基于异构代价敏感决策树的分类器算法

阮晓宏[1,2] 黄小猛[1] 袁鼎荣[1] 段巧灵[1]

(广西师范大学计算机科学与技术学院 桂林 541004)[1] (来宾市职业技术学校 来宾 546100)[2]

摘 要 代价敏感学习方法常常假设不同类型的代价能够被转换成统一单位的同种代价,显然构建适当的代价敏感属性选择因子是个挑战。设计了一种新的异构代价敏感决策树分类器算法,该算法充分考虑了不同代价在分裂属性选择中的作用,构建了一种基于异构代价的分裂属性选择模型,设计了基于代价敏感的剪枝标准。实验结果表明,该方法处理代价机制和属性信息的异质性比现有方法更有效。

关键词 决策树分类,代价敏感学习,异构代价敏感

中图法分类号 TP18 **文献标识码** A

Classification Algorithm Based on Heterogeneous Cost-sensitive Decision Tree

RUAN Xiao-hong[1,2] HUANG Xiao-meng[1] YUAN Ding-rong[1] DUAN Qiao-ling[1]

(College of Computer Science and Information Technology, Guangxi Normal University, Guilin 541004, China)[1]

(School of Vocational and Technical, Laibin 546100, China)[2]

Abstract Usually, cost-sensitive learning assumes that different types of cost can be converted into a unified units of the same price. Apparently how to construct appropriate cost-sensitive attribute selection factor is a challenge. In this paper, a kind of heterogeneous cost-sensitive decision tree algorithm was designed, which fully considers the different cost in selecting split attribute, constructs an attribute selection model based on heterogeneous cost-sensitive, designs the price sensitive pruning strategy based on cost-sensitive. The experimental results show that this method is effective and more efficient than the present other methods.

Keywords Decision-tree classification, Cost-sensitive learning, Heterogeneous cost-sensitive

1 引言

决策树是机器学习和数据挖掘中的重要研究课题,在实际应用中获得广泛而成功的应用,如ID3算法、CART算法和C4.5算法[1]。早期的决策树分类算法旨在提高分类的精确性,最大限度地减少误分类,这样必然导致分类器内在地偏向于主要的类,而忽略精度影响较小但分类结果影响重大的少数类,即忽略错误分类所带来的代价,例如一只羊被错误地分入一群狼中的代价只是损失一只羊,而一只狼被错误地分入一群羊中的代价则是损失一群羊,因此Elkan[4]等提出代价敏感分类学习策略。随后,代价敏感决策树学习算法引起了广泛的研究兴趣并取得丰硕的研究成果[7-12]。当前,代价敏感决策树学习集中于不同类型代价中的决策模型研究,比如误分类代价、测试代价、计算代价、教育代价等,但是实际应用中代价是综合的,比如医疗诊断中,不仅仅存在误分类代价,同时还存在检测、会诊等其它方面的代价,因此多源异构代价敏感决策成为决策树学习中新的挑战。

分裂属性的选择是决策树构建的一个关键但基本的过程,最流行的属性选择方法侧重于测量属性的信息增益(或者基尼系数),但是具有多值的属性更容易取得较大值的信息增益,因此(Mitchell,1997年)[1]在C4.5算法的分列属性选择中用平均信息增益率替换信息增益。当错误分类所引起的代价不容忽视时,很自然地把降低代价机制和属性信息结合起来作为分裂属性选择的标准,属性选择的目标是最小化误分类总代价。最小化误分类总代价被称为基于CAI的分裂属性选择,但是现有的基于CAI分裂属性选择方法基于单一代价机制,对于异构多种代价中的分裂属性选择显得无能为力。针对异构代价的应用需要以及单一代价敏感机制的技术局限性,张(2010年,2012年)[2,3]提出基于HCAI的异构代价敏感决策中的分裂属性选择方法,该方法改进了基于CAI的分裂属性选择策略,使其能够联合不同的代价和属性信息,成为分裂属性选择标准。文章认为基于HCAI分裂属性选择的异构代价能够清楚地放大/缩小异构代价对分裂属性选择的影响,同时异构属性还能够放大/缩小平均信息增益率对属性的影响,但是代价和属性之间的平衡性未得到很好的解决,从而影响分类精度和整体代价。张在文献[3]中避免了文献[2]中测试代价被放大的可能,但分裂属性信息依然存在因为过小而被忽略的风险以及误分类代价过大而被放大的可能。针对异构代价平衡性问题,提出一种解决方法,即首先标准化各种不同代价,联合各种代价和属性信息作为分裂属性选择标准,

