北京化工大学

生产实习课程设计报告

课程设计题目：基于深度学习的编程竞赛成绩预测算法研究

组 长：孟铃翔

组 员：黄帅旺

2020年7月

摘 要

近年来，机器学习和深度学习作为人工智能的分支，发展十分火热。为了了解一般机器学习、深度学习项目的详细流程，学习具体的模型知识，并利用这些知识对蓝桥杯成绩进行预测，以帮助学校集训队更好地了解学生训练成果，更加科学地、有针对性地安排训练，我们选择了基于深度学习的编程竞赛成绩预测算法研究作为课题。

在课程设计中，我们进行了数据爬取、模型分析、数据可视化这三项工作。首先，我们利用爬虫技术，采用python库中的BeautifulSoup对BUCTOJ和acmer.site网站的数据进行爬取，结合已有的往年获奖记录数据，构造和设计了用于机器学习和深度学习的模型向量。接下来，我们使用了传统机器学习模型和深度学习模型，包括AdaBoost、决策树、GDBT、KNN、逻辑回归、SVM和全连接神经网络模型，对构建好的数据进行拟合，同时得到对即将来临的蓝桥杯获奖结果的预测。最后，我们根据已有数据设计了一些研究问题。利用数据可视化的方法，结合Matplotlib和Pyecharts等python第三方库，对这些问题进行了探究。

经过我们的实验，根据对往年获奖数据的分析，我们发现OJ上面问题的提交数以及准确率与最后的获奖结果无线性相关关系，但是CodeForces积分能够很好地反映最终的得奖结果。CodeForces积分更高的同学往往有着更高的得奖概率，这一结论也符合我们的常识和预期。其次，对比不同模型得到的准确率，在其他条件相同的情况下，我们发现特征向量的维度对机器学习模型的准确率有着一定的影响。就目前结果来看，降低特征向量的维度能够显著地提高KNN模型的准确率，同时能够明显地降低AdaBoost模型的准确率。但是由于本次实验数据集的规模较小，数据降维对于机器学习模型在大规模数据集上拟合预测得到结果的准确率的影响还需要我们深入探究。另外，根据对神经网络梯度下降过程的可视化，我们也发现对于特征向量数值域特征的归一化，能够加速我们对于神经网络模型的训练过程，同时提高神经网络预测结果的准确率。这一结论之前在深度学习领域被证明过，这次实验结果也再次证明了这一理论的正确性。

最后，我们相信学校集训队也可以参考我们的实验结果，更加科学、系统、有针对性地进行ACM集训队的建设，激发大家的编程兴趣和专业热情，推动我们学校不断进步，一步步取得更好的成绩。

1. 项目背景及意义
   1. 项目背景及研究现状

机器学习是实现人工智能的一个途径，即以机器学习为手段解决人工智能中的问题。机器学习在近30多年已发展为一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、计算复杂性理论等多门学科。机器学习算法是一类从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法。因为学习算法中涉及了大量的统计学理论，机器学习与推断统计学联系尤为密切，也被称为统计学习理论。

深度学习是机器学习中一种基于对数据进行表征学习的算法。深度学习的好处是用非监督式或半监督式的特征学习和分层特征提取高效算法来替代手工获取特征。

目前，机器学习和深度学习应用广泛，已广泛应用于数据挖掘、计算机视觉、自然语言处理等领域。显而易见，机器学习和深度学习也可以应用我们的竞赛成绩预测中。

目前在我们学校的竞赛以及集训队建设中，人工数据分析的结果、老师之前的经验仍是我们安排训练和比赛建设的主要参考。在如今大数据和互联网时代的背景下，这一方法不仅数据利用率低，同时也占用了大量的人力和实践成本。结合新的技术对这一方法进行改进，是我们尽早也是必须要做的。通过利用机器学习的统计和深度学习中表征学习的概念，对我们得到的大量数据进行分析和可视化探究，能够帮助我们入门机器学习和深度学习，加深我们对于机器学习和深度学习的理解，也能提高我们收集到的数据的利用率，为我们进行集训队建设提供更加科学的指导，进而提高训练质量，达到事半功倍的效果。

