

基于遗传算法的易于理解的分分类规则构造

赵雷, 朱文兴

(福州大学数学与计算机科学学院, 福州 350002)

摘要: 提出了一种分类规则易于理解性的新的定义, 并给出了应用属性信息增益计算分类规则可理解性的方法。分析了遗传算法发现分类规则的过程。乳腺癌统计数据上的实验表明, 它可以发现准确和易于理解的分分类规则。

关键词: 分类规则; 遗传算法; 信息增益; 易于理解的规则

Comprehensible Classification Rules Construction Based on Genetic Algorithm

ZHAO Lei, ZHU Wenxing

(College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350002)

【Abstract】 This paper puts forward a new definition for the comprehensibility of a classification rule. The paper also gives a method for computing a rule's comprehensibility by using attribute's information gain. The method of mining classification rules using GA is also analyzed. The experiment on breast cancer statistical data shows that it can find accurate as well as comprehensible rules.

【Key words】 Classification rules; Genetic algorithm; Information gain; Comprehensible rules

分类(Classify)是数据挖掘的一种重要的数据分析形式, 可以用于提取描述数据类的模型或预测未来的数据趋势。常用分类算法主要有决策树方法、贝叶斯方法、人工神经网络方法、粗糙集方法和遗传算法等。目前应用遗传算法求取分类规则的研究还处于原型阶段, 然而却日趋流行^[4]。

现有的基于遗传算法的分分类规则挖掘主要考察分分类规则的准确性(accuracy)、有趣性(interestingness)和易于理解性(comprehensibility)^[1]。文献[1]基于规则的简单度对规则的易于理解性进行了度量, 认为分分类规则的长度越短或形式越简单, 则规则越容易理解。但简单规则存在着不能给用户提提供足够知识的缺陷, 而在诸如医疗或科研等领域, 用户通常希望规则给他们提供更多和更深入的知识便于理解与分析, 此时规则简单只是其次考虑的因素。文献[2]则认为“IF_THEN”形式的分分类规则是易于理解的, 这也仅仅是从规则的外在形式对易于理解性进行的衡量。国内文献目前还没有关于分分类规则易于理解性的论述, 其中文献[8]提出了一种基于增量式遗传算法发现分分类规则的方法, 但也主要是强调发现规则的准确性。

本文提出了一种基于遗传算法发现易于理解的分分类规则的方法: 给出了分分类规则易于理解性的新的定义, 基于规则所提供的知识和信息来考察规则的易于理解性, 不同于仅从简单性进行度量的方法; 提出了使用属性信息增益计算分分类规则的易于理解性的方法。在此基础上, 构造了相应的适应度函数; 最后以乳腺癌统计数据为例, 应用小生境遗传算法构造乳腺癌重发与不重发的易于理解的分分类规则。

1 基于遗传算法的分分类规则发现

1.1 分分类规则的模型表示及编码

分类是在已有数据的基础上构造一个分类函数或分分类模式(或称为分分类规则)。该函数或模型能够把数据库中的数据记录映射到给定类别中的一个, 从而应用于数据预测^[7]。

最简单的情况, 进行挖掘的类属性只有一个, 分分类规则模型表示为: IF ($A_1 = I_1 \ A_2 = I_2 \ , \dots \ A_k = I_k$) THEN ($D=J_j$)。其中, IF部分成为规则的前提, THEN部分是结论。

我们采用Michigan编码方法^[1], 即单个记录对应单条分分类规则。并且, 参于进化的初始群体由同类个体组成, 算法的一次运行得到一类个体对应的分分类规则^[2]。参加预测特征属性的数目为n, 分别记为: A_1, A_2, \dots, A_n , 编码方案如图1所示。

A_1				A_i				A_n			
A_1	O_1	V_1	G_1	A_i	O_i	V_i	G_i	A_n	O_n	V_n	G_n

图1 一个染色体的编码表示

由图1知, 数据集中的一个记录编码成遗传的染色体, 特征属性 A_i 编码成了基因, 每个基因由 W_i, O_i, V_i ^[2]和 G_i 4个域构成。其中 W_i 是权值域, 为特征属性上当前值的个体数占有所有个体的百分比, 可通过设置权值域的阈值生成变长的分分类规则^[2]。 O_i 是运算符, 对离散属性而言, 取“=”和“≠”符号; 对于连续属性, 取“≤”和“>”符号。 V_i 是属性的值域, 采用二进制编码方法。对于离散属性, 二进制的位数为属性的离散值的数目。对于连续值属性, 则可以直接用连续值的二进制来表示^[1]。我们添加 G_i 作为属性的信息增益域, 增益值在遗传算法执行前计算并存储在 G_i 中。