本文受国家自然科学基金项目(61170131),广西创新团队项(GXNSFGA060004),广西师范大学项目资助。

阮晓宏 男,硕士生,讲师,主要研究方向为机器学习;**黄小猛** 男,硕士生,主要研究方向为数据挖掘;**袁鼎荣** 男,博士,教授,主要研究方向为机器学习、数据挖掘。

然后在剪枝过程中依然利用代价大小决定剪枝标准，最后通过实验验证我们的方法。

2 代价敏感决策学习基础

代价敏感学习(CSL)是传统决策树学习的扩展，它是一个寻找最小代价过程的分类器。Elkan(2001年)定义CSL如下[4]。

定义1(代价矩阵) 设训练数据集 T 拥有 n 个属性 a_1，a_2，…，a_n，分裂属性 $a_i \in \{a_1, a_2, \cdots a_n\}$ 拥有 m 个不同的类标识 $l_1, l_2, \cdots, l_m$，c_{ij} 表示第 j 类数据分为第 i 类的代价，如果 $i=j$ 为正确分类，则 $c_{ij}=0$，否则为错误分类 $c_{ij} \neq 0$，其值由用户给定，我们称矩阵 $C=(c_{ij})$ 为代价矩阵。

定义2(预测代价) 对于任意一个样例 x，如果将其分为第 i 个类，那么可能的总代价为：

$$L(x,i)=\sum_j p(j|x)C(i,j)$$

我们称 $L(x,i)$ 为将 X 分为第 i 类的预测总代价，简称预测代价。如果确定将 X 分为第 i 类，则 $L(x,i)$ 称为预测代价。

定义3(误分类代价敏感决策) 决策树学习过程中，由属性信息特征和预测代价 $L(x,i)$ 共同决定分裂属性，我们称这样的决策为误分类代价敏感决策。

定义4(异构代价敏感决策) 决策过程中除了错误分类所带来的代价以外，还存在测试成本、教育成本和介入成本等其它方面的代价，分裂属性选择过程中，同时考虑所有不同类型的代价，我们称这样的决策为异构代价敏感决策。

分裂属性选择是决策树递归中一个基本的过程。在这一过程中，随着分裂属性的确立，属性信息特征和代价相应变化，下面给出属性信息特征和代价变化的相关概念。

定义5(信息增益) 文献[3]依据基尼指数确定分裂属性，基尼指数表示为 $G(T)$，定义为：

$$G(T)=1-\sum_j (l_j)^2 \tag{1}$$

式中，$p(l_j)$ 是相关样例属于 l_j 类的概率。

在基尼指数的基础上，属性 A 作为分裂属性的信息增益可表示如下：

$$Gain(A)=G(T)-G(A,T) \tag{2}$$

式中，$G(A,T)$ 是当属性 A 作为分裂属性分裂后在所有类中剩余的基尼系数。

定义6(误分类代价减损) 在分类代价基础上，属性 A 作为分裂属性的代价减损可表示如下：

$$Reduce(A)=\frac{Mc-\sum_{i\in ClassSet(A)} Mc(A_i)}{Mc} \tag{3}$$

式中，Mc 表示分裂属性分裂前的代价和，$\sum_{i\in ClassSet(A)} Mc(A_i)$ 表示按分裂属性 A 分裂后的所有子类代价总和，其中 $ClassSet(A)$ 是分裂属性 A 分裂后的所有类的集合。

定义7(测试代价) 分裂属性 A_i 的测试成本表示为 $TC(A_i)$，标准化为：

$$TC(A_i)_{nomal}=\frac{\mathrm{Max}(1, TC(A_i))}{\mathrm{Max}(1, TC(A_1), TC(A_2), \cdots, TC(A_n))} \tag{4}$$

