

基于增量式遗传算法的粗糙集分类规则挖掘

何 明, 冯博琴, 马兆丰, 傅向华

(西安交通大学电子与信息工程学院, 710049, 西安)

摘要: 从规则获取和优化两个方面研究了基于遗传算法(GA)的增量式粗糙集分类规则挖掘方法. 通过研究决策表和决策规则系数,建立了基于粗糙集表示和度量的知识理论,将 GA 和粗糙集分类规则挖掘算法相结合,在保持原有知识完备的前提下,利用 GA 对以增量形式获得的分类规则进行优化,获取最优分类规则. 试验结果表明,执行增量式 GA 所需时间较执行一般 GA 所需时间要少,可有效完成分类规则优化的任务,同时还可提高分类的精度,使分类结果具有更好的可理解性.

关键词: 粗糙集;数据挖掘;增量式遗传算法;分类规则

中图分类号: TP18 **文献标识码:** A **文章编号:** 0253 - 987X(2004)06 - 0579 - 04

Rough Set Classification Rules Mining Based on Incremental Genetic Algorithm

He Ming, Feng Boqin, Ma Zhaofeng, Fu Xianghua

(School of Electronics and Information Engineering, Xi an Jiaotong University, Xi an 710049, China)

Abstract: The rough set classification rules mining based on incremental genetic algorithm (GA) is studied from two aspects: decision rules acquisition and optimization. Knowledge theory based on rough set representation and measure is constructed according to coefficients of the decision rule and decision table. To acquire optimal classification rules, the proposed method combines GA with the rough set classification rules mining algorithm. Furthermore, the rules, in incremental form, acquired by GA are optimized. Experimental results show that it performs well in the task of optimization. Comparing with the general GA it enhances the classification precision, performs task with less run time, and more understandable result can be obtained.

Keywords: rough set; data mining; incremental genetic algorithm; classification rule

粗糙集理论^[1]是由波兰科学家 Pawlak 在 1982 年提出的,主要是处理模糊和不确定性问题,该理论近年来已成功地应用于人工智能、机器学习、数据挖掘、模式识别和智能信息处理等领域^[2~4]. 分类是数据挖掘的主要研究内容之一,由分类规则组成的类别描述可以对未知数据进行分类预测. 分类规则的挖掘目前主要采用的方法有: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(GA)等.

GA 是一种全局寻优的有效方法,它具有鲁棒性、隐含并行性和全局搜索等特点,也很容易与其他

技术结合. 本文利用遗传算法对粗糙集分类规则集进行优化,以求得最优分类规则集.

1 基本概念

根据粗糙集理论,决策表 $S = (U, C, D)$ 的定义^[1,5]为,给定决策表 S ,每一个对象 $x(x \in U)$ 都对应一个序列 $c_1(x), \dots, c_m(x), d_1(x), \dots, d_n(x)$,其中 $\{c_1, \dots, c_m\} = C, \{d_1, \dots, d_n\} = D$. S 中的每一行对应一个决策规则,可以表示为 $c_1(x), \dots, c_m(x) \rightarrow d_1(x), \dots, d_n(x)$,或者简记为 $C_x \rightarrow D_x$. 每一个决策规则 $C_x \rightarrow D_x$ 都有 3 个系数与之相

收稿日期: 2003 - 09 - 12. 作者简介: 何 明(1975~),男,博士生;冯博琴(联系人),男,教授,博士生导师. 基金项目: 国家高技术研究发展计划资助项目(2003AA1Z2610).

关,即规则的强度、确定度和覆盖度因子。

定义1 决策规则 C_x D_x 的强度 $\text{str}_x(C, D) = \frac{|C(x) \cap D(x)|}{|U|}$, $|X|$ 为集合 X 的基数。

定义2 决策规则 C_x D_x 的确定度 $\text{cer}_x(C, D) = \frac{|C(x) \cap D(x)|}{|C(x)|} = \frac{\text{str}_x(C, D)}{C(x)}$, $C(x) = \frac{|C(x)|}{|U|}$ 。

定义3 决策规则 C_x D_x 的覆盖度 $\text{cov}_x(C, D) = \frac{|C(x) \cap D(x)|}{|D(x)|} = \frac{\text{str}_x(C, D)}{D(x)}$, $D(x) = \frac{|D(x)|}{|U|}$ 。

