研究问题：自动定义题目难度

代码开源地址：<https://github.com/caposerenity/Data_Science>

研究方法：

就自定义题目难度这一问题，为我们使用了test\_data这一较全面的数据集，在下载所有测试数据集后，我们首先抽样查看了约一百份不同题型题目、不同同学的提交代码，发现其中存在使用C++等python外的语言编码、面向测试用例编程以刷分等现象。因此我们在评估题目难度之前首先对数据进行了预处理，清洗了面向测试用例的提交的分数记录。

在预处理筛出面向用例提交记录中，我们采取了正则表达式等特征匹配方法。针对先前抽样中出现的面向用例特征进行筛除，主要包含了：连续的if()print（）；else if（）print（）的情况；直接以if（）print（）开头且代码长度短于设置阈值的情况；仅有print（）内容且print中内容未包含运算的情况等若干种面向用例特征。通过对这些行为进行一定的赋分，达到阈值即认为存在面向用例的作弊行为，将这次提交标记为【cheat】并将分数改为0，同时在该同学这道题目的case记录下增加cheat标签。在对挑出的另外20次面向用例提交和30次正常提交记录进行的测试中，我们实现了95%的面向用例检测率和0%的误诊率。基于此，我们对test\_data中的所有提交记录进行了面向用例和是否为python代码的预处理过程。

数据预处理结束后，我们考察了每道题的分数特征（平均分，AC率，1A率（即第一次提交就满分占比））及时间特征（平均debug时间和提交次数），综合量化得出每道题目的难度系数。

对于每道题的分数，我们发现在经过预处理后的数据集中，出现的分数有大量的0分和大量的满分，因此仅仅以平均分作为判断难度的依据有失偏颇。于是我们将每道题的AC率及1A率也纳入考虑范围。

对于每道题的时间因素，我们考察了平均debug时间以及提交次数。对于debug时间，我们提取学生首次提交和最后一次提交的时间跨度作为考量特征，考虑到时间间隔内不全为debug行为，我们将权重设置为非线性分布，并因这些时间特征具有比较大的偶然性，我们将时间的基础分设为由分数分析得到的基础难度系数，得到最终难度系数

在进行时间分析时，我们认为标记为【cheat】的时间方面记录不具有难度参考价值，因此不做分析。下附权重表。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 提交时间跨度/min | t=0 | 0<t<=10 | 10<t<=30 | 30<t<=90 | 90<t<=300 | 300<t<=600 | 600<t<=1440 | t>1440 |
| 权重系数 | 0.8 | 0.4 | 0.6 | 0.7 | 0.75 | 0.85 | 0.9 | 1 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 提交次数 | n=0 | 0<n<=5 | 5<n<=10 | 10<n<=15 | 15<n<=20 | 20<n<=40 | n>40 |
| 权重系数 | 0.8 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1 |

得到这些时间特征和分数特征后，我们通过查阅资料选择了改进后的模糊层次分析法[1]，得到最终的计算指标权重如下（详细计算过程在附录），带入计算得到最终题目难度。

ωＣＯ ＝（ω分数, ω时间 ）＝（0.65,0.35），

ωＰＣ１ ＝（ω AC率 ,ω1A率 ,ω均分 ）=(0.383,0.317,0.3)

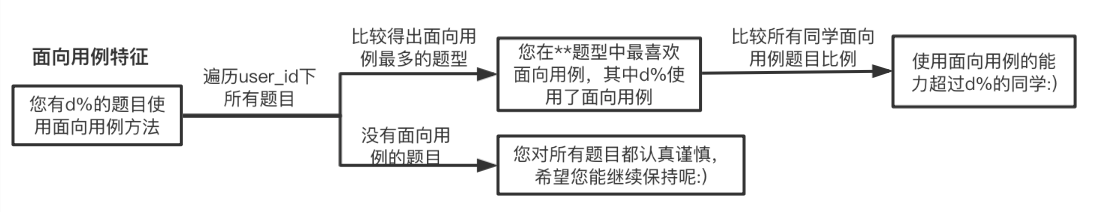
ωＰＣ２ ＝（ω debug时间 ,ω提交次数）=(0.55,0.45)

