改进的自适应遗传算法应用研究

雷 亮[1,2] 汪同庆[1] 彭 军[2] 杨 波[2]
(重庆大学光电技术及系统教育部重点实验室 重庆 400044)[1]
(重庆科技学院电子信息工程学院 重庆 400050)[2]

摘 要 图像数据挖掘是目前国际上数据库、图形图像技术和信息决策领域最前沿的研究方向之一。近年来,许多学者开始致力于图像挖掘算法的研究。首先介绍了传统的双种群遗传算法(AGA算法)实现关联规则提取的执行过程,然后针对IAGA算法容易产生停滞现象、造成局部收敛等问题,改进了遗传算子,设计出了新的遗传算法(NAGA算法),最后将NAGA算法成功地运用到遥感图像挖掘,实现了图像关联规则的提取。实验证明,这种改进的自适应遗传算法是一种稳定的、性能优越的算法。

关键词 图像挖掘,自适应遗传算法,遥感图像
中图法分类号 TP 391.41 **文献标识码** A

Application & Research of Improved Adaptive Genetic Algorithm

LEI Liang[1,2] WANG Tong-qing[1] PENG Jun[2] YANG Bo[2]
(Key Lab of Optoelectronic Technology&System Ministry of Education, Chongqing University, Chongqing 400030, China)[1]
(School of Electronic Information Engineering, Chongqing University of Science and Technology, Chongqing 400050, China)[2]

Abstract Image data mining is an active research area in databases, graphics, images and information technology. Recently, many researchers have employed algorithm study for image mining, and some improved algorithms were proposed about it. In the practice, the popularly used IAGA is easily stagnant, resulting in partial convergence. To solve those problems, the paper studied the traditional two-population genetic algorithm (AGA algorithm) to extract association rules, improved genetic operators, designed a new self-adaptable genetic algorithm (NAGA algorithm) based on improved genetic operator. Lastly, the NAGA algorithm was applied successfully to mine remote sensing images, and those image association rules were distilled by the new methods. The results presented in the paper demonstrate that the new genetic algorithm is a stable, superior algorithm.

Keywords Image mining, Self-adaptable genetic algorithm, Remote image

1 引言

随着数字图像的广泛应用,挖掘非标准化和多媒体数据是数据挖掘的发展趋势之一。图像挖掘为多媒体的数据挖掘提供了很好的方法和技术。图像挖掘(Image Mining,简称IM)是在图像数据库中抽取隐含的、先前未知的、潜在有用的知识;图像数据关系的非平凡过程,是集中了计算机视觉、图像处理、图像检索、数据挖掘、机器学习、模式识别、数据库和人工智能等技术的多学科交叉的研究领域[1]。由于图像数据挖掘的复杂性,目前的理论和技术还远未成熟,尚处于初始阶段。

与此同时,数据挖掘作为解决"信息爆炸,知识贫乏"的手段一经出现,就得到了迅速发展,大批学者把目光集中到这个领域,做了大量的工作,成绩斐然。很多理论技术在实际中得到应用,但大多数工作集中在事务型(商业型)数据库上。近年来图像数据挖掘才得到了重视,例如,美国计算机协会(ACM)自2000年起,每年举行一次多媒体数据挖掘国际学术会议;Simon Fraser大学开发的Multimedia Miner可以使用三维可视化技术清楚地显示关联规则。Krzysztof Kopersk等研制开发了用于遥感图像数据挖掘和统计分析的VisiMine(GeoBrowse)系统。目前IM与其他数据(如地面和天气信息)结合已初步应用于科学研究、农业、医学、生物、气象、资源勘探、自然灾害预测、监测与防灾减灾等领域。例如,Osmar R等[2]将数据挖掘应用于医学领域,提高了对乳腺疾病识别的准确率,为医护工作者提供有价值的参考;Eklund、Huang、邸凯昌、布和敖斯尔等,分别将遥感图像数据挖掘用于盐碱地分类、湿地分析、土地利用分类、土壤盐度分析,提高了分类精度,取得了很好的效果。Mitsuru Kakimoto等以人的脑功能图像为挖掘数据,得到了讲话者的手指活动与讲话动作的关系规则。Rie Honda等将时间序列关联规则用于卫

