装备科技译著出版基金

高新科技译丛

Pattern Classification Using Ensemble Methods

模式分类的集成方法

【以色列】Lior Rokach 著　　黄文龙 王晓丹 王毅 肖宇 译

World Scientific
Connecting Great Minds

国防工业出版社
National Defense Industry Press

高新科技译丛

装备科技译著出版基金

模式分类的集成方法

Pattern Classification Using Ensemble Methods

[以色列] Lior Rokach 著

黄文龙 王晓丹 王 毅 肖 宇 译

国防工业出版社

·北京·

著作权合同登记　图字：军-2014-040 号

图书在版编目（CIP）数据

模式分类的集成方法/（以）罗卡赫（Rokach, L.）著；黄文龙等译. —北京：国防工业出版社，2015.11

（高新科技译丛）

书名原文：Pattern Classification Using Ensemble Methods

ISBN 978-7-118-10397-7

Ⅰ. ①模… Ⅱ. ①罗… ②黄… Ⅲ. ①模式分类－方法研究

Ⅳ. ①O235

中国版本图书馆 CIP 数据核字（2015）第 260837 号

※

国防工业出版社出版发行

（北京市海淀区紫竹院南路 23 号　邮政编码 100048）

北京京华虎彩印刷有限公司印刷

新华书店经售

*

开本 710×1000　1/16　**印张** 11½　**字数** 213 千字

2015 年 11 月第 1 版第 1 次印刷　**印数** 1—2000 册　**定价** 69.90 元

（本书如有印装错误，我社负责调换）

国防书店：（010）88540777　　发行邮购：（010）88540776

发行传真：（010）88540755　　发行业务：（010）88540717

译者前言

机器学习这门学科所关注的问题是计算机程序如何随着经验积累自动提高性能。集成分类思想的提出是伴随着模式识别处理问题日益复杂而产生的。在模式识别系统中，最终的目标是得到尽可能好的识别性能。为了实现这一目标，传统的做法是对目标问题分别采用不同的分类方法处理，然后选择一个最好的分类器为最终的解决方案。但随着目标复杂度的增加以及新算法的开发，人们发现尽管分类器性能有所差异，但被不同分类器错分的样本并不完全重合。即对于某个分类器错分的样本，运用其他分类器有可能得到正确的类别标签，即不同分类器对于分类的模式有着互补信息。如果只选择性能最优的分类器作为最终的解决方案，那么其他分类器中一些有价值的信息就会被丢弃。于是人们开始研究不同分类器的分类互补信息是否能被充分利用，集成分类思想就是在这种条件下提出来的。传统的基于统计学习的分析手段在集成学习的理论分析上已经显出其局限性，需要考虑引进新的数学工具来研究集成学习的理论。同时，集成学习的应用性研究也受到越来越多的重视。

目前，集成学习已经被成功应用于遥感数据分类、雷达目标识别、手写体识别、人脸识别、时间序列预测、蛋白质序列辨识、语音识别、图像处理、文本分类、网络入侵检测、疾病诊断等许多实际问题。尽管如此，与生物认知系统相比，模式识别系统的识别能力和鲁棒性还远不能让人满意。模式识别还有许多的基础理论和基本方法等待人们解决，新问题也层出不穷。为此，相关人员很需要一本关于这一领域的高水平学术著作，既要包括基础知识的介绍，也要包括本领域研究现状以及未来发展的展望等。《模式分类的集成方法》正是这样一本经典著作，是智能信息处理领域难得一见的优秀著作。

Lior Rokach 教授是一位国际上公认的智能信息系统领域的专家，是该领域多个方向上的领军人物，在国际主流期刊和会议发表学术论文 100 余篇（如 Machine Learning, Machine Learning Research, Data Mining and Knowledge Discovery, IEEE Transactions on Knowledge, and Data Engineering and Pattern Recognition）。编辑和出版了十余部专著，都深受广大读者喜爱，非常畅销。这本《模式分类的集成方法》出版后也是广受好评，必将对该领域产生深远的影响。如牛津大学出版社的马克列文对该书的评价为：“这是一本非常及时的出版物，该书是一本全面和详细的参考书，该书对于希望拓宽他们在这一领域的知识或正在此领域或相关领域做