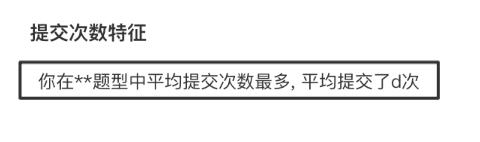
* 学生特征分析

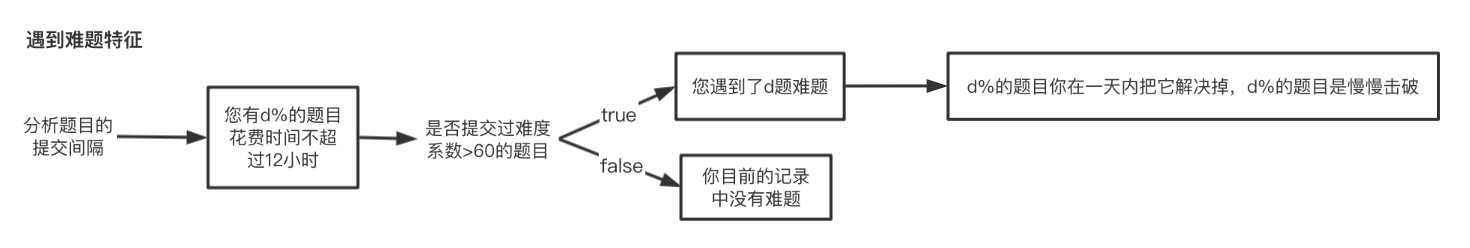
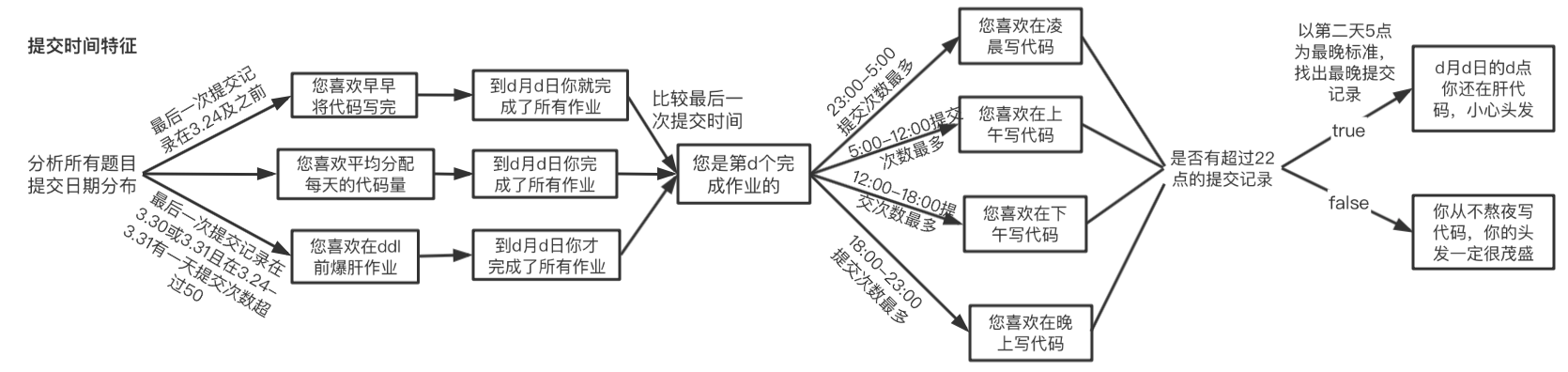
首先，我们考察学生对某类题目的掌握程度。但由于每道题目的难度有差异，因此直接用分数评判掌握程度不太合理。因此我们借助了偏态量化模型[2],根据学生在该题中的位次排名及题目的综合难度量化得出一个偏态量化分数，以此来判断学生对该题目的掌握程度。

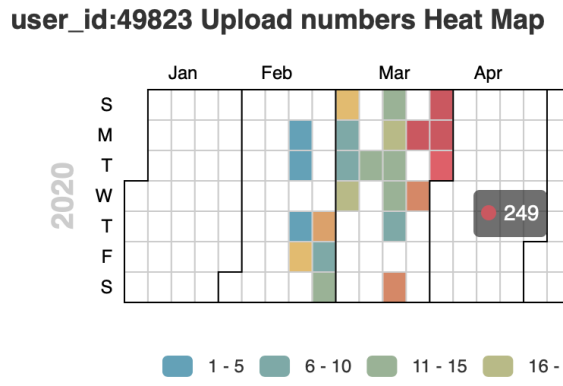
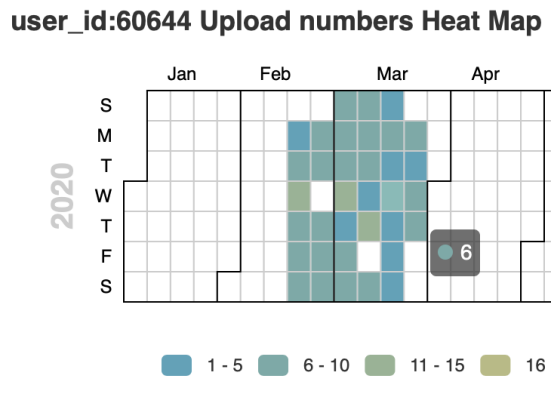
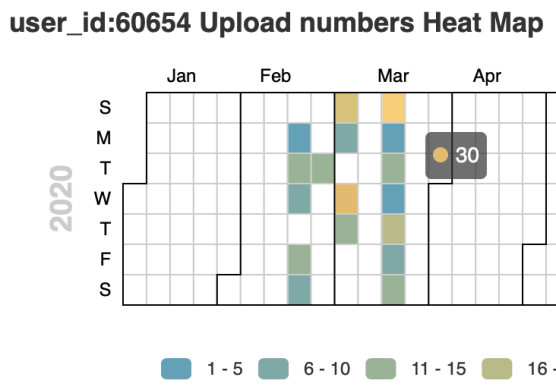
我们希望通过分析题目的提交时间、一道题目的提交间隔、面向用例题目数，得到学生做题的习惯特征，并反馈给学生，以此让同学能清楚地了解自己的做题习惯并希望他们能在此基础上养成良好习惯。

