# Ensemble Learning

胡求[1] MG1633031 huqiu00@163.com

[1](南京大学 计算机科学与技术系,南京 210023)

**摘 要:** 本次作业完成集成学习的内容，首先实现一个朴素贝叶斯分类器，然后以朴素贝叶斯作为基分类器，实现 Adaboost 算法，并在两个数据集上进行测试。在实现朴素贝叶斯分类器时，同时处理了离散值属性和连续值属性，在结果表中给出了 Adaboost 算法应用在两个数据集上分别的准确率均值和方差。

**关键词:** 朴素贝叶斯; 集成学习; Adaboost; 提升方法; 交叉验证

## 1 朴素贝叶斯算法简介

朴素贝叶斯算法(Naïve Bayes)算法是基于贝叶斯定理和特征条件独立的假设的一种分类方法。对于给定的训练数据集，首先基于特征条件独立假设学习输入/输出的联合概率分布，然后基于此模型，对给定的输入 x，利用贝叶斯定理求出后验概率最大的输出 y. 朴素贝叶斯法实现简单，并且也有不俗的学习与预测的效率，是一种常用的方法，也被列入“机器学习十大算法”之一。

## 2 离散特征和连续特征

利用朴素贝叶斯法进行分类我们采用极大似然法对条件概率和先验概率进行估计，然后利用此条件概率和先验概率对测试集样本进行预测，所以我们需要得到 $P(Y=c_k)$和 $P(X^j=a_{jl} \mid Y=c_k)$，然而，对于离散值和连续值而言两个条件概率的处理方法是不同的。

### 2.1 离散值特征的处理

在离散特征中，由于属性的取值情况总数固定，我们可以采用频数进行估计，并做一定的拉普拉斯平滑，即有

$$P_\lambda(X^{(j)}=a_{jl} \mid Y=c_k)=\frac{\sum_{i=1}^{N} I(x_i^{(j)}=a_{jl}, y_i=c_k)+\lambda}{\sum_{i=1}^{N} I(y_i=c_k)+S_j\lambda}$$

式中 $\lambda \geqslant 0$. 等价于在随机变量各个取值的频数上赋予一个正数 $\lambda>0$. 当 $\lambda=0$ 时就是极大似然估计. 常取 $\lambda=1$，这时称为拉普拉斯平滑（Laplace smoothing）. 显然，对任何 $l=1,2,\cdots,S_j$， $k=1,2,\cdots,K$，有

$$P_\lambda(X^{(j)}=a_{jl} \mid Y=c_k)>0$$

$$\sum_{l=1}^{S_j} P(X^{(j)}=a_{jl} \mid Y=c_k)=1$$

进行拉普拉斯平滑后，先验概率的贝叶斯估计为

$$P_\lambda(Y=c_k)=\frac{\sum_{i=1}^{N} I(y_i=c_k)+\lambda}{N+K\lambda}$$

### 2.2 连续值特征的处理

在处理连续值特征时，假定连续值属性服从均值为 μ，标准差为 σ 的高斯分布，由下式定义

$$g(x,\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

则 $P(x_k \mid C_i)=g(x_k,\mu_{c_i},\sigma_{c_i})$

只要计算出训练样本中连续值属性属于各个类别的数据样本的均值和标准差，我们就可以利用上述公式算得一个相对的概率，虽然上式的 P 可能不在 0 到 1 之间，但是并不影响最后后验概率的比较结果。

## 3 朴素贝叶斯以及提升方法的实现

### 3.1 提升方法

提升方法的基本思想是提高前一个弱分类器分错的样本的权重，使接下来的弱分类器能够学习到这个“残差”，相当于把分类问题交给多个弱分类器分而治之，每个分类器把前面分类器不擅长的数据学好，最后大家一起表决，表决时错误率低的分类器的分类结果应该占以更大的比重，通过这种方式，得到“三个臭皮匠赛过诸葛亮”的效果。Adaboost 也是一直以来被使用最广泛的提升算法。其他的 boosting 方法还有 gradient boost，boosting tree 等。

Adaboost 算法的伪代码如下

**Algorithm** *AdaBoost*(Data Set: $\mathcal{D}$, Base Classifier: $\mathcal{A}$, Maximum Rounds: $T$)
**begin**
  $t=0$;
  **for** each $i$ initialize $W_1(i)=1/n$;
  **repeat**
    $t=t+1$;
    Determine weighted error rate $\epsilon_t$ on $\mathcal{D}$ when base algorithm $\mathcal{A}$
        is applied to weighted data set with weights $W_t(\cdot)$;
    $\alpha_t=\frac{1}{2}\log_e((1-\epsilon_t)/\epsilon_t)$;
    **for** each misclassified $\overline{X_i}\in\mathcal{D}$ **do** $W_{t+1}(i)=W_t(i)e^{\alpha_t}$;
      **else** (correctly classified instance) **do** $W_{t+1}(i)=W_t(i)e^{-\alpha_t}$;
    **for** each instance $\overline{X_i}$ **do** normalize $W_{t+1}(i)=W_{t+1}(i)/[\sum_{j=1}^{n}W_{t+1}(j)]$;
  **until** $((t\geq T)$ OR $(\epsilon_t=0)$ OR $(\epsilon_t\geq 0.5))$;
  Use ensemble components with weights $\alpha_t$ for test instance classification;
**end**

