Report of Deep Learning for Natural Langauge Processing

Weida Chen

908715799@qq.com

Abstract

本研究使用金庸的十六部经典武侠小说作为语料库，通过 LDA 模型对文本进行建模，并将每个段落表示为主题分布以进行分类。本文探讨了设定不同主题个数的影响，并比较了以“词”和以“字”作为基本分析单元对分类性能的影响。此外，本研究还评估了不同段落长度（短文本和长文本）对模型性能的具体影响。

Introduction

LDA 模型是一种用于文本分析的概率模型，它最早由 Blei 等人在 2003 年提出，旨在通过对文本数据进行分析，自动发现其隐藏的主题结构。被广泛应用于文本挖掘、信息检索、自然语言处理等领域。基于 LDA 模型，本次实验将要研究以下几个问题。

从给定的语料库中均匀抽取 1000 个段落作为数据集（每个段落可以有 K 个 token，K 可以取 20，100，500，1000，3000），每个段落的标签就是对应段落所属的小说。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（分类器自由选择），分类结果使用 10 次交叉验证（i.e.900 做训练，剩余 100 做测试循环十次）。实现和讨论如下问题：

（1）在设定不同的主题个数 T 的情况下，分类性能是否有变化？

（2）以"词"和以"字"为基本单元下分类结果有什么差异？

（3）不同的取值的 K 的短文本和长文本，主题模型性能上是否有差异？

Methodology

本实验首先基于训练集文本数据训练 LDA 主题模型，获取每部小说的主题分布；然后通过训练好的 LDA 模型，将训练集的文本转换为主题概率分布，这些分布作为分类器的输入特征向量；最后使用从 LDA 模型得到的特征向量训练分类器，以便将这些向量映射到对应的标签。

M1: LDA 模型构建

隐含狄利克雷分配（Latent Dirichlet Allocation，LDA）是一种统计模型，用于发现文档集合中隐藏的主题信息。它是一种无监督的机器学习和自然语言处理技术，广泛应用于文本挖掘和文本分析领域。

LDA 模型的核心思想是将文本表示为一组概率分布，其中每个文档由多个主题混合而成，每个主题又由多个单词组成。LDA 模型的基本原理是先假设一个文本集合的生成过程为：首先从主题分布中随机选择一个主题，然后从该主题的单词分布中随机选择一个单词，重复上述过程，直到生成整个文本。具体来说，LDA 模型的生成过程包括以下三个步骤：

1）对一篇文档的每个位置，从主题分布中抽取一个主题；每个文档的主题分布遵循一个狄利克雷分布。

2）从上述被抽到的主题所对应的单词分布中抽取一个单词；这个分布也遵循一个狄利克雷分布。

3）重复上述过程直至遍历文档中的每一个单词。

在数学上，LDA 可以表述为：设α和β是狄利克雷分布的参数。每个文档d有一个主题分布θ_d，$\theta_d \sim Dir(\alpha)$；每个主题$k$有一个词分布$\phi_k$，$\phi_k \sim Dir(\beta)$。对于每个文档中的每个词$w_{d,n}$，首先选择一个主题 $z_{d,n} \sim Multinomial(\theta_d)$，然后从这个主题对应的词分布中选择一个词 $w_{d,n} \sim Multinomial(\phi_{z_{d,n}})$。

LDA 模型提供了一个强大的框架来处理自然语言文本中的隐含结构，但它也有局限性，比如对参数选择敏感，且在处理非常短的文本时效果不佳。尽管如此，它仍然是文本分析中一种非常有价值的工具。

M2：随机森林分类器

随机森林是一种强大的机器学习算法，属于集成学习方法的一种，它通过组合多个决策树的预测结果提高整体性能和准确性。随机森林算法能够用于分类和回归任务，并且因其出色的准确性、鲁棒性和易用性而广泛应用于各种领域。

随机森林由多棵决策树构成。每棵树都是在数据集的一个随机子集上训练得到的，这种技术称为自助聚合。在构建每棵树时，随机森林算法还会随机选择一部分特征进行分裂，这增加了模型的多样性，减少了过拟合的风险，并提升了模型的准确性。对于分类问题，随机森林分类器的预测结果是基于所有树的预测结果进行投票或多数决的。每棵树给出一个预测结果，整个模型的输出将是最多树同意的类别。

