BDT方法分析

崔震崴[1]

1. 北京大学，北京市海淀区颐和园路5 号，100871；

多元变量分析（MVA），作为一种挖掘数据特征和关系的工具，广泛用在各行各业中。MVA包含Fisher线性判别器，超长方体分割，k-NN聚类，MLP神经元网络分析等多种分析方法，广泛用在高能物理的分析当中。

分类是一种重要的数据分析形式，是提取刻画重要数据类的模型，一般分为两个阶段：学习和分类。学习阶段就是通过分析，或从训练集“学习”，建立描述预先定义的数据类或概念集的分类器[1]。训练集，在物理分析中一般指蒙卡信息或部分了解全面的实验数据。训练集中样本点X 一般有n个属性，$X=(x_1+x_1+x_1\cdots x_n)$,在进行训练之前，会对全部的样本做标准化等预处理[2]，其中标准化处理后的新变量为：

$$e_j^*=(\frac{x_{1j}-\bar{x}_j}{s_j},\frac{x_{2j}-\bar{x}_j}{s_j},\cdots,\frac{x_{Nj}-\bar{x}_j}{s_j})$$

其中，x_{ij}代表第i个样本的第j个分量，$\bar{x}_j$为第j个分量的平均值，s_j为第j 个分量的标准差。在高能物理中常用的ROOT软件中，样本预处理中除了标准化还有，主成分分解去相关等方法。

经过预处理后的数据相关性减弱，具有一般性，可以进行多变量分析的主要过程——“学习和分类”，这里需要注意在不同的文献中“学习和分类”，又叫做“训练和应用”等。

BDT是现有的一种分类速度快，精度高的MVA方法。利用BDT分析实验数据中常见的情形为信号和本底的区分。• 求解这类问题的过程，首先是利用一个训练样本集来构建（训练）一个决策树的过程。决策树，是一个二叉树，每个节点上包含信号和本底的数目信息。训练从根节点开始，在每一个节点，通过某种优化步骤，寻找样本属性空间中某一个变量及其阈值，使得在这一节点的判选中能最有效地区分信号和本底。通过该节点的判选，输入事例被区分为“类信号事例”和“类本底事例”两部分，其中“类信号事例”中信号事例的比率高于判选前的信号事例的比率：而“类本底事例”部分则相反．这两部分事例作为下一层节点的输入进行进一步的判选．这一过程一直延续下去，直到满足某种终结条件时停止。

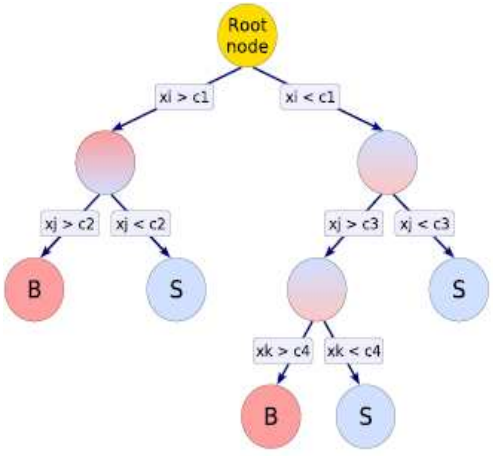

图 1 BDT分析示意图。

通过上图可以看出决策树建立，对节点阈值有严格的要求。一般情况下要保证判别指数I增量最大化[2]：

作者简介：崔震崴，粒子物理与核物理专业，导师：李强，E-mail: lovekeyczw@yahoo.co.jp

$\Delta I = I - (\frac{n_1}{n_{int}} I_1 + \frac{n_2}{n_{int}} I_2)$

其中$n_{int} = n_1 + n_2$,I, I_1, I_2为母节点和两个子节点的判别指数，n_{int}, n_1, n_2分别为母节点和两个子节点的输入事例数。判别指数I 是用来估价信号/本底判别能力的标准。一般有如下四种：

Gini指数： p(1-p)

交叉熵： -plnp-(1-p)ln(1-p)

误判误差： 1-max(p,1-p)

统计显著性： $n_S/\sqrt{n_S + n_B}$

上式中p为信号纯度，$p = n_S/(n_S + n_B)$,这四种定义在不同的问题中均有应用[1]。

决策树结束的条件主要有四种：第一，设定一个最大的叶节点数，当训练过程已经形成的叶节点数等于大于该数值则训练停止．第二，设定一个最小的事例数，当输入事例数小于，该节点的训练停止。以上两种做法看起来缺乏理论依据，并且对于不同的问题需要根据经验确定适当的具体数值．第三种做法是当一个节点的输入事例为同一类事例时，该节点的训练终止．第四种做法是根据所有节点的增量值来决定训练是否终止。

利用节点分割条件和终止条件确定的决策树，可能会出现很多节点的情况。节点过多会影响决策树的准确度：第一，决策树的错误率随节点数的增加而减小，但存在一个最佳节点数的决策树，它的错误率达到极小；当决策树的节点数大于该值时，错误率反而增加，所以决策树的节点数并非越多越好．第二，过长的决策树训练得到的名义误判率往往低于误判率的真实值，这种导致低估误判率的分叉过长的决策树训练称为过度训练。所以需要对决策树进行“剪枝”，常用的剪枝方案为L Breiman提出的最小复合费用修剪方案[3]。

考虑到MVA处理的问题一般都有明显的随机性，单一的决策树存在偶然性。组合分类方法是提高分类准确率降低偶然性的主要方法，具体包括装袋(Bag)，提升(Boosting)和随机森林[2]等,其中随机树林法（BDT）是一种综合装袋与提升的方法，被广泛用在物理分析中。

参考文献：

[1] 《数据挖掘——概念与技术》(《Data Mining–Concepts and Techniques》) Jiawei Han, Micheline Kamber，Jian Pei 著，范明，孟小峰 译. 机械工业出版社.

[2] 《实验数据多元统计分析》朱永生 著. 科学出版社.

[3] Classification and regression trees. California: Waldsworth International Group, Belmont, 1984.