第43卷 第6A期
2016年6月

计 算 机 科 学
Computer Science

Vol. 43 No. 6A
June 2016

基于GEP的分类规则挖掘

付红伟

(军事经济学院 武汉 430035)

摘 要 分类规则挖掘方法和回归问题的区别在于分类规则挖掘的目标属性是离散的标称值,而回归问题的目标属性是连续和有序的值。主要介绍了用GEP实现分类规则挖掘的两种主要方法,并分析了如何对适应度函数进行改进以挖掘易于理解的分类规则。

关键词 GEP,分类规则,适应度函数

中图法分类号 TP391 **文献标识码** A

Classification Rule Mining Based on GEP

FU Hong-wei

(Military Economy Academy, Wuhan 430035, China)

Abstract Classification rule mining method and regression problems differ in that the target attribute of mining classification rules is discrete nominal value, and the target attribute of regression problem is the continuous and orderly value. This article mainly introduced two main methods of classification rule mining implementd by GEP, and analyzed how to improve the fitness function to mine classification rules which is easy to understand.

Keywords GEP, Classification rule, Fitness function

分类规则挖掘算法主要包括决策树方法(Decision Tree)、贝叶斯方法[1]、神经网络方法、最近邻方法(Nearest Neighbor)、粗糙集方法(Rough Set)、基于案例的方法(Case-based)和遗传算法等。评估分类规则的5个标准为预测的准确率、计算速度、鲁棒性、可伸缩性和可解释性。虽然已有许多比较不同分类算法的方法,但仍未形成一个准则,该问题仍然是一个研究课题。

1 GEP分类方法

1.1 基本的GEP分类方法

基本的GEP分类方法是由C. Ferreira所提出的。C. Ferreira采用一种0/1舍入阈值的方法,具体思想如下:设阈值为Rt,当染色体的输出等于或大于Rt时把该输出转换成1,否则转换成0,也就是当表达式返回的值大于这个阈值时,就认为它属于分类的目标类别,否则不属于目标类别。

函数符号可以采用基本的算术运算符,也可以采用类似于大于、小于、小于或等于、等于或大于、等于以及不等于这样的逻辑函数符号。终点集由所有的属性集构成。所使用的适应度函数非常简单。个体的适应度f_i对应正确分类的训练样本的个数。

对于多类别分类问题,必须对数据进行重新整理。将数据分类成n个不同类C时需要将数据处理成n个独立的0/1分类问题,然后分别根据n个划分进化得到n个不同的模型,再将这n个模型结合起来,就得到了最终的分类规则。这种方法的一个优点是思想简单,将所有的分类问题分解成二元分类问题;缺点是该适应度函数对于数据样本中某些样本数量较少的类来说,很容易陷入局部最优解[2]。

Candida也采用了另外一些适应度函数,比如能够较好地解决数据不平衡问题的适应度函数是敏感性/特效性适应度函数。该函数考虑到了数据样本中数据的分布问题,根据分类规则将分类后的数据分为4种,其中真正样本(Ture Postive, TP)是该样本本身是属于该类也被正确分类在该类种的样本;真负样本(Ture Negtive, TN)是该样本本身不属于该类而被正确分在不属于该类种的样本;假正样本(Fasle Postive, FP)是该样本本身属于该类而被误分类在该类种的样本;假负样本(False Negtive, FN)是该样本本身不是属于该类而被误分类在属于该类种的样本。敏感性/特效性(Sensitivity/Specificity)计算公式如下:

$$Sensitivity = TP/(TP+FN)$$

$$Specificity = TN/(TN+FP)$$

1.2 改进的GEP分类方法

改进的GEP分类方法是芝加哥大学伊利诺伊州分校的ChiZhou等人提出的。对于二分类问题,只需要一个规则就可以将两个类别分开,对于n个规则,则采用n个二分类问题来做等价处理,然后将这n个规则合并一个规则集。

ChiZhou采用单基因染色体,放弃了C. Ferreira提出的头尾结构,设定一个固定的基因长度,采用平等对待函数符号和终点元素的办法,在生成初始种群之后,判断种群中每个个体的合法性,保留合法的个体,剔除不合法的个体,这样能够保证种群中的个体都是合法的;这种方法也能保证种群中个体表达式树的长度是动态变化的。

