1997年 火山地质与矿产 VOLCANOLOGY & MINERAL RESOURCES 第十八卷 第一号 Vol. 18 No. 1

·工作方法·

67-75

CART—分类与回归树方法介绍

张松林①

(南京地质矿产研究所,南京 210016)

0212.1 P628

摘要 分类与回归树(CART—Classification and Regression Trees)是一种非常有趣并且十分有效的非参数分类和回归方法。它通过构建二叉树达到预测目的。该方法是四位美国统计学家耗时十多年辛勤劳动的成果。在他们所著的"Classification and Regression Trees(1984)"一书中有该方法的详细说明。笔者认为该方法在古生物化石分类和鉴定以及矿产资源预测等方面颇有用武之地。本文只对该方法作简单介绍,希望能对从事预测工作的读者有所启发。美国加利福尼亚统计软件公司已开发了CART方法软件。

关键词 分类 回归 预测 二叉树

矿产资源

1 分类和回归定义

分类和回归首先利用已知的多变量数据构建预测准则,进而根据其它变量值对一个变量进行预测。在分类中,人们往往先对某一客体进行各种测量,然后利用一定的分类准则确定该客体归属那一类。例如,给定某一化石的鉴定特征,预测该化石属那一科、那一属,甚至那一种。另外一个例子是,已知某一地区的地质和物化探信息,预测该区是否有矿。回归则与分类不同,它被用来预测客体的某一数值,而不是客体的归类。例如,给定某一地区的矿产资源特征,预测该区的资源量。

分类和回归的预测准则是在对已知数据集进行系统分析的基础上构建的。该数据集不但包括被预测变量的数据,还包含了预测准则中用到的其它相关变量的数据。例如,在化石鉴定中,数据集由许多已知种类的化石鉴定特征组成。在资源量预测中,数据集由许多资源量已知地区的资源特征和资源量组成。在获得了一个数据集后,分类和回归的关键问题在于如何利用这些数据构建预测准则。

在构建预测准则时,需要考虑以下两个目标:

① 收稿日期:1996—12—16

作者介绍:张松林,男,1964年生。硕士,副研究员。数学地质专业。主要著作有《床板珊瑚形珊瑚属种鉴定的微机处理系统》,地质出版社,1991。

(1)构建尽可能准确的预测准则;

(2)构建的预测准则应该给出对问题最高度的认识。

第二个目标的意思是,通常在预测工作中,许多变量被测量,那么究竟是哪些预测变量给出了重要的预测信息、它们是如何给出这些信息的。

上述两个目标并不矛盾。因为一个简单的、不准确的预测准则不太可能给出对问题的高度认识。但是,当问题的目标是理解所研究的对象,并且理解各个变量对预测的影响,那么如果一个易于理解和解释的预测准则与一个数学上复杂的预测准则具有相当的准确性,前者显然更加可取。

分类中通常使用的统计学方法为判别分析或逻辑回归。回归中的常用方法是某种形式的线性回归。在这两种情况下,预测准则均以代数表达式的形式给出。但这些表达式通常难以理解和解释,对缺乏统计学背景知识的人更是如此。

CART 采用了与传统统计学完全不同的方式构建预测准则。它所构建的预测准则以二叉决策树的形式给出,非常容易理解、使用、说明和解释。由 CART 方法构建的预测树很简单,但它在很多情况下比常用的统计方法构建的代数学预测准则更加准确。事实上,数据越复杂,变量越多,CART 比其它方法的优越性就越显著。

2 CART 预测树

先考虑一个假设的例子。假定我们对许多客体中的每一个测量了 25 个特征 X(1),……,X(25),我们想把这些客体的每一个划归为三个类中的某一类。这一问题的一颗可能的 CART 树如图 1 所示。

图 1 中的树可被用来进行如下的分类:假设我们有某个客体的测量数据向量,其中 X(1)=5.2,X(3)=1.4,X(8)=7.3,X(11)=0.52,在第一次分化时,它会向右,第二次分化时向左,第三次时向右,进入类 2 终结点而告终。图 1 中树的最终的类框称为终结点。

