机器学习纳米学位

毕业项目 文档分类

关力宁 2018.05.16

1. 问题概述

1.1 项目概述

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。

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模块就很好奇，现在自己也可以用学到的知识来实现一个文档分类程序了。

本项目将使用使用经典的 20 类新闻包作为训练和测试数据集，该新闻包大约有 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，可利用 sklearn 包下载。

1.2 问题陈述

本项目要解决的是一个文档多分类问题，数据集利用 20 类新闻包，文档多分类问题即可采用传统机器学习方法，也可采用深度学习神经网络的方法解决。

文本分类的基本流程是读取数据、清洗数据、特征提取、模型训练、模型评估，难点主要在特征提取和模型训练环节。我将利用 tf-idf 词袋子模型和 Word2Vec 模型提取文档特征，利用监督学习中决策树、朴素贝叶斯、支持向量机和深度学习中 CNN 四种模型进行分类训练，并以词袋子+决策树作为基准模型比较训练结果。我希望最后能在上述分类方法中找到一个预测准确率达到 80%以上的模型。

1.3 评价指标

20 类新闻包数据集是一个分类平衡的数据集，因此我将采用准确率作为评价指标。准确率是指模型预测正确的结果所占的比例，对于本项目，计算公式为：

Accuracy = total_valid_predicition / total_sample

total_valid_predicition 是指分类预测正确文档数之和；total_sample 是指所有样本文档数之和。

2. 分析

2.1 数据的探索

分类文本数据使用经典的 20 类新闻包，利用 *sklearn* 工具包下载，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，此外 sklearn 工具包下载的数据集已经按 60%和 40%的比例分成了训练集和测试集。

下面分析两个新闻文档，baseball 分类下 99971 号：

```
From: admiral@jhunix.hcf.jhu.edu (Steve C Liu)
Subject: Baseball Stats
Organization: Homewood Academic Computing, Johns Hopkins University, Baltimore, Md, USA
Lines: 17
Distribution: usa
Expires: 5/5/93
NNTP-Posting-Host: jhunix.hcf.jhu.edu
Summary: 1992 EWB II Stats wanted

        Hello, my friends and I are running the Homewood Fantasy Baseball
League (pure fantasy baseball teams). Unfortunely, we are running the league
using Earl Weaver Baseball II with the Comm. Disk II and we need the stats
for the 1992 season. (Preferably the 1992 Major League Stat Disk) We have
the '92 total stats but EWB2 needs the split stats otherwise we have 200
inning games because the Comm. Disk turns total stats into vs. L's stats
unless you know both right and left -handed stats.

        So, if anyone has the EWB2 '92 Stat Disk please e-mail me!

_____________________________________________________________________________
|Admiral Steve C. Liu        Internet Address: admiral@jhunix.hcf.jhu.edu|
|"Committee for the Liberation and Intergration of Terrifying Organisms  |
|and their Rehabilitation Into Society" from Red Dwarf - "Polymorph"      |
|****The Bangles are the greatest female rock band that ever existed!****|
|   This sig has been brought to you by... Frungy! The Sport of Kings!    |
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

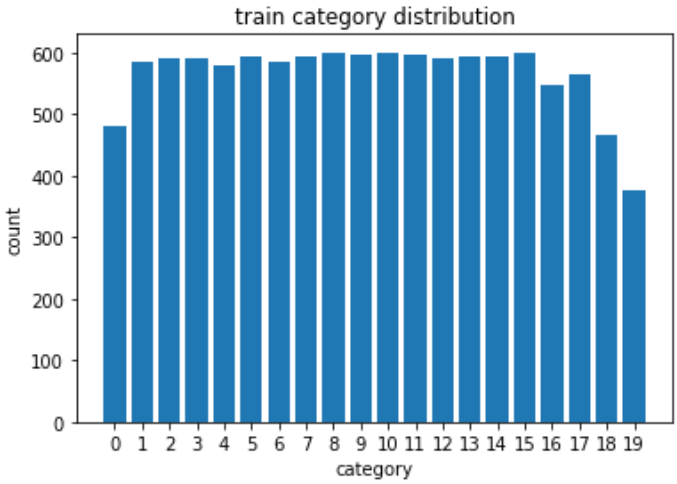
Sci.space 下新闻：

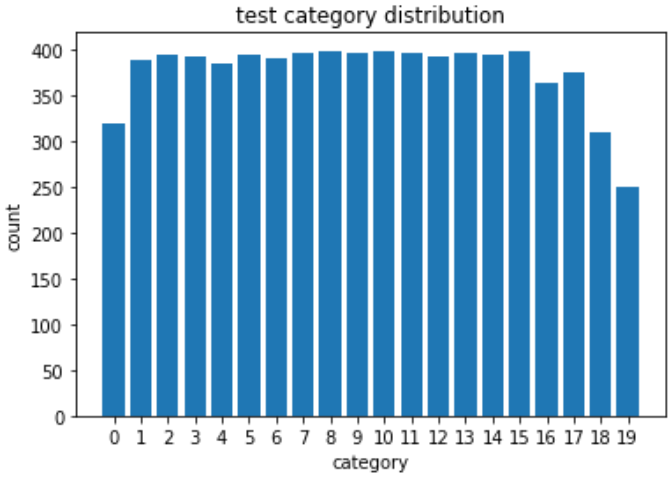
```
From: xrcjd@mudpuppy.gsfc.nasa.gov (Charles J. Divine)
Subject: Space Station Redesign Chief Resigns for Health Reasons
Organization: NASA/GSFC Greenbelt Maryland
Lines: 12

