小样本下的自适应学习策略研究

***

2018年12月06日

目录

- 目前需求
- 实验进展
- 实验计划

目前需求

大纪元数据（2万条）：负样本比较多，
占70%左右，目标是把负样本挑出来

人民日报数据（2483条）：作为正样本

实验设计

- 正面样本2138条(有效)
- 负面样本20000条(有效)

- 模型一训练集4000条（正面负面各2000条）
- 模型二训练集169条（69条正面样本100条负面样本）
- 模型二测试集169条（69条正面样本100条负面样本）

实验设计

数据格式：

- 模型一：Context:”新闻正文数据”, Label:”0/1” (0表示负面，1表示正面)


```
{"id": "0", "words": ["大", "纪元", "2018", "年", "11", "月", "14", "日讯", "（", "大", "纪元", "记者", "许
{"id": "1", "words": ["本报", "北京", "8", "月", "20", "日电", " ", "（", "记者", "李心萍", "）", "记者", "2
{"id": "1", "words": ["新", "唐人", "电视台", "将", "在", "感恩节", "期间", "播出", "神韵", "交响乐团", "演出",
{"id": "2", "words": ["372", "潜艇", "起航", "出港", "，", "向着", "目标", "海域", "隐秘", "进发", "。", "吴奔
{"id": "2", "words": ["大", "纪元", "2018", "年", "11", "月", "14", "日讯", "（", "大", "纪元", "记者", "吴
{"id": "3", "words": ["原", "标题", "：", "两年", "助", "31", "万", "“”", "黑户", "”", "办理", "身份证", "制图
{"id": "3", "words": ["大", "纪元", "2018", "年", "11", "月", "14", "日讯", "（", "大", "纪元", "记者", "吴
{"id": "4", "words": ["责编", "：", "羊", "金", "、", "袁勃", "人民日报", "客户端", "下载", "手机", "人民网",
{"id": "4", "words": ["大", "纪元", "2018", "年", "11", "月", "14", "日讯帛", "琉", "总统", "雷蒙杰", "索应"
{"id": "5", "words": ["北约", "“”", "蟒蛇", "－", "2018", "”", "军演", "近日", "在", "波兰", "拉开帷幕", "。"
{"id": "5", "words": ["大", "纪元", "2018", "年", "11", "月", "14", "日讯退", "团队", "声明", "认清", "中共"
{"id": "6", "words": ["美国", "《", "空军", "时报", "》", "网站", "21", "日", "报道", "称", "，", "当地", "时
{"id": "6", "words": ["大", "纪元", "2018", "年", "11", "月", "14", "日讯", "（", "大", "纪元", "记者", "张
{"id": "7", "words": ["图为", "8", "月", "19", "日", "，", "江西", "新干", "丰圣", "竹业", "有限公司", "员工"
{"id": "7", "words": ["大", "纪元", "2018", "年", "11", "月", "14", "日讯", "（", "大", "纪元", "记者", "张
{"id": "8", "words": ["图为", "蒙华", "铁路", "汉江", "特", "大桥", "。", "金", " ", "伟摄", "（", "新华社",
{"id": "8", "words": ["大", "纪元", "2018", "年", "11", "月", "14", "日讯", "大家", "好", "，", "欢迎", "大
```

