# Ensemble Learning

黄 欢

（计算机科学与技术系 计 33 2013011331）

## 1 Design

**Data：**

1803 spam hosts

4411 normal hosts

per host:

96 content features

138 transformed link features

**Task:**

**ensemble learning algorithms:**

Bagging

AdaBoost.M1

**base classifiers: （使用 scikit-learn: machine learning in Python）**

SVM

Decision Tree

Naïve Bayes**(optional)**

**Main.py：Compare 6 combinations**

1. 训练数据：随机选取 80%的数据作为训练集。
2. 特征选取：onlyContent，onlyLink，contentAndLink。**(optional)**
3. 处理数据不平衡问题：先进行分层抽样，将数据分为训练集和测试集的时候，两类数据独立采样，保证训练集和测试集中两类数据的比例与原数据集一致。再随机选取训练集中的 spam 类数据进行复制，使得训练集中两类数据一样多。 **(optional)**
4. 调用 sklearn.preprocessing.normalize(x, axis = 0)函数，对数据进行归一化，以提高性能。**(optional)**
5. 调整 ensemble learning 的迭代次数，即弱分类器的个数，分析迭代次数对性能的影响。Base 分类器的性能就是迭代次数为 1 时的性能。**(optional)**
6. Bagging 算法实现：对 x,y 进行 bootstrap 抽样，用 base classifier 训练抽样后的数据，将新训练出来的 base classifier 插入到分类器列表，再用分类器列表中的每个分类器对测试集进行分类，并投票决定最终分类结果。
7. AdaBoost.M1 算法实现：首先将所有样本的权值设为 1/N，用 base classifier 对有

权值的训练集进行训练，计算错误率 Et 和 beta，修改权值，保存当前分类器的 voting weight，再用分类器列表中的每个分类器对测试集进行分类，由投票权值的大小决定最终分类结果。

# 2 Results

根据如下公式：

$$Accuracy = \frac{\text{number of correctly classified records}}{\text{number of test records}}$$

$$precision: P = \frac{\sum_k (\frac{a_k}{a_k + b_k})}{K}$$

$$recall: R = \frac{\sum_k (\frac{a_k}{a_k + c_k})}{K}$$

$$f - score: F = \frac{2PR}{P + R}$$

可以得到结果，保存在 src/result 中。

# 3 Analysis

a) **在使用所有特征并对数据进行归一化的情况下，6 种组合迭代 6 次：**

| Round = 6 | | accuracy | precision | recall | F1 |
|---|---|---|---|---|---|
| Bagging | dTree | 0.8944 | 0.831884 | 0.795014 | 0.813031 |
| | SVM | 0.846215 | 0.663883 | 0.908571 | 0.767189 |
| | NB | 0.761475 | 0.571942 | 0.857143 | 0.686084 |
| AdaBoost.M1 | dTree | 0.859592 | 0.788732 | 0.742706 | 0.765027 |
| | SVM | 0.687753 | 0 | 0 | 0 |
| | NB | 0.761475 | 0.571942 | 0.798246 | 0.626866 |

Bagging 和 dTree 的 accuracy，precision 和 F1 值最大。

Bagging_SVM 的 recall 最大。

b） 特征选取对性能的影响

| accuracy | Only Content | Only Link | Content And Link |
|---|---|---|---|
| Bagging_dTree | 0.863994 | 0.867306 | 0.8944 |
| Bagging_SVM | 0.801626 | 0.844316 | 0.846215 |
| AdaBoost.M1_dTree | 0.840729 | 0.858167 | 0.859592 |
| AdaBoost.M1_SVM | 0.279412 | 0.710462 | 0.687753 |