Writer Kathy Sawyer reported in today's Washington Post that Joseph Shea, the
head of the space station redesign has resigned for health reasons.

Shea was hospitalized shortly after his selection in February.  He returned
yesterday to lead the formal presentation to the independent White House panel.
Shea's presentation was rambling and almost inaudible.

Shea's deputy, former astronaut Bryan O'Connor, will take over the effort.

Goldin asserted that the redesign effort is on track.
--
Chuck Divine
```

可以看到文档结构主要分为三部分：文档头、文档内容、文档尾，有些文档没有包含尾信息，如 baseball 分类下 102585。文档头信息中有作者、主题、机构等信息，其中头部的 from、subject 和 organization 等信息和类别的关联性很高，这对于提高预测率很有帮助，而预测一个文档的分类时，头部当然也是文档的一部分，分别用去除头尾部和完整文档训练基准模型(tf-idf+决策树)，完整文档的准确率为 0.571，要比去除头尾部高出将近 7 个百分点，因此选择保留文档头尾部。

2.2 探索性可视化

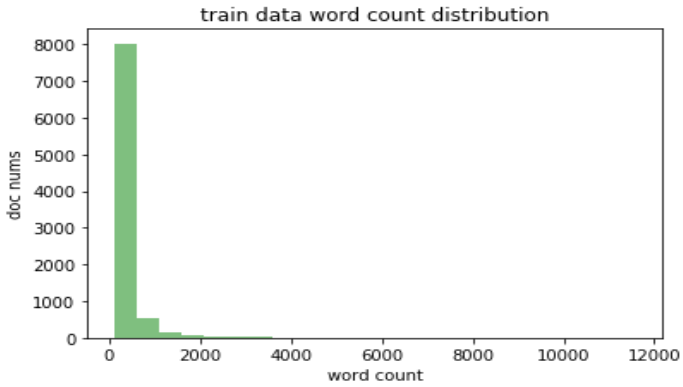
首先，利用可视化探究 20 类新闻包是否是一个分类平衡的数据集，因为这影响我是否可以将准确率作为最终评价指标。

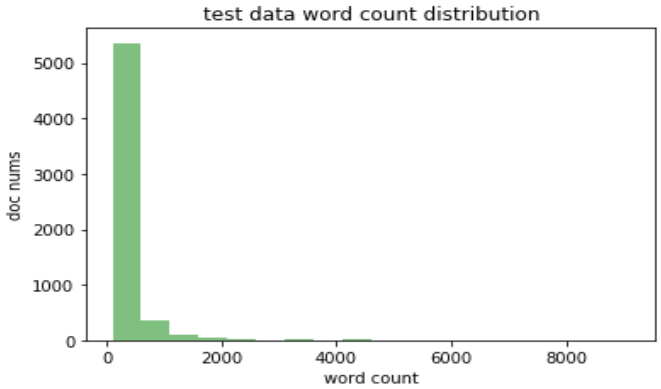




通过可视化，可以确认无论训练集还是测试集都是分类平衡的，因此可以用准确率做评价指标。

其次，因为最终是基于文档中单词进行分类预测的，所以文档中单词数也是一个很重要的因素，利用可视化来探察所有文档的单词数分布是否比较集中。





可以看出，绝大部分文档的单词数都小于 500，一小部分大于 500 小于 1000，极小比例文档单词数大于 1000。

2.3 算法和技术

- 清洗数据

  首先我将去掉文档中的 headers、footers、quotes，只保留新闻内容主体，然后去掉内容主题中的标点符号、停用词，最后将单词统一为小写形式。

- 提取文档特征
  - 词袋 tf-idf 模型

    词袋模型是一种统计某个词在一份文档中出现次数的算法。统计所得的词频数据可以用于比较文档并测量其相似性，具体应用包括搜索、文档分类、主题建模等。词频-逆文档频率（TF-IDF）是另一种根据文章中包含的词来判断文章主题的方法。TF-IDF 为词赋予权重——TF-IDF 测量的是相关性，而非频率。因此，在整个数据集中，词频都会被 TF-IDF 分值所取代。TF-IDF 会测量某一特定文档中的词的出现次数（即词频，term frequency）。出现某一个词的文档数量越多，这个词作为信号的价值就越小。这样做的目的是仅留下独特的高频词用作标记。每个词的 TF-IDF 相关性是一种标准化的数据格式，总和也是 1。

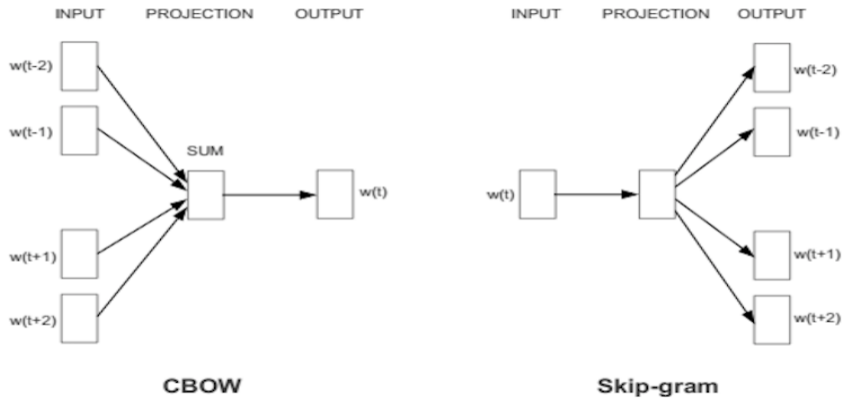
$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

- Word2Vec 模型

