作业报告二

Weijia Long
wjial@buaa.edu.cn

Abstract

从语料库中均匀抽取 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，分类结果使用 10 次交叉验证。针对在设定不同的主题个数 T 的情况下，分类性能的变化，以"词"和以"字"为基本单元下分类结果的差异以及不同的取值的 K 的短文本和长文本，主题模型性能的差异开展实验。除此之外，本实验额外测试随机森林、SVM、朴素贝叶斯三种分类器下的模型差异。

Part 1

针对主题个数 T 分别设置为 5，10，15，20，25，30，50, 100，200，300，1000，3000 进行实验，在设定 k 为 3000 下以及使用随机森林分类器，并以字为基准的情况下观察分类性能的变化。

Experimental 1

在设定 k 为 3000 下以及使用随机森林分类器的情况下分类性能的变化如下表所示

主题个数	训练准确度	测试准确度
5	0.35	0.28
10	0.36	0.42
15	0.42	0.52
20	0.47	0.43

25	0.42	0.47
30	**0.49**	0.41
50	0.47	**0.53**
100	0.45	0.52
200	**0.49**	0.44
300	0.42	0.48
1000	0.37	0.39
3000	0.44	0.36

表 1. 不同主题个数下的分类性能表现

可以看出随着主题个数的增加，训练准确度和测试准确度在增高，但在高于 50 以后，准确度反而会下降，因此 T 选择 30-50 之间较为合适。

Part 2

针对以字和词的基础上分别进行实验，设定主题数 T 为 50，k 为 3000，使用随机森林分类器的基准下观察分类性能的变化。

Experimental 2

在主题数 T 为 50，k 为 3000，使用随机森林分类器的情况下分类性能的变化如下表所示

字/词	训练准确度	测试准确度
词	0.13	0.12
字	**0.47**	**0.53**

表 2. 不同基本单元下的分类性能表现

可以看出以字为基本单元准确度较高。

Part3

针对不同的取值的 K 的基础上分别进行实验，K 的取值分别为 20，100，500，1000，3000. 设定主题数 T 为 50，基本单元为字，使用随机森林分类器的基准下观察分类性能的变化。

Experimental 3

在主题数 T 为 50，基本单元为字符，使用随机森林分类器的情况下分类性能的变化如下表所示

K	训练准确度	测试准确度
20	0.22	0.29
100	0.30	0.30
500	0.37	0.42
1000	0.40	0.52
3000	**0.47**	**0.53**

表 3. 不同文本长度的分类性能表现

可以看出随着 k 的增加，训练准确度和测试准确度在增高，因此 K 选择 3000 最为合适。

Part4

针对不同分类器分别进行实验，设定 K 为 3000，主题数 T 为 50，基本单元为字，使用随机森林分类器的基准下观察分类性能的变化。

Experimental 3

在 K 为 3000，主题数 T 为 50，基本单元为字符，使用不同分类器的情况下分类性能的变化如下表所示

分类器	训练准确度	测试准确度
随机森林	**0.47**	**0.53**
SVM	0.24	0.37
朴素贝叶斯	0.25	0.33

表 4. 不同分类器下模型的分类性能表现

可以随机森林分类器表现最好，因此选择随机森林最为合适。