Proposal-CapstoneOfUdacity-NLP

April 4, 2017

1 毕业项目开题报告：自动文档分类

1.1 1. 项目背景

在互联网时代，越来越多的信息被上传到互联网上，下至个体上至各类组织，在进行大大小小各种决策时，都不可能忽视从互联网这个渠道获取的信息。面对浩如烟海的信息，要在更短的时间内得到更多、更准确的信息，仅靠人力查看与整理完全不现实；从而，各种相关的工具、技术应时而生，典型的例子包括搜索引擎、邮件过滤器、问答系统、消费者意见与情感分析技术（ref: http://52opencourse.com/222/斯坦福大学自然语言处理第六课-文本分类（text-classification））。而这些工具、技术的重要基础之一就是与文本分类相关的技术。

文本分类是基于文本内容将待定文本划分到一个或多个预先定义的类中的方法（ref: http://c.xml.org.cn/blog/uploadfile/20076211443809.PDF ），包括文本表示（预处理、索引、统计、特征表示）、分类器训练、评价与反馈等（ref: http://c.xml.org.cn/blog/uploadfile/20076211443809.PDF ）。文本表示方面的工作，包括词汇层面的独热编码（one-hot representation）、N 元语法（N-gram）模型，句子、段落层面的词袋模型（Bag-of-words model）（ref: https://zh.wikipedia.org/wiki/词袋模型）等，也有词嵌入（Word Embedding）模型（ref: http://forum.yige.ai/thread/70 ）（ref: https://www.zhihu.com/question/32275069 ）（ref: http://weibo.com/3121700831/BsCvWgmPs ）如词汇层面的 Word2Vec，句子、段落层面的 Sentence2Vec、Doc2Vec 等（ref: http://www.cnblogs.com/maybe2030/p/5427148.html ）（ref: http://blog.csdn.net/wangongxi/article/details/51591031 ）；分类器的训练方法则包括 SVM、KNN、贝叶斯、基于有监督学习器的集成学习器等常见的有监督学习方法（ref: http://59.108.48.5/course/mining/12-13spring/参考文献/04-04%20 基于机器学习的文本分类技术研究进展.pdf ）（ref: http://c.xml.org.cn/blog/uploadfile/20076211443809.PDF ）；评价方法则包括对于各分类器分类效果的查准率 P、查全率 R、F1 度量，以及对于总体而言的宏观平均（Macroaveraging）（给予每个分类同等权重从而求算术平均值，计算所有分类器的综合效果；用于测量小分类的效果）与对于总体而言的微观平均（Microaveraging）（给予每篇文档同等权重从而求算术平均值，计算每篇文档分类结果的综合效果；用于测量大分类的效果）（ref: http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-text-classification-1.html ）（ref: 周志华西瓜书 2.3.2 ）

本文作者对搜索引擎设计技术与情感计算技术感兴趣，因此选择研究此课题，以便为后续相关研究与设计做准备。

1.2 2. 问题描述

总的来说，本项目要解决的问题：

如何设计一套文本分类模型，能够把 18000 多条新闻较准确地分配到 20 个主题类别中？

问题分解为 3 个部分：

1. 从文本表示的角度看：已知的两种文本表示方法，即词袋模型（Bag-of-Words）方法和词嵌入模型（Word Embedding）方法，如何使用这两种方法为文本建立表示模型？哪一种模型更适合用于建立表示模型？
2. 从训练分类器的角度看：这是一个有监督学习问题。那么对于给定的文本表示模型，应该选择哪种监督学习方法进行训练，分类效果更好？
3. 从评估结合上述 2 个问题后得到的最终模型效果的角度看：经过训练后，哪种文本表示模型与哪种监督学习方法结合后训练得到的分类器的分类效果更好？

1.3 3. 输入数据

问题中涉及的数据集包括下述 2 份

1.3.1 3.1 text8

这是 gensim 在训练 word2vec 中所建议的一份数据。参考 http://www.mattmahoney.net/dc/textdata 与 https://groups.google.com/forum/#!topic/word2vec-toolkit/q02SKIqtmvU 。该数据将用于训练词向量（词袋模型的离散型词向量，或词嵌入模型的连续型词向量），从而实现对文本数据的规范表示