  Word2Vec 是一个用于处理文本的双层神经网络。它的输入是文本语料，输出则是一组向量：该语料中词语的特征向量。虽然 Word2vec 并不是深度神经网络，但它可以将文本转换为深度神经网络能够理解的数值形式。Word2vec 的目的和功用是在向量空间内将词的向量按相似性进行分组。它能够识别出数学上的相似性。

  Word2Vec 神经网络的输出是一个词汇表，其中每个词都有一个对应的向量，可以将这些向量输入深度学习网络，也可以只是通过查询这些向量来识别词之间的关系。Word2Vec 的方式有两种，一种是用上下文预测目标词（连续词袋法，简称 CBOW），另一种则是用一个词来预测一段目标上下文，称为 skip-gram 方法，我们使用 skip-gram 方法。



- 分类模型
  - 决策树

    决策树是一种预测模型，代表的是一种对象特征属性与对象目标值之间的一种映射关系。决策树分为分类树和回归树两种，分类树对离散变量做决策，输出是样本的预测类别；回归树对连续变量做决策，输出是一个实数 ，本项目应采用分类树。构造决策树的关键是如何确定树中每个节点，这里主要依赖信息熵和信息增益理论，具体实现算法有 ID3、C4.5、CART，scikit-learn 使用 CART 算法的优化版本。
  - 朴素贝叶斯

朴素贝叶斯分类是一种十分简单的分类算法，它的思想基础是这样的：对于给出的待分类项，求解在此项出现的条件下各个类别出现的概率，哪个最大，就认为此待分类项属于哪个类别。贝叶斯定理如下：

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

朴素贝叶斯分类器，计算每个特征属性划分对每个类别的条件概率：

$$P(y_i|x) = \frac{P(x|y_i)P(y_i)}{P(x)}$$

$$P(x|y_i)P(y_i) = P(a_1|y_i)P(a_2|y_i)...P(a_m|y_i)P(y_i) = P(y_i)\prod_{j=1}^{m} P(a_j|y_i)$$

- 支持向量机

  支持向量机（support vector machines，SVM）是一种分类模型，它的基本模型是定义在特征空间上的间隔最大的线性分类器，间隔最大使它有别于感知机；支持向量机还包括核技巧，这使它成为实质上的非线性分类器。简单的说就是把数据投影到高维空间进行生维，使得在低纬度无法可分的问题在高维空间变得可分。

- 深度学习 卷积神经网络(CNN)

  深度学习最初在图像和语音领域取得巨大成功，一个很重要的原因是图像和语音原始数据是连续和稠密的，有局部相关性，而传统的表示文本的 one-hot 词向量维度过高过于稀疏是不适用于神经网络的。但随着 word2vec、GloVec 等文本表达方式的提出，应用深度学习解决大规模文本分类问题又成为了可能。在应用卷积神经网络时，由于图像是二维数据，而经过词向量表达的文本是一维数据，因此此处卷积层用一维卷积，一维卷积带来的问题是需要设计通过不同 filter_size 的 filter 获取不同宽度的视野。激励层采用 relu，它收敛快，求解梯度简单。然后通过池化层进行降维，最后通过全连接层连接所有特征并把输出值传递给 softmax 分类器。应用卷积神经网络时，由于共享卷积核，对于处理高维数据压力不大，还有一个很棒的优点是无需手动选取特征。

2.4 基准模型

以词袋 tf-idf 和决策树作为基准模型，决策树易于理解和实现，是分类问题中最常见的模型，基准模型在测试集上的准确率 0.571。

3. 方法

3.1 数据预处理

利用正则表达式来移除文档中的数字、标点符号和其他非法字符，正则\W 用于匹配非数字、字母、下划线字符，\d 用于匹配数字字符；利用 nltk 包移除掉文档中的停用词。

文档处理前：

