从熵均值决策到样本分布决策*

何劲松+, 郑浩然, 王煦法

(中国科学技术大学 计算机科学与技术系,安徽 合肥 230026)

Decision Varied from Entropy to Parametric Distribution

HE Jin-Song+, ZHENG Hao-Ran, WANG Xu-Fa

(Department of Computer Science and Technology, University of Science and Technology of China, Hefei 230026, China)

+ Corresponding author: Phn: 86-551-3607519, E-mail: hjss@ustc.edu.cn

http://www.ustc.edu.cn

Received 2001-11-14; Accepted 2002-04-10

He JS, Zheng HR, Wang XF. Decision varied from entropy to parametric distribution. *Journal of Software*, 2003,14(3):479~483.

Abstract: In order to improve the predictive accuracy of inductive learning, a heavy analysis about the demerit of C4.5 is given, and the reason why there are many debates and compromise between standard method and meta algorithms is pointed out. By the method of estimating the probability distribution of training examples, a new and simple method of decision tree is turned out. Experimental results on UCI data sets show that the proposed method has good performance on accuracy issue and faster computing speed than C4.5 algorithm.

Key words: machine learning; inductive learning; decision tree; pattern recognition; parametric estimation

摘 要: 为了研究归纳学习的判决精度问题,分析了 C4.5 算法的不足以及标准算法与亚算法之间争论和妥协的根本原因,从估计训练样本的概率分布的角度出发,给出了一种简单而新颖的决策树算法.基于 UCI 数据的实验结果表明,与 C4.5 算法相比,该方法不仅具有比较好的判决精度,而且具有更快的计算速度.

关键词: 机器学习;归纳学习;决策树;模式识别;参数估计

中图法分类号: TP18 **文献标识码**: A

1 归纳学习标准算法与亚算法的争论和妥协

作为机器学习诸多方法中的一种,归纳学习在判决精度问题上并不占据优势.但作为智能问题求解中的特定层次,归纳学习可以产生规则的特性却是不可或缺的.归纳学习标准算法与亚算法的争论和妥协的关键正是这二者之间的矛盾与调和.

发现归纳学习的一个基础理论——Occam 算法[1]的不足并倡导研究决策森林(即亚算法)的起源是实验方法[2],而实验所采用的学习算法是 ID3[3]及其衍生 C4.5[4].因此,1990's 标准算法与亚算法的争论和妥协实际上存在于 C4.5 与亚算法的典型代表 Boosting[5]以及 Bagging[6]之间.然而,发现亚算法不足的研究仍然是实验方法[7].

亚算法的提出和发展是为了修补标准算法在学习样本集上的过确定问题和不恰当的决策构造方法或训练

* 第一作者简介: 何劲松(1967－),男,安徽南陵人,博士,主要研究领域为模式识别,机器学习,进化计算.

方法的单一性问题.尽管亚算法在发展的初期缺乏应有的理论根据[7],但它们却取得了实验结果上的优势.而文献[7]认为,亚算法的不足除了缺乏必要的理论根据和计算量很大之外,最根本的问题是它们不能对已知的样本给出完全正确的分类,而且在判决精度方面提高的程度也有限,并不总是优越的.既然双方各有优劣,妥协的结果自然是相互认同、继续研究.

在上面对标准算法和亚算法的对比叙述中,我们所关注的一个基本现象是亚算法正在成为一种合成系统的框架,而标准算法却正成为其框架中的基本成分.在实验结果上来看,亚算法的成功程度仍然与它所选择的标准算法有关.从研究的角度出发,我们所需要讨论的是其中的两个基本问题.即:(1) 标准算法作为亚算法必不可少的基本成分,它的不足究竟在哪里?除了 ID3 及其衍生 C4.5 之外,是否还有其他可行的解决方案?(2) 亚算法作为一种框架结构,它们能否回到基于规则表示的决策形式而不是表决?能否减少计算量而不影响其判决精度?能否可以对已知的例子或学习样本作出完全正确的判断?显然,对于以上两方面问题的解决需要持续性的研究.本文着重考虑的是对前者的研究.