项目的研究者或工程技术人员是非常有用的。它对于那些想将集成学习方法用于数据挖掘工作的科技人员也是非常有用的”。

该书主要讨论了集成学习的概念、构成、作用及其最新研究成果，重点介绍了最新的、高效且实用的集成学习算法，给出了分类识别的大量应用实例，总结了作者近年来在模式识别中的理论和应用研究成果。除了介绍许多重要经典的内容以外，书中还包括了最近十几年来刚刚发展起来的并被实践证明有用的新技术、新理论。并将这些新技术应用于模式识别当中，同时提供这些新技术的实现方法和 JAVA 程序源代码及相应的实验数据，对于读者的自学和算法验证非常有利。针对其中最具有代表性的几种算法，对其工作机制进行深入研究，并利用大量的数值试验对算法的性能进行多方面的考查。主要内容包括：模式分类概述；集成学习的基本理论；分类器组合；经典的集成方法；集成分类的各子模块方法；集成多样性；集成选择；集成算法的评价。这些探讨不仅对集成学习领域的研究具有非常重要的理论意义，而且也具有很强的实用参考价值。

该书可作为高等院校人工智能、模式识别、信息系统、统计分析和管理、电子科学与技术、计算机科学与技术、生物医学工程、控制科学与工程及其他领域的有关专业和研究方向的研究生、本科高年级学生作为关于信息分析、检测、识别的教材或教学参考书。也可作为从事模式识别等相关领域并具有一定理论基础的科研人员、工程技术人员学习、借鉴和参考。

本著作得到国家自然科学基金（61573375，61273275，61402517）、航空科学基金（No.20151996015）宇航动力学国家重点实验室开放基金项目（2012ADL-DW0202）、中国博士后基金（2013M542331，2015M572778）和陕西省自然科学基金（2013JQ8035）的资助支持，特此感谢！

译　者

2015 年 7 月

序　言

集成方法的本质是模仿人类在面对多个选项时作决策的智能行为。其核心原理是对多个个体模式分类器进行权衡处理，并将其整合到一起形成一个分类器，这个新的分类器的性能要好于任何一个单一的分类器。

来自不同学科领域的研究人员，如模式识别、统计分析和机器学习等，在过去的数十年里对集成方法进行了深入的研究。随着对这一方向的持续关注，每个特定领域的研究人员就产生了种类繁多、适合其具体应用环境的集成算法。《模式分类的集成方法》这本书的主旨是对各种集成方法提供一个条理分明、结构合理的介绍，即通过整合这一领域各种各样的思想，提出一个基本一致和统一的思想框架，来介绍有关集成的算法、理论、趋势、挑战和应用。

该书信息丰富、贴近实际，为科研人员、大学师生和应用领域的工程师提供了一个有关集成方法的综合、简明和便利的参考资料。这本书详细介绍了一些经典的集成算法，以及近年来出现的一些扩展方法和新方法。在介绍每种方法的同时，也解释了该方法适用的环境以及会产生怎样的结果和需付出的代价。这本书专注于集成方法领域的研究成果，涵盖了该领域最重要和最受关注的各子方向的研究对象。

该书共七章。第 1 章对模式识别的基本理论进行了概要介绍。第 2 章介绍了构建一个集成分类器的基本算法框架。第 3 章～第 6 章介绍了设计和应用集成方法涉及到的各类子问题。最后，第 7 章讨论了如何对一个集成算法进行评价。

除了对集成方法的理论进行介绍之外，丰富的应用实例纵贯全书，以对相应的理论进行解释说明和论证，包括人造数据和真实的数据。并且对广泛使用的经典实例进行了重点验证。书中用到的数据和算法的 JAVA 程序可以从相应网站获得，非常有利于读者对各种算法的理解和应用。

该书为模式识别、信息系统、计算机科学、统计分析和管理等领域的研究人员提供了一个相当有用的集成方法的参考资料。另外，这本书也对那些从事社会科学、心理学、医学、遗传学和其他需处理复杂数据的科研人员和工程技术人员具有很高的参考价值。

这本书的主要素材来源于以色列本古里昂大学本科和研究生课程的核心内容。此书可作为研究生和高年级本科生的模式识别、机器学习和数据挖掘等课程的参考教材。将集成方法用于实际的数据挖掘项目的读者也会对此书产生独特的

兴趣。此书运用大量的数学语言来描述问题和解决方案，所以行文比较严谨枯燥。尽管如此，对于此书的读者来说只要具备基本的概率论和计算机科学（算法）方面的基础知识也就足够了。

