基于遗传算法的图像特征选择

陈卫东,刘素华
CHEN Wei-dong,LIU Su-hua

河南工业大学 信息科学与工程学院,郑州 450001
College of Information Science and Engineering,Henan University of Technology,Zhengzhou 450001,China
E-mail:chenweid@haut.edu.cn

CHEN Wei-dong,LIU Su-hua.Image features selection based on genetic algorithm.Computer Engineering and Applications, 2007,43(28):78-80.

Abstract: Aimed to the problem that original feature is mass and redundancy in pattern recognition,a method of feature optimal based on genetic algorithm is proposed.This paper describes the general idea of genetic algorithm.Then explains and designs a fitness function and genetic operators.Simulations results show that this method has good performance in both the quality of obtained feature subset and efficiency.

Key words: feature selection;Genetic Algorithm(GA);pattern recognition;store-product pest

摘 要:针对模式识别时,提取的特征参数量大而又有冗余的现象,提出了基于遗传算法的特征选择方法。介绍了遗传算法的基本原理,阐述并设计了适应度函数和遗传算子。仿真实验表明,该方法在求解的效率和解的质量方面都达到了令人满意的效果。

关键词:特征选择;遗传算法;模式识别;仓储物害虫
文章编号:1002-8331(2007)28-0078-03 **文献标识码**:A **中图分类号**:TP391

1 引言

昆虫分类学是一门古老的学科,昆虫研究人员一直在努力寻找一种科学、快速、准确的虫种鉴定方法。近年来,随着计算机技术的不断发展,计算机数字图像处理技术、模式识别等方法取得了突破性进展, 利用计算机技术对昆虫进行自动识别,是昆虫研究领域希望解决的重要课题之一。《谷物害虫实时监测与分类识别系统》的研究就是利用数字图像处理技术和模式识别技术,对获取的昆虫图像进行预处理后,再提取昆虫图像的特征,然后送入分类器[1]中进行识别,即建立一种无需人工干涉的昆虫种类模式识别系统。

特征是决定相似性与分类的关键[2]。由于用计算机识别昆虫, 故从昆虫的图像上提取了大量的原始特征以期提高识别率。但从提取方法上看,许多特征不是独立的,即这些特征具有冗余度,影响模式识别的速度和准确性。所以,需要对原始特征进行选择,丢掉那些模棱两可、不易判别或相关性强的特征,这个特征选择问题,实质是一个组合优化问题。

常规的优化算法,如解析法,只能得到局部最优而非全局最优解,且要求目标函数连续及可微;枚举法虽然克服了这些缺点,但计算效率太低。即使很著名的动态规划法,也会遇到"指数爆炸"问题,它对于中等规模和适度复杂性的问题,也常常无能为力。而遗传算法是目前比较理想的优化方法,由于它易于跳出局部次优解和无需建立优化方程,且具有良好的隐并行性和稳健性,已成为信息科学、计算机科学和人工智能等诸多学科所关注的焦点。

2 遗传算法

遗传算法[3](Genetic Algorithm,GA)是建立在自然选择和遗传变异基础上的自适应概率性搜索算法,在该算法中,个体是二进制字符串编码,每一编码字符串为一候选解,这种个体有多个,即有一群候选解。个体是主要的进化对象,象生物进化一样有繁殖、交叉和突变三种现象。在每一代中,保持一定数目 M 为定值的解群,经过对各解的适应度值计算,使解群中各解得到评价。从进化的角度看,新一代的群体对环境的平均适应度比双亲的一代要高。

3 仓储物害虫图像特征优化设计

3.1 个体编码

在特征选择[4]问题中,个体的编码方式采用二进制,二进制位串表达简单,操作方便,可代表较广范围的不同信息。若原始特征有 15 个,则个体的长度 L=15,个体的每一个基因对应相应次序的特征,即当个体中的某一个基因为"1"时,表示该基因对应的特征项被选用;反之,为"0"时,表示该特征项未被选用。例如,个体 110010000101100 表示第 1、第 2、第 5、第 10、第 12 个、第 13 个特征项被选用。

3.2 初始种群的生成

利用随机函数产生 M 个个体组成初始群体, 也可以根据

基金项目:河南省科技公关计划(the Key Technologies R&D Program of Henan (Province),China under Grant No.0624220006)。
作者简介:陈卫东(1972-),男,讲师,研究方向:信号处理与计算机应用;刘素华(1972-),女,副教授,研究方向:人工智能和软件工程。

专家经验选出 M 个个体作为初始群体,后者寻优速度更快,收敛到最优个体的时间短。一般来说,初始群体素质都很差,遗传算法的任务就是从这些群体出发,模拟进化过程,择优汰劣,最后得出非常优秀的群体,其中的佼佼者就是问题的解。

3.3 评估函数(适应度函数)的确定

如何将遗传算法中的不断进化的个体与现实问题中的优劣选择相联系是该算法成功的关键。一般的方法是构造一个与现实问题相联系的评估函数。这个评估函数的作用类似于自然界中生物适应环境能力的度量,用它来进行优胜劣汰。

特征选择的目的是找出分类能力最强的特征组合,因此需要一个定量准则来衡量各个体对应的特征组合的分类能力[5]。各样本之所以能分开,是因为它们位于特征空间的不同区域。显然,如果不同类别之间的距离越大、同一类别内各样本相互间的距离越小,则分类效果越好。所以,我们将群体中个体的适应度函数设为:

$$J(x)=S_b-S_w \tag{1}$$

其中,S_b 类间距离,S_w 表示类内距离。它们的计算方法如下:

两模式之间距离的表示方法有多种,如海明距离、欧氏距离等,本文采用如下的欧氏距离:

$$N(A,B)=1-\frac{1}{\sqrt{n}}\sqrt{\sum_{i=1}^{n}(\boldsymbol{\mu}_A(x_i)-\boldsymbol{\mu}_B(x_i))^2} \tag{2}$$

其中,$\boldsymbol{\mu}_A(x_i)$ $\boldsymbol{\mu}_B(x_i)$分别是代表模式 A 和模式 B 的特征向量.(注意,这里用的特征向量均是归一化后的特征值,而不是用各种算法直接提取的特征值)。

在计算类间距离时,$\boldsymbol{\mu}_A(x_i)$ $\boldsymbol{\mu}_B(x_i)$代表 A 和 B 两类类中心的均值向量,均值向量可通过式(3)求出:

$$c_i=\frac{1}{n_i}\sum_{x\in w_i}x \quad i=1,2,\cdots,k \tag{3}$$

其中:$w_1,w_2,\cdots,w_k$ 为 k 个类别,$\boldsymbol{c}_i$ 为第 i 个类别的类中心特征向量,在 w_i 类别中有 n_i 个模式[1]。

对任意两类均要用式(2)计算类间距离,然后相加即得 S_b。

计算类内距离时,$\boldsymbol{\mu}_A(x_i)$、$\boldsymbol{\mu}_B(x_i)$是同一类内的模式 A 和 B 的特征向量。对每一类内的各模式间均要计算类内距离,然后各类的类内距离相加即得 S_w。

利用式(1)、式(2)可以得出每个个体的适应度 ji,依据适应度 ji 对群体进行下面的遗传操作。

3.4 遗传操作设计

3.4.1 繁殖(选择)算子

繁殖的基础是适应度值,适应度高的个体在下一代具有较多的繁殖机会,从而有较多的后代,而适应度较低的个体则产生较少的后代,最后逐渐被淘汰。

如何选择适应度值高的个体作为繁殖下一代的双亲呢?目前已用许多方法可以达到这个目的,如比例选择、择优保存和排序选择等。本文采用通用的方法——类似轮盘赌的比例选择法。

首先,根据每个个体的适应度 ji,计算群体的总适应度 $J=\sum_{i=1}^{M}j_i$,然后用式(4)和式(5)计算每一个个体的选择概率和累计概率。

$$P_i=\frac{j_i}{J} \qquad i=1,2,\cdots,M \tag{4}$$

$$q_i=\sum_{j=1}^{i}P_j \quad i=1,2,\cdots,M \tag{5}$$

在轮盘赌上按各 P_i 大小分成不等的扇形区域,再将旋转轮盘赌 M 次即可选出 M 个个体来。在计算机上实现的步骤为:产生[0,1]之间的随机数 r,若 $r<q_1$,则第一个个体 v_1 入选,否则依次选 v_i($2\leqslant i\leqslant m$),使满足($q_{i-1}<r\leqslant q_i$)。

非常显然,适应度越大的个体在轮盘上占的面积就越大,相应入选概率就越高。(这种选择是放回去的选择,即 j 大的个体可再次选中,而较差的个体被淘汰)。这就模拟了自然界适者生存法则。

3.4.2 交叉算子

在生物进化过程中,两个染色体通过交配而重组,形成新的染色体,从而产生新的个体或物种,在遗传算法中模仿这个环节的,即交叉算子。交叉算子的设计分两步,第一步是将新繁殖产生的染色体随机两两配对;第二步是随机地选择交叉点和交叉长度。交叉算子有单点较差、双点交叉和多点交叉。

将随机两两配对的父代染色体按交叉概率 P_c 进行如下的多点交叉操作:

设置 n 个多点交叉位,在[1,L-1]的范围内(L 为染色体的长度),随机生成这 n 个交叉点,在交叉点之间间续地相互交换,从而生成两个新染色体作为下一代的成员,这就实现了交叉算子,考虑如下长度为 15 的染色体:

父染色体 1:100110010100011

父染色体 2:001101001100101

交叉位置:3,8,12

交叉后的新染色体为:

子染色体 1:101 1 0 1 0 101001 0 1

子染色体 2:000 1 1 0 0 011000 1 1

多点交叉的思想源于控制染色体特定行为的信息部分无须包含于近邻的子串中,同时,多点交叉的破坏性可以促进解空间的搜索,而不是促进过早的收敛,因此使得搜索更健壮。

3.4.3 变异算子

尽管繁殖和交叉操作很重要,在遗传算法中是第一位的,但不能保证不会遗漏一些重要遗传信息。在遗传算法中,变异算子用来防止这种不可弥补的遗漏。变异就是某个个体的某一位偶然地(概率很小的)随机改变,即在某些特定位置上简单地把 1 变成 0 或反之。变异是沿着个体空间的随机移动。当它有节制地和交叉算子一起使用时,它就是一种防过渡成熟而丢失重要信息的保险策略。

若取变异概率为 P_m,则群体中可能变异的位数的期望值为 $P_m\times m\times M$(m 为染色体长度,M 为群体大小),每一位以等概率变异,具体步骤为:

(1)对每一染色体中的每一位产生[0,1]间的随机数 r,若 $r<P_m$,则该位变异;

(2)实施变异操作:即原来为 0 的变为 1,原来为 1 的变为 0。

至此,如果新染色体数达 M 个,则说明已形成一个新的群体,进行下一代的遗传;否则继续本代的繁殖、交叉、变异操作。

3.5 终止条件

每经过一代遗传,问题的解便朝着最优解的方向前进一步,只要这个过程一直进行下去,最终将走向全局最优解,而每

一步的操作却是很简单的。但由于遗传算法没有利用目标函数的梯度等信息,从而无法用传统的方法来判定算法的收敛与否以终止遗传过程。常用的方法是通过控制参数来实现算法的终止,如运算到指定的最大代数,到达后即停止;或者当相邻几代的平均适应度差值小于某个阈值 ε 时就可以终止遗传操作,也就找到了最优特征组合。

4 特征选择的遗传算法描述

步骤 1 输入控制参数;

步骤 2 随机生成初始群体 G_t, 计算群体中各个体的适应度、选择概率、群体的平均适应度(初始时繁殖代数 T=0,表示第 0 代);

步骤 3 应用遗传算子(即繁殖、交叉、变异)产生新一代群体 G_{t+1};

步骤 4 对新一代群体进行评估,即计算各个体的适应度、选择概率、群体的平均适应度;遗传代数 T 增 1。

步骤 5 判断繁殖代数 T 大于规定代数 (或运行时间大于规定时间)吗?若是,转步骤 6;否则,转步骤 3。

步骤 6 按适应度值将结果代个体排序, 选择适应度值最高的作为最优解。

5 结果与讨论

本次研究中所采用的实验对象分为训练样本集和验证样本集两部分训练样本集中的昆虫图像分别为 4 个种类,16 幅(如图 1),验证样本集中的昆虫图像为 5 个种类,30 幅。

图 1 部分害虫样本图像

上述昆虫图像均是通过数码照相机在相同的实验条件下获得,实验条件如下:摄像头焦距为 9 cm,自然光照射,目标背景为白色每幅昆虫限定尺寸为 148×256 像素点阵。

首先,对采集到的害虫图像进行增强、去噪等预处理然后,对预处理后的图像提取其一阶灰度值统计量特征 6 个(均值、方差、偏度、峰值、能量和熵);灰值游程距阵纹理特征 16 个(0°、45°、90°、135°四个不同方向的短游程长度、长游程长度、灰度值的不均匀度量和游程长度的百分率);几何特征 3 个(区域面积、圆度和长宽比),共 25 个特征。

因上述 25 个特征的量纲不同,它们之间不具有可比性,本文采用如下式(6)将数据压缩到[0,1]之间。

$$\bar{\alpha}_i=\frac{\alpha_i}{2\alpha} \quad i=1,2,\cdots,25 \qquad (6)$$

其中:$\alpha_1,\alpha_2,\cdots,\alpha_{25}$ 是样本的 25 个特征值,α 是 $\alpha_1,\alpha_2,\cdots,\alpha_{25}$ 的均值。

用遗传算法进行优化,实验中控制参数取:

M=80(种群规模);

P_c=0.75(交叉概率);

P_m=0.001(变异概率,即每一个千位的传送中,只变异一位);

T=600(遗传代数,终止遗传操作的条件。根据这里的问题,600 代以后算法基本收敛,再多的遗传代数,其所得的解没有多大的变化,如图 2)。

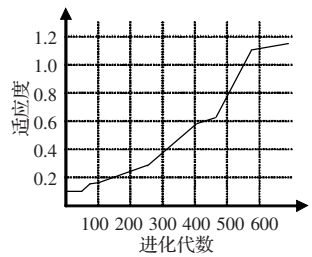

图 2 遗传算法收敛曲线

采用 VC 语言,在普通微机上进行实验,用 25 个原始特征和优化后的 17 个特征实验,实验结果比较如表 1 所示。

表 1 实验结果比较

	识别率	误识率	拒识率	学习时间
25 个特征	86.33%	10%	3.66%	约 1 小时 20 分钟
17 个特征	83.33%	12%	4.66%	约 30 分钟

表面上看,利用遗传算法对特征优化后,分类器的正确识别率反而下降了,事实上,这是由于如下的原因造成的:(1)客观条件的限制——仓储物昆虫样本太少以及样本不规范造成的;(2)遗传算法用于特征优化还有待于改进,已有人在这方面进行研究[8]。只要拥有足够的样本,先利用遗传算法进行特征优化,然后设计分类器,可使学习时间和分类时间大大下降。

当然,遗传算法还存在一些问题,如群体的大小,交叉、变异概率均是实验参数,但很难确定。同样,遗传算法不一定总是获得最优解,这一问题称为成熟前收敛,其主要发生于解的适应值停止提高,即适应度值达到平衡时,还没有达到最优解。现在已有文献解决这一问题。

6 结束语

在《谷物害虫实时监测与分类识别系统》的研究中,针对从图像直接提取的特征参数量大而有冗余的现象,本文利用遗传算法强大的并行全局寻优能力,剔除原始特征集合中的冗余特征,将生成的优化特征集用于分类器的训练,从而提高了系统的识别率。

虽然遗传算法已经产生了较好和有意义的结果,但是对于其原理的理解和应用还处于初期, 还有许多亟待完善之处,如遗传算法群体的大小、繁殖方式、变异概率的选择等。但可以相信,遗传算法为组合优化提供了一个独特的方法,它必将在更广的领域得到更广泛的应用。(收稿日期:2007 年 1 月)

参考文献:

[1] 刘素华,侯惠芳.基于模糊理论的仓储物害虫的模式识别分类研究[J].计算机工程与应用,2004,40(5):227-231.

[2] 边肇祺,张学工.模式识别[M].北京:清华大学出版社,2000.

[3] 陈国良.遗传算法及其应用[M].北京:人民邮电出版社,1996.

[4] Handels H,Ross Th.Feature selection for optimized skin tumor recognition using genetic algorithms[J].Artificial Intelligence in Medicine,1999,16:283-297.

[5] Rudolph G.Convergence analysis of canonical genetic algorithm[J].IEEE Trans on Neural Networks,1994,5(1):96-101.