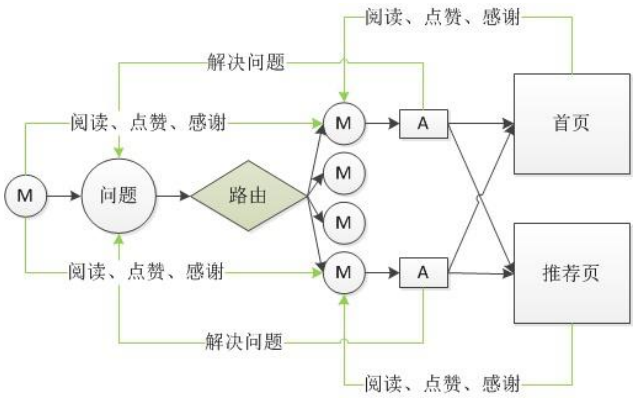

图 1: 知乎问题路由工作机制

以创作者推荐问题为例，问题路由主要包括排序、召回、特征落地三个流程，系统首先根据用户 ID 向问题路由发起请求，通过用户画像获取用户特征（年龄、兴趣）；然后使用召回模块召回用户潜在能够回答的问题，从数百万的问题中找到几百个问题供排序模块使用；接下来通过特征模块获取问题的相关特征，据此对问题进行精准排序，并将头部问题返回至用户，排序中产生落地的特征会记录在日志中和用于训练模型。本次比赛所解决的问题主要与排序流程相关，而在排序过程中找到用户 U 回答问题 Q 的概率将至关重要。本次比赛先对比赛中提供的数据集进行了分析和处理，然后对处理后的数据进行特征处理，并通过对比选择了适合进行本次预测的 LightGBM 模型进行模型训练与测试，最后由 AUC 来对模型进行评估。

1. 数据集介绍

本次比赛一共提供了 9 个数据集，他们分别是:

1. single_word_vectors_64d.txt

每一行代表一个单字及其 64 维 embedding 的表示，每一行有 65 列，第一列以 SWxxx 表示单字编

码序号， 其后是 64 个浮点数，代表 64 维 embedding 向量，列之间采用 /tab 分隔符分隔。这个 embedding 不是用比赛数据集训练出来的，是用知乎全站的数据集训练出来的。

2. word_vectors_64d.txt

每一行代表一个词及其 64 维 embedding 的表示，每一行有 65 列，第一列以 Wxxx 表示词编码序号， 其后是 64 个浮点数，代表 64 维 embedding 向量，列之间采用 /tab 分隔符分隔。

3. topic_vectors_64d.txt

每一行代表一个话题及其 64 维 embedding 的表示，每一行有 65 列，第一列以 Txxx 表示话题 ID 编码序号，其后是 64 个浮点数，代表 64 维 embedding 向量， 列之间采用 /tab 分隔符分隔。

4. question_info_0926.txt

包含邀请数据集 (数据集 7 和 8) 及回答数据集 (数据集 5) 表中涉及到的所有问题列表，每一行代表一个问题的相关信息， 每一行有 7 列， 列之间采用 /tab 分隔符分割。

问题数据集由问题 ID、问题创建时间、问题标题的单字编码序列、问题标题的切词编码序列、问题描述的单字编码序列 、问题描述的词编码序列、问题绑定的话题 ID 等特征。

5. answer_info_0926.txt

为邀请数据集 (数据集 7 和 8) 中用户最近 2 个月内的所有回答，每一行代表一个回答的相关信息， 每一行有 20 列， 列之间采用 /tab 分隔符分割。

回答数据集包含用户近两个月以来对问题的回答情况，主要包括回答创建时间、回答内容的单字编码序列、回答内容的切词编码序列、回答是否被标优、回答是否被推荐、回答是否被收入圆桌、是否包含图片、是否包含视频、回答字数、点赞数、取赞数、评论数、收藏数、感谢数、举报数、没有帮助数、反对数。

