Report of Deep Learning for Natural Language Processing

谢承罡 SY2406119

xiechenggang@buaa.edu.cn

Abstract

本研究究基于 LDA 主题模型与带高斯核的 SVM 分类器，探讨了主题数量、基本单元和段落长度对中文小说文本分类性能的影响。通过 10 次交叉验证实验，发现主题数量增加显著提升了分类性能，但存在饱和点，尤其是在长文本中，主题数量和准确率的关系存在先增加再轻微下降。段落长度与分类性能呈正相关，长文本因信息丰富，显著提升了分类准确率。研究结果为 LDA 模型在文本分类中的参数优化提供了实证支持。

Introduction

LDA 模型作为一种无监督学习方法，在文本分类任务中具有广泛应用，但其性能受到主题数量、文本长度和基本单元选择等因素的显著影响。本研究以中文小说为研究对象，通过系统分析这些关键因素对分类性能的作用机制，旨在为 LDA 模型的参数配置提供实践指导。与此同时，带高斯核的支持向量机（SVM）作为经典的监督学习算法，在分类和回归任务中展现出独特优势。该算法通过高斯核函数将原始数据映射到高维特征空间，有效解决了非线性可分数据的分类难题，能够找到最优的分类超平面。高斯核函数的带宽参数决定了映射的复杂度，使其能够灵活适应不同的数据分布特征。得益于强大的非线性建模能力，高斯核

SVM 在图像分类、文本分类等领域取得了优异的表现，但其性能对参数选择较为敏感，且计算复杂度相对较高，这些特点在实际应用中需要特别注意。

Methodology

本研究从 16 部金庸小说中均匀抽取 1000 个段落构建实验数据集，每个段落以其所属小说作为标签，并设置 20、100、500、1000 和 3000 个 token 五种不同长度，以确保数据具有充分的多样性和代表性。在数据处理环节，首先对文本进行清洗，然后根据"字"或"词"的基本单元选择进行分词处理，同时过滤停用词以降低噪声干扰。实验设计方面，采用 LDA 模型生成主题分布，通过设置 5、10、20、50 和 100 五个不同主题数，将每个段落转换为对应的主题概率分布向量，作为后续分类的特征输入。为全面评估模型性能，研究采用 10 折交叉验证方法，每次迭代使用 90%的数据训练高斯核 SVM 模型，剩余 10%用于测试，最终通过逻辑回归分类器计算平均准确率及其标准差，从而系统评估不同参数配置下的分类效果及其稳定性。

Experimental Studies

从表 1 中可以看出，段落中 token 数量对分类准确率有显著影响，段落 token 数量越多，分类准确率总体呈现上升趋势，表明长段落因包含更丰富的语义信息，能够更好地支持分类任务。同时，主题数量对分类准确率的影响存在一个饱和值，在饱和值以下，随着主题数量的增加，分类准确率逐步提升，这是因为更多的主题能够更好地捕捉文本的多样性；然而，当主题数量超过饱和值后，准确率不再显著提升，甚至可能下降，过多的主题会引入噪声，导致模型过拟合，从而影响分类效果。因此，选择适中的主题数量和较长的段落长度是优化分类性能的关键。

Table 1 “word”分类准确率

Args	ACC
K=20 T=8	0.0390
K=20 T=24	0.0320
K=20 T=48	0.0440
K=100 T=8	0.1150
K=100 T=24	0.0600
K=100 T=48	0.0690
K=500 T=8	0.1590
K=500 T=24	0.1730
K=500 T=48	0.1730
K=1000 T=8	0.2830
K=1000 T=24	0.3110
K=1000 T=48	0.3300
K=3000 T=8	0.4520
K=3000 T=24	0.6560
K=3000 T=48	0.6850

