文章编号: 1001-0920(2003)01-0096-03

采用重复剪辑近邻法提高决策树算法的性能

叶晨洲, 杨 杰, 姚莉秀, 陈念贻
(上海交通大学 图象处理及模式识别研究所, 上海 200030)

摘 要: 决策树算法易受训练样本集中噪声和混杂区域的影响。重复剪辑近邻法能消除样本集中符合某些先决条件的噪声, 清除混杂区域中后验概率较小的类别所包含的样本, 并在各类样本间形成符合 Bayes 分类准则的界线。用它对合适的训练样本集进行筛选, 可在不损害分类准确率的同时明显地减小决策树的规模, 有助于增强决策树的可理解性和可用性, 从而提高决策树的性能。

关键词: 数据挖掘; 决策树; 重复剪辑近邻法; 样本筛选
中图分类号: TP18 **文献标识码**: A

Improving performance of decision trees with multi-edit-nearest-neighbor algorithm

YE Chen-zhou, YANG Jie, YAO Li-xiu, CHEN Nian-yi
(Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai 200030, China)

Abstract: Noises and overlapped regions existing in training samples hurt the simplicity and generality of decision trees. To solve this problem, a sample selection algorithm based on multi-edit-nearest-neighbor rule is proposed. This algorithm, under ideal conditions, can eliminate the noise satisfying some prerequisites, purify the overlapped region according to its members' posterior probabilities, and finally form a Bayesian boundary between samples of different classes. When applied to an appropriate training dataset, it obviously cuts down the size of resulting decision trees without sacrificing the accuracy. This improves both the understandability and generality of decision trees.

Key words: Data mining; Decision tree; Multi-edit-nearest-neighbor algorithm; Sample selection

1 引 言

决策树[1~5]是数据挖掘领域中应用最为广泛的方法之一。将它作为工具, 可从某些生产过程积累的历史数据中, 发现控制因素与产出或故障现象与原因间的规律。决策树的生成过程采用从上至下、分而治之的策略。随着迭代深度的增加, 算法考虑的样本数不断减少。这样虽然能降低算法的时间复杂度, 但也使算法在较深层次的样本划分中, 专注于训练样本集某个子集的统计信息, 而忽视各类样本的整体分布情况, 造成了对噪声敏感、对交遇区(两类样本相互混杂的区域)处理不合理的缺点[6]。这些缺点使得生成的决策树在结构上非常复杂(节点数目庞大), 从而失去了可理解性和可用性, 即对训练集有较好分类效果, 而对新样本的分类能力则不佳。对单棵决策树进行删减[1], 或将不同训练方式获得的多棵决策树加以组合[7], 是解决这类问题的两种常用方法。但它们必须在决策树建立之后才能进行。

本文尝试使用重复剪辑近邻法, 在决策树生成

收稿日期: 2001-07-11; **修回日期**: 2001-09-03。
基金项目: 国家 863 高技术计划基金资助项目(863-511-945-005, 863-306-ZD 13-05-6)。
作者简介: 叶晨洲(1974—), 男, 上海人, 博士生, 从事故障诊断、数据挖掘的研究; 杨杰(1964—), 男, 上海人, 教授, 博士生导师, 从事计算智能、模式识别等研究。

前对其训练样本进行筛选, 从而弥补决策树的原有不足。虽然决策树可处理含有非数值型属性的样本, 但本文限定所有样本仅包含数值型属性(事实上, 非数值型属性可通过多值逻辑转换为数值型属性)。

2　剪辑近邻法和重复剪辑近邻法

2.1　剪辑近邻法

剪辑近邻法[2] 的决策过程由剪辑和分类两步组成。剪辑在先, 分类在后。

在剪辑步骤中, 按一定方法将包含 N 个样本的给定样本集 X^N 分成包含 NT 个样本的考试集 X^{NT} 和包含 NR 个样本的参考集 X^{NR}, 并且 $NT + NR = N$。

