基于多目标微粒群优化的异质数据特征选择

巩敦卫,胡 滢,张 勇
(中国矿业大学信息与电气工程学院,江苏徐州 221116)

摘 要: 环境和测量仪器精度的影响,使得采样数据的不同特征具有不同的质量.对这类异质数据进行特征选择,需要同时考虑特征子集确定分类器的准确度和可靠性,从而增加了特征选择的难度.本文研究异质数据的特征选择问题,提出一种基于多目标微粒群优化的特征选择方法.该方法首先以特征选择的概率为决策变量,将具有离散变量的特征选择问题,转化为连续变量多目标优化问题;然后,采用微粒群优化求解时,基于高斯采样,产生微粒的全局引导者,以提高 Pareto 解集的分布性;最后,依据储备集中元素更新的速度,确定需要扰动的微粒,以帮助微粒群跳出局部最优.将所提方法应用于多个典型数据集分类问题,实验结果表明了所提方法的有效性.

关键词: 特征选择; 异质数据; 多目标优化; 微粒群优化; 高斯采样
中图分类号: TP301 **文献标识码:** A **文章编号:** 0372-2112 (2014)07-1320-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2014.07.012

Feature Selection of Heterogeneous Data Based on Multi-Objective Particle Swarm Optimization

GONG Dun-wei, HU Ying, ZHANG Yong
(*School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou, Jiangsu* 221116, *China*)

Abstract: Different features of a sampling datum have different quality as a result the influence of the environment and the equipment precision. For the feature selection of this kind of heterogeneous data, both the accuracy and the reliability of the classifier determined by a feature subset are required to simultaneously consider, which enhances the difficulty of selecting features. The problem of the feature selection of heterogeneous data is focused on in this paper, and a method of selecting features is presented based on multi-objective particle swarm optimization. In this method, the above problem is first converted to a multi-objective optimization problem by regarding the probability of selecting a feature as the decision variable. When particle swarm optimization (PSO) is employed to solve the converted problem, the global guider of particles is generated by Gaussian sampling so as to improve the performance of Pareto solutions in distribution. In addition, the particle to be disturbed is determined according to the speed of updating a particle in the archive to help the swarm jump out of local optima. The proposed method is applied to classify several benchmark data sets, and the experimental results demonstrate its effectiveness.

Key words: feature selection; heterogeneous data; multi-objective optimization; particle swarm optimization; Gaussian sampling

1 引言

特征选择,是指从数据集的多个特征中,选择一部分特征,使得设定的性能指标达到最优.按照机器学习和分类器结合方式的不同,特征选择方法大体分为过滤算法和封装算法等 2 类[1,2].由于特征子集的选择和机器学习同时执行,封装算法得到的分类器的精度与泛化能力均优于过滤算法,但是,反复的机器学习需要花费大量的时间.为了解决上述问题,近年来,很多学者采用启发式搜索方法,如遗传算法[3]、微粒群优化(Particle Swarm Optimization, PSO)[4]、模拟退火[5],以及禁忌搜索[6],寻找特征子集.

上述特征选择方法均针对数据各特征的质量相同的情况,也即每个特征的值均是完全可信的.环境和测量仪器精度的影响,使得不同特征的质量往往相差很大.对于含有多个特征的数据,一部分特征的值来自高

收稿日期:2013-04-15;修回日期:2013-05-17;责任编辑:孙瑶
基金项目:国家自然科学基金(No.61005089);江苏省自然科学基金(No.BK2011215);高等学校博士学科点专项科研基金(No.20100095120016);中国博士后科学基金(No.2012M521142)

精度仪器的测量结果,相应的,该部分特征的质量较好,从而决策者对它们的信任程度也较高;由于测量另一部分特征的仪器精度低,使得这部分特征的质量较差,从而决策者对它们的信任程度也较低.这类不同特征的值具有不同质量的数据,称为异质数据;相应的特征选择问题,称为异质数据特征选择问题.对于上述问题,用于特征质量无差别数据的已有特征选择方法不再适用.为了解决该问题,需要同时考虑特征子集确定的分类器的精度和可靠性.

微粒群优化是一种受鸟群或鱼群觅食行为启发产生的全局搜索方法.用于解决多目标优化问题的微粒群优化,称为多目标微粒群优化(Multi-Objective PSO, MOPSO).由于具有参数少、易于实现,以及不依赖优化问题的梯度等优点,多目标微粒群优化已被用于解决许多实际问题,如动态传感器设计[7]、电网经济调度[8],以及柔性工件加工调度[9].但是,该方法尚没有用于异质数据特征选择.