到稿日期:2008-07-22 返修日期:2008-12-30 本文受国家科技支撑计划(2007BAG06B06)资助。
雷 亮(1973-),男,博士研究生,主要研究方向为图像处理、模式识别和数据库技术等,E-mail:cqlei.l@163.com;汪同庆(1949-),男,博士生导师,教授,主要研究方向为光机电一体化技术、计算机自动识别技术及应用的研究等;彭 军(1970-),男,博士,教授,主要研究方向为图像处理、混沌密码学研究等;杨 波(1973-),男,博士后,副教授,主要研究方向为图像处理、图像识别等的研究和应用。

星云图分类，效果很好。Qin Ding 等利用研究眼遥感图像中不同波段之间的关联关系，得到了波段反射值与农业产量之间的关系，为农业增产和产量预测提供依据。Thanapat Kangkachit 等研究了数字图像中实体之间和特征与实体之间的关联规则问题[3,4]。

对图像序列中关联规则的提取，一般采用遗传算法。传统双种群遗传算法采用固定的遗传算子，不利于保证两个种群间的群体差异性。本文将使用一种改进的双种群遗传算法，使遗传算子能自适应地调整，提高算法的收敛性。挖掘图像序列之间的强关联规则，通过关联规则的提取，实现图像数据挖掘。

2 传统的双种群遗传算法

标准遗传算法是针对一个宏观的种群进行选择、交叉、变异 3 种操作，类似于人类进化过程。一群人随着时间的推移不断地进化，具备越来越多的优良品质。然而，由于他们的生长、演化、环境和原始祖先的局限性，经过相当长的时间后，将逐渐进化到某些特征处于相对优势的状态。当一个种群进化到这种状态，称之为平衡态，这个种群的特性就不会再有很大的变化。双种群遗传算法是一种并行遗传算法，它使用多种群同时进化，并交换种群之间优秀个体所携带的遗传信息，以打破种群内的平衡态，达到更高的平衡态，跳出局部最优。

双种群遗传算法在提取关联规则时，首先建立两个遗传算法群体，即种群 A 和种群 B，分别独立地进行自然选择、染色体交叉、染色体变异操作，且交叉概率、变异概率固定。当每一代运行结束以后，产生一个随机数 num，分别从 A，B 中选出最优染色体和 num 个染色体进行杂交，以打破平衡态。

这种双种群遗传算法求解过程如下(分别针对两个不同种群同时进行)：

构造染色体

产生初始种群

For(进化代数=0；进化代数<=进化最大代数；进化代数++)

{可行化过程

求出各染色体对应的适应值

自然选择

染色体交叉

染色体变异种群交叉

}

根据这种双种群遗传算法的求解过程，我们将其用于实现在图像挖掘中关联规则的提取的步骤归纳为 8 个主要步骤：

(1)构造染色体，产生初始种群。在遗传算法中，染色体的表现形式通过编码机制来实现。编码机制是遗传算法的关键步骤。

同时，设置关联规则的支持度阈值 S 和可信度阈值 C。

(2)可行化过程。计算所有规则的支持度 S' 和可信度 C'。

(3)适应度计算。将计算出来的规则的支持度 S' 与给定支持度阈值 S 之商作为适应度函数，即 $f(A)=S'/S$。根据适应度值对个体进行筛选：如果 $f(A)>1$，则保留该规则进入下一代，否则删除，并计算保留下来的个体数 M。如果 $M<N$，则随机生成 $(N-M)$ 个个体。

(4)判断停止进化的条件。判断迭代的代数是否为要求代数。若是，停止进化，选适应度值 $f(A)>1$ 的规则输入，否则继续执行(5)。

(5)自然选择。在自然选择环节，既要保证最优个体可以生存到下一代，给适应度较大的个体较大的机会进入下一代，又要避免个体间因适应值不同而被选入下一代的机会悬殊，因此采用比例选择与精华模型相结合的选择策略。即将每代种群 n 个染色体按 $f(A)$ 值排序，将值最大的染色体复制一个，直接进入下一代。下一代种群中剩下的 $n-1$ 个染色体用轮盘选择法产生。