* 1. 本课题主要研究内容
* 数据获取

利用爬虫技术，结合BeautifulSoup, Pandas等第三方Python库，对BUCTOJ和acmer.site网站数据进行爬取，得到实验需要的正样本、负样本、要预测的数据所对应的特征向量。

* 模型预测

在得到模型向量后，结合Scikit-Learn和百度PaddlePaddle框架，使用机器学习和深度学习模型对构建好的数据进行训练测试和预测。在得到多个模型预测的结果后，通过投票的方式，得到最终的预测结果。

* 数据可视化

在进行实验前，我们根据研究的任务和背景，设计了需要探究的问题。在实验过程中，结合Matplotlib和Pyecharts等第三方库，我们通过对于实验数据的可视化，对设计的问题进行了探究，得到了实验结论。这些问题得到的结论对未来集训队的建设提供了一些指导，有很大参考价值。

1. 项目实施方法设计
   1. 方法框架

* 特征工程：设计和构造数据，得到特征向量
* 模型预测：将得到的特征向量放入不同的机器学习和深度学习模型进行预测，得到预测结果
* 数据可视化：通过数据可视化，对研究问题进行研究
  1. 特征工程
* 获取数据

1. 正样本：根据已有的历史获奖数据，结合Pandas库，获取到基础正样本。
2. 负样本：结合爬虫技术，使用爬虫工具BeautifulSoup对BUCTOJ和acmer.site数据进行爬取，获取到了用户数据。去掉正样本和预测样本，去掉用户不为学号的用户，去掉无提交记录的用户后，得到基础负样本。

* 构造特征

1. 历史各级奖项的次数：通过已有数据统计该用户历史的各个奖项的获奖次数。包括国家级一等奖获奖次数、国家级二等奖获奖次数等等。
2. 编程年份：根据已有数据，正样本的编程年份为获奖年份减去入学年份。负样本的编程年份计算较为复杂。首先通过爬虫，对负样本的BUCTOJ提交记录进行爬取，得到最后一次提交时间。计算出的编程年份为最后提交时间减去入学年份。
3. 编程语言：利用历史获奖数据，根据不同用户参加比赛的报名年份统计编程语言。编程语言主要用于获取用户的OJ提交记录，负样本编程语言默认为C++。
4. OJ提交数、OJ正确数、OJ准确率：结合爬虫技术对BUCTOJ数据进行爬取，根据用户使用的编程语言对OJ上正样本、负样本的提交数和正确数进行统计，同时计算出用户在OJ上做题的准确率。
5. CF Rating、CF Times、all\_cf\_aftersolve、correct\_cf\_aftersolve：结合爬虫技术，根据acmer.site提供的api接口，统计与用户相关的CF训练数据。
6. 其他训练数据：使用爬虫获取包括计蒜客、牛客网、AtCoder等其他训练数据，根据往年数据获取用户的性别、专业特征。

* 处理和加工数据

1. 编程语言：编程语言只是用来统计相应OJ数据，统计完之后可直接丢弃。
2. OJ相关数据与CF相关数据：
   * 缺失值：一般机器学习更倾向于使用平均数、中位数等统计量对缺失值进行填充。本实验中由于数据规模小，且带有缺失值特征较多，经过小组讨论后，决定用相应获奖类别最小值与最大值之间的随机数对缺失的进行合理填充。
   * 值过小：值过小的情况有两种。一是用户报名时使用的编程语言提交记录较少，在这种情况下用所有语言的提交数和正确数代替特定语言提交数和正确数。另一种情况为用户刷题量较小，在该种情况下用剩余值的中位数代替值过小的特征。
3. 其他数据：

