数学无处不在 – 血液检测的数学模型和理论（三）

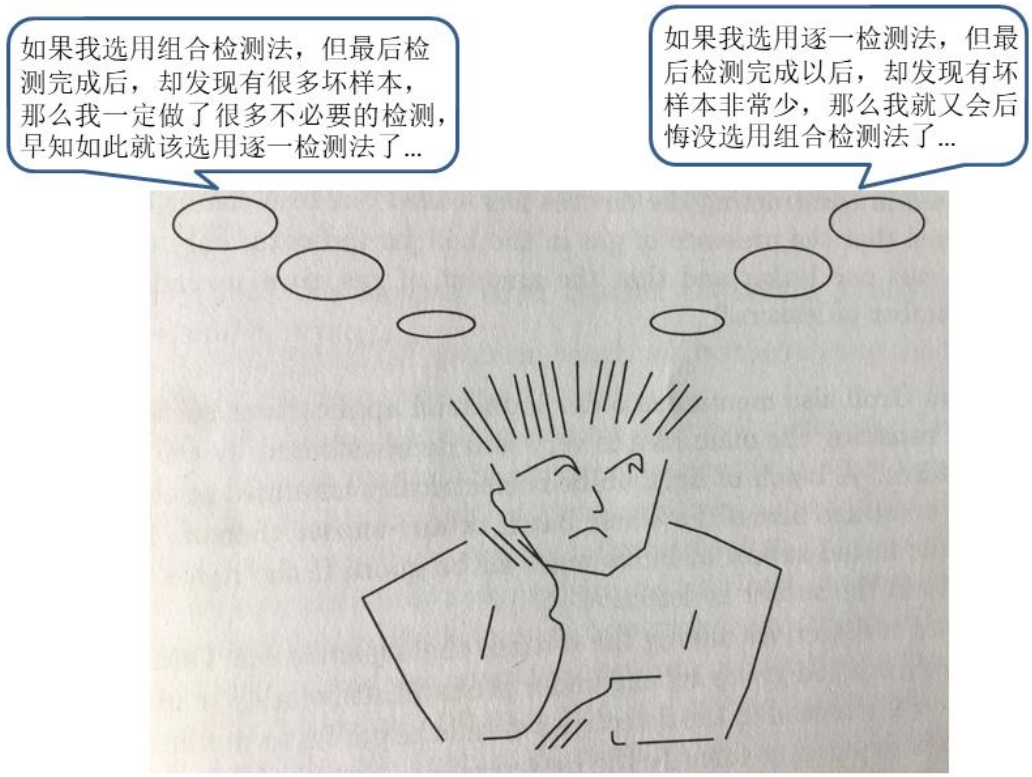
胡晓东
中国科学院数学与系统科学研究院

今天我继续给大家介绍与血液检测相关的数学模型和理论：如何运用数学的方法，用最少的成本（检验次数和时间），在大量的血液样本中准确地检验出哪些是阴性的（好的），哪些是阳性的（坏的）。

前两次我先后讲了两个模型，**概率模型**：假设事先已经知道所需检测的 **N** 个样本中有 pN 个是阳性的，其中 $0 \le p < 1$；**组合模型**：假设事先已经知道所需检测的 **N** 个样本中有 d 个是阳性的，其中 $0 < d < N$。同时介绍了相应的两个理论结果，当 p 或者 d 比较小的时候，组合检测法需要的检测次数比逐一检测法需要的检测次数少。

