一种基于$(\mu+\lambda)-ES$进化策略的特征选择方法

冯 林[1,2] 原永乐[1]
(四川师范大学计算机科学学院 成都 610101)[1]
(可视化计算与虚拟现实四川省重点实验室 成都 610068)[2]

摘 要 特征选择是模式识别和数据挖掘等研究领域的一个热点。提出了一种新的特征选择方法FeBES(Feature Selection Based on $(\mu+\lambda)-ES$ Evolutionary Strategy),它以遗传算法为基础,以定义的最优特征集的评价准则为适应度函数,采用$(\mu+\lambda)-ES$进化策略挑选出一组较高质量的特征子集。仿真实验结果表明了该方法的有效性。

关键词 粗糙集,特征选择,遗传算法,支持向量机
中图法分类号 TP18 **文献标识码** A

Approach for Feature Selection Based on $(\mu+\lambda)-ES$ Evolutionary Strategy

FENG Lin[1,2] YUAN Yong-le[1]
(College of Computer Science, Sichuan Normal University, Chengdu 610101, China)[1]
(Sichuan Key Laboratory of Visualization Computing and Virtual Reality, Chengdu 610068, China)[2]

Abstract Feature selection is one of the hot spots in the field of pattern recognition and data mining etc. A novel feature selection method, termed FeBES(Feature Selection Based on $(\mu+\lambda)-ES$ Evolutionary Strategy), was proposed. Under the rule of optimization evaluation of features subset and $(\mu+\lambda)-ES$ evolutionary strategy, a subset of features based on genetic algorithm was selected. Experimental results illustrate that the FeBES is effective for feature selection.

Keywords Rough sets, Feature selection, Genetic algorithm, SVM

1 前言

特征选择能有效降低特征向量的维数,简化分类器的设计和提高识别率。它在模式识别、数据挖掘、机器学习等诸多领域有着十分广阔的应用,引起了众多研究人员的广泛关注[1-3]。

特征提取主要有两个目的:一是缩小数据集;另一个是集中那些具有显著类别差异的分类信息[4]。特征选择的关键问题有两个:一个是最优特征集的评价准则函数(适应度函数)的选择;另一个是快速高效的搜索策略。众多学者对这些问题进行了深入研究,提出了多种类别可分离性准则和最佳特征搜索算法,如距离准则、信息熵准则、重要性评价准则等等。但这些方法均存在这样或那样的不足,尤其是这些准则函数没有考虑所选特征子集的维数,需要在特征选择前先确定特征子集维数或者进行试探法确定维数[5],所以特征选择问题一直是人工智能研究领域关注的一个热点。

特征选择是一个典型的组合优化问题,对于一个有m个特征的特征选择问题,可能的解有$C_m^r(m=1,2,\cdots,r)$个。因此,计算特征选择的复杂性是随着决策表中特征个数的增长呈指数增长,是一个典型的NP问题。而遗传算法(Genetic Algorithm,GA)具有全局搜索和隐含并行性等优点,能够处理传统优化方法难以解决的问题,因此将它用于特征选择可以避免因特征个数多而产生"组合爆炸"问题。但传统GA算法在实际应用中存在迭代次数多、收敛速度慢、易陷于局部极值和过早收敛等一些不足。因此,如何改进传统GA算法的缺陷,解决实际应用中的优化问题,受到了众多学者的广泛关注。如文献[6]提出了一种$(\mu+\lambda)-ES$进化策略,并已经证明$(\mu+\lambda)-ES$进化策略能够以近似1的概率来"处处收敛到最优解"。文献[7]把这一思想方法应用于实际问题求解,取得了很好的实验结果。

本文基于遗传GA算法及其$(\mu+\lambda)-ES$进化策略,提出了一种特征选择的新方法FeBES。它以新的类别可分离性定义为最优特征集评价准则函数,采用有效的$(\mu+\lambda)-ES$的搜索策略,挑选出一组符合条件的有效特征,较好地解决了在类条件概率分布密度未知的情况下分析特征的有效性比较困难的问题。本文最后给出了仿真实验结果与分析。