> 注：http://www.mattmahoney.net/dc/textdata 上提供的链接 http://www.mattmahoney.net/dc/text8.zip 对应文件已经损坏（md5 校验值与网上给的不同）。通过下载 enwik9 并使用修改过的 Perl 脚本处理 enwik9 后得到 text8（md5 值与 http://www.mattmahoney.net/dc/textdata 给出的一致）（原始脚本见 http://www.mattmahoney.net/dc/textdata#appendixa ，仅修改 `s/{{[^}]*}}//g;` 为 `s/\{\{[^}]*\}\}//g;`，因为我的 Perl 提示了 `Unescaped left brace in regex is deprecated, passed through in regex; marked by <-- HERE in m/{{ <-- HERE [^}]*}}/ at old_wikifil.pl line 34.`）

这是一份对原始的英文维基百科于 2006 年 3 月 3 日的 dump 文件（参考 http://www.mattmahoney.net/dc/textdata：The test data for the Large Text Compression Benchmark is the first 109 bytes of the English Wikipedia dump on Mar. 3, 2006. http://download.wikipedia.org/enwiki/20060303/enwiki-20060303-pages-articles.xml.bz2）进行清洗后得到的数据。

清洗的步骤至少包括：保留了常规正文文本、图像说明，丢弃了表格、超链接（转为普通文字）、引用、脚注、标记符号（如 `<text ...>`、`</text>`、`#REDIRECT`、`[`、`]`、`{{`、`}}`……）外国语言（英语以外的语言）版本，并将数字用英文拼写出来，将大写字母转换为小写字母等。经过上述清洗处理后，文本中只包含：由小写字母 a-z 组成的单词、单一空格（将不在 a-z 之间的字符也一律转换为空格）

如下为 text8 中前 2000 字节：

```
anarchism originated as a term of abuse first used against early working class radicals including t
```

1.3.2 3.2 20 Newsgroups

这是已经分为 20 类主题的 18000 条新闻文本，参考 http://www.qwone.com/~jason/20Newsgroups/ ，选择下载了其中的 http://www.qwone.com/~jason/20Newsgroups/20news-bydate.tar.gz 。在本项目中，该数据将用于训练文本分类器、评估分类器效果

具备如下特征：

- 每条新闻均被研究人员标注为 20 个主题中的一个
- 总数据集被分割为 2 个子集：训练集（占总数据 60%）+ 测试集（占总数据 40%）
- 剔除了跨主题的新闻（即任何一份新闻都只在单一主题下），提出了新闻组相关辨认标识（如 Xref, Newsgroups, Path, Followup-To, Date）

如下为其中的一份训练样本：

```
From: bed@intacc.uucp (Deb Waddington)
Subject: INFO NEEDED: Gaucher's Disease
Distribution: Everywhere
Expires: 01 Jun 93
Reply-To: bed@intacc.UUCP (Deb Waddington)
Organization: Matrix Artists' Network
Lines: 33


I have a 42 yr old male friend, misdiagnosed as having
 osteopporosis for two years, who recently found out that his
 illness is the rare Gaucher's disease.

Gaucher's disease symptoms include: brittle bones (he lost 9
 inches off his hieght); enlarged liver and spleen; internal
 bleeding; and fatigue (all the time). The problem (in Type 1) is
 attributed to a genetic mutation where there is a lack of the
 enzyme glucocerebroside in macrophages so the cells swell up.
 This will eventually cause death.

Enyzme replacement therapy has been successfully developed and
 approved by the FDA in the last few years so that those patients
 administered with this drug (called Ceredase) report a remarkable
 improvement in their condition. Ceredase, which is manufactured
 by biotech biggy company--Genzyme--costs the patient $380,000
 per year. Gaucher's disease has justifyably been called "the most
 expensive disease in the world".

NEED INFO:
I have researched Gaucher's disease at the library but am relying
 on netlanders to provide me with any additional information:
**news, stories, reports
**people you know with this disease
**ideas, articles about Genzyme Corp, how to get a hold of
```

```
   enough money to buy some, programs available to help with
   costs.
**Basically ANY HELP YOU CAN OFFER

Thanks so very much!

Deborah
```

