



《文本分析与挖掘》

**2021/2022(1)**

期末综合实验报告

报告题目 警情数据分析并实现类别预测

学生班级 计智1901班

学生姓名 杨宇圳

学生学号 201906061428

**计算机科学与技术学院**

目录

[1 问题分析 3](#_Toc92288061)

[2 对警情数据进行统计分析 3](#_Toc92288062)

[2.1 警情数据总体情况 3](#_Toc92288063)

[2.2 各个区域警情情况 3](#_Toc92288064)

[2.3 类别词云分析 5](#_Toc92288065)

[2.3.1 对数据进行预处理 5](#_Toc92288066)

[2.3.2 显示词云 5](#_Toc92288067)

[3 基于textcnn实现类别预测 6](#_Toc92288068)

[3.1 使用原始数据 6](#_Toc92288069)

[3.2 对文本进行适当的预处理 7](#_Toc92288070)

[3.3 使用SMOTE、ADASYN降低类别不平衡影响 9](#_Toc92288071)

[3.3.1 SMOTE 9](#_Toc92288072)

[3.3.2 ADASYN 10](#_Toc92288073)

[3.3.3 使用SMOTEENN进行尝试 10](#_Toc92288074)

[3.3.4 使用SMOTETomek进行尝试 10](#_Toc92288075)

[3.3.5 结论 11](#_Toc92288076)

[4 讨论与总结 11](#_Toc92288077)

基于testcnn方法的警情类别分类问题

1. 问题分析

近年来，随着警情信息量逐年的爆发式增长,公安警务工作日趋繁重,警情信息管理工作压力日益增大,而警情可以根据案件情况和类型进行分类，警方则根据不同的类别对案件进行针对性地处理，科学合理配置警力，以便高效解决案件，避免不必要的警力浪费。所以，根据警情信息对案件进行分类就尤为重要。

当前的警情分类功能是接警人员根据报警内容进行人工标注得到的,人工分类不仅工作量较大,分类结果也会受到接警人员的主观影响。

所以本实验基于真实的警情数据集，针对短文本实现数据的挖掘与分析，对警情进行数据统计与关键词提取，而有助于宣传以及防范同类别事件的发生；本实验也对警情案件实现自动归类，以提高警务人员工作效率和准确率。