从表 2 中可以观察到，段落长度与分类准确率之间存在显著的正相关关系。随着段落中 token 数量的增加，分类性能呈现持续提升的趋势，这表明较长的文本片段能够提供更丰富的语义信息和上下文特征，从而为分类任务带来更可靠的判别依据。在主题数量方面，研究发现了典型的"先升后稳或降低"变化规律：这种现象揭示了过多的主题可能导致特征空间过度细分，不仅增加了计算复杂度，还可能引入噪声干扰，最终影响模型的泛化能力。这些发现提示我们，在实际应用中应当综合考虑段落长度和主题数量的平衡，通常选择中等偏长的段落配合适中的主题数量，可以在保证分类效果的同时避免模型过拟合的风险。

Table *2* “char”分类准确率

Args	ACC
K=20 T=8	0.0760
K=20 T=24	0.0840
K=20 T=48	0.0660
K=100 T=8	0.1780
K=100 T=24	0.1950
K=100 T=48	0.1730
K=500 T=8	0.3810
K=500 T=24	0.5400
K=500 T=48	0.5630
K=1000 T=8	0.5320
K=1000 T=24	0.7180
K=1000 T=48	0.7710
K=3000 T=8	0.6710
K=3000 T=24	0.8580
K=3000 T=48	0.8830