由于集成方法的范围非常广泛，所以不可能在一本书中涵盖所有的技术和算法。感兴趣的读者可以参考一些其他的经典著作，如由 Ludmila Kuncheva (John Wiley & Sons, 2004)所著的《模式分类器：方法和算法》，以及其他一些期刊和会议集。《信息融合》和《信息融合前沿》等期刊都包含了大量的集成方法。此外，许多关于模式识别、机器学习和数据挖掘的期刊也收录关于集成技术的研究论文。另外，一些重要的国际会议，如“多分类器系统国际研讨会（MCS）”和“信息融合国际会议（FUSION）”也是获取相关参考资料的重要来源。

我的许多同事对此书的初稿不吝赐教，提出了宝贵的建议。在这中间，特别要指出的是学者 Alon 博士，他的建议详尽而深刻。作者要特别感谢 Oded Maimon 教授对于此书给出的建议。感谢 Horst Bunke 和 Patrick Shen-Pei Wang 教授，本书的内容吸纳了他们在机器感知和人工智能方面的系列论述。作者还要对世界科学出版的编辑 Steven Patt 先生以及其他工作人员表示感谢，我们在本书的出版过程中合作的非常愉快。

最后，但当然不是微不足道的，作者还要特别感谢我的家人和朋友们给予的耐心、时间、支持和鼓励。

Lior Rokach
以色列贝尔·谢瓦内盖夫本古里安大学
2009 年 9 月

目　录

第 1 章　模式分类概述

模式识别是一门理工类的学科，其目的旨在将模式（也称为示例、数组和样本）划分为不同的类别，即分类或类别标识。通常来说，分类一般基于统计模型，该模型可以由一组已知类别的模式推导出来。另外，也可以利用领域专家的知识进行分类。

一个模式通常由描述某个对象的测量值（特征值）组成。例如，假设我们要对鸢尾花（Iris）的属进行分类（如将其分为 Iris Setosa, Iris Versicolour 和 Iris Virginica 三个子类）。对此模式的特征值可由鸢尾花萼和花瓣的长度和宽度组成。每个样本的标识由 Iris Setosa, Iris Versicolour 和 Iris Virginica 中的一个来确定。另外，样本标识也可由 1，2，3 或 a,b,c，或者任何 3 个明显不同的值来表示。

模式识别另一个常用的例子是光学字符识别（OCR）。此应用实例是将扫描文本转换为机器可编辑文本，以便于存储和检索。每个文本要经历三个步骤。第一步，对文本进行光学扫描，将其转换为位图格式。第二步，对扫描位图中的字符进行分割，以使每个字符相互分隔开，然后，利用特征提取器对每个字符进行特征提取，如空白区域、封闭形状、对角线和交叉线等。第三步，将扫描字符的特征与其对应的字符特征相关联。关联过程通过某个模式识别算法来实现。这样，标识/种类/类别的集合就变成了字符特征的集合，即字母、数字、标点符号等。

1.1　模式分类

在典型的统计模式识别应用中，模式集 S 也称为训练集，是事先给定的。集合 S 中的样本的标识是已知的，目的是通过训练集来构建一个算法对新的样本赋予标识。分类算法也称为诱导器，由特定训练集构建的诱导器的实例也称为分类器。

训练集的表示方法很多，最常见的是每个样本由一个向量描述。每个向量属于一类，并与其类别标识相关联。这样，训练集被存储在一个表中，表的每一行代表一个不同的样本。令 A 和 y 分别表示 n 个特征集合 $A=\{a_1,a_2,\cdots,a_i,\cdots,a_n\}$ 和类别标识。

样本特征也指属性，根据类别标识类型通常可分为以下两类。

（1）语义性的标识。该类别标识是一个无序集中的成员。在这种情况下，域值通常表示为 $\mathrm{dom}(a_i)=\left\{v_{i,1},v_{i,2},\cdots,v_{i,|\mathrm{dom}(a_i)|}\right\}$，式中，$\left|\mathrm{dom}(a_i)\right|$ 是域的有限集的势。