设 $x_j \in X^{NT}$, 且 $y'(x_j) \in X^{NR}$ 是它在参考集中的最近邻样本。利用参考集中的样本对考试集中的每个样本用近邻法[2,3] 进行分类, 剪辑掉 X^{NT} 中不与 $y'(x_j)$ 同类的样本 x_j, 然后将 X^{NT} 中剩余的样本构成剪辑样本集 X^{NTE}。分类时, 利用剪辑样本集 X^{NTE} 和最近邻规则对新样本 x 进行分类决策。

可以证明[2], 当 $N \to +\infty$ 时, 剪辑近邻法的错误率 P^E 总是小于等于未剪辑近邻法的错误率 P。当 P 很小时, $P^E \approx P^*$, P^* 为 Bayes 错误概率。

2.2　重复剪辑近邻法

既然剪辑过程可以降低近邻法的错误概率, 那么只要样本数足够多, 便可重复执行剪辑程序, 以进一步提高近邻法的性能。执行步骤如下[2]:

1) 从当前样本集合中随机取出一定比例的样本作为参考集, 将当前集合作为考试集。

2) 执行剪辑过程: 当本次剪辑完成后, 若有样本被剔除, 则令 $I = 0$, 转 1); 若没有样本被剔除且 I 小于设定值 MAX-I, 则令 $I = I + 1$, 转 1); 若没有样本被剔除且 I 等于设定值 MAX-I, 则剪辑过程结束。

可以证明[2], 用经过 M 次剪辑的样本集对原分布样本用最近邻规则进行分类, 其渐近条件误识率为

$$p_M(e|x) = 1 - \frac{p(\omega_1|x)^{2^M+1} + p(\omega_2|x)^{2^M+1}}{p(\omega_1|x)^{2^M} + p(\omega_2|x)^{2^M}} \tag{1}$$

当 $M \to +\infty$ 时, 则

$$\lim_{M\to+\infty} p_M(e|x) = \min[p(\omega_1|x), p(\omega_2|x)] = p^*(e|x) \tag{2}$$

其中 $p^*(e|x)$ 为 Bayes 条件误识率。因此剪辑近邻法渐近地具有 Bayes 最优性质。

3　用重复剪辑近邻法筛选决策树训练样本

3.1　重复剪辑近邻法对噪声的影响

定义样本 x 被标为 ω 且确实属于 ω 的概率为 $p(t_i|x)$, 而 x 被标为 ω 但实际上并不属于 ω(x 为噪声) 的概率为 $p(m_i|x)$。则有如下结论:

结论 1　若 $p(\omega|x) > p(\omega_j|x)$ 且

$$\frac{p(m_i|x)}{p(\omega|x)} < \frac{p(m_j|x)}{p(\omega_j|x)}$$

$$i, j \in \{1, 2\}, \quad i \neq j$$

则重复剪辑近邻法可消除 x 处噪声的影响。

证明　k 次剪辑后, 样本 x 被认为是 ω($l = 1, 2$) 类的概率由两部分组成, 即

$$p_k(\omega|x) = p_k(t_l|x) + p_k(m_l|x) \tag{3}$$

显然

$$\sum_{l=1}^{2} [p_k(t_l|x) + p_k(m_l|x)] = 1 \tag{4}$$

对所有类别而言, 此时 x 为噪声的概率为

$$p_k(n|x) = \frac{p_k(m_1|x) + p_k(m_2|x)}{[p_k(t_1|x) + p_k(m_1|x)] + [p_k(t_2|x) + p_k(m_2|x)]} \tag{5}$$

则 $k+1$ 次剪辑后, 样本 x 为噪声的概率为

$$p_{k+1}(n|x) = \frac{p_k(m_1|x)[p_k(t_1|x) + p_k(m_1|x)] + \rightarrow}{[p_k(t_1|x) + p_k(m_1|x)]^2 + \leftarrow \frac{p_k(m_2|x)[p_k(t_2|x) + p_k(m_2|x)]}{[p_k(t_2|x) + p_k(m_2|x)]^2}} \tag{6}$$