我们可以看出,这颗树的使用和理解均非常简单。尽管我们总共有 25 个变量,但仅有几个变量在各次分化中使用。并且,我们很容易看出这几个变量是如何影响分类的。例如,树最左侧的终结点类 1 由这样一些客体说明,它们的三个变量 X(8)、X(3)和 X(22)的值均比较低。

下面我们考虑回归的情况。再一次假设我们对许多客体测量了 25 个特征,我们需要预测与每一客体有关的某一数值 y。这一问题的一颗可能的 CART 树如图 2 所示。

图 2 中的 CART 回归树与图 1 中的 CART 分类树非常相似,除了终结点中的内容是具体的数值而不是某一类别。当用这一回归树进行预测时,我们可以根据某一客体的诸多变量值决定它沿树的走向,最终到达终结点,该客体便获得该终结点对应的 y 值。

二叉分类和回归树既简单又直观,也许有人会认为它们过于简单而不可能有其它复杂方法如判别分析、逻辑回归或线性回归等准确。但实际上它们的误差率非常低、很多情况下要比复杂方法的误差率低。特别当处理由许多独立变量组成的复杂数据时,由 CART 方法产生的二叉树的误差率要比通常的参数方法的误差率低得多。

CART 方法不仅是非参数的,它还考虑了这样一个事实,即在数据的不同部分,变量间

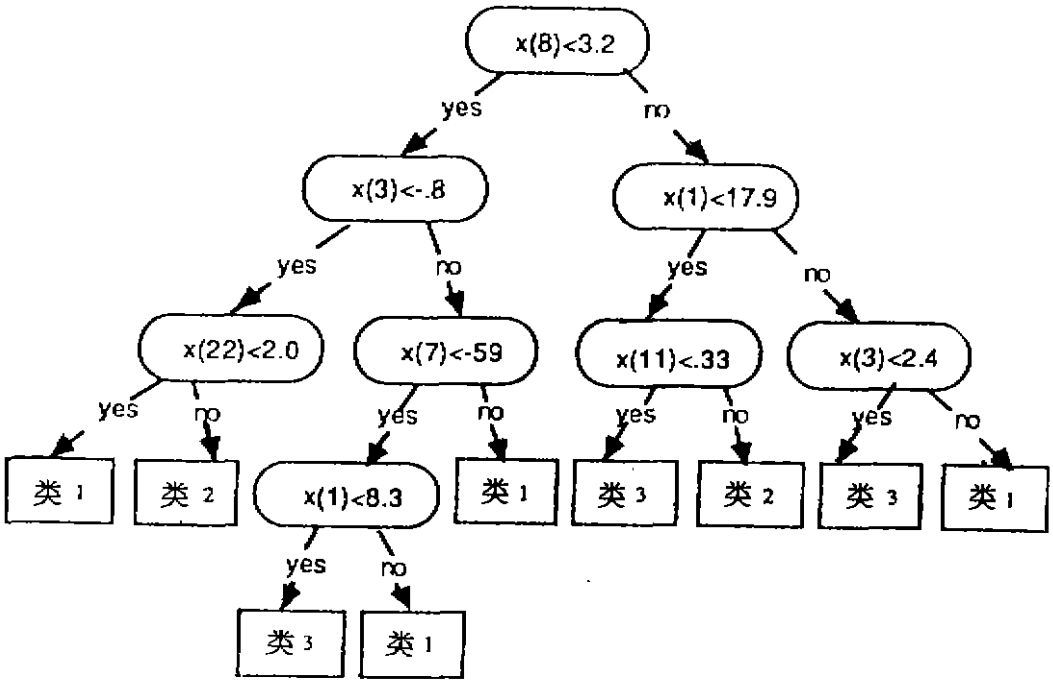

图 1 CART 分类树例子

Fig. 1 An example of CART classification trees

的关系有可能不同。例如，一旦数据被分成两部分，则由于这两部分数据变量之间的关系不同，左边部分数据进一步分化通常与右边部分数据进一步分化不同。另外，CART 方法还非常稳健，它受少数异常数据的影响非常小，而通常的参数方法受少数异常数据的影响则非常大。