ChiZhou等人不采用舍入阈值的思想,而是直接考虑表

付红伟(1982—),男,讲师,主要研究方向为计算机技术及应用、数据挖掘,E-mail:hwfu123@163.com。

达式返回值的符号问题。如果表达式的返回值大于0，则认为该样本属于该类；否则不属于该类。

ChiZhou在终点集中加入了1和其它几个素数常数{2，3，5，7}等，从理论上说来，只要有了这几个常数，将其组合就可以得到任何有理数。但是前面关于符号回归的实验说明对大部分问题而言，加入随机常数对于问题解的质量并不会带来提高。

采用的适应度函数为完整性和一致性函数。其中完整性表示规则所覆盖的正例的样本个数与整个训练数据集中的正例个数的比例。而一致性则表示规则所覆盖的正例与规则所覆盖的整个样本的总数[4]

一致性增益(Consistence Gain)的定义为：

$$consig_R=(\frac{p}{p+n}-\frac{P}{P+N})*\frac{P+N}{N}$$

其中，p和n是规则R所覆盖的正例(Positive Example)和反例(Negative Example)的个数，P和N是训练集中所有的正例和反例的个数。这个定义考虑到了训练集中正例与反例的分布问题。

完整性定义为：

$$compl_R=\frac{p}{P}$$

规则的适应度函数为：

$$fitness=consig_R*\exp(compl_R-1)$$

该公式中所采用的指数函数能够让规则更加倾向于一致性增益这一项，同时能够将整个适应度函数值控制在$[0,1]$这个范围之内。

对于分类冲突和分类失败，采取的策略为：将所产生的表达式按照其适应度值降序排列，如果两个表达式具有相同的适应度值，则将覆盖训练集中较多的那个表达式排在前面。对于每一个新的输入样本，将n个表达式逐个运用于这个样本上，直到其中一个产生正值，然后将该样本赋予一个满足的表达式所代表的类，剩余的表达式就不需要继续计算。假设对于某一个样本，已经对所有的表达式都求了值，但仍然没有找到满足的表达式，则对该样本赋予一个缺省类。其中该缺省类选取未分类的样本中含有最多的样本个数的类。

ChiZhou还采用覆盖策略(Covering Strategy)在一个数据集上产生多个分类规则，并采用MDL规则来对规则集进行剪枝，这在理论上确实能够避免过拟合的问题；而且ChiZhou的方法考虑到了将属性为名词型的数据进行离散化的问题，能够解决属性值不全为数值型的问题，使算法具有更广泛的适应性，这是有效改进的一个方面。

2 适应度函数的改进

李曲提出了一种挖掘简洁分类规则的GEP分类方法，下面先介绍其主要思想。

由于整个规则的长度取决于GEP中Karva表达式的长度(Karva Expression Length，KEL)而不是染色体的长度(Chromosome Length，CL)，因此为了让得到的规则更简洁，一个很自然的想法是让表达式的长度作为一个罚函数项，使得在分类精度不受损失的情况下，搜索过程能更倾向于简单的规则。我们称之为“简洁GEP”(Parsimonious Gene Expression Programming，PGEP)分类算法[3]。

基于这种思想，李曲提出如下的适应度函数：

$$Fitness_i=N_i-\frac{KEL_i}{CL_i}$$

其中，N_i是规则能够正确分类的样本个数，KEL_i是规则所对应的Karva表达式的长度，CL_i是染色体的长度。

该适应度函数的好处在于N_i是一个整数值，而$KEL_i<CL_i$，因此第二项是一个$[0,1]$之间的小数。所以一方面如果一个规则能够分类n个样本，具有较短的表达式长度，另一个规则能够分类$n+1$个样本，但是表达式较长，则很显然算法会选择第二个规则；而如果两个规则都能够正确分类m个样本，但其中第一个规则较短，则很显然这个较短的规则会被选中。