欲使

$$p_{k+1}(n|x) < p_k(n|x) \tag{7}$$

成立, 须有

$$[p_k(\omega_1|x) - p_k(\omega_2|x)] \times [p_k(\omega_1|x)p_k(m_2|x) - p_k(\omega_2|x)p_k(m_1|x)] > 0 \tag{8}$$

成立。由 k 的任意性以及每次剪辑对相关概率密度函数的影响知: 如果原始样本集中 $p(\omega_1|x) > p(\omega_2|x)$ 且

$$\frac{p(m_2|x)}{p(\omega_2|x)} > \frac{p(m_1|x)}{p(\omega_1|x)}$$

或 $p(\omega_1|x) < p(\omega_2|x)$ 且

$$\frac{p(m_1|x)}{p(\omega_1|x)} > \frac{p(m_2|x)}{p(\omega_2|x)}$$

则式(8) 成立, 从而式(7) 成立。设 q($q \geq 1$) 次剪辑后, x 为噪声的概率为

$$p_q(n|x) =$$

$$\frac{\{p(m_1|x)[p(t_1|x)+p(m_1|x)]\}^{2^{q-1}}+}{[p(t_1|x)+p(m_1|x)]^{2^q}+}\rightarrow$$
$$\leftarrow\frac{\{p(m_2|x)[p(t_2|x)+p(m_2|x)]\}^{2^{q-1}}}{[p(t_2|x)+p(m_2|x)]^{2^q}} \quad (9)$$

(各概率密度函数的下标 0 略去) 因为式(8) 成立, 所以若 $p(\omega_1|x) > p(\omega_2|x)$, 则 $p(m_1|x) < 1$; 若 $p(\omega_2|x) > p(\omega_1|x)$, 则 $p(m_2|x) < 1$。在此条件下不难证明, 当 $q \rightarrow +\infty$ 时, $p_q(n|x) \rightarrow 0$。□

3.2 重复剪辑近邻法对交迭部分的影响

特征空间中由非噪声的两类样本混杂形成的区域称为交遇区。对此有如下结论:

结论 2 对交遇区中的样本 x, 若 $p(\omega_i|x) > p(\omega_j|x)$, $i, j \in \{1, 2\}$, $i \neq j$, 则重复剪辑近邻法可清除该处 ω_j 类样本。

3.3 采用重复剪辑近邻法筛选决策树训练样本

为了减少训练样本集中噪声和交遇区对决策树的不利影响, 本文采用重复剪辑近邻法对原始训练样本进行处理。将算法结束时尚未被剪辑的样本组成新的训练集, 对决策树生成算法进行训练。根据结论 1 和结论 2, 适于采用重复剪辑近邻法处理的样本集应具备以下条件:

1) 具有合理的样本数;

2) 包含两类样本, 它们的组成和分布与实际基本相符;

3) 样本的类别标注基本正确。

4 实验结果

本文通过实验比较了采用重复剪辑近邻法在训练样本剪辑前后, 对 ID3[4], C4.5[1,4], OC1[5] 及树状分段线性分类器[2] 生成的决策树(参考集) 的规模(非叶子节点数) 和准确率。所有决策树都未采用删减步骤。实验 1 和实验 2 中使用的重复剪辑近邻法按每 4 个随机抽一的方式, 从当前样本集中抽取约占总数 1/4 的样本组成参考集; 实验 3 考虑到样本数较少, 每次随机抽取 95% 的样本组成参考集。采用欧氏距离度量样本间的差别。当测试样本与多个参考样本的距离都为最小值时, 认为它与次序在先的参考样本同类。为便于比较, 同时列出了剪辑前后近邻法使用的参考集大小及分类正确率。