### 3.2 提升方法权重调整

由于提升方法有一个数据样本的权重调整过程，我们需要探索如何将权重对模型的训练的作用体现出来，有如下两种方法可以达到效果。

一是重采样法，对于一些无法接受带权样本的及学习算法，适合用“重采样法”进行处理。方法大致过程是，根据各个样本的权重，对训练数据进行重采样，初始时样本权重一样，每个样本被采样到的概率一致，每次从 N 个原始的训练样本中按照权重采样 N 个样本作为训练集，然后计算训练集错误率，然后调整权重，重复采样，集成多个基学习器。

二是重赋权法，对每个样本附加一个权重，对离散值来说，在计算条件概率的贝叶斯估计时，不再是简单将其出现次数相加，而是带有各个样本权重的相加；对于连续值来说，样本权重的增大表现为均值向该样本做更多的偏移，实际上可以直接提高该样本该属性该连续值的数值大小来实现。

笔者开始采用的是重采样法实现 Adaboost 训练，因为这种方式简单，不需要改变数据的值，但是后来发现，这种方式做出来效果并不好，甚至随着基分类器的增多，Adaboost 的效果呈下降趋势！仔细检查过代码，应该没有问题，翻阅资料得知，两种方法除了有固定的可使用的场景外，还有各自的适用场景，样本集重采样法(resampling)对于不稳定的算法能够取得很好的效果，不稳定算法指的是样本集的微小变动就能够对结果有很大的影响的算法，很遗憾，朴素贝叶斯是稳定的算法，并且其估计的参数少，对缺失数据都不是很敏感。还有一些算法如 Linear Model, SVM, kNN 等都是稳定的算法，而不稳定算法的代表是决策树和神经网络两位。

最后，迫于效果太差，于是又转用了重新赋权的方式，将权重嵌入到数据中，最终结果显示还好，Adaboost 较单个朴素贝叶斯基分类器有提升，虽然提升并不是很大。

代码中两种实现方式都有，最终测试使用的 adaboost_weight(), get_model_weight(),get_err_rate_weight() 等方法，即重赋权的方法。

## 4 实验结果

最终在两个数据集上都做了 10 折交叉验证实验，实施交叉验证前程序都将整个数据集顺序打乱，所以每次运行程序的结果都会有变化，但是总体上还是能够体现结果，选取某次实验结果如下表

表 1. 交叉验证某次结果

| 数据集 | | 交叉验证平均准确率 (mean acc) | 交叉验证准确率方差 (variance) |
|---|---|---|---|
| **breast-cancer-assignment5.txt** | 单个 NB 分类器 | 0.7291 | 0.001400 |
| | Adaboost (best) | 0.7617 | 0.003646 |
| **german-assignment5.txt** | One NB classifier | 0.7410 | 0.000934 |
| | Adaboost (best) | 0.7580 | 0.000636 |

为了更加直观地观察 Adaboost 的提升效果，将准确率随 Adaboost 基分类器数的增加呈现的趋势通过图像展示出来，方便观察，下面图 1，图 2 分别给出在两个数据集上的提升情况

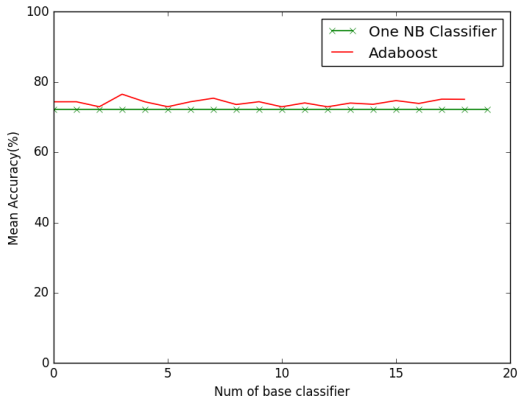


图 1. breast-cancer 数据集提升状况

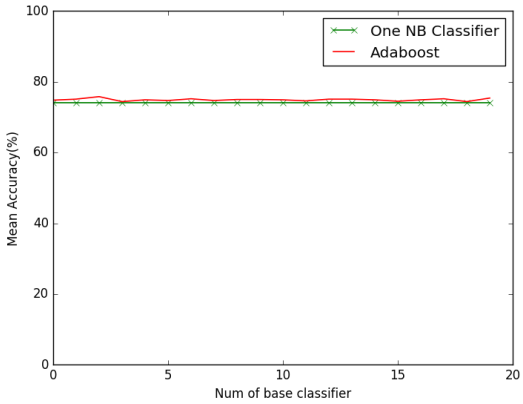


图 2. german-assignment 数据集提升状况

## 5 结束语

本次实验完成了集成学习的实验，以朴素贝叶斯作为基分类器，实现了经典的提升算法 Adaboost，交叉验证实验结果证明 Adaboost 能够以朴素贝叶斯分类器进行集成学习，并提升单个朴素贝叶斯分类器的效果，但是由于朴素贝叶斯已经是一个不错的分类器，单个朴素贝叶斯分类器的交叉验证平均准确率已经能够达到 0.74-0.75，尤其是第二个数据集，由于数据集较大，所以概率估计和分类效率都较为稳定，使得单个 NB Classifier 准确率能够达到很高，所以提升效果显得不那么明显，Adaboost 可能更适合于提升较弱的分类器的效果。

另外，虽然朴素贝叶斯的准确率很高，但是他是有瓶颈的，它的模型完全是根据数据集而得来，所以感觉没有什么需要学习的，完全是统计的方法，并且它有一个固有的弱点就是做了属性独立性假设，而现实条件往往不能满足属性独立性，尤其是数据属性冗余性大的时候，朴素贝叶斯的效果不会太好。

**References**:

[1] 李航.统计学习方法.

[2] Jiawei Han et. al.数据挖掘：概念与技术.

[3] 周志华.机器学习.