如下为其中的一份测试样本：

```
From: banschbach@vms.ocom.okstate.edu
Subject: Re: Candida(yeast) Bloom, Fact or Fiction
Lines: 68
Nntp-Posting-Host: vms.ocom.okstate.edu
Organization: OSU College of Osteopathic Medicine

In article <1r9j33$4g8@hsdndev.harvard.edu>, rind@enterprise.bih.harvard.edu (David Rind) writes
> In article <1993Apr22.153000.1@vms.ocom.okstate.edu>
>  banschbach@vms.ocom.okstate.edu writes:
>>poster for being treated by a liscenced physician for a disease that did
>>not exist.  Calling this physician a quack was reprehensible Steve and I
>>see that you and some of the others are doing it here as well.
>
> Do you believe that any quacks exist?  How about quack diagnoses?  Is
> being a "licensed physician" enough to guarantee that someone is not
> a quack, or is it just that even if a licensed physician is a quack,
> other people shouldn't say so?  Can you give an example of a
> commonly diagnosed ailment that you think is a quack diagnosis,
> or have we gotten to the point in civilization where we no longer
> need to worry about unscrupulous "healers" taking advantage of
> people.
> --
> David Rind

I don't like the term "quack" being applied to a licensed physician David.
Questionable conduct is more appropriately called unethical(in my opinion).
I'll give you some examples.

    1. Prescribing controlled substances to patients with no
       demonstrated need(other than a drug addition) for the medication.

    2. Prescribing thyroid preps for patients with normal thyroid
       function for the purpose of quick weight loss.

    3. Using laetril to treat cancer patients when such treatment has
       been shown to be ineffective and dangerous(cyanide release) by
       the NCI.
```

```
These are errors of commission that competently trained physicians should
not committ but sometimes do.  There are also errors of omission(some of
which result in malpractice suits).  I don't think that using anti-fungal
agents to try to relieve discomfort in a patient who you suspect may be
having a problem with candida(or another fungal growth) is an error of
commission or omission.  Healers have had a long history of trying to
relieve human suffering.  Some have stuck to standard, approved procedures,
others have been willing to try any reasonable treatment if there is a
chance that it will help the patient.  The key has to be tied to the
healer's oath, "I will do no harm".  But you know David that very few
treatments involve no risk to the patient.  The job of the physician is a
very difficult one when risk versus benefit has to be weighed.  Each
physician deals with this risk/benefit paradox a little differently.  Some
are very conservative while others are more agressive.  An agressive
approach may be more costly to the patient and carry more risk but as long
as the motive is improving the patient's health and not an attempt to rake
in lots of money(through some of the schemes that have been uncovered in
the medicare fraud cases), I don't see the need to label these healers as
quacks or even unethical.

What do I reserve the term quack for?  Pseudo-medical professionals.
These people lurk on the fringes of the health care system waiting for the
frustrated patient to fall into their lair.  Some of these individuals are
really doing a pretty good job of providing "alternative" medicine.  But
many lack any formal training and are in the "business" simply to make a
few fast bucks.   While a patient can be reasonably assured of getting
competent care when a liscenced physician is consulted, this alternative
care area is really a buyer's beware arena.  If you are lucky, you may find
someone who can help you.  If you are unlucky, you can loose a lot of
money and develop severe disease because of the inability of these pseudo-
medical professional to diagnose disease(which is the fortay of the
liscened physicians).

I hope that this clears things up David.

Marty B.
```

1.4 4. 解决办法

1. 【特征抽取与文本表示】从文本数据集中抽取表示文本所需的特征，然后在文本数据集上用这些抽取出的特征重新表示文本数据集
2. 【分类器训练】对已经建立的表示模型，在每个模型上分别使用一些有监督学习方法在训练数据上分别训练出一定数量的分类器。
3. 【性能评估】对上述分类器进行如下评估：
    1. 对于上述每种表示模型：比较同一文本表示模型下不同训练方法的训练效果
    2. 在每种表示模型的语境下各选出分类效果最好的那个分类器，并进行比较