2 从 ID3 和 C4.5 引出的策略问题

ID3 及其衍生 C4.5 被人们认为是标准决策树学习算法中最优秀的算法,并且大量的改进性研究也都是基于它们的核心策略进行的.可以说,C4.5 延续的 8 个版本在改进问题上已经做得很充分了.我们在此对其核心进行分析的目的不是为了对 ID3 或 C4.5 作直接的改进,而是通过分析找出其未涉及或难以涉及的问题.

ID3 及 C4.5 的核心策略是信息熵.熵是一种对信息量的度量,也是一种对不确定问题的度量.显然,对分类问题而言容易确定总比难以确定好.因此,ID3 和 C4.5 的基本策略是熵最小度量准则.即,对于任意一个训练集中的样本 e_i,它在第 j 个特征属性的概念或取值为 e_{ij}.对于 NP 问题,令 e_{ij} 属于 N 类的概率为 $P_{\mathrm{N}}(e_{ij})$,属于 P 类的概率为 $P_{\mathrm{P}}(e_{ij})$,则对于 e_{ij} 的不确定性计算可以用熵来度量.有

$$I(e_{ij})=-\{P_{\mathrm{N}}(e_{ij})\log_2[P_{\mathrm{N}}(e_{ij})]+P_{\mathrm{P}}(e_{ij})\log_2[P_{\mathrm{P}}(e_{ij})]\}. \tag{1}$$

对于 ID3,第 j 维或第 j 个属性的总熵值的计算方法为

$$I_j=\sum_{All\ i}P(e_{ij})I(e_{ij}), \tag{2}$$

其中 $P(e_{ij})$ 为 e_{ij} 出现的概率.而对于 C4.5,第 j 维或第 j 个属性的总熵值的计算方法为

$$I_j=\sum_{All\ i}P(D_{ij})I(D_{ij}), \tag{3}$$

其中 D_{ij} 为 e_{ij} 的取值区间,$P(D_{ij})$ 实际上就是 $P(e_{ij})$ 的置换.熵最小选择策略就是在所有的属性熵中选择其中 I_j 最小值对应的那一个属性进行样本分离.对熵最小策略的改进之一是熵增益最大策略.即令样本集合的先验熵为 E,则熵增益的计算方法为

$$Gain_j=E-I_j. \tag{4}$$

那么,ID3 和 C4.5 的根本缺陷在哪里呢?样本的轴向平行分离方法是标准决策树算法的固有特点和缺陷,标准算法无法对此进行改进.在此问题上,OC1[8]和 GTO/SVM[9]的分离方法使用的是斜超平面分离方法.就标准决策树算法而言,我们无须对此作更多的分析.除此之外,ID3 和 C4.5 的根本缺陷在于它们没有或只是部分涉及到样本的分布问题.

3 熵分布和样本分布的复杂性

我们先考虑熵的分布问题.对于标准的 ID3 算法而言,其熵最小选择策略无论从数学表达上看,还是从计算方法上看都是均值计算.毫无疑问,熵是一种非常好的数学工具,ID3 的计算方法也非常简单.然而对于稍微复杂的样本集合,情况却并非尽如人意.虽然人们已经对此类问题讨论很多,我们在此还是需要给出一个更直观的解释.图 1 和图 2 分别是我们所记录下的 Heart Disease 数据的熵值分布和样本分布的情况.其中,图 2 的样本取值已经被归一化到[0,1]区间.可以看出,当熵值分布并不比样本分布更显得简单时,将概念归纳方法用于数值计算时效果很难奏效.同样,使用熵增益的计算方法也有类似的情况.例如,当 NP 两类的先验概率相等时,样本集合的先验熵为 1.此时,熵最小准则与熵增益最大准则没有本质区别.

我们对熵选择策略的另一种实验观察结果见表1.对于表1的数据我们无须作更多的解释,它所表达的结论是显然的.即,0熵值与预测精度之间没有必然的联系.由于C4.5的计算方法不是对样本点而是对样本点所占据的数值区间的计算,所以,C4.5的判决精度要高于ID3.在逻辑结论上,ID3和C4.5都可以很好地解释各自的计算理论,但它们之间对熵的计算关系却是似是而非的.超越这一问题进行解释,C4.5 的优势在于,它部分地考虑了样本分布问题.然而,C4.5 所考虑的区间既不是一种类内或类间距离,也不是某种概率距离.也就是说,C4.5 算法的不足之一是没有顾及样本分布的复杂性问题.

