流行病学研究中AID聚类分析的应用

莫世华[1] 谢淑云[1] 蒋廷魁[1] 胡松禅[2] 陈敏玲[2]

AID（Automatic Interaction Detection）聚类，是一种分裂法聚类，它不受线性条件限制，能处理各种离散型变量和连续型变量资料。它在聚类中每次样品的分裂，都是因变量按因子大小排序基础上有序的最优二分割，而因子的入选，又是以实现因变量最优二分割前提下所作的选择，因此构造了因子与因变量之间，在样品每次分裂中的“自动交互检测（AID）”关系。所以，它不但具有合理分类的功能，而且还有一定的因子筛选能力。同时，利用这两方面功能，又可建立分类预测模型。由于AID聚类有较广的适应范围和应用领域，笔者曾在以往的流行病学实践中多次试用，均有较满意的结果，现将其中有代表性的实例，择要报告如下。

在疾病流行区域分类中的应用：例1，根据某地NAG腹泻11年监测资料，其发病似有地区性，为研究地区分布特征，笔者运用AID聚类划分流行区域，并筛选有关的地理因子。

1.方法：分别计算该地全部194个乡镇历年NAG腹泻检出率秩次之和，作为聚类因变量Yj（j=1，2，…，194）；取每个乡镇中心点的经度、纬度、海拔高度和与海岸线最近距离，分别作为聚类待选因子Xi（i=1，2，3，4；j=1，2，…，194）。以上四个因子和一个因变量，构成5×194原始数据阵进行AID聚类，并对聚类最后形成的各个类的$\bar{Y}$分别作两两比较，如有$F<F\alpha$，则对作比较的两个类进行合并，由此最终确定各类流行区的$\bar{Y}$及所包含的乡镇（样品号）。

2.结果：该地194个乡镇分为五类。第一类分布在与H海湾相邻的C江入海口地区；第二类为H海湾沿岸其他乡镇；第三类均为第一、二类毗邻乡镇；第四类为其余平原地区；第五类为山区。NAG腹泻检出率11年秩和，以第一类为最高，以后依次下降。聚类入选因子，主要为离海岸线距离，其次为海拔高度和经度。这三个因子再用AIRD聚类进一步筛选，又剔除了“经度”因子。以上结果基本反映了NAG腹泻的地理流行病学分布特征，与该地后来NAG生存环境研究的结果十分吻合。

在流行预测中的应用：例2，根据A群流脑流行周期性、季节性特征，笔者利用某地29年疫情资料作AID分类预测和因子筛选。

1.方法：取预测年上三年每年的流脑发病率、预测年上年发病时间的$\bar{a}$、r、Mo、Md值及（5～6月/12～1月）与（4～6月/11～1月）两个时间的发病数比值，计算预测年周期趋势值等10个因子为预测待选因子；取对应的流脑年发病率（已扣除菌苗保护率）为预测建模的因变量，由此作AID聚类并建立预测树图。此后只要代入预测年的各入选因子值，依图检索该年归属的类，然后按该类各样品年的年发病率均值及上下限，对预测年的年发病率作点值与区间值预报。

2.结果，浙江省80年代均用此法预报，获满意结果，特别是利用聚类入选因子构造全省分县市的真值图预报，也有较好的结果，其中1980、1985、1986三年公开预报，县市符合率分别在72.78%、85.39%及80.22%。

（收稿：1993-05-14）

1 浙江省卫生防疫站 310009 杭州市
2 浙江省计算技术研究所