本文研究异质数据的特征选择问题,提出一种基于多目标微粒群优化的特征选择方法.主要贡献体现在:(1)给出了异质数据特征选择问题的转化方法,将具有离散变量的特征选择问题,转化为连续变量多目标优化问题,从而能够采用已有的连续变量优化方法求解;(2)提出了微粒群优化求解转化后问题时,微粒全局引导者的产生方法,提高了 Pareto 解集的分布性;此外,给出了需要扰动微粒的确定方法,保证了微粒群跳出局部最优;(3)将所提方法应用于多个典型数据集分类问题,通过实验验证了所提方法的有效性.

2　相关工作

2.1　微粒群优化

微粒群优化是一种全局搜索方法,该方法采用简单的速度-位移模型,将群体中每个个体看成搜索空间没有体积的微粒,并以一定的速度飞行.首先,初始化一个微粒群,并赋予群中每一微粒一个随机的速度;然后,每个微粒根据自身和同伴的飞行经验,动态调整速度和位置.记微粒群中第 i 个微粒的位置为 $\boldsymbol{X}_i=(x_{i1},x_{i2},\cdots,x_{iD})$,速度为 $\boldsymbol{V}_i=(V_{i1},V_{i2},\cdots,V_{iD})$,其中,$D$ 为 $\boldsymbol{X}_i$ 包含的分量个数;此外,记 $\boldsymbol{X}_i$ 经历的最优位置,称为微粒个体引导者,为 $\boldsymbol{P}_i=(p_{i1},p_{i2},\cdots,p_{iD})$;微粒群经历的最优位置,称为微粒全局引导者,为 $\boldsymbol{P}_g=(p_{g1},p_{g2},\cdots,p_{gD})$.在第 $t+1$ 代,微粒 i 的速度和位置更新如下:

$$v_{id}(t+1)=w\cdot v_{id}(t)+c_1\cdot r_1\cdot(p_{id}(t)-x_{id}(t))+c_2\cdot r_2\cdot(p_{gd}(t)-x_{id}(t)) \quad (1)$$

$$x_{id}(t+1)=x_{id}(t)+v_{id}(t+1),\quad i=1,2,\cdots,n,d=1,2,\cdots,D \quad (2)$$

式中,t 为迭代次数,n 为微粒群的规模,c_1 和 c_2 是学习因子,r_1 和 r_2 是[0,1]之间的随机数,w 为惯性权重.

2.2　基于微粒群优化的特征选择

近年来,采用微粒群优化求解特征选择问题,成为模式分类中非常有潜力的研究方向.为了提高文本挖掘的速度,朱颢东等[2]提出一种并行微粒群优化特征选择方法.Mohammad 等[10]将微粒群优化与支持向量机结合,基于关联规则,选择特征子集时,利用微粒群优化寻找支持向量机核函数的最优参数.Chuang 等[4]将混沌二进制微粒群优化用于特征选择问题,分别通过帐篷映射和逻辑映射修改惯性权重,并说明前者得到的特征子集具有更好的分类效果.Alper 等[11]研究特征选择问题时,利用逻辑映射模型,动态调整特征子集里相互关联的特征.此外,Jin 等[12]提出反馈神经网络与微粒群优化相结合的混合方法,用于特征选择问题.但是,上述方法均针对单性能指标的特征选择问题,无法用于异质数据的特征选择.

3　提出的方法

针对异质数据的特征选择问题,将特征选择方法与多目标微粒群优化相结合,提出一种基于多目标微粒群优化的异质数据特征选择方法.该方法首先以特征选择的概率为决策变量,将具有离散变量的特征选择问题,转化为连续变量多目标优化问题;然后,采用微粒群优化求解时,基于高斯采样,产生微粒的全局引导者,以提高 Pareto 解集的分布性;最后,依据储备集中元素更新的速度,确定需要扰动的微粒,以帮助微粒群跳出局部最优.

3.1　问题转化

鉴于微粒群优化通常解决的是连续变量优化问题,相应的,通过对连续变量编码,形成微粒.因此,采用微粒群优化求解具有离散变量的特征选择问题时,需要首先采用合适的方法,将特征选择问题转化为连续变量优化问题.

不失一般性,记数据集的特征个数为 N,在进行特征选择时,如果某一个特征被选中,相应的变量记为 1;否则,记为 0,那么,特征选择问题可以建模为一个决策变量取值为 0 或 1 的组合优化问题,也即具有离散变量的优化问题.