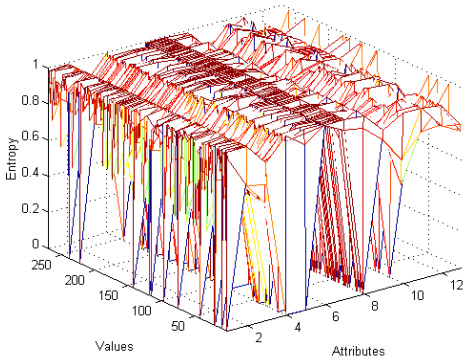

Fig.1 Entropy distributing of Heart Disease

图1 Heart Disease数据的熵值分布

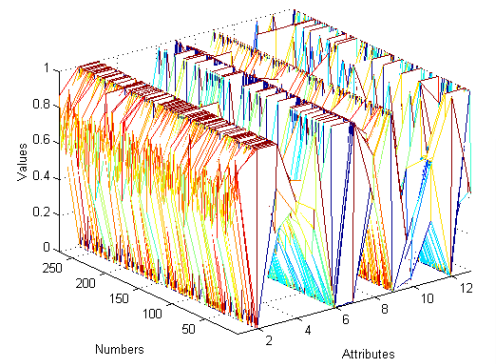

Fig.2 Example distributing of Heart Disease

图2 Heart Disease数据的样本分布

Table 1 Statistics of certainty or uncertainty offered by examples

表1 样本的不确定性或确定性统计

Datasets	Total examples	Total examples with zero entropy	Error rate of ID3 (%)	Error rate of C4.5 (%)
Heart Disease	270	209	27.78	22.96
Pima Indians	768	674	35.47	30.21
Breast Cancer	699	189	11.9	5.66
Votes-1984	435	0	6	5.29
Liver Disorder	345	156	42	38.9

对于样本分布的复杂性问题,从模式识别的角度看,它是特征提取环节传递过来的,从归纳学习的角度看,它是由数学方法的不充分引起的.在概率统计理论中,由于人们在理论分析和证明问题上的要求以及数学分析上的易行考虑,概率分布或样本分布的基本工具都是类似于 Gauss 函数的参数估计方法.如何利用这些有限的理论工具处理复杂的样本分布问题,在现有的技术条件下其可行的方法之一是将理论工具和概念分析结合起来.下面,我们对所形成的计算策略进行评价的讨论.

4 对所估计的参数进行评价的策略

模式识别的根本目标是希望分类的错误率尽可能小,而对错误率产生影响的因素取决于特征提取和特征选择的好坏.特征选择对应的是决策树归纳学习中的属性选择策略,特征提取决定了特征空间的样本分布.对于特征提取而言,所抽取的特征呈单峰分布总是比呈多峰分布要好.原则上,所采用的特征选择方法应当考虑特征提取问题.即,我们需要估计特征空间的样本分布.

Bias 隐含的前提是各维特征变量是相互独立的,即决策树算法是基于1维特征的计算.对于包含 M 个模式类别的特征空间$\Re^N$,记模式符号为ω.令 $x_{i,j,k}$ 表示第 i 类第 k 个样本在第 j 维的取值,S 为各模式类总的训练样本数目,S_i 为第 i 类总的样本数目,则第 ω_i 类的先验概率 $P_i = S_i/S$.第 i 类在第 j 维的均值 $\mu_{i,j}$ 可以作如下估计:

$$\mu_{i,j} = \frac{1}{S_i}\sum_{k=1}^{S_i} x_{i,j,k}. \tag{5}$$

设各维数据已经被归一化到[0,1]区间,则 $0 \le \mu_{i,j} \le 1$.将所有模式类在第 j 维的均值按从小到大的升序进行排列,并记为 $\mu_j = \{\mu_{1,j},...,\mu_{i,j},...,\mu_{M,j}\}$.理想情况下的特征分布应当为 $\mu_{1,j} = 0$, $\mu_{M,j} = 1$,并且任意两个相邻的均值之

间的距离 $\mu_{i+1,j}-\mu_{i,j}=1/(M-1)$.由于所提取的特征的均值并非完全理想分布,作为一种比较,我们需要定义一种计算方法来度量实际分布与理想分布之间的距离.令 $d_{\mu,j}$ 表示第 j 维的各模式类均值分布的统计量,则

$$d_{\mu,j}=\sum_{i=1}^{M-1}\left[(\mu_{i+1,j}-\mu_{i,j})-\frac{1}{M-1}\right]^2. \tag{6}$$

