基于决策树改进CART算法的决策支持与分析技术

李春鑫[1] 李天伟[2]

(海军大连舰艇学院研究生15队 大连116018)[1] (海军大连舰艇学院航海系 大连116018)[2]

摘 要 针对决策支持与分析技术,提出了基于决策树的改进CART算法。该算法由树生长和树剪枝两部分构成,具有辨识相关输入的能力,由于引入了递归最小二乘估计器,对线性模型可降低计算量,并采用模糊技术处理不连续边界问题。我们给出了该算法的应用实例,由于隐含权值归一化,该算法能够快捷地对自适应神经模糊推理系统进行结构辨识。

关键词 决策树,CART算法,递归最小二乘估计器,自适应神经模糊推理系统

Decision Support and Analysis Technique Based on Improved CART Arithmetic of Decision Trees

LI Chun-Xin[1] LI Tian-Wei[2]

(Dalian Naval Academy, Dalian 116018, China)[1] (Dept Navigation, Dalian Naval Academy)[2]

Abstract Aiming at decision support and analysis technique, improved CART arithmetic of decision trees is put forward. The arithmetic consists of developing trees and cutting trees, because recursive LSE is introduced, it can reduce calculation for linear model, besides, to deal with the question of discrete boundary, fuzzy technique is used, and it also has the ability of identifying relevant input. The application of the arithmetic is also presented. Due to the normalization of connotative weight, the arithmetic can identify the structure of ANFIS conveniently.

Keywords Decision trees, CART, RLSE, ANFIS

1 引言

决策是针对某一问题,根据确定的目标及当时的实际情况制定多个候选方案,然后按一定标准从中选出最佳方案的思维过程。在许多情况下,决策一旦作出就无法挽回,因而如何快速准确地作出决策显得极其重要。基于此我们提出了决策树的改进CART算法。

2 CART算法的原理

决策树是把数据集的输入空间划分为互斥区域,每个区域赋予一个标识、一个值和一个表示该区域内数据点特色的动作。为了构造一个合适的决策树,CART首先基于采样数据集广延地生长树,然后,基于最小复杂性代价准则[1],再回头修剪这棵树,这样得到一系列不同大小的树,最后所选择的那

李春鑫 硕士研究生,主要从事军事航海信息及控制关键技术的研究。

格[7,8]等并不违背,相反,OBSA是上述两种技术的一种具体的实现方式。

参 考 文 献

1 Uschold M, Gruninger M. Ontologies: Principles, Methods and Applications. Knowledge Engineering Review, 1996, 11(2): 93~155

2 Decker S, Erdmann M, Fensel D, et al. ONTOBROKER: Ontology based Access to Distributed and Semi-Structured Information. In: R. Meersman et al. eds. Semantic Issues in Multimedia Systems Boston: Kluwer Academic Publiser, 1999

3 Klein M, Fensel D, Harmelen F, et al. The Relation between Ontologies and XML Schemata. In: Proc. of the (ECAI)'00 Workshop on Applications of Ontologies and Problem-Solving Methods, Berlin, 2000

4 Kifer M, Lausen G, Wu J. Logic Foundations of Object-Oriented and Frame-Based Languages. Journal of the ACM. 1995, 42(4): 741~843

5 Erdmann M, Studer R. Ontologies as Conceptual Models for XML Documents. In: Proc. of the KAW '99 12th Workshop on Knowledge Acquisition, Modelling and Management

6 张维明. 语义信息模型及应用. 电子工业出版社, 2002. 77~79

7 Cannataro M, Talia D. Knowledge Grid—An Architecture for Distributed Knowledge Discovery. CACM, 2003, 46(1): 89~93

8 Goble C A, Roure D D, Shadbolt N R, et al. Enhancing Services and Applications with Knowledge and Semantics. In: The Grid2: Buleprint for a New Computing Infrastructure, New York: Morgan Kaufmann Publishers, 2004. 431~458

棵树为用另一组独立的数据时具有最好性能的树。因而，CART 算法由两部分组成：树生长和树剪枝。

2.1 树生长

通过将训练数据划分为不相连的子集的一个个分叉(决策边界)，CART 生长为一棵树。从包括所有训练数据的根结点开始，为求最能减少误差指标的分叉，做一次穷尽搜索。一旦确定最佳分叉，数据集相应地划分成不相连的子集；这些子集用源于根结点的子结点表示。然后，再对子结点实施同样的划分。当与一个结点有关的误差值小于某个阈值时，或当进一步划分树，误差的减少不超过某个阈值时，这个递归过程终止。下面详细说明递归树的生成。