2 适应度函数的确定

在特征选择中,最优特征集评价准则(适应度函数)的选择是特征选择研究中的一个关键问题。特征的性能通常与不同类别的距离有关,距离越大,分类错误概率的上限越小。人们通常采用马氏距离来衡量类与类之间的距离,虽然这种距离能较好逼近分类错误概率,但当特征不服从正态分布时,它的评价效能一般较差,并且协方差矩阵的逆矩阵有可能不存

到稿日期:2010-09-09 返修日期:2011-01-13 本文受可视化计算与虚拟现实四川省重点实验室科研基金(J2010N01),四川省教育厅科研基金(09ZC079)资助。

冯 林(1972—),男,博士,副教授,主要研究方向为粗糙集理论、数据挖掘等,E-mail:scfengyc@126.com;原永乐(1986—),男,硕士生,主要研究方向为粗糙集理论。

在[4]。针对这一问题，本节给出最优特征集评价函数，并以此作为遗传算法中适应度计算的依据。

为了叙述方便，先给出如下约定：

设样本数据集有 n 个样本、m 个特征、c 个类别。

$X_r^k=[x_{rk}^1 \quad x_{rk}^2 \quad \cdots \quad x_{rk}^m]$ 表示 k 类样本在 m 维特征上的第 r 个样本向量；

$Y_s^k=[y_{sk}^1 \quad y_{sk}^2 \quad \cdots \quad y_{sk}^\pi]^T$ 表示 k 类上的 π 个样本在第 s 个特征上的特征向量，并记 $E(Y_s^k)$ 表示 Y_s^k 的期望值；

$Z_k=[E(Y_1^k) \quad E(Y_2^k) \quad \cdots \quad E(Y_m^k)]$ 表示 k 类样本分别在 m 维特征上的期望值向量。

定义 1 第 i 类的类内聚集度 C_{ii} 定义为

$$C_{ii}=\frac{1}{N_i}\sum_{r=1}^{N_i}\{||X_r^i-Z_i||_2\} \tag{1}$$

式中，N_i 为第 i 类的样本数。

定义 2 第 i 类与第 j 类的距离 d_{ij} 定义为

$$d_{ij}=||Z_i-Z_j||_2 \tag{2}$$

定义 3 在 c 个类别上，各类别之间的“散度”D 定义为

$$D^2=\frac{1}{\binom{2}{c}}\sum_{i=1}^{c-1}\sum_{j>i}^{c}(d_{ij}-\bar{m})^2 \tag{3}$$

式中，$\binom{2}{c}$ 表示在类别 c 上的组合数，$\bar{m}$ 表示各个类别之间距离的期望值。

定义 3 指出了在 m 个特征上各类别之间的“发散”程度，它的优点很明显。我们以图 1 为例来说明这个问题。在图 1 中，有 C_1，C_2 及 C_3 3 个类别，“+”代表各个类别的中心，可以看出：在图 1(a)中的各类别之间的距离和小于图 1(b)中各类别之间的距离和，但很明显，图 1(a)的“散度”更好。因此，与文献[4,5]中的方法单以各类别之间的距离和来度量各类别之间的“分离度”相比，定义 3 用各类之间距离的方差来度量各类之间的“发散”是合理的。