$d_{\mu,j}$ 越小,则表示各均值点在[0,1]区间内分布得越理想.特别地,对于 2 类问题可以简单地有

$$d_{\mu,j}=(\mu_{2,j}-\mu_{1,j})^{-2}, \tag{7}$$

即 $d_{\mu,j}$ 越小,则表示 $\mu_{1,j}$ 与 $\mu_{2,j}$ 之间的距离越接近 1.

另外,方差也是统计分析的重要特征量,它表示样本与均值之间的平均平方距离.考虑到各模式类的类内散射度的总的计算方法问题,我们使用样本与均值之间总的平方距离作为类内样本散射度的度量准则.令 $d_{\lambda,j}$ 表示第 j 维各模式类总的样本散射度,则

$$d_{\lambda,j}=\sum_{i=1}^{M}\left[P_i\sum_{k=1}^{S_i}(x_{i,j,k}-\mu_{i,j})^2\right]. \tag{8}$$

$d_{\lambda,j}$ 的值越小,则表示在总体上各模式类的样本在第 j 维越集中在各类的均值附近.即,第 j 维样本的总的散射度越小,则该维特征越理想.

根据式(6)和式(8),令

$$d_j=d_{\mu,j}+d_{\lambda,j}, \tag{9}$$

我们希望 d_j 越小越好.因此,我们在集合 $d=\{d_1,...,d_j,...,d_N\}$ 中选择其中最小值所对应的那一维特征作为形成决策树的 Bias.基于式(9)的决策树算法描述如下:

Step 1. 将当前窗口 W 中的数据 D 归一化到[0,1]区间中的 D'.

Step 2. 对 D'用公式(9)得到集合 $d=\{d_1,...,d_j,...,d_N\}$.

Step 3. 选择 $\min\{d_1,...,d_j,...,d_N\}$ 对应的那一维特征作为候选节点,对 D 进行分割.

Step 4. 对分割出的各个子窗口 W_s 进行判别.

① 若子窗口 W_s 下的样本为同一类样本,则该子节点为决策树的叶节点.

② 否则,在子窗口 W_s 重复 Step 1,直到满足①的条件.

5 熵选择策略 vs 样本分布选择策略

我们采用 UCI[10] 机器学习数据库中的 7 个常用数据 Heart Disease,Pima Indians,Votes-84,Liver Disorder,Breast Cancer(*W*),Iris Plants 和 Wine 进行对比性测试.实验中的测试方法为随机测试法(不是交叉验证 CV 法).我们随机取 90%的样本作为训练例,其他 10%的样本则作为测试数据.对每个数据集,我们都做 100 次随机取样,并以 100 次平均错误率作为评判的依据.表 2 中的$\mu\lambda$是指本文给出的决策树算法,CPU 耗时指的是形成决策树所消耗的平均计算时间.

Table 2 Performance testing of classification and waste time

表 2 分类性能和计算速度测试

Algorithm / Datasets	C4.5		$\mu\lambda$	
	Error rate (%)	Waste time of CPU (s)	Error rate (%)	Waste time of CPU (s)
Heart Disease	22.96	129.2	20.14	6.264
Pima Indians	30.21	3 081	30	39.54
Votes-1984	5.29	210.1	4.89	26.257
Liver Disorder	38.9	47.3	35	4
Breast Cancer	5.66	297.1	5.36	55.06
Iris Plants	5.9	7.00	4.2	0.681
Wine	28.5	31.15	16.83	0.972

我们需要对测试结果作一些必要的说明.在表 2 中,ID3 和 C4.5 的错误率统计结果是随机测试法得到的结果,而不是交叉验证法(CV).一般而言,100 次 90%随机取样测试相当于 10 次 CV10 测试.这两种方法的测试结果略有差异.例如,Iris 数据,10 次 CV10 测试的 C4.5 算法的错误率为 4.8%,而 100 次 90%随机取样测试结果为 5.9%.其他数据集的测试结果也有类似情况.选择同样测试方法,实验结果才有可比性.由于测试偏差肯定会存在,因

此测试次数越多,其结果就越恒定.