实验设计

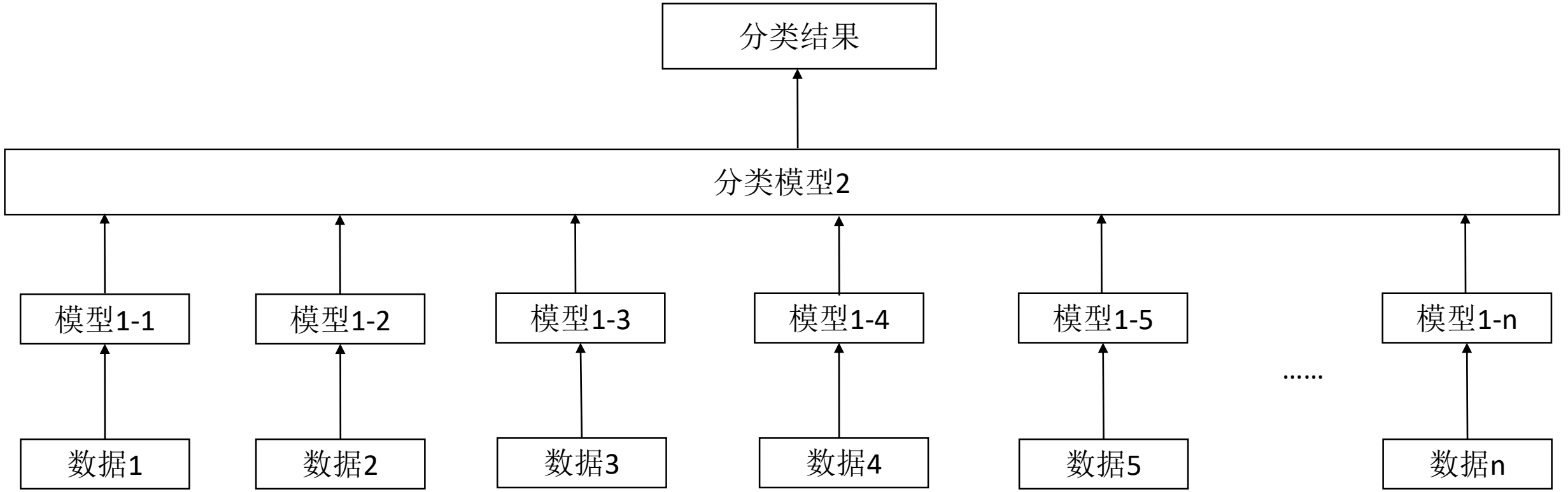
数据格式：

- 模型二：

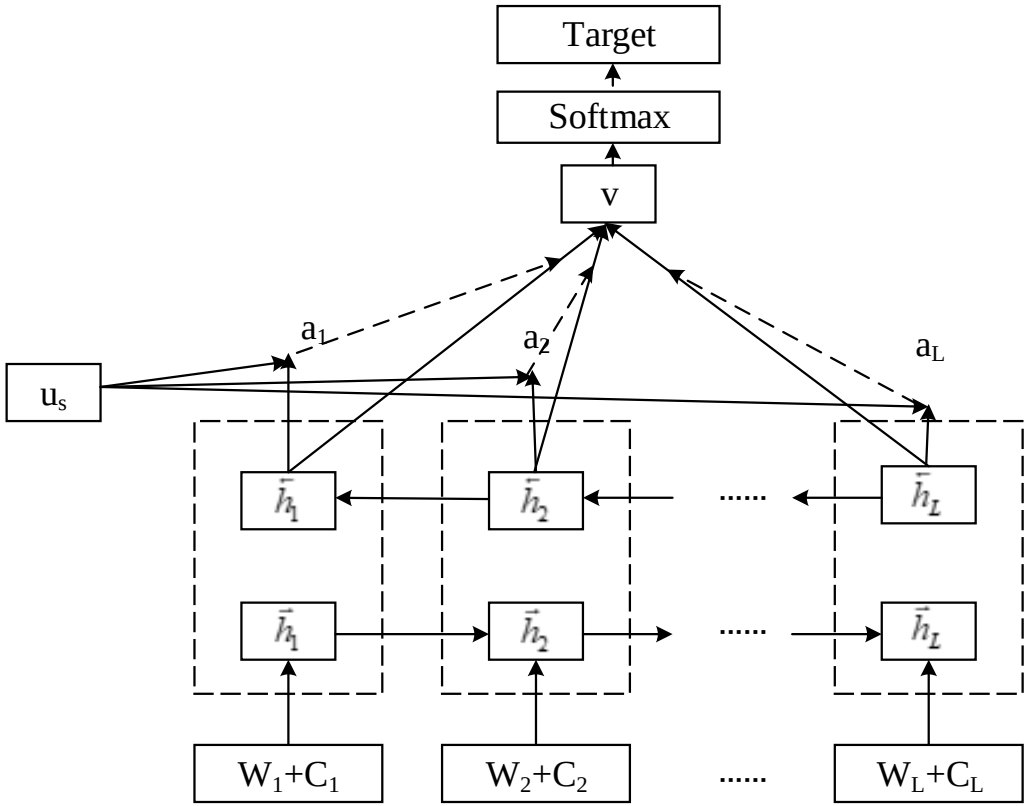
```
test2,test3,test4,test5,test6,test7,test8,test9,
1,0,1,1,1,1,0,0,1,0,0
0,0,1,1,1,1,0,0,1,0,1
1,0,1,0,0,1,0,0,1,0,0
0,0,0,0,1,0,0,1,1,0,1
0,0,0,0,0,0,0,0,1,0,0
1,1,1,1,1,1,0,1,1,0,1
1,0,1,0,0,0,0,0,1,0,0
0,0,1,0,1,1,0,1,1,0,1
0,0,1,0,0,0,0,0,1,0,0
1,1,1,1,1,1,1,1,1,0,1
0,0,0,0,0,0,0,0,1,0,0
1,1,1,1,1,1,1,1,1,0,1
1,0,1,0,0,0,0,0,1,0,0
0,0,0,0,0,0,0,1,1,0,1
0,0,0,0,0,0,0,0,1,0,0
1,1,1,1,1,1,0,1,1,0,1
0,0,0,0,0,0,0,0,1,0,0
```

