基于 LDA 主题模型的文本分类

刘张辰宇
ZY2203703

Abstract

以 16 本武侠小说作为语料库，首先对语料库进行预处理，将除中文以外的字符去掉，并将语料库分别以字和词为单位，切割成训练集和测试集，通过 LDA 主题模型，对训练集进行主题建模，得到每篇小说的主题概率分布以及每个主题的字(词)分布，并计算测试集的主题概率分布，通过选择字(词)出现概率最大的主题概率分布确定文本最终的主题概率分布，使用欧氏距离作为文本相似的衡量指标，最终对测试集文本作分类。

Introduction

LDA，是一种文档主题生成模型，它可以将文档中每篇文档的主题按照概率分布的形式给出。也称为一个三层贝叶斯概率模型，包含词、主题和文档三层结构。所谓生成模型，就是说，我们认为一篇文章的每个词都是通过“以一定概率选择了某个主题，并从这个主题中以一定概率选择某个词语”这样一个过程得到。文档到主题服从多项式分布，主题到词服从多项式分布。

LDA 是一种非监督机器学习技术，可以用来识别大规模文档集（document collection）或语料库（corpus）中潜藏的主题信息。它采用了词袋（bag of words）的方法，这种方法将每一篇文档视为一个词频向量，从而将文本信息转化为了易于建模的数字信息。但是词袋方法没有考虑词与词之间的顺序,这简化了问题的复杂性,同时也为模型的改进提供了契机。每一篇文档代表了一些主题所构成的一个概率分布,而每一个主题又代表了很多单词所构成的一个概率分布。

LDA 的核心思想是寻找到最佳的投影方法，将高维的样本投影到特征空间(feature space)，使得不同类别间的数据“距离”最大，而同一类别内的数据“距离”最小。

Methodology

本次实验通过 LDA 主题模型，对语料库切割得到的训练集进行主题建模，并对测试集进行主题建模后的分类。

M:LDA

本次实验的算法步骤：

1. 对语料库进行预处理，剔除无关字符
2. 对语料库分别进行以字、词为单位分割，并分别对每本小说均匀提取 13 个段落，每个段落大小分别为 1000、500，得到训练集和测试集
3. 确定主题数量 topic_nums
4. 通过 LDA 对训练集进行建模，得到每本小说的主题概率分布及每个主题的字(词)分布，并不断迭代更新主题概率分布，直至收敛不变
5. 使用训练集得到的主题字(词)分布，计算训练集的主题概率分布，不断迭代直至收敛

6. 使用欧式距离作为评价指标，计算训练集与每本小说主题概率分布之间的差值，取最小值对应的小说作为训练集的小说类别

Experimental Studies

A. 以字为单位：

当主题数量 topic_nums = 20 时，
输出结果为：

```
正确分类的数量: 21
错误分类的数量: 187
正确分类率 0.10096153846153846
训练迭代次数 4
测试迭代次数 4
```

当主题数量 topic_nums = 40 时，
输出结果为：

```
正确分类的数量: 28
错误分类的数量: 180
正确分类率 0.1346153846153846
训练迭代次数 5
测试迭代次数 4
```

当主题数量 topic_nums = 80 时，
输出结果为：

```
正确分类的数量: 31
错误分类的数量: 177
正确分类率 0.14903846153846154
训练迭代次数 5
测试迭代次数 4
```

当主题数量 topic_nums = 150 时，
输出结果为：

```
正确分类的数量: 33
错误分类的数量: 175
正确分类率 0.15865384615384615
训练迭代次数 5
测试迭代次数 4
```

当主题数量 topic_nums = 300 时，
输出结果为：

```
正确分类的数量: 16
错误分类的数量: 192
正确分类率 0.07692307692307693
训练迭代次数 7
测试迭代次数 4
```

B.　　以词为单位：

当主题数量 topic_nums = 20 时，
输出结果为：

```
正确分类的数量: 120
错误分类的数量: 88
正确分类率 0.5769230769230769
训练迭代次数 6
测试迭代次数 5
```

当主题数量 topic_nums = 40 时，
输出结果为：

```
正确分类的数量: 164
错误分类的数量: 44
正确分类率 0.7884615384615384
训练迭代次数 7
测试迭代次数 4
```

当主题数量 topic_nums = 80 时，
输出结果为：

```
正确分类的数量: 171
错误分类的数量: 37
正确分类率 0.8221153846153846
训练迭代次数 6
测试迭代次数 5
```

当主题数量 topic_nums = 150 时，
输出结果为：

```
正确分类的数量: 179
错误分类的数量: 29
正确分类率 0.8605769230769231
训练迭代次数 6
测试迭代次数 5
```

当主题数量 topic_nums = 300 时，
输出结果为：

```
正确分类的数量: 68
错误分类的数量: 140
正确分类率 0.3269230769230769
训练迭代次数 6
测试迭代次数 4
```

以字和词为单位分类，在不同主题数下的分类效果：

	Topic_nums				
	20	40	80	150	300
以字为单位	0.1009	0.1346	0.1490	0.1586	0.0769
以词为单位	0.5769	0.7884	0.8221	0.8605	0.3269

对于本问题，通过上述表格可以看出以词为单位的分类正确率远大于以字为单位的分类正确率，而随着主题数 topic_nums 增大，分类正确率会先升高后降低。

由于以字为单位进行分类时，语料库都属于武侠小说，故文本之间的相似程度很高，分类正确率低；而以词为单位进行分类时，可以得到每本小说的人名、地名、招术名等，故分类正确率会高。

主题数量在一定范围内增大时，可以更好地划分文本之间的差异，故正确率提高；而当主题数量过多时，模型严重过拟合，故分类正确率降低。

Conclusions

LDA 主题模型可以对文本进行建模并分类，对于主题不同类的文本应具有较好的分类能力；对同一类型的文本，以词为基本单元分类选择合适的主题数量，也能具有较好的分类能力。