3 CART 树的生成

下面介绍 CART 树的生成方法。CART 树是根据对“学习集”数据的运算生成的。以前面三个类的划分为例，假定我们已经检查了 300 个客体，对每一客体，我们不仅知道它 25 个特征的测量值，而且还知道它归那一类。为简单起见，我们假定每一类各有 100 个客体。我们称上述数据的任一子集为一个结点，这样一开始我们便有一个由 300 个客体全部数据组成的结点，其中各有 100 个客体属于某一类。

CART 树的形成过程如下：首先，检查所有形如 X(1)<C 的分化，这里 C 是一个常数，它的变化范围自 X(1)的最小值到 X(1)的最大值。具体的分化是这样的，假设 C=1.1，那么所有 X(1)<1.1 的客体被分化到左边，其余的则到右边。计算左右两边分属类 1、类 2 和类 3 的客体数目，假设左边是 57、41、65，右边是 43、59、35。这一分化如图 3 所示。

注意这一分化并没有把三个类很好地分开，但也许其它的一些 C 值会给出更好的分化。CART 方法通过变化 C 值检查所有可能的分化，然后选择其中最好的一个，假设为 X(1)<10.7。最好的分化给出了类别划分的最好结果。CART 利用数值准则评价每一分化在

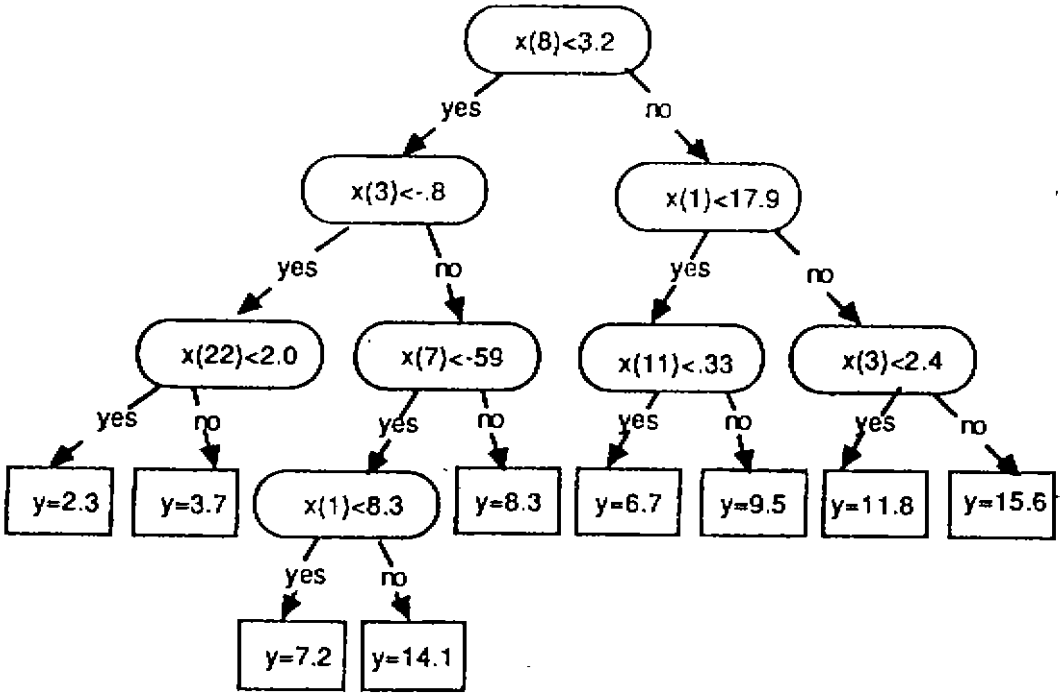

图 2 CART 回归树例子

Fig. 2 An example of CART regression trees

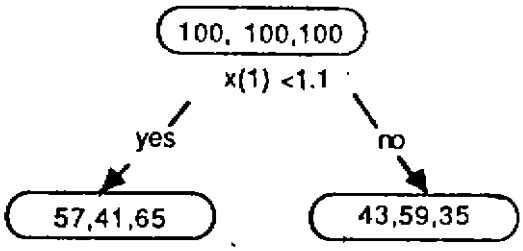

图 3 基于变量 X(1)的一个分化

Fig. 3 A split based on variable X(1)

