信息论在信息学竞赛中的简单应用

侯启明

引言

信息论是计算机科学中很重要的一个分支。虽然有关信息论的内容很少在以往各种关于信息学竞赛的材料里出现，但实际上，信息论作为证明某些问题解法最优性的理论基础，如果能在竞赛中适当地运用，往往可以取得事半功倍的效果。那么，什么是信息论呢？

信息论简介

信息论是关于信息的本质和传输规律的科学的理论。但是在竞赛中用到的主要是关于不确定性和信息的关系的知识。通过它可以很方便地得到某些交互式问题的一个较好的步数下界（这就是某些文献中所说的“信息论下界”）。不过，具体应该怎么做呢？让我们先来看一些信息论的核心理论：

定义：如果一个随机变量 x 共有 n 种取值，概率分别为 $p_1,p_2,.....,p_n$，则其熵为 $H(x)=f(p_1,p_2,.....,p_n)=\Sigma -Cp_i\log p_i$ （C 为正常数，一般取 1）

定理 1：在得到关于随机变量 x 的一个熵为 h 的信息后，x 的熵将会减少 h。

定理 2：当一个随机变量的各种取值概率相等时，它的熵最大。

这些理论看上去和某些题目关系密切，不是吗？那么，具体应该如何运用呢？让我们来看一些例子：

实际应用

例 1：验证一下定理 1。

我们宿舍二楼到三楼之间楼梯的窗户外面是相邻的一个平房的房顶。在那一带栖息着三只浑身雪白，一只蓝眼睛，一只绿眼睛的——猫！三只猫分别是一只胖猫，一只瘦而尾巴健全的猫和一只瘦而尾巴不健全的猫。在天冷的时候，它们喜欢趴在楼内的暖气上。于是，每只猫就有了两种状态——在屋内和在屋外。显然，三只猫的状态共有 8 种可能情况，假设它们是等概率的。现在，我在一楼的小卖部。由于种种原因，我希望知道猫当时的状况，因此，我往上看了一眼，结果发现在这个位置只能知道屋内猫的只数……

问题 1：把所有猫的情况作为一个随机变量 x，则当我在小卖部的时候，x 的熵是多少？

解答 1：由于 8 种情况的概率相等，所以 H(x)=f(1/8,1/8,1/8,1/8,1/8,1/8,1/8,1/8)=log8

问题 2：我看一眼所得到的信息 y 的熵是多少？

解答 2：由于猫的只数共有 0,1,2,3 四种情况，概率分别为(1/8,3/8,3/8,1/8)，所以：

H(y)=f(1/8,3/8,3/8,1/8)=-

(1/8*log(1/8)+3/8*log(3/8)+3/8*log(3/8)+1/8*log(1/8))=log8-6log(3/8)

问题 3：我看完之后，x 的熵 H'(x)是多少？

解答 3：此时猫的只数为 0,1,2,3 的四种情况的概率依次是(1/8,3/8,3/8,1/8)，而每种情况的熵分别为(0,log3,log3,0)，所以此时 H'(x)的数学期望为：

H'(x)=1/8*0+3/8*log3+3/8*log3+1/8*0=6log3

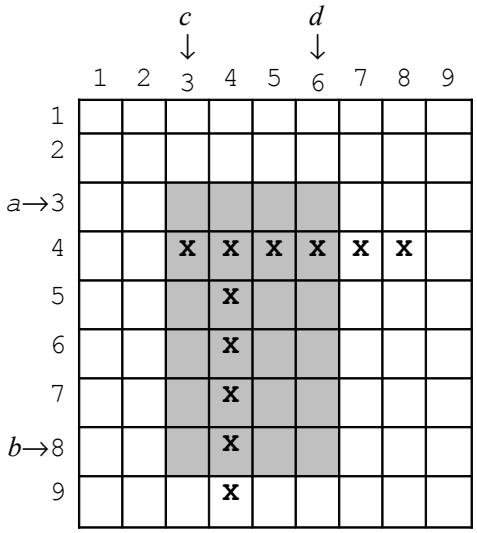
可以发现 H(x)=H(y)+H'(x)。

例 2：Rods(IOI2002)

一个Rod是一个由至少2个单位正方形连成的水平或竖直的长条。在一个N*N的方阵中，放了水平和竖直两个Rod。在图1中，Rod用X表示。两个Rod可能有公共方格，比如在图1中，