学生特征分为四个方面：面向用例、提交次数、做题时间的特征分析和遇到难题的特征分析，详情如下图。

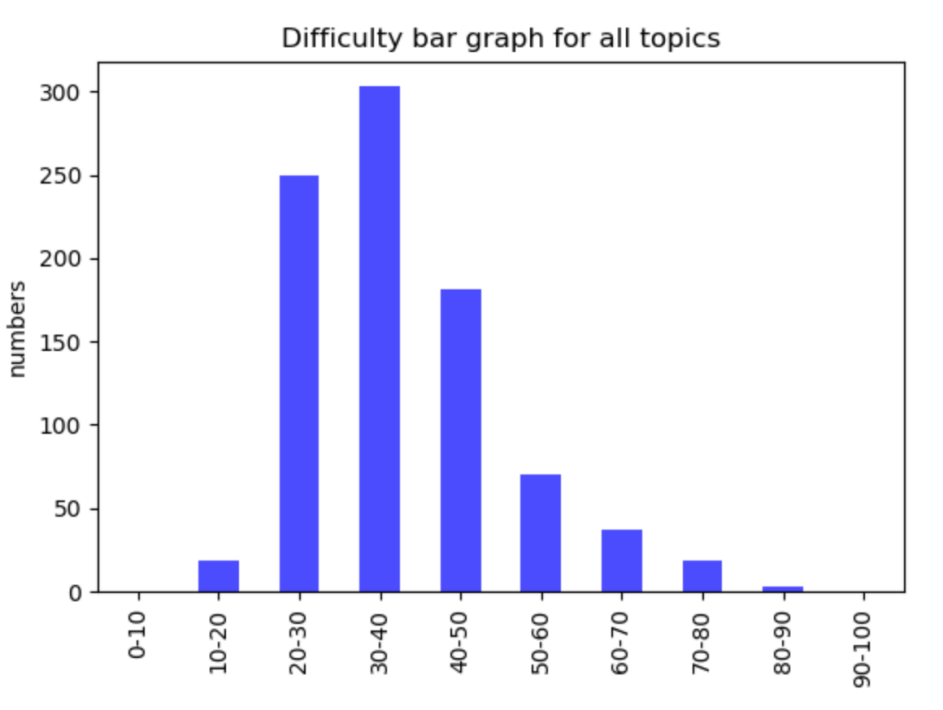




 其中，在分类提交时间特征时，我们利用提交时间热力图（src/visual/upload\_num\_heatmap），分类成“早早将代码写完”“平均分配每天代码量”“ddl前爆肝作业”三种不同模式（如下图），如果学生最后一次提交记录在3月24日及之前，判定为“早早将代码写完”；如果最后一次提交记录在3月30日或3月31日且在3.24-3.31中有一天提交总次数超过50判定为“ddl前爆肝作业”；其余的学生类型就判断为“平均分配每天代码量”。



在确定是否为难题时，我们也通过对所有题目的难度掌握（如下图）得出了一个划分难度系数60分，即难度系数超过60就视为难题。



**参考文献**

[1]刘立，王明艳，顾宝程，基于模糊层次分析法的软件编程能力评估模型，吉首大学学报（自然科学版），第３９卷 第２期， ２０１８年３月

[2] 李光兴、余洪琼，考试成绩转换成偏态分布量化成绩的算法，高师理科学刊第40卷第一期，2020年1月