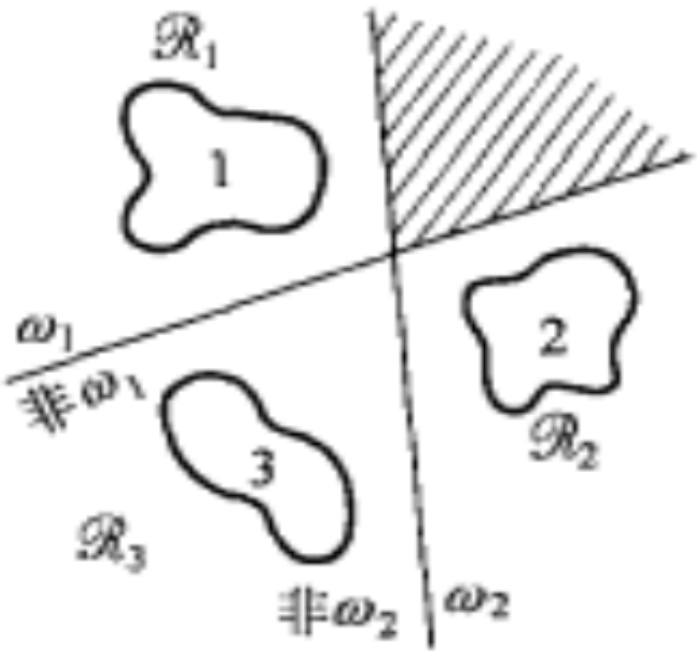
多类线性分类器

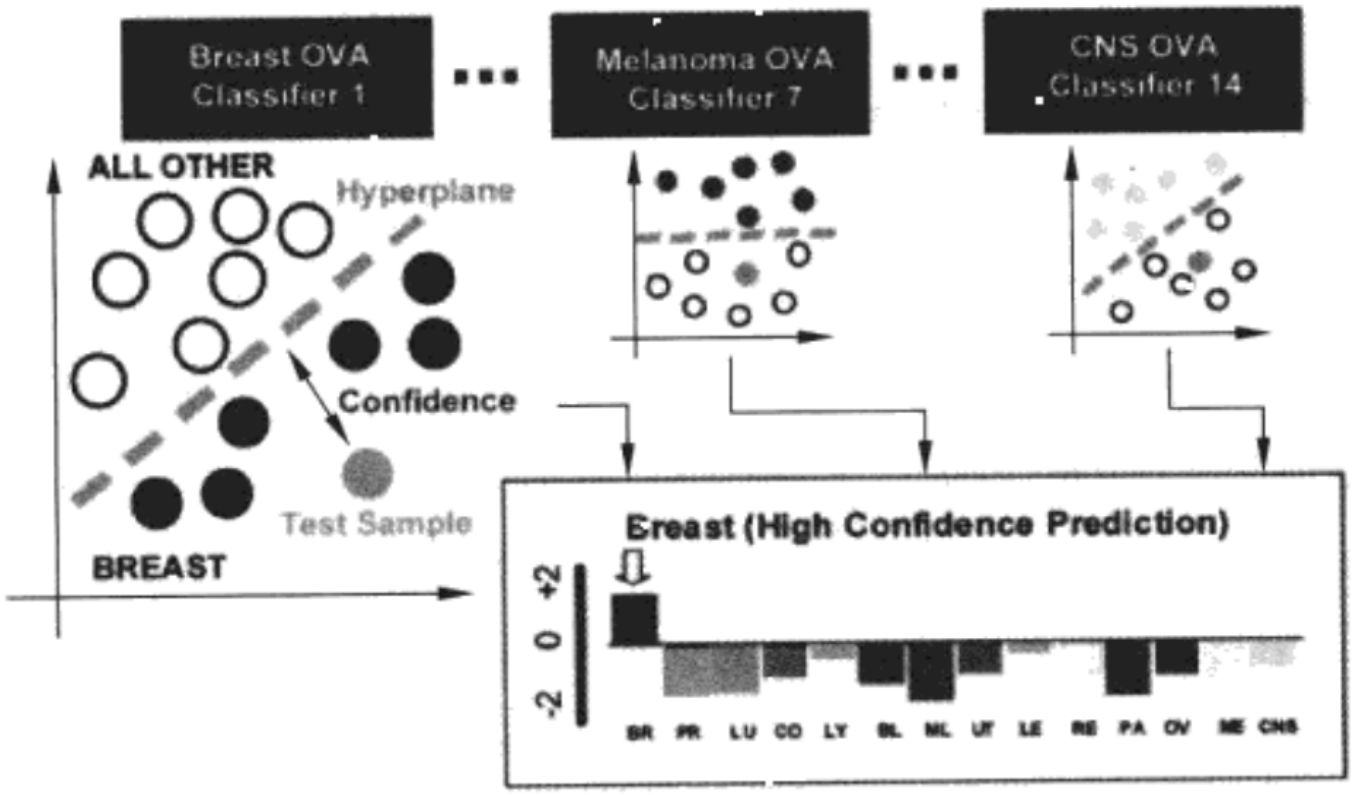
4.7.1 多个两类分类器的组合

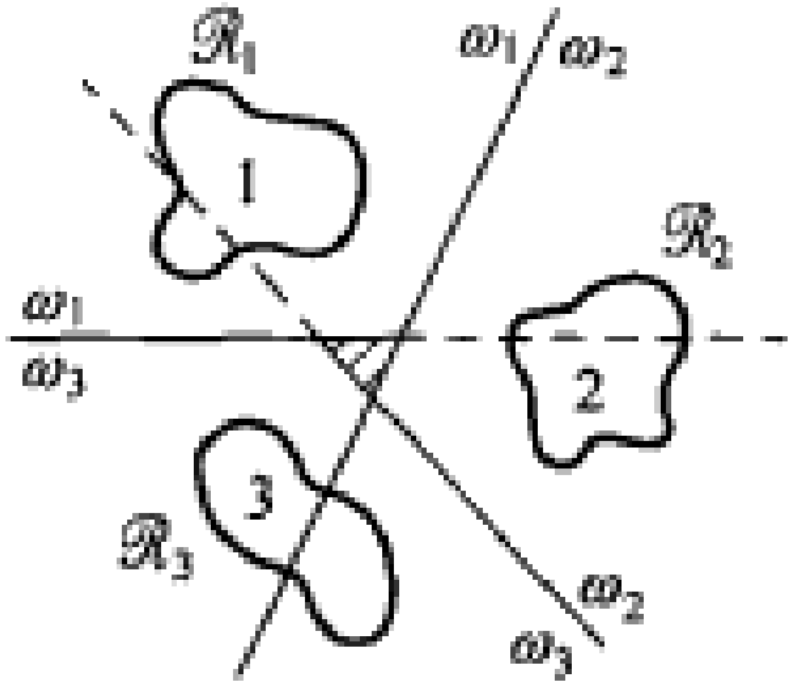
第一种做法叫做“一对多”的做法,英文可以叫one-vs-rest或者one-over-all。 假设共有c个类,我们共需要c-1个两类分类器就可以实现c个类的分类。



这种做法可能会遇到两方面的问题。一个问题是，假如多类中各类的训练样本数目相当,那么,在构造每个一对多的两类分类器时会面临训练样本不均衡的问题,即两类训练样本的数目差别过大。使得多数错误发生在样本数小的一类上。
另一个问题是,用c-1个线性分类器来实现c类分类,就是用c-1个超平面来把样本所在的特征空间划分成c个区域,一般情况下,这种划分不会恰好得到c个区域,而是会多出一些区域,而在这些区域内的分类会出现歧义。

第二种是 逐对 (pairwise)分类，对多类中的每两类构造一个分类器





共需要C_n^2个两类分类器。显然,这种做法要比一对多的做法多用很多两类分类器。但是,逐对分类不会出现两类样本数过于不均衡的问题,而且决策歧义的区域通常要比一对多分类器小。问题是，有多个classifier涉及第i类，分类器在最后的分类决策前得到的是一个连续的量,分类是对这个量用某个阈值划分的结果，比如所有线性分类器都是最后转化为一个线性判别函数g(x)=w'x+wo与某阈值(通常是0)比较的问题。