实验数据集包含两列非文本信息：发生地域编号和报警人性别，以及两列文本信息：报警内容和出警情况，再加上一个文本信息：最终分类结果。警情一共分成 6 个警情类别。

1. 对警情数据进行统计分析
   1. 警情数据总体情况

对警情数据集进行导入和拼接后，在进行统计后使用pandas\_bokeh()进行画图。得到如图2-1的统计图。

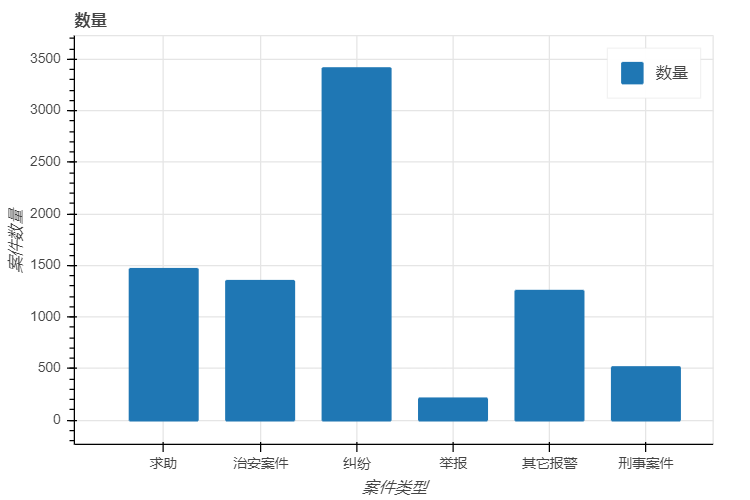


图2-1

通过对上图的分析，可以发现该警情数据集存在大量的纠纷案件，举报案件较少，警方应对纠纷案件加以重视，加强防范。

* 1. 各个区域警情情况

根据警情类别和发生地域进行分类和统计，利用pandas\_bokeh()得到如下图2-2-1和图2-2-2：

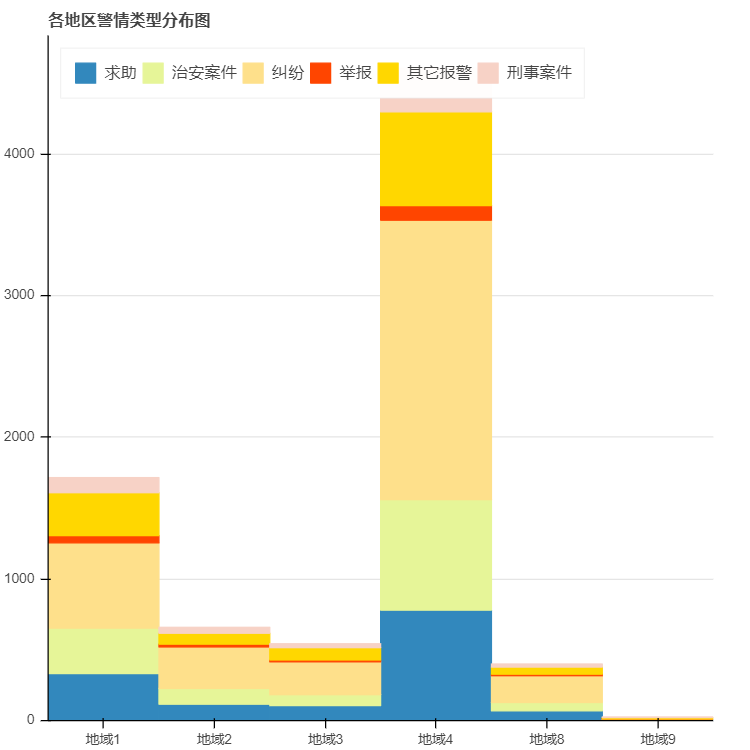


图 2-2-1

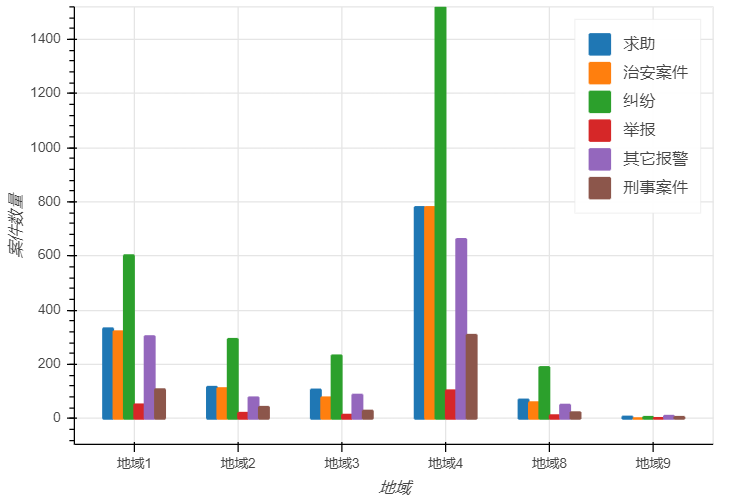


图 2-2-1

警方可以根据这两张统计图获取相关信息，如本次实验警情数据多发生在地域4、其次为地域1，应对这两地多进行管理防范。

* 1. 类别词云分析

基于各个类别可能存在不同的关键词，本实验使用词云技术过滤掉大量的文本信息，形成“关键词云层”或“关键词渲染”，对不同类别文本中出现频率较高的“关键词”进行视觉上的突出，实现类别数据的可视化，方便警务人员从大量的警情数据中获取类别关键词，提高以后进行类别分类的准确率和效率。

* + 1. 对数据进行预处理

本实验对报警内容和出警情况的文本进行拼接后，使用jieba进行分词得到每个类别的分词文本结果，发现存在大量无用词语和通用词语，如日期时间、脱敏符号“\*”、“民警”和“辅警”等，对词云的显示造成了干扰。所以对分词后的结果进行清洗，更好地展示图云效果。

* + 1. 显示词云

经过数据预处理后，进行词云显示，以下为词云的结果

|  |  |
| --- | --- |
| 纠纷 | 举报 |
| 其他报警 | 求助 |
| 刑事案件 | 治安案件 |

可以通过词云展示，粗略地展示关键词。如纠纷类别的关键词有纠纷和协商。举报的关键词有赌博、棋牌室等。

1. 基于textcnn实现类别预测
   1. 使用原始数据

首先尝试不对报警内容和出警情况做任何清洗处理，使用原始数据进行训练。

1）预处理

将 “报警内容”和 “出警情况（反馈内容）”的文本内容进行拼接，使用jieba进行分词，并对把每个文本转成长度统一为 max\_len=100 的向量，太长的截断，太短的后面补零，把每个文档生成为一个长度为100的向量。

