中国好创意
垃圾短信基于文本内容识别
竞赛报告书

lifematrix
stevenliucx@gmail.com

2015.12

数据集

id	label	content
11	0	乌兰察布丰镇市法院成立爱心救助基金
12	1	(长期诚信在本市作各类资格职称（以及印 /章、牌、 ……等。祥：x x x x
13	1	《依林美容》三．八．女人节倾情大放送活动开始啦！！！！超值套餐等你拿，活动时间x
14	0	品牌墙/文化墙设计参考
15	0	苏州和无锡两地警方成功破获了一起劫持女车主的案件
16	0	自然之友苏州小组今日下午按原计划举办小组读书活动暨"我为城市量体温"启动仪式
17	0	共查扣违法三轮车304辆、残疾证23本

训练集 80万条
正样本率: 10%

实际的垃圾短信率不会这么高，估计做过向上取样(upsample)

id	label	content
800001	?	.x月xx日推出凭证式国债x年期x.xx.xx%，x年期x.xx%到期一次还本付
800002	?	x强度等级水泥的必要性和可行性进行深入研究
800003	?	Don'tSellaProduct
800004	?	以上比赛规则由江苏科技大学教职工摄影协会负责解释
800005	?	坐12个小时飞机身体已经疲惫不堪
800006	?	为什么不能是你③以多数人的努力程度
800007	?	地址位于天津市滨海新区响罗湾旷世国际大厦A座1801室

测试集 20万条

预测

任务

目标

基于短信文本内容，准确地、完整地识别出垃圾短信、正常短信。用准确率、查全率、效率（每秒处理条数）衡量。

评估指标

判定数据 / 原始数据	正常短信	垃圾短信	合计
正常短信	A	B	A+B
垃圾短信	C	D	C+D
合计	A+C	B+D	-

垃圾短信准确率 = D/(B+D)，垃圾短信判正确的占全部判垃圾短信的比率；

垃圾短信查全率 = D/(C+D)，垃圾短信判正确的占全部短信的比率；

正常短信准确率 = A/(A+C)，正常短信判正确的占全部判正常短信的比率；

正常短信查全率 = A/(A+B)，正常短信判正确的占全部短信的比率；

$$F_{垃圾} = 0.65 * 垃圾短信准确率 + 0.35 * 垃圾短信查全率$$

$$F_{正常} = 0.65 * 正常短信准确率 + 0.35 * 正常短信查全率$$

$$F_{总分} = 0.7 * F_{垃圾} + 0.3 * F_{正常}$$

问题的性质

- 有督管的机器学习任务
- 二元分类问题(binary classification)
- 自然语言处理
  - 涉及中文，需分词
- 短文本分析

问题的难点

- **短文本，feature可能不够**
  - 需构造多样化的features
- **记录数不算多，100万条**
  - 防止过度拟合
- **需要预测Label 0/1，而非概率**
  - 评估指标不是通常的AUC或logarithmic loss
  - 分类器预测出概率后，还要寻找模型最优的threshold
- **性能要求**
  - 要满足单机或多机使用

