代价敏感的GEP分类算法实现

张 赪，蔡之华

(中国地质大学计算机学院 武汉 430074)

【摘要】在数据挖掘领域中，通常以分类精度作为分类算法效果的评估标准。这一标准是建立在假设任意一实例被误分类为任意类时都具备同样代价的基础上的。当此假设不成立时，直接使用传统分类方法就无法取得良好的分类和预测效果。针对这一问题，通过改进编解码方法以及在适应度函数中集成样本的不同误分类代价，提出了一种基于基因表达式程序设计的代价敏感分类算法(CSC-GEP)，并在三个UCI数据集上对该算法进行了测试，实验结果表明CSC-GEP是一种有效的代价敏感分类算法。

关 键 词 分类； 代价敏感； 误分类代价； 基因表达式程序设计
中图分类号 TP18 文献标识码 A

Cost-Sensitive Classification by Gene Expression Programming

ZHANG Cheng，CAI Zhi-hua

(School of Computer Science, China University of Geosciences Wuhan 430074)

Abstract In data mining reseach, the classification algorithms generally pursue more highly accuracy. It is based on the assumption that all misclassifications have the same cost. However, the assumption is not correct in the real world, so that the normal classification algorithms do not perform well. By improving the encode/decode methods and taking different misclassification cost into account, this paper concerns a new cost-sensitive algorithm called CSC-GEP based on Gene Expression Programming (GEP). The experimental results show that the new algorithm is effective.

Key words classification; cost-sensitive; cost of misclassification; gene expression programming

在数据挖掘中，分类的任务通常是建立一个期望误分类数量最小的分类器，比较典型的例子如C4.5的决策树算法，在传统上都是基于分类错误具有相同的代价的假定上的。而在现实中，不同的分类错误通常会导致不同的代价。以UCI数据集中的Heart disease数据集为例，错误地将健康的人分类为患病的代价仅仅是使健康的人接受更多的检查，但如果将患病的人分类为健康，则有可能延误病人的治疗。很明显，第二种错误分类会令使用者付出远比第一种错误分类更大的代价。仍以Heart disease数据集为例，在这一数据集上建立分类器的目的是为了检测出患病的就诊者。对某一特定疾病而言，可能存在的情况是全部就诊者中仅有极少数是患病者，则在最糟糕的情况下，使用传统分类方法产生的分类器只需要牺牲全部患病者的实例作为误分类即可获得一个在传统意义上的高精度分类结果。但此时这个分类器也就丧失了存在的意义。因此在处理误分类代价不统一的数据集时，单纯地以分类精度作为分类器的评估标准就变得不恰当了。一种合理的解决方法是以代价敏感(Cost Sensitive Classification，CSC)的分类取代精度敏感的分类。

本文通过在GEP的适应度函数中加入代价矩阵，提出了基于基因表达式编程的代价敏感分类算法CSC-GEP。通过在UCI数据集中的Heart disease数据集、Sick数据集和Credit数据集上的分类算法与传统算法进行实验对比，结果证明CSC-GEP是一种有效的代价敏感分类算法。

1 代价敏感分类方法

由于在传统分类算法中通常将不同误分类的代价视作相等，因此这类基于分类精度的分类算法无法直接用于CSC问题中。在现有的众多分类算法中，有相当一部分致力于CSC问题的解决。但总体说来这些算法都遵循下面两种思路：(1) 重构训练样本

收稿时间：2007 – 09 – 07
基金项目：“ 十一五 ” 民用航天项目(C5220061318)；湖北省自然科学基金(2003ABA043)
作者简介：张 赪(1981 –)，男，硕士生，主要从事数据挖掘、数据仓库以及商业智能方面的研究；蔡之华(1964 –)，男，教授，博士生导师，主要从事数据挖掘与数据仓库、演化计算方面的研究.

集，再使用基于分类精度的传统分类算法；(2) 不改变训练样本集，建立专用的代价敏感分类算法。

第一种思路包括增加高误差代价样本数量和减少低误差代价样本数量，但这种方法无法准确控制各类误差的权重，同时还有可能造成过度学习。同时人为减少低误差样本也可能导致信息和规则的丢失[1-5]。

第二种思路以权值矩阵为主，将不同误分类代价赋以不同的权值并建立相应的权值矩阵，同时将权值矩阵加入到分类算法中以达到使分类算法以误分类代价为分类标准的目的[6-7]。