区分类别方面的优劣，具体的表达式和说明可以参考[1]的第 4 章。在获得关于 X(1)的最优分化后，CART 继续检查所有形如 X(2)<C 的分化，并且找出其中最好的一个，假设为 X(2)<−6.2。接着 CART 以同样的方式，确定 X(3)，……，X(25)对应的最优分化。这样我们便获得了 25 种不同的分化，其中每一种为某一个变量对应的最优分化。根据同样的评价准则，我们从这些分化中挑选出一个最好的分化。假设这一最好的初始分化是 X(8)<3.2。以三个类别的数目形式表达，这一分化如图 4 所示。

然后，相似的方法步骤被应用于左边结点的数据。即检查所有的变量，确定一个最优分化，给出左边结点中客体类别的最优划分，假设这一分化是 X(3)<−0.8。接着对右边结点

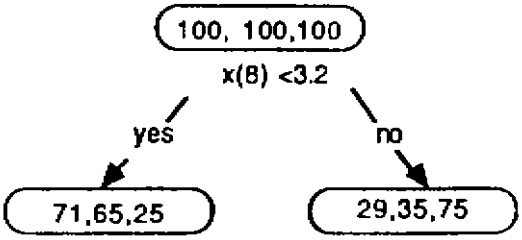

图 4 最优初始分化

Fig. 4 The best initial split

的数据做同样的工作，假设最优分化是 X(11)<17.9。目前的结果如图 5 所示。

如果我们不想继续进行客体的分化，那么便可得到一颗具有 4 个终结点的分类树。树中最左边的终结点被定为类 2 结点，因为其中类 2 客体有 53 个，占了大多数，而类 1 客体只有 17 个，类 3 客体仅有 8 个。根据同样的理由，左边第二个终结点被定为类 1 结点，第三个为

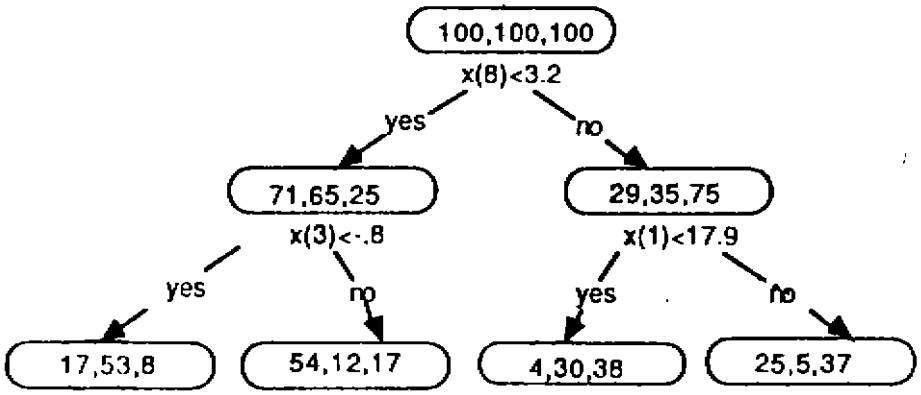

图 5 三次最优分化后的结果

Fig. 5 Results after three best splits

类 3 结点，最右边的终结点也是类 3 结点。

但上面的树不是一颗非常准确的分类树。例如，如果用组成学习集的 300 个已知客体来检验这颗树的分类准确性，那么，落到最左边终结点的 78 个客体中，25 个类 1 或类 3 客体将被预测为类 2 客体，导致 32%的误差率。最右边终结点对应的误差率更高，达 45%。我们称用学习集客体进行检验获得的误差率为视误差率。