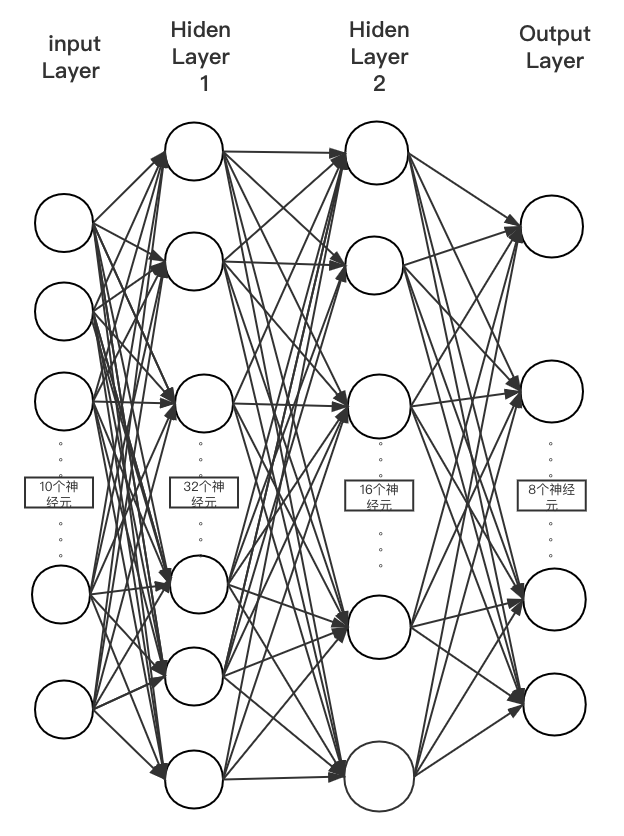
开始时设计的特征向量包括性别、专业特征。但是由于往年数据大部分获奖者性别为男性，专业为计算机，故认为这些特征对于整体模型的准确率无较大影响。剩余像计蒜客、牛客网、AtCoder的训练数据，由于缺失值过多，故选择去掉这些特征，只保留CodeForces有关的训练数据。

* 1. 模型预测

结合Scikit-Learn和百度PaddlePaddle框架使用以下机器学习和深度学习模型对特征进行测试、参数调优、预测得到预测结果：

* AdaBoost
* DesicionTree（决策树）
* GDBT（梯度提升树）
* KNN（最近邻模型）
* LogisticRegression（逻辑回归）
* SVM（支持向量机）
* FCN（全连接神经网络）

其中神经网络结构图如下图所示：



我们使用的网络为一个4层的神经网络。其中两个隐藏层分别有32个神经元和16个神经元，使用ReLu作为激活函数。输出层的8个神经元分别代表得到8种奖项对应的概率，使用Sigmoid作为激活函数。神经网络使用的优化器为RMSProp优化器。

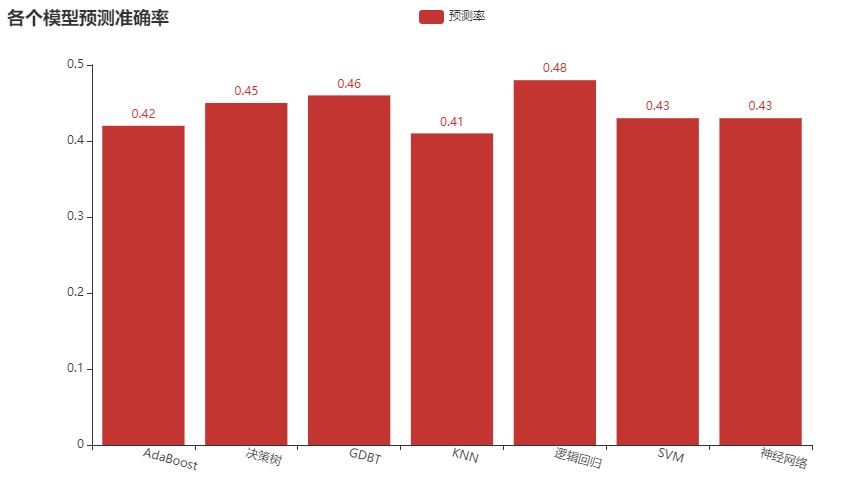
使用这样神经网络的原因有以下几个。首先我们使用的是一个全连接神经网络，因为这样的网络在对表格数据分类时，相比其他结构的神经网络具有更高的准确率。其次我们使用网络的层数较少，这是因为我们数据集的规模较小。如果网络层数较深的话，那么会有许多参数。在小规模数据集上这些参数得不到足够的训练，导致模型在占用大量计算资源的情况下，得到较低的预测准确率，资源利用率较低。因此我们选择这样的神经网络作为我们的深度学习模型。

