比较全面的Adaboost算法总结（一）

原创 石头 机器学习算法那些事 2018-12-04

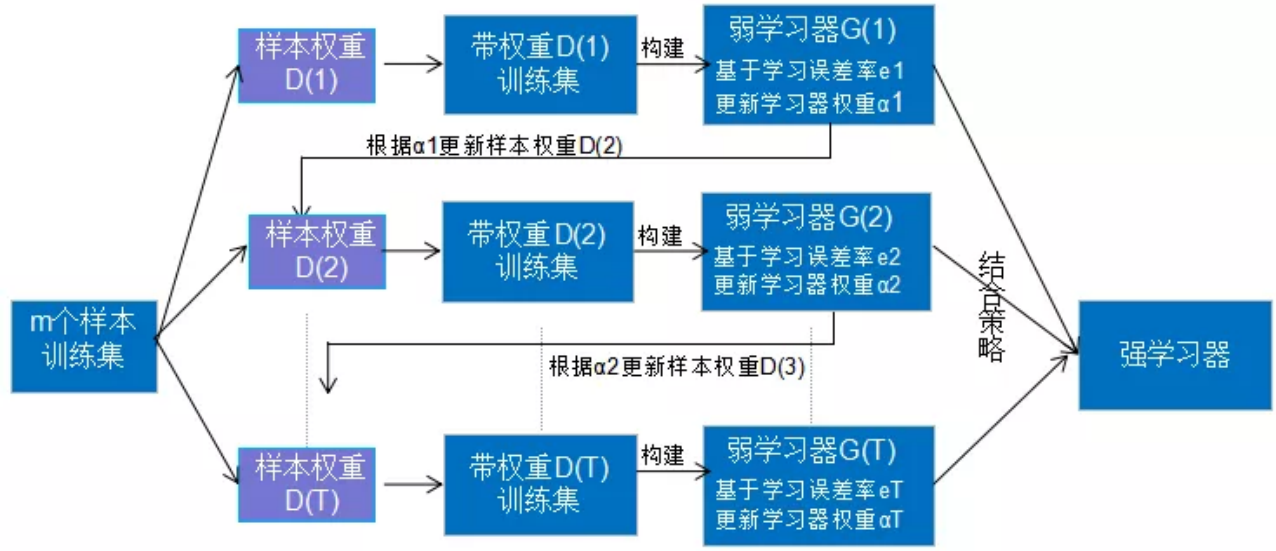
前言

集成学习的Boosting算法串行生成多个弱学习器并按一定的结合策略生成强学习器，AdaBoost算法是Boosting系列算法中的一种，本文详细总结了AdaBoost算法的相关理论。

目录

Boosting的算法流程

Boosting算法是一种由原始数据集生成不同弱学习器的迭代算法，然后把这些弱学习器结合起来，根据结合策略生成强学习器。



如上图，Boosting算法的思路：

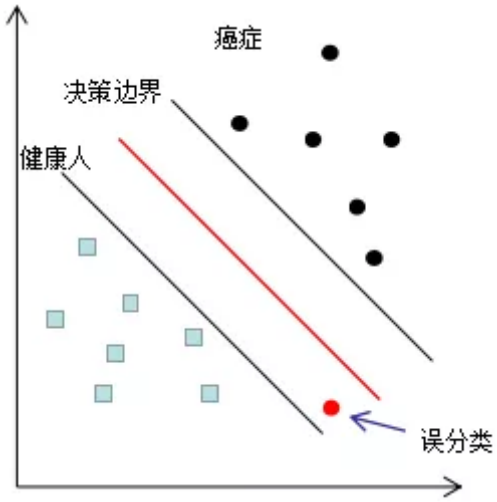
（1）样本权重表示样本分布，对特定的样本分布生成一个弱学习器。
（2）根据该弱学习器模型的误差率e更新学习器权重α。
（3）根据上一轮的学习器权重α来更新下一轮的样本权重。
（4）重复步骤(1)(2)(3)，结合所有弱学习器模型，根据结合策略生成强学习器。

Boosting算法的权重理解

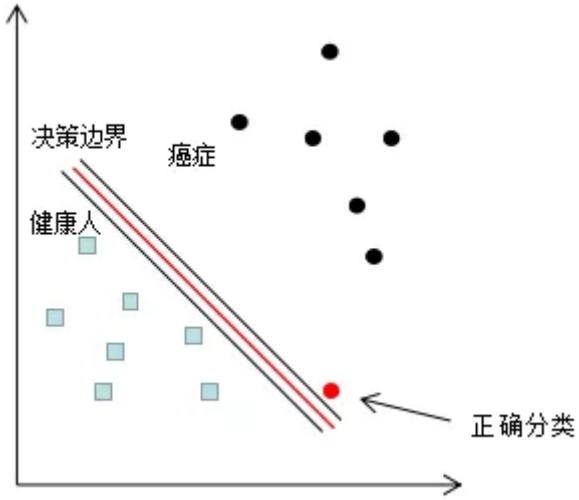
Boosting算法意为可提升算法，可提升方法具体表现在（一）改变训练数据的概率分布（训练数据的权值分布），（二）弱分类器权重的生成。理解这两个原理是理解AdaBoost算法的基础。