因此，我们应该继续进行客体的分化，以获得准确的分类树。但问题在于什么时候应该终止分化。如果我们一直进行分化直到不能再分为止，我们最终将得到具有 300 个终结点的一颗树，每一个终结点只包含一个客体。这一颗分类树的视误差率为零。但是，如果拿 300 个已知客体以外的若干客体来检验，真正的误差率不太可能为零。这颗大树的真误率可能比上面只有四个终结点的小树的误差率要高。

因此，构建分类树的关键在于选择大小适当的树，即选择一定数目的分化使真误差率达

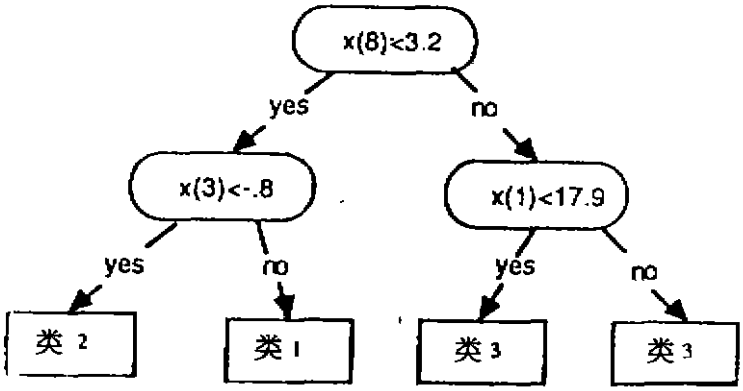

图 6 三次最优分化形成的分类树

Fig. 6 The classification tree formed by three best splits

到最小。这一问题的解决方法放在下一节介绍。

还有两个问题需要说明。第一个是 CART 怎样构建一颗回归树。再次假设我们有 300 个客体，每一客体有 25 个特征的测量值，并且我们还知道每一客体的 y 变量值。我们想要构建一颗预测树来预测客体的 y 变量值。与前面分类时的情形相同，这颗树的顶结点包含 300 个数据向量。我们首先计算所有 y 值的平均值和方差，假设均值为 9.3，方差为 51.5。

构建一颗回归树的基本方法是寻找这样的一些分化，它们逐次将数据分成两个结点，每一结点中的 y 值尽可能地一致。更具体地说，假设有一个分化，形式为 X(1)<C，它将 142 个客体分进左边结点，158 个客体分进右边结点。计算左边结点中 142 个 y 值的方差，比如说 46.7，并计算右边结点中 158 个 y 值的方差，比如说 49.3。这一结果表明左右两个结点中 y 值的离散程度与原来 300 个 y 值的离散程度相若，因此这一分化从降低离散程度方面考虑效果并不好。

假设另外有一个分化，比如说 X(7)<78.2，这一分化的左结点中 y 值的方差为 20.3，右结点中 y 值的方差为 26.0。这一分化减少了大约 50%的方差，它的左右结点中的 y 值比顶结点中的 y 值更紧靠它们的平均值。这是一个较好的分化，因为它产生的两个子结点中，y 值的离散程度比父结点中 y 值的离散程度要低得多。

因此，分化的优劣根据它产生的子结点中 y 值的离散程度来评价。更确切地说，对任何一个分化，计算其左右结点中 y 值的方差的加权平均，权值与左右结点中客体的数目成正比。最小的加权平均值决定最好的分化。CART 对所有变量搜索所有可能的分化，在每一个结点找出一个最好的分化。例如，经过最初的 3 次最优分化后，我们可能得到的一颗回归树如图 7 所示。

如果这颗树即是我们想要的回归树，那么，当一个 y 值未知的客体沿此树向下走时，如果它进入最左边的终结点，它的 y 值便被预测为该终结点中学习集客体的 y 值的平均值，即 3.1。如果该客体进入最右边的终结点，它的 y 预测值便为 13.3。