除了神经网络，我们也在设计实验的其他部分时参考了实际的经验。

在进行预测时，考虑到以往的现实情况，大部分参与比赛的同学都拿到了奖励，因此我们使用了两种策略对结果进行预测。分别是：使用负样本进行训练和不使用负样本进行预测。

在使用负样本进行预测的情况中，由于负样本数据规模是正样本的5倍，为了保证模型的数据均衡，每次随机从负样本中抽取与正样本相同数目的数据，对于同种模型训练5次。选取准确率最高的模型代表该种模型预测结果。最后再利用不同种模型投票，得到最终的预测结果。

不使用负样本进行预测也使用类似方法，为了保证准确性。每种模型也训练5次，选取准确率最高的模型，最终投票决定。



* 1. 数据可视化

利用不使用负样本的数据，对以下问题进行探究并进行可视化：

* 获得各个不同类型奖项同学的各个特征数值
* 数据降维对不同模型准确率的影响
* 数据归一化对于深度学习模型的影响
  1. 得到最终预测结果

根据不同数据处理方法的模型准确率，结合上文提到的模型预测方法，选择在模型上准确最高的数据处理，进行多模型预测。具体方法如下：

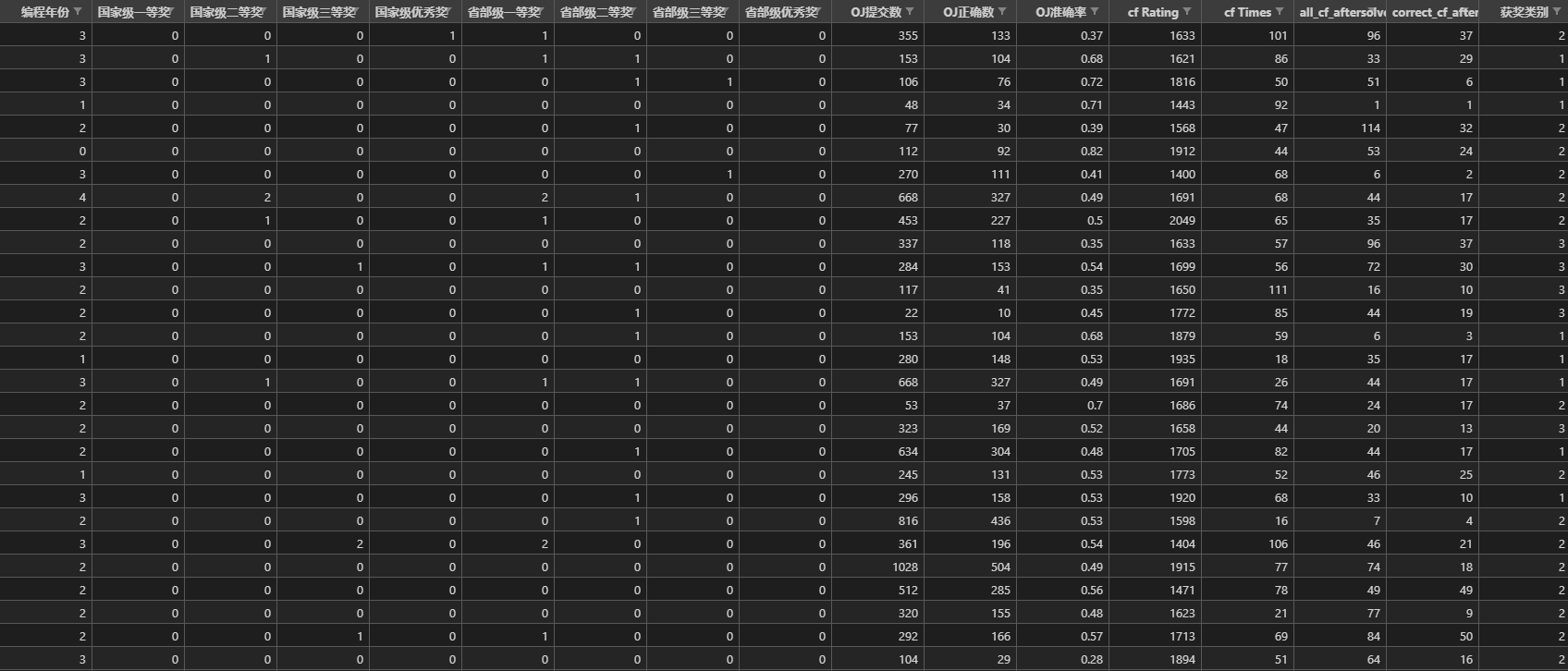
* 传统机器学习模型：KNN模型使用降维后准确率最高的模型进行预测，AdaBoost模型使用降维前准确率最高的模型进行预测。其余模型由于降维前后准确率无较大差异，故使用准确率最高的模型进行预测，降低参数复杂性。
* 深度学习模型：对全连接神经网络，使用归一化后的数据进行训练，获得学生获得每种奖励的预测概率。最后选择预测概率最高的奖励类别作为我们最终的预测结果。