定义一个概率向量,记为 $\boldsymbol{p}=(p_1,p_2,\cdots,p_N)$,其中 $0\leqslant p_i\leqslant 1$,表示第 i 个特征被选中的概率,$i=1,2,\cdots,N$.如果 $p_i\geqslant 0.5$,那么,第 i 个特征被选中;否则,该特征不选择.例如,对于具有 3 个特征的数据集,如果 $\boldsymbol{p}=(0.3,0.7,0.4)$,那么,仅选择第 2 个特征,形成特征子集.这样一来,就能够将特征是否被选择的问题,转化为相应的概率分量是否大于 0.5 的问题,从而将原来的

特征选择问题,转化为以概率向量为决策变量的连续优化问题,使得采用微粒群优化求解该问题成为可能.

鉴于此,本文采用微粒群优化求解转化后的问题.此时,将概率向量编码,形成微粒的位置,记为 $\boldsymbol{X}=(x_1,x_2,\cdots,x_N)$,其中 x_i 为 p_i 的编码,一般采用实数编码.相应的,微粒的速度也为实数向量,但分量的取值不限于[0,1].

3.2 微粒适应值计算

鉴于数据的不同特征具有不同的质量,因此,在特征选择时,考虑特征子集形成的分类器的可靠性,是非常必要的.不失一般性,采用[0,1]之间的数值,反映某特征的质量,这样一来,数据集的质量可以用分量取值在[0,1]之间数值的向量表示,记为 $\boldsymbol{E}=[e_1,e_2,\cdots,e_N]$,$e_i\in[0,1]$,$i=1,2,\cdots,N$,且 e_i 越大,数据的第 i 个特征的质量越好;特别的,当 $e_i=1$ 时,该特征的质量最好,是完全可靠的.由于考虑了数据特征的可靠性,因此,采用微粒群优化进行特征选择时,评价微粒的性能,除了根据特征子集形成的分类器的精度之外,还要考虑该特征子集的可靠性.

首先,考虑微粒对应的特征子集的可靠性.根据第3.1节的编码方式,微粒 $\boldsymbol{X}$ 解码后,为包含 N 个分量的概率向量,这些分量的取值在[0,1]之间;进一步,根据概率向量与特征子集的对应关系,$\boldsymbol{X}$ 与某些(或全部)分量取值为0或1的向量对应,记之为 $\boldsymbol{Y}=(y_1,y_2,\cdots,y_N)$,其中,当 $x_i\geqslant 0.5$ 时,$y_i=1$,表示该特征被选择;反之,$y_i=0$,表示该特征不被选择.从这个意义上讲,$\boldsymbol{Y}$ 是 $\boldsymbol{X}$ 的函数,因此,记之为 $\boldsymbol{Y}(\boldsymbol{X})$.这样一来,微粒 $\boldsymbol{X}$ 对应特征子集的可靠性可以表示为:

$$f_1(\boldsymbol{X})=\frac{\boldsymbol{Y}(\boldsymbol{X})\cdot\boldsymbol{E}^{\mathrm{T}}}{|\boldsymbol{Y}(\boldsymbol{X})|}=\frac{\sum_{i=1}^{N}y_i(x_i)\cdot e_i}{\sum_{i=1}^{N}y_i(x_i)} \tag{3}$$

然后,考虑微粒对应特征子集形成的分类器的精度.本文采用留一交叉验证法,计算分类器的精度,此时,对于含有 K 个数据的数据集而言,将每个数据分别作为一次验证集,剩余 $K-1$ 个数据作为训练集.以第 i 个数据作为验证集为例,首先,用除第 i 个数据之外的其他 $K-1$ 个数据作为训练集,对微粒对应特征子集形成的分类器训练,从而得到一个分类模型;然后,利用第 i 个数据,测试分类模型的性能.如果分类模型能够正确预测该数据所属的类别,那么,$S_i=1$;否则,$S_i=0$.鉴于 S_i 的取值与 $\boldsymbol{X}$ 相关,因此,记之为 $S_i(\boldsymbol{X})$.综合考虑所有 K 个数据,微粒 $\boldsymbol{X}$ 对应特征子集形成的分类器的精度可以表示为:

$$f_2(\boldsymbol{X})=\frac{1}{K}\sum_{i=1}^{K}S_i(\boldsymbol{X}) \tag{4}$$

综合式(3)和(4),微粒 X 的适应值,记为 $F(\boldsymbol{X})$,可以表示为:

$$F(\boldsymbol{X})=\max(f_1(\boldsymbol{X}),f_2(\boldsymbol{X})) \tag{5}$$

3.3 储备集更新

本文借鉴文献[13]的方法更新储备集,思想如下:首先,利用占优关系,选出微粒群的非被占优微粒,复制到储备集中,并删除储备集的重复解;然后,判断储备集中解的个数与储备集容量的关系,如果前者大于后者,那么,计算储备集中所有解的拥挤距离,并按降序排列,删除拥挤距离小的部分个体后,形成新的储备集.