实验 1 数据 1 的维数为 2。训练集和测试集生成规则如下: ω_1 类样本和 ω_2 类样本各自均匀分布在两个面积相等且部分重叠的矩形区域 R_1 和 R_2 内, 其左下与右上坐标分别为((2, 2), (21, 21)) 和((2, 16), (21, 34)), $p(x|\omega_1) \times P(\omega_1) < p(x|\omega_2) \times P(\omega_2)$。训练集含 1 类样本 124 个, 2 类样本 268 个, 各自含有 1% 的噪声。测试集含 1 类样本 107 个, 2 类样本 280 个, 未加入噪声。

比较结果列于表 1, 其中 OC1 的实验结果是运行 10 次的平均值, 括号内为方差。

表 1 数据 1 的实验结果

	剪辑前		剪辑后	
	决策树规模	准确率/%	决策树规模	准确率/%
ID3	40	87.1	1	90.4
C4.5	50	87.9	1	90.4
OC1	40.6(1.4)	84.7(0.4)	1.0(0.0)	89.3(0.6)
分类器	2	82.2	1	89.4
近邻法	392	81.4	298	89.9

实验 2 数据 2 的维数为 9, 取自 UCI repository of machine learning databases。共含有 683 个有效样本, 其中 1 类(良性癌) 样本 444 个, 2 类(恶性癌) 样本 239 个。采用 10-fold Cross-validation 进行测试。比较结果列于表 2, 其中括号外为运行 10 次的平均值, 括号内为方差。

表 2 数据 2 的实验结果

	剪辑前		剪辑后	
	决策树规模	准确率/%	决策树规模	准确率/%
ID3	26.1(1.1)	94.4(1.9)	6.2(1.0)	95.2(1.3)
C4.5	30.2(2.18)	94.9(1.6)	6.7(0.9)	95.5(1.7)
OC1	11.5(2.1)	94.0(1.4)	4.9(1.6)	95.2(1.6)
分类器	15.0(1.6)	94.6(0.5)	6.0(0.0)	94.8(0.7)
近邻法	615.0(0.0)	96.2(1.3)	579.2(3.0)	96.9(1.5)

实验 3 数据 3 的维数为 16, 取自某钢铁厂, 用于分析炼钢过程各项参数与钢铁含碳量间的关系。样本总数为 204, 其中 1 类(含碳量低) 样本 106 个, 2 类(含碳量高) 样本 98 个, 样本数较少且含噪声。采用 10-fold Cross-validation 进行测试。比较结果列于表 3, 其中括号外为运行 10 次的平均值, 括号内为方差。

表 3 数据 3 的实验结果

	剪辑前		剪辑后	
	决策树规模	准确率/%	决策树规模	准确率/%
ID3	24.0(1.9)	71.3(3.2)	22.2(2.7)	72.3(5.0)
C4.5	28.4(3.4)	69.2(8.2)	25.0(1.1)	73.3 (5.0)
OC1	11.4(1.4)	67.3(6.7)	9.8(1.7)	67.4(8.7)
分类器	53.0(12.6)	58.38(14.8)	40.4(3.9)	58.38(12.2)
近邻法	183.8(0.4)	84.1(8.0)	173.0(6.4)	85.1(9.5)

(下转第 102 页)

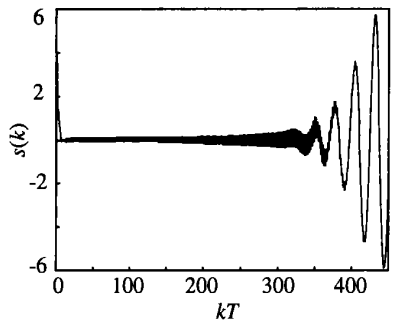

图 3 切换函数 s(k) 的运动轨线

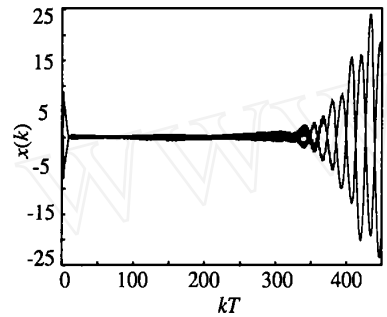

图 4 系统状态 $x_1(k)$ 和 $x_2(k)$ 的运动轨线

制律设计方案。由于该方法的步长修正项与上一采样时刻穿越 $s(k) = 0$ 的幅度成反向关系, 可有效地抑制预估误差大时引起的发散振荡, 在不确定性比文献[6]方法大得多的情况下, 仍能保证准滑动模态存在和系统稳定, 仿真结果也证实了这一点。另外, 在不确定性大时, 可用加大 α 来加以克服。