2 增量式 GA 的粗糙集分类规则挖掘

2.1 基本思想

本文提出一种基于增量式 GA 的分类规则挖掘方法,其实现方法如下。

(1) 从数据库中接收一批数据作为训练集,并调用相应的 GA 进行优化,得到最优分类规则集。

(2) 系统接收增量数据,并用目前最优的分类规则集对其分类。

若分类错误,则将原有数据集和增量数据一起作为训练实例,并再次调用 GA 搜索覆盖这两个数据集的最优分类规则集,此时的 GA 初始群体应包含已有的最优规则集(这一部分与原基本 GA 是不同的),若分类正确,则不必调用 GA。按照上述方法,不断地用新得到的规则集对下一次接收到的增量数据再分类。显然该方法是一种批量式与增量式相结合的方法^[6]。

2.2 GA 优化分类规则集

(1) 分类规则编码。根据 Goldberg 的标准 GA, 每一个个体代表问题的一组解,因此每一个个体应含有表达全部解的一组规则集。每一条规则对应于一个染色体,并由条件部分和结论部分组成。一条分类规则可以看作是由合取范式构成的逻辑公式,如“IF A_1 AND NOT A_2 THEN C_2 ”,也可以用二进制串“100”编码表示。其中,最左边的两个二进制位分别代表属性 A_1 和 A_2 ,最右边的二进制位代表类别。

(2) 适应度函数的确定。构造分类规则的适应度函数为 $F(r) = \text{str}(r) + \text{cer}(r) + \text{cov}(r)$ 。

(3) 规则匹配。利用一种限制性交配策略,只允许同类的规则进行交叉,而且对于同一结论的规则,只允许其条件部分进化。如果分类规则集中的两条分类规则中仅有一个特征属性的描述有所不同,则这两个描述归纳为一个更范化的描述。

(4) 遗传算子。选择算子采用赌轮选择法^[7];交

叉算子是将个体与个体之间的同类规则的共同基因位进行交叉;变异算子按照一定的概率(一般比较小)改变染色体中某些基因的值,突变的范围需根据实际情况而定。

2.3 算法流程

算法1 基于基本 GA 的分类规则优化算法,其输入为分类规则集和遗传算法运行参数,输出为优化后的分类规则集,步骤如下。

步骤1: 选择问题的一个编码,给出一个有 N 个染色体的初始群体 $\text{pop}(t)$, $t = 1$ 。

步骤2: 对群体 $\text{pop}(t)$ 中的每一个染色体 $\text{pop}_i(t)$ 计算其适应度函数 $f_i = \text{fitness}(\text{pop}_i(t))$ 。

步骤3: 若满足停止规则,则算法停止;否则计算概率

$$p_i = f_i / \sum_{j=1}^N f_j, \quad i = 1, 2, \dots, N \quad (1)$$

并以概率分布(1)从 $\text{pop}(t)$ 中选择一些染色体构成一个种群,即

$$\text{newpop}(t+1) = \{\text{pop}_j(t) \mid j = 1, 2, \dots, N\}$$

步骤4: 通过交叉,交叉概率为 p_c , 得到一个有 N 个染色体的 $\text{corsspop}(t+1)$ 。

步骤5: 以一个较小的概率 p , 使得一个染色体的一个基因发生变异,生成 $\text{mutpop}(t+1)$; $t = t+1$, 生成一个新的群体 $\text{pop}(t) = \text{mutpop}(t)$; 返回步骤2。

算法2 基于增量式 GA 的分类规则优化算法,其输入为增量数据,输出为多次优化后的分类规则集,步骤如下。

步骤1: 根据近似空间 K 所提供的知识,确定论域中各对象的条件属性 C (特征属性)和决策属性 D (分类属性),并将其转化为决策表 $S = (U, C, D)$, 选取数据集 D_0 。

步骤2: 对 D_0 进行预处理,包括数据清理和连续属性值的离散化,得到 D_C 。

步骤3: 调用算法1,得到一组最优分类规则集并保存在分类规则库中,同时算法1被更新(含有遗传优化的群体)。

步骤4: 如果不更新数据集,转到步骤8;否则,执行步骤5。

步骤5: 读取增量数据集并进行预处理(数据集的清理和离散化),得到 D_C 。

步骤6: 用当前分类规则库中的规则对 D_C 进行分类,如果达到了预期效果(大于预先设定的分类规则的阈值),则将 D_C 与 D_C 合并,即 $D_C = D_C$

D_C , 并返回步骤 4; 否则, $D_C = D_C \cup D_C$, 执行步骤 7.

步骤 7: 调用修改后的算法 1 (初始群体包含最优分类规则集) 进行分类规则的优化, 得到更新的分类规则集, 并返回步骤 4.

步骤 8: 输出当前分类规则库中的规则集.