```
From: xrcjd@mudpuppy.gsfc.nasa.gov (Charles J. Divine)
Subject: Space Station Redesign Chief Resigns for Health Reasons
Organization: NASA/GSFC Greenbelt Maryland
Lines: 12

Writer Kathy Sawyer reported in today's Washington Post that Joseph Shea, the
head of the space station redesign has resigned for health reasons.

Shea was hospitalized shortly after his selection in February.  He returned
yesterday to lead the formal presentation to the independent White House panel.
Shea's presentation was rambling and almost inaudible.

Shea's deputy, former astronaut Bryan O'Connor, will take over the effort.

Goldin asserted that the redesign effort is on track.
--
Chuck Divine
```

文档处理后：

```
from xrcjd mudpuppy gsfc nasa gov charles j divine subject space station redesign chief resigns for health reasons or
ganization nasa gsfc greenbelt maryland lines writer kathy sawyer reported in today s washington post that joseph she
a the head of the space station redesign has resigned for health reasons shea was hospitalized shortly after his sele
ction in february he returned yesterday to lead the formal presentation to the independent white house panel shea s p
resentation was rambling and almost inaudible shea s deputy former astronaut bryan o connor will take over the effort
goldin asserted that the redesign effort is on track chuck divine
```

3.2 执行过程

- 获取文档特征。首先我利用 sklearn 的 TfidfVectorizer 来求得单词的 tfidf 权重值。接着我利用 20 类新闻和 text8 作为语料库分别训练出 Word2Vec 模型，对于单词 post，20 类新闻的 word2vec 模型预测出的最相近单词组是

