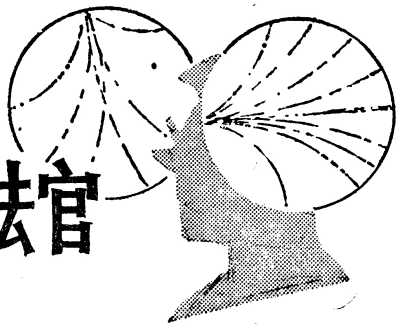
模式识别——判断基本粒子的法官



王 泰 杰

在高能物理实验中，鉴别粒子的类型十分重要，但又谈何容易．高能粒子在探测器内留下的踪迹，用“电光石火”之类的词汇还不足以形容，能够感知它们的存在已经是很艰难的事情，更何况确切地判断它们的“身份”呢!

判断粒子的“身份”是实验数据离线分析中的任务之一，判断的手段是一种叫做模式识别的方法．

从鉴别照相说起

判断粒子的“身份”与判断人的身份确有相似之处．让我们先设想一下物理学家判断粒子“身份”时手里掌握的是什么：经过对数据的初步分析，选出了许多粒子，而且测定了每个粒子的一些性质．现在的任务是根据它们的性质判断它们的种类——“身份”．这好比是手头有一叠人象照相，要根据每张照相的外貌特征来鉴别它们的身份．难怪要将判断粒子“身份”的方法——模式识别比作“法官”了．

模式识别（Pattern Recognition）属于人工智能这一科学范畴．医学上识别不同的细胞，军事上识别高速飞行中的导弹种类，机器人识别语音，这些都是模式识别用武的天地．判断粒子的“身份”，只是模式识别小试牛刀而已．

人在识别客体的时候，总是根据客体的某些特征来作判断．如果我们的法官（或公安人员）能将照片中的外国人识别出来，他无非是根据一些身体特征：皮肤白、鼻梁高、头发卷而黄的是外国人；皮肤黄、鼻梁低、头发直而黑的是中国人．用数字的语言来说，皮肤的颜色、鼻梁的高低或头发的卷曲程度是不同的坐标，由这些坐标构成一个特征空间．每人的几项特征相应于特征空间的一个点．可以想象，所有的中国人在特征空间内的相应点会形成比较靠近的一团，而外国人形成另一团．模式识别的任务是用定量的方法将特征空间划分成两部分，根据一个人的特征点落在哪一部分，就可认定他的身份．

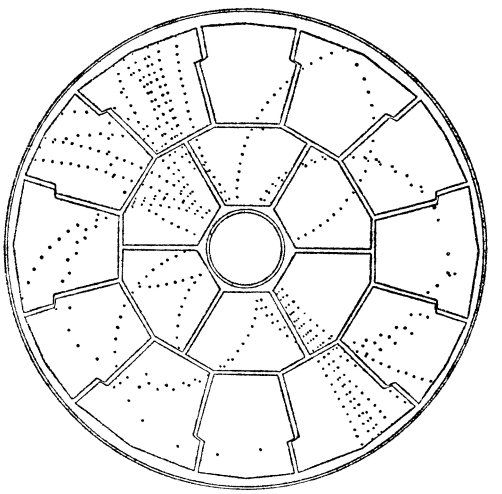
其实，物理学家面临的问题还要困难一些．粒子是看不见摸不着的，我们之所以能探测到它们，是靠它们与探测器内物质相互作用留下的“痕迹”，也就是这些相互作用的产物．一个大型探测器由许多部分组成，粒子击中哪一部分，哪一部分就将这种“痕迹”转化为脉冲信号．因此，从探测器收集到的不是粒子，而是“信号”．一次初级的高能粒子碰撞可能产生包含几十个粒子的终态，其中的每个粒子可能产生几十个甚至几百个“信号”，那么探测器一次收集到的将是上千个“信号”．数据分析要做的第一件事，不是辨认粒子的“身份”，而是辨认粒子——辨认哪些“信号”构成一个粒子的径迹．这就好象我们的法官拿到的不是一叠完整的照相，而是一堆撕得粉碎的照相碎片．聪明的法官先得决定哪个鼻子配到哪个眼睛下面，将一张张照相复原．

照相复原术

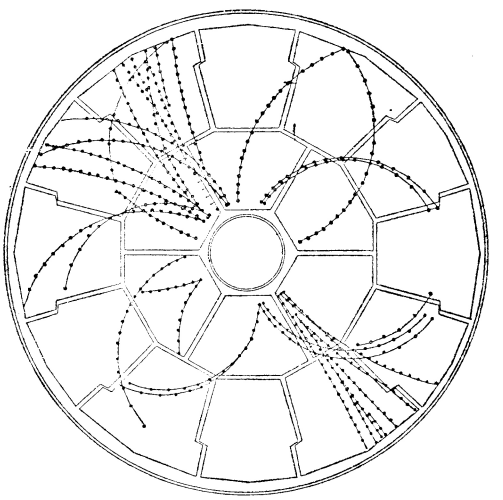
高能实验数据分析中的照相复原术称为“径迹重建”，用的还是模式识别的方法．不过，现在特征空间的坐标便是每个探测器的输出数值，一组“信号”在特征空间内相应于一个点．在照相复原时，如果你乱点鸳鸯谱，随便将眼睛、鼻子和耳朵拚到一起，结果一定十分滑稽，一眼就可看出不象一个真人．与此相类似，在经迹重建时，只有某些特殊的“信号”组合才有可能是真正的粒子径迹，它们在特征空间内也集中为一团．