3 试验结果与分析

3.1 数据集选取

采用 Kdd in China 网站 (<http://master.chinaren.net/~sinokdd>) 提供的一个关于汽车信息的实验数据集, 对基于 GA 的挖掘结果和基于增量式 GA 的挖掘性能进行了测试. 用 Iris 数据库的数据集 (美国加州大学 Irvine 分校的机器学习数据库, <http://www.sgi.com/Technology/mlc/db/>) 验证本文的增量式 GA 对粗糙集分类规则挖掘的效果.

3.2 增量式 GA 性能测试

(1) 挖掘结果. 群体初始规模为 50, 适应度函数 $F(r) = str(r) + cer(r) + cov(r)$, 其中的 $str(r)$ 、 $cer(r)$ 、 $cov(r)$ 分别对应决策规则的强度、确定度和覆盖度. 用 GA 对 1 000 个数据对象进行操作, 挖掘结果如表 1 所示.

表 1 分类规则挖掘结果

Rule ₁	maint (high med low) lug boot (med big) safety (high) \Rightarrow unacc (0.18, 0.91, 0.34)
Rule ₂	maint (high med) safety (high) \Rightarrow unacc (0.06, 0.94, 0.13)
Rule ₃	maint (very high) person (2) \Rightarrow unacc (0.23, 1.00, 0.43)
Rule ₄	buying (very high med) safety (low med) \Rightarrow unacc (0.41, 0.98, 0.87)
Rule ₅	buying (high med) safety (high) \Rightarrow unacc (0.03, 0.95, 0.06)
Rule ₆	lug boot (small) safety (low med) \Rightarrow unacc (0.09, 0.97, 0.17)

(2) 性能测试. 增量式 GA 性能测试数据集的大小分别选为 500、800 和 1 100 个元组, 增量数据集大小选为 200、300 和 600 个元组. 对原数据集执行基本 GA, 即算法 1; 对加入增量数据后的数据集执行增量式 GA, 即算法 2. 图 1 给出了基本 GA 的性能测试结果, 图 2 给出了增量式遗传算法的性能测试结果.

从如图 2 可以看出, 执行增量式 GA 所需时间比执行基本 GA 所需时间要少.

3.3 增量式遗传分类算法的优化

基于 Iris 数据集, 采用 C5.0 (<http://www.ces.>

unsw.edu.au/~quinlan/) 生成规则集, 见表 2.

经过优化的规则与 C5.0 具有相同的形式, 不同的是分类规则的精度. 表 3 是 C5.0 和经过增量式 GA 优化后的规则在训练集和测试集上的精度.

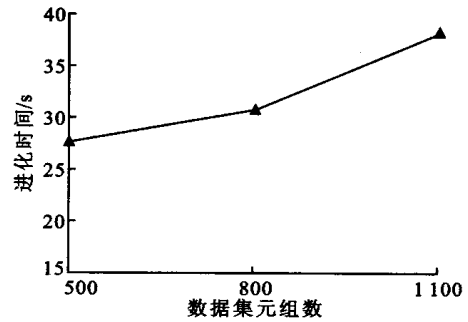


图 1 基本 GA 性能测试结果

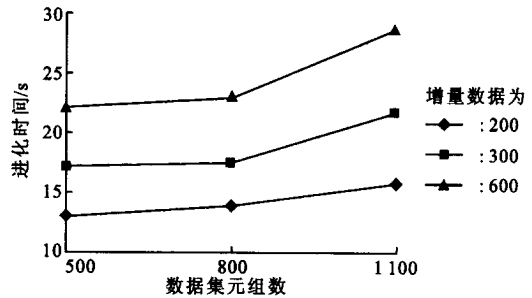


图 2 增量式 GA 性能测试结果

表 2 基于 C5.0 的 Iris 规则集

Rule ₁	Petal.Length \leq 1.9 class Iris-Setosa (覆盖 35 个对象)
Rule ₂	Petal.Length > 1.9 Petal.Length \leq 5 Petal.Width \leq 1.6 class Iris-Versicolor (覆盖 32 个对象)
Rule ₃	Petal.Width > 1.6 class Iris-Virginica (覆盖 29 个对象)
Rule ₄	Petal.Length > 5 class Iris-Virginica (覆盖 28 个对象)
默认类别	Iris-Setosa

表 3 基于 Iris 数据集和增量式 GA 优化后的规则精度

Id/ 个	C5.0 规则集		增量式 GA 优化规则集	
	Tr/ 个	Te/ 个	Tr/ 个	Te/ 个
Virginica	25	24	25	25
Setosa	25	25	25	25
Versicolor	25	23	24	24
精度	98.67 %	96.00 %	98.67 %	98.67 %

注: Id 表示待识别数据集数量; Tr 表示训练集数量; Te 表示测试集数量.

