Report of Deep Learning for Natural Langauge Processing

ZY2303217 郑茗畅
zy2303217@buaa.edu.cn

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。LDA模型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

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。其核心思想是利用多个弱学习器（决策树），通过集成它们的预测结果，构建一个更加强大和稳健的模型。这个过程包括三个关键步骤：首先是随机采样。随机森林从训练数据集中随机抽取多个不同的子数据集，这样每棵决策树都会在一个独立的数据子集上进行训练，增加了模型的多样性和泛化能力。接着是构建决策树。针对每个数据子集，随机森林构建一棵决策树。在构建的过程中，通常会采用随机选择特征子集来进行节点分裂，这样可以增加每棵树的独特性，进一步提高整体模型的泛化能力。最后是投票或平均。对于分类问题，每棵决策树都会给出一个类

别，而对于回归问题，每棵决策树都会给出一个预测值。随机森林通过投票（分类问题）或平均（回归问题）的方式，汇总所有决策树的预测结果，得到最终的分类或回归结果。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，包括医疗诊断、金融风险评估、客户流失预测等。

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: 支持向量机分类器

支持向量机（Support Vector Machine，SVM）是一种强大的监督学习算法，用于分类和回归分析。其主要思想是在特征空间中找到一个最优的超平面，将不同类别的数据分隔开来。具体来说，SVM的关键概念包括：首先是超平面，它是一个\(N-1\)维的线性子空间，对于二维空间就是一条直线，对于三维空间就是一个平面。在高维空间中，它是一个超平面。SVM的目标就是找到一个最优的超平面，使得两个不同类别的数据点到这个超平面的距离尽可能地远，从而实现良好的分类。其次是支持向量，它是离超平面最近的数据点，这些数据点对确定超平面起着关键作用。支持向量机的决策边界由这些支持向量完全决定，因此它们在确定分类结果上起着至关重要的作用。

SVM有不同的核函数，用于处理非线性可分的数据。常见的核函数包括线性核函数、多项式核函数和高斯核函数等，它们可以将数据映射到高维空间，从而使得在原始空间中线性不可分的问题在新的高维空间中变得线性可分。SVM的优点包括：在高维空间中有效地处理线性和非线性可分问；通过引入核函数，可以灵活地处理各种类型的数据；在处理小样本、高维度数据和非线性问题时表现良好；由于其最优化的特性，SVM对于泛化能力较强，对于数据量不大的情况也有较好的性能。支持向量机是一种强大的分类器，适用于许多不同的领域，包括文本分类、图像识别、生物信息学等，其优秀的性能和理论基础使其成为机器学习领域中的重要算法之一。

M4: 多项式朴素贝叶斯分类器

贝叶斯网络，又称贝叶斯信念网络或贝叶斯有向无环图（DAG），是一种概率图模型，用于表示一组随机变量及其条件依赖关系。它被广泛应用于机器学习和人工智能领域，用于建模不确定性并根据观察数据进行预测。贝叶斯网络由有向无环图（DAG）和条件概率分布表（CPT）组成。节点之间的有向边表示一个随机变量对另一个随机变量的条件依赖关系。通常情况下，如果节点 A 指向节点 B，则表示变量 B 在给定变量 A 的条件下的条件概率。每个节点都有一个条件概率分布表，用于描述该节点在其父节点给定的情况下的条件概率。例如，如果节点 B 有父节点 A，则节点 B 的 CPT 就描述了在给定 A 的取值情况下 B 的取值的概率分布。贝叶斯网络主要适用于以下三种条件：数据量足够大，可以准确地估计条件概率分布表；变量之间的依赖关系是稳定的，不会频繁变化。

多项式贝叶斯基于原始的贝叶斯理论，但假设概率分布是服从一个简单多项式分布。多项式分布来源于统计学中的多项式实验，这种实验可以具体解释为：实验包括n次重复试验，每项试验都有不同的可能结果。在任何给定的试验中，特定结果发生的概率是不变的。

