文档归类-项目报告

# 问题的定义

## 项目背景

随着互联网时代的蓬勃发展，人们积累了大量数字文本文档，如何有效的组织这些数据并从中进行信息挖掘，成为一个课题。最早的文本分类研究可以追溯到20世纪六十年代[1]，当时还以规则处理为主。近些年，伴随着计算机计算能力的大幅提升，以及统计机器学习和深度学习的快速发展成型，文本分类的方法也一直在变化。

文档归类，也称文本分类，指对文本数据集按照一定的分类体系或标准进行分类的标记。文本分类问题与其它分类问题没有本质上的区别，其方法可以归结为根据待分类数据的某些特征来进行匹配，选择最优的匹配结果，从而完成文档所属的一个或多个分类的标记[2]。

本项目将利用机器学习算法，从文档中抽取特征，完成文档的自动归类工作。将采用Usenet的20新闻组分类数据[3]进行模型的训练和验证。

## 问题描述

20新闻组分类问题是一个有监督多分类问题。每篇新闻属于且仅属于一个类别。通过对每篇文档进行特征提取，基于训练集的标注信息，对模型进行训练，并在测试集上验证模型效果，最终得到一个有效的模型可以对输入的文档完成正确分类信息的输出。模型将通过预测的分类结果的准确率来进行评估。

特征提取方面，将尝试TfIDF[4]，词向量[5]以及字符特征等方式来建模。模型方面，将尝试朴素贝叶斯，决策树，SVM以及深度学习等分类模型完成分类工作。

最终，会得到一个有效的模型，输入一篇新闻文档，模型可以准确预测出该文档所属的类别。

## 评价指标

通过数据类别分析发现，各个类别的数据量相对比较均衡。因此，模型将基于准确率指标进行效果分析，即

yi 是第i个样本的真实分类标签，是第i个样本的预测分类标签，如果真实分类与预测相同，则I=1，否则I=0

另外，模型的训练时间也会是一个评估指标。

# 分析

## 数据的探索

20新闻组数据可以通过sklearn提供的函数sklearn.datasets.fetch\_20newsgroups直接下载，总共有18846个文档。按照函数提供的训练集、测试集分割方式，即基于某个日期进行数据的划分，训练集有11314个文档，测试集有7532个文档。训练集会进一步按20%的比例分割出验证集。训练集将用于模型的训练。验证集用于模型调参。测试集用于模型最终效果的评估。每个文档独立成为一个文件，类似电子邮件格式，包含头部，正文和尾部三个区块，头部包含发件人、标题等信息，正文是具体内容以及引用的内容，尾部为一些署名信息。

正文内容看起来如下：

|  |
| --- |
| Subject: Re: Lexan Polish?  From: jeff@mri.com (Jonathan Jefferies)  Expires: Sun, 8 Aug 1993 07:00:00 GMT  Organization: Microtec Research, Santa Clara, California, USA  Keywords: Lexan, Plastic  Summary: Scratches in Plastic  Lines: 27  In article <C41soE.M62@ns1.nodak.edu> wilken@plains.NoDak.edu (Scott Wilken) writes:  >A couple of years ago I replaced the stock windscreen on my Interceptor  >with a higher one from National Cycle. The thing happens to be made of  >Lexan.  >  >Can anyone recommend a polish to use on it that is safe for lexan? Its  >starting to show a few scratches, and id like to polish them out..  >Go FAST! | Internet: wilken@plains.nodak.edu | AMA #587126  >Take Chances! | UUCP: ..!uunet!plains!wilken | DoD #0087  >VF700F Interceptor | Bitnet: WILKEN@PLAINS |  Suggest McQuires #1 plastic polish. It will help somewhat but nothing  will remove deep scratches without making it worse than it already is.  McQuires will do something for fine or light stuff.  Also suggest calling your local plastic shop. In Calif. "TAP PLASTIC" is  a chain that carries most of what is needed for repair and sometimes  replacement of plastic bits. Telephone in the Bay area is 415-962-8430.  I'm not sure how amenable they are to shipping. I have found that they  have several excellent products for cleaning, and removing crap from  windscreens and face shields. Also they have one called "lift-it" which  works real well in removing sticky stuffs such as adhessives from plastic  wihtout scratching same.  Luck,  Jonathan Jefferies, jeff@mri.com |