同济大学
TONGJI UNIVERSITY

实验设计：整体结构

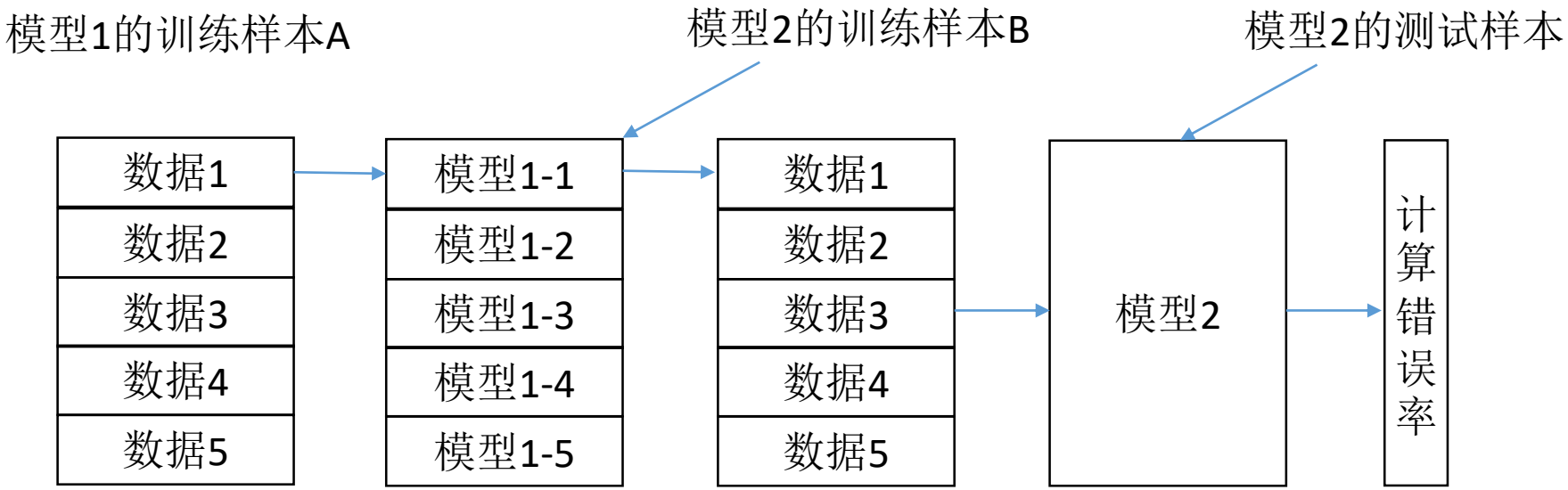


实验设计：模型一结构



同济大学
TONGJI UNIVERSITY

实验设计



实验结果：5个模型

每个模型的输入数据：
模型一数据集：40000/5=8000条 （训练集6400，测试集1600）

模型二数据集：169条 （训练集136，测试集33）

模型测试数据：169条

```
pred_result:[0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0
 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1
 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0
 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0]