```
[('repost', 0.576711118221283),
 ('posted', 0.5601168274879456),
 ('followup', 0.5511782169342041),
 ('posts', 0.5496430397033691),
 ('newsgroup', 0.5399307608604431),
 ('postings', 0.5394784212112427),
 ('summarize', 0.5338574051856995),
 ('comment', 0.5240219831466675),
 ('answers', 0.518371045589447),
 ('topic', 0.5101109743118286)]
```

Text8 的 word2vec 预测出的单词组是

```
[('pre', 0.4663783013820648),
 ('beaumanor', 0.4652716815471649),
 ('audial', 0.447038832932662964),
 ('office', 0.4347648024559021),
 ('reisendorf', 0.402309238910675O5),
 ('louvois', 0.3989852070808410 6),
 ('dc', 0.3935057520866394),
 ('continuation', 0.3869740962982178),
 ('bellorum', 0.384490430355072),
 ('kmelnytsky', 0.3835177421569824)]
```

- 利用 train_test_split 将 tf-idf 表示的训练集切分为训练数据集和验证数据集，验证集比例是 0.1，并且定义了用于评估模型准确率的函数 evaluate_model。
- 利用 sklearn 包，分别用决策树、朴素贝叶斯、支持向量机三种模型进行训练。其中贝叶斯采用 MultinomialNB，原因是词向量特征是离散数据，MultinomialNB 相比于高斯朴素贝叶斯更适合；支持向量机采用 LinearSVC，原因是词向量特征维度很高，线性核函数处理效率较高。测试结果如下：

模型	验证集准确率	测试集准确率
决策树(基准模型)	0.672	0.578
朴素贝叶斯	0.875	0.809
支持向量机	0.924	0.845

- 最后我利用 keras 建立神经网络，并分别用 20 类新闻和 text8 得到的 word2vec 作为输入训练模型，经比较 20 类新闻得到的 word2vec 效果更好。对于 embedding 层，考虑到大多数文档长度在 500 以下，我将文档长度定位 300，根据常用词的规模将词库数量设置为 15000，词向量维度设置为 200。接着利用一层卷积层和池化层来缩小向量长度，再通过 Flatten 层把二维向量转换为 1 维，最后用一层 Dence 将输出收缩到 20 个分类，测试准确率是 0.602 和 0.533。利用 20 类新闻得到的 Word2Vec 要比 text8 的效果更好，在词量上 20news 比 text8 少了 4 万单词，但同一单词 most_familiar，20news 显然比 text8 更贴近 20 类新闻。