这篇文章属于rec.motorcycles类目，它提到repair，plastic等词，跟摩托车有些关联。可以看到，头部信息应该只有标题（Subject）值得保留，发信人信息可能导致模型错误的建模，比如部分用户只活跃在某些类目，导致模型习得该特征。引用信息中，第一行的引用信息也应该予以删除，避免类似的习得错误特征的问题。类似的，尾部的署名信息也应该删除。

所有的20个分类名称如下，涵盖从计算机，到科学，汽车，运动，政治等各种题材。

alt.atheism, comp.graphics, comp.os.ms-windows.misc, comp.sys.ibm.pc.hardware, comp.sys.mac.hardware, comp.windows.x, misc.forsale, rec.autos, rec.motorcycles, rec.sport.baseball, rec.sport.hockey, sci.crypt, sci.electronics, sci.med, sci.space, soc.religion.christian, talk.politics.guns, talk.politics.mideast, talk.politics.misc, talk.religion.misc

## 探索性可视化

首先针对训练集和测试集的类别分布做可视化。



图表 2‑1 训练集类别标签分布



图表 2‑2 测试集类别标签分布

从图表 2‑1和图表 2‑2可以观察到，数据在各个类别的分布较为平均，个别类别数据量稍少，因此采用前面的准确率公式可以有效评估模型性能。

## 算法和技术

特征提取方面，会尝试使用TFIDF，词向量，字符等特征来建模。

采用TFIDF来进行特征提取，主要是考虑该方法对于在某篇文章中出现频率高，但是在所有文档中出现次数低的词给予较高的权重，也就是倾向于筛选出文章具有代表性的关键词。

词向量（Word2vec）特征可以压缩语义空间，保留语义信息的同时降低特征维度，可以使模型具备一定的语义理解能力，对于未登录词，使用0初始化特征。

字符特征则进一步简化特征维度，并通过深度模型自动加工提取特征。针对英语字母表，最终选择了69个字符来作为特征表示，包括26个小写字母，10个数字，32个其他字符和一个换行符，采用独热编码[10]进行建模。

在基于词向量的特征提取方案下，考虑到训练集样本规模有限，不足以训练语言模型，将引入text8[6]数据集使用word2vec[7]来训练词向量，该数据集为英文维基百科经过清洗和截断构造而成，每个单词均为小写并空格分隔，一共一行，截断到100MB大小而成。另外，也会使用基于维基百科训练的Glovec向量[8][9]。

模型方面，由于TFIDF是词袋模型，没有考虑文本顺序信息，因此无法用于CNN等基于位置信息建模的模型，另外，此项目是文章分类，属于长文本，RNN在这种场景下也不适合直接应用。最后，词向量和字符特征把文档表示为2维特征，也无法应用于传统机器学习模型。

综上，最终计划采用如下的方案来实现文档归类：

|  |  |  |
| --- | --- | --- |
| 序号 | 特征提取 | 模型 |
|  | TFIDF | 决策树[11] |
|  | TFIDF | 朴素贝叶斯[12] |
|  | TFIDF | SVM[13] |
|  | 词向量 | CNN (TextCNN[14]) |
|  | 词向量 | CNN+LSTM (C-LSTM [15]) |
|  | 字符 | CNN (Char-CNN [16]) |

其中，决策树，朴素贝叶斯和SVM模型将使用sklearn中的实现，深度模型将使用keras+Tensorflow来完成。

在LSTM之前，我们先介绍另一个如雷贯耳的名词RNN(Recurrent Neural Networks)，中文名是递归神经网络。LSTM是对RNN隐含层的改进。RNN最主要的用途是处理序列数据，且该序列的当前输出与前面的输出也有关。比如，要分析一段文本的含义时，当前词汇与上一个词汇是有关系的，并不是孤立存在的，上文对当前的输出产生了影响。因此RNN在NLP（Natural Language Processing）中获得非常成功的应用。

LSTM因为《Long short-term memory》这篇论文为世人所熟知，LSTM的出现是为了解决RNN无法有效处理长期记忆（Long Term Dependencies）的问题。

