机器学习工程师纳米学位

毕业项目：文档分类

2017.04.30

开题报告

项目背景

自然语言处理是计算机科学领域与人工智能领域中的一个重要方向。在自然语言处理处理面临很多挑战，其中一项任重的任务就是文本分类。文本分类一般包括了文本的表达、 分类器的选择与训练、 分类结果的评价等过程。本项目的研究任务是探究如何使用合理的自然语言处理技术进行文档分类。

本项目将利用20新闻组数据[1]，探索不同的文本表示方式，其中包括词袋子模型[2]，利用词向量模型（如word2vec[3]、glove[4]等）构建对文本的表示。并尝试使用不同的机器学习模型，如逻辑回归、SVM和神经网络等，利用上述不同的文本表示对文档进行分类。

问题描述

本项目目的就是利用上述自然语言处理技术结合所学机器学习知识对文档进行准确分类。需要利用不同的文本表示方式来表示每篇文档（如词袋子模型、词向量模型），并用所采用的文档表示结合机器学习算法对文档进行分类。将以分类准确率作为评价标准进行衡量。

数据或输入

该项目使用使用经典的20类新闻包，里面大约有20000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。该数据可利用*sklearn*工具包[5]下载。对于新闻数据，文本具有清晰的结构和规范的语法，我们使用sklearn中的函数自动过滤掉标题、注脚和引用（由于这些内容与分类问题具有较强的关联性，而本项目的目的在于关注文档内容对分类的影响）。同时还将使用标准的文本预处理步骤，如分词、去除标点符号、转换成小写、去除不常用词、词性转换等。

进行预处理之后，在每类新闻中我们将选择一部分文档作为训练集，另一部分作为测试集。

此外，针对上述提出的词向量模型的训练也需要大量数据，20类新闻数据样本量可能不足以训练出较好的词向量模型，为此我们将可以采用Mikolov曾经使用过的text8[6]数据包进行训练词向量模型。

解决方法描述

本项目将采用词袋子模型和词向量模型两种方式对文档进行表示。

- 使用词袋子模型来表示每篇文档，常见的一种思路是首先将文本进行分词，也就是将一个文本文件分成单词的集合，建立词典。这里不打算尝试多元组构成词库，因为使用多元组构建词典会使得词典非常大，构建文本的特征表示时向量维数非常大不利用学习。构建词典之后，将每篇文档表示成长度为词典大小的一维的向量，其中词典的每个词对应向量中的一个位置，其数值使用该词对应的tf-idf值，这样能有效减少常见词的权重，增加特异词的权重。从而可以得到文档的特征表示。接下来，就可以方便利用机器学习分类模型进行训练。

- 利用Word2Vec方式即词向量模型表示每篇文档，这里面包含两部分主要工作：
    - 利用文本数据使用skip-gram的方法[3]进行训练，将每个单词表示成向量形式。词向量训练后需要进行简单评测，比如检验一些单词之间相似性是否符合逻辑等。为此将是用T-SNE技术将词投影在二维平面进行可视化。观察相邻词之间是否意思相近。
    - 如何将词向量转化成整个文本的表示需要进一步探讨。这里将采用两种方式。
        - 将文档中的词向量进行平均得到整个文档的向量表示。t同样地，使用机器学习分类模型进行训练。
        - 直接将文本表示成词向量序列，利用神经网络模型（如LSTM，1d-CNN）处理词向量序列，进行分类。

在获得文本的表示方式之后，我们将利用机器学习模型训练数据进行分类。采用的机器学习模型有：逻辑回归、SVM和神经网络。

基准模型

基准模型将使用词袋子模型构建文档表示，并使用多分类问题的logistic regression作为分类模型。该基准模型能够清晰的定义并容易实现，可作为后续复杂模型的参照。

评估标准

该项目要求比较不同算法表示方式和不同算法在测试集上的分类准确率。以分类准确率作为评估标准。

项目设计

- 获取数据。并对文本数据数据进行预处理，出去新闻文档的标题、注脚和引用。对单词的大小写、同一词不同形式（如单复数）进行统一、并去除标点符号。
- 探讨文本的表示方式：
    - 构建词袋子模型
    - 训练词向量模型，并利用词向量构建对文本的表示。
- 探讨不同的机器学习模型在不同的文本表示上的表现效果。
- 以测试集上的分类准确率作为评估标准，比较不同的模型和算法。

参考文献

[1] http://www.qwone.com/%7Ejason/20Newsgroups/

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

[3]http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

[4]https://nlp.stanford.edu/pubs/glove.pdf

[5]http://scikit-learn.org/stable/datasets/twenty_newsgroups.html

[6]http://mattmahoney%5B.NET%5D(http://lib.csdn.net/base/dotnet/dc/text8.zip