但是，从图 7 中可以看出，一些终结点的方差值还相当高，这说明我们得到的预测值的误差可能会比较大。因此我们可以继续进行分化，以减小终结点的方差。

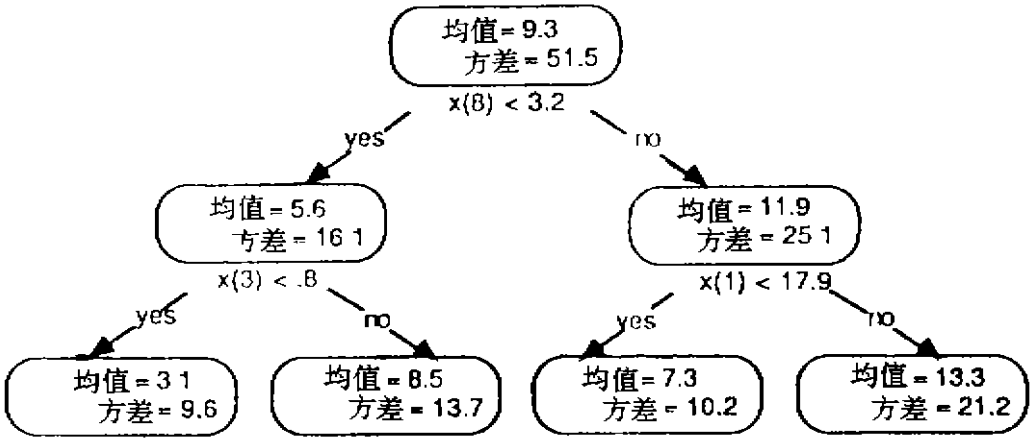

图 7 三次最优分化形成的回归树

Fig. 7 The regression tree formed by three best splits

随着分化次数的增多，各个终结点的方差将越来越小，最终当每个终结点中只有一个客体时，所有的方差均为零，相应的最大回归树的视误差率也为零。这时，我们便面临与前面分类时遇到的同样问题，即找出真误差率最小的回归树。

有一点需要说明，y 值的方差并非 y 值离散性的唯一度量。事实上，方差对异常数据非常敏感，即几个远离平均值的 y 值将对 y 值的方差产生非常大的影响。

一个不太敏感的度量是各个 y 值与 y 值的中位数之差的绝对值的平均值。当这一度量被用来说明 y 值的离散性时，各个终结点对应的预测值将是该结点中 y 值的中位数。上面的两个度量均可用于生成 CART 回归树。

第二个问题是，在树的生长过程中，CART 如何处理非数值变量。上面我们使用的测量特征均被假设为数值变量。但我们经常还会遇到这样一些特征，它们没有任何自然的数值顺序。例如，在资源量预测问题中，某一特征可能会有这样一些值：砂岩，泥岩，页岩，灰岩。我们称在一无序集中取值的变量为类型变量。

如果 X(1)是一个数值变量，那么 CART 将搜索所有形式为 X(1)<C 的分化。但当 X(1)是类型变量，取值范围为(R,S,T,U,V)时，CART 将搜索所有这样形式的分化：X(1)是否属于(R,S,T,U,V)的某一子集？例如，CART 将评价 X(1)=R?,X(1)=S?,……,X(1)=R 或 S?,X(1)=R 或 T?,……，等等。在这样一个例子中，CART 总共将检查 15 种可能的分化。

4 选取大小合适的树

上一节中已经指出，如果我们构建的树太小，那么它的误差率会比较高。如果树太大，尽管用学习集检验获得的视误差率很小，但它的真误差率可能还是比较大。因此，我们需要构建一颗大小适当的树，它的真误差率为最小。

CART 方法解决这一问题的具体步骤如下：首先，不断对学习集的客体进行分化，直到形成一颗非常大的树。这颗树的每一终结点中一般只有几个客体。终结点中客体的具体数目可根据客体的总数确定。通常，如果客体总数小于 1000，那么每个终结点的客体不应多于

