**boosting和bagging算法学习**

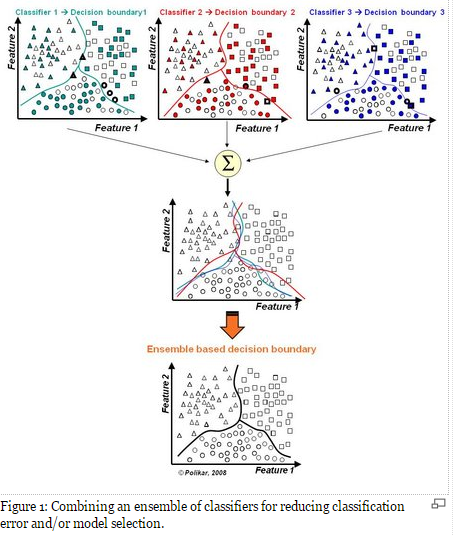
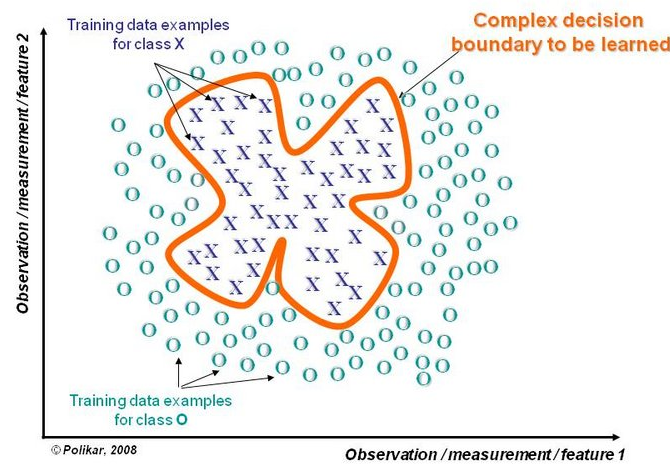
**https://blog.csdn.net/u010659278/article/details/44527437**

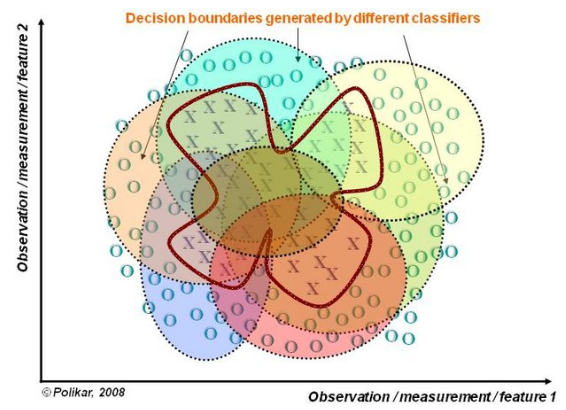
### ensemble learning

boosting和bagging都是集成学习（ensemble learning）领域的基本算法[[2](http://zh.wikipedia.org/zh-cn/AdaBoost)]。   
集成学习是指将若干弱分类器组合之后产生一个强分类器。弱分类器（weak learner）指那些分类准确率只稍好于随机猜测的分类器（error rate < 50%）。   
集成算法成功的关键在于能保证弱分类器的**多样性（diversity）**。集成不稳定的学习算法能得到更明显的性能提升。

#### ****为什么要集成？****

以下是Polikar给出的解释 [[3](http://www.scholarpedia.org/article/Ensemble_learning)]：

1. 模型选择（Model Selection）   
   假设各弱分类器间具有一定差异性（如不同的算法，或相同算法不同参数配置），这会导致生成的分类决策边界不同，也就是说它们在决策时会犯不同的错误。将它们结合后能得到更合理的边界，减少整体错误，实现更好的分类效果。   
   
2. 数据集过小或过大（Too much or too little data）   
   数据集较大时，可以分为不同的子集，分别进行训练，然后再合成分类器。   
   数据集过小时，可使用自举技术（bootstrapping），从原样本集有放回的抽取m个子集，训练m个分类器，进行集成。
3. 分治（Divide and Conquer）   
   若决策边界过于复杂，则线性模型不能很好地描述真实情况。因此先训练多个线性分类器，再将它们集成。   
   
4. 数据融合（Data Fusion）   
   当有多个不同数据源，且每个数据源的特征集抽取方法都不同时（异构的特征集），需要分别训练分类器然后再集成。



1. 数据融合（Data Fusion）

置信度估计（Confidence Estimation）   
可以为集成分类器计算分类结果的置信度。如使用多个弱分类器分类某个未知样本，且经过多数投票后得到的标签为y。如果投y的分类器越多，这个结果的置信度越高；反之则越低。

### bagging方法

bagging也叫自举汇聚法（bootstrap aggregating），是一种在原始数据集上通过有放回抽样重新选出S个新数据集来训练分类器的集成技术。也就是说这些新数据集是允许重复的。   
使用训练出来的分类器集合来对新样本进行分类，然后用多数投票或者对输出求均值的方法统计所有分类器的分类结果，结果最高的类别即为最终标签。

### boosting方法

分类器集合是在迭代中串行地产生的。训练时着重关注训练集中那些不容易区分的样本。AdaBoost是一种常见的boosting算法，下面是对算法的描述。

#### AdaBoost

即Adaptive boosting，是一种迭代算法。每轮迭代中会在训练集上产生一个新的分类器，然后使用该分类器对所有样本进行分类，以评估每个样本的重要性（informative）。   
具体来说，算法会为每个训练样本赋予一个权值。每次用训练完的新分类器标注各个样本，若某个样本点已被分类正确，则将其权值降低；若样本点未被正确分类，则提高其权值。权值越高的样本在下一次训练中所占的比重越大，也就是说越难区分的样本在训练过程中会变得越来越重要。   
整个迭代过程直到错误率足够小或达到一定次数为止。

**训练过程**

为每个样本初始化权值*w*=1*n*

；开始迭代，在第*t*轮迭代中：   
1. 使用训练集训练分类器*Ct*，训练误差*e*=所有被分类错误样本的权值之和   
2. 计算分类器的权值为*α*=12*ln*(1−*ee*)   
3. 更新样本当前的权值*wt*：   
- 若分类正确，则减少权值，*wt*+1=*wt*∗*e*−*α*；   
- 若分类错误，则加大权值，*wt*+1=*wt*∗*eα*

4. 将所有样本的权值归一化，使其相加为1

**分类过程**   
用生成的所有分类器预测未知样本X，最终结果为所有分类器输出的加权平均。

**分析**   
优点：泛化错误率降低，易实现，无需调整参数，可以处理不均衡的样本集。   
缺点：对于离群点很敏感，易过拟合。

参考了   
july的博客 <http://blog.csdn.net/v_july_v/article/details/40718799>   
<http://zh.wikipedia.org/zh-cn/AdaBoost>   
<http://www.scholarpedia.org/article/Ensemble_learning>   
《机器学习实战》