1. 实验设计与结果分析

3.1 实验设计

3.1.1 实验数据介绍

实验数据分为三类，包括正样本（往年获奖数据）、参赛样本（今年的报名数据）、负样本（历年除获奖用户和今年参赛用户之外的未获奖用户。各个样本包含的特征有编程年份、之前获得的各级别奖项个数、OJ提交数、OJ正确数、OJ准确率、cf积分、cf比赛次数、cf补题数、cf正确补题数。部分实验数据如下：



3.1.2 研究问题

* RQ1：获得各奖项同学的各个特征与获奖类别之间的关系
* RQ2：数据降维对于不同机器学习模型准确率的影响
* RQ3：数据归一化对于训练深度神经网络模型梯度下降收敛速度的影响

3.1.3 评价指标

RQ1：

根据图中各个特征的相关性对结果进行评价。通常可以通过各模型对于最终结果影响的趋势看出特征是否与结果相关。

RQ2：

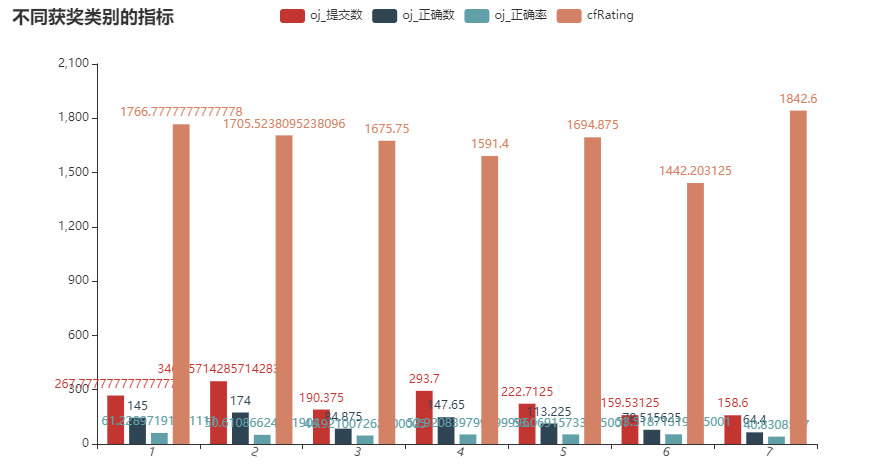
在其他条件相同的情况下，根据降维前后模型准确率的比较来判断对数据进行降维是否对同一机器学习的准确率有影响

RQ3：

通过梯度下降训练相同次数比较loss或者达到同样loss需要的训练次数来判断数据归一化对于梯度下降训练速度的影响。同时可通过折线图十分直观地看出归一化对梯度下降训练地影响程度。