(6)染色体交叉重组。对(5)所产生的新种群，按选择概率 p_c 选择个体对，进行交叉重组，共进行 $n/2$ 次。文献[5]表明交换概率 $p_c=0.6\sim0.8$ 之间时，进化性能较好，交叉规则采用 PMX 法。

(7)染色体变异。染色体按照变异概率 p_m 进行变异操作。变异决定了物种的多样性，因此变异是可以使没有被选择对象被选中的重要途径。

(8)种群交叉。将两个种群中的最优解取出，再在每个种群中随机选取 num 个染色体，将这 $num+1$ 个染色体互换，进入对方种群。

3 改进的双种群遗传算法

在实际应用传统的双种群遗传算法(AGA)的过程中，由于交叉概率、变异概率固定不变，容易出现过早收敛而仅得到局部最优解的现象[6]。因此需要对算法进行改进，让交叉概率和变异概率能够自适应调节，使算法寻优速度加快，而且不易陷入局部最优解。

文献[7-10]提出了一种改进的自适应遗传算法(IAGA)，其交叉概率 P_c 和变异概率 P_m 分别如式(1)和式(2)：

$$P_c=\begin{cases}P_{c1}-\dfrac{(P_{c1}-P_{c2})(f'-f_{avg})}{f_{max}-f_{avg}}, & f'\geqslant f_{avg}\\ P_{c1'}, & f'\leqslant f_{avg}\end{cases} \tag{1}$$

$$P_m=\begin{cases}P_{m1}-\dfrac{(P_{m1}-P_{m2})(f_{max}-f)}{f_{max}-f_{avg}}, & f\geqslant f_{avg}\\ P_{m1'}, & f\leqslant f_{avg}\end{cases} \tag{2}$$

其中，f_{max} 代表群体中最大适应度；f_{avg} 代表每代群体的平均适应度；f' 代表要交叉的两个个体中较大适应度；f 代表要变异个体的适应度。

但是，运用本算法，较差个体的变异能力较低，容易产生停滞现象。而精英保留策略虽然起到了保护和推广优秀个体的作用，但是其个体数目不宜过大，否则会使种群进化陷入停滞不前，造成局部收敛。

为此，本文对交叉概率 P_c 和变异概率 P_m 进行了改进，提出了一种新的自适应遗传算法(NAGA)。为了更好地描述这种算法，我们引入 N_1 和 N_2。

$$N_1=\frac{f_{max}-\bar{f}}{f_{max}-f_{min}+\epsilon} \tag{3}$$

$$N_2=\frac{\bar{f}-f_{min}}{f_{max}-f_{min}+\epsilon} \tag{4}$$

利用 N_1 和 N_2，计算交叉概率 P_c 和变异概率 P_m。

$$p_c=\begin{cases}0.9 & N_1(z)\leqslant N_1(z-1),z\in[1,gen]\\ 0.5 & N_1(z)>N_1(z-1),z\in[1,gen]\end{cases} \tag{5}$$

$$p_m=\begin{cases}0.3 & N_2(z)\leqslant N_2(z-1),z\in[1,gen]\\0.01 & N_2(z)>N_2(z-1),z\in[1,gen]\end{cases}\quad(6)$$

在式(3)和式(4)中 f_{max} 代表群体中最大适应度；f_{min} 代表群体中最小适应度；$\bar{f}$ 表示本代群体的平均适应度；ϵ 是一个无穷小正数，主要是为了防止分母为 0。

在式(5)和式(6)中，z 表示遗传的当前代，$z-1$ 表示上一代；gen 表示进化的总代数。

改进后的交叉概率和变异概率不但能够随适应度自动改变，而且使种群中最大适应度值的个体的交叉概率和变异概率不为零，这就相应地提高了种群中表现优良的个体的交叉概率和变异概率，使得它们不会处于一种近似停滞不前的状态，从而使算法跳出局部最优解。