M5: KNN分类器

K最近邻（KNN）是一种基本的分类和回归方法，其原理简单直观。对于一个未知样本，KNN算法会在训练集中找出与该样本最相似的K个样本，然后根据这K个样本的类别进行投票（分类问题）或者计算平均值（回归问题），来确定该样本的类别或者值。KNN主要适用于一下三种情况：由于KNN算法在预测时需要计算未知样本与所有训练样本之间的距离，因此数据集较大时计算开销会很高，不太适合使用KNN算法；KNN算法在高维空间中容易受到维度灾难的影响，因此适合处理维度较低的数据集；KNN算法假设样本分布均匀，即相似的样本在特征空间中聚集在一起。

M6: 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：语料库为金庸的 16 部小说集，因此从语料库中均匀抽取 1000 个段落作为数据集时，每部小说选取 63 个 K 词段落即可。选择时可从前往后进行选取，确保选择不重复的段落。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 sklearn 库中内置的 CountVectorizer 和 LatentDirichletAllocation 函数对语料库进行训练，提取主题特征向量作为段落的表示，将文本转换为主题分布。

4. 分类器训练：使用随机森林分类器、支持向量机分类器、多项式朴素贝叶斯分类器、KNN 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5、10、20、30、50、100、200、300、500和1000，段落长度K依次为20、100、500、1000和3000，依次以字、词为单位，依次使用随机森林分类器、支持向量机分类器、多项式朴素贝叶斯分类器、KNN分类器。在验证集上得到的分类平均准确率如下所示。

Table 1 随机森林分类器以字为单位验证集分类平均准确率

T \ K	20	100	500	1000	3000
5	0.1	0.09	0.19	0.17	0.05
10	0.03	0.08	0.09	0.19	0.19
20	0.03	0.04	0.14	0.25	0.27
30	0.04	0.03	0.18	0.19	0.4
50	0.07	0.05	0.16	0.22	0.35
100	0.06	0.08	0.15	0.18	0.75
200	0.05	0.09	0.11	0.39	0.88
300	0.1	0.08	0.2	0.41	0.63
500	0.06	0.21	0.24	0.56	0.7
1000	0.08	0.2	0.26	0.65	0.65

Table 2 支持向量机分类器以字为单位验证集分类平均准确率

T \ K	20	100	500	1000	3000
5	0.04	0.06	0.13	0.04	0.04
10	0	0.08	0.03	0.04	0.07
20	0.02	0.04	0.21	0.17	0.26
30	0.04	0.07	0.17	0.14	0.36
50	0.08	0.1	0.21	0.15	0.31
100	0.09	0.07	0.22	0.12	0.77
200	0.09	0.1	0.17	0.44	0.82
300	0.06	0.11	0.19	0.54	0.48
500	0.09	0.24	0.2	0.52	0.54
1000	0.11	0.19	0.3	0.56	0.69

Table 3 多项式朴素贝叶斯分类器以字为单位验证集分类平均准确率

T \ K	20	100	500	1000	3000
5	0.04	0.06	0.13	0.03	0.03
10	0.03	0.08	0.03	0.03	0.04

20	0.03	0.03	0.07	0.12	0.18
30	0.05	0.07	0.09	0.14	0.3
50	0.07	0.08	0.08	0.11	0.29
100	0.09	0.09	0.07	0.11	0.64
200	0.1	0.1	0.11	0.38	0.65
300	0.08	0.12	0.21	0.45	0.45
500	0.05	0.16	0.22	0.41	0.46
1000	0.09	0.2	0.27	0.58	0.65

Table 4 KNN分类器以字为单位验证集分类平均准确率

T \ K	20	100	500	1000	3000
5	0.12	0.06	0.21	0.19	0.14
10	0.09	0.08	0.1	0.19	0.16
20	0.03	0.07	0.14	0.15	0.31
30	0.09	0.06	0.09	0.16	0.46
50	0.05	0.11	0.1	0.16	0.43
100	0.11	0.12	0.1	0.2	0.67
200	0.07	0.09	0.13	0.51	0.78
300	0.06	0.17	0.25	0.49	0.48
500	0.1	0.16	0.28	0.56	0.57
1000	0.06	0.19	0.27	0.58	0.69