参考文献(References):

[1] Dot Y, Holf R G. Microprocessor based sliding mode controller for dc motor drives[A]. *Pres Ind Appl Society Annual Meeting*[C]. Cincinnati, 1980.

[2] Furuta K. Sliding mode control of a discrete systems [J]. *Syst Contr Lett*, 1990, 14(2): 145-152.

[3] Sapturk S Z, Istefanopulos Y, Kaynak O. On the stability of discrete-time sliding mode control systems[J]. *IEEE Trans Autom Contr*, 1987, 32(10): 930-932.

[4] 高为炳. 变结构控制的理论及设计方法[M]. 北京: 科学出版社, 1998.

[5] 高为炳. 离散时间系统的变结构控制[J]. 自动化学报, 1995, 21(2): 154-161.
(Gao Weibing. Discrete-time system variable structure control[J]. *Acta Autom Sinica*, 1995, 21(2): 154-161.)

[6] 于双和, 傅佩琛, 强文义. 不确定离散时间系统的变结构控制[J]. 控制理论与应用, 2000, 17(1): 85-87.
(Yu Shuanghe, Fu Peichen, Qiang Wenyi. Variable structure control of uncertain discrete-time system [J]. *Control Theory Appl*, 2000, 17(1): 85-87.)

(上接第 98 页)

实验结果表明, 采用重复剪辑近邻法对原始训练样本进行处理后, 能有效减少训练集中的样本数量, 3 组实验中的减少幅度分别为 24.0%, 5.8%和 5.9%; 而不同算法获得的决策树规模也较处理前明显减小。所有决策树在规模上的减小幅度超过了训练样本数的减少幅度。各组实验中所有算法的分类准确率(或平均准确率)较处理前也有不同程度的提高, 4 种决策树算法的平均提高幅度分别为 2.0%, 3.1%, 2.3%和 3.0%。

5 结 语

噪声和交遇区容易对决策树生成算法产生不利影响。重复剪辑近邻法能消除样本集中符合某些条件的噪声, 清除交遇区中后验概率较小的类别所包含的样本, 并在两类样本间形成符合 Bayes 分类准则的界线。采用它进行样本筛选, 可在决策树建立之前, 依靠全部训练样本减少甚至消除噪声和交遇区的不利影响, 从而在不损害分类准确率的同时明显减小决策树的规模。

参考文献(References):

[1] Chidanand Apte, Sholom Weiss. Data mining with decision trees and decision rules[J]. *Future Gener Comp Syst*, 1997, 13(2): 197-210.

[2] 边肇祺, 张学工. 模式识别(第 2 版)[M]. 北京: 清华大学出版社, 1999. 122-124, 145-152.

[3] Andrew Webb. *Statistical Pattern Recognition*[M]. New York: Oxford University Press Inc, 1999.

[4] Quinlan J R. Induction of decision trees[J]. *Machine Learning*, 1986, 1(1): 81-106.

[5] Sreeramma K Murthy, Simon Kasif, Steven Salzberg. A system for induction of oblique decision trees[J]. *J Artif Intell Res*, 1994, 2(8): 1-32.

[6] Krishnan R, Sivakumar G, Bhattacharya P. Extracting decision trees from trained neural networks[J]. *Pattern Recog*, 1999, 32(12): 1999-2009.

[7] Thomas G Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization [J]. *Machine Learning*, 2000, 40(2): 139-157.