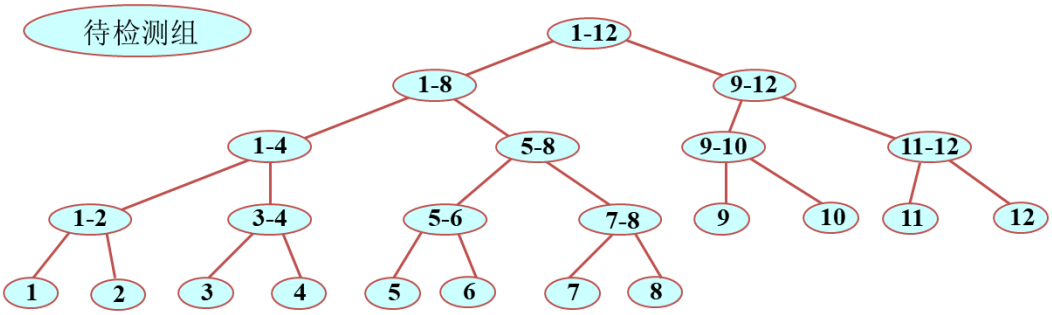
今天我来介绍，上述的概率模型和组合模型不再适用的情况，亦即，我们有 **N** 个（血液）样本需要检测，但既不知道其中好的样本多还是坏的样本多，更不知道究竟有多少个是坏的样本，我们该不该采用组合检测方法呢？针对这种情形，1993 年 D.-Z. Du（堵丁柱）与 F. K. Hwang（黄光明）[1] 合作建立了**竞争模型**：我们只有在检测出所有的坏样本以后，才能知道 **N** 个样本中有 d 个坏样本（可能 $d=0$）。

根据前两次的介绍和分析，大家已经知道，如果 **N** 个（血液）样本中含有的坏样本比较少，那么应该采用组合检测方法；否则应该采用逐一检测法。因而，当我们没有关于 **N** 个（血液）样本中坏样本个数的任何信息时，就会面临如下两难的选择：



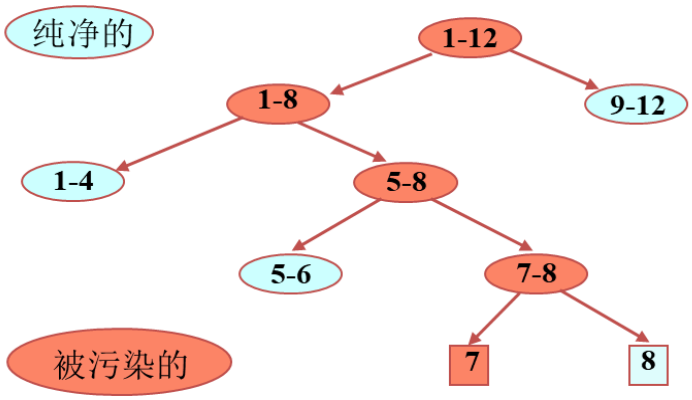
面对上述困境，堵丁柱和黄光明两位教授是如何在竞争模型下设计有效的组合检测法，并对其进行分析的呢？他们的方法是基于如下一棵二叉搜索树：

- 根节点在第一层（检测组含有所有样本），下面是第二层、第三层等等。
- 用**广度优先搜索**，在同一层的节点由左到右排序。
- 每一层的待检测组用**二分法**设计。在每一步，若检测出一个被污染的组 **X**，则将其进行二分，得到两个子组 **X'**和 **X"**，然后分别检测它们，其中 **X'**含有 $2^{\lceil \log|X| \rceil - 1}$ 个样本，而 **X" = X \ X'**。



上图给出了一个例子，其中 **N=12**，一共有 **23** 个待检测组。显然，当所有的样本都是好的时候，只需要检测第 **1** 组（因为该组含有所有样本，检测结果显示阴性，因而不需要再检测其他组）；然而，当所有的样本都是坏的时候，需要检测所有的 **23** 个组（因为每次检测结果，都显示阳性），需要的检测次数远远大于逐一检测法做需要的 **12** 次检测！

下图给出了当只有❼号样本是坏样本时，需要检测的 **9** 个组。（大家可以练习一下，当仅有❾号样本或者❿号样本是坏的时候，两种情形都需要做 **7** 次检测。）