4 应用实例

本实例的图像数据来源于文献[11]提供的遥感图像数据。

(1)特征提取

在对图像进行预处理、特征提取后，得到我们关心的 4 个属性特征数据：植被覆盖度(Coverage，简称 C)、坡度(Slope，简称 S)、耕地分布(Farm，简称 F)和土壤侵蚀强度(Eros，简称 E)。这 4 个属性特征数据的属性值分布范围如表 1 所列。

表 1 各维属性取值范围

	C	S	F	E
范围	[0-100]	[0-90]	[0,1]	[1-6]

(2)属性划分

对上述 4 个属性分割之后，得到 l0 个原始项，如表 2 所列。

表 2 属性分割后得到的 10 个原始项

	项名	范围	含义
1	C_low	0～30%	植被覆盖率低
2	C_mid	31%～60%	植被覆盖率中等
3	C_high	61%～100%	植被覆盖率高
4	S_low	0^0～25^0	坡度小
5	S_mid	26^0～50^0	坡度中等
6	S_high	51^0～90^0	坡度大
7	F_true	1	耕地
8	F_false	0	非耕地
9	E_low	1～3	轻度侵蚀
10	E_high	4～6	强侵蚀

(3)关联规则挖掘

根据分割后得到的 10 个原始项，运用上面介绍的改进的双种群遗传算法得到频繁项目集，然后根据支持度和可信度的定义(最小支持度取 3%、最小可信度取 65%)，得到一些强关联规则，如表 3 所列。

表 3 关联规则

序号	条件 1	条件 2	条件 3	结果	支持度(%)	可信度(%)
1	C_mid			E_low	13.2	95.6
2	F_false			E_low	62.1	67.3
3	F_true			E_high	3.8	70.1
4	S_low			E_low	21.5	82.7
5	C_low	S_low		E_low	19.3	82.4
6	C_low	F_true		E_high	4.0	69.4
7	C_low	S_high		E_high	4.3	65.2
8	F_false	C_mid		E_low	11.9	92.1
9	F_false	S_low		E_low	18.2	84.7
10	S_low	C_mid		E_low	3.3	97.9
11	S_mid	C_mid		E_low	7.6	95.8
12	S_mid	F_true		E_high	3.1	80.3
13	F_false	C_low	S_high	E_high	4.3	65.7
14	F_false	C_low	S_low	E_low	12.4	70.1
15	F_false	S_low	C_mid	E_low	3.1	83.2
16	F_false	S_mid	C_mid	E_low	8.3	94.5

通过挖掘这些强关联规则，我们可以得到一些重要的、可用于决策的知识：

(1)耕地一般导致强度侵蚀(规则 3,6,12)，非耕地一般为轻度侵蚀(规则 2,8,9,14,15,16)。

(2)低坡度的地区大部分侵蚀强度较轻(规则 4,5,10,14,15)。

(3)植被覆盖率高的一般不会产生强度侵蚀(规则 1,8,10,11,15,16)。

(4)非耕地只有在低植被覆盖和高坡度的情况下才会与高侵蚀产生强关联(规则 13)。

实验证明，本文提出的这种新的染色体交叉重组和变异策略，能够满足遗传算法中对于种群多样性的需求。同时，这种 NAGA 算法在收敛性上改善了现有的一些自适应遗传算法的性能，其计算效果比较理想，是一种有效的、稳定的、十分实用的算法。

结束语 图像数据挖掘是目前国际上数据库、图形图像技术和信息决策领域最前沿的研究方向之一。本文主要介绍了图像挖掘中涉及的相关概念，详细介绍了 AGA 算法实现关联规则提取的求解过程，对 AGA 算法中存在的不足进行了分析；针对 IAGA 算法在实际应用中存在的问题，提出了 NAGA 算法，并将其成功应用到遥感图像挖掘，提取出了植被覆盖度、坡度、耕地分布和土壤侵蚀强度之间的强关联规则。实验证明，运用这种改进的遗传算法，不会处于一种近似停滞不前的状态，从而使算法跳出局部最优解，无论在收敛性还是在稳定性上均改善了 IAGA 算法的性能。