标准RNN在进行长期记忆计算会产生‘梯度消失或者梯度膨胀问题’（原作者描述为“vanishing or exploding gradients”），即RNN的梯度数据会在若干次递归计算后，由于反向多次求导的递归过程中产生的计算误差的叠加，导致梯度消失或者梯度膨胀。粗略的说，‘梯度消失’的表现为很多个小于1的项连乘就很快的逼近零。

LSTM是如何做到避免梯度消失，简单的说，LSTM在神经元的状态转移时，不对状态值添加scale参数或矩阵变换。同时，引入‘门函数’，如输入门，输出门，遗忘门，控制输入信息、状态转移，以及输出信息的有选择性地表达。

## 基准模型

早期的论文[17]，其中提到的PrTFIDF模型在测试集的准确率为90.3%。论文中使用了标题和正文文本，并随机分割33%的数据作为测试集，其余数据作为训练集进行训练，由于没有引入验证集调参，论文中的模型可能有过拟合的风险。因此最终选择Stanford的分类器[18]的指标，准确率81.7%，该文献也是基于时间来分割测试集，与本文的数据分隔方法一致，可以作为对比。

# 方法

## 数据预处理

对于文本的处理主要有几种方式：

大小写转换，统一转换成小写。

去除标点符号。由于是做文档分类，而标点符号主要是些断句或者情感信息，对于分类没有帮助，因此需要去掉。但是在字符特征的建模下，需要保留符号作为分隔符。

每篇文档的头（header），脚注（footer）以及引用（quote）信息，也需要选择性去除。因为，这部分元信息对于文档分类的帮助可能不大，而且容易造成过拟合（比如某些用户只活跃在部分新闻组类别下，导致模型可能会基于用户名而不是文档内容来进行分类），因此需要有选择的去除。最终，元信息保留头部的标题（Subject），以及引用信息，因为标题可能有重要的分类信息，引用文本在回复类文档中，是重要的上下文信息。

去除停用词，由于没有语义信息，也需要去除。

低频词，对于出现次数小于一定阈值的词也可以去掉，减少长尾词汇，缩减词表大小。

词干提取（stemming）与词形还原（lemmatization），对于单词的变体进行归一便于统一语义信息来完成分类。

## 执行过程

## 完善

# 结果

## 模型的评价与验证

## 合理性分析

# 项目结论

## 结果可视化

## 对项目的思考

## 需要作出的改进

# 参考文献

1. 文本分类概述. <http://blog.csdn.net/chl033/article/details/4733647>
2. Document Classification. https://en.wikipedia.org/wiki/Document\_classification
3. 20 Newsgroups. <http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html>
4. TF-IDF. https://en.wikipedia.org/wiki/Tf%E2%80%93idf
5. Word embedding. Wikipedia. <https://en.wikipedia.org/wiki/Word_embedding>
6. Text8. <http://mattmahoney.net/dc/textdata>
7. Word2vec. <https://en.wikipedia.org/wiki/Word2vec>
8. Pennington, J., Socher, R. and Manning, C.D. (2014) Glove: Global Vectors for Word Representation. Proceedings of the Empiricial Methods in Natural Language Processing.
9. Pretrained Glove 6B tokens. https://nlp.stanford.edu/projects/glove/
10. 独热编码. https://en.wikipedia.org/wiki/One-hot
11. DecisionTree. <https://en.wikipedia.org/wiki/Decision_tree>
12. Naïve Bayes Classifier. <https://en.wikipedia.org/wiki/Naive_Bayes_classifier>
13. Cortes C, Vapnik V. Support-Vector Networks[J]. Machine Learning, 1995, 20(3): 273-297.
14. Kim Y. Convolutional Neural Networks for Sentence Classification[J]. empirical methods in natural language processing, 2014: 1746-1751.
15. Zhou C, Sun C, Liu Z, et al. A C-LSTM Neural Network for Text Classification[J]. arXiv: Computation and Language, 2015.
16. Zhang X, Zhao J J, Lecun Y, et al. Character-level convolutional networks for text classification[J]. neural information processing systems, 2015: 649-657.
17. Joachims T. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization[C]. international conference on machine learning, 1997: 143-151.
18. Stanford Classifier. <https://nlp.stanford.edu/wiki/Software/Classifier/20_Newsgroups>