区间值属性决策树学习算法*

王熙照 洪家荣

(哈尔滨工业大学计算机科学系 哈尔滨 150001)

E-mail: wangxz@hbu.edu.cn

摘要 该文提出了一种区间值属性决策树的学习算法.区间值属性的值域不同于离散情况下的无序集和连续情况下的全序集,而是一种半序集.作为ID3算法在区间值意义下的推广,算法通过一种分割信息熵的极小化来选取扩展属性.通过非平稳点分析,减少了分割信息熵的计算次数,使算法的效率得到了提高.

关键词 机器学习,归纳学习,决策树,区间值属性.

中图法分类号 TP18

学习算法通常使用一个启发式来引发在属性值和类组合而成的空间上的搜索.用训练样本产生决策树时,一种较有效且常用的启发式是依据类信息熵极小化来选择属性的,它最早由J. R. Quinlan[1]在他的ID3算法中提出.迄今为止,用ID3算法产生决策树已有了多种改进和推广,如文献[2~6].其中,在文献[2]中,J. Cheng等人通过属性值和类表示关系的修订改进了原有的决策树,推广了使用范围;在文献[3]中,J. R. Quinlan注意到,决策树产生结果是明确的,故不能表达分类过程中潜在的不确定性,而属性值的微小变化可能导致分类的剧烈变化,于是,建议采用一种概率方法来构造决策树;在文献[4]中,洪家荣等人从示例学习最优化的角度分析了决策树归纳学习的优化原则,提出了一种基于信息增益的扩展属性选取方法;在文献[5]中,U. M. Fayyad等人使用极小化信息熵,提出了一种产生连续值属性决策树的学习算法;在文献[6]中,Y. Yuan等人在一定程度上处理了产生决策树时存在的不确定性,提出了一种模糊决策树归纳学习算法.

从众多的决策树归纳学习算法中可以发现,学习问题中的属性一般被认为有两种:一种称之为"符号"值属性,另一种称之为"连续"值属性.符号值属性的值域为一个没有序关系的有限集合,而连续值属性值域则是一个具有全序关系的集合例实数或整数的子集.本文研究了一种属性取值为区间的决策树学习问题.区间值属性的值域既不象符号值那样,一般没有序,也不象连续值那样,构成全序,而是具有一种偏序关系.故这种学习问题可认为是介于符号值和连续值之间的一种.文中提出了基于类信息熵极小来选择区间值扩展属性的学习算法,通过对平稳点和非平稳点的分析,提高了算法的效率.示例表明了算法的实用性.

1 区间值属性决策树学习算法

在决策树产生过程中,一个结点处的一个区间值属性常把它的值域分成3个部分.对所考虑的一个区间值属性C(取值为有限闭区间)而言,选定一个阈值T.T值将区间值属性C的值域按$T>C$,$T<C$和$T\in C$分割成3个分支,其中对C的取值$[x,y]$而言,$T>C$指$T>y$;$T<C$指$T<x$;而$T\in C$指$T\in[x,y]$.将这样的一个阈值T称之为割点.这样,在每一个结点处,对此结点上的训练例子选定一个属性的一个割点来进行树的扩展(此属性称之为扩展属性).这种扩展过程直到叶结点,即结点中的例子同属于一类为止,最终一个区间值属性决策树便会生成.

扩展属性的选取对树的扩展过程起着极其重要的作用.假定我们选定了一个结点准备进行树的扩展,此结点上有N个训练例子,所有训练例子共分为k类.设N个例子的取值为$e_i=[e_i^-,e_i^+](i=1,2,\ldots,N)$,将$N$个区间值的端点按从小到大的顺序排列,得到一个由$2N$个点组成的序列.取任两个相邻端点的中点(无重复端点时,有$2N-1$个)作为备选割点.于是,对每一个区间值属性将有1个至多$(2N-1)$个备选割点组成的序列.对选定结点上的每一个区间值属性的每一个备选割点,训练例子集将被分割成3部分.下面,我们来计算分割的信息熵.

---

* 本文研究得到河北省自然科学基金资助.作者王熙照,1963年生,副教授,主要研究领域为机器学习,模糊信息处理.洪家荣,1939年生,1997年逝世,教授,博士导师.

本文通讯联系人:王熙照,保定 071002,河北大学数学系

本文1996-11-05收到原稿,1997-07-18收到修改稿

设结点上的训练例子集 S 共有 k 类：$L_1,L_2,\ldots,L_k$，S 中第 i 类例子所占的比例记为 $P(L_i,S)=Card(L_i\cap S)/Card(S)(i=1,2,\ldots,k)$，其中 $Card(\cdot)$ 表示一个集合的元素个数. 则集合 S 的类信息熵定义为

$$E(S)=-\sum_{j=1}^{k}P(L_j,S)\log P(L_j,S),$$

其中对数函数可选取任何方便的底.