本次实验中，设定决策树的数量为 100，其它参数均使用默认值。

M3：文本分类整体方法

本研究的整体方法如下：

1. 数据清洗：以金庸 16 部小说集为语料库，首先进行数据清洗，除去停用词和非中文字符。若以词为单位，则还需基于 jieba 库对文本进行分词。
2. 抽取段落：首先统计每篇小说的长度，然后将每篇小说的长度与语料库总字符数的比值作为采样权重，并根据这些权重预先将小说分成段落。此后根据采样权重随机选择小说，并确保选择不重复的段落。最后，将选定的段落的段落编号、文章标签和段落内容等信息保存下来。
3. LDA 模型训练：首先根据步骤 2 中抽取的段落构建字典和语料库。然后，基于 gensim 库中内置的 LDA 模型对语料库进行训练，提取主题特征向量作为段落的表示。
4. 随机森林分类器训练：使用随机森林分类器对段落进行分类，并使用十折交叉验证方法计算分类器的平均准确度。

Experimental Studies

本实验设定 LDA 主题个数 T 依次为 10、20、50 和 100，段落长度 K 依次为

20、100、500 和 1000，依次以字、词为单位。在验证集上得到的分类平均准确率如下表所示。

表 1：基于以字为单位的验证集的分类器平均准确度

T \ K	20	100	500	1000
10	0.116	0.238	0.541	0.673
20	0.124	0.295	0.708	0.803
50	0.125	0.319	0.760	0.880
100	0.132	0.357	0.765	0.895

表 2：基于以词为单位的验证集的分类器平均准确度

T \ K	20	100	500	1000
10	0.114	0.179	0.308	0.394
20	0.121	0.162	0.335	0.512
50	0.124	0.145	0.321	0.651
100	0.144	0.155	0.415	0.744

根据以上表格可以总结出以下规律。

1. 设定不同的主题个数 T，分类性能会有一定程度的变化。在一定范围内，T 越大分类器的分类性能越好。这可能是因为随着主题数量的增加，从小说中提取的语义特征的表达能力越强，进而被分类器正确分类的概率越大。
2. 设定不同的段落长度 K，分类性能会有显著的差异。在一定范围内，K 越大分类器的分类性能越好。这可能是因为随着段落长度的增加，LDA 模型可以观察到更多单词的组合和上下文信息，使得模型更容易捕捉文本中隐藏的主题结构和关联性。这种差异对于基于 LDA 模型的文本分类性能的影响是显著的。
3. 分别以词和以字为基本单元，分类结果差异较大。对比分类结果可以发现，以字为单元的分类性能远远高于以词为单元。其中，以字为单元的分类准确率最高能够达到 90%左右，而以字为单元的分类准确率最高仅有 75%左右。这可能是因为基于字的主题分布更能反映出各小说之间在语言风格和文学特征上的不同之处，而基于词的主题分布可能受到文学特征的影响较小。

Conclusions

本实验基于 LDA 模型，实现了金庸小说的语段分类任务。实验中，首先进行数据清洗，去除停用词和非中文字符，并利用 jieba 库对文本进行分词。接着，通过统计每部小说的长度并根据其占语料库总字符数的比例确定采样权重，以此权重将小说分成不同的段落。采样时根据这些权重随机选择段落，并记录段落编号、文章标签及其内容。此后，使用 gensim 库的 LDA 模型基于这些段落训练得到主题特征向量。最后，利用随机森林分类器对段落进行分类，并通过十折交叉验证评估分类器的平均准确度。

本实验还探究了不同主题数量 T、不同段落长度 K 以及选定不同基本单元（字或词）对文本分类性能的影响。实验结果表明，在其它条件不变时，T、K 越大，模型的分类性能越好；且以字为基本单元的分类性能普遍高于以词为基本单元。

References

[1] https://zhuanlan.zhihu.com/p/658568949