6. member_info_0926.txt

用户数据集包含邀请数据集 (数据集 7 和 8) 中用户相关特征信息，每一行代表一个用户的相关信息， 每一行有 21 列， 列之间采用 /tab 分隔符分割。

用户数据集由用户 ID、性别、创作关键词的编码序列、创作数量等级、创作热度等级、注册类型、注册平台、访问频率（有五种取值 [new | daily | weekly | monthly | unknow] ， 分别对应为 [新用户 | 日活跃用户 | 周活跃用户 | 月活跃用户 | 未知]）、用户二分类特征 A、用户二分类特征 B、用户二分类特征 C、用户二分类特征 D、用户二分类特征 E、用户分类特征 A、用户分类特征 B、用户分类特征 C、用户分类特征 D、用户分类特征 E、用户的盐值分数、用户关注的话题、用户感兴趣的话题等 21 个特征组成。

7. invite_info_0926.txt

用户邀请数据集包含用户最近 1 个月的邀请数据，每一行代表一个问题邀请的相关信息， 每一行有 4 列， 列之间采用 ／tab 分隔符分割。

用户邀请数据集包括邀请的问题 ID、被邀请用户 ID、邀请创建时间、邀请是否被回答等特征。

8. invite_info_evaluate_0926.txt

邀请数据测试集包含未来 7 天的问题邀请数据，每一行代表一个问题邀请相关信息，每一行有 3 列， 列之间采用 ／tab 分隔符分割。这三列分别为邀请的问题 ID、被邀请用户 ID、邀请创建时间。

9. invite_info_evaluate_2_0926.txt

比赛的最终验证集，与数据集 8 格式相同。

2. 特征工程

特征工程指的是把原始数据转变为模型的训练数据的过程，它的目的就是获取更好的训练数据特征，使得机器学习模型逼近数据和特征决定的机器学习的上限。特征工程能使得模型的性能得到提升。

2.1. 数据预处理

在工程实践中，我们得到的数据会存在有缺失值、重复值等，在使用之前需要进行数据预处理。数据预处理没有标准的流程，通常针对不同的任务和数据集属性的不同而不同。数据预处理的常用流

程为：去除唯一属性、处理缺失值、属性编码、数据标准化正则化、特征选择、主成分分析。

由于本次得到的数据是知乎平台已经处理过的数据集，因此不存在“脏数据”，无需进行数据清洗，我们只需去除唯一属性，进行特征编码等处理即可。

2.2. 数据分析

通过对用户数据集 member_info_0926.txt 的每个特征进行统计，我们可知用户数据集共 21 个特征，其中 5 个特征（创作关键词、创作数量等级、创作热度等级、注册类型、注册平台）在数据集中只有一个取值，说明这 5 个特征是完全不具备区分作用的单值特征，可以直接去掉。

将问题数据集 question_info_0926.txt 与用户数据集 member_info_0926.txt 合并后，利用直方图分析特征对结果的影响。发现性别、访问频率、所有的用户二分类特征等特征具有较强的区分作用。其中我们将数值不同的盐值进行分桶，发现不同区间的盐值对结果也将产生不同的影响。

为了保护用户隐私，所有问题、回答以及邀请的时间均进行了相应的偏移，偏移后的时间给出到小时的精度， 时间格式为“D×-H×”，其中 D 代表天数，H 代表小时，表示的是第 X 天的第 X 个小时。日期格式为 D1-H3， 含义为 day 1 的 3 点。在后续的数据处理阶段我们同样要将时间格式进行转换，可转换成天和时间两个特征。

3. 数据处理

首先，在数据分析阶段我们分析出的用户数据集中创作关键词的编码序列、创作数量等级、创作热度等级、注册类型、注册平台等五个特征是不具备区分作用的单值特征，因此要将它们进行删除。