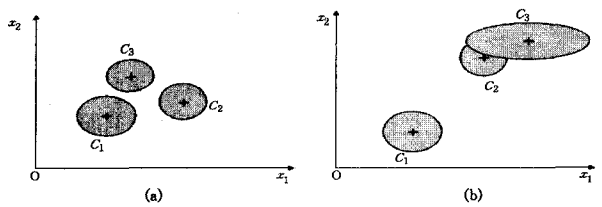

图 1 类别 C_1、C_2 与 C_3 在二维空间(x_1,x_2)中的分布情况

根据以上定义，我们给出评价特征集质量的准则函数。

定义 4 给定 n 个样本、m 个特征、c 个类别的数据集，并设 m 个特征用一个二进制编码 X 来表示，则评价特征集 X 质量的准则函数 $f(X)$ 定义为

$$f(X)=\frac{D}{\sum_{i=1}^{c}C_{ii}} \tag{4}$$

定义 4 较好地体现了不同类别之间的“离散”程度以及同一类别的“聚集”程度。$f(X)$的值说明了使各类别之间的“发散”程度越大的那些特征组合就是所选择的特征子集。

3 FeBES 特征选择方法

我们就 FeBES 方法中涉及的个体编码、变异方式及初始群体的产生等方面做如下详细介绍。

3.1 个体编码

本文采用二进制编码方式，每一个基因位为 0 或 1。如果某位为 1，表示选择该特征，否则为可约去特征。个体编码(基因链码)按特征顺序依次表示为 $a_1a_2\cdots a_m$，其中 m 为属性个数，$a_h\in\{0,1\}$。

3.2 初始群体的产生

随机产生 N 个个体，每个个体由 3.1 节中个体编码方式表示。

3.3 交叉与变异操作

遗传算法中，搜索性主要通过交叉和变异实现。如何通过交叉与变异使 GA 的种群多样性得以保持，以提高 GA 的搜索能力，是 GA 研究和应用中要解决的重要问题。本文我们采用的交叉算子是单点交叉，变异算子为基本变异算子，即针对任一父本，随机挑选 γ 个基因位置并对这些基因位置的基因值以变异概率 P_m 取反，即 0 变成 1，1 变成 0。

3.4 选择策略

在特征选择中，快速高效的优化方法是特征选择研究中的另一个关键问题。本文采用$(\mu+\lambda)-ES$ 进化策略[6]，即从 μ 个父代个体和由其所产生的 λ 个子代个体中选择适应度最高的个体$(\lambda\geqslant\mu)$。在进行变异时，由每个父代产生 P 个新个体，在$(\mu P\cup\mu)$个个体中选择 μ 个适应度最高的个体，将其保留到子代群体中。

3.5 算法描述

算法 1 特征选择 FeBES 方法

输入：m 个特征的数据集

输出：特征选择的结果

Step1 初始化相关参数，包括初始种群规模 μ、特征长度 m、终止代数 g、单点交叉概率 p_1、基本变异概率 p_2；

Step2 按照 3.2 节要求，随机产生 μ 个(偶数个)、特征长度为 m 的初始父本 $S=\{X_1,X_2,\cdots,X_\mu\}$；

Step3 按照式(4)分别计算 μ 个初始父本的适应度 $f(X_i)(i=1,2,\cdots,\mu)$；

Step4 由任两个父本分别按交叉操作产生 2 个子代，则 μ 个父本产生 μ 个子代个体 $S_1=\{Y_1',Y_2',\cdots,Y_\mu'\}$；

Step5 按变异率 p_2 所决定的变异次数 δ，从 S_1 中随机挑选 δ 个染色体进行变异操作，并用产生的新染色体代替原染色体，得 $S_2=\{X_1',X_2',\cdots,X_\mu'\}$；

Step6 按照式(4)分别计算 $X_1',X_2',\cdots,X_\mu'$适应度，$f(X_i')(i=1,2,\cdots,\mu)$；

Step7 按照$(\mu+\lambda)-ES$ 的进化策略，选择出新个体集合$\{X_i,X_i'|\max(f(X_i),f(X_i'))(i=1,2,\cdots,\mu)\}$，即从 μ 个父代和 μ 个子代个体的集合中选择 μ 个适应度函数值最大的个体组成新父本集合 S_3，并代替 S，即 $S=\{X_1,X_2,\cdots,X_\mu\}$；