本文的研究属于第二种。通过对GEP的编码和解码方式的改进使其适应分类算法所处理的数据的特殊性，同时在适应度函数中加入集成样本不同误分类代价的代价矩阵，提出了CSC-GEP的设计以解决CSC问题，同时通过在UCI数据集[8]中Heart disease、Sick和Credit数据集对算法的有效性进行了测试。

2 基因表达式程序设计

基因表达式程序设计(Gene Expression Programming，GEP)[9-10]是葡萄牙科学家Cândida Ferreira发明的一种基于基因组(Genome)和表现型组(Phenome)的新的遗传算法。它与遗传算法(Genetic Algorithms，GAs)和遗传程序设计(Genetic Programming，GP)的根本区别在于它们所采用的个体的性质不同：在GAs中个体是固定长度的线性串(染色体)；在GP中个体是长度和形状不同的非线性实体(分列树)，在基因表达式程序设计中个体被编码成固定长度的线性串(基因组或者染色体)，然后被表达成不同长度和形状的非线性实体(简单图示或者表达式树)。

基因表达式程序设计的实现主要包括编码方式和染色体构成、遗传算子的选择、适应度函数的选择等四个部分。

3 基于基因表达式编程的敏感代价分类算法

由于演化算法具有强大的全局搜索性，因此将演化算法用于机器学习领域一直是一个极具吸引力的研究方向。对于CSC问题，已有将代价矩阵引入GAs、GP以图建立CS分类算法的论文发表[11]。GEP在算法设计上结合了GAs、GP两者的长处，同时与常规算法相比，GEP的工作原理是在广域搜索空间中搜索最适合的解集,不需要复杂的构建步骤，只需要定义需要获得的结果。

在演化算法中实现分类器一种通用的办法是将分类器视作分类规则的集合，每一个基因表达一条分类规则。基于这种原则，CSC-GEP主要的设计如下：

3.1 编码方式

GEP的原始编码在设计上比较适合数值性的计算，因此为了实现分类规则集，GEP的符号集设计为{OR, AND, NOT, $A\{\}$, $B\{\}$, $C\{\}$,…}，其中OR、AND、NOT是逻辑运算符与、或、非。$A\{\}$、$B\{\}$、$C\{\}$表示训练集各属性。如$A\{\}$代表属性A中所拥有的几类值。如A属性是离散特征值则直接加入$A\{\}$；如A属性是连续值或多类别常规离散值则进行适当离散化或模糊化再加入$A\{\}$。在算法实现时，应使用多符号集方法，即符号集为{OR, AND, NOT}、{A_1, A_2, A_3, …}…，其中A_1、A_2、A_3是将训练集如前所述进行预处理后所得的对应属性值。

3.2 解码方式

由于分类规则的效果评价与常规函数有极大的差异，因此在解码方式上也应使用不同的方法。使用前序表达式解码以避免建立表达式树所产生的时间代价，同时遍历产生的表达式不使用括号。NOT运算符只对表达式中的下一位起作用，如为运算符则OR变为AND，AND变为OR，NOT则取消。如下一位为属性值则在对应的属性集中随机选取一个新属性值加以替换。

3.3 适应度函数计算

解码完成后依次读入训练集数据，对应每一个基因及其所属的默认类，对匹配的样本数据对比其真实类别，并根据相应的代价矩阵进行累加再除以与基因匹配的样本总数。适应度函数定义为：

$$f_{\mathrm{cs}} = \mathrm{mc\_count} \times \mathrm{cost}(i, j) / \mathrm{m\_count}$$

式中 $\mathrm{mc\_count}$ 是与基因匹配的全部样本中的误分类样本数；$\mathrm{mc\_count}$ 是与基因匹配的全部样本数；$\mathrm{cost}(i, j)$ 是将i类样本误分类为j类的代价。

3.4 交叉/重组操作

基本操作不变，但由于基因的复杂性和交叉/重组操作可能导致矛盾基因，因此在完成此类操作后启动基因检查算法，对新基因有效部分中的非操作符进行比对，当发现同属性中的不同属性值则强行将不同值转化为相同值。这一检查可有效避免无效基因的产生。

3.5 变异操作

基本操作不变，但当确定变异位置时进行检查。如为操作符则变异为操作符集{OR,AND,NOT}中的任意其他一个操作符；如为属性值则变异为对应属性集中的其他一个随机值。

