Report of Deep Learning for Natural Langauge Processing

Yiming Xu
1953999172@qq.com

Abstract

本文对金庸的16篇小说进行了语料分析，运用 LDA（Latent Dirichlet Allocation）模型进行建模，以此构建文本分类器，旨在探究不同主题数量、分词单位（词与字）以及段落长度对分类性能的影响。

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

Part1：LDA模型建模

LDA（Latent Dirichlet Allocation，潜在狄利克雷分配）是一种统计模

型，用于发现大规模文档集合中的隐藏主题结构。它假设文档是由一系列不可见的主题生成的，而每个主题又由一组词语的概率分布所定义。LDA 通过概率统计方法对文档集进行建模，旨在揭示文档间的潜在主题关系以及主题内部的词汇关联。

文档（Document）：LDA 处理的对象是文档集合，每个文档由一系列词语组成。

主题（Topic）：主题是文档中词汇出现模式的抽象概括，由一组词语及其对应的概率分布构成。例如，一个关于“科技”的主题可能包含“AI”、“机器学习”、“算法”等词语，每个词语都有其在该主题下的概率。

词袋（Bag of Words）：在 LDA 模型中，文档被视为无序的词袋，即忽略词语的顺序和语法关系，仅关注文档中出现的词语及其频次。

生成过程（Generative Process）：对于每个文档从全局主题分布中抽取主题比例向量 θ （Dirichlet 分布）对于文档中的每个词语：从文档的主题比例向量 θ 中抽取一个主题 z（多项式分布）从该主题对应的词语分布 φ 中抽取一个词语 w（多项式分布）

参数估计（Parameter Estimation）：LDA 模型的实际应用中，我们已知的是文档集合 D，但 θ、φ 和 z 都是未知的。通常使用 Gibbs Sampling、Variational Inference 或者其他近似推断方法来估计这些隐变量。

主题表示（Topic Representation）：对于每个文档，LDA 模型输出其主题分布 θ，即该文档在各个主题上的概率分布。这样，每个文档就可以被表示为其主题分布，而不是原始的词序列。

本研究采用定量分析方法，通过构建 LDA 主题模型并结合文本分类技术，对文学作品段落进行分类，以探究主题数量、分词单位以及段落长度对分类性能的影响。

Part2：SVM建模

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。

Experimental Studies

LDA 模型用于对抽取的段落进行主题建模。具体步骤如下：

1) 文本预处理：对每个段落进行分词（按词或字），形成词袋表示。

2) LDA 模型训练：使用 LDA 模型对预处理后的文本数据进行训练，指定主题数量 T。训练完成后，每个段落将得到一个 T 维的主题分布向量，表示该段落在各个主题上的概率权重。

3) 分类器训练：将每个段落的主题分布向量作为特征，其所属小说的标签作为目标变量，使用你选择的分类器进行训练。这里，分类器的任务是学习如何根据段落的主题分布将其正确分类到对应的小说类别中。

4) 交叉验证与性能评估：使用 10 次交叉验证对分类器进行评估，每次保留 10%的数据作为测试集，其余 90%作为训练集。计算每次交叉验证的分类性能指标（如准确率、F1 分数等），并分析不同主题数量 T、分词单位（词或字）以及段落长度（K 值）对分类性能的影响。

Model	Tokens	Test Accuracy
SVM(Char)	20	0.1941
SVM(Char)	100	0.1994
SVM(Char)	500	0.2002
SVM(Char)	1000	0.1986
SVM(Char)	3000	0.1986
SVM(Words)	20	0.1969
SVM(Words)	100	0.1990
SVM(Words)	500	0.1942
SVM(Words)	1000	0.1981
SVM(Words)	3000	0.1976

通过对比两种模型的实验结果，我们可以看到在某些情况下，SVM（Words）模型在测试准确率上略高于 SVM（Char）模型，尤其是在最大标记数为 20 时。然而，在其他最大标记数下，两种模型的性能差异不太明显，而且它们的准确率在不同最大标记数下都有相似的波动范围。要深入理解模型的性能和泛化能力，可能需要进一步的实验和分析。

Conclusions

通过本次实验，我们能够深入理解 LDA 用于发现大规模文档集合中的隐藏主题结构。