SVM也是这样一种分类器。因此可以把分类器的输出值看作是对样本属于某一类别的一种打分,如果分值大于零(或其他阈值)则判断样本属于该类,而且分值越高对此分类越确信,反之决策不属于该类。

利用这种分类器,可以用c个一对多的两类分类器来构造多类分类系统,即每个类别对应一个分类器,其输出是对样本是否属于w;类给出一个判断。在多类决策时，如果只有一个两类分类器给出了大于阈值的输出,而其余分类器输出均小于阈值,则把这个样本分到该类。更进一步,如果各个分类器的输出是可比的,而且根据类别的定义知道任意样本必定属于且仅属于c个类别中的一类,那么可以在决策时直接比较各个分类器的输出,把样本赋予输出值最大的分类器所对应的类别。(但是需要注意,对很多分类器来说，如果它们是分别训练的,其输出值之间并不一定能保证可比性。)

能够根据类别间的内在关系把它们分级合并成多个两类分类问题,则可以用二叉树来构建多个两类分类器。比如假如我们的目标是分出a、b、c.d、e、f六个类,如果发现这些类别的概念间有内在的关系,比如e、f两个类关系比较紧密,同属于一个更高层次的概念,c、d同属于一个概念,b和c、d又关系比较紧密,等等，则可以把问题分解成{a}对{b,c,d,e,f}、{b,c,d}对{e,f}.{b}对{c,d}、{c}对{d}、{e}对{f}这五个两类分类问题。