递归树用于解决递归问题，用一个对象的多个属性确定其一个或多个数值属性。对于一个递归树，结点误差指标常取为拟合节点数据集的局部模型的平方误差或残差：

$$E(t)=\min_{\theta}\sum_{i=1}^{N(t)}(y_i-d_t(x_i,\theta))^2 \tag{1}$$

式中$\{x_i,y_i\}$是典型的数据点，$d_t(x_i,\theta)$是结点 t 的局部模型(θ 可变)。

把结点 t 分解成 t_l 和 t_r 的任意分叉 s，误差测度的变化可表示为：

$$\Delta E(s,t)=E(t)-E(t_l)-E(t_r) \tag{2}$$

最好的分叉 s^* 为误差测度降低最多的分叉：

$$\Delta E(s^*,t)=\max\Delta E(s,t) \tag{3}$$

生成递归树的策略是反复地分叉结点，这样最大限度地减少递归树的整体误差测度 $E(T)$。因此，递归树的目标是：以一步超前、贪婪的方式，递归地分解分叉结点，使给定的合理误差测度最小。

2.2 树剪枝

由以上算法生成的树常常规模很大，而且与训练数据集有偏差，必须进行剪枝处理，基于最小代价复杂性或最弱子树收缩原理是最有效的方法之一，该方法第一步是产生一棵充分张开的树 T_{max}，这棵树拟合训练数据相当好，但规模较大，因此我们要寻找其中的最弱子树进行剪枝。考虑训练误差测度和终结点数目，即考虑树的复杂性指标，就可以找到最弱子树。

(1)对于任意子树 $T\subset T_{max}$，定义其复杂度为 T 中的终结点数目 $|\overline{T}|$。那么代价复杂性测度 $E_\alpha(T)$ 定义为：

$$E_\alpha(T)=E(T)+\alpha|\overline{T}|,\alpha\text{ 是代价复杂度参数} \tag{4}$$

(2)对于每个 α，对应于给定的 $E_\alpha(T)$，可以找到一个最小子树 $T(\alpha)$：

$$E_\alpha(T(\alpha))=\min_{T\subset T_{max}}E_\alpha(T) \tag{5}$$

(3)当 α 值增大时，$T(\alpha)$一直保持最小，直到到达一个跳跃点 α'，此时，树 $T(\alpha')$成为新的最小树。设 T_{max}有 L 个终结点。采用逐步向上进行树剪枝的思想使得满足：

$$\{t_1\}=T_1\subset\cdots\subset T_{L-1}\subset T_L=T_{max} \tag{6}$$

式中 T_i 有 i 个终结点

(4)求树 T 的下一棵最小树。对于 T 中的每个内结点 t，求使得 $T-T_t$ 为下一棵最小树的 α 值，记为 α_t：

$$\alpha_t=\frac{E(t)-E(T_t)}{|\overline{T}_t|-1} \tag{7}$$

(5)选择具有最小 α_t 的内结点作为剪枝的目标结点。树的剪枝过程为：

①计算 T_i 中每个内结点 t 的 α_t 值；

②求最小 α_t 并选择 $T-T_t$ 为下次最小树；

③判断是否只有一个根结点，若不是，则转①；

④用独立测试(检验)数据集的方法选择最优规模树。

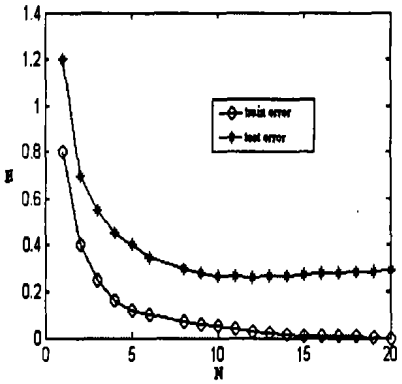

图1 相对于树规模的误差测度

图1是由上述的树剪枝过程得到的关于候选树($T_1,T_2,\cdots,T_L,L=20$)的树误差测度相对于树规模的典型模式，T_8到 T_6的跃变表明 T_8中缩减的子树有两个以上的终结点，由图可知优化规模为12。E 表示误差测度；N 表示终结点数。

3 改进的 CART 算法

为构造有恒定输出的终结点的递归树，先前描述的 CART 算法可以辨识出适当规模的树，并确定树不需要的无关输入，但如果终结点是用线性方程来描述其特征的，为求相关输入，需要更大的计算量，这是很不方便的。为此，我们提出降低计算量的方法－RLSE(递归最小二乘估计器)，即在新数据和新参数适应过程中递归地得到最小二乘估计器。

对于矛盾方程 $A\theta=y$(其中$[AMy]$的第 k 行表示为$[a_k^TMy_k]$，$1\leqslant k\leqslant m$)，最小二乘估计器为：

$$\theta_k=(A^TA)^{-1}A^Ty \tag{8}$$

为了利用已得的 θ_k，以最小代价计算 θ_{k+1}，而不是使用所有可得的数据重新计算，我们引入递归最小二乘估计器，它可按下式计算：

$$\begin{cases} P_{k+1}=P_k-\dfrac{P_k a_{k+1}a_{k+1}^T P_k}{1+a_{k+1}^T P_k a_{k+1}} & (9) \\ \theta_{k+1}=\theta_k+P_{k+1}a_{k+1}(y_{k+1}^T-a_{k+1}^T\theta_k) & (10) \end{cases}$$

其中$1\leqslant k\leqslant m-1$，最终的$\theta$等于$\theta_m$，即使用所有$m$个数据对的估计器。

此外，在使用CART算法构造树的整个输入输出映射中，会产生不希望的不连续边界，为了光滑每个分叉点上的不连续边界，我们采用模糊集的方法，将决策树转化为模糊问题进行处理。