Table 5 随机森林分类器以词为单位验证集分类平均准确率

T \ K	20	100	500	1000	3000
5	0.12	0.14	0.42	0.53	0.69
10	0.1	0.23	0.58	0.67	0.89
20	0.11	0.34	0.71	0.75	0.97
30	0.14	0.33	0.76	0.86	0.98
50	0.11	0.42	0.81	0.9	0.98
100	0.11	0.46	0.85	0.88	0.92
200	0.12	0.45	0.77	0.77	0.9
300	0.1	0.46	0.71	0.83	0.92
500	0.03	0.47	0.61	0.83	0.94
1000	0.04	0.31	0.76	0.76	0.9

Table 6 支持向量机分类器以词为单位验证集分类平均准确率

K / T	20	100	500	1000	3000
5	0.1	0.18	0.36	0.45	0.56
10	0.12	0.2	0.54	0.7	0.85
20	0.13	0.31	0.68	0.8	0.96
30	0.13	0.39	0.83	0.88	1
50	0.11	0.4	0.81	0.93	1
100	0.13	0.42	0.77	0.89	0.99
200	0.13	0.46	0.85	0.88	0.99
300	0.1	0.48	0.77	0.86	0.99
500	0.03	0.43	0.74	0.73	0.96
1000	0.03	0.36	0.79	0.67	0.8

Table 7 多项式朴素贝叶斯分类器以词为单位验证集分类平均准确率

K / T	20	100	500	1000	3000
5	0.07	0.18	0.35	0.47	0.48
10	0.05	0.18	0.47	0.66	0.79
20	0.07	0.3	0.67	0.8	0.98
30	0.09	0.32	0.75	0.87	1
50	0.09	0.38	0.76	0.92	1
100	0.08	0.45	0.82	0.89	0.99
200	0.09	0.46	0.82	0.86	1
300	0.06	0.48	0.76	0.84	0.97
500	0.03	0.44	0.75	0.78	0.99
1000	0.03	0.36	0.83	0.77	0.83

Table 8 KNN分类器以词为单位验证集分类平均准确率

K / T	20	100	500	1000	3000
5	0.07	0.2	0.38	0.47	0.65
10	0.09	0.24	0.52	0.62	0.86
20	0.1	0.29	0.58	0.8	0.98
30	0.09	0.28	0.73	0.89	0.99
50	0.1	0.29	0.76	0.88	1
100	0.1	0.29	0.65	0.84	0.98
200	0.09	0.28	0.71	0.75	0.94
300	0.13	0.29	0.63	0.76	0.94

500	0.1	0.25	0.45	0.68	0.82
1000	0.1	0.29	0.64	0.6	0.68

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。然而超过范围后，T 继续增大，分类器的分类性能反而变差。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。而主题数量过多，分类问题复杂性较大，分类效果变差。因此 T 选择 30-200 之间较为合适。
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。因此 K 选择 3000 最为合适。
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，以字为基本单元的分类性能低于以词为基本单元。这可能是因为基于词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，而基于字的主题分布可能受到文学特征的影响较小。
4. 在以字为基本单元进行分析时，随机森林分类器的表现效果最好，在验证集上得到的分类平均准确率最高可达 0.88。在以词为基本单元进行分析时，支持向量机分类器、多项式朴素贝叶斯分类器和 KNN 分类器均可在合适的 K、T 取值下在验证集上得到的分类平均准确率达到 1。

Conclusions

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。然而超过范围后，T 继续增大，分类器的分类性能反而变差。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。而主题数量过多，分类问题复杂性较大，分类效果变差。因此 T 选择 30-200 之间较为合适。
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。因此 K 选择 3000 最为合适。
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，以字为基本单元的分类性能低于以词为基本单元。这可能是因为基于词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，而基于字的主题分布可能受到文学特征的影响较小。
4. 在以字为基本单元进行分析时，随机森林分类器的表现效果最好，在验证集上得到的分类平均准确率最高可达 0.88。在以词为基本单元进行分析时，支持向量机分类器、多项式朴素贝叶斯分类器和 KNN 分类器均可在合适的 K、T 取值下在验证集上得到的分类平均准确率达到 1。

References

[1] https://www.jianshu.com/p/09bc46ffdac6
[2] https://www.cnblogs.com/liuyihai/p/8309019.html
[3] https://www.cnblogs.com/hjk-airl/p/16457435.html
[4] https://blog.csdn.net/D802366y/article/details/108366499
[5] https://blog.csdn.net/autocyz/article/details/46786469