2）设置模型

a.设置 embedding 层输出维度 embed\_dim=200；

b.卷积层设置三种卷积核大小 kernel\_size=[2,3,4]，每种大小用64 个卷积核，即 filters=64；

c.倒数第二的全连接 dense 层输出 64 维，Relu 激活；

d.最后一个输出二分类结果，6 个节点，softmax 激活。

3）模型的训练

设置损失函数为sparse\_categorical\_crossentropy，优化器采用adam，训练50个epoch，并设置以8轮为限制的早停回调函数。

4）训练结果

最终在训练集的准确率为0.99，验证集的准确率为0.80，如下图3-1-1所示



图 3-1-1 训练结果

在测试集的准确率为0.78，测试集结果如下图3-1-2所示：

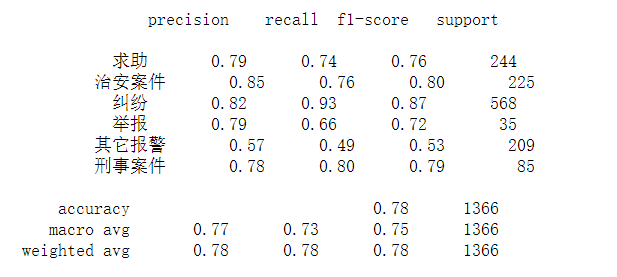


图 3-1-2 测试集结果

5）分析结果

未对文件进行任何清洗，存在很多无效字符，猜测可能会影响到训练结果，且模型过拟合现象严重。

* 1. 对文本进行适当的预处理

1）数据载入

如3.1一致，将 “报警内容”和 “出警情况（反馈内容）”的文本内容进行拼接。

2）文本的清洗

查看原始文本，如下图3-2-1所示：

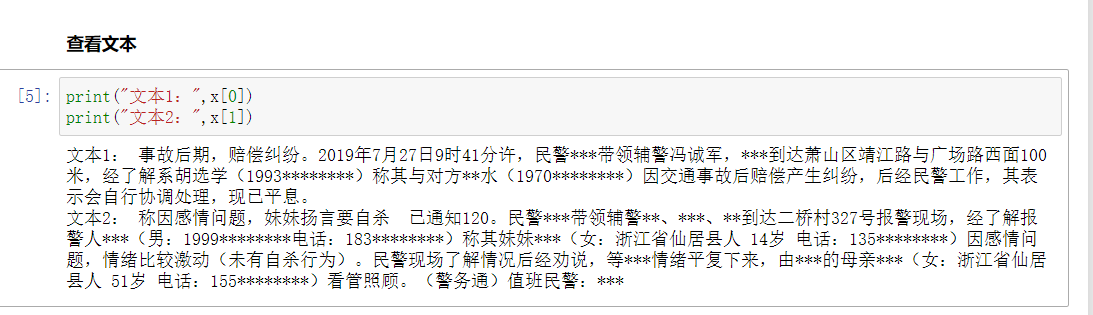


图 3-2-1 原始文本数据

有大量的无用字符"\*"，而且括号中的信息多为身份信息、日期和电话等，也对分类无效，应当进行清除。

清洗后的文本如下图3-2-2所示：