通过上面的例子我们发现，这个组合检测法在所给样本中坏样本比较少的时候，用的检测次数比逐一检测法少，否则它比逐一检测法用的次数多。一个自然的问题就是，这个组合检测法到底好不好呢？应该如何分析它的有效性呢？为此堵丁柱和黄光明两位教授采用了竞争分析（competitive analysis）方法。

给定 **N** 个（标了号的）样本，$0 \le d \le N$，用 $S(N, d)$表示 **N** 个样本中有 ***d*** 个坏样本的所有可能的实例组成的集合。对于一个检测算法 **A**，用 $N_A(s)$表示它检测样本实例 **s** 所需要的检测次数，并定义

$$M_A(d \mid N) = \max \{N_A(s) \mid s \in S(N, d)\},$$

称算法 **A** 是 ***α*-竞争算法**如果存在一个常数 ***c*** 使得对任意 $0<d<N$，都有

$$M_A(d \mid N) \le \alpha M(d, N) + c,$$

其中 $M(d, N)$表示针对（未检测就）已知 **N** 个样本中含有 ***d*** 个坏样本的所有实例，**最优检测算法**所需要做的检测次数。（这里我们排除了 $d=0$ 和 $d=N$ 这两种情形，因为 $M(0, N)=0$ 和 $M(N, N)=0$ 而 $M_A(0 \mid N)>0$ 和 $M_A(N \mid N)>0$）。堵丁柱和黄光明两位教授证明了如下定理（大意是：他们的算法在事先不知道 ***d*** 的值所需要的检测次数不超过事先知道 ***d*** 的值所需要的检测次数的 **2** 倍。）

定理　**D**u-Hwang 算法 B 满足 $M_B(d \mid N) \le 2M(d, N) + 5$。

上述结果中的系数 **2** 先后又被黄光明与合作者[3]和堵丁柱与合作者[4]用更加精细的倍增（doubling）组合技巧改进到了 **1.65**。

实际上，竞争分析方法是 1985 年由 D. D. Sleator 和 R. E. Tarjan [2]最先提出，用于分析求解在线问题（online problems）的在线算法（online algorithm）的性能。他们假设已知若干种事件都有可能会发生，但事先又不知道最终哪一种事件将会发生，随着事件进程的不断发展，相关的信息逐步显现，在这个过程中，如何即时地做出一系列决策，使得事件的整个过程结束时，所得收益都不会太差或者费用不会太高。

股票交易可以视为一个在线问题。每天股市开盘后，经纪人很难准确地知道接下来的股市走势，但是他还需要根据已完成股票交易的信息，随时做出决策（不能等到收盘），并在收盘时给客户带来比较稳定的、比较好的收益。要做到这一点，他需要很好地考虑和平衡各种可能出现的交易进程，不能顾此失彼。

最后，竞争分析的核心也可以视为一个二人博弈：一方是在线玩家，另一方是邪恶的庄家。庄家一次出一张“牌”，玩家看到一张牌以后，就要做出选择（他不知道庄家以后会出什么牌）。庄家出牌的目的是，使得玩家做出的一系列选择所产生的成本越高越好，同时自己所需要的成本越低越好。

下一次我将给大家介绍一下，当允许检测出现错误：对一个没有坏样本的组进行检测，结果显示阳性（有坏样本），或者对一个有坏样本的组进行检测，结果却显示阴性（没有坏样本），我们又该如何快速地检测出坏样本？

参考文献

[1] D.-Z. Du and F. K. Hwang, Competitive group testing. Discrete Applied Mathematics, 45 (1993), 221-232.
[2] A. Bar-Noy, F. K. Hwang, I. Kessler and S. Kutten, Competitive group testing in high speed networks, Discrete Applied Mathematics, 52 (1994), 29-38.
[3] D.-Z. Du, G.-L. Xue, S.-Z. Sun and S.-W. Cheng, Modifications of competitive group testing, SIAM Journal on Computing, 23 (1994), 82-96.
[4] D. D. Sleator and R. E. Tarjan, Amortized efficiency of list updates and paging rules. Communications of the ACM, 28 (2) (1985), 202-208.