Report of Deep Learning for Natural Langauge Processing

BaobaoHeZY2354234
hebaobao@buaa.edu.cn

Abstract

本文通过对金庸的 16 部小说进行语料分析，均匀抽取 1000 个段落作为数据集（每个段落有 K 个 token,K 分别选取 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使用 SVM 进行分类，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）。讨论不同的主题个数 T 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。

Introduction

LDA 由 Blei,DavidM.、Ng,AndrewY.、Jordan 于 2003 年提出，用来推测文档的主题分布。它可以将文档集中每篇文档的主题以概率分布的形式给出，从而通过分析一些文档抽取出它们的主题分布后，便可以根据主题分布进行主题聚类或文本分类。

LDA 采用词袋模型。所谓词袋模型，是将一篇文档，我们仅考虑一个词汇是否出现，而不考虑其出现的顺序。在词袋模型中，“我喜欢你”和“你喜欢我”是等价的。与词袋模型相反的一个模型是 n-gram，n-gram 考虑了词汇出现的先后顺序。

支持向量机（SVM）中的线性 SVM：只有当数据完全线性可分时，我们才能使用线性 SVM。完全线性可分意味着可以使用一条直线（如果是二维）将数据点分为 2 类。

Experimental Studies

在设定不同的主题个数 T 的情况下，分类性能如表 1 所示：分类准确度随着主题个数增加而增加。

表 1

K	T	Classifier	Analyzer	Training Accuracy	Test Accuracy
3000	10	SVM	Word	0.112122347	0.07
3000	25	SVM	Word	0.117690387	0.08
3000	50	SVM	Word	0.091872659	0.07
3000	100	SVM	Word	0.107640449	0.11
3000	5	SVM	char	0.208526841	0.2
3000	10	SVM	char	0.229787765	0.21
3000	25	SVM	char	0.230973783	0.29
3000	50	SVM	char	0.269051186	0.29
3000	100	SVM	char	0.244444444	0.31

在以"char"和以"word"为基本单元下，分类性能如表 2 所示：以“char”为单位的分类准确性优于“word”。

表 2

K	T	Classifier	Analyzer	Training Accuracy	Test Accuracy
3000	5	SVM	char	0.208526841	0.2
3000	5	SVM	Word	0.122147316	0.06
3000	10	SVM	char	0.229787765	0.21
3000	10	SVM	Word	0.112122347	0.07
3000	25	SVM	char	0.230973783	0.29
3000	25	SVM	Word	0.117690387	0.08
3000	50	SVM	char	0.269051186	0.29
3000	50	SVM	Word	0.091872659	0.07
3000	100	SVM	char	0.244444444	0.31
3000	100	SVM	Word	0.107640449	0.11

不同的取值的 K 的短文本和长文本，分类性能如表 3 所示：总体来看，K 越大，分类性能越好。

表 3

K	T	Classifier	Analyzer	Training Accuracy	Test Accuracy
20	100	SVM	Word	0.124382022	0.06
100	100	SVM	Word	0.134419476	0.09
500	100	SVM	Word	0.14906367	0.11
1000	100	SVM	Word	0.136616729	0.07
3000	100	SVM	Word	0.107640449	0.11
20	100	SVM	char	0.232059925	0.19
100	100	SVM	char	0.232072409	0.25
500	100	SVM	char	0.287041199	0.34
1000	100	SVM	char	0.32062422	0.32
3000	100	SVM	char	0.244444444	0.31

Conclusions

增加主题数量，使用“char”作为单位，增加段落长度通常能提升分类性能，提高分类准确性。

References

[1] https://zhuanlan.zhihu.com/p/31470216

[2] https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/