方格（4，4）无法确定是仅属于1个Rod还是同时属于两个Rod。因此，我们在这种情况下假定它同时属于两个Rod。这样，图中竖直Rod的上端点是（4，4）而不是（5，4）。

图1



最初我们并不知道两个 Rod 的位置，所以你的任务是编程序把它们的位置找出来。你只能通过库函数 rect(*a*,*b*,*c*,*d*)来定位两个 Rod。该函数检验矩形[*a*,*b*] [*c*,*d*](如图 1 中阴影区域)。注意参数的顺序，该函数要求 *a* *b*，*c* *d*。如果至少一个属于某个 Rod 的方格落在矩形[*a*,*b*] [*c*,*d*]内的话，rect 返回 1，否则返回 0。

考试当时我很快想到了一个最大步数为 $6\log_2 n+C$（某个常数）的方法，但是因为这个数差不多刚好达到步数限制，所以我在这时就开始试图优化步数中 $\log_2 n$ 的系数，结果徒劳无功，反而耽误了时间，最后才发现丢分不是因为步数超限而是因为少考虑了一些特殊情况。因此，看过答案以后，我试着从信息论的角度分析了一下这个问题：

由于题目中没有涉及到概率，因此

假设所有情况都是等概率的。所以，设 Rod 的摆放方法为随机变量 x，x 所有可能的取值数为 f(n)，那么 x 的熵 H(x)就等于 log(f(n))。而由于库函数只有两种可能的返回值，其熵最大为 $H_{max}(y)=\log 2$。因此，询问次数的信息论下界就是 $L=H(x)/H_{max}(y)=\log(f(n))/\log 2=\log_2 f(n)$。

下面讨论 f(n)的值：在 n*n 的方阵中放 1 个 Rod（无论横竖）共有 n*C(n+1,2)种方案，放两个相交的 Rod 共有 C^2 (n+2,3)种方案，所以：

$f(n)=(n^2(n+1)/2)^2-((n+2)(n+1)n/6)^2=(2n^6+3n^5-n^4-3n^3-n^2)/9$

当 n 充分大时：$L=\log(f(n))/\log(2)>\log_2(2n^6/9)$约$=6\log_2 n-2.2$

由于各种原因，不一定总是使两种返回值概率相等，所以最坏情况下的调用次数往往达不到信息论下界，两者大约相差一个常数，因此，可以认为 $6\log_2 n+C$ 是 rect 函数最大调用次数的下界。这样，在得到一个这样的算法之后，就没有什么必要再去徒劳地优化步数，完全可以把时间花在检查对特殊情况的处理上或干脆开始做别的题了。

例 3：Coins(选手推荐题 0024，推荐者饶向荣)

以前下面的难题曾被提议为地方数学奥林匹克竞赛：

有一堆 15 个硬币，其中有 14 个好的，一个坏的。所有好的硬币的质量是相同的，但坏的硬币的质量却不一样，现在告诉你某一枚是好的，要你用一架天平称出哪个是坏的硬币。

看你是不是可以在3次比较之内，找出坏的硬币。

后来此题引申为：

有一堆n个硬币，其中有n-1个好的，一个坏的。所有好的硬币的质量是相同的，但坏的硬币的质量却不一样，现在告诉你某一枚是好的，要你用一架天平称出哪个是坏的硬币。看你是不是可以在k次比较之内，找出坏的硬币。

输入n和k，如果能在k此比较中找到n枚硬币中的哪枚为坏的，就输出'POSSIBLE'，否则输出'IMPOSSIBLE'。

两年前我的一位远房亲戚曾给我出过一个类似的题目（n=12，k=3，但开始时不知道哪一枚硬币是好的），当时我苦苦思索了一晚上，终于想出来一个可行的解法。于是，那位亲戚加大了数据规模（n=1100，k=7，IMPOSSIBLE），我想了大概一周，觉得好像无解，但苦于无法证明我的解法的最优性，始终不能理直气壮地回答“IMPOSSIBLE”，只好回答说我最多做到n=1093，k=7。后来她给了我一个“说明”，但我始终觉得不太严密；拿来问我们班的IMO金牌，他的回答是“显然”，我觉得也不严密:(。于是，这件事就成了我这两年的一个遗憾。现在，有了信息论的武器，这个遗憾终于得到了解决！