1.2 适应度函数设计

对于前提为A, 结论为C的分分类规则, 数据中存在4类不同的个体, 见表1。

为度量规则的准确性, 进行如下定义:

定义1 规则的置信度^[1] $confidence = pp/(pp+pn)$

基金项目: 国家自然科学基金资助项目(10301009)

作者简介: 赵雷(1974-), 男, 硕士、助教, 主研方向: 人工智能, 知识发现; 朱文兴, 博士、教授

收稿日期: 2005-12-06 **E-mail:** :zl@fzu.edu.cn

定义2 规则的覆盖度^[1] complement = pp/(pp+np)

定义3 规则的准确度^[1] accuracy = confidence × complement

定义4 适应度函数^[1] fitness = accuracy

表1 数据库中的4类个体及数目

规则	IF A THEN C	IF A THEN NOT C	IF NOT A THEN C	IF NOT A THEN NOT C
个体数目	pp	pn	np	nn

2 基于遗传算法的易于理解分类规则发现

2.1 易于理解分类规则的概念

现有文献[1, 2, 5]认为规则的易于理解性反比于规则的长度或规模,即规则越短,则规则的可理解性越强。我们认为,上述观点仅从规则简单程度上对易于理解性进行解释,没有深入其概念的实质,因此这里称其为简单度(simplicity),有别于我们对规则易于理解度的定义。如果规则的简单度反比于其长度k,可定义为

定义5 规则的简单度 simplicity=1/k

依定义5构造的适应度函数不一定得到易于理解分类规则。如下发烧科门诊数据库,记录数目为20,特征属性取值“T”表示有对应症状,“F”表示无对应症状,发烧、腹痛、头疼、腹泻为特征属性,非典为类属性。数据表及代表规则如表2所示。

表2 一个医院的发烧门诊病人数据

ID	发烧	腹痛	头疼	腹泻	非典	代表规则
1, 2	T	F	F	F	YES	IF 发烧=T THEN 非典="YES"
3-10	T	T	T	T	YES	IF (发烧 腹痛 头疼 腹泻)=T THEN 非典="YES"
11	T	F	F	F	NO	IF 发烧=T THEN 非典="NO"
12-20	F	F	F	F	NO	IF (发烧=F 腹痛=F 头疼=F 腹泻=F) THEN 非典="NO"

为分析非典病的成因,按定义4计算对“非典=YES”类的个体规则的简单度如表3所示。

表3 基于准确性的适应度函数计算

个体	代表规则	pp	pn	np	simplicity
1, 2	(1)IF 发烧=T THEN 非典=YES	10	0	1	0.909
3-10	(2)IF (发烧 腹痛 头疼 腹泻)=T THEN 非典=YES	8	0	2	0.8

计算知,规则(1)的简单度高于规则(2),按文献[1]观点,规则(1)的易于理解性应优于规则(2)。然而,常识告诉我们,医务工作者们普遍认为规则(2)能提供更多与非典病相关的信息,规则(2)的易于理解性反而应该优于规则(1)。所以不能仅仅把规则的易于理解性等价于规则的简单度。

2.2 基于属性信息增益的规则可理解性的定义

针对上述缺陷,本文对“规则易理解性”提出了新的定义,即:规则的易理解性是指一条规则能够提供给用户更多相关领域的知识和信息。并且,我们引入了属性信息增益(Information Gain)来计算一条规则的易于理解性。设对一个给定样本分类所需的信息为 $I(s_1, s_2, \dots, s_m)$,属性A的熵为 $E(A)$,则属性A的信息增益按下式计算^[4]:

$$InfoGain(A) = I(s_1, s_2, \dots, s_m) - E(A) \quad (1)$$

在分类中,信息增益技术最常见于决策树算法中。决策树算法基本上属于贪心算法,它采用自顶向下的方法构造决

策树。最大信息增益的属性位于树顶,在每一层都要找到剩余属性中最高增益的属性,作为当前节点的测试属性。该属性使得对结果划分中的样本分类所需的信息量最少,并反映划分的最小随机性或“不纯性”^[4]。高信息增益的属性是给定集中具有高区分度的属性,我们希望分类规则中用于预测的特征属性的信息增益值都尽可能高,从而对应分类规则所包含的信息量也就越多。因此,我们定义基于属性信息增益的分类规则易于理解性如下:

定义6 规则的易于理解性

$$comprehensibility = \frac{\sum_{j=1}^k InfoGain(A_j)}{\sum_{j=1}^n InfoGain(A_j)}$$

其中k是规则的长度,即参与预测的特征属性的数目。我们定义分类规则的易于理解性等于规则中参与预测的特征属性信息增益的和与所有特征属性的信息增益的和的比值。

2.3 对适应度函数的修改

为了得到准确和易于理解分类规则,我们修改了适应度函数,定义其为准确度和可理解度的加权和:

定义7 fitness = w₁ × accuracy + w₂ × comprehensibility

w₁和w₂的值介于0,1之间,且w₁ + w₂=1。这样,适应度值在0~1间变化。当w₁=1, w₂=0时,即为分类规则的准确度。例如,对于数据表2,设w₁=0.5, w₂=0.5,按定义7计算个体适应度值如表4所示。

表4 基于易理解性的分类规则适应度的计算

个体	代表规则	pp	pn	np	适应度
1, 2	(1) IF 发烧=T THEN 非典=YES	10	0	1	0.65
3~10	(2) IF (发烧 腹痛 头疼 腹泻)=T THEN 非典=YES	8	0	2	0.9

显然,适应度函数的更改使规则(2)成了优势个体,这也是我们所希望的。

综上所述,我们构造的适应度函数,使得最后得到的规则,既不会太长,因为规则只是尽量包含信息增益最大的一些属性;又不会太短,因为在定义时没有基于规则的简单度,从而体现了规则准确性和易于理解性的良好结合。

2.4 算法流程

我们采用了小生境遗传算法^[6]进行实验,步骤如下:

Step1 确定分类所需特征属性和类属性,随机生成M个记录组成的初始训练数据集T。

Step2 对T预处理,包括数据清理,将连续属性离散化,计算各特征属性的信息增益值,对数据进行编码,得到编码后的初始群体P(t)。设置进化代数计数器t=0,并求各个体的适应度F_i(i=1,2,...,M)。

Step3 按个体的适应度降序排序,记忆前N个个体(N<M)。

Step4 选择运算。对群体P(t)进行比例选择运算,得到P'(t)。

Step5 交叉运算。对选择出的个体集合P'(t)作两点交叉运算,得到P''(t)。

Step6 小生境淘汰。将Step5得到的M个个体和Step2所记忆的N个个体合并在一起,得到一个含有M+N个个体的群体;对这M+N个体,求每两个个体X_i和X_j之间的相异度^[4],适应度较低的个体处以罚函数。

Step7 依据这M+N个个体的新适应度对各个个体降序

排序，记忆前 N 个个体。

Step8 终止条件判断，若不满足终止条件，则：更新进化代数计数器： $t \leftarrow t+1$ ，并将 Step7 排序中的前 M 个个体作为新的下一代群体 P(t)，然后转到 Step3；若满足终止条件，则输出适应度值最大的分类规则，算法结束。

3 实验及分析

数据源选UCI网站的Breast cancer data^[3]，此表统计了乳腺癌的复发状况。数据分为 recurrence-events 和 no-recurrence-events 两类，包含 9 个特征属性和 1 个类属性。参数配置如下：种群大小 pop_size = 286，染色体长度 l_chrom = 9，最大进化代数 max_gen = 200，杂交概率 Pc = 0.8，变异概率 Pm = 0.05，权值域阈值 Wi (i=1...9) = 0.2，适应度函数权值 W1, W2 = 0.5。图 2 为输出界面。

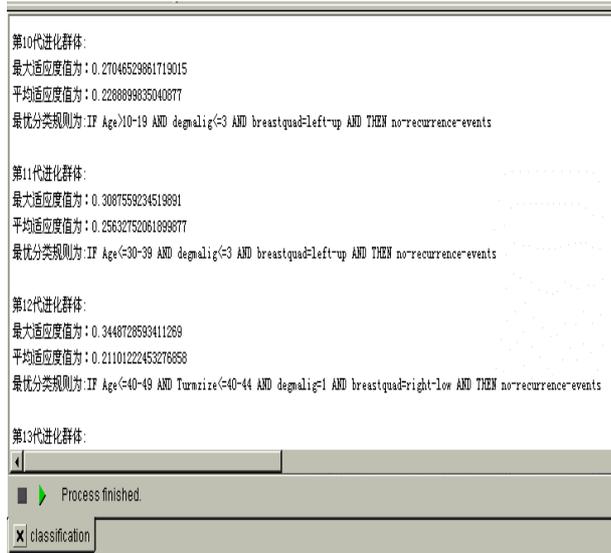


图 2 算法输出

由表 5 知，有关 Breast Cancer 是否复发的规则能提供用户更多相关疾病的背景知识，是医疗工作者所期望的。也可以通过修改 w_1 和 w_2 的值得到不同易于理解性的分类结论。