1.5 5. 评估指标

综合考虑下述 3 个指标：

- F1：同时考虑查准率（Precision）$P = \frac{TP}{TP+FP}$ 与查全率（Recall）$R = \frac{TP}{TP+FN}$
- 训练时间 t_{train}：训练分类器达到标准所耗费的时间
- 分类时间 t_{test}：训练出的分类器在测试数据上进行分类所耗费的时间

在最终评估分类器性能时，使用下述公式来综合考虑这 3 个指标：

- $score(F_1, t_{train}, t_{test}) = \frac{F_1}{1+t_{train}*t_{test}}$

1.6 6. 基准模型

参考 A Comparative Study on Different Types of Approaches to Text Categorization 和 Representation and Classification of Text Documents: A Brief Review 的 Table 1: Comparative Results Among Different Representation Schemes and Classifiers obtained on Reuters 21578 and 20 Newsgroup Datasets，选取其中以 20 Newsgroup 为数据集、且与本项目待测方法有关的实验结果如下表：

Results reported by	Representation Scheme	Classifier Used	Micro F1	Macro F1
[Ko et al., 2004]	Vector representation with different weights	Naïve Bayes	83.00	83.30
	K-NN	81.04	81.20	
	SVM	86.10	86.00	
[Tan et al., 2005]	Vector representation	Naïve Bayes	0.835	0.835
	K-NN	0.848	0.846	
	SVM	0.889	0.887	
[Mubaid and Umair.,2006]	Vector representation	SVM	84.62	78.19
[Lan et al., 2009]	VSM with term weighting schemes	SVM	0.808	0.808
	K-NN	0.691	0.691	

表中的参考文献如下：

[Ko et al., 2004]
Ko, Y. J., Park, J., and Seo, J. 2004. Improving text categorization using the importance of senten
International Journal Information Processing and Management, Vol. 40, pp. 65 - 79.

[Tan et al., 2005]
Songbo, T., Cheng, X., Ghanem, M. M., Wnag, B., and Xu, H. 2005. A novel refinement approach for tex
categorization. In the Proceedings of Fourteenth ACM International Conference on Information and K
476.

[Mubaid and Umair.,2006]
Mubaid, H. A., and Umair, S. A. 2006. A New Text Categorization Technique Using Distributional Clust
and Learning Logic. IEEE Transactions on Knowledge and Data Engineering, Vol 18 (9), pp. 1156 -
1165

[Lan et al., 2009]
Lan, M., Tan, C. L., Su. J., and Lu, Y.2009. Supervised and Traditional Term Weighting Methods for A
735

考虑到文献在 7.2. Benchmarks for Text Categorization 这一节提到的：

> In general, different sets of experiments may be used for cross-classifier comparison only if the experiments have been performed (1) on exactly the same collection (i.e., same documents and same categories); (2) with the same "split" between training set and test set; (3) with the same evaluation measure and, whenever this measure depends on some parameters (e.g., the utility matrix chosen), with the same parameter values

也就是说，仅当这些实验满足下述条件时，各分类器是可比的：(1) 实验数据集相同；(2) 在数据集上的划分相同（训练集，测试集）；(3) 使用相同的方法测量性能，且每当测量依赖于某些参数时，参数必须取相同值

上述 4 篇文章能满足全部 3 个条件的只有 [Lan et al., 2009] 所进行的部分实验，即 V-A 中的 Figure 3 与 Figure 5 对应实验。参考该文章，整理表格如下：

Results reported by	Representation Scheme	Classifier Used	Micro F1	Macro F1
[Lan et al., 2009]	VSM with term weighting schemes	SVM	0.808	0.808
	K-NN	0.691	0.691	

参考 Machine Learning in Automated Text Categorization，该文章在 7.3 Which text classifier is best? 这一小节的讨论中尝试得出一些结论：