1. 训练数据的权重理解

我们对癌症病人和健康人作一个定性的分析，目的是理解Boosingt算法训练数据权重更新的思想。如下图为分类器G(1)的分类情况，假设样本数据的权重相等。



癌症误分类成健康人的结果很可能是丧失生命，因此这种误分类情况肯定不能出现的，若我们对该误分类点的权重增加一个极大值，以突出该样本的重要性，分类结果如下图：



因此，增加误分类样本的权重，使分类器往该误分类样本的正确决策边界方向移动，当权重增加到一定值时，误分类样本实现了正确分类，因为训练样本的权重和是不变的，增加误分类样本权重的同时，也降低了正确分类样本的权重。这是Boosting算法的样本权重更新思想。

2. 弱学习器的权重理解

Boosting算法通过迭代生成了一系列的学习器，我们给予误差率低的学习器一个高的权重，给予误差率高的学习器一个低的权重，结合弱学习器和对应的权重，生成强学习器。弱学习器的权重更新是符合常识的，弱学习器性能越好，我们越重视它，权重表示我们对弱学习器的重视程度，即权重越大，这是Boosting算法弱学习器权重的更新思想。

AdaBoost的算法流程

第一节描述了Boosting算法的流程，但是没有给出具体的算法详细说明：

（1）如何计算弱学习器的学习误差；

（2）如何得到弱学习器的权重系数α；

（3）如何更新样本权重D；

（4）使用何种结合策略；

我们从这四种问题的角度去分析AdaBoost的分类算法流程和回归算法流程。第k轮的弱分类器为 $G_k(x)$，且训练数据集在第K轮训练样本的权重分布为：

$$D(k)=(w_{k1},w_{k2},\dots,w_{km});w_{1i}=\frac{1}{m},i=1,2,\dots,m$$

其中m表示训练数据集的大小，w_{1i}表示训练数据集的初始权重

w_{ki}表示训练数据集第k轮迭代的样本权重，

1. AdaBoost的分类算法流程

我们假设是二分类问题，输出为{-1,1}。第K轮的弱分类器为 $G_k(x)$

1) 计算弱分类器的分类误差

在训练集上的加权误差率为：

$$e_k=P(G_k(x_i)\neq y_i)=\sum_{i=1}^{m}w_{ki}I(G_k(x_i)\neq y_i)$$

其中，符号I表示指示函数

2) 弱学习权重系数α的计算

$$\alpha_k=\frac{1}{2}\log\frac{1-e_k}{e_k}$$

由上式可知，学习器误差率越小，则权重系数越大。

3) 下一轮样本的权重更新

$$w_{k+1,i}=\frac{w_{ki}}{Z_k}\exp(-\alpha_k G_k(x_i)y_i)$$

其中Zk是规范化因子，使每轮训练数据集的样本权重和等于1。

$$Z_k = \sum_{i=1}^{m} w_{ki} \exp(-\alpha_k G_k(x_i) y_i)$$

当样本处于误分类的情况，$G_k(x)y(i) = -1$，则 $w_{k+1,i} > w_{k,i}$，该误分类样本的权重增加；

当样本是处于正确分类的情况，$G_k(x)y(i) = 1$，则 $w_{k+1,i} > w_{k,i}$，该正确分类样本的权值减小。

4） 结合策略，构建最终分类器为：

$$G(x) = sign(\sum_{m=1}^{K} \alpha_m G_m(x))$$

2. AdaBoost的回归算法流程

1) 计算弱学习器的回归误差率：

a) 计算训练集上的最大误差：

$$E_k = \max |y_i - G_k(x_i)|, i = 1,2,...m$$

b) 计算每个样本的相对误差：

如果是线性误差，则

$$e_{ki} = \frac{|y_i - G_k(x_i)|}{E_k};$$

如果是平方误差，则

$$e_{ki} = \frac{(y_i - G_k(x_i))^2}{E_k^2}$$

如果是指数误差，则

$$e_{ki} = 1 - \exp(\frac{-|y_i - G_k(x_i)|}{E_k})$$

c) 计算回归误差率

$$e_k = \sum_{i=1}^{m} w_{ki} e_{ki}$$

（2） 弱学习权重系数α的计算

$$\alpha_k = \frac{e_k}{1 - e_k}$$

（3）下一轮样本的权重更新

$$w_{k+1,i} = \frac{w_{ki}}{Z_k} \alpha_k^{1-e_{ki}}$$

Zk是规范化因子，使样本权重的和为1，

$$Z_k = \sum_{i=1}^{m} w_{ki}\alpha_k^{1-e_{ki}}$$

（4）结合策略，构建最终学习器为：

$$f(x) = \sum_{k=1}^{K}(\ln\frac{1}{\alpha_k})g(x)$$

其中，$g(x)$是所有$\alpha_k G_k(x)$的中位数，$k = 1,2,...,K$

AdaBoost算法的训练误差分析

过程就不推倒了，可参考李航《统计学习方法》P142~P143，这里就只给出结论。

AdaBoost的训练误差界：

$$\frac{1}{N}\sum_{i=1}^{N} I(G(x_i) \neq y_i) \leq \exp(-2M\gamma^2)$$

其中 $\gamma > 0$，M表示弱分类器个数，N表示训练样本集的个数

由上式可知，AdaBoost的训练误差是以指数速率下降的，即AdaBoost算法随着迭代次数的增加，训练误差不断减小，即模型偏差显著降低 。

本文倾向于入门AdaBoost算法，下一篇文章会发散思维，介绍AdaBoost算法的相关性质，

参考：

https://www.cnblogs.com/pinard/p/6133937.html

李航《统计学习方法》

推荐阅读文章

集成学习原理总结

比较全面的随机森林算法总结

【实践】随机森林算法参数解释及调优

长按二维码关注

机器学习算法那些事

微信：beautifulife244

砥砺前行 不忘初心