true_result:[0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0
 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1
 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0
 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
err_ratio:0.005917159763313609
```

测试集五次测试集结果：

第一次实验 错误率：
0.4082
0.4082
0.0000
0.0000
0.0000

第二次实验 错误率：
0.0237
0.0059
0.0059
0.0059
0.0059

TONGJI UNIVERSITY

实验结果：10个模型

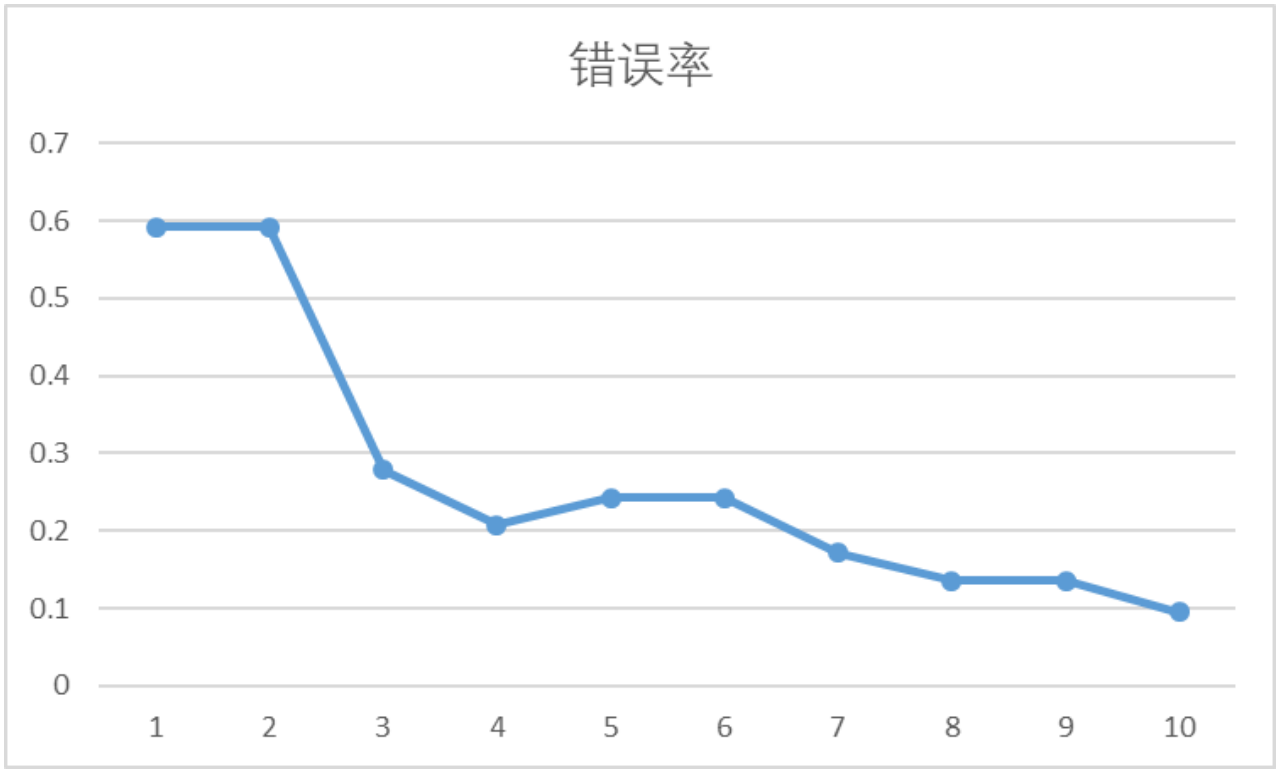
每个模型的输入数据：
模型一数据集：40000/10=4000条 （训练集3200， 测试集800）

