优达学城文档归类开题报告

2018 年 4 月 9 日

肖志铭

1 课题背景

随着统计自然语言处理的发展，机器学习已成为处理自然语言问题的常见技术。而文本分类问题作为自然语言处理中最常见、最基础的问题之一，也早已有了多种基于机器学习技术的解决方案。在本项目中，我将尝试在这个问题上做一些相对深入的研究，锻炼自己运用机器学习技术的能力，并同时加深对 NLP 领域的了解。

2 问题描述

文本分类是指使用电脑对文本集按照一定的标准进行的自动分类。它的应用相当广泛，比如垃圾过滤,新闻分类等。当前主要有两类文本分类方法，一类是基于知识工程技术，还有一类便是基于统计和机器学习技术，在本项目中使用到的分类方法就属于第二类。

3 数据集

本项目中使用的数据集为 20 newsgroups 数据集（它的下载地址请参看 README.txt）。该数据集收集了近 20000 条新闻组文档，均匀的分为 20 个不同主题的新闻组集合。其中一些新闻组的主题比较相似，还有一些却几乎无关。这样的特性能有效的测试文本分类方法的性能，在主题无关和主题类似两种不同情况下的分类效果。这个数据集共有三个不同的版本，分别为：第一个完全没有修改过的版本，第二个按时间顺序分为训练(60%)集和测试(40%)集两部分的版本（不包含重复文档和新闻组名），以及第三个不包含重复文档，且只有来源和主题的版本。根据下载页面的推荐，本文将使用第二个版本的数据集作为分类器的训练集和测试集。

4 方案描述

本项目中的文本分类方法基于统计原理和机器学习技术实现。大体思路是首先通过 word2vec 得到数据集中每个词的词向量，再结合 tf-idf 得到每篇文档的向量表示。最后将文档向量作为特征，使用朴素贝叶斯或 SVM 等监督学习方法，训练得到一个用于文本分类的模型。

5 基准模型

词袋法是最为经典的文档表示方法之一，原理简单却高效实用。不过相比 word2vec，它存在维度容易过高、无法表达词语之间关系等缺点。在此将使用词袋模型作为基准，与使用 word2vec 的模型作对比。根据对比结果，考察两种模型的文本分类性能。

6 评估标准

对于本模型的评估，除了准确率外还需要考虑召回率，因此使用 F1 值来均衡衡量以上两个指标。F1 值的定义如下：

$$F1 = 2 * accuracy * recall/(accuracy + recall)$$

其中 accuracy 代表了准确率，而 recall 代表了召回率。

本项目中使用的数据集有多个子类别。对项目中使用到的子类别，须分别计算它们的 F1 值。并在此基础上，计算它们的宏平均值。

7 项目设计

该项目的工作流程分为以下几步：

(1) 数据的获取和预处理。本项目将直接通过 sklearn 获取 20 newsgroups 数据集，并进行必要的数据预处理。

(2) 建立基准模型。根据词袋法抽取特征，并以此训练文本分类器，作为基准模型。

(3) 获取 word2vec 词向量。该步将使用 gensim 来完成，得到文档中词语的 word2vec 向量表示。

(4) 训练基于 word2vec 词向量的文本分类器。基于上一步得到的 word2vec 词向量，结合 tf-idf，得到文档的向量表示。并以此进行训练，得到新的文本分类器。

(5) 模型对比。设定评估指标，分析分类器的性能，和基于词袋法的基准模型进行对比。

(6) 参数调优。调试分类器参数，尝试获取更佳的性能，并分析和评价参数调整的效果。

以上是我的文本分类项目的大体流程，希望通过该项目，加深自己在机器学习上的理论认知，并增强自己在机器学习上的实践能力。