"记住"训练样本, 就是所谓的"机械学习" [Cohen and Feigenbaum, 1983], 或称"死记硬背式学习", 参见 1.5 节.

能力. 如果仅仅把训练集中的瓜"记住", 今后再见到一模一样的瓜当然可判断, 但是, 对没见过的瓜, 例如"(色泽=浅白) ∧ (根蒂=蜷缩) ∧ (敲声= 浊响)"怎么办呢?

我们可以把学习过程看作一个在所有假设(hypothesis)组成的空间中进行搜索的过程, 搜索目标是找到与训练集"匹配"(fit)的假设, 即能够将训练集中的瓜判断正确的假设. 假设的表示一旦确定, 假设空间及其规模大小就确定了. 这里我们的假设空间由形如"(色泽=?) ∧ (根蒂=?) ∧ (敲声=?)"的可能取值所形成的假设组成. 例如色泽有"青绿""乌黑""浅白"这三种可能取值; 还需考虑到, 也许"色泽"无论取什么值都合适, 我们用通配符"∗"来表示, 例如"好瓜 ↔ (色泽= ∗) ∧ (根蒂=蜷缩) ∧ (敲声=浊响)", 即"好瓜是根蒂蜷缩、敲声浊响的瓜, 什么色泽都行". 此外, 还需考虑极端情况: 有可能"好瓜"这个概念根本就不成立, 世界上没有"好瓜"这种东西; 我们用 ∅ 表示这个假设. 这样, 若"色泽""根蒂""敲声"分别有3、2、2 种可能取值, 则我们面临的假设空间规模大小为 $4 \times 3 \times 3 + 1 = 37$. 图 1.1 直观地显示出了这个西瓜问题假设空间.

这里我们假定训练样本不含噪声, 并且不考虑"非青绿"这样的 $\neg A$ 操作. 由于训练集包含正例, 因此 ∅ 假设自然不出现.

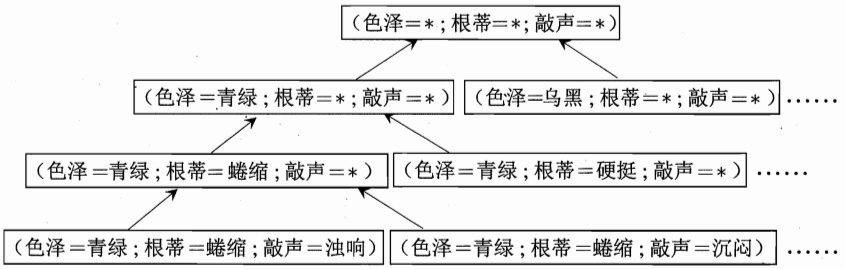

图 1.1 西瓜问题的假设空间

可以有许多策略对这个假设空间进行搜索, 例如自顶向下、从一般到特殊, 或是自底向上、从特殊到一般, 搜索过程中可以不断删除与正例不一致的假设、和(或)与反例一致的假设. 最终将会获得与训练集一致(即对所有训练样本能够进行正确判断)的假设, 这就是我们学得的结果.

有许多可能的选择, 如在路径上自顶向下与自底向上同时进行, 在操作上只删除与正例不一致的假设等.

需注意的是, 现实问题中我们常面临很大的假设空间, 但学习过程是基于有限样本训练集进行的, 因此, 可能有多个假设与训练集一致, 即存在着一个与训练集一致的"假设集合", 我们称之为"版本空间"(version space). 例如, 在西瓜问题中, 与表 1.1 训练集所对应的版本空间如图 1.2 所示.