同时,我们还列出了平均计算耗时的统计结果,直观地显示了本文算法在算法的计算速度上具有很好的性能.其原因是显然的,因为在决策树的节点计算中,本文算法只需要乘法计算,而 ID3 和 C4.5 则需要对数运算.

6 相关问题讨论与展望

基于熵决策方法的 ID3 和 C4.5 算法是目前标准决策树算法中最著名的算法,它们有很多优点,也有很多不足.如果熵能够描述所有的决策和分类问题,归纳学习的问题则会简单得多.遗憾的是,近 10 年来持怀疑观点者和证据越来越多.盲目地认同和盲目地怀疑都不是严肃认真的态度.

我们结合概率统计分析理论中的几个基本概念,给出了一种基于参数估计和参数分布度量的学习策略.方法非常简单,实验结果也相当不错.这对于概率统计分析理论方面的研究者来说似乎太简单,但对于归纳学习却有进一步研究的意义.

持亚算法观点的研究者认为标准算法的策略太单一.诚如所述,我们通过对熵决策算法的分析也发现了这个问题.熵决策的核心思路是,选择容易确定的属性总比选择不易确定的属性好.我们在本文中给出的另一种思路是,选择样本呈单峰或接近单峰分布的属性总比呈复杂分布的属性要好.客观地说,这两种思路各有千秋.如何将这两种思路结合起来才是最重要的,也是我们将要继续研究的.一个开放性的论题是,究竟能有多少种可行的、能有效地针对样本空间某种度量方法的归纳学习方法?究竟它们能否统一在一个具体的学习系统中而不失去规则的表达形式?

归纳学习还存在难以突破的障碍.XYZ∨AB 逻辑数据是发现 Occam 算法的著名实验数据之一.我们采用遗传算法中的 Pittsburgh 学习方法,当规则限定在 *k*-term DNF 为两个正类子句时,预测的错误率总为 0.此时,结论与 Valiant 关于 PAC 学习理论完全吻合.然而,在实际问题中我们却不能确定 *k*-term DNF 究竟为多长.

再回到熵决策和参数距离决策这两个问题上.就这两种决策树所形成的策略而言,尽管参数距离决策在我们所列举的 7 个 UCI 数据的实验中比熵决策的错误率要低,但错误率的统计平均测试方法毕竟只反映了问题的某一方面.正如概率统计分析理论中的均值参数,它只是众多的总体衡量方法中的一种.我们在大量的实验观察中注意到,对于某一个特定的随机抽取的训练数据集,熵决策和参数距离决策得到的结论和效果并不相同.多次错误率统计平均测试标准掩盖了一些局部的差异,使我们容易忽略某些具有研究价值但在总体测试上又不是很好的技术细节.在实验观察上,熵选择策略与参数距离选择策略具有互补性,但如何将它们合理地进行融合,还有待于继续研究.

References:

[1] Blummer A, Ehrenfeucht A, Haussler D, Warmuth MK. Occam's Razor. Information Processing Letters, 1987,24:377~380.

[2] Murphy PM, Pazzani MJ. Exploring the decision forest. In: Proceedings of the Computational Learning and Natural Learning Workshop. Provincetown, MA, 1993. 10~12.

[3] Qualian JR. Induction of decision trees. Machine Learning, 1986,1:81~106.

[4] Qualian JR. C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

[5] Freund Y, Schapire RE. Experiments with a new boosting algorithm. In: Proceedings of the 13th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Pulishers, 1996. 148~156.

[6] Breiman L. Bagging predictors. Machine Learning, 1996,24:123~140.

[7] Qualian JR. Bagging, boosting, and C4.5. In: Proceedings of the 13th National Conference Artificial Intelligence. Portland, Ore., 1996. 725~730.

[8] Murthy S, Kasif S, Salzberg S. A system for induction of oblique decision trees. Journal of Artificial Intelligence Research, 1994,2: 1~32.

[9] Bennett KP, Blue J. Hybird extreme points Tabu search. R.P.I. Math Report, No.240, Troy, New York: Rensselaer Polytechnic Institute, 1996.

[10] University of California. Irvine repository of machine learning database, obtainable by anonymous FTP. ftp://ics.uci.edu in the /pub/machine-learning-databases directory.