将备选割点 T 训练例子集，分割成 3 个子集 S_1,S_2,S_3，分割的信息熵定义为 $S_i(i=1,2,3)$ 的类信息熵的加权平均.

定义 1. 设有一组例子 S，一个区间值属性 C，一个备选割点 T. S_1,S_2,S_3 是 S 的 3 个子集，满足 $\bigcup_{j=1}^{3}S_j=S$、$S_i\cap S_j=\varnothing(i\neq j)$，则割点 T 引出的分割信息熵定义为

$$E(C,S,T)=\sum_{j=1}^{3}\frac{Card(S_j)}{Card(S)}E(S_j).$$

在属性 C 的所有备选割点中，选择一个使得 $E(C,S,T)$ 达到最小的 T，记为 T_C，称之为属性 C 的最优割点.

对每一个区间值属性都选出其最优割点后，寻找一个最优割点引出的分割信息熵最小的属性作为此结点上的扩展属性，进行决策树的扩展. 重复这种扩展过程，直到结点中的例子同属于一类为止.

考虑算法的复杂性. 一个具有 N 个训练例子的结点上，任何一个区间值属性一般至多有 $2N-1$ 个备选割点，上述讨论表明，为求出最优割点，需要对每一个备选割点计算其分割的信息熵. 实际上，这是不必要的. 下一节关于非平稳点的分析表明了由备选割点引出的分割信息熵总是在非平稳处达到. 因此，在计算过程中，如果一个备选割点是平稳的，则由它引出的分割信息熵不必计算，而仅需计算作为非平稳点的备选割点的分割信息熵. 对于平稳点相对集中的情况，这样可以大大减少熵的计算次数，从而提高算法的效率. 综上所述，一个基于分割信息熵极小启发式的区间值属性决策树学习算法描述如下.

(1) 选定一个结点，任取一个区间值属性 C；

(2) 统计此结点上关于属性 C 的非平稳的备选割点；

(3) 计算非平稳备选割点引出的分割信息熵，筛选出最小者来确定最优割点；

(4) 在所有区间值属性中，筛选出其最优割点引出的分割信息熵最小的属性，以此作为此结点上的扩展属性，进行树的扩展；

(5) 重复上述过程，直到分类结束.

例 1：考虑如表 1 所示的有两类构成的示例学习问题.

表 1　某医院治疗视网膜脱落手术的若干病历记录

序号	网脱时间(月) C_1	网脱范围(象限) C_2	裂孔大小(Pd) C_3	术后效果
1	[8.00,12.0]	[3.50,4.00]	[0.20,0.40]	N
2	[4.00,6.00]	[1.50,2.00]	[0.20,0.80]	P
3	[9.00,13.0]	[2.50,3.00]	[0.30,0.60]	N
4	[1.00,2.00]	[0.00,1.00]	[0.10,0.50]	P
5	[7.00,14.0]	[3.50,3.60]	[0.20,0.30]	N
6	[0.00,1.00]	[1.50,2.00]	[0.40,0.60]	P
7	[5.00,10.0]	[3.00,3.00]	[0.50,0.70]	N
8	[1.00,3.00]	[1.00,1.20]	[0.40,0.50]	N
9	[9.00,13.0]	[2.50,3.50]	[0.20,0.25]	P
10	[0.00,1.00]	[0.50,2.00]	[0.80,1.20]	P
11	[4.00,6.00]	[1.80,2.00]	[0.20,0.90]	P
12	[3.00,5.00]	[1.50,4.00]	[0.20,0.50]	N

表中区间值属性有 3 个，即网脱时间、网脱范围和裂孔大小，依次记为 C_1,C_2 和 C_3. 术后效果一栏里，P(正例类)表示“复位”，N(反例类)表示“好转”. 采用第 1 节提出的算法对此 12 个样本进行训练，得到的决策树如图 1 所示.

2　非平稳割点分析

定义 2. 设某结点上有 N 个例子，其 $2N$ 个端点从小到大依次排列为 $a_1,a_2,\ldots,a_{2N}$(不妨先设无重复值)，任意两相邻端点取中值产生的 $2N-1$ 个备选割点，记为 $p_1,p_2,\ldots,p_{2N-1}$. 对任意的 $i(1<i<2N-1)$，则如果产生 p_{i-1},p_i,p_{i+1} 的 4 个端点同为左端点(或右端点)，且 4 个端点所对应的例子同属于一类割点 p_i 称为是平稳的.