(上接第 210 页)

罚函数中的 2 个参数，L 为二叉树的大小，即节点的个数。于是，适应度计算公式为

$$fitness = penalty * corrcoeff \left(\sqrt{diag(Q)} * Y, \sqrt{diag(Q)} * Y_m \right)$$

corrcoeff 是关于 Y 、 Y_m 的相关系数。

对每一代中的每一个模型个体计算所得的适应度 fitness 按照表 1 所定义的 GP 演化参数中的竞赛选择方法，计算得到这一代适应度最大的个体作为下一代演化的初始个体，反复迭代，最终得到适应度最大的个体作为描述 IP 业务流量预测的模型结构。

4 结论

本文通过在函数集中引入常微分方程构建新模型对文献[1]中 IP 业务流量预测的 GP 模型加以改进，减少了 GP 模型结构在流量行为分析对预测精度方面存在的不足，从而提高了模型的动态描述能力；同时针对 GP 演化建模中关键参数的设定，对个体适应值变化的影响进行分析评价，并将该 GP 算法中适应性函数的计算做了详细论述，以此增强模型结构在预测方面的外推性能，改善模型的复杂度。

进一步的研究将重点考虑采用并行计算的方法以减少

表 5 算法在 Breast Cancer 数据集上运行后得到的分类规则

类别	最优分类规则	适应度值
Recurrence-events	IF Tumorsize>20-24 AND degmalig=3 AND breastquad=right-low	0.456
No-recurrence-events	IF Age<=50-59 AND invnodes=12-14 AND degmalig<=2	0.730

4 结束语

本文对基于遗传算法的分类规则发现进行了较为全面的分析，在此基础上，进一步研究了易于理解分类规则的构造技术，提出了基于属性信息增益的评价分类规则易于理解性的方法，算法发现的分类规则包含有更多的分类信息，优化了规则的易于理解性。

参考文献

- Freitas A A. A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery[M]. Advances in Evolutionary Computation. Springer-Verlag, 2000.
- Fidelis M V, Lopes H S, Freitas A A. Discovering Comprehensible Classification Rules with A Genetic Algorithm[C]. Proc. of Congress on Evolutionary Computation, La Jolla, CA, USA, 2000-07: 805-810.
- Murphy P M, Aha D W, Irvine U C I. Repository of Machine Learning Databases[D]. Department of Information and Computer Science, University of California, 1994.
- Han J. Data Mining Concepts and Techniques[M]. New York, USA: Morgan Kaufmann Press, 2001.
- Ryan M D, Rayward-Smith V J. The Evolution of Decision Trees[C]. Proc. of the 3rd Annual Conf. on Genetic Programming, 1998: 350-358.
- 徐金梧, 刘纪文. 基于小生境技术的遗传算法[J]. 模式识别与人工智能, 1999, 21(3):104-107.
- 杨炳儒. 知识工程与知识发现[M]. 北京: 冶金工业出版社, 2000: 196-258.
- 邢乃宁, 孙志挥. 基于增量式遗传算法的分类规则挖掘[J]. 计算机应用研究, 2001, 18(11): 13-15.

GP 演化计算所花费的 CPU 时间，加快搜索速度；同时引入小波分析，将 IP 业务流量的高频噪声部分和低频部分分开，用以减小数据复杂度，去除掺杂的噪声，使得到的新模型更能实时反映 IP 业务流量变化趋势。

参考文献

- 曹阳, 王治, 杨艳. 基于遗传程序设计的 IP 业务流量长期预测[J]. 计算机学报, 2003, 26(12): 1-5.
- Silva S. GPLAB A Genetic Programming Toolbox for MATLAB[Z]. 2004. <http://gplab.sourceforge.net>.
- Mad'ar J, Abonyi J, Szeifert F. Genetic Programming for the Identification of Nonlinear Input-Output Models[Z]. 2005. <http://citeseer.ist.psu.edu/716907.html>.
- Luke S, Spector L. A Comparison of Crossover and Mutation in Genetic Programming[C]. Proceedings of the 2nd Annual Conference on Genetic Programming, 1997.
- Koza J R, Bennett III F H, Andre D, et al. Genetic Programming III: Darwinian Invention and Problem Solving[M]. Morgan Kaufmann, 1999: 120-1140.