3 异构代价敏感决策树构造方法

异构代价敏感决策树分类器类似传统决策树分类器，其步骤如下：(1)设计异构代价敏感的分裂属性选择策略；(2)设计异构代价敏感的决策树构建方法；(3)设计剪枝策略；(4)测试所构建的异构代价敏感决策树性能。伪代码描述如表1所列。

表1 代价敏感决策树分类器算法伪代码

训练阶段：
规范每个属性并判断叶子节点是否继续可分
如可分则：(1)进行分裂属性选择
(2)基于选择分裂属性构建异构代价敏感决策树
(3)减枝
测试阶段：测试给定的样例

3.1 异构代价敏感的分裂属性选择

我们设计异构代价敏感的分裂属性选择因子 ASF(Attribute Selection Factor)如下：

$$ASF(A_i)=\frac{(2^{Averagegain(A_i)}-1)}{TC(A_i)_{normal}} * Reduce(A_i) \tag{5}$$

式中，$Averagegain(A_i)$ 表示平均信息增益，$TC(A_i)_{normal}$ 表示标准化测试成本，$Reduce(A)$ 表示误分类代价减损。

我们提出的属性选择因子综合考虑属性信息、测试代价和误分类代价对属性选择的影响。使用平均增益取代传统的信息增益进行分裂属性选择，旨在加强信息增益的分类能力。标准化测试代价指在避免了因分母过大而忽略属性信息对分裂属性选择因子的影响。误分类代价减损指将通过分裂属性选择所引起的误分类代价降低，加强误分类代价变化对属性选择因子的影响，因此我们约定：设数据集的属性标签为 A_1，A_2，…，A_n，将属性选择因子取最大值的属性作为分裂属性。

3.2 构建异构代价敏感决策树

首先，根据式(5)计算各候选分裂属性的属性选择因子，然后选取最大的选择因子所对应的属性作为分裂属性，如果遇到分裂因子具有相同的值，则按照应用需要依次考虑误分类减损大小，测试代价大小。建树过程中如果满足以下两个条件之一将被停止。

(1)在一个节点中的所有样例属于相同类；

(2)所有的属性都耗尽。

异构代价敏感决策树构建算法伪代码描述如表2所列。

表2 异构代价敏感决策树构造算法

1. 创建根节点 N
2. 如果训练集为空，则返回节点 N 标记为 Failure；
3. 如果训练集中的所有记录都属于同一个类别，则以该类别标记节点 N；
4. 如果候选属性为空，则返回 N 作为叶节点，标记为训练集中最普通的类；
5. for each 候选属性 attribute_list
6. 依据式(6)从候选属性 attribute_list 中选择分裂属性 SplitA
7. 标记节点 N 为属性 SplitA；
8. for each 属性 SplitA 的已知值 $SplitA_i$
9. 由节点 N 长出一个条件为 $SplitA=SplitA_i$ 的分支；
10. 设 si 是训练集中 $SplitA=SplitA_i$ 的训练样本的集合；
11. if si 为空
12. 加上一个树叶，标记为训练集中最普通的类；
13. else 递归调用本算法

第5步是本算法的核心，也是异构代价敏感决策树与传统的决策树构造算法的区别所在，其它步骤类似于IC4.5。

3.3 剪枝策略

决策树剪枝的目的是防止决策树学习过度拟合，当前主要策略包括：

(1)减少误分类，如悲观性错误剪枝和最小错误剪枝。悲观性错误剪枝通过比较剪枝前和剪枝后的错分样本数来判断

是否剪枝，旨在减少错分样本数。最小错误剪枝即指通过剪枝得到一棵相对于独立数据集来说具有最小期望错误率的决策树。

(2)减少计算代价，如代价复杂性剪枝和最少描述长度剪枝。代价复杂性剪枝在剪枝过程中因子树 Tt 被叶节点替代而增加的错误分类样本总数称为代价，剪枝后子树 Tt 减少的叶节点数称为复杂性。最少描述长度剪枝根据决策树的编码代价大小进行剪枝，目标是使得训练样本的大多数数据符合这棵树，把样本中不符合的数据作为例外编码，使得编码决策树所需的比特最小和编码例外实例所需的比特最小。