（2）数值性的标识。该类别标识是实数。数值特征可是无限集的势。

类似地，$\mathrm{dom}(y)=\{c_1,c_2,\cdots,c_k\}$ 构成标识的集合。表 1.1 显示了鸢尾花数据集的一个片段。Iris 数据集是模式识别领域最为经典的数据集之一。它首先由 R. A. Fisher 于 1936 年引入模式的示例中。此例的目的是将鸢尾花按其典型特征划分为子属类。

表 1.1　Iris 数据集（4 个数值特征和 3 个子类）

萼片长度	萼片宽度	花瓣长度	花瓣宽度	类别（Iris 类型）
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
6.0	2.7	5.1	1.6	Iris-versicolor
5.8	2.7	5.1	1.9	Iris-virginica
5.0	3.3	1.4	0.2	Iris-setosa
5.7	2.8	4.5	1.3	Iris-versicolor
5.1	3.8	1.6	0.2	Iris-setosa
⋮	⋮	⋮	⋮	⋮

此数据集包含 3 类，分别对应于鸢尾花的 3 种类型：即 $\mathrm{dom}(y)=\{\mathrm{IrisSetosa},\mathrm{IrisVersicolor},\mathrm{IrisVirginica}\}$。每个模式由 4 个数字特征来描述（cm）：$A=\{\text{萼片长度},\text{萼片宽度},\text{花瓣长度},\text{花瓣宽度}\}$。

示例空间（所有可能样本的集合）定义为所有输入特征域的笛卡儿乘积的形式：$X=\mathrm{dom}(a_1)\times\mathrm{dom}(a_2)\times\cdots\times\mathrm{dom}(a_n)$。全体实例空间（或标识示例空间）$U$ 定义为所有输入特征域和目标特征域的笛卡儿乘积，即 $U=X\times\mathrm{dom}(y)$。

训练集表示为 $S(B)$，由 m 个元组组成，即

$$S(B)=\left(\langle x_1,y_1\rangle,\cdots,\langle x_m,y_m\rangle\right) \tag{1.1}$$

式中：$x_q\in X$；$y_q\in\mathrm{dom}(y)$；$q=1,2,\cdots,m$。

通常，假设训练集是随机产生的，并且是依照空间 U 上的某种固定的、未知的联合概率分布的独立分布。需指出的是在用一个函数进行有监督的分类中，这是一种通常的设置。

如上所述，诱导器的目的是产生分类器。通常，一个分类器按照训练集中模式的标识来划分示例空间。由诱导器构建的分隔区域的界线称为边界线，以使新的样本可划分到正确的区域。具体来说，就是由上面定义的未知固定分布

D 给定一个带有输入特征 $A=\{a_1,\cdots,a_i,\cdots,a_n\}$ 的训练集 S 和一个语义目标特征 y，来推导出一个具有最小泛化误差的最优分类器。

泛化误差由分布 D 上的误分类率来定义。令 I 表示一个诱导器。定义由 I 从训练集产生的 $I(S)$ 为分类器。用分类器 $I(S)$ 对一个样本 x 进行分类表示为 $I(S)(x)$。在目标特征为名词性标识的情况下，泛化误差表示为

$$\varepsilon\left(I(S),D\right)=\sum_{\langle x,y\rangle\in U}D(x,y)\cdot L\left(y,I(S)(x)\right) \tag{1.2}$$

式中 $L\left(y,I(S)(x)\right)$ 是 0/1 损失函数，即

$$L\left(y,I(S)(x)\right)=\begin{cases}0 & ,y=I(S)(x)\\ 1 & ,y\neq I(S)(x)\end{cases} \tag{1.3}$$

在数值标识的情况下，求和算子替换为求积分算子。

1.2 诱导算法

诱导算法或简称为诱导器（也称为学习器），是基于已知训练集来构建模型的算法，以使输入特征与目标特征之间产生某种关联关系。例如，一个诱导器可由特定的输入训练样本集及其相应的类别标识来构建，并产生一个分类器。

令 I 表示一个诱导器。定义由 I 从训练集产生的 $I(S)$ 为分类器。用 $I(S)$ 可对一个样本 x 预测其标识，表示为 $I(S)(x)$。

对模式分类领域丰富的研究成果和最新的研究进展略作探究，便不难找到一些适用于初学者的成熟的诱导算法。

大多数分类的核心构成是模型，该模型标明了如何对一个新的样本进行分类。不同诱导器的模型表示形式是不同的。例如，C4.5 算法[Quinlan (1993)]将模型表示为一个决策树的形式，而朴素贝叶斯[Duda and Hart (1973)]诱导器以概率的形式来表示模型。诱导器可以是确定性的（如 C4.5 算法），也可以是随机性的（如反向传播算法）。