模型二数据集：169条 （训练集136，测试集33）

模型测试数据：169条

```
pred_result:[1 1 1 0 0 1 1 1 1 1 0 1 1 0 0 1 0 1 0 1 0 1 1 1 0 1 0 1 1 1 1 1 1 1 1 1 1
 1 1 0 0 1 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 0 1 0 1 1 0 1 1 0 1 0 1 1 0 0 1
 0 1 0 1 1 1 1 0 1 0 0 1 0 1 0 1 1 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0
 1 0 0 0 1 1 1 0 1 1 1 0 1 0 1 0 1 0 1 0 0 0 1 0 1 0 0 0 0 0 0 0 1 0 1 0 0
 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0]
true_result:[0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0
 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1
 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0
 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
err_ratio:0.24260355029585798
```

实验结果：10个模型



测试集十次测试集结果：

0.5917

0.5917

0.2781

0.2071

0.2426

0.2426

0.1716

0.1360

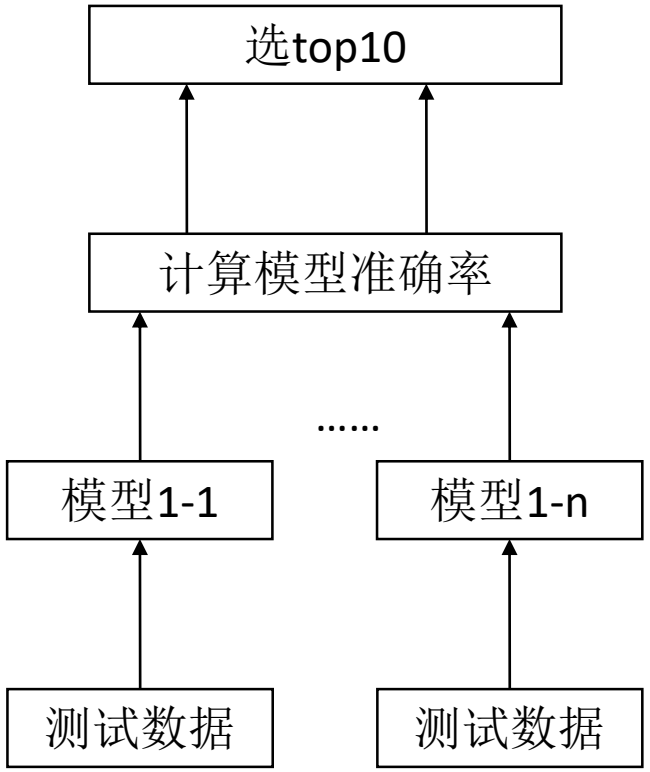
0.1360

0.0946

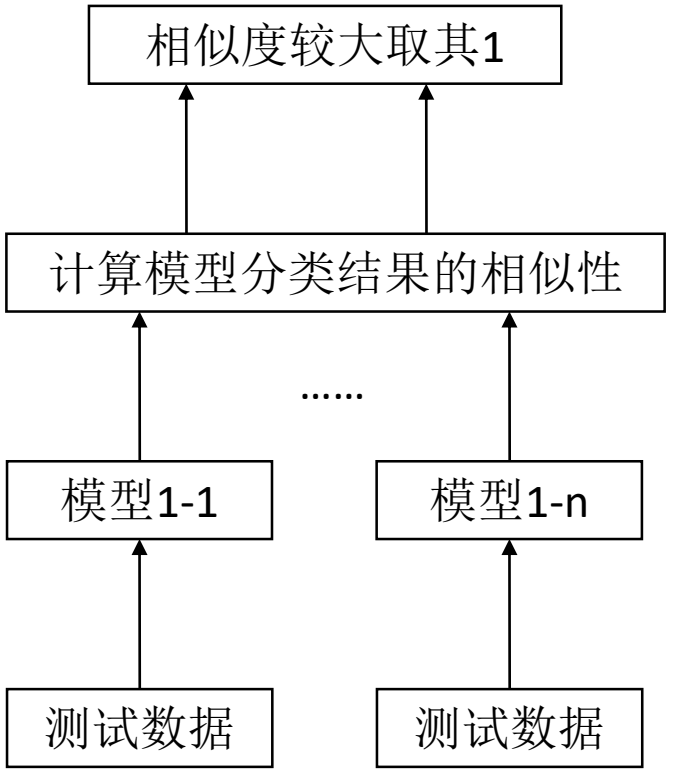
全部的**0.0313**

实验计划

1. 筛选model1模型



根据得分进行筛选

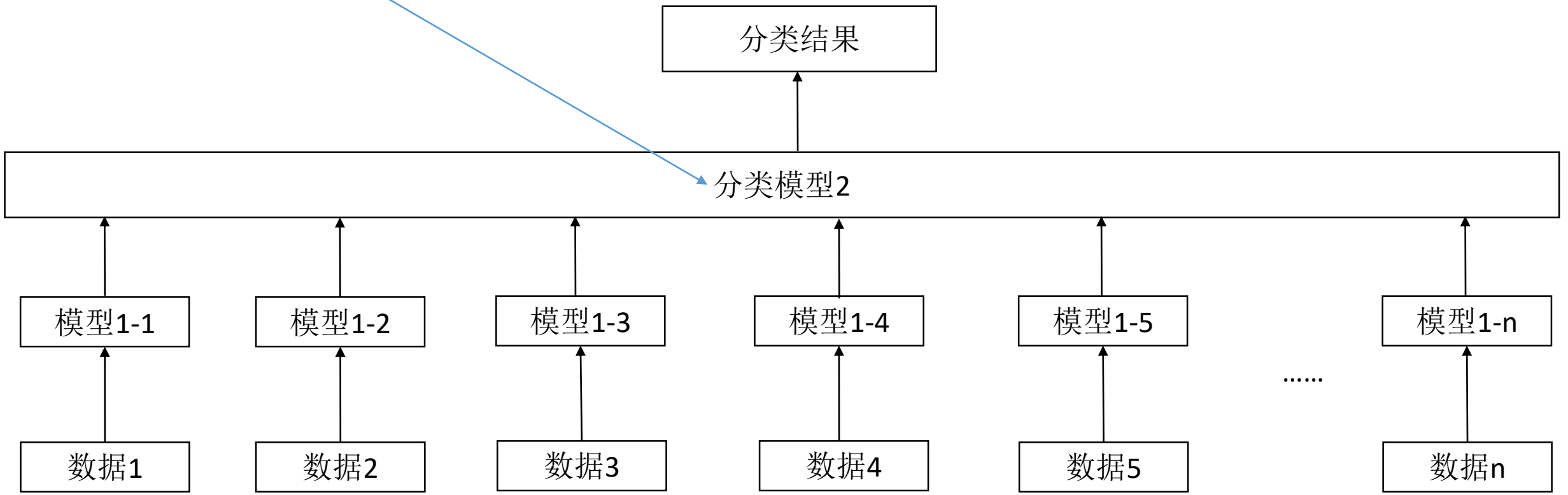


根据结果的相似性进行筛选

实验计划

2. 修改model2模型从MLP改为决策树或者其他模型

3. 增加强化学习策略，训练模型2



谢谢！O_O

***

2018年12月06日