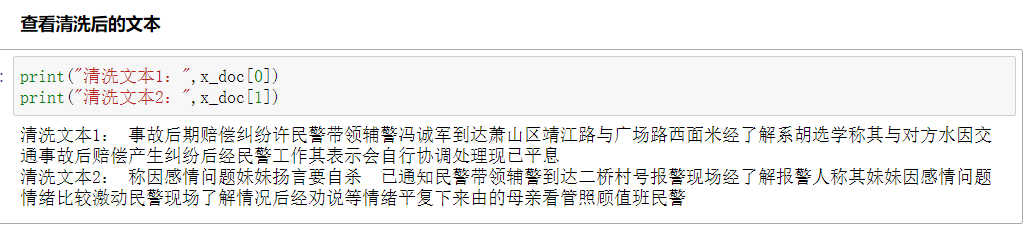


图 3-2-2 清洗后的文本

通过观察发现，众多无用信息已被删除。

3）数据的预处理与3-1一致

4）模型的配置与训练也与3-1保持一致

5）训练结果

训练结果如下图3-2-3所示，训练集正确率为0.96，验证集为0.82



图 3-2-3 训练结果

测试结果如下图3-2-4所示

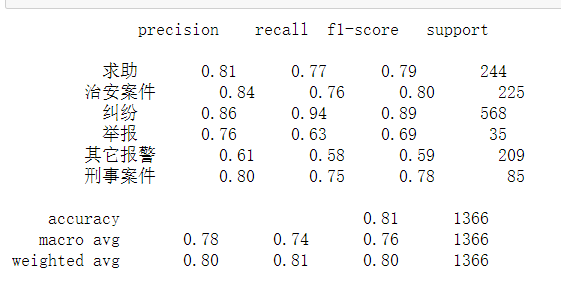


图 3-2-4 测试集结果

相比于3-1未对文本进行的清洗的模型，准确率提高了0.02，反映出文本清洗也是有一定的必要性。

6）参数调优

* 调整词向量维度为400，句长为300
* 调整模型结构，使用3个一维卷积，卷积核大小 kernel\_size=[3,4,5]，每个卷积层有128 个卷积核
* 调整dropout层比例为0.6，降低过拟合。

进行训练后，得到如下图3-2-5结果

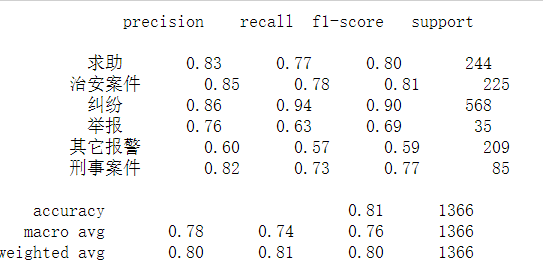


图 3-2-5调优后结果

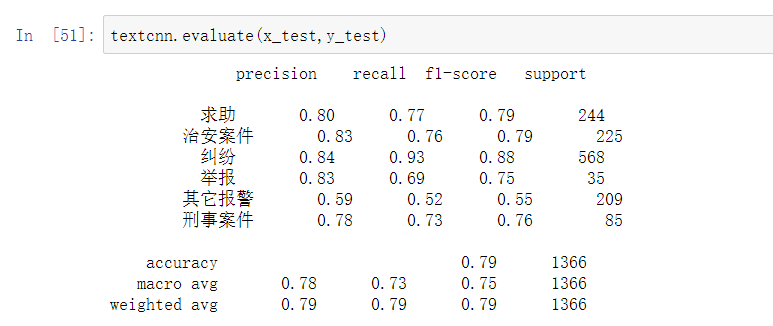
虽然总体准确率未上升，但各个类别的f1-score值有小幅上升，提高一些模型的性能。

* 1. 使用SMOTE、ADASYN降低类别不平衡影响

查看警情数据，发现各个类别的比例存在不平衡的现象。故使用SMOTE、ADASYN这两种方法对预处理过后的数据进行过采样处理，由于3-2的模型性能更加，所以模型的配置与训练设置均与3-2保持一致。