当然，我们也考虑过采用罚函数压力更大一些的情况，比如

$$Fitness_i=N_i-\frac{KEL_i}{CL_i}*N_i$$

的情况。实验证明，若罚函数压力过大，会造成整个搜索构成太过倾向于找到更短的规则，降低规则的分类能力。

本文也提出了对适应度函数的改进策略，希望得到的分类规则在具有准确性的同时还具有好的易于理解性，因此对适应度函数进行了改进，定义其为准确度和易于理解性的加权和：

$$fitness=w_1*accuracy+w_2*comprehensibility$$

其中，$accuracy$表示规则正确分类的样本数占总样本数的比例，即$\frac{N_i}{N}$，N_i表示规则能够正确分类的样本个数，N表示样本总数；$comprehensibility$可以简单地用$\frac{1}{KEL_i}$表示，即

$$fitness=w_1*\frac{N_i}{N}+w_2*\frac{1}{KEL_i}$$

w_1和w_2的值介于0和1之间，且$w_1+w_2=1$。这样，适应度的值也在0和1之间变化。用户还可以通过修改权值的大小得到具有不同易于理解性的分类规则，当$w_1=1$且$w_2=0$时，适应度函数即等于分类规则的准确度。

综上所述，所构造的适应度函数总体上考虑了规则的准确度和易于理解性。

3 GEP挖掘分类规则实验

3.1 实验参数设置

本文引用美国加利福利亚大学机器学习知识库中的标准数据集对GEP分类规则挖掘算法进行了实例测试。

在数据预处理的部分，对于本身为二分类的数据，仅进行一次分类；对于n分类问题，采用前面提到的将n分类问题分解为n个二分类问题，每次解决一个二分类问题。

为了保证两种算法都能充分发挥效力，达到最好的效果，实验中参数设置如下：对于所有的测试，采用最大演化代数为100，群体规模为100个个体，头部长度为8，基因个数为3。函数集为基本算术运算符$F=\{+,-,*,/,Q,s,c\}$。Q，s，c分别代表开方、sin、cos运算。终点符号为每个测试数据集的所有属性。

各算子的选择概率设置如下：变异概率取0.033，单点杂交算子概率取0.7，两点杂交算子概率取0.3，IS变换和RIS变换概率均取0.22，基因交叉概率取0.1。

这里，在原GEP的基础上对个体头部的生成进行了改

(下转第460页)

Software Validation[J]. Communication of the ACM, 1978:21(1):1048-1064

[3] Goguen J A, Thatcher J W, Wagner E G, et al. Abstract data-types as initial algebras and correctness of data representations [C]// Proc. Conf. on Comptr. Graphics, Pattern Recognition and Data Structure. 1975

[4] Zilles S N. Abstract specifications for data types[R]. IBM Res. Lab., San Jose, Calif., 1975

[5] Meyer B. Applying "Design by Contract"[J]. Computer, 1992, 25(10):40-51

[6] Findler R B, Felleisen M. Behavioral Interface Contracts for Java [OL]. http://www.researchtate.net/publication/2245179—Behavioral—Interface—Contracts—for—Java

[7] Findle R B, Felleisen M. Contracts for Higher-Order Functions [J]. ACM Sigplan. Notices, 2002, 37(9):48-59

[8] Cheon Y, Leavens G T, Sitaraman M, et al. Model Variables: Cleanly Supporting Abstraction in Design By Contract[J]. Software-practice & Experience, 2003, 35(6):583-599

[9] Hoffman D, Strooper P. State Abstraction and Modular Software Development[M]// SIGSOFT 95. Washington, DC, USA, 1995

[10] Hatcliff J, Leavens G T. Behavioral Interface Specification Language[J]. ACM Computing Surverys, 2012, 44(3):1-58

[11] Jacobs B, Piessens F. Inspector Methods for State Abstraction: Soundness Proof[J]. Journal of Object Technology, 2007, 6(5): 55-75

[12] Dallmeier V, Wasylkowski A, Bettenburg N. Identifying Inspectors to Mine Models of Object Behavior[C]// ICFEM. 2004

[13] Jacobs B, Piessens F. Inspector Methods for State Abstraction: Soundness Proof[R]. CW Reports, 2007

[14] Grunwald D, Gladisch C. Generating JML Specifications from Alloy Expressions[M]// Hardware and Software: Verification and Testing. 2014

[15] Agostinho S, Moreira A. Contracts for Aspect-Oriented Design [C]// SPLAT 2008. ACM, 2008

[16] Kumar A, Bandyopadhyay. Modeling of State Transition Rules and its Application[J]. ACM SIGSOFT Software Engineering Notes, 2010, 35(2):1-7

[17] Polikarpova N, Furia C A. What Good Are StrongSpecifications? [C]// ICSE 2013. San Francisco, CA, USA, 2013