一个割点不是平稳的,则称为是非平稳割点.直观上看,非平稳割点是这样一种点,它的左右两侧的训练例子同属一类且由同一种端点产生.当 $2N$ 个端点有重复值时,重复端点左右相邻的两个备选割点应按非平稳割点处理.

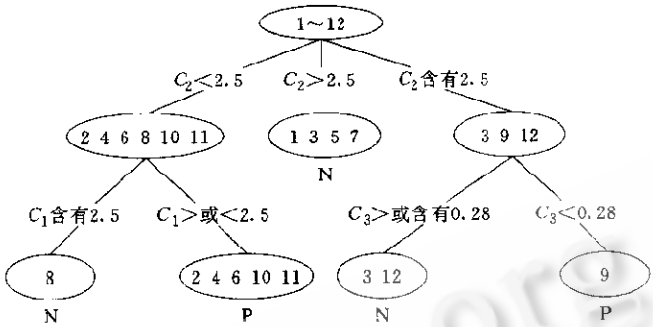
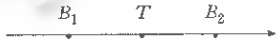


图1　一个关于区间值属性C_1,C_2,C_3的决策树

定理 1. 如果割点 T 使得分割信息熵 $E(C,T,S)$达到极小,则 T 一定是非平稳割点.

证明:为书写简洁,我们不失一般性地假设训练例子集 S 中只有两类,即正例类和反例类.如下所示,考虑平稳割点 T 和它最近的左右两个非平稳割点 B_1 和 B_2.



割点 T 将 S 分成 3 个子集,记小于 T 的子集为 S_1,其中正例 p_1 个,反例 n_1 个;含有 T 的子集为 S_2,其中正例 p_2 个,反例 n_2 个;大于 T 的子集为 S_3,其中正例 p_3 个,反例 n_3 个.设 S 中共有 N 个训练例子,P 个正例,于是由割点 T 引出的分割信息熵 $E(C,T,S)$可表示为

$$f(p_1,p_2,p_3,n_1,n_2,n_3)=-\sum_{j=1}^{3}\frac{n_j+p_j}{n}\left(\frac{p_j}{n_j+p_j}\log\frac{p_j}{n_j+p_j}+\frac{n_j}{n_j+p_j}\log\frac{n_j}{n_j+p_j}\right),$$

其中 $p_1+p_2+p_3+n_1+n_2+n_3=N$. $p_1+p_2+p_3=P$.令 $g=Nf$,于是

$$g(p_1,p_2,p_3,n_1,n_2,n_3)=\sum_{j=1}^{3}(-p_j\log p_j+(p_j+n_j)\log(p_j+n_j)-n_j\log n_j).$$

注意到 B_1,B_2 之间的割点都是平稳的,按定义,这些割点同是左(或右)端点产生的,且端点相应的例子同属于一类,故当平稳割点在开区间(B_1,B_2)内从小到大变化时,函数 g 中仅有 1 个变元变动,不妨设为 p_1.由于约束的存在函数 g 中的变元并非是独立的,将 g 作为 p_1 的函数改写为

$$\begin{aligned}g(p_1)=&\sum_{j=1}^{2}(-p_j\log p_j+(p_j+n_j)\log(p_j+n_j)-n_j\log n_j)\\&-(N-(p_1+p_2+n_1+n_2+n_3))\log(N-(p_1+p_2+n_1+n_2+n_3))\\&+(N-(p_1+p_2+n_1-n_2))\log(N-(p_1+p_2-n_1+n_2))-n_3\log n_3,\end{aligned}$$

其中除 p_1 外,其他变元视为常量.注意到$\frac{\mathrm{d}}{\mathrm{d}x}(x\log x)=1+\log x$,$\frac{\mathrm{d}^2}{\mathrm{d}x^2}(x\log x)=\frac{1}{x}$,于是

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}g(p_1)=-\frac{1}{p_1}+\frac{1}{p_1+n_1}-\frac{1}{N-(p_1+p_2+n_1+n_2+n_3)}+\frac{1}{N-(p_1+p_2+n_1+n_2)}.$$

从 $n_1>0,n_3\geqslant 0$ 或 $n_3>0,n_1\geqslant 0$ 可知,$\frac{\mathrm{d}^2}{\mathrm{d}x^2}g(p_1)<0$.它表明 g 作为 p_1 的函数是下凸的.由于下凸函数的极小值一定在某端达到,故函数 g(即分割的信息熵)的局部极小值比在非平稳点 B_1 或 B_2 达到. □

3　算法的效率分析

将 N 个训练例子的一个属性的 $2N$ 个端点从小到大排列后,即可产生 $2N-1$ 个备选割点.在没有引入非平稳割点之前,需要计算 $2N-1$ 次割点信息熵.假设所考虑的学习问题有 k 个类,则在引入非平稳割点后仅需计算 s 个非平稳割点的信息熵.注意到 $k-1\leqslant s\leqslant 2N-1$,而在一般情况下,$k\ll N$,故在实际问题中,期望割点信息熵的计算次数有明显减少.