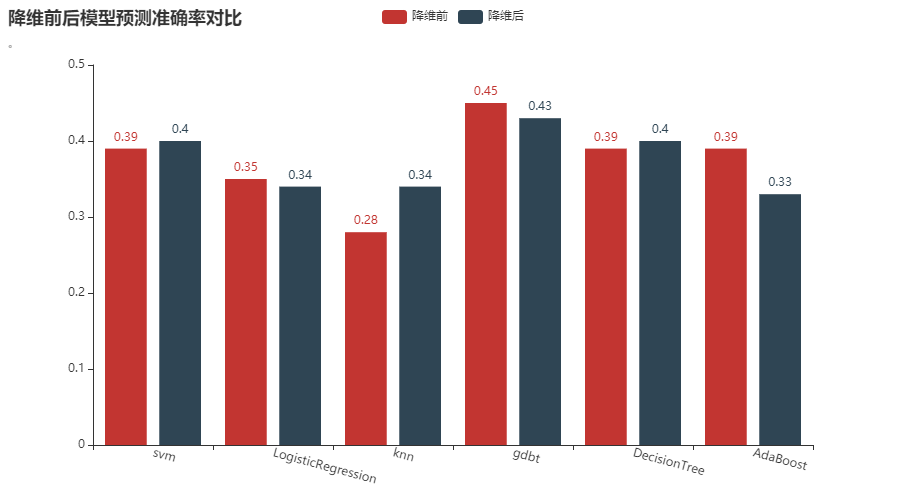
3.2 实验结果分析

3.2.1 RQ1



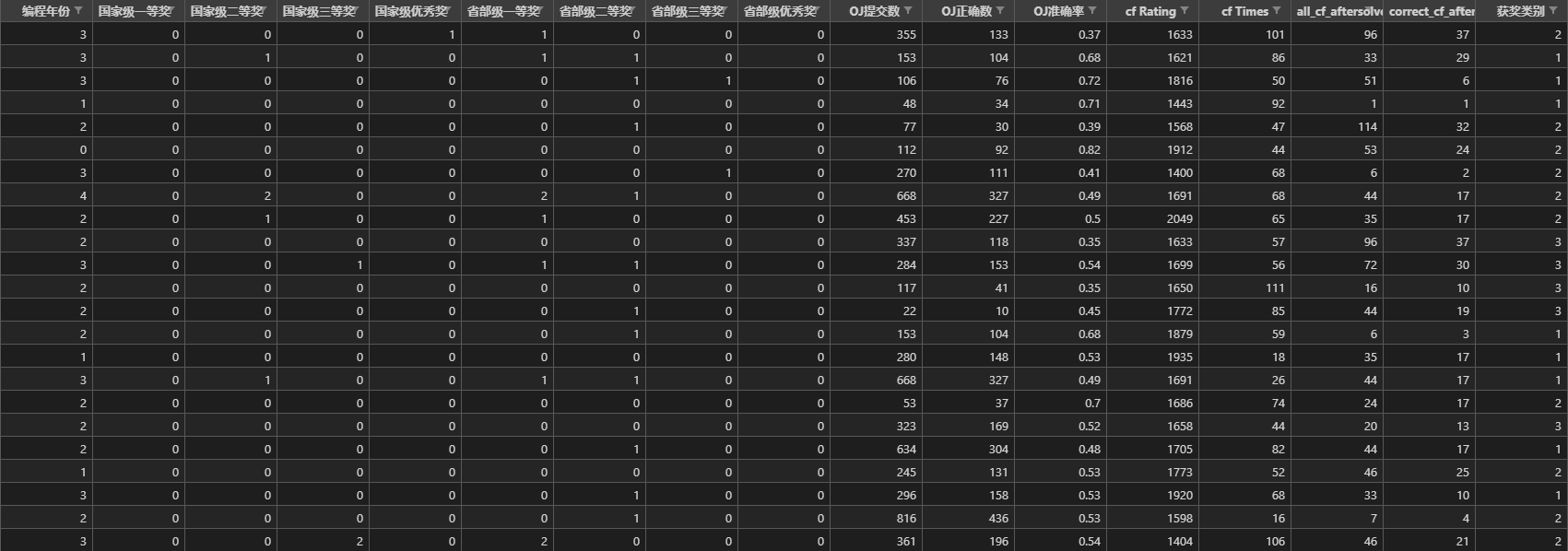
如图所示，可看出OJ提交数与准确率对于最终成绩有着一定的影响。但是影响并不显著。同时可以看到CodeForces积分与得奖的类别有着更高的相关性。CodeForces积分更高的同学，得到更高奖励的概率更大。