Step8 进化代数增 1，如果进化代数小于 g，则转 Step3，否则转 Step9。

Step9 输出特征选择的结果，结束。

4 实验及结果分析

我们采用 UCI 中 IRIS 数据集[8]、UCI 中 WINE 数据集[8]来评估算法的性能。由于遗传算法与概率方法密切相关，因此我们也使用随机数据集 RANDDATA 来测试本文方法的性能。IRIS 数据集包括 3 个模式类，每个样本包括一个 4 维的特征向量，每个特征被归一化为[0,1]，样本数为 150；WINE 数据集包括 3 个模式类，每个样本包括一个 13 维的特征向量，每个特征被归一化为[0,1]，样本数为 178；随机数据集 RANDDATA 包括 2 个模式类，每个样本包括一个 25 维的特征向量，随机产生的样本数为 1000。

4.1 实验 1:FeBES 方法的性能测试

实验 1 的主要目的是验证 FeBES 方法的效果。因此,本文给出了利用 SVM 分类器的识别准确率来验证特征选择效果的方法。实验中采用 10 折交叉验证的方法分别在上述 3 个数据集上进行实验。

实验步骤如下:

Step1(方法 1) 使用本文中 FeBES 方法分别对上述 3 种数据集进行特征选择,并把特征选择后的数据集作为 SVM 分类器的输入,并输出识别结果,每次实验进行 3 次,结果为 3 次的平均值。

Step2(方法 2) 把上述 3 种未作特征选择的各数据集(原始数据集)作为 SVM 分类器的输入,并输出识别结果。

实验中 IRIS、WINE 与 RANDDATA 数据集的初始种群规模、单点交叉概率、基本变异概率分别为 12,1,0.05;25,1,0.01;15,1,0.02。支持向量机的参数设置为 SVM Type:C_SVC,Kernel Function:RBF,Multiclass Method:one-against-one。

实验结果见表 1。

表 1 FeBES 方法的性能测试

数据集	原始特征个数	实验次数	迭代次数	FeBES 方法特征个数	分类准确率(%)	方法一实际分类准确率(%)	方法二分类准确率(%)
IRIS	4	1	53	3	96.67	96.45	97.33
		2	58	2	96.00		
		3	49	3	96.67		
WINE	13	1	1011	5	97.75	97.75	96.07
		2	996	5	97.75		
		3	957	5	97.75		
RANDDATA	25	1	5042	12	87.31	87.01	86.14
		2	4859	13	86.42		
		3	5311	12	87.31		

从表 1 看出,FeBES 方法能降低原始数据集中特征的个数。一方面,在 FeBES 方法中,WINE 与 RANDDATA 数据集用较少的特征子集获得了比方法 2 较高的分类准确率,虽然 IRIS 数据集在 FeBES 方法中获得的分类准确率稍低于方法 2,但此时特征个数已减少。实验结果说明,FeBES 方法在处理上述数据集的特征选择方面是有效的。

4.2 实验 2:FeBES 方法与粗糙集特征选择方法的比较

粗糙集理论在属性约简(特征选择)方面取得了很大成功[9,10]。实验 2 的主要目的是把本文特征选择方法与相关粗糙集理论特征选择方法做比较,实验采用了重庆邮电大学计算机科学与技术研究所开发的 RIDAS 测试平台[11]来实现,且同样利用 SVM 分类器的识别准确率来验证特征选择效果,实验中采用 10 折交叉验证的方法分别在上述 3 个数据集上进行实验。实验中的参数设置同实验 1。

实验步骤如下:

Step1 使用本文中 FeBES 方法分别对上述 3 种数据集进行特征选择,并把特征选择后的数据集作为 SVM 分类器的输入,并输出识别结果,每次实验进行 3 次,结果为 3 次的平均值。