4 改进CART算法的决策支持与分析技术的应用

ANFIS(自适应神经模糊推理系统)的学习法则以及任何其他的参数自适应方法仅设计参数辨识，在进行任何参数调节过程之前，我们还需要用结构辨识方法确定一个初始的ANFIS结构，有了可靠的结构和参数辨识方法，我们方可完成模糊建模的周期。在训练之前，我们可用改进CART算法求出ANFIS规则的数目和隶属函数的初始位置，从而完成结构辨识。而改进CART算法用于ANFIS的结构辨识的优势在于隐含权值的归一化定理。

命题：CART构造的ANFIS网络中隐含权值归一化。

即在把决策树转化为推理系统时，如果(1)$\mu_{x>a}(x)+\mu_{x\leqslant a}(x)=1$，其中$x$是任意的输入向量，$a$是$x$的任意分叉点，$\mu$为隶属函数；(2)T-范式算子取乘积算子，用于计算每条规则的激励强度，则每条规则的激励强度的总和恒为1。我们用数学归纳法予以证明。

证明：令n为规则数，$\omega_i(i=1,2,\cdots,n)$是第i条规则的激励强度。

当$n=2$时，因为ω_1和ω_2分别是和的隶属度，所以我们有$\omega_1+\omega_2=1$成立；

当$n=k$时，假设$\sum_{i=1}^{k}\omega_i=1$成立。则当$n=k+1$时，不失一般性，我们假设新产生的规则是k和$k+1$，它们是上次终结点k分叉的结果，因此，我们有

$$\sum_{i=1}^{k+1}\omega_i=\sum_{i=1}^{k-1}\omega_i+\omega_k+\omega_{k+1}=\sum_{i=1}^{k-1}\omega_i+\bar{\omega}_k(\mu_{x<a}(x)+\mu_{x\geqslant a}(x))=\sum_{i=1}^{k-1}\omega_i+\bar{\omega}_k=1$$

$\bar{\omega}_k$是规则k分叉之前的激励强度。证毕。

我们以二叉决策树为例，该决策树可等效于一组确定的规则集，如图2所示。我们利用改进的CART算法，得出ANFIS结构如图3所示。由于图3中的ANFIS结构的隐含权值归一化，在整个训练过程中保持不变，不仅消除了归一化层的需要，并且减少了训练和应用的计算时间以及舍入误差。

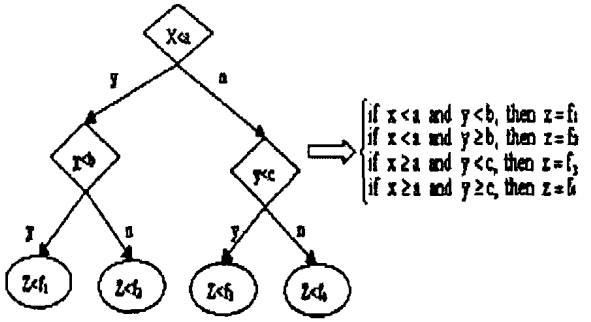

图2 二叉树等效成模糊集

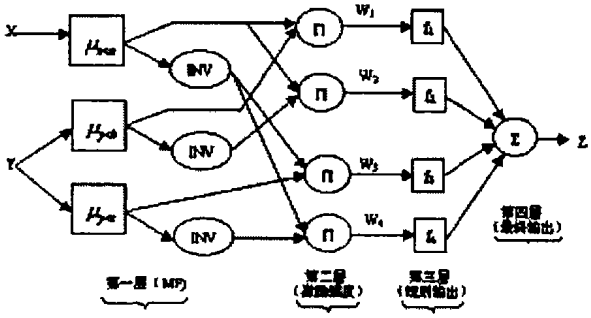

图3 ANFIS的结构

结束语 决策树机理透明，便于我们按照树结构解释如何做出一个决策。而CART算法是最具有代表性的决策树归纳方法，CART算法能迅速地确定一个模糊推理系统的大致结构，然后对没有归一化层的有效ANFIS结构，选择合适的隶属函数和输出函数。CART能选择相关的输入，并对输入空间进行树划分，而ANFIS可改善其划分结果，并使该结果处处光滑连续。可以说CART和ANFIS在功能上是互补的。我们这里只着眼于递归问题，类似的方法也能用于分类问题。

参考文献

1 Breiman L, Friendman J H, Olshen R A, Stone C J. Classification and regression trees. Wadsworth, Inc., Belmont, Californiz, 1984

2 Jang J-S R. Structure determination in fuzzy modeling: a fuzzy CART approach. In: Proc. of IEEE Intl. Conf. on Fuzzy Systems, Orlando, Florida, June 1994

3 李书涛. 决策支持系统原理与技术[M]. 北京：北京理工大学出版社，1996

4 王永庆. 人工智能原理与方法[M]. 西安：西安交通大学出版社，1998

5 高洪深. 决策支持系统(DSS)[M]. 北京：清华大学出版社，2000