然后，由于在模型中我们并没用到问题数据集中的问题标题单字编码，问题标题切词编码，问题描述单字编码，问题描述切词编码，问题绑定话题，关注话题，感兴趣话题，问题创建时间，邀请创建时间等特征，因此数据处理阶段同样将它们进行删除。

训练集 invite_info_0926.txt 是在知乎平台中选取的连续一个月的邀请回答数据，我们将处理后的用户和问题特征添加到训练集 invite_info_0926.txt 中，以便进行下一步的处理。

测试集 invite_info_evaluate_2_0926.txt 选取的则是训练集后连续一个星期的数据，同样的，将处理后的用户和问题特征添加到测试集 invite_info_evaluate_2_0926.txt 中。

在回答数据集 answer_info_0926.txt 中删除没有用到的回答问题的单字、切词编码序列。随后在回答数据集中结合问题数据集 question_info_0926.txt，计算出回答距提问的天数，并划分时间窗口。

通过查看邀请数据集可知时间跨度为 3838-3867，回答数据集中的时间跨度为 3807-3867，测试邀请数据集中时间跨度为 3868-3874。由于我们要预测未来七天的邀请是否被回答，所以构造时间跨度为 3861-3867 的训练集，结合时间跨度为 3838-3860 的邀请数据和时间跨度为 3810-3860 的回答数据，主要思路是通过让模型学习邀请发出之前问题被回答的情况，预测未来七天邀请被响应的情况，那么模型学习到的参数同样适用下一个七天的邀请，也就是测试集发出的邀请。但是用于预测测试集的特征的时间跨度也需要顺延七天，也就是说测试集需要结合时间跨度为 3845-3867 的邀请数据和时间跨度为 3817-3867 的回答数据。

最后，将所有数据集的时间格式“D×-H×”进行统一处理，提取出 D 后数据 x 作为 day 特征，H 后 X 作为 hour 特征添加到数据集中，并删除原本的时间特征“D×-H×”。

4. 特征处理

对用户 id，问题 id，性别，访问频度，用户多分类特征 a，用户多分类特征 b，用户多分类特征 c，用户多分类特征 d，用户多分类特征 e 等离散型的二分类特征值和多分类特征值用 Label Encoder 进行数字编码，将其类别特征值转换为表示类别的整型数字，方便与下一步模型的训练。

对具有很好区分度的用户 id，问题 id，性别，访问评率，用户二分类特征 a，用户二分类特征 b，用户二分类特征 c，用户二分类特征 d，用户二分类特征 e，用户多分类特征 a，用户多分类特征 b，用户多分类特征 c，用户多分类特征 d，用户多分类特征 e 等特征，进行单特征计数。

对特征进行分组，分组后计算这些特征的 mean、sum、std、count。

5. 模型选择

本次实验选择的模型是 LightGBM，它是基于 Boosting 框架的主流集成算法，和 XGBoost 一样是对 GBDT 的高效实现，LightGBM 和 XGBoost 在各数据挖掘比赛中都具有突出表现。但相对来说 LightGBM 具有更突出的表现。

算法	accuracy score	auc score	执行时间（S）
LightGBM	0.861501	0.764492	0.283759
XGBoost	0.861398	0.764284	2.047220

表 1：LightGBM 和 XGBoost 的性能比较

从上述的性能对比结果来看，LightGBM 对比 XGBoost 的准确率和 AUC 值都有所提升。并且，一个至关重要的差别是模型训练过程的执行时间。LightGBM 的训练速度几乎比 XGBoost 快 7 倍，并且随着训练数据量的增大差别会越来越明显。这证明了 LightGBM 在大数据集上训练的巨大的优势，尤其是在具有时间限制的对比中。

除此以外，LightGBM 还具有低内存使用、更高的准确率、支持并行化学习、可处理大规模数据、支持直接使用 category 特征等突出优势，以及基于 Histogram 的决策树算法、带深度限制的 Leaf-wise 的叶子生长策略、直方图做差加速、直接支持类别特征(Categorical Feature)、Cache 命中率优化、基于直方图的稀疏特征优化、多线程优化等特点。因此在这里选择 LightGBM 来训练模型。