* + 1. RQ2

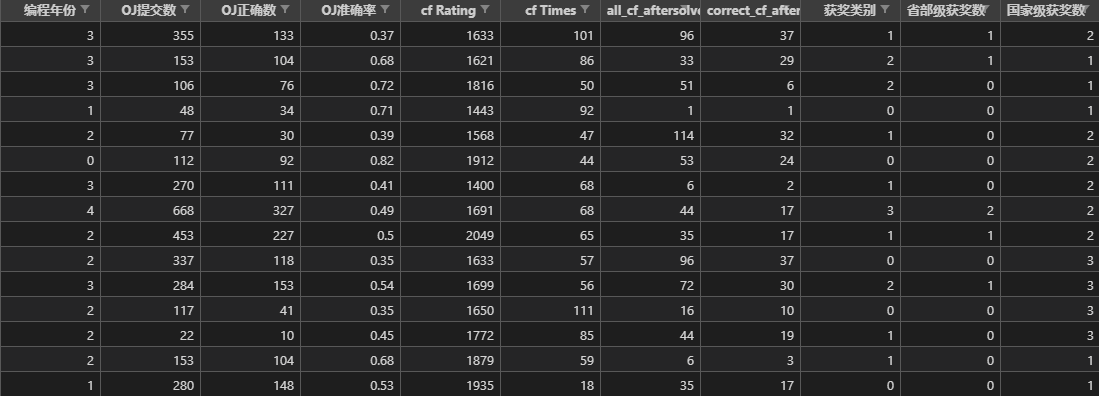


主要进行的降维操作是将之前获得各个奖项的个数统一为了之前所获国家级奖励及省部级奖励的个数，实现数据降维。如图所示，可以看到数据降维对不同的模型有着不同的效果。数据降维明显改善了KNN模型的准确率，将其准确率提升了0.06.同时另一方面，降维也明显降低了AdaBoost模型的准确率。降维对其他模型整体影响不大。但由于本次实验数据集规模较小，后续仍需在大规模的数据集中验证本次实验的数据。另外也能看出，在所有传统的机器学习模型中，GDBT（即梯度提升树）对于表格模型的拟合及分类预测效果最好。

