第 2 次作业报告

Siyuan Jin
20373680@buaa.edu.cn

Abstract

本报告探讨了利用 LDA 模型进行文本分类的方法，首先概述了 LDA 模型的基本原理，并详细描述了使用 LDA 模型进行文本分类的流程。报告进行了三个实验，分别是比较不同主题个数下的分类性能、比较以词和字为基本单元的分类性能差异，以及比较不同段落长度下的分类性能。实验结果表明，适当增加主题数、以词为基本单元以及增加段落长度可以显著提升分类性能。总体而言，LDA 模型可以有效地进行文本分类，并且通过调整相关参数可以获得更好的分类效果。

Introduction

LDA 模型

LDA（Latent Dirichlet Allocation）是一种用于主题建模的概率图模型 [1]。它用于发现文档集合中隐藏的主题结构。LDA 假设每个文档可以被看作是不同主题的混合，而每个主题则是由不同单词的分布组成的。通过分析文档中出现的词语并推断它们属于哪些主题，LDA 可以揭示文本数据中的潜在主题信息。

LDA 的基本思想是：对于给定的文档集合，它假设每个文档都是通过从一组主题分布中随机选择主题，并从每个主题的词语分布中随机选择词语而生成的 [2]。换句话说，LDA 将文档生成的过程想象为一个双重的随机过程，其中文档的内容由主题决定，而主题又由词语决定。

在实际应用中，LDA 可以用于诸如文本分类、主题分析、信息检索等任务。通常，使用 LDA 模型的步骤包括数据预处理（如去除停用词、词干提取等）、构建词袋模型或 TF-IDF 矩阵、指定主题数量、训练模型、以及解释和评估结果等。

LDA 被广泛应用于自然语言处理领域，是理解大规模文本数据中主题结构的有力工具之一。

Methodology

M1：数据获取和预处理

1. 导入语料文件；
2. 对语料库的数据去除停用词（可选）；
3. 从语料库中均匀抽取 1000 个段落作为数据集（每个段落有 K 个 token）；
4. 随机打乱抽取到的段落数据集，并按照 9：1 分为训练集和测试集。

M2：利用 LDA 模型和 SVM 分类器对文本建模

1. 指定主题数为 T，利用经过预处理的训练集作为训练样本训练 LDA 模型；
2. 利用训练样本的主题分布和特征向量训练一个线性 SVM 分类器；

3. 通过训练好的分类器得到训练样本预测标签，与真实标签比较得到文本分类准确率。

M3：通过实验对比不同参数对实验结果的影响

1. 比较在设定不同的主题个数 T 的情况下，分类性能的变化；
2. 比较以“词”和“字”为基本单元下的分类性能差异；
3. 比较在设定不同的段落的 token 数 K 的情况下，分类性能的变化。

Experimental Studies

M1：在设定不同的主题个数 T 的情况下，分类性能变化

实验设定：K=1000；比较主题数 T=10,50,100,150,200,250,300,350,400,450,500 的分类性能变化。

实验结果如下：

表 1: 实验 1 结果统计表

实验序号	主题数	基本单元	每段字/词数	是否去除停用词	训练集准确率	测试集准确率
1	10	词	1000	是	24.8%	18.4%
2	50	词	1000	是	47.2%	47.8%
3	100	词	1000	是	48.7%	44.5%
4	150	词	1000	是	49.0%	33.6%
5	200	词	1000	是	58.1%	56.5%
6	250	词	1000	是	62.1%	44.5%
7	300	词	1000	是	59.5%	45.6%
8	350	词	1000	是	59.6%	43.4%
9	400	词	1000	是	63.8%	59.7%
10	450	词	1000	是	63.3%	52.2%
11	500	词	1000	是	68.2%	51.1%

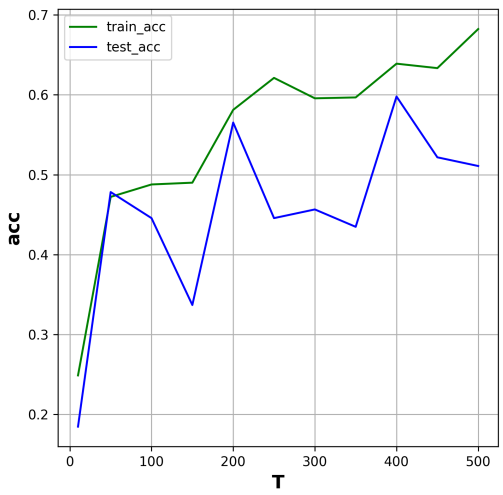

图 1: 实验 1 训练结果

M2：比较以“词”和“字”为基本单元下的分类性能差异

实验设定：段落长度 K==1000，主题数 T=200，比较以“词”和“字”为基本单元下的分类性能差异。

实验结果如下：

表 2: 实验 2 结果统计表

实验序号	段落长度	基本单元	主题数	是否去除停用词	训练集准确率	测试集准确率
1	1000	字	200	是	57.5%	56.5%
2	1000	词	200	是	62.7%	61.2%

M3：设定不同的段落 token 的情况下，分类性能的变化

实验设定：T=200；比较 K=20,100,500,1000,3000 的分类性能变化。

表 3: 实验 3 结果统计表

实验序号	段落长度	基本单元	主题数	是否去除停用词	训练集准确率	测试集准确率
1	20	词	200	是	33.3%	27.1%
2	100	词	200	是	36.5%	38.0%
3	500	词	200	是	49.6%	44.5%
4	1000	词	200	是	55.8%	59.7%
5	3000	词	200	是	68.7%	68.4%

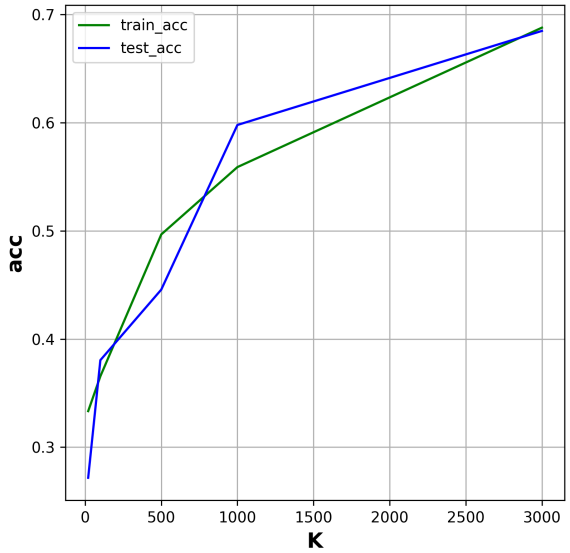

图 2: 实验 3 训练结果

Conclusions

通过实验可以得到以下结论：1) 适当增加主题数可以显著提升分类效果，但超过一定范围后因为过拟合导致效果提升有限；2) 以词为基本单元的分类效果明显优于以字为基本单元；3) 增加

段落长度有助于显著提升分类效果。

参考文献

[1] 陈运文. 一文详解 LDA 主题模型 — zhuanlan.zhihu.com. https://zhuanlan.zhihu.com/p/31470216, 2018. [Accessed 09-05-2024].

[2] 通俗理解 LDA 主题模型. https://blog.csdn.net/v_july_v/article/details/41209515, 2024. [Accessed 09-05-2024].