6. 模型训练与测试

模型训练阶段选取 LightGBM 模型，并结合 K 折交叉验证进行。K 折就是将数据集 D 划分为 K 个大小相似的互斥子集，每次将其中一个子集作为测试集 test，剩下 k-1 个子集作为训练集 train 进行训练，循环交替进行 k 次训练和测试，最终返回的是 K 个测试结果的均值。

在机器学习中，样本量不充足时，通常使用交叉训练验证。在将原始数据集划分为 K 部分的过程中，有很多不同的采样方法，在这里我们使用针对非平衡数据的分层采样 StratifiedKFold。

StratifiedKFold 分层采样交叉切分，确保训练集，测试集中各类别样本的比例与原始数据集中相同。

在实验中我们选取 K=5，进行 5 折交叉验证。以及 StratifiedKFold 分层采样。

7. 评估指标

AUC（Area under curve）是 ROC 曲线下面积，指随机给定一个正样本和一个负样本，分类器输出该正样本为正样本的概率值比分类器输出该负样本为正的那个概率值要大的可能性。用来评估分类器的性能。

$$AUC = \frac{\sum_{ie\ positiveClass} rank_i - \frac{M(1+M)}{2}}{M \times N}$$

AUC 计算方式如上式所示，M 为正样本数，N 为负样本数，$rank_i$ 为 第 i 个正样本所在的位置。

Log loss 是在机器学习构建分类模型的任务中经常使用的损失度量方法，公式为：

$$-\sum_{i}^{N}\sum_{j}^{M} y_{ij} log(p_{ij})$$

其中 N 对应于我们的样本数或者输入的实例的数量，i 对应于某一个样本或者实例；M 表示我们的样本可能的分类数量，j 表示某一个分类；yij 表示对于某个样本 i，其属于分类 j 的标签，通常是 0 或者 1，i 只属于一个分类；pij 表示的是样本 i 预测为 j 分类的概率。

本次比赛最终仅采用 AUC 进行评估，但在实际模型训练过程中我们涉及到 AUC 和 Log loss 两种评估方式，因此在这里全部列出。

总结

在本次比赛中，首先对比赛中提供的数据集进行了分析和处理，然后对处理后的数据进行特征处理，并通过对比选择了适合进行本次预测的 LightGBM 模型进行模型训练与测试，最后平台通过 AUC 来对我们训练的模型进行评估。在比赛最终的验证集上，本模型取得了 0.81019 的最终成绩。

这次比赛让我动手去实践数据分析，了解数据挖掘的过程，并且此次是在我们日常生活中经常接触的知乎上进行的数据分析过程，让我们了解知乎的推荐系统的工作流程，开拓了视野，同时也体会到了数据挖掘的乐趣，认识到自身知识和能力的不足，收获颇丰。

致谢

首先应该感谢尹建华老师的数据挖掘课程，虽然由于本人基础薄弱，知识储备不足，很多地方一知半解，但是依然要感谢尹建华老师的辛勤付出和生动讲解。

其次我要感谢实验室同学们的无私帮助与指导，尤其感谢王健同学提供的思路和不厌其烦的讲解。

参考资料

[1] 周志华.机器学习[M].北京.清华大学出版社
[2] LightGBM 介绍及参数调优 [EB/OL].https://www.cnblogs.com/jiangxinyang/p/9337094.html
[3] 机器学习算法之 LightGBM https://www.biaodianfu.com/lightgbm.html
[4] 详解知乎问题路由推荐系统 [EB/OL].https://zhuanlan.zhihu.com/p/93030223
[5] 智源 - 看山杯专家发现算法大赛 2019 数据及介绍 [EB/OL].https://biendata.com/competition/zhihu2019/data/