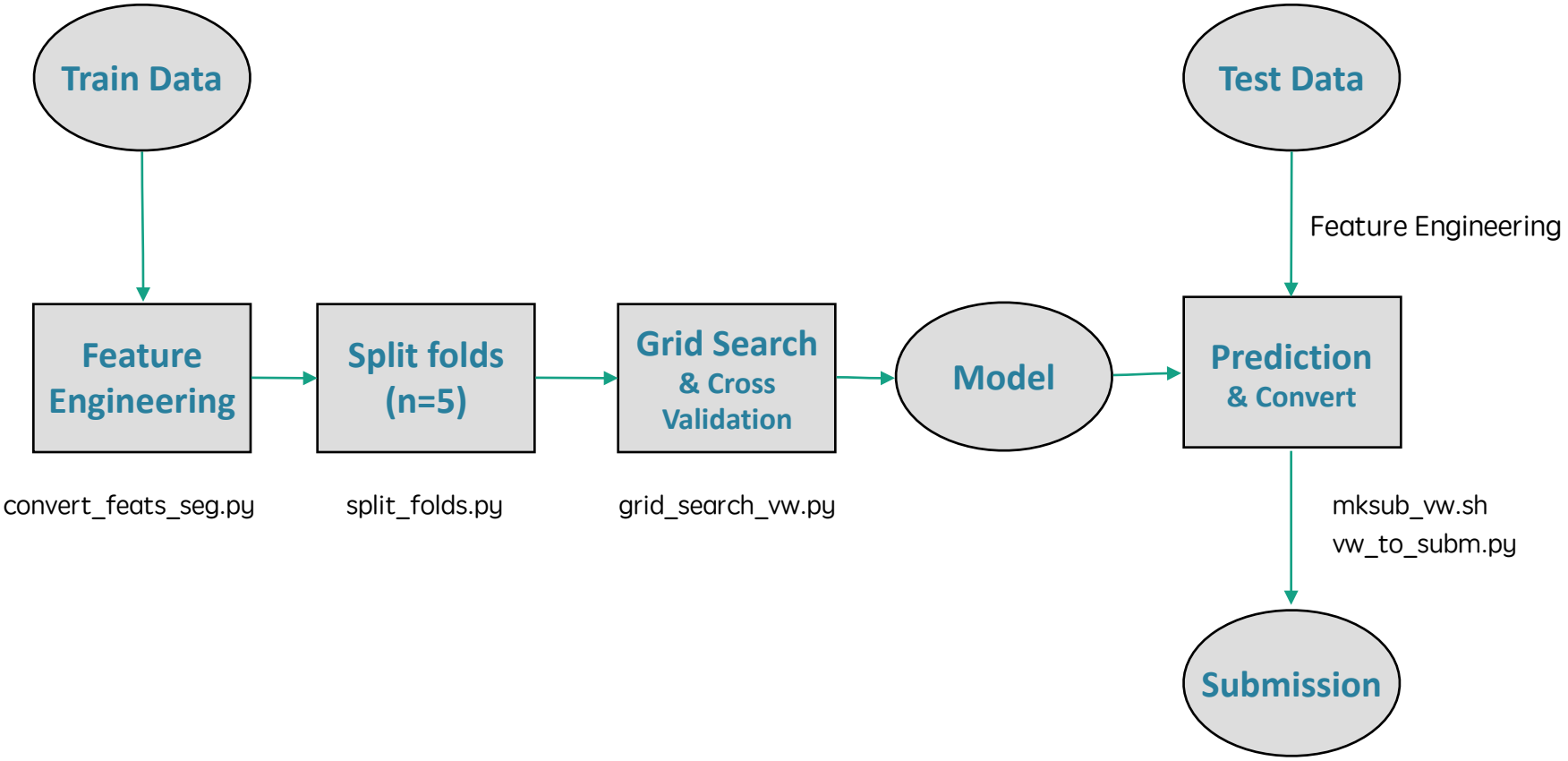
求解思路

- 1. 对文本分词，构造描述性强的feature
- 2. 训练分类器
  - 备选的分类器：线性分类器(logistic等)、SVM和树模型(CART)等
  - 由于性能的要求，优先选择**线性分类器**
- 3. 优化模型
  - 通过grid search和交叉校验，寻找最优的参数组合
- 4. 模型组合(Ensamble)
  - Bagging / Boosting
  - 尝试其它模型(分类树)，进行ensamble。 (因时间不够，本步实施)

选择的技术与工具

- 1. 结巴分词
  - https://github.com/fxsjy/jieba
- 2. Python 2.7.6
  - Feature构造和程序控制
- 3. Vowpal Wabbit 8.8.1，在线学习工具
  - https://github.com/JohnLangford/vowpal_wabbit

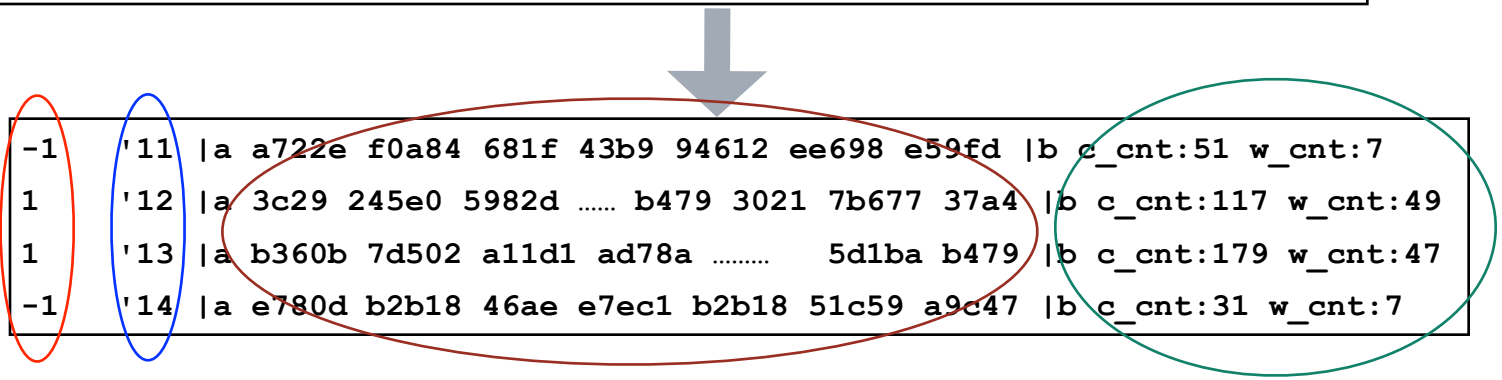
方法


Train Data
Test Data
Feature Engineering
Feature Engineering
Split folds (n=5)
Grid Search & Cross Validation
Model
Prediction & Convert
convert_feats_seg.py
split_folds.py
grid_search_vw.py
mksub_vw.sh
vw_to_subm.py
Submission