如何区分出特征空间内这最有意义的一团有各种不同的处理方法．比较直观的，如“通路试探法”．在探测器最内层和最外层各选一个“信号”，把它们连成一条“道路”，如果这条“道路”上确有足够数量的“信号”，就接受其为候选径迹．比较抽象的，如“特征分析法”．因为真正的径迹在特征空间内集合为一团，有可能找出一个超曲面来包容它们．对任意选取的一组“信号”，只需做一些数字变换，便可检查它是否落在这个超曲面内．以此来判断这个组合是不是象男人胡子和女人头发相配合那般地阴错阳差．总之，“八仙过海，各显神通”，不同的模式识别方法总能程度不同地完成径迹重建的任务．为了给读者一个径迹重建的印象，让我们

看图 1. 这是一种大型探测器的投影示意图. 其中(a)图画出一个终态事例产生的"信号". 尽管人的肉眼也可以在这令人眼花缭乱的图象中分出一些明显的径迹来，但至少有两个困难是用肉眼不能克服的. 第一，一个实验每秒钟收集几个或几十个事例，人的肉眼绝对没有这么高的处理速度. 第二，图中右上和左下



(a) 重建前



(b) 重建后

图 1 某大型探测器上一个事例的投影图

这两部分粒子特别密集，靠肉眼判断哪个"信号"属于哪个粒子，一定会有许多似是而非的错误. 图 1 的(b)是用计算机作模式识别的结果，尽管不能担保百分之百正确，多数的判断是达到或超过了肉眼的能力. 到此为止，法官手中的照相已经复原完毕，可以进行"身份"鉴定了.

国境线上的"身份"鉴别法

首先让我们假定法官手中的那叠照片是从国境线上边防站里拿来的，在那里进出的中国人、外国人都有相当比例。要是法官选取了比较恰当的一、两个特征例如头发和眼睛，就足以将绝大多数的中国人和外国人区别开了. 当然，判错的也会有，把黑眼珠的外国人当做中国人，把卷头发的中国人当成外国人，但总是个别情况. 十之八九总是判得对的.

高能实验中鉴别荷电的 π 介子和质子 p 就类似这种情形. 对每种粒子至少测定了两个量，譬如动量和电离能损. 在某一范围内动量相同的粒子，π 产生的电离能损较小，p 产生的电离能损较大. 图 2 是某探测器测量得到的 4GeV/C 动量的 π 和 p 的平均电离能损分布，横坐标是平均电离能损的大小，纵坐标是具有一定平均电离能损的粒子个数. 根据平均电离能损可以判断大多数粒子是 π 还是 p. 当然在两根曲线重叠的部分，判断就难了. 那是卷头发中国人和黑眼珠外国人的区域.

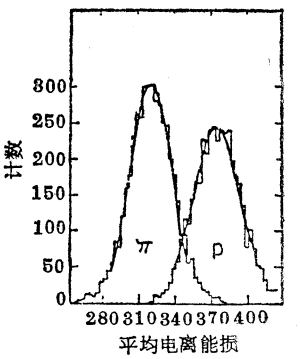


图 2 平均电离能损分布

要增加判断的可信程度，可以多采用一、两个特征，就好比在只用头发和眼睛这两个特征不能判断人的身份时，就把鼻子和皮肤也加上. 图 3 为某探测器鉴别粒子用的散射图. 探测器对一个粒子测量上百个电离能损值，电离能损有涨落，上百个测量值可以形成一条分布曲线. 用这上百个值，除了可以计算平均电离能损，还可以计算电离能损分布的阶矩，二阶矩就是表征分布情形的量. 具有相同动量的不同粒子，不但它们的电离能损平均值不同，而且电离能损的分布也不同，图 3 的横坐标是平均电离能损，纵坐标是电离能损的二阶矩，图中每个点表示一个粒子的测量值. 图上荷电 π 介子、电子 e 和质子 p 相应的点明显地分成三团. 因此不难将图中的坐标平面(特征空间)划分为分属于 π 介子、电子或质子的区域. 这比图 2 只用

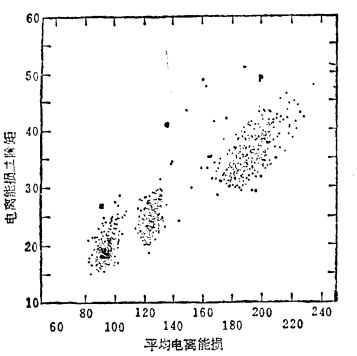


图 3 电离能损测量散射图

两个特征(动量和电离能损)的判断自然更可信了.