[18] Polikarpova N, Furia C A, Meyer B. Specifying reusable components[C]// VSTTE. LNCS, vol. 6217, 2010:127-141

[19] Wei Y, Furia C A, Kazmi N, et al. Inferring better contracts[C]// ICSE. 2011:191-200

[20] Wei Y, Roth H, Furia C A, et al. Stateful testing: Finding more errors in codeand contracts[C]// ASE. 2011:440-443

(上接第 453 页)

进,即以一定概率选择函数集和终止符集中的元素来组成头部。具体地,若依概率 P 挑选函数集中的元素,则依概率 $(1-P)$ 挑选终止符集中的元素。实验表明,当挑选函数集中元素的概率略大于挑选终止符集中元素的概率时,群体的整体质量较好。本文中 p 取 0.65。

采用的适应度函数公式如下:

$$fitness=w_1 * \frac{N_i}{N}+w_2 * \frac{1}{KEL_i}$$

其中,w_1、w_2、KEL_i 等具体含义见第 2 节。

每个算法运行 25 次,最后得到的结果为 25 次运行结果的平均值。

3.2 实验结果

因 Iris 数据集是 3 类别分类问题,故对数据集进行相应处理,得到 Iris-setosa、Iris-versicolor、Iris-virginica 这 3 个数据集。通过反复实验,对 w_1 和 w_2 分别进行不同的组合,得到的实验结果如表 1 所列。

表 1 改进 GEP 分类结果

数据集		$w_1=1.0$ $w_2=0.0$	$w_1=0.9$ $w_2=0.1$	$w_1=0.8$ $w_2=0.2$	$w_1=0.7$ $w_2=0.3$
Iris-setosa	分类精度(%)	100	100	100	100
	KEL 平均值	26	7	6	6
Iris-versicolor	分类精度(%)	97.4	97.3	96.7	95.8
	KEL 平均值	28	22	18	14
Iris-virginica	分类精度(%)	97.6	97.4	97.3	96.7
	KEL 平均值	20	16	12	10

从表 1 可以看出,随着 w_1 逐渐变小,各数据集的分类精度有所下降,但染色体的有效长度 KEL 相应也变小,即分类精度下降而规则的易理解性上升。当 $w_1=1.0$ 时,适应度函数等于分类规则的准确度,故所产生的规则分类精度很高,但由于未考虑染色体的有效长度,导致所产生的规则的 KEL 平均值偏高;当 $w_1=0.9$ 时,适应度函数综合考虑分类准确度和 KEL 长度,但通过给予分类准确度更高的权重来实现准确度优先考虑而 KEL 长度次之考虑,所产生的规则准确度有一定程度的下降,KEL 平均值有明显的下降;依此类推,w_1 取 0.8、0.7 时,分类准确度逐渐降低,而 KEL 平均值逐渐减小即复杂度降低。之所以没有考虑 $w_1=0.1$、0.2、0.3、0.4 等情况,是由于对于分类规则挖掘而言,应该更强调分类准确度而非规则的简单性,如果 w_1 取以上值,会导致规则简短但分类精度达不到要求,故不予考虑。

但通过分析可以看出,w_1 取值 1.0 和 0.9 的情况下,分类准确度的下降微乎其微,而 KEL 平均值的下降却显而易见,对于表中 3 个数据集,如果分类准确度只考虑 2 位数,则 w_1 取值 1.0 和 0.9 是没有区别的。故在对分类准确度的精度要求不是特别高的情况下,可以通过设置 w_1 的值来兼顾准确度和规则简单性。

综上所述,可以看出,在进行分类规则挖掘时,采用本文所提出的适应度函数,可以通过对 w_1 和 w_2 的值进行相应设置,达到综合考虑分类准确度和易于理解性的目的。

参 考 文 献

[1] Peilikan M. A Simple Implementation of Bayesian Optimization Algorithm[OL]. http://core.ac.uk/display/22715852

[2] 付红伟,刘园园,陈金鑫,等. 改进的 GEP 算法在演化建模中的应用[J]. 重庆工学院学报,2009(3):167-170

[3] 付红伟,毛亚梅,罗炜,等. 混合决策树/遗传算法的数据挖掘[J]. 软件导刊,2009,4(1):157-159

[4] 姜大志,吴志健,康立山,等. 基因表达式程序设计的 GRCM 方法[J]. 系统仿真学报,2006,6(18):1466-1468