考虑文献[7]提供的关于睡眠状态知识自动获取这一典型的示例学习问题.其共有 1 236 个例子、11 个属性和 6 个类.将睡眠的前 3 个状态视为正例类,后 3 个状态视为反例类,并将属性所取的整数值随机延拓成区间形式,则有

$$\frac{\text{计算全部备选割点信息熵的次数}}{\text{计算非平稳割点信息熵的次数}} \approx 2.18.$$

其结果从一个侧面证实了计算量减少的显著性. 从上面的分析可见,就割点信息熵的计算次数而言,最好的情况是端点排列后自然形成 $k-1$ 个非平稳割点,最坏的情况是 $2N-1$ 个割点均是非平稳的情况,它与没有引入非平稳割点前的计算量是一样的.

一般情况下,非平稳割点的个数主要取决于相邻几个端点是否属于同一类. 假设各类是等可能出现的,一个自然的问题就是除去端点的左右因素外,连续 4 个端点值属于同一类的概率有多少? 考虑最坏的情况,即端点值与例子之间是独立的的情况,此时,连续 4 个端点值属于同一类的概率应近似为 $\sum_{i=1}^{k}\left(\frac{C_i}{N}\right)^4$,其中 C_i 为第 i 类例子个数. 注意到等可能性 $C_i/N \approx 1/k$,于是有

$$Prob(\text{相邻 4 端点同类}) \geqslant \sum_{i=1}^{k}\left(\frac{1}{k}\right)^4 = \frac{1}{k^3}.$$

此概率随着端点值与例子之间相关程度的增加而增加. 由于大部分实际问题中,例子与它的端点值之间是密切相关的,故实际连续相邻 4 个端点属于同一类的频率远大于此概率值. 上述关于睡眠状态知识获取的实验结果也显示出这一点.

4 结束语

本文提出了一种基于分割信息熵极小来选择区间值扩展属性的学习算法,通过对平稳点和非平稳点的分析,提高了算法的效率. 示例表明了算法的实用性. 本文的算法可以方便地推广到多维矩形值属性情况,但一般集值属性决策树的学习算法还有待进一步的探讨.

参考文献

1 Quinlan J R. Induction of decision trees. Machine Learning, 1986,1:81～105

2 Cheng J, Fayyad U M, Irani K B *et al*. Improved decision trees: a generalized version of ID3. In: Dietterich T ed. Proceedings of the 5th International Conference on Machine Learning. San Mateo, CA: Morgan Kaufmann Publishers, 1988. 100～108
Quinlan J R. Probabilistic decision trees. In: Kodratoff Y, Michalski R eds. Machine Learning: An Artificial Intelligence Approach. Vol 3. San Matyeo, CA: Morgan Kaufmann Publishers, 1990

3 洪家荣, 丁明峰, 李星原等. 一种新的决策树归纳学习算法. 计算机学报, 1995,18(6):470～474
(Hong Jia-rong, Ding Ming-feng, Li Xing-yuan *et al*. A new algorithm for decision tree induction. Chinese Journal of Computers, 1995,18(6):470～474)

5 Fayyad U M, Irani K B. On the handling of continuous-valued attributes in decision tree generation. Machine Learning, 1992, 8:87～102

6 Yuan Y, Shaw M J. Induction of fuzzy decision trees. Fuzzy Sets and Systems, 1995,69:125～139

7 Michalski R S, Mozetic I, Hong Jia-rong. The multipurpose incremental learning system AQ15. In: Revist M ed. Proceedings of the 5th National Conference on Artificial Intelligence. Philadelphia, PA: Morgan Kaufmann, 1986. 1041～1045

Learning Algorithm of Decision Tree Generation for Interval-Valued Attributes

WANG Xi-zhao HONG Jia-rong

(*Department of Computer Science Harbin Institute of Technology Harbin* 150001)

Abstract The authors present a learning algorithm of decision tree generation for interval-valued attributes. With regard to range of value, a nominal attribute is not ordered and a continuous-valued attribute is linearly ordered, but the interval-valued attribute is partially ordered. As a generalization of ID3-algorithm on intervals, this algorithm uses minimal information entropy of partitioning to select the extended attributes. The efficiency of the algorithm is improved by analyzing unstable cut points.

Key words Machine learning, induction, decision trees, interval-valued attributes.