机器学习纳米学位

毕业项目：文档分类

优达学员 关力宁
2018.05.11

开题报告

项目背景

自然语言处理（NLP）是机器学习技术重要应用范畴之一，而文档分类应该是自然语言处理中最普遍的一个应用，例如文章自动分类、邮件自动分类、垃圾邮件识别、用户情感分类等等，因此一个好的文档分类程序有非常大的应用意义。想要开发一个号的文档分类程序，主要需要解决两个问题，即文档特征提取和分类算法模型。文档特征提取方法目前有基于统计学的 ngram 模型、基于词向量的词袋子模型、Word2Vec 等模型，而分类模型可以应用传统监督学习中常用的分类算法，如决策树、支持向量机等，近年来基于深度学习神经网络的算法来进行文档分类预测的效果更好，如 CNN。

自己选择这个项目的主要原因是自己曾经参与过一个新闻推送系统，系统中就有对新闻进行分类的模块，当时是由大数据和算法部门的同事负责的，自己那时对这个分类的原理就很好奇，现在终于自己有机会用学到的知识也实现一个文档分类程序了。

问题陈述

本项目要解决的问题是如何设计好一个文档多分类程序，即针对输入的文档，程序可以预测出该文档的分类标签，并且预测的准确率要较高，该程序开发完成后是可重复使用的。此问题是可以通过选择合适的文档特征提取模型以及分类器模型来解决的。

数据集和输入

本项目将使用经典的 20 类新闻包作为数据集，这个新闻包里面大约有 18846 条新闻，比较均衡地分成了 20 类，是比较常用的文本数据之一，数据集可通过 sklearn 工具包下载，并且 sklearn 中的工具包已经切分好了训练集和测试集。

通过程序分析，在类别分布方面，数据集在这 20 个分类下基本成均匀分布，只有 3 个种类的新闻数少于 500，其他种类的新闻数均在 500 到 600 区间；在文档长度分布上，数据集成正偏态分布，大部分文档的单词数小于 500；通过观察，在文档内容上，一些文档中包含能明显确定分类的数据，如 alt.atheism 分类的 49960 文档，第二行 Subject 中包含了分类信息 Alt.Atheism，因此在清洗数据步骤要移除文档中这种对分类导向性过强的数据，以及文档中一些对分类预测没有相关性的头尾部内容。除此，还要做单词统一为小写、去除标点符号、去除停用词等工作。最后，我将用训练集对模型进行训练，用测试集测试模型。

解决方法

第一步，我将对每篇文档内容做预处理；第二步，我将分别使用词袋子模型和 Word2Vec 模型来构建文档的表示方式；第三步，我将分别使用训练集数据训练决策树、朴素贝叶斯、支持向量机、神经网络 CNN 四种模型，并用测试集测试训练效果；最后，我会依据准确率及训练时间做出我最后的结论。

基准模型

我将以 tf-idf 词袋模型构建文档的表示、以决策树模型作为分类模型的组合作为基准模型，基准模型的预测准确率大致在 0.65 到 0.67。这样选择的原因是决策树模型易于理解和实现，并且训练速度快，是监督学习中进行分类预测的常用方法，而词袋子模型适用于传统机器学习方法，并且也有易于理解和实现的优点。

评价标准

我将以模型分类预测的准确率作为主要评价标准（原因参见参考文献 5），准确率的定义是对于给定的测试数据集，分类器正确分类的样本数与总样本数之比，此项目的准确率算法就是分类器正确分类的文档数与总样本数之比。如果准确率得分相近，将参考训练时间作出综合评价。

项目设计

1. 获取数据。通过在网站直接下载的方式获取数据集（网站地址在参考资料 3）
2. 清洗数据。我会用 numpy、sklearn 等工具包统一文档单词为小写形式、统一单词为单数形式、去掉标点符号、去掉停用词、去掉特殊符号，以此来提高数据质量。
3. 特征提取。我会分别采用词袋子模型和词向量模型(Word2Vec)构建文档的特征表示。
4. 模型训练。我会用经词袋子模型处理的训练集对决策树、朴素贝叶斯、支持向量机进行训练，经 Word2Vec 处理的训练集对深度学习 CNN 模型进行训练，并用测试集分别进行测试。
5. 模型评估。最后根据准确率以及训练时间评估基准模型和其他三种模型，得出最优的文档分类程序。
6. 表达方式。为了更好的表达我的研究报告，我将会在各个步骤进行图形或表格的可视化处理。

参考资料

[1]https://blog.csdn.net/ahmanz/article/details/51273500

[2]http://www.cnblogs.com/platero/archive/2012/12/03/2800251.html

[3]http://www.qwone.com/~jason/20Newsgroups/

[4]https://www.cnblogs.com/sddai/p/5696870.html

[5]http://www.docin.com/p-1448732393.html

[6]https://zhuanlan.zhihu.com/p/33925599

[7]https://blog.csdn.net/Scotfield_msn/article/details/72904092

[8]https://download.csdn.net/download/qq_33394807/9934498