一般来说，运用 Content 和 Link 两种特征的准确率最高。Only Link 比 Only Content 的准确率高。

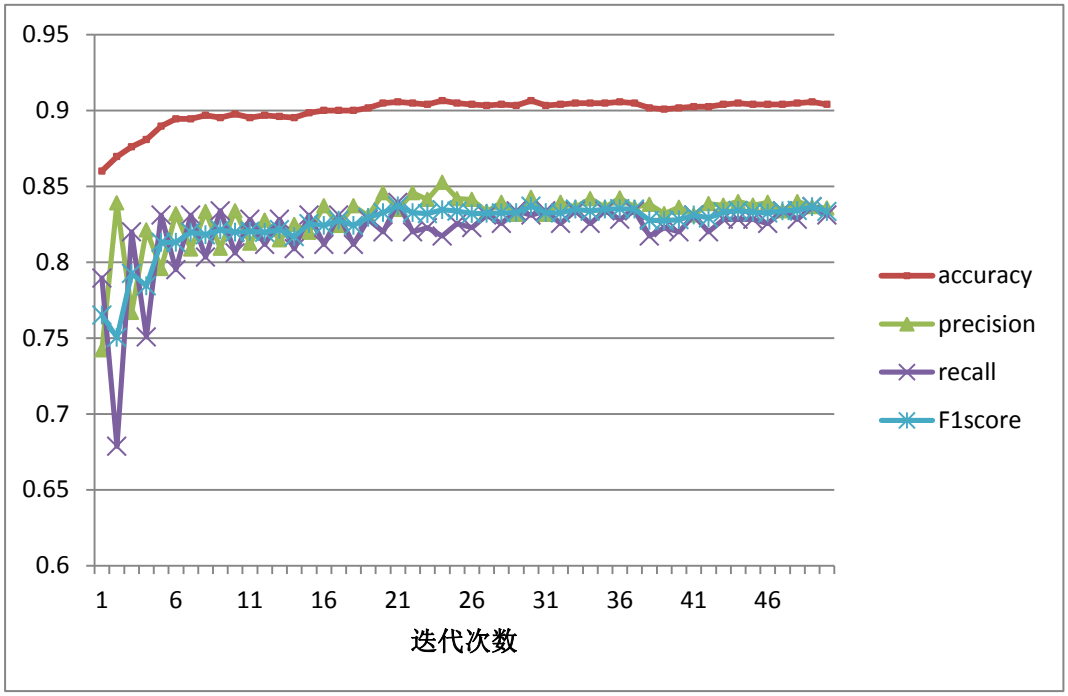
c） 数据归一化对性能的影响

| accuracy | 未归一化 | 归一化 |
|---|---|---|
| Bagging_dTree | 0.860258 | 0.8944 |
| AdaBoost.M1_dTree | 0.887658 | 0.859592 |

一般来说，归一化的准确率更高，但是 dTree 不依赖于特征的线性组合，所以没有太明显的改进。

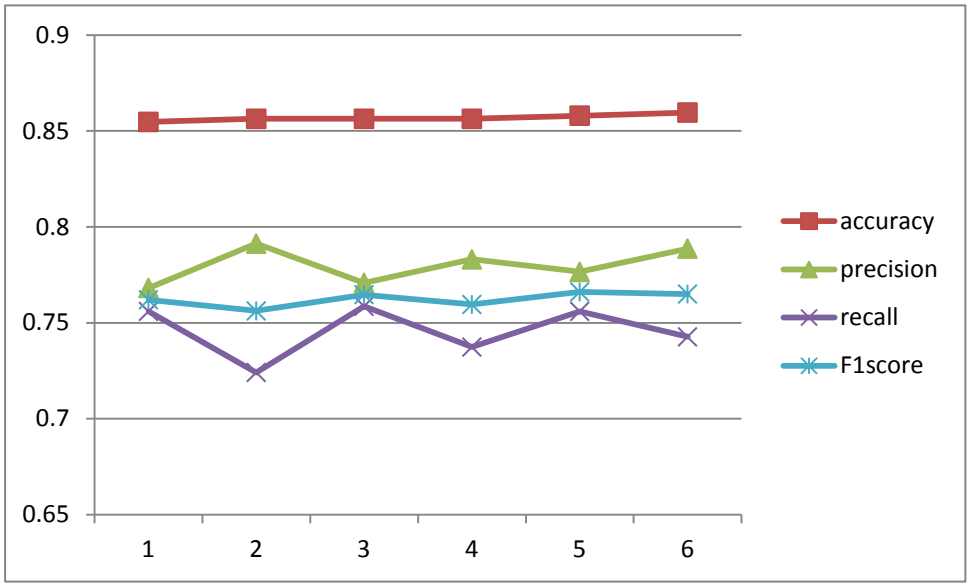
d） 迭代次数对性能的影响

1. bagging_decisionTree_contentAndLink_normalize




随着迭代次数的增大，各种结果效果也越来越好，最后 20 次迭代后趋于稳定。

2. AdaBoostM1_decisionTree_contentAndLink_normalize




随着迭代次数的增大，各种结果效果基本稳定。