我们的策略优先考虑误分类减损，然后考虑测试代价，即对于用户给定的正数 α,β 满足：

(1) $Reduce(A)<\alpha$；

(2) $TC(A_i)>\beta$。

剪枝的条件首先要满足尽可能使代价减损达到用户指定条件，然后满足测试代价降低到用户要求。

4 相关研究

自 Quinlan(1986 年)提出 ID3 算法以来，决策树学习引起研究者极大的兴趣，并获得广泛的应用。由于 ID3 固有的局限性，Quinlan(1997 年)在 ID3 算法基础上提出 C4.5 算法，该算法采用信息增益率进行分裂属性选择，并扩展 ID3 能处理连续值的属性和包含缺失值的训练样本，但 ID3 和 C4.5 依然不能解决错误分类所引起的代价问题，因此 Elkan(2001 年)提出代价敏感决策树学习[4]，Elkan 只考虑误分类代价敏感问题，忽略了测试、计算等不同代价敏感问题，因此近年来异构代价敏感成为研究的热点。张等(2010，2012 年)对异构代价敏感学习做了卓有成效的工作[3,4]，张(2010 年)首先构造了分裂属性选择标准，见式(6)：

$$SAS_{CTS}(A)=(2^{AverageGain(A,T)}-1)\frac{ReduceMC(A)}{TC(A)+1} \quad (6)$$

张在随后的研究中发现该式因为分裂属性信息过小而可能被忽略，测试代价存在被无限放大的可能，因此将其标准化后变成如下形式：

$$SAS_{CTS}(A)=(\frac{2^{AverageGain(A,T)}-1}{TC(A)+1})_{normal}ReduceMC(A) \quad (7)$$

式(7)避免了测试代价被放大的可能，但分裂属性信息依然存在因过小而被忽略的风险和误分类代价过大而被放大的可能。因此我们提出一种新的方法来防止属性信息被忽略的风险和误分类代价被放大的可能，并对相关概念和定义进行了详细的讨论。

5 实验研究

为了评估所构建的异构代价敏感决策树(记为“HCSDT”)的性能，我们选择了最流行的方法进行比较，如“SAS”(张等(2010，2012 年)对异构代价敏感学习)、“GR”(分裂属性选择，不考虑代价)、“MTC”(考虑了误分类代价和测试代价带有不同的单位，但没考虑分裂属性的分类准确度)。GR 类似 C4.5 算法，但是没有代价敏感的功能。在 UCI 数据集上，我们使用所有的测试方法把无限的测试代价和有限的测试代价用来评估派生最小的误分类代价。

5.1 不同缺失值情况下的实验

从图 1 来看，所有方法误分类代价随着缺失值的增大而增大。由于数据集中引入了更多的缺失值，误分类的错误将会增加。即缺失数据率越高，建造的模型效果就越差。比较 HCSDT 方法和其他方法，在不同的缺失率下，HCSDT 具有最低的误分类代价。比较 HCSDT 算法和 SAS 算法，我们可以容易地得出结论，HCSDT 算法在代价敏感学习中设置不同单位的代价进行异构学习是可行的、合理的。

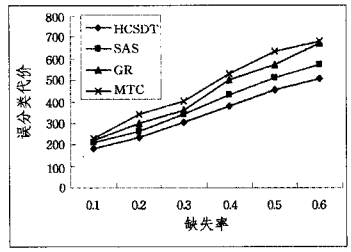

图 1 缺失值与误分类代价关系

5.2 不同测试代价下的实验

实验的测试代价的域是在 100 到 600 之间，在 y 轴表示误分类代价。研究结果也表明 HCSDT 算法在有限的测试代价方面比其他的算法更好。从图 2 可知：HCSDT 算法在最大的增益率和各种资源上最小的总代价都是最好的。此外 GR 算法(传统的方法)得到的是最坏的结果。