首先，对硬币用1到n进行编号，设坏硬币的编号为x。同样假设x的所有取值情况概率相等：

∴H(x)=logn。

∵用天平称一次的结果y只有3种可能情况（左边较重，右边较重，平衡）

∴Hmax(y)=log3。

∴从n个硬币中通过天平找出一个坏硬币至少需要H(x)/Hmax(y)=log3n步（结论1）

虽然在本题的条件下，这个信息论下界是达不到的，但是如果没有这个结论，真正最优的解法的最优性就无从证明。得出这个结论后，证明中的困难就迎刃而解了。

进一步的分析：

设在有足够多的已知的好硬币的情况下，k次比较最多从g(k)个硬币中找出一个重量未知的坏硬币。

当k=1时，通过枚举可以发现，g(1)=2。

当k>1时：

考虑第一次比较结果是平衡的情况，设t为这次没上天平的尚未确定好坏的硬币的个数，由于可以通过剩下的k-1次比较把坏硬币从这t个硬币中找出来，所以t<=g(k-1)。

现在考虑第一次比较结果不是平衡的情况，此时可以确定坏硬币在上了天平的g(k)-t个硬币中，同样由于可以通过剩下的k-1次比较把坏硬币从这g(k)-t个硬币中找出来，根据结论1，得到：$g(k)-t<=3^{k-1}$

综合起来，得到$g(k)<=g(k-1)+ 3^{k-1}$

现在通过构造一种称法来证明$g(k)=g(k-1)+ 3^{k-1}$：

第一次比较第1到3^{k-1}号硬币和3^{k-1}个好硬币，分以下情况讨论：

平衡：说明坏硬币在剩下的g(k-1)个硬币中，由g的定义，可以在k-1步内找出。

好球较轻：说明坏硬币在这3^{k-1}个硬币中，且坏硬币较重。此后每次把所有硬币分成三等份，比较其中两份，如果平衡，说明坏硬币在第三份中，否则坏硬币就在重的一份中，这样就把坏硬币的范围缩小成了原来的三分之一。这个步骤重复k-1次之后，刚好找出坏硬币。

好球较重：与上一种情况类似，不再赘述。

构造完毕

这样，根据$g(k)=g(k-1)+ 3^{k-1}$，计算得出：$g(k)=(3^k+1)/2$

于是，得到结论2：

结论2　在有足够多的已知的好硬币的情况下，k次比较最多从$g(k)=(3^k+1)/2$个硬币中找出一个重量未知的坏硬币。

现在，开始解决原问题：

设在有一个已知的好硬币的情况下，k次比较最多从f(k)个硬币中找出一个重量未知的坏硬币。

显然，$f(k)<=g(k)$。

现在通过构造一种称法，来证明$f(k)=g(k)$：

设硬币1为已知好硬币：

第一次比较1号到$(3^{k-1}+1)/2$号硬币和$(3^{k-1}+1)/2+1$号到$3^{k-1}+1$号硬币：

平衡：比较$3^{k-1}+2$号到$2*3^{k-1}+1$号硬币和1到3^{k-1}号硬币，后面步骤参考有足量的好硬币的情形

不平衡：比较$(3^{k-1}+1)/2+1$号到$(3^{k-1}+3^{k-2})/2$号硬币、$(3^{k-2}+1)/2+1$号到$(3^{k-1}+1)/2$号硬币和1号到$(3^{k-2}+1)/2$号硬币加上足量的好硬币

平衡：:转化成k小1的问题

不平衡：:确定坏硬币在某3^{k-1}个硬币中，且知道是它比好硬币轻还是重

构造完毕。原问题解决，k次最多在$n=f(k)+1=(3^k+3)/2$个硬币中称出一个坏硬币。

总结

细心的读者应该会注意到，本文中的例题不用信息论的知识都可以解决。那么，信息论在这里的意义是什么呢？实际上，作为一种纯粹的理论，信息论并非是解决具体题目的手段，而是用来对一类问题进行分析的通用工具。它可以为我们的解法提供强有力的理论依据，更可以通过估计上下界来指导解法的构造。综上所述，信息论在信息学竞赛中是大有用武之地的。