AdaBoost

摘要：本文主要讲述的是数据挖掘的一种分类算法 AdaBoost 的背景，应用场景以及算法的主要训练过程。同时也对 AdaBoost 算法的特点和优点进行了分析。

一、 背景介绍：

AdaBoost 是一种监督学习的方法，同时 AdaBoost 也是一种元算法（元算法是对其他算法组合的一种方式），Boosting 算法是一种把若干个分类器整合为一个分类器的方法。Boosting 分类的结果是基于所有分类器的加权求和结果的，boosting 中的分类器的权重并不相等，每个权重代表的是其对应分类器在上一轮迭代中的成功度。

二、 主体内容：

2.1 应用范围：

AdaBoost 主要解决的问题有：两类问题，多类单标签问题，多类多标签问题，回归问题。

2.2 运行过程：

AdaBoost 是 adaptive boosting（自适应 boosting）运行过程如下：（基于错误提升分类器的性能）

训练数据中的每个样本，并赋予其一个权重，这些权重构成了向量 D。一开始，这些权重都初始化成相等值。

首先在训练数据上训练出一个弱分类器并计算该分类器的错误率，然后在同一数据集上再次训练弱分类器。在分类器的第二次训练当中，将会重新调整每个样本的权重，其中第一次分对的样本的权重将会降低，而第一次分错的样本的权重将会提高。

为了从所有弱分类器中得到最终的分类结果，AdaBoost 为每个分类器都分配了一个权重值 alpha，这些 alpha 值是基于每个弱分类器的错误率进行计算的，其中错误率 ε 的定义为：

$$\varepsilon = \frac{\text{未正确分类的样本数目}}{\text{所有样本数目}}$$

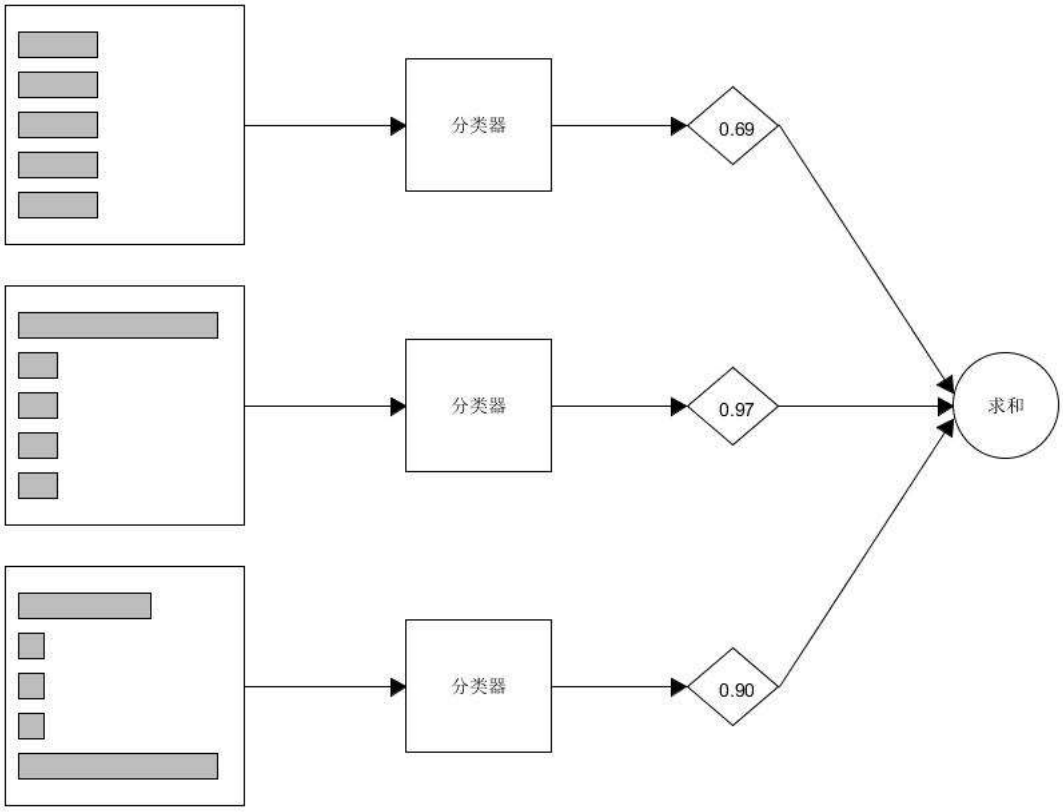
而 alpha 的计算公式如下：

$$\alpha = \frac{1}{2}\ln(\frac{1-\varepsilon}{\varepsilon})$$

2.3 算法运行流程：

AdaBoost 算法的流程如下图所示：

(1) 左边为数据集，其中直方图的不同宽度表示每个样例上的不同权重。在经过一个分类器之后，加权的预测结果会通过菱形中的 alpha 值进行加权。每个菱形中输出的加权结果在圆形中求和，从而得到最终的输出结果



(2) 计算出 alpha 值之后，可以对权重向量 D 进行更新，以使得那些正确分类的样本的权重降低二错分样本的权重升高。D 的计算方法如下：

A. 如果某个样本被正确分类，那么该样本的权重更改为：

$$D_i^{(t+1)} = \frac{D_i^{(t)} e^{-\alpha}}{Sum(D)}$$

B. 而如果某个样本被分错，那么该样本的权重更改为：

$$D_i^{(t+1)} = \frac{D_i^{(t)} e^{\alpha}}{Sum(D)}$$

(3) 在计算出 D 之后，AdaBoost 又开始进入下一轮迭代。AdaBoost 算法会不断地重复训练和调整权重的过程，直到训练错误率为 0 或者弱分类器的数目达到用户的指定值为止。

2.4 AdaBoost 算法特性

1. 训练的错误率上界，随着迭代次数的增加，会逐渐下降；

2. AdaBoost 算法即使训练次数很多，也不会出现过拟合的问题。

2.5 AdaBoost 算法分析

AdaBoost 的**特点**可以总结如下：

1. 每次迭代改变的是样本的分布，而不是重复采样；

2. 样本分布的改变取决于样本是否被正确分类；

3. 最终的结果是弱分类器的加权组合。

AdaBoost 的**优点**可以总结如下：

1. AdaBoost 是一种高精度的分类器；

2. 可以使用各种方法构建子分类器，AdaBoost 算法提供的是框架；

3. 当使用简单分类器时，计算出的结果是可以理解的，而且若分类器构造非常简单；

4. 不用担心过拟合问题。

三、 结语：

Adaboost 算法是一种实现简单，应用也很简单的算法。Adaboost 算法通过组合弱分类器而得到强分类器，同时具有分类错误率上界随着训练增加而稳定下降，不会过拟合等的性质，应该说是一种很适合于在各种分类场景下应用的算法。

四、 引用：

赵兴丽 西南大学计算机与信息科学学院重庆 AdaBoost 算法概述

http://blog.sina.com.cn/s/blog_6354bd9f0100y6cb.html

《机器学习实战》 [美] Peter Harrington