两种形式都可以对一个新的样本进行分类。分类器可以明确的将某个类别标识赋予给定样本（清晰分类器），或者，分类器可以产生一个条件概率向量，该向量表明给定样本属于每一类的概率（概率分类器）。在这种情况下，就可以对一个观测值 x_q 估计其条件概率 $\hat{P}_{I(S)}\left(y=c_j\left|a_i=x_{q,i};i=1,2,\cdots,n\right.\right)$。为了区别于实际条件概率，这里用“尖帽（∧）”符号来表示条件概率估计。能够构建概率分类器的诱导器称为概率诱导器。

下面简要回顾通用的方法，以便于概念的学习，包括：决策树、神经网络、遗传算法、基于示例的学习、统计方法、贝叶斯方法和支持向量机。这些方法在本书的后续章节中都进行了详细的讨论。

1.3 规则推导

规则推导算法产生一组 if-then 规则，用于描述分类的过程。此方法的主要优点是高可理解性。即规则可被当作一个连贯英语形式的条件语言采集器，很易于应用。大多数规则诱导算法基于分隔和克服范例[Michalski (1983)]。因此，这类算法具有以下优点：① 能够发现简单的轴向平行边界；② 很好的适合于语义领域；③ 易于处理不相关的属性特征。然而，规则诱导算法在非轴向平行边界的情况下并不适用。另外，这类算法也易遭受断片问题，即在诱导过程中可用数据会出现退化现象[Pagallo, Huassler (1990)]。另一个需避免的问题是小规模脱节或过拟合问题。这种问题是由于规则只覆盖了一小部分训练样本，因此模型对训练数据拟合得很好，但对新样本的分类却造成很高的误差[Holte, et al. (1989)]。

1.4 决策树

决策树是通过对样本空间进行递归分区构建的一类分类器。其模型被描述为一棵有根的树，即带有一个节点、没有前向边的方向树被称为“根”。所有其他节点都有一个前向边。有一个后向边的节点称为“内部节点”或“测试节点”。其余的所有节点称为“叶节点”（也称为“终节点”或“决策节点”）。在一个决策树中，每个内部节点依据输入特征值的某个离散函数，将示例空间划分为两个或多个子空间。最简单也最常见的情况是每次测试只考虑一个特征，然后依据特征值来划分示例空间。如果特征值是数字型的，就是指一个范围。

每个叶节点按照最合适的目标值划归为某一类。或者，叶节点也可依据一个概率向量（亲和度向量）来表示类的归属，向量中的每个元素表示目标特征属于某一类的概率。图 1.1 为一个决策树的例子，描述了表 1.1 中鸢尾花识别任务的解决方案。

图 1.1 中内部节点表示为圆圈，叶节点表示为三角形。每个内部节点（非叶节点）可以有两个或多个后向分支。每个节点相应于某种特性，分支相应于值的范围。这些值的范围必须能够对给定特性的值的集合进行划分。

样本通过一条从根节点到叶节点的贯穿整个决策树的路径来分类，此路径由每个节点划分条件的结果来确定。具体来说，从根节点出发，测试根节点的特性，然后找出给定特性的观测值相应于哪一个后向分支。下一个节点就是所选择分支的末端对应的节点。对新节点重复以上操作并依次贯穿整个树，直至到达叶节点为止。