大街上的"身份"鉴别法

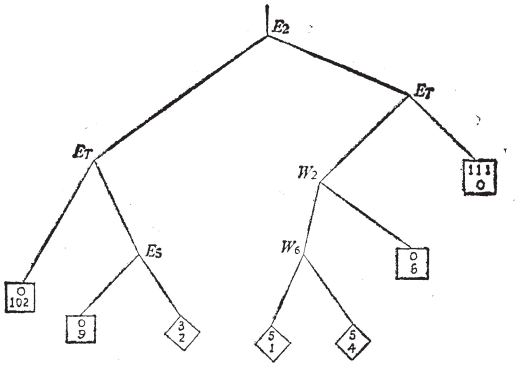
再让我们假定法官手中的照相是在熙熙攘攘的前门大栅栏拍摄的. 那里绝大多数是中国人，而外国人则是凤毛麟角. 这时我们的法官可得特别小心：不管他采用的是什么特征，他选出来的"外国人"，十有八九是选错了的中国人. 这并不奇怪，如果一万张照相中只有一个外国人，而法官用白皮肤来判别外国人，一定会找出几百个"面如冠玉"的中国人来，比真正的外国人多几百倍、高能实验中遇到的情况有时比这还头痛. 有的实验要找的是电子，而终态中的 π 介子数是电子数的一百万倍. 即使每一万个 π 介子中只有一个被错判为电子，最终收集到的"电子"中还有百分之九十九是判错了的 π 介子.

当然还是可以用老办法来处理——增加用作鉴别的特征. 不过，用十几个特征作鉴别与用二、三个特征作鉴别可大不相同，再也没有机会用图 2 或图 3 那样直观的分辨方法了. 高维空间的几率密度函数很难估计，要用几率密度为依据来划分特征空间就更难了.

让我们回头来看看 π 介子和电子的鉴别问题. 如果是用电磁簇射计数器的数据来作鉴别，电子在计数器内产生大信号，π 介子产生小信号，本来可以用信号大小来区分它们. 可是 π 介子和电子产生的信号大小又都有涨落，个别电子会产生小信号，个别 π 介子会产生大信号. 难就难在 π 介子的数目比电子的多得多. 百分之几的大信号 π 介子，就象"面如冠玉"的中国人一样，把问题搅得一团糟. 于是不得不添上簇射的纵向特征和横向特征. 一般来说，电子产生的簇射信号，在其开头部分就会有较大的能量沉积，而 π 介子即使产生大信号，也不大会在开头部分就有大的能量沉积. 电子的簇射信号又有明显的横向展开. 把这些特征综合起来，当然足以区分 π 介子和电子了. 有一个实验，将簇射计数器分为七层，收集七层的沉积能量（脉冲大小），七层的横向展开，再加上总能量，一共用十五个量来作判断. 应该说特征是足够多了，但困难在于这十五个量怎么用法，是一起用，还是有先有后？每一个量的判选界限在哪里？于是要求特殊的处理方法。

有一种叫做两叉决策树 (Binary Decision Tree) 的方法，可以简便地解决用多个特征量作鉴别的问题.

先选用一定数量的训练样品，在目前讨论的实验中，是用 124 个电子和 124 个 π 介子，也就是说，它们是"身份"已明的粒子. 用它们的特征量，构成一个决策树. (图 4) 树的每一个分叉点，都用某一个特征量作判断来决定一个样品应当往左或往右. 例如图 4 中第一个分叉点，用第二层的沉积能量 (E_2) 作判断，事先确定一个切割值，样品的 E_2 大于切割值便向右走，小于切割值向左走. 每个样品经过若干个分叉点，在每



E_T 总能量
E_i 第 i 层的脉冲大小
W_i 第 i 层的横向展开

图 4 一个分辨 π 介子和电子的两叉决策树

一个分叉点用一个特征量来决定走向，最后到达一个终端站. (图 4 的方框和菱形)在终端站统计到达样品的"身份"，图 4 中每个终端站内的两个数字，上方的为电子样品个数，下方为 π 介子个数. 如果某种粒子的个数占压倒优势，则称这个终端站为决策站，(方框)如果两种粒子的个数相差不远，则为悬案站. (菱形)

有了决策树，可以判断"身份"不明的对象. 一个实际测到的粒子，凭它的各个特征量，在决策树的每个分叉点决定去向，最后进入一个终端站. 如果进入一个决策站，则可认定为 π 介子或电子. 如果进入一个悬案站，表明对它的"身份"作判断要冒犯错误的风险，那么比较稳妥的方法是不作判决，任其身份不明. 因此两叉决策树方法，虽然能简便地判断多数的电子和 π 介子，但是有一定的损失率. 在图 4 显示的例子中，"真"电子得以判对的效率为 $111/124=89.5\%$. 可以说，有点"不放过一个坏人，但要冤枉一些好人"的味道.

总之，不管哪一种模式识别的方法，都不是断案如神的"包青天". 法网恢恢，疏而有漏，有时"放过坏人"，有时"冤枉好人"，但总能保证对粒子的鉴别足够可靠，使物理学家从而得出科学的结论.