表3的结果表明,增量式GA优化后的分类精度要优于单个分类算法的精度,根据试验结果可以对多个算法进行组合优化,得到一个综合的分类规则。

4 结束语

用粗糙集理论实现高效的分类规则挖掘是一个值得研究的课题。本文以粗糙集理论为基础,针对GA实现技术,结合数据挖掘中分类规则挖掘问题的具体情况,将分类技术与进化方法结合起来,把GA作为一种优化技术应用于分类器上。在此基础上,针对数据挖掘的可扩展性要求,进一步研究了增量式GA的实现技术,并将其应用于分类规则挖掘问题。试验结果表明,这样的结合和扩展可以得到较好的分类效果,它比一般算法更精确或更易理解。利用分类技术处理原始训练数据,可得到初步的分类规则集,而增量式GA作为不同规则的组合器,则可以从一个更高的层次上对规则集进行优化,从而实现了增量挖掘的功能。

(上接第557页)

的近似质量、近似分类精度 d 和信息熵就越小;同时,如果决策属性的属性值划分得越细,则其核属性集不小于划分前的核属性集。因此,在对决策表离散化时,要求决策属性离散化的程度要适宜,即决策细化的程度要适宜。对广义决策表的决策细化问题,将在另文中讨论。

3 结束语

本文基于Rough集理论,研究了决策系统中决策值细化的程度问题,探讨了决策系统中决策值细化程度与核属性、规则近似质量、近似分类精度和信息熵之间的关系,并给出了相应的理论证明。研究决策值细化问题,对研究决策表属性约简、决策规则的形成和规则的有效性及其实用性等问题的研究具有实际意义。同时,该文的研究思想对医学、气象、化工等领域的数据挖掘和系统决策也有借鉴作用。

参考文献:

- [1] 侯利娟,王国胤,聂能,等. 粗糙集理论中的离散化问题[J]. 计算机科学,2000,27(12):89~94.
- [2] 尹旭日,周志华,何贵洲,等. 一种基于Rough集理论的数据过滤方法[J]. 计算机研究与发展,2000,37(9):1082~1086.
- [3] 刘清,黄兆华,刘少辉,等. 带Rough算子的决策规

参考文献:

- [1] Pawlak Z. Rough sets [J]. International Journal of Computer and Information Science,1982,11(5):341~356.
- [2] Pawlak Z. Rough sets: theoretical aspects of reasoning about data [M]. Norwell, USA: Kluwer Academic Publishers,1991.
- [3] Skowron A. Rough sets and Boolean reasoning [M]. New York: Physica-Verlag,2001. 95~124.
- [4] Ziako W. Rough sets:trends, challenges, and prospects [A]. Rough Sets and Current Trends in Computing [C]. Berlin: Springer-Verlag,2001. 1~7.
- [5] Pawlak Z. Theorize with data using rough sets [A]. Computer Software and Application Conference, Oxford, England, 2002.
- [6] 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):13~21.
- [7] Goldberg D E. Genetic algorithms in search, optimization, and machine learning [M]. New York: Addison Wesley,1989.

(编辑 苗凌)

则及数据挖掘中的软计算[J]. 计算机研究与发展,1999,36(7):800~804.

- [4] 张应山. 多边形矩阵理论[M]. 北京: 中国统计出版社,1993.
- [5] Dunsch I, Gediga G. Simple data filtering in rough set systems[J]. International Journal of Approximate Reasoning,1998,18(1-2):93~106.
- [6] Pawlak Z. Rough sets[M]. Norwell, USA: Kluwer Academic Publishers,1991.
- [7] Pawlak Z. Rough sets [J]. International Journal of Information and Computer Science,1982,11(5):341~356.
- [8] Dunsch I, Gediga G. Uncertainty measures of rough set prediction [J]. Artificial Intelligence,1998,106(1):109~137.
- [9] Pawlak Z, Slowinski R. Rough set approach to multi-attribute decision analysis[J]. European Journal of Operational Research, 1994, 72(3): 443~459.
- [10] Lee T L, Tsai C P, Jeng D S, et al. Neural network for the prediction and supplement of tidal record in Taichung Harbor, Taiwan[J]. Advances in Engineering Software, 2002, 33(6):329~338.
- [11] Beaubouef T, Petry F E. Information theoretic measures of uncertainty for rough sets and rough relational databases [J]. Information Sciences,1998,109(1-4):185~195.

(编辑 苗凌)