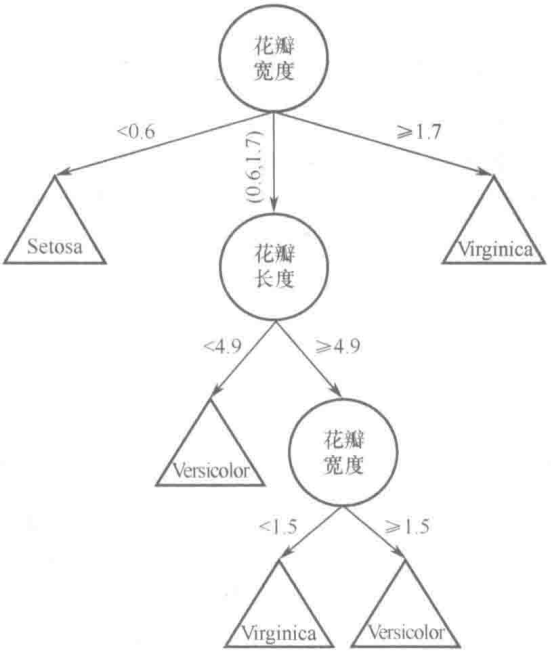

图 1.1　求解鸢尾花分类任务的决策树

决策树可用于名词性的和数值性的特征。在数值性特征的情况下，决策树在几何上可被解释为一个超平面的集合，每个特征与某个坐标轴是正交的。

自然地，对于决策树的构建者而言，都期望得到的决策树尽量简单，而不是更加复杂。此外，按照 Breiman[Breiman, et al. (1984)]的研究，决策树的复杂度对其性能有着重要的影响。通常，通过对数据的过拟合而获得的决策树规模比较庞大，而其泛化性能就比较差（与规则分类器一样）。然而，如果不是通过对数据的过拟合而获得的一个大的决策树，其对新样本的泛化性能就比较好。树的复杂程度可以通过用于构建树的停止标准和删减策略来控制。考查树的复杂程度通常包括：① 节点的总数；② 叶节点的总数；③ 树的深度；④ 所使用的特征的总数。

决策树的构建与规则的推导紧密相关。一个决策树中从根到叶的每一条路径可以转换为一条规则，即将对路径的测试作为规则的前件，将叶节点对类的预测作为规则的后件。所得到的规则集可以被简化以提高它的可理解性，并提高其精度[Quinlan (1987)]。

决策树诱导器是由给定数据集自动构造一个决策树的算法。通常，其目标是通过最小化泛化误差来找到最优的决策树。但是，也有其他类型的一些目标函数，例如，最小化节点数或最小化树的平均深度。

基于给定数据集来构造最优决策树是一项相当困难的工作。Hancock [Hancock, et al. (1996)]指出对于给定数据集，找到最优决策树是一个 NP-难题，而 Hyafil[Hyafil, Rivest (1976)]证明了在必需的期望数目的测试样本下，为了对

一个未知样本进行分类，构建一个最小二叉树的工作是 NP-完全问题。甚至对于一个给定的决策树找到其对应的最小决策树[Zantema, Bodlaender (2000)]，或者从已知的决策表中构造最优决策树[Naumov (1991)]，都是 NP-难题。

这些结果表明，最优决策树算法仅仅适合于非常小的数据集和非常少的特征数。因此，启发式方法就成为解决这类问题所必需的。大体上来说，这些方法可以分为两大类：从上至下的方法和从下至上的方法，已有文献中采用第一种的偏多。

现有许多从上至下的决策树诱导器，如 ID3 算法[Quinlan (1986)]，C4.5 算法[Quinlan (1993)]，CART 算法[Breiman, et al. (1984)]等。其中有些诱导器包括两个阶段：增长和剪枝（如 C4.5 算法和 CART 算法），而其他一些诱导器仅包含增长阶段。

图 1.2 给出了包含增长和剪枝阶段的从上至下诱导算法来构造一个决算树

TreeGrowing (S, A, y, SplitCriterion, StoppingCriterion)
式中：
S——训练集；
A——输入特征集；
y——目标特征；
SplitCriterion——评价某个划分的方法；
StoppingCriterion——停止增长过程的标准。
创建一个带有单个根节点的新树 T 。
if StoppingCriterion (s) THEN
将 T 标记为一个具有 S 中最普遍类标 y 的叶节点.
else
$\forall a_i \in A$ 搜索 a ，使其能够获得最好的 SplitCriterion (a_i, S) .
以 a 标识 t
for each outcome v_i of a:
设置 $\text{Subtree}_i = \text{TreeGrowing}\left(\sigma_{a=v_i}, S, A, y\right)$.
将 t_T 的根节点与 Subtree_i 相连，连接边标识为 v_i
end for
end if
return TreePruning (S, T, y)
TreePruning (S, T, y)
式中：
S——训练集；
y——目标特征；
T——要裁剪的树。
do
选择 T 中的一个节点 t ,使得裁掉该节点可以最大地提升某个评价指标
if $t \neq \phi$ then $T = \text{pruned}(T, t)$
until $t = \phi$
return T

图 1.2　决策树诱导器从上至下算法框架