3.3 完善

- 我利用网格搜索法和 k 折交叉法来优化朴素贝叶斯和支持向量机模型，优化后朴素贝叶斯的提升较为明显，在测试集上的准确率提高到了 0.828，支持向量机并没有提升，测试集准确率仍未 0.845。
- 对于卷积神经网络，在卷积层我使用了[2,3,4,5]四个 kernal size 卷积窗口来模拟 ngrams 进行拼接，在 flattern 层后增加了一层 dropout 进行去拟合，在 dropout 后

再增加一层 Dence 层，通过两个全连接层把输出收缩到 20 个分类，优化后准确率 t 提高到了 0.836。

4. 结果

4.1 模型的评价与验证

下表展示了各个模型的训练结果以及优化后的训练结果：

	基准模型	朴素贝叶斯	贝叶斯优化后	SVM	优化后 SVM	CNN 20newsgroup	CNN text8
训练集准确率	0.672	0.875	0.914	0.924	0.925	0.939	0.940
测试集准确率	0.578	0.809	0.828	0.845	0.845	0.836	0.819

由上表可知，优化后的 SVM 模型分类准确率最高，比基准模型的准确率高出很多，但是优化效果不明显。卷积神经网络在采用多个 kernal_size 训练后分类效果也很好，但训练时间相较于传统机器学习模型过长。贝叶斯模型优化后准确率得分也不错，且训练时间最短。

虽然朴素贝叶斯的模型会失去词语之间的顺序信息，但考虑到有些独立假设在各个分类之间的分布都是均匀的所以对于似然的相对大小不产生影响；即便不是如此，也有很大的可能性各个独立假设所产生的消极影响或积极影响互相抵消，最终导致结果受到的影响不大，所以贝叶斯模型的效果确非常明显，事实证明，贝叶斯模型确实是简单、实用且强大的。

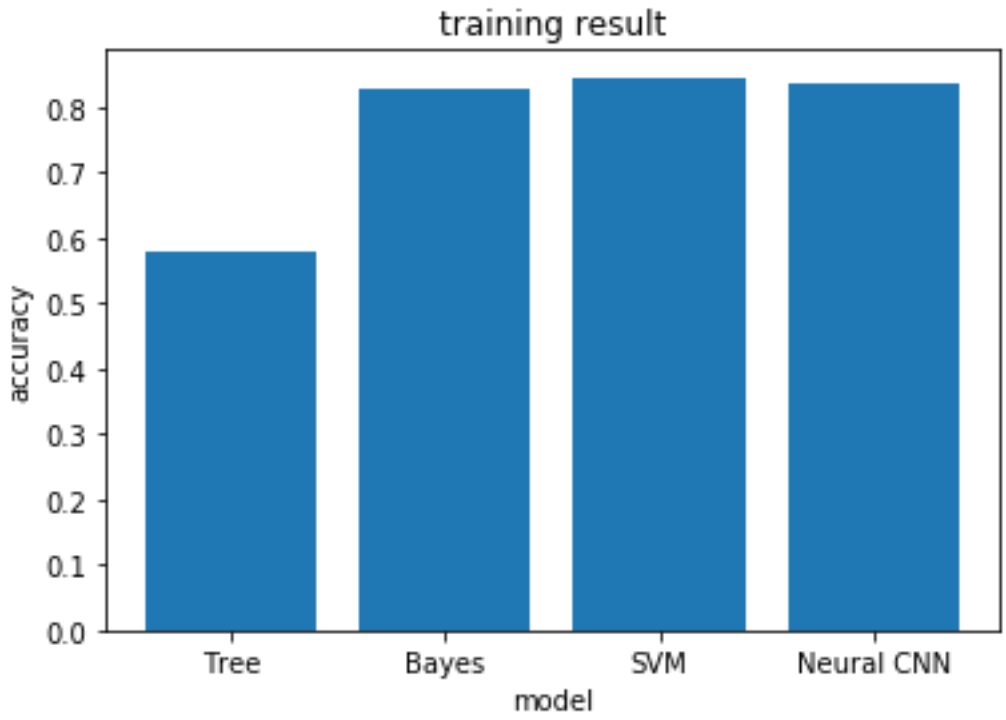
4.2 合理性分析

最终选择的 tf-idf 词向量结合支持向量机的模型准确率最高，接近 85%的分类预测准确率，这可以满足在生产环境处理简单的文档分类要求了。支持向量机的核心是把特征映射到一个高维空间中，这样就可以转化为一个线性可分的问题，泛化能力也较强。而朴素贝叶斯模型和深度学习神经网络的准确率相比 SVM 差距其实并不大。

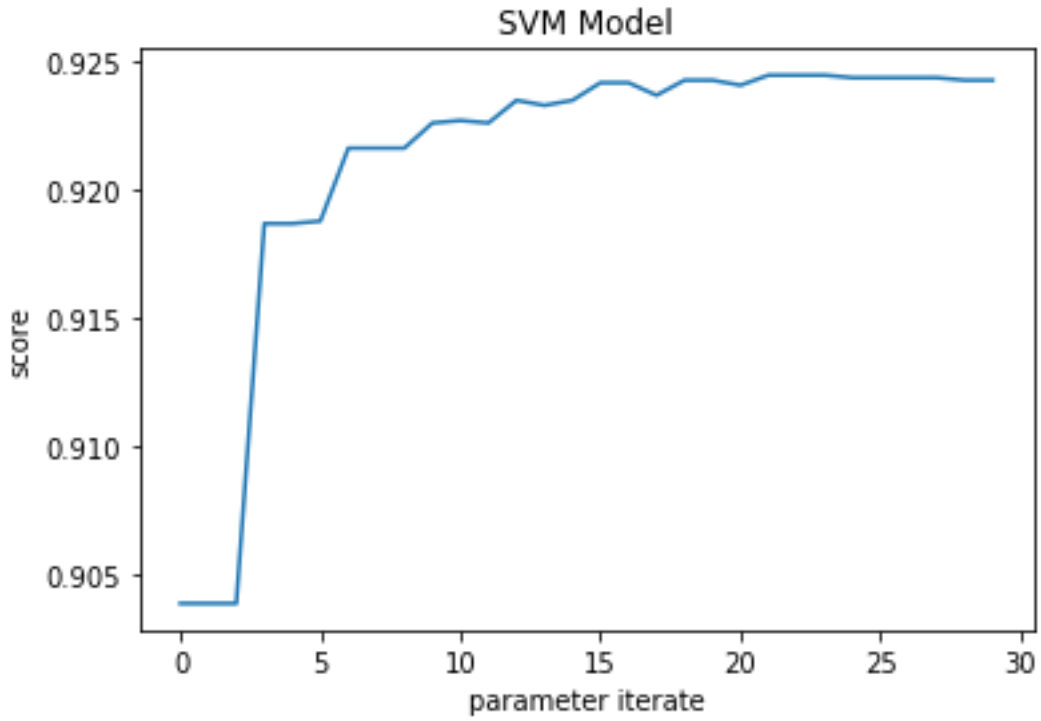
5. 项目结论

5.1 结果可视化

各个模型调优后的最高准确率如下图所



再看看效果最好的支持向量机，不同参数值对它的影响



可以看到随着参数的迭代，准确率逐渐升高，并在最后慢慢趋于稳定。C 值越小，泛化能力越强，但分类效果容易越差，C 值越大，越容易出现过拟合；t 值为容错率。从最优参数可以看到 C 值是迭代区间中心偏右的值，证明模型尽量保证了分类效果并且避免出现过拟合情况，t 值选择最小的迭代参数，保证较高精度准确率，是合理的。

5.2 对项目的思考