Step2 使用粗糙集理论的一般属性约简、基于信息熵的属性约简[12]对上述 3 种数据集作特征选择,并把特征选择后的数据集作为 SVM 分类器的输入,并输出识别结果。实验采用的离散化方法均为 Nguyen 改进的贪心算法[11]。

实验结果见表 2。

表 2 FeBES 方法与粗糙集特征选择方法的对比

数据集	原始特征个数	粗糙集相关特征选择方法				FeBES 方法分类准确率(%)
		一般属性约简特征个数	分类准确率(%)	基于信息熵属性约简特征个数	分类准确率(%)	
IRIS	4	4	97.33	4	97.33	96.45
WINE	13	5	97.75	5	97.75	97.75
RANDDATA	25	13	85.73	13	85.73	87.01

从表 2 并结合实验 1 可以看出,FeBES 方法与粗糙集理论方法在 WINE 数据集上计算的结果相同。而在 RANDDATA 数据集上,FeBES 方法选择的特征子集包含着粗糙集理论特征选择的结果;对 IRIS 数据集而言,使用粗糙集理论没有降低特征的维数,这主要与离散化策略有关。通过实验我们也发现,其它离散化策略可以降低特征的维数,但分类准确率较低。因此,由于篇幅限制,本文没有列出此类情况的计算结果。实验结果说明,FeBES 方法有效降低了数据集特征的维数,并且取得了与基于粗糙集理论的特征选择方法差不多的结果。

结束语 特征选择是近年来模式识别、数据挖掘、机器学习等领域研究的一个热点。本文提出了一种基于遗传算法的新的特征选择方法 FeBES。首先对类间分离度的计算方法进行了改进,并指出了改进后方法的优点,并以此作为特征子集的最优评价准则;其次,采用$(\mu+\lambda)-ES$进化搜索策略,避免了传统遗传算法收敛性速度不足的问题。与相关基于粗糙集理论特征选择方法的对比实验也表明,本文的特征选择方法是有效的。同时,本文也为基于遗传算法的特征选择问题提供了又一新方法。

参 考 文 献

[1] Mitra P, Murthy C A, Pal S K. Unsupervised feature selection using feature similarity[J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2002, 24(3): 301-312

[2] 宋国杰,唐世渭,杨冬青,等. 基于最大熵原理的空间特征选择方法[J]. 软件学报,2003,14(9):1544-1550

[3] Zhang Ge-Xiang, Hu Lai-Zhao, Jin Wei-Dong. Quantum computing based machine learning method and its application in radar emitter signal recognition[J]. Lecture Notes in Artificial Intelligence, 2004, 3131: 92-103

[4] 吕铁军,王河,肖先赐. 新特征选择方法下的信号调制识别[J]. 电子与信息学报,2002,24(5):661-666

[5] 张葛祥. 雷达辐射源信号智能识别方法研究[D]. 成都:西南交通大学,2005

[6] 郭崇慧,唐焕文. 演化策略的全局收敛性[J]. 计算数学,2001,23(1):105-110

[7] 商琳,万琼,姚望舒,等. 一种连续值属性约简方法 ReCA[J]. 计算机研究与发展,2005,42(7):1217-1224

[8] Blake C, Keogh E, et al. UCI repository of machine learning databases[DB/OL]. http://www.ics.uci.edu/~mlearn/MLRepository.html, 2009-10-28

[9] 冯林,李天瑞. 基于 SQL 的属性核与约简高效计算方法[J]. 计算机科学,2010,37(1):236-238

[10] 王国胤,姚一豫,于洪. 粗糙集理论与应用研究综述[J]. 计算机学报,2009,32(7):1229-1246

[11] Wang Guo-yin, Zheng Zhen, et al. RIDAS-A Rough Set Based Intelligent Data Analysis System[C]// Proceedings of ICMLC 2002. Beijing: IEEE Press, 2002: 646-649

[12] 王国胤,于洪,杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报,2002,25(7):759-766