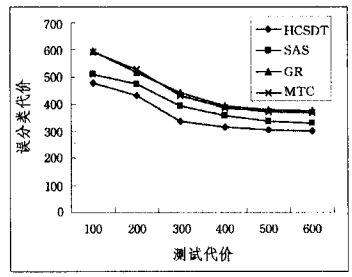

图 2 测试代价与误分类代价关系

5.3 相关分类器算法精度比较的实验

实验的分类精度在表 3 中很容易看出，GR 算法在代价敏感决策树分类区中的分类精度是最低的，同时它造成的误分类代价无疑也是最大的，HCSDT 的分类精度是其中最好的，SAS 稍微差一点点。

表 3 不同算法分类精度与误分类表

算法名称	分类精度	误分类率
GR	80.07±2.78	19.93±2.78
MTC	83.23±1.36	16.77±1.36
SAS	89.67±1.82	10.33±1.82
HCSDT	93.89±1.91	6.11±1.91

结束语 本文提出了一种基于异构代价敏感的分裂属性选择策略，其在属性选择过程中充分考虑了属性信息特征、误分类代价和测试代价，避免了基于 HCAI 分裂属性选择中的因属性信息特征值过小而被忽略的风险和误分类代价过大而被放大的可能。我们进行了各种实验，以评估所提出的方法和现有的方法在 6 个 UCI 数据集上的性能。实验结果表明，该方法在处理异构代价的问题上比现有的方法更稳健有效。

参 考 文 献

[1] Mitchell T M. Machine Learning[M]. McGraw Hill，1997

[2] Zhang S C. Cost-sensitive classification with respect to waiting cost[J]. Knowledge-Based Systems，2010，23：369-378

(下转第 146 页)

快，其中对于 f_1 和 f_2，HV-DPSO 在收敛速度上明显高于 PSO；对于 f_1，f_2，f_3 和 f_4，收敛次数和收敛速度 DPSO 和 HV-DPSO 有较大优势，随着测试函数的复杂度增加，DPSO 的优势逐渐下降，而 HV-DPSO 的优势基本保持不变；对于 f_3 和 f_4，杂交变异算子的引入，体现了在多维复杂函数方面，进一步提高了算法的性能。实验表明，在综合性能方面，改进后的 DPSO 算法和 HV-DPSO 算法都好于 PSO 算法，HV-DPSO 算法要好于 DPSO 算法。

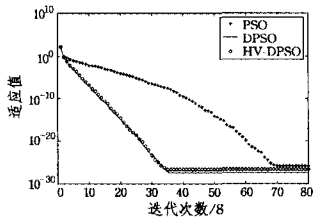

图 2　Sphere 函数收敛对比分析

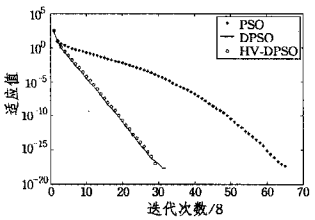

图 3　Griewan 函数收敛对比分析

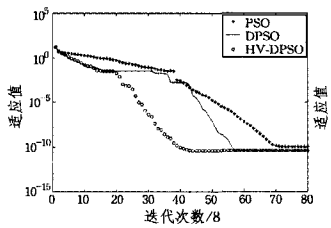

图 4　Ackley 函数收敛对比分析

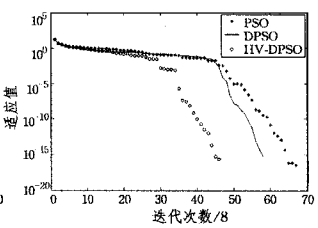

图 5　Rastrigin 函数收敛对比分析

结束语　粒子群算法中，较低的种群多样性会导致算法早熟收敛，陷入局部最优解。本文引入动态变化的惯性权重以及杂交变异规则来增加其多样性，控制种群多样性处于高效的范围内，通过设置比例系数，控制迭代次数与惯性权重的关系、其他粒子与最优粒子之间距离的关系。在保持种群多样性的基础上，实现算法全局与局部搜索的平衡。采用 4 个常用基准函数进行了试验，并与标准 PSO 做了比较，结果表明改进算法相对于标准 PSO 算法，克服了早熟收敛而且具有较好的收敛精度。