5个。

然后，运用一定的算法对这颗大树的树枝不断进行修剪。在整个修剪过程中将给出一列越来越小的树，形成一个修剪树序列。算法的原理是，这一序列中的每一颗树与其它大小相同(指终结点数目相同)的子树比较，具有更小的视误差率。因此，从某种意义上讲，这一修剪树序列是一个最佳序列。

下一个问题是，如何从这一列树中挑选出一颗所需要的树。如果我们知道每一修剪树的真误差率，那么情况就非常简单，我们只需要挑选那颗真误差率最小的树。依靠视误差率是没有用的，因为树越大，视误差率就越小。我们所需要的是，真误差率的更为可靠的估计。CART 提供了两种方法解决这一问题。

一种较为简单快捷的方法是利用检验集。如果我们有许多已知客体，那么，我们可以随机地选取其中的一部分，比如说 1/3，留作检验集。其余的客体则被用来构建大树和修剪树序列。

下一步便是用检验集去检验修剪树序列中的每一颗子树。这会产生每一颗子树的真误差率的一个估计。例如，在分类中，估计值是子树对检验集中客体作出错误分类的比例。在回归中，估计值是检验集中的 y 值与子树对它们的预测值的平均平方差(如果中位数被用来作为预测值，那么估计值是检验集中的 y 值与子树对它们的预测值的平均绝对偏差)。

通常，一颗非常大的子树的真误差率的估计值会非常大，而当子树变小时，这一值也会变小。但当子树变得很小时，这一值又会增大。CART 方法将选取那颗估计值最小的树。

上述方法只在具有很多已知客体时才适用。当已知客体的数目不大时，我们往往希望用全部的数据来构建树。在这一情况下，CART 采用交叉证实(Cross—Validation)的方法来估计真误差率。这一方法将每一客体既用于构建树，又用于估计误差率。但它牵涉到构建主树以外的一系列辅助树，因此需要很多的计算机运行时间。在[1]的第三章有这一方法的详细说明。

上述两种方法经实践检验均非常可靠。从误差率的角度考虑，用它们选取的树总是非常接近最佳的树。

从上面的介绍可以看出，CART 方法的基本思想非常简单。从顶结点开始，CART 在每一结点选取最优变量的最优分化，形成一颗大树。然后修剪这一颗大树获得一系列子树，并从中挑选出真误差率的估计值为最小的子树。最后根据这颗树对客体进行预测。通常，即使学习集很大，即学习集包含许多客体，并且每一客体有许多变量的测量值，CART 树的大小仍会比较适中。因为在很多情况下，一小部分变量即包含了绝大部份的预测信息，CART 树利用这些变量即可设法捕获绝大部分的预测信息。

CART 方法还具有许多其它的特性，例如，对数据缺失的处理，基于变量线性组合的分化，变量重要性的排序以及对误分类费用的考虑等等，这些特性使得 CART 方法更加灵活有效。

5 参考文献

1 Breiman L, Friedman J H, Olshen R A, et al. Classification and Regression Trees. Wadsworth, Inc. 1984

2 California Statistical Software, Inc. An Introduction to CART Methodology. 1985

AN INTRODUCTION TO THE METHODOLOGY OF CART—CLASSIFICATION AND REGRESSION TREES

Zhang Songlin
(*IGMR, Nanjing*, 210016)

Abstract

CART—Classification and Regression Trees—is an interesting and powerful non-parametric methodology in classification and regression. It arrives at predictions by constructing binary trees. The methodology is the result of over 10 years of testing and development by four American statisticians, and is explained fully in their book "Classification and Regression Trees (1984)". It is thought to be very useful by the writer of this paper in dealing with jobs such as the classification and identification of fossils and the prediction of mineral resources. This present introduction is intended as a short and simplified description of what the CART does and how it does it, and to enlighten readers who are engaging in prediction pursuits. The computer program for the methodology has been developed by the California Statistical Software Incorporation and is available there.

Key words classification regression prediction binary trees