本项目两个主要的难题是怎样表示和提取文档的特征以及选择哪种分类模型，解决过程中我把单词作为文档特征，并采用词向量来描述这个特征，tfidf 结合了词频和逆文本频率能很好的识别出那些能有效识别某一类别文档的单词，此外还采用了 skip-gram 模式的 word2vec，word2vec 能够将稀疏、维度高的词向量矩阵进行降维，而 skip-gram 模式可以基于单词推测出近义的语句，这很符合分类新闻的需求。

分类模型分别选择了传统机器学习的决策树、朴素贝叶斯和支持向量机(决策树作为基准模型)和卷积神经网络，经优化朴素贝叶斯、支持向量机、CNN 均能达到 80%左右的准确率，可以看出传统的机器学习方法在文档分类问题上还是由很好的表现的，在一些场景下，应用传统机器学习方法能更简单、快速的解决实际需求，同时深度学习还有非常大的潜力可以挖掘！

5.3 需要作出的改进

深度学习中除了用 CNN 来进行文本分类，还可以利用 RNN。RNN 与 CNN 不同点在于，CNN 中隐藏层的节点之间是不连接的，而 RNN 隐藏层之间的节点是由连接的，隐藏层的输入不仅包括输入层的输出，还包括上一时刻隐藏层的输出。RNN 常用 LSTM 层或者 GRU 层实现，我尝试了用 GRU 层训练，测试集上得到的准确率是 0.834，效果也很不错。

6. 参考资料

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

[2] http://mattmahoney.net/dc/test8.zip

[3] https://developers.google.com/machine-learning/crash-course/classification/accuracy

[4] http://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_20newsgroups.html

[5] https://deeplearning4j.org/cn/bagofwords-tf-idf

[6] https://deeplearning4j.org/cn/word2vec

[7] http://www.cnblogs.com/leoo2sk/archive/2010/09/17/naive-bayesian-classifier.html

[8] https://pythonspot.com/nltk-stop-words/

[9] http://scikit-learn.org/stable/modules/feature_extraction.html

[10] https://radimrehurek.com/gensim/models/word2vec.html

[11] https://github.com/keras-team/keras/blob/master/examples/pretrained_word_embeddings.py

[12] https://keras-cn.readthedocs.io/en/latest/

[13] https://zhuanlan.zhihu.com/p/25928551