贝叶斯

朴素贝叶斯 + 对位置信息进行编码。

|  |
| --- |
|  |
| 图1 朴素贝叶斯的分类效果 |

模型的正确率约为50%，共有1702个汉字分错，从分错的汉字中抽100个样本进行观察，发现存在以下几种错误

|  |
| --- |
|  |
| 图2 错误分类的样本分析 |

拟议的后续改进的方案：

1. 针对包含错误，可以考虑**新增一个汉字笔画的特征**，新增后P(庞|abc笔画=3)<P(龙|abc笔画=3)，理论上可行。
2. 针对位置错误，一种基于常识的规则是：一个汉字，与他位置信息（比如左右结构，上下结构，全包围半包围结构）相似的字比与他内容信息（包含的部首）相似的字少，即部首编码信息在分类时应占有更大的权重，**可以考虑基于特征加权的贝叶斯分类器。**

其他方案

1. 转化为结构化数据，一个汉字对应一个样本。为了保留部首的位置信息，对一个部首（如一）映射到27个位置：分别代表大中小、左中右、上中下。数据集中一共440个部首，部首编码到443，构造出的特征11961列，高度稀疏，尝试使用Light GBM自带的互斥特征合并方法进行降维，效果很差PR均低于1%，目前没想到更好的转化方法。
2. **尝试Knn，定义一个针对该任务的距离衡量方案**。

后续方案

（1）特征构造（2）尝试基于特征加权的贝叶斯分类器（3）Knn