* + 1. SMOTE
       1. 使得6个类别的比例近似为1

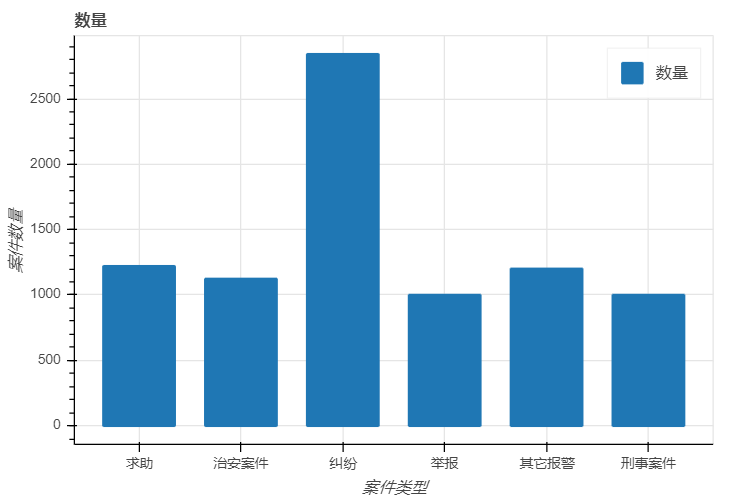
下图为测试结果：



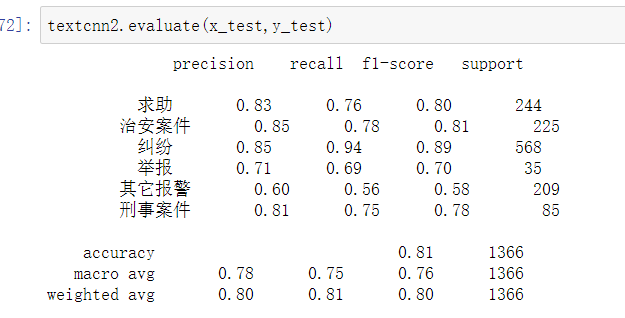
准确率为0.79，有些下降，但小类别“举报”类警情的f1-score上升了。

* + - 1. 适当调整各个类别的比例

调整后的比例如下图所示。



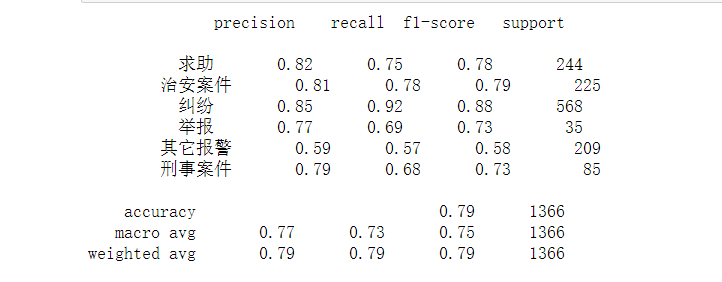
测试集结果如下图所示



总体的准确率上升至0.81。

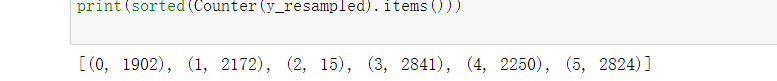
* + 1. ADASYN

以下为测试结果，与SMOTE类似



* + 1. 使用SMOTEENN进行尝试

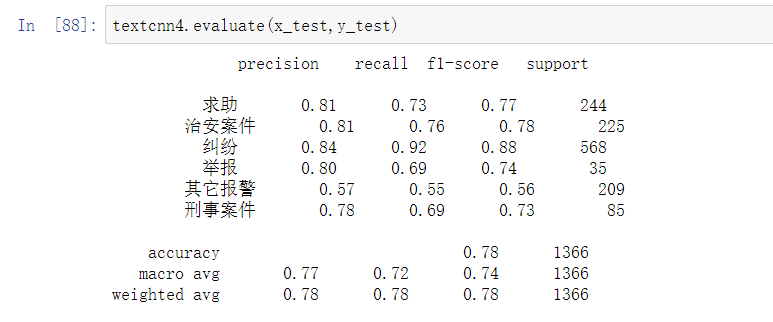
使用SMOTEENN进行尝试，查看各个类别情况，如下图



发现原本的第二大类仅剩余15个数据，猜测SMOTEENN将这些数据都当成噪声删除了，SMOTEENN方法不使用于本数据集。

* + 1. 使用SMOTETomek进行尝试

测试结果如下图所示



* + 1. 结论

使用了多个过采样的方法进行尝试，测试集的结果显示，预测性能未能有很好的提升，反而有些下降，尝试提升预测性能失败，SMOTE、ADASYN方法对该数据集不是十分有效。

1. 讨论与总结

在数据集分析的模块的中，为了能有更好地可视化展示，使用了pandas\_bokeh进行绘图，其相比于普通的plt绘图，其优势在于统计展示的图像可以进行交互，方便警员进行查看。

在类别预测的模块中，通过实验结果发现，使用文本清洗的数据，最中模型准确率性能较未清洗的有一定提升。尝试使用smote、adasyn过采样方法未能解决样本不均衡问题，模型的性能未能有显著提升。

经过一系列的调整，最终，预测类别的测试集准确率在0.81，对实际的警情分类预测有一定的参考价值。