参 考 文 献

[1] Zhang J, Hsu W, Lee M L. Image mining issues, frameworks, and techniques[C]// Proceedings of the Second International Workshop on Multimedia Data Mining(MDM/KDD 2001). San Francisco, CA, USA, 2001: 13-20

[2] 孙庆先，方涛，郭达志. 图像数据挖掘中的关联规则[J]. 计算机工程，2006，3：50

[3] Stanchev P. Using Image Mining for Image Retrieval [C]// IASTED. Computer Science and Technology. Proceedings of the IASTED International Conference on Computer Science and Technology. Cancun, Mexico, May 2003. Calgery-Alberta, T3B OM6, Canada: Int. Assoc. of Science and Technology for Development, 2003: 214-218

[4] Gibson S, et al. Intelligent mining in image databases, with applications to satellite imaging and to web search[M]. Data Mining and Computational Intelligence. Heidelberg, Germany: Physica-Verlag GmbH, 2001: 309-336

[5] (加) Han Jiawei, Kamber M. 数据挖掘概念与技术(第 2 版)[M]. 范明，孟小峰，译. 北京：机械工业出版社，2007：151-154

[6] 曾凡超，朱征宇，邓欣，等. 车辆路径问题的改进的双种群遗传算法[J]. 计算机工程与设计，2007，28(20)：4999

(下转第 247 页)

本文基于增量式概念格算法，选取 70 个对象，按照表 1 给出的多值属性背景表，生成的三维概念格如图 2 所示。取最小支持度阈值 minsupport=0.1，最小置信度阈值 minconfidence=0.8，挖掘出岩性空间分布规律空间邻近规则 250 条。由于篇幅所限，仅列出其中 10 条规则，如表 2 所列。

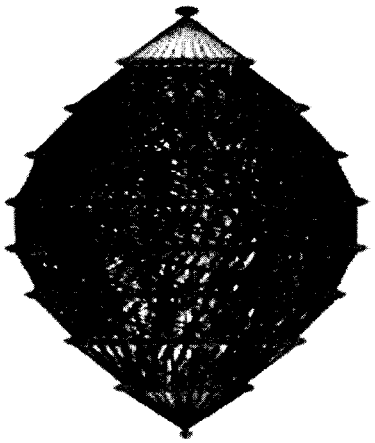

图 2　生成的概念格

表 2　岩性空间分布规律

规则编号	规则	支持度	置信度
1	右上方位置 T_1b^2→左下方和右下方位置均为 T_1b^1	0.17	0.92
2	右下方位置 T_1j^2→右上方位置 T_1j^3	0.14	0.83
3	正上方位置 T_1b^2→右下方位置 T_1b^1	0.18	0.92
4	正上方位置 T_1b^2→正下方位置 T_1b^1	0.18	0.92
5	中心位置 T_1j^2，右下方 T_1j^2→右上方 T_1j^3	0.12	0.81
6	中心位置 T_1j^2，左下方位置 T_1j^3→左上方位置 T_1j^2，正下方 T_1j^3	0.14	0.9
7	左下方和正下方位置均为 T_1j^2→右上方和正上方位置均为 T_1j^3	0.14	0.9
8	中心和左下方位置均为 T_1j^3→右上方和正上方位置均为 T_1b^1	0.12	0.9
9	中心位置 T_1j^3，右上方位置 T_1b^1→左下方位置 T_1j^3，正上方位置 T_1b^1	0.12	0.9
10	左下方位置 T_1j^3，正上方位置 T_1b^1→右上方位置 T_1b^1，正下方位置 T_1j^3	0.12	0.81

由于篇幅所限，本文仅取挖掘出的规则中的几条进行解释。

规则：右下方位置 T_1j^2⇒右上方位置 T_1j^3。该规则的含义是：若南东方向的地层是嘉陵江组二段，则可推断北东方向的地层必为嘉陵江组三段；

规则：正上方位置 T_1b^2，正左方位置 T_1b^1⇒右上方位置 T_1b^2，右下方位置 T_1b^1，左下方位置 T_1b^1，正下方位置 T_1b^1。该规则的含义是：若正北方向的地层是巴东组二段，正西方向的地层是巴东组一段，则可推断北东方向的地层为巴东组二段，南东方向、南西方向、正南方向的地层均为巴东组一段。

可见挖掘出的规则很好地说明了地层的空间临近关系和空间分布规律，并且与实际情况相符，从而为岩性的智能解译提供了重要的判据。

结束语　本文针对三峡库区地形复杂、地质灾害频繁、土壤植被发育的情况分析和挖掘出岩性空间分布的规律；通过将遥感影像与地质图叠加，选择各地层边缘的像素点，分析其中心、正上方、正下方、正左方、正右方、左上方、右上方、左下方、右下方 9 个方向上的岩性；基于概念格算法和规则提取，挖掘出三峡库区嘉陵江组二段 T_1j^2、嘉陵江组三段 T_1j^3、巴东组一段 T_2b^1、巴东组二段 T_2b^2、大冶组 T_1d 等地层的岩性邻近规则和空间分布规律。通过对岩性空间分布规律的研究，能够为三峡库区岩性的智能分类和解译提供重要的信息和先验知识。下一步工作将研究挖掘出的规则应用于岩性的智能分类和解译。

参 考 文 献

[1] Hunt G R. Spectroscopic Properties of Rocks and Minerals in Handbook of Physical Properties of Rocks[M]. Volume I. Boca Raton: CRC Press, 1982

[2] Clark R N, Roush T L. Reflectance Spectroscopy: Quantitative Analysis Techniques for Remote Sensing Applications[J]. Journal of Geophysical Research, 1984, 89(B7): 6329-6340

[3] Rowan L C, Simpson C J, Mars J C. Hyperspectral Analysis of the Ultramafic Complex and Adjacent Lithologies at MordNT [J]. Austrlia. Remote Sensing of Enviroment, 2004, 91(3): 419-431

[4] 赵建华，杨树锋，陈汉林. 基于分形纹理的遥感图像岩性识别方法[J]. 遥感信息，2004(2)：1-4

[5] 黄颖端，李培军. 基于地统计学的图像纹理在岩性分类中的应用[J]. 国土资源遥感，2003，3：45-49

[6] 马超飞，马建文. 应用多源数据提取高植被覆盖地区岩性信息-以湖南衬阳地区为例[J]. 地质科学，2002，37(3)：365-371

[7] 库向阳，薛惠锋，雷学武，等. 基于分类规则挖掘的遥感影像分类研究[J]. 遥感学报，2006，10(3)：332-338

[8] 孙庆先，方涛，郭达志，等. 空间数据挖掘技术中的划区效应及在矿山中的应用[J]. 煤炭学报，2007，32(8)：804-807

[9] 秦昆. 基于形式概念分析的图像数据挖掘研究[D]. 武汉：武汉大学，2004

(上接第 205 页)

[7] 张玲，刘勇，何伟. 自适应遗传算法在车牌定位中的应用[J]. 计算机应用，2008，28(1)：185

[8] Gao Li, Dai Shangping, et al. Using Genetic Algorithm for Data Mining Optimization in an Image Database[C]// Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). 2007

[9] Koperski K, Han J. Discovery of spatial association rules in geographic information databases[C]// Proc. of International Symposium on Advance in Spatial Databases, SSD, LNCS. vol. 951, Springer Verlag, 1995: 47-66

[10] Chen G, Wei Q. Fuzzy association rules and the extended mining algorithms[J]. Information Sciences, 2002, 147: 201-228

[11] 马超飞，刘建强. 遥感图像多维量化关联规则挖掘[J]. 遥感技术与应用，2003，18(4)：244-246