参 考 文 献

[1] Eberhart R C, Kennedy J. A new optimizer using particle swarm theory[C]//Proceedings of the Sixth International Symposium on Micro Machine and Human Science. Japan: Na-goya, 1995: 39-43

[2] Kennedy J, Eberhart R C. Particle swarm optimization[C]// Proceedings of the IEEE International Conference on Neural Networks. Piscataway: IEEE, 1995: 1942-1948

[3] 纪震，廖慧连，吴青华. 粒子群算法及应用[M]. 北京：科学出版社，2010

[4] 田雨波，朱人杰，薛权祥. 粒子群优化算法中惯性权重的研究进展[J]. 计算机工程与应用，2008，44(23)：39-41

[5] 唐忠. 粒子群算法惯性权重的研究[J]. 广西大学学报：自然科学版，2009，34(5)：640-644

[6] Shi Y, Eberhart R C. A modified particle swarm optimizer[C]// IEEE World Congress on Computational Intelligence. Piscataway: IEEE, 1998: 69-73

[7] 刘建华，樊晓平，瞿志华. 一种惯性权重动态调整的新型粒子群算法[J]. 计算机工程与应用，2007，43(7)：68-70

[8] 徐玉杰，仇雷，刘清. 自适应惯性权重的混沌粒子群算法研究[J]. 南京师范大学学报：工程技术版，2012，12(1)：64-69

[9] 王克华，牛慧，张亚南，等. 一种参数自适应调整和边界约束的粒子群算法[J]. 电子设计工程，2011，19(21)：46-49

[10] 盛桂敏，薛玉翠，张博阳. 动态自适应粒子群优化算法[J]. 绥化学院学报，2011，31(6)：190-192

[11] 张顶学，关治洪，刘新. 一种动态改变惯性权重的自适应粒子群算法[J]. 控制与决策，2008，23(11)：1253-1257

[12] 龙文，梁昔明，董淑华，等. 动态调整惯性权重的粒子群优化算法[J]. 计算机应用，2009，29(8)：2240-2242

[13] 祝洪博，徐刚刚，海冉冉，等. 基于云自适应梯度粒子群算法的无功优化[J]. 电网技术，2012，36(3)：162-167

(上接第 142 页)

[3] Zhang S C. Decision tree classifiers sensitive to heterogeneous costs[J]. The Journal of Systems and Software, 2012, 85: 771-779

[4] Elkan C. The foundations of cost-sensitive learning[C]//Proceeding of the Seventeenth International Joint Conference of Artificial Intelligence. Morgan Kaufmann, Seattle, August 2001: 973-978

[5] Nunez M. The use of background knowledge in decision tree induction[J]. Machine Learning, 1991, 6: 231-250

[6] Tan M, Schimmer J. Cost-sensitive concept learning of sensor use in approach and recognition[C]//Proceedings of the 6th International Workshop on Machine Learning. Ithaca, New York, 1989: 392-395

[7] Freitas A, Costa-Pereira A, Brazdil P. Cost - sensitive decision trees applied to medical data[C]//Proceedings of DaWak-2007, LNCS 4654. 2007: 303-312

[8] Davis J V, Ha J, Rossbach C J, et al. Cost-sensitive decision tree learning for forensic classification[C]//Proceedings of the 17th European Conference on Machine Learning (ECML). 2006: 622-629

[9] Zhang S C, Jin Z, Zhu X F. Missing Data Imputation by Utilizing Information within Incomplete Instances[J]. Journal of Systems & Software, 2011, 84: 452-459

[10] Wang W D, Miao S, Yang J Y. Classifier Algorithm on Orthogonal Projection on[J]. Computer Scienc, 2011, 38(5): 1-4

[11] Wu H R, Qin J, Zheng B J. Anti-attack Ability Based on Costs in Complex Network[J]. Computer Scienc, 2012, 39(8): 1-4

[12] Jan T K, Lin C H, Wang D W, et al. A Simple Methodology for Soft Cost-sensitive Classification KDD[Z]. Beijing, 2012, 8