1. 表现最好的学习器：集成学习器，支持向量机（SVM）≈ 决策树，kNN
2. 次优的学习器：神经网络
3. 表现最差的学习器：朴素贝叶斯

文章随后也提及，上述结论不是绝对的，例如实际使用环境中某写「语境」具备的特征可能与训练语料中的性质大为不同，而不同的分类器对这些性质的响应又不同（It is important to bear in mind that the considerations above are not absolute statements (if there may be any) on the comparative effectiveness of these TC methods. One of the reasons is that a particular applicative context may exhibit very different characteristics from the ones to be found in Reuters, and different classifiers may respond differently to these characteristics）。尽管如此，仍不妨以上述结论与数据为参考之一。

关于 Word2Vec 的性能，Deeplearning4j 的文档中是这样陈述的：

> Word2vec 很适合对文档进行深入分析，识别文档的内容和内容子集。它的向量表示每个词的上下文，亦即词所在的 n-gram。词袋法适合对文档进行总体分类。

估计 Word2Vec 对文档进行总体分类的效果或许不如 TF-IDF。再考虑到Distributed Representations of Sentences and Documents 的 Table 3 中表明在 Distributed Representations 下的错误率

有相对 TF-IDF 情况下 32% 左右的改善，相应可认为正确率有 32% 左右的改善，从而设定标准如下，所有提及的学习器的训练效果将不小于下述基准：

文本表示方案	分类器	微 F_1（Micro F_1）	宏 F_1（Macro F_1）
TF-IDF	集成学习器	0.85	0.85
SVM	0.80	0.80	
决策树	0.80 ± 0.02	0.80 ± 0.02	
kNN	0.75	0.75	
神经网络	0.70	0.70	
朴素贝叶斯	0.65	0.65	
Word2Vec	集成学习器	0.85 ± 0.10	0.85 ± 0.10
SVM	0.80 ± 0.10	0.80 ± 0.10	
决策树	0.80 ± 0.12	0.80 ± 0.12	
kNN	0.75 ± 0.10	0.75 ± 0.10	
神经网络	0.70 ± 0.10	0.70 ± 0.10	
朴素贝叶斯	0.65 ± 0.10	0.65 ± 0.10	

1.7 7. 设计大纲：你的解决方案如何实现，如何获取结果（1 页）

1. 【特征抽取与文本表示】
   - 使用 **TF-IDF** 方法抽取特征，建立表示模型 1
   - 使用**词嵌入**（Word embedding）方法（在这里，具体使用 Word2Vec）抽取特征，建立表示模型 2
   - 考虑到 Word2Vec 的对标是 LSI，可能会使用 LSI 或 LDA 建立表示模型 1 或表示模型 3
2. 【分类器训练】
   - 文本表示模型建模工具
     - gensim
     - scikit-learn
   - 学习算法：对上述 2~3 个模型，在每个模型上分别使用下述有监督学习方法在训练数据上训练出一组分类器：
     - 神经网络
     - 逻辑回归
     - 决策树
     - 支持向量机（SVM）
     - k 近邻（k-NN）
     - 朴素贝叶斯
     - 集成学习
       * 基于上述方法（决策树以外、神经网络以外）的集成学习（AdaBoost）
       * 随机森林
   - 学习工具：
     - tensorflow：用于训练神经网络模型
     - scikit-learn：用于训练下述学习算法
       * 逻辑回归
       * 决策树

    * 支持向量机（SVM）
    * k 近邻（k-NN）
    * 朴素贝叶斯
    * 集成学习
        · 基于上述方法的集成学习（AdaBoost）
        · 随机森林

3. 【性能评估】

    1. 评估流程
        1. 在每个文本表示模型 m 的语境下训练每一个分类器 c 时，就记下分类器 c 被训练到不低于基准要求的水平时所耗费的时间 t_{train}
        2. 对于上述每种文本表示模型：对同一种文本表示模型，对测试集文本数据进行分类，并得到 F_1 与实际分类时间 t_{test}，比较所有方法的效果
        3. 从每个表示模型对应的所有分类器中选出效果最好的一个分类器 $c(i)$，比较这些分类器的性能
    2. 评估指标：见上述「5. 评估指标」

In [ ]: