文章编号:1001 - 9081(2006)10 - 2372 - 03

基于改进的 RS-GA 图像特征选择方法

张杰慧,何中市,黄丽琼
(重庆大学 计算机学院,重庆 400044)
(jiehui_zhang@163.com)

摘 要:针对目前图像识别中原始特征数量大、不相关特征多以及冗余等现象,提出了一种图像特征选择方法。将遗传算法(GA)与粗集(RS)思想有机结合进行特征选择,引入粗集中相关属性依赖度的定义,设计了适应度函数和遗传算子,以提高算法时间效率和获得最佳搜索结果,并将该特征选择方法应用于图像,实验表明,基于改进的 RS-GA 图像特征选择方法达到了较好的效果,并具有较高的算法效率。

关键词:粗集理论;属性约简;图像特征选择;遗传算法;相关属性依赖度
中图分类号: TP391.41 **文献标识码**:A

Image feature selection method based on improved RS-GA

ZHANG Jie-hui, HE Zhong-shi, HUANG Li-qiong
(*College of Computer Science, Chongqing University, Chongqing* 400044, *China*)

Abstract: With regard to the problem that original feature in image classification is mass and redundancy, a new image feature selection method was presented. This method combines the Rough Set(RS) theory with Genetic Algorithm(GA) properly to select feature. To improve the efficiency of this algorithm and get the optimal searching result, definition of relative attribute dependency of rough set theory was introduced, and fitness function and genetic operators were designed. Then, this proposed method was applied to image feature selection. Experimental results show that it has better performance and higher algorithm efficiency.

Key words: rough set theory; attribute reduction; image feature selection; genetic algorithm; relative attribute dependency

0 引言

在图像识别领域,“特征维数灾难”是一个普遍现象,特征维数高导致了学习算法时间增加,有时甚至使识别率下降,因此对于图像识别特征选择是一个至关重要的问题。特征选择是从一组数量为 N 的原始特征中选出数量为 $M(M < N)$ 的一组最优特征。目前特征选择方法主要是一些传统的特征选择方法[1,2]。

1982 年,波兰华沙理工大学 Z. Pawlak 教授等提出了用粗集理论(Rough Sets,RS)研究不完整数据、不精确知识的表达、学习、归纳等方法[3]。粗糙集的核心内容是属性重要性的度量和属性约简。其中,约简是应用粗集理论的基础,其内涵即为去掉多余的属性,因此,粗集非常适合于用来处理特征选择的问题,其中特征选择就是原始属性的约简,即选出给定数据集中最理想的条件属性集,但是寻找最小属性约简是一个 NP 难问题[3,4]。

遗传算法(Genetic Algorithm,GA)是借鉴生物界自然选择和自然遗传机制的自适应搜索算法,具有良好的全局搜索性能,减少了限于局部最优解的风险,鲁棒性强,适用于并行处理,因此可以将遗传算法用于最优属性约简的搜索,目前已有一些用遗传算法进行粗集属性约简的研究[5,6],但效率方面还有待提高。

围绕以上两个问题,本文首先简单介绍了粗糙集理论,然后将遗传算法与粗集思想有机结合起来,针对粗集约简 NP 难引入相对属性依赖度的定义,设计了新的适应度函数和遗传算子,提出了一种改进的基于粗集和遗传算法的特征选择方法,最后结合图像对提取的图像特征进行试验和分析。

1 粗集理论概述

粗集将分类与知识联系在一起,认为知识源于人类以及其他物种的分类能力,并用等价关系形式化表示分类,下面给出一些相关定义[3,4]:

定义 1 一个信息系统可以表达为知识系统 S:

$S = < U, R, V, F >$

其中:U 是研究对象的集合;$R = C \cup D$ 是属性集合,子集 C 和 D 分别称为条件属性集和结果属性集;$V = \cup v_r, v_r$ 是属性 r 的值域;$f: U \times R \to V$ 定义了一个信息函数:即它指定 U 中每一个对象 x 的各属性值。若 $C \cap D = \varnothing$,则知识表达系统可以表示为一个决策表。

定义 2 对于属性集 $P \subset R$,对象 $X, Y \subset U$,P 上的不可分辨关系记为 $ind(P)$。

$ind(P) = \{(X,Y) \subset U: \forall a \in P, f(X,a) = f(Y,a)\}$

关系 $ind(P)$ 构成了 U 的一个划分,用 $U/ind(p)$ 表示,也可表示为 U/R。$U/ind(p)$ 中的任何元素称为等价类。

定义 3 对于任何 $X \subseteq U$ 和属性子集 $R \subseteq A$,X 的 R 下近似集:

收稿日期:2006 - 04 - 29;**修订日期**:2006 - 06 - 19 **基金项目**:国家自然科学基金资助项目(60173060)

作者简介:张杰慧(1982 -),女,湖南邵阳人,硕士研究生,主要研究方向:数字图像处理; 何中市(1965 -),男,四川广安人,教授,博士生导师,主要研究方向:自然语言处理、数字图像处理; 黄丽琼(1983 -),女,江西南昌人,硕士研究生,主要研究方向:自然语言处理.

$R\_(X) = \cup \{Yi \subset U \mid ind(R) : Yi \subset X\}$

X 的 R 上近似集:

$R^{-}(X) = \cup \{Yi \subset U \mid ind(R) : Yi \cap X \neq \varnothing\}$

定义 4　对 $P \subseteq R, Q \subseteq R$,$Q$ 的 P 正域记为 $POS_P(Q)$,定义为:

$POS_P(Q) = \cup P\_(Q), (X \in U \mid Q)$

是论域中所有通过用分类 $U \mid P$ 表达的知识能够确定地划入 $U \mid Q$ 的对象的集合。

定义 5　知识 Q 对知识 P 的依赖度 k,定义为:

$$k = \gamma_P(Q) = \frac{card(POS_P(Q))}{card(U)} \tag{1}$$

其中 $card(\cdot)$ 表示了该集合元素的数目。$\gamma_P(Q)$ 可以看做 Q 和 P 间依赖性的量度,也可解释为对对象分类的能力。

定义 6　在数据协调的决策表中,若存在属性集 $P \subset C$,称为 P 相对于决策属性集 D 的条件属性集 C 的约简,当且仅当:

1) $POS_P(D) = POS_C(D)$

2) 不存在 $r \in P$,使得:$POS_{P-\{r\}}(D) = POS_C(D)$

定义 7　所有 C 的属性约简的交称为 C 的核:

$core_D(C) = \cap red_D(C)$

2　基于 RS-GA 的特征选择算法设计

遗传算法中,参数编码、初始群体的设定、适应度函数的设计、遗传操作设计、控制参数设定五个要素组成了遗传算法的主要内容,其中适应度函数的设计是整个 GA 算法的核心步骤。本文结合粗集理论属性约简的思想对遗传算法的几个主要点进行研究设计,并在传统属性约简的思想上引入了相关属性依赖度的概念。

2.1　属性的相关依赖度

先假设决策表是相容的,即 $\forall t,s \in U$,如果 $f_D(t) \neq f_D(s)$,则 $\exists q \in C$,使 $f_q(t) \neq f_q(s)$。

定义 8　0 映射:对于 $P \subseteq C \cup D$,U 在 P 上的映射记为 $\prod_P(U)$,是 U 的一个子集,满足以下两点:1) 消除属性 $C \cup D - P$;2) 合并所有的难以识别的元组(决策表的行)[7]。

定义 9　相关依赖度:对于 $Q \subseteq C$,决策属性集 D 对属性集 Q 的相关依赖度记为 $\kappa_Q(D)$,定义为[6]:

$$\kappa_Q(D) = \frac{\mid \prod_Q(U) \mid}{\mid \prod_{Q \cup D}(U) \mid} \tag{2}$$

其中 $\mid \prod_X(U) \mid$ 是等价关系 $U/IND(X)$ 中等价类的数目。根据参考文献[7], 对于完全协调的决策表,$\kappa_Q(D)$ 与 k 在属性重要性的度量上是等价的。

2.2　具体实现

根据粗集理论,将粗集思想结合遗传算法,对遗传操作的设计如下:

1) 个体编码

在图像识别中,将图像特征做为决策表的条件属性,分类结果作为决策表的决策属性。

遗传思想中的个体编码采用二进制位,染色体的长度等于条件属性集的数量(即原始特征的数目),其中个体的每一个基因对应条件属性集的相应次序的属性(特征),当基因为“1”时表示对应的属性被选用,反之未被选用。因此,不同的染色体表示了不同条件属性的组合。

2) 初始种群的生成

一般在初始群体选择时就随机产生初始群体,但是根据定义 7,决策表的所有约简都包括决策表的核,利用核做为启发式信息,令核对应的属性编码为“1”,且在整个 GA 演化过程中保持不变,个体的其他基因位随机产生,连续产生这样的个体 PopSize 个。这样产生的初始群体接近问题解,能有效地减少遗传算法的搜索空间并能缩短求解时间[6]。其中决策表的核可利用可辨识矩阵较易求得[8],时间复杂度为 $O(m * n * \log_2 n)$。

3) 适应度函数的设计

特征选择的目的是降低特征维数的同时,找出分类能力最强的特征组合。因此,适应度函数设计的两大准则为:a)分类质量尽可能的高;b)染色体中值为 1 的位数尽可能低,即约简所包含的属性个数尽量小。

由于几个遗传算子都依赖于染色体的适应度值,因此适应度函数的设计目标,在很大程度上决定着迭代收敛的方向。为体现上述两个目标,本文使用相对属性依赖度 $\kappa_Q(D)$,定义了新的适应度函数:

$$score(P) = \frac{m - L_P}{m} + \kappa_P(D) \tag{3}$$

其中 P 为染色体所对应的条件属性集,m 为染色体的长度,$K_P(D)$ 代表属性集 P 对决策属性集的依赖度,L_P 代表该条染色体中值为 1 的基因数(即染色体所对应特征子集的特征数)。

该适应度函数将属性依赖度引入适应度函数,反映了属性的分类能力,同时又使用了 L_P 来控制了染色体的长度,同时体现着两大准则。

整个遗传算法的计算时间复杂度为 $O(t_{\max} * popsize *$ 适应性的计算复杂度),其中适应度值的复杂度计算决定着整个遗传操作的效率,在公式(3) 中属性依赖度使用 $\kappa_P(D)k = \gamma_P(D)$,在较大程度上提高了整个 GA 的运行效率,因为公式(2) 只需对每个染色体属性集的等价关系中等价类的数目进行计算,而公式(1) 要通过其等价类计算其正域。

4) 选择和交叉

利用堆排序对适应函数值进行从大到小排序,计算当代群体的适应度值总和,根据种群个人的适应值采用轮盘赌选择法进行选择。

采用最优保存策略,即当前群体中适应度最高的个体不参与交叉和变异运算,而是用它来替换本代群体中经过交叉和变异等一串操作后产生的适应度最低的个体,这样可以保证最终结果是整个搜索过程中的最优。

对选择的个体随机形成两两配对,对配对个体采用单点交叉,以一定的交叉概率 P_C 选择个体参与交叉操作。

5) 变异

对个体进 x 以变异率 P_m 随机选择其中一个基因(特征)发生变异,由于图像特征数据具有维数高等特点,对特征集(条件属性集) 进行属性约简的目标就是寻求长度最小的约简,本文提出了如下变异规则:

a) 如果该位为 0,且该染色体个体内无其他的 1 位,则将该位的值变为 1;

b) 如果该位为 0,且该染色体个体还有其他的位为 1,则另外随机选择值为 1 的位与该位值交换。例如:

1101111000

1111110000

c) 如果该位为 1, 则将该位值改为 0。

6) 终止条件

如果迭代次数大于指定的最大迭代次数 t_{max}，或相邻几代的平均适应度差值小于某个阈值 ε 时，则终止遗传操作。

2.3 算法框架

针对以上具体方案，本文提出的基于改进的 RS-GA 图像特征选择算法的框架如下：

1）个体编码和控制参数初始化。设置进化代数计数器 $t=0$、初始群体个数 $popsize$，最大进化代数 t_{max}、遗传算子、变异 P_m 概率，交叉概率 P_C 和适应度阈值 ε 等。

2）初始群体的产生。用上文中的方法产生初始群体 $popsize$ 个。

3）个体评价。根据公式(3)计算群体 $P(t)$ 中各个个体的适应度值以及适应度总和。

4）使用本文方法对群体 $P(t)$ 经过选择、交叉、变异运算后得到下一代群体 $P(t+1)$。

5）终止条件判断。若满足终止条件，则 $t \leftarrow t+1$，转到步骤3)；否则，以进化过程中得到的具有最大适应度的个体作为最优输出，终止计算。

3 在图像特征选择上的实现

针对图像的特点，对图像特征用基于 RS-GA 的特征选择算法进行特征选择，其具体的操作流程如下：

1）图像预处理，提取图像特征。对于像素高的图像，每一幅图像的数据量太大，可以用图像分割技术先将图像进行分块，再对分块后的图像进行特征提取。

2）数据离散归一化。由于粗集只处理定性数据或概念类的对象，首先对提取的特征值进行离散归一化[4]。

3）归一化的数据建立决策表。我们把图像实例作为对象，提取的图像特征作为决策表的条件属性，类别结果作为决策属性，提取有用特征的过程实质就是对决策表中属性的约简。

4）考察决策表的协调性。如果决策表不协调的，则将表分解为两部分，一个为完全不协调的表，另一个为完全协调的数据表，再对完全协调的表进行如下操作。

5）使用可辨别矩阵求决策表的核。

6）进行属性约简。设定初始控制参数，利用上述 RS-GA 算法进行属性约简至迭代终止。

7）选择最终结果中适应度值最高的个体作为最优解。

4 实验结果与分析

为了检验本文提出方法的有效性，进行仿真实验，实验平台采用 WindowsXP，Matlab 语言编程环境和 RSES2 软件[9]做为辅助实验工具。

所用数据集采用 UCI 数据库中的图像特征数据 Image Segmentation data 和医学图像数据 breast-cancer-wisconsin[10] 中 WPBC 和 WDBC 数据集，其中 image 数据是对先将图像分割为 3×3 像素的小子块再进行特征提取；WPBC 是 Wisconsin 州乳癌预测的图像特征数据，预测癌症是复发的(R：recurrent)或是非复发的(N：nonrecurrent)；WDBC 则是用于乳癌诊断的图像特征，分类结果包括良性(B)和恶性(M)。实验数据说明见表1。

本文首先采用等频率法对数据集进行离散归一化，再对归一化的数据采用本文方法多次反复实验进行特征选择，由于所用数据集结构不同，计算参数有一定的调整；基本设置 pc=0.7，pm=0.1。

根据以上选择的最优特征，分别用 Decomposition Tree、LTF-C(Local Transfer Function Classifier)、Naive Bayes 分类器进行分类验证，因为遗传算法具有的一定的随机性，将使用实验中某次得到的最优特征子集分类与使用全部特征进行对比。实验结果如表2所示，单次适应度曲线如图1所示。

表1 实验数据说明

数据库名称	特征数	类别数	实例数量
Image	19	6	2100
WPBC	33	2	198
WDBC	31	2	569

表2 相关分类算法的分类结果

数据集	特征集(维数)	Decomposition Tree		LTF-C	NaiveBayes
		正确率(%)	覆盖率(%)	正确率(%)	正确率(%)
WPDC	All (33)	76.9	92.9	73.2	77.5
	Opt (5)	75	84.1	85.7	77.5
WDBC	All (30)	93	92.9	40.2	93.7
	Opt (4)	89	94.7	89.3	91.6
Image	All (19)	71.4	92.1	13.2	91.2
	Opt (4)	72.4	93	18.2	91.4

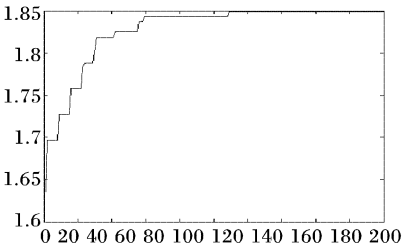

图1 遗传进化适应度曲线

图1中每次迭代的最佳个体适应度值能以较快的速度上升并达到平衡，说明了该方法具有很好的收敛效果。表2显示了在使用了 Decomposition Tree 和 Naive 分类器时，对特征选择后分类的正确率没有明显改变，但是使用 LTF-C 分类器进行分类时，正确率有明显的改善。本文方法中利用属性核作为启发式信息降低了遗传算法搜索空间的大小，引入了相关属性依赖度降低了适应度函数的时间复杂度，并且降低了特征维数以至图像识别中搜索空间的降低，因此特征选择和识别的效率都有很大幅度的提高，节约了运行时间。

参考文献：

[1] DASH M, LIU H. Feature selection for classification [J]. Intelligent Data Analysis - An International Journal, Elsevier, 1997, 1 (3).

[2] 张丽新. 高维数据的特征选择及基于特征选择的集成学习研究[D]. 北京：清华大学，2004.

[3] 曾黄麟. 智能计算[M]. 重庆：重庆大学出版社，2004.4-69.

[4] 王国胤. Rough 集理论与知识获取[M]. 西安：西安交通大学出版社，2001.

[5] 陶志，许宝栋，汪定伟，等. 基于遗传算法的粗糙集知识约简方法[J]. 系统工程，2003，21(4)：116-122.

[6] 何明，冯博琴，马兆丰，等. 一种改进的 Rough 集属性约简启发式遗传算法[J]. 西安石油大学学报(自然科学版)，2004，(5)：80-87.

[7] HAN JC. Feature Selection Based on Rough Set and Information Entropy[J]. Granular Computing, 2005 IEEE International Conference on Volume 1, 2005, 1: 153-158.

[8] 刘文军，谷云东，冯艳宾，等. 基于可辨识矩阵和逻辑运算的属性约简算法的改进[J]. 模式识别于人工智能，2004，17(1)：119-123.

[9] http://www.cs.wisc.edu/~olvi/uwmp/cancer.html[CP]，2006.

[10] http://logic.mimuw.edu.pl/~rses/[DB]，2006.