Feature Engineering

- 转换成VW可读的格式
- 分词 + hash
- 添加features：character & word count

id	label	content
11	0	乌兰察布丰镇市法院成立爱心救助基金
12	1	（长期诚信在本市作各类资格职称（以及印 /章、牌、 ……等。祥：x x x x
13	1	《依林美容》三．八．女人节倾情大放送活动开始啦！！！！超值套餐等你拿，活动时间x
14	0	品牌墙/文化墙设计参考



```
-1   '11 |a a722e f0a84 681f 43b9 94612 ee698 e59fd |b c_cnt:51 w_cnt:7
1    '12 |a 3c29 245e0 5982d …… b479 3021 7b677 37a4 |b c_cnt:117 w_cnt:49
1    '13 |a b360b 7d502 a11d1 ad78a ……   5d1ba b479 |b c_cnt:179 w_cnt:47
-1   '14 |a e780d b2b18 46ae e7ec1 b2b18 51c59 a9c47 |b c_cnt:31 w_cnt:7
```

label
{0,1}->{-1,1}

instance
id

feature namespace a
分词的hash值

feature namespace b
char count, word count

Grid Search 结果

主要的改进步骤

#	模型	说明	交叉验证得分	A 榜得分	B 榜得分	
1	基本模型(缺省参数) *loss func=logisitcs, lr = 1.0, threshold=0.5*	未分词	*0.99170*	*0.99029*	*0.98969*	
2	*ngram=2, threshold=0.6*	二元模型	*0.99475*	*0.99502*	*0.99477*	
3	全分词，*ngram=2, threshold=0.4*	全分词	*0.99530*	*0.99546*	*0.99504*	
4	*loss func=hinge, lr=0.15, threshold=0.3, l2=1e-6*	损失函数 *hinge*	*0.99620*	*0.99644*	*0.99625*	
5	*loss func=hinge, ngram=2, lr =0.65, threshold=0.34, l2=1e-5, bits=31*	增加*hash bits*空间 最终调优	*0.99630*	*0.99656*	*0.99646*	**最优 最终提交**

注：前4个是在VW 7.7下调优。最后结果(5)是在8.8.1最新版下，调优并提交。

性能

#	步骤	说明	耗时
1	转换测试集	80万条	3m48s
2	转换训练集	20万条	57s
4	训练模型	hash bits = 31	1m38s
5	预测、生成提交文件	预测速度2.5万/s	8s

训练和测试不到2分钟

测试环境：Intel(R) Xeon(R) CPU E5-2665 0 @ 2.40GHz

- 由于极高的训练速度，也减少了实验和grid search的时间，可以尽快发现最优模型

总结

- **本方案对大规模文本分类问题，有很好的针对性**
  - 极高准确性， 得分在第5
  - 极高速度，100万记录训练和预测时间，不超过2分钟
  - 支持hash trick，降低维度
  - 支持多种过滤器，在一个模型中训练
  - 支持在线学习
  - 工业级工具

- **所用方法&所建模型，清晰明了**
  - 易于理解，便于软件开发中的维护

获得的经验

- 二元模型(ngram=2)效果最佳
  - ngram>2效果下降。skip-k-gram也未有效果。原因可能是，短信文本与通常文本(邮件/网页)不同，多用简语。太复杂的模型，反而引入噪声。
- 中文分词有效
  - 时间有限，没有比较不同分词工具。最好针对短信建立专门的语料库
- hinge损失函数更适合文本处理
- 学习率的调优，显著提高效果
- 交叉校验避免过度拟合，保证了模型的推广性
  - 在训练集交叉校验得分提升的模型，在测试集(A&B)上也有同步提升
- Bagging效果不明显

今后的改进

- **改进正确性**
  - 尝试其它分词工具
    - 必要时针对短信文本建立特定的语料库
  - 尝试其它模型，建立ensamble
    - 树模型
- **改进性能**
  - 将feature转换改用CPython
  - 采用多机Allreduce方式
- **实用性**
  - 用更大更实际的数据集

参考资料

Hash Tricks	*K.Weinberger, A.Dasgupta, J.Attenberg, J.Langford, and A.Smola,* ***Feature Hashing for Large Scale Multitask Learning****, ICML 2009.* *http://arxiv.org/abs/0902.2206*
E-mail Spam	*J. Attenberg, K.Weinberger, A. Smola, A. Dasgupta, M. Zinkevich,* ***Collaborative Email-Spam Filtering with the Hashing-Trick****.* *http://www.cs.cornell.edu/~kilian/papers/ceas2009-paper-11.pdf*
Vowpal Wabbit	*John Langford,* ***Technical Tricks of Vowpal Wabbit.*** *http://www.slideshare.net/jakehofman/technical-tricks-of-vowpal-wabbit*
	Zhen Qin, ***Tutorial on Recent Practical Vowpal Wabbit Improvements****.* *https://github.com/JohnLangford/vowpal_wabbit/wiki/Zhen.pdf*
Skip-gram Modelling	*D.Guthrie, B.Allison, W.Liu, L.Guthrie, Y.Wilks,* ***A Closer Look at Skip-gram Modelling****.* *http://homepages.inf.ed.ac.uk/ballison/pdf/lrec_skipgrams.pdf*

谢 谢!