# 题目自动标注模型实验报告——智慧学伴

1. 实验目的

本次实验希望实现针对题目的自动标注功能，即能够支持系统根据题干信息自动识别题目最可能的知识点分类及能力层级分类。

1. 实验原理

本实验依据分类语料（即智慧学伴题目数据，共计4685道）与分类体系（包括知识点分类与能力层级分类）作为初始数据集及分类规则，借助TF-IDF统计方法进行关键性特征提取，采取四种机器学习模型（SVM，KNN，Decision Tree，MLP）进行分类训练实验，并尝试进行调参及添加辅助性信息规则以提高模型的准确率。

1. 实验过程

本实验原始数据集来源于智慧学伴系统内现有题目，共计4685道。

题目分为选择题，简答题和主观题，针对选择题提取题干外加正确选项作为输入，针对简答题和主观题提取题干和正确答案作为输入。

能力层级分类共采用了5种分类方式：按“一级知识点”进行分类，共分7类；按“二级知识点”进行分类，共分23类；按“三级知识点”进行分类，共分95类；按“一级能力层级”（即A、B、C层级），共分3类；按“二级能力层级”（即A1、A2、A3、B1……C3层级），共分9类。

实验过程中首先采用分词工具对原始数据集进行分词操作，并运用TF-IDF统计方法进行关键性特征提取，其次选取4种机器学习模型，即SVM（支持向量机分类器）、kNN（K最近临分类算法）、DT（Decision Tree，决策树预测模型）以及MLP（多层感知机，即人工神经网络），利用原始数据集的80%作为训练集进行分类训练实验，同时采取十折交叉验证来测试算法准确性，最后利用剩余20%的原始数据集作为测试集进行模型准确度验证。

1. 实验结果
2. 按“一级知识点”进行分类时，利用4种机器学习模型进行训练的训练集及测试集准确率分别如下：
3. 训练集10-flod交叉验证准确率

* SVM: 0.88 0.87 0.85 0.89 0.91 0.89 0.86 0.87 0.88 0.89
* MLP：0.89 0.88 0.86 0.90 0.89 0.88 0.86 0.87 0.89 0.90
* DT: 0.82 0.83 0.80 0.83 0.81 0.78 0.79 0.78 0.78 0.80
* KNN: 0.77 0.76 0.76 0.72 0.76 0.80 0.76 0.77 0.78 0.79

1. 测试集准确率

* SVM: 0.89
* MLP: 0.90
* DT: 0.82
* KNN: 0.77

1. 按“二级知识点”进行分类时，利用4种机器学习模型进行训练的训练集及测试集准确率分别如下：
2. 训练集10-flod交叉验证准确率

* SVM: 0.77 0.77 0.75 0.81 0.81 0.77 0.79 0.76 0.80 0.79
* MLP：0.79 0.78 0.75 0.76 0.80 0.75 0.76 0.76 0.81 0.77
* DT: 0.67 0.69 0.65 0.71 0.71 0.67 0.65 0.68 0.65 0.66
* KNN: 0.61 0.63 0.59 0.60 0.63 0.67 0.63 0.65 0.67 0.63

1. 测试集准确率

* SVM: 0.79
* MLP: 0.80
* DT: 0.67
* KNN: 0.65

1. 按“三级知识点”进行分类时，利用4种机器学习模型进行训练的训练集及测试集准确率分别如下：
2. 训练集10-flod交叉验证准确率

* SVM: 0.60 0.63 0.60 0.64 0.65 0.64 0.61 0.62 0.64 0.63
* MLP：0.61 0.61 0.58 0.64 0.62 0.61 0.59 0.63 0.61 0.65
* DT: 0.52 0.46 0.48 0.54 0.49 0.50 0.46 0.51 0.52 0.51
* KNN: 0.42 0.43 0.42 0.40 0.44 0.44 0.42 0.43 0.46 0.42

1. 测试集准确率

* SVM: 0.0.62
* MLP: 0.61
* DT: 0.49
* KNN: 0.44

1. 按“一级能力层级”进行分类时（共分3类），利用4种机器学习模型进行训练的训练集及测试集准确率分别如下：
2. 训练集10-flod交叉验证准确率

* SVM: 0.65 0.63 0.61 0.62 0.64 0.67 0.61 0.64 0.62 0.64
* MLP：0.62 0.60 0.56 0.60 0.62 0.67 0.58 0.61 0.61 0.61
* DT: 0.58 0.58 0.59 0.60 0.60 0.59 0.58 0.62 0.61 0.56
* KNN: 0.57 0.52 0.52 0.56 0.52 0.57 0.54 0.50 0.57 0.58

1. 测试集准确率

* SVM: 0.63
* MLP: 0.61
* DT: 0.59
* KNN: 0.55

1. 按“二级能力层级”进行分类时（共分9类），利用4种机器学习模型进行训练的训练集及测试集准确率分别如下：
2. 训练集10-flod交叉验证准确率

* SVM: 0.51 0.50 0.48 0.51 0.50 0.51 0.39 0.51 0.48 0.50
* MLP：0.46 0.45 0.44 0.47 0.50 0.47 0.41 0.47 0.44 0.48
* DT: 0.40 0.41 0.34 0.37 0.40 0.35 0.33 0.39 0.39 0.38
* KNN: 0.36 0.34 0.35 0.38 0.35 0.36 0.30 0.36 0.36 0.39

1. 测试集准确率

* SVM: 0.49
* MLP: 0.45
* DT: 0.41
* KNN: 0.34