提升方法 AdaBoost 法 1: 算法介绍

2021 年 3 月 23 日

1 基本思路与概念

1.1 强可学习

在概率近似正确 (probably approximately correct，PAC) 学习框架中，如果一个类存在一个多项式算法能够学习它，并且正确率很好，那么就称这个概念为强可学习。

1.2 弱可学习

相反，如果能够学习它，但是正确率仅仅比随机猜测好一点点，那么就称这个概念为弱可学习。

1.3 弱和强是等价的

Schapire 证明了强可学习和弱可学习是等价的。

1.4 两个问题

对于分类而言，给定一个训练样本，求比较粗糙的分类算法比求更加精确的分类算法要简单很多。提升算法就是从弱分类出发，然后构造一系列弱分类算法，并将这些弱分类器组合成一个强分类器。这一系列弱分类器大多是通过改变样本的分布或者称为权值（一个意思），来构建。

1.4.1 问题 1

每一轮如何改变权值或者分布？

1.4.2 问题 2

最后如何将这些弱分类器进行组合，使其称为强分类器？

2 AdaBoost 算法

假设为二分类问题。

Algorithm 1 AdaBoost 算法.

Input:

训练数据集 $T = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$ 其中 $x_i \in X = R^n, y_i \in Y = \{-1, +1\}, i = 1, 2, ..., N$; 弱学习算法 $G_1(x)$

Output:

强学习算法 $F(x)$。

1: 初始化权值分布

$$D_1 = (w_{1,1}, w_{1,2}, ..., w_{1,N}), \quad w_{1,i} = \frac{1}{N}, \quad i = 1, 2, ..., N$$

其中 D_m 表示第 m 次训练弱分类器对应的样本分布（权值），$w_{m,i}$ 表示每个样本的权值，初始化为均匀初始化。

2: 对每一次（m）的弱分类器训练都执行如下步骤：

(A).

使用当前数据分布 D_m 的数据，训练弱分类器 $G_m(x)$

$$G_m(x) : X \to \{1, -1\}$$

(B).

计算所得弱分类器 $G_m(x)$ 的分类误差：

$$\begin{aligned} e_m &= \sum_{i=1}^{N} P(G_m(x_i) \neq y_i) \\ &= \sum_{i=1}^{N} w_{wi} I(G_m(x_i) \neq y_i) \\ &= \sum_{G_m(x_i \neq y_i)} w_i \end{aligned}$$

误差率就是将分类错误的那些样本的权重相加即可。

(C).

由误差率计算本次的弱分类器在最终分类器中的系数 α_m, 其表示在最终该弱分类器 $G_m(x)$ 在最终分类器中的重要程度，准确率越高权重越大。

$$\alpha_m = \frac{1}{2} \log \frac{1 - e_m}{e_m}$$

可以看出，是由误差率求出的。

得出本次的 $f_m(x) = \alpha_m G_m(x)$ (D).

为下次训练 $G_{m+1}(x)$ 调整数据的权重或者分布, 得到 D_{m+1}，调整策略根据样本预测的误差率来进行，分错的权重调大，分正确的权重调小。

$$w_{m+1,i} = \frac{w_{m,i}}{Z_m} \exp\left(-\alpha_m y_i G_m(x_i)\right), \quad i = 1, 2, ..., N$$

其中可以看出，$y_i G_m(x_i)$ 在分正确的情况下是等于-1 的，分错误等于 1。最终会将分正确的样本权重下调，错误的权重上调。

3: 构建基本分类器的线性组合：

$$F(x) = \sum_{m=1}^{M} \alpha_m G_m(x)$$

Input:

Output:

如果 $F(x)$ 的分类错误率达到要求，则终止，否则继续第 2 部，直到 $F(x)$ 达到要求。

OVER