Report of Deep Learning for Natural Langauge Processing

Tuo Zhao
975352925@qq.com

Abstract

本报告对包含 16 部小说的中文语料库抽取 1000 个段落作为数据集，利用 LDA 模型在给定的语料库上进行文本建模，主题数量为 T，并把每个段落表示为主题分布后进行分类，分类器选择随机森林，分类结果使用 10 次交叉验证。通过实验结果讨论了以下问题：（1）在设定不同的主题个数 T 的情况下，分类性能的变化。（2）以"词"和以"字"为基本单元下分类结果的差异。（3）以"词"和以"字"为基本单元下分类结果的差异。

Introduction

LDA，即 Latent Dirichlet Allocation（潜在狄利克雷分配），是一种生成式概率模型，用于发现文档集合中的主题结构。它是一种无监督学习方法，最初由 David Blei、Andrew Ng 和 Michael I. Jordan 于 2003 年提出。LDA 的基本思想是假设每个文档包含多个主题，而每个主题又由多个单词组成。在模型中，每个主题都是一个概率分布，表示单词的集合，而每个文档则是在这些主题上的概率分布。LDA 的目标是根据给定的文档集合，推断出主题的分布以及每个文档的主题分布。LDA 模型的推断过程通常使用变分推断或 Gibbs 采样等方法来实现。通过这些推断方法，可以得到文档集合中的主题结构，进而进行文本主题分析、文本摘要、信息检索等任务。LDA 主题模型不关心文档中单词的顺序，通常使用词袋特征（bag-of-word feature）来代表文档。LDA 模型认为主题可以由一个词汇分布来表示，而文章可以由主题分布来表示。所以想要生成一篇文章，可以先以一定的概率选取上述某个主题，再以一定的概率选取那个主题下的某个单词，不断重复这两步就可以生成最终文章。

随机森林（Random Forest）是一种集成学习方法，用于解决分类和回归问题。它由多个决策树组成，每棵树都是独立生成的，而且树中的特征选择是随机的。在进行分类时，随机森林会输出每棵树的分类结果，并通过投票或者平均值来确定最终的分类结果。在生成每棵树时，随机森林会随机选择一部分特征进行节点分裂，这有助于减少过拟合，并提高模型

的泛化能力。由于每棵树都是在随机选择的特征子集上进行训练的，因此随机森林对于过拟合的抵抗能力很强。

本文利用 LDA 模型在给定的语料库上进行文本建模并把每个段落表示为主题分布后进行通过随机森林方式进行分类，完成了本次实验。

Methodology

具体来说，分类流程有以下步骤：

1. 准备语料库：选择一个中文文本作为实验的语料库。本报告中，将 16 部金庸小说进行合并得到完整文本。
2. 文本预处理：对语料库进行预处理。删除所有的隐藏符号、非中文字符和标点符号。使用结巴分词和字符级标记化对文本进行预处理。
3. 提取段落：从预处理后的文本中提取段落，构建训练集和测试集，并将其转换为词袋表示。
4. 训练LDA模型：使用训练集来训练LDA模型以提取主题特征
5. 验证性能：使用随机森林分类器对主题特征进行分类，并评估模型的性能。

Experiment Results

Task1：不同主题个数T对于分类性能的影响

本次实验主题数量设置如下：5，20，100，500。在 token 为 1000 的条件下，分别以 word 和 char 为基本单元进行实验。实验结果如下表：

	5	20	100	500
word	0.32	0.48	0.69	0.70
char	0.50	0.80	0.89	0.85

表 1 不同主题数量下的分类准确率

从表 1 中可以看出，以 word 为基本单元，分类准确率均随主题数量的增加而增加。以 char 为基本单元时，Topic 达到 100 时准确率最高，之后随着 Topic 的增加，准确率进行下

降。分析原因可能如下：增加特征数量通常会增加模型的复杂度，从而有助于提高模型的性能。在 LDA 模型中，增加主题数量可以理解为增加了特征数量，因为每个主题都可以被视为一个特征，反映了文档集合中的不同主题或话题。当主题数量增加时，模型也更有可能捕捉到文档中隐藏的信息或语义，从而提高了对文本特征的表达能力，有助于改善分类性能。但过大的主题数量也可能引入过拟合的风险，进而降低分类性能，特别是在数据量较少或噪声较多的情况下。

Task2：以"词"和以"字"为基本单元下分类结果差异

本次实验主题数量设置如下：5，20，100，500。Token 设置如下：20，100，500，1000。分别以 word 和 char 为基本单元进行实验。实验结果如下，横向为 Token 变化，纵向为 Topic 变化：

word	20	100	500	1000
5	0.06	0.10	0.23	0.32
20	0.07	0.11	0.24	0.48
100	0.07	0.11	0.41	0.69
500	0.13	0.14	0.50	0.70

表 2 以词为基本单元下分类准确率

char	20	100	500	1000
5	0.09	0.20	0.39	0.50
20	0.09	0.24	0.72	0.80
100	0.10	0.36	0.76	0.89
500	0.14	0.29	0.57	0.85

表 3 以字为基本单元下分类准确率

从表2和表3中可以看出，在相同实验条件下，以字为基本单元的分类准确率均高于以词基本单元的分类准确率。分析可能原因如下：以字为基本单元时，每个字符都被视为一个独立的特征，这可能更能捕捉到词语中的细微差异和语义信息，尤其是对于一些较短的文本片段或者词语组合。相比之下，以词为基本单元时，每个词被视为一个特征，可能会丢失一些字符级别的信息，导致特征表达不够细致。

Task3：短文本和长文本在主题模型性能上的差异

从表2和表3中可以看出，在Topic不变的情况下，随着Token的增加，以字和词为基本单元的分类准确率均增加。分析可能原因如下：随着Token数量的增加，可以提供更多的信息用于训练分类模型。以字为基本单元时，每个字符都被视为一个特征，而以词为基本单元，每个词被视为一个特征。增加Token数量会增加特征的丰富程度，从而更好地描述文本的内容和特征。同时可以涵盖更多的语义信息和上下文信息，有助于提高模型对文本的理解和分类能力。特别是在以词为基本单元时，增加Token数量会涵盖更多的词语组合和语义信息，从而提高模型的性能。

Conclusion

本次实验结果表明，通常情况下以字为基本单元的分类准确率大于以词为基本单元的分类准确率。Topic 数量在一定范围内增加时，分类准确率会增加，但需避免 Topic 过多所带来的过拟合问题。Token 数量在一定范围内增加时，分类准确率会增加。

Reference

[1] https://blog.csdn.net/weixin_41168304/article/details/122389948