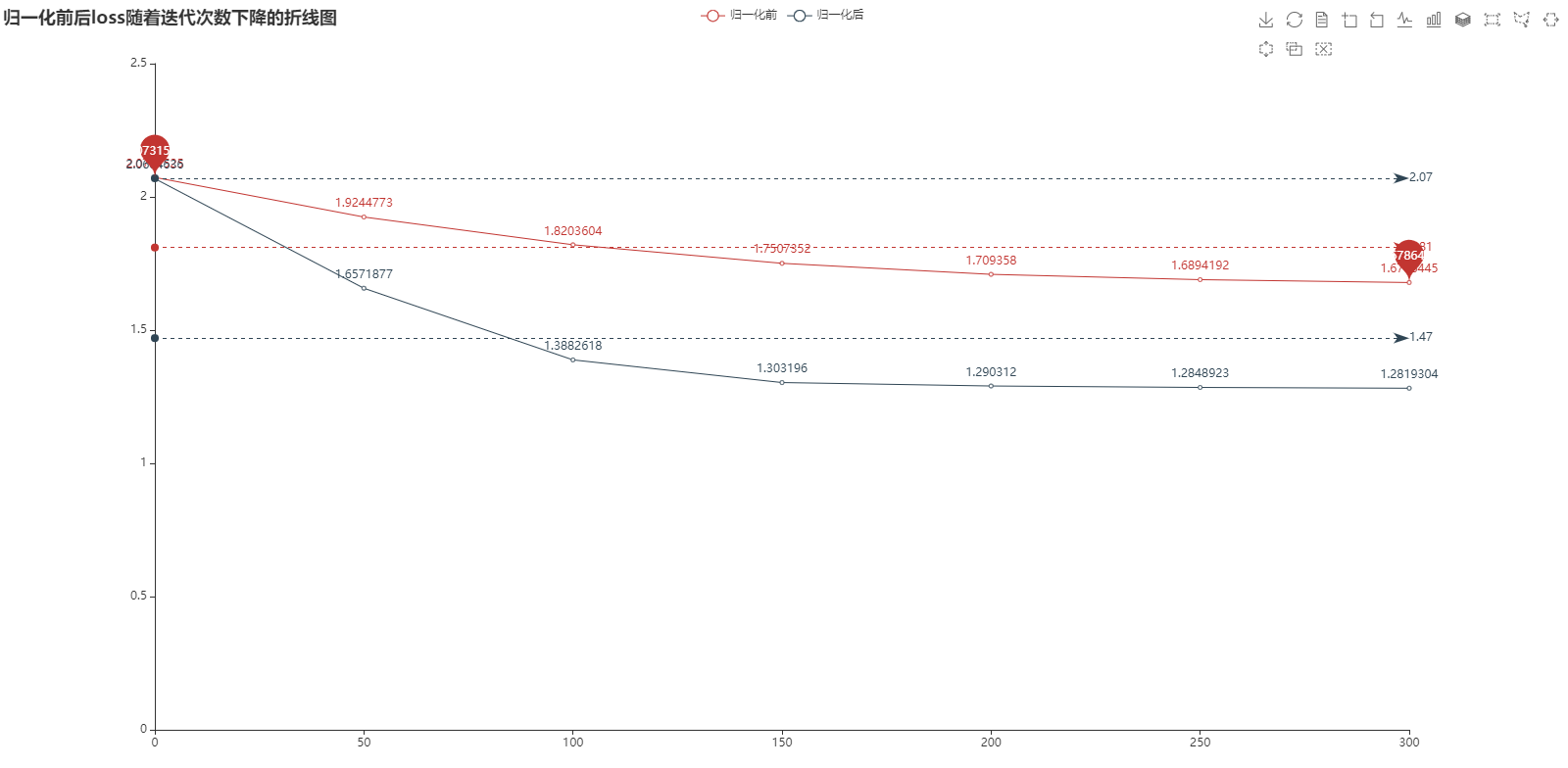
降维前数据如图所示：



降维后数据如图所示：



* + 1. RQ3



如图所示，可以明显观察到，在学习率一致的情况下，对数据进行归一化，将所有数据统一到[-1, 1]的区间范围内之后，神经网络梯度下降的收敛速度加快。同时可看出使用归一化数据模型收敛时的loss更小，模型准确率更高。说明归一化能够加快在梯度下降中训练神经网络地收敛速度，同时提高深度学习模型进行分类预测的准确率。

1. 结论与展望

本次实验通过使用多个模型对学生成绩进行预测, 并对影响学生成绩的各个因素进行了分析。实验分析结果可以为学生、老师提供一些意见与建议。但是因为数据集数量的限制，以及我们对于机器学习与深度学习理论知识的浅显理解，模型准确率不是特别理想。后续可以收集更大量的、更多种类的数据, 使得数据更加合理。并在经过进一步对机器学习与深度学习知识的研究与实践，采用更合适的模型与参数优化预测结果。在此基础上, 还可以针对模型结果为学生制定学习计划, 提出不同的训练建议。由此看来, 机器学习和深度学习值得我们更加深入地研究，来助其进一步发展，应用在更多领域。

1. 参考资源
2. 机器学习中的特征工程:

<https://blog.csdn.net/asialee_bird/article/details/84863410>

1. 机器学习中的数据清洗与特征处理综述:

<https://tech.meituan.com/2015/02/10/machinelearning-data-feature-process.html>

1. 完整机器学习项目的工作流程:

<https://zhuanlan.zhihu.com/p/28288050>

1. PaddleFluid与TensorFlow基本使用概念对比:

<https://zhuanlan.zhihu.com/p/38290380>

1. BUCT生产实习入门实践——鸢尾花分类:

<https://aistudio.baidu.com/aistudio/projectdetail/732131>

1. 《基于高校论坛数据的成绩预测和学生心理状况分析》电子科技大学，唐厚强
2. 《基于机器学习的高校学生成绩预测方法研究》山东大学，马玉玲
3. 组员课题完成情况详细说明

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 组员一 | 班级 | 姓名 | 学号 | 贡献度 |
| 计科1703 | 孟铃翔 | 2017040334 | 60% |
| 贡献描述 | 数据预处理与特征工程：60%  模型预测：80%  数据可视化：50%  报告与PPT制作：50% | | | |
| 组员二 | 班级 | 姓名 | 学号 | 贡献度 |
| 计科1703 | 黄帅旺 | 2017040338 | 40% |
| 贡献描述 | 特征预处理与特征工程：40%  模型预测：20%  数据可视化：50%  报告与PPT制作：50% | | | |