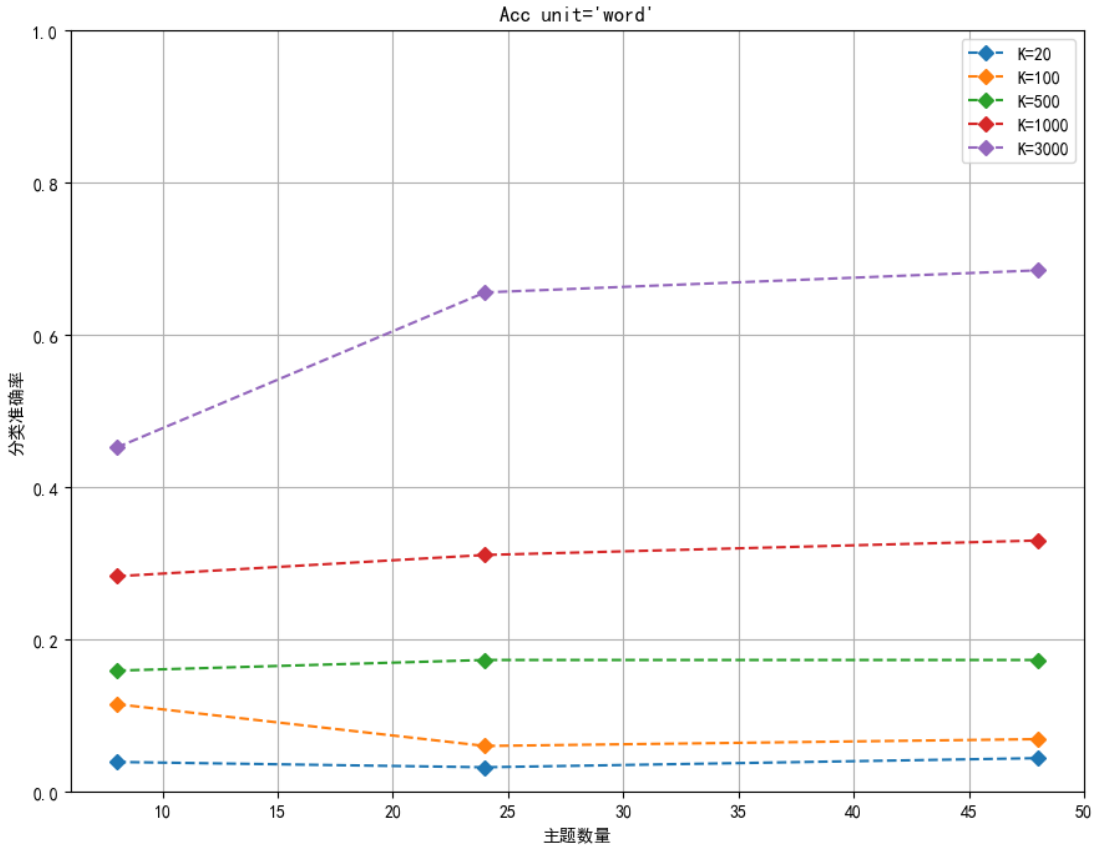

Figure *1* “word”分类准确率

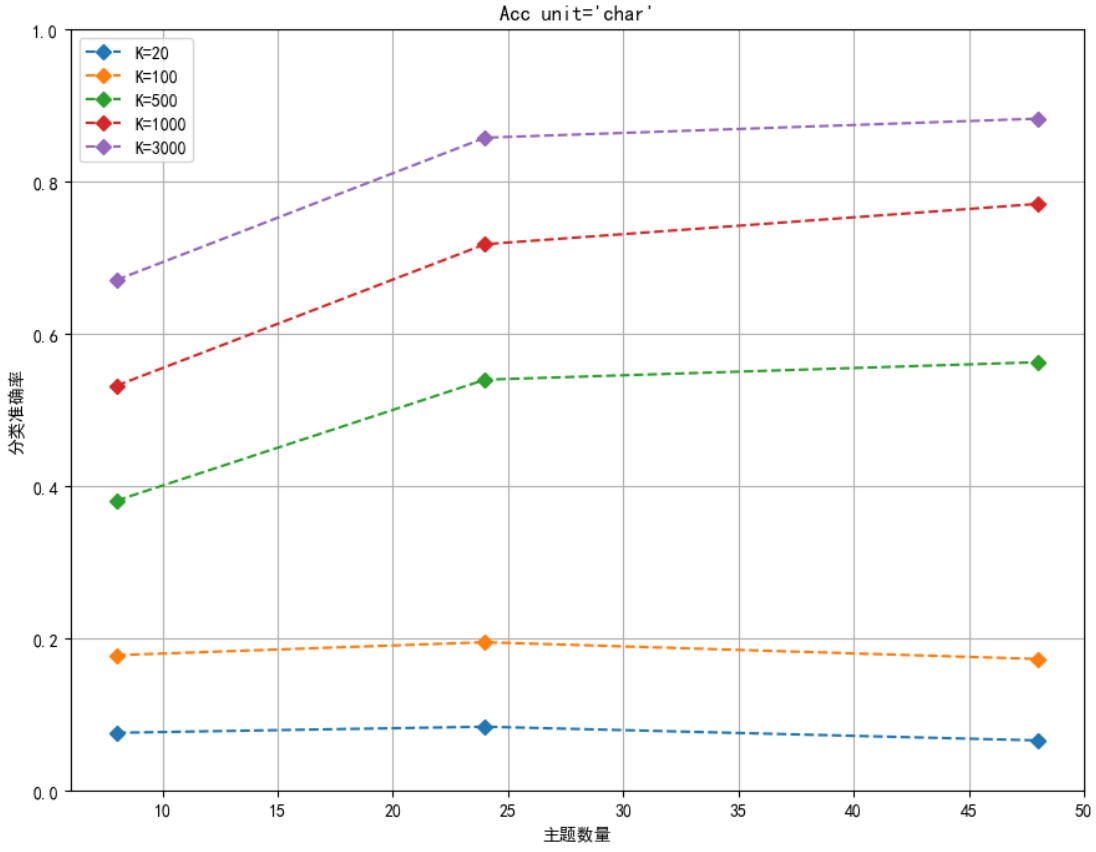

Figure *2* “char”分类准确率

Conclusion

研究表明，在 LDA 分类任务中，段落长度与分类准确率呈正相关，较长的文本片段因其更丰富的语义信息而显著提升分类效果。同时，主题数量存在最优区间：在达到临界值前，增加主题数量能通过更细致的文本表征提高准确率；但超过该值后，准确率增长停滞甚至下降，过多的主题会引入噪声导致过拟合。因此，采用较长段落并选择适中的主题数量是优化分类性能的有效策略。