4　实　验

在实验中，考虑到现实应用中对稀有样本的正确分类往往更加重要，所以在代价矩阵的设计中将错误分类多数类的代价设为1，错误分类稀有类的代价设为$f(f>1)$。为保证算法的有效性，令f=2，5，10。每个代价因子上的实验数据都运行10次并取平均值，且最后报告的结果取3个代价因子实验数据的平均值。GEP部分参数设置初始种群大小为500；演化代数为200；交叉概率为0.9；变异概率为0.05。实验结果如表1所示。

表1　总分类代价

数据集	C4.5	Cost-UBoost	CSC-GEP
Heart-disease	16.57	9.57	9.53
Sick	14.24	7.39	7.61
Credit	32.24	22.50	22.36

表2　高代价类别误分类率

数据集	C4.5	Cost-UBoost	CSC-GEP
Heart-disease	2.25	0.80	0.18
Sick	2.01	0.78	0.61
Credit	4.83	2.18	1.20

表3　误分类率

数据集	C4.5	Cost-UBoost	CSC-GEP
Heart-disease	6.06	8.29	9.29
Sick	4.99	5.72	6.32
Credit	9.93	18.63	20.14

表1、2和表3分别从分类总代价，稀有类别的误分类率以及总的误分类率列出了C4.5算法，Cost-UBoost算法以及本文所提出的CSC-GEP在Heart-disease、Sick和Credit数据集上的测试结果。表1的数据说明CSC-GEP所生成的分类器在总分类代价上明显低于C4.5，其代价与Cost-UBoost基本相当。表2和表3则从总误分类率和高代价类别误分类率两方面说明CSC-GEP通过牺牲总体分类精度作为代价，以提升高代价类别的分类精度并最终达到降低分类器总代价的目的。实验结果表明CSC-GEP是一种有效的代价敏感分类算法。

5　结　论

当不同类别的样本被错误分类产生的代价不相同时，以分类精度为评估标准的传统分类算法就无法使问题得到良好的解决。基于标准的GEP算法，本文通过将代价矩阵引入GEP并对编/解码方式以及遗传算子加以改进，提出了CSC-GEP的设计方法。实验结果表明，CSC-GEP虽然在分类精度方面表现不佳，但是却有效地降低了分类代价，同时也明显提高了高代价类别的分类正确率。

为验证CSC-GEP的有效性，本文在三个二分类的UCI数据集Heart-disease、Sick和Credit上，对算法进行了测试并给出了测试结果。将CSC-GEP的应用范围扩展到多类别数据集，提高算法的运行速度和收敛成功率是下一步的主要工作。

参　考　文　献

[1] ZADROZNY B, LANGFORD J , ABE N. Cost-sensitive learning by cost-proportionate example weighting[C]// Proceedings of the 3rd International Conference on Data Mining. Melbourne: ACM Press, 2003: 204-213.

[2] DOMINGOS P. Metacost: a general method for making classifiers cost-sensitive[C]//Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining. San Diego: ACM Press, 1999: 155–164.

[3] FAN W, STOLFO S, ZHANG J, et al. AdaCost: misclassification cost-sensitive boosting[C]//Proceedings of the 16th International Conference on Machine Learning. Bled: ACM Press, 1999: 97-105.

[4] CHAN P, STOLFO S. Toward scalable learning with nonuniform class and cost distributions[C]//Proeedings of the 4th International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 1998: 164-168.

[5] TING K M, ZHENG Z. Boosting cost-sensitivetrees [C]//Proeedings of the 1st International Conference on Discovery Science. Fukoka: Springer Press, 1998: 245-255.

[6] 郑恩辉, 李　平, 宋执环. 基于支持向量机的代价敏感挖掘[J]. 信息与控制, 2006, 35(3): 294-298.

[7] LING C X, SHENG V S. A comparative study of cost-sensitive classifiers[J]. 计算机学报, 2007, 30(8): 1203-1211.

[8] BLAKE C, KEOGH E, MERZ C J. UCI repository of machine learning databases[DB/OL]. http://www.ics.uci.edu/mlearn/MLP~epository.html, 2007-6-10.

[9] FERRERA C. Gene expression programming: a new adaptive algorithm for solving problems[J]. Complex Systems, 2001, 13(2): 87.

[10] 张烈超, 蔡之华, 陈安升. SGA,GP,GEP的研究概述[J]. 微计算机信息, 2006, 22(1-2): 185-187.

[11] LI Jin, LI Xiao-li, YAO Xin. Cost-sensitive classification with genetic programming[C]//In Proceedings of the Congress on Evolutionary Computation. [S.l.]: Springer Press, 2005: 133-141.

编　辑　张　俊