3.4 微粒全局引导者选择

本文基于高斯采样,选择微粒的全局引导者,思想是:首先,将目标空间划分成若干网格,并从存在储备集解的网格中,寻找包含解最少的网格;然后,从该网格中,随机选择储备集的一个解,作为高斯采样的均值,并基于储备集中与该网格最近的2个解之间的距离,形成高斯采样的方差;最后,基于上述均值和方差,应用高斯采样,得到微粒的全局引导者.

基于高斯采样选择微粒的全局引导者,能够使微粒群的进化,不但受储备集中解的邻域密度的影响,而且使微粒群从不同方向逼近问题的 Pareto 前端,从而保证了 Pareto 解集的分布性.

下面,详细阐述基于高斯采样的微粒全局引导者选择方法.

记 $f_1^{\min}$ 和 $f_1^{\max}$ 分别为微粒群进化到某代为止,特征子集可靠性的最小和最大值,类似的,$f_2^{\min}$ 和 $f_2^{\max}$ 分别为特征子集形成分类器精度的最小和最大值;此外,记 n_1 和 n_2 分别为可靠性和分类器精度等分的个数,那么,对于可靠性而言,每一等分的宽度为:

$$\Delta_1=\frac{f_1^{\max}-f_1^{\min}}{n_1} \tag{6}$$

类似的,分类器精度等分的宽度为:

$$\Delta_2=\frac{f_2^{\max}-f_2^{\min}}{n_2} \tag{7}$$

按照上述方法,可以将可靠性和分类器精度等分成不同的部分,由这些部分能够形成不同的网格.

为了确定储备集中不同解在目标空间的位置,仅需给出该解所属的网格即可.考虑储备集中的某一解,记为 a,由式(5)可以得到 a 的适应值 $F(a)=(f_1(a),f_2(a))$.为了确定 a 所属的网格,记为 $C(c_1(a),c_2(a))$,仅需采用下式,给出 $c_1(a)$和 $c_2(a)$的值即可:

$$\begin{aligned}c_1(a)&=\left\lfloor\frac{f_1(a)-f_1^{\min}}{\Delta_1}\right\rfloor+1\\ c_2(a)&=\left\lfloor\frac{f_2(a)-f_2^{\min}}{\Delta_2}\right\rfloor+1\end{aligned} \tag{8}$$

式中,$\lfloor\cdot\rfloor$为向下取整函数.

考虑存在储备集解的网格,从这些网格中,选择包含解最少的,记为 $C_{\min}$.如果该网格仅包含储备集中的一个解,那么,选择该解作为高斯采样的均值;否则,从中任选一个.记得到的均值为 a^*;然后,分别在网格 $C_{\min}$的两侧,寻找储备集中距离 $C_{\min}$最近的解,记为 a_1 和 a_2;最后,以 a^* 作为均值,$\frac{\| a_2 - a_1 \|}{2}$作为方差,实施高斯采样,得到某微粒的全局引导者.

3.5 微粒扰动

微粒群优化过程中,随着迭代次数的增加,很多微粒倾向于在小范围内局部搜索,从而限制了微粒群的探索性能,增大了微粒群陷入局部最优的风险.为了提高微粒群的全局搜索性能,对部分微粒进行扰动.此时,需要确定:(1)扰动微粒的选择;(2)微粒扰动的策略.

首先,选择需要扰动的微粒.考虑进化到某代的微粒群在第 3.4 节划分网格中的分布,对于网格 $C(i,j)$,微粒群进化一代后,该网格中被更新掉的微粒有 2 种情况:(1)更新掉的微粒较多;(2)更新掉的微粒较少,甚至没有.如果某网格中更新掉的微粒较多,说明该网格所在的区域不是问题的局部极值点,此时,无需对该网格中的微粒进行扰动;否则,说明该网格所在的区域有可能是问题的局部极值点,为了避免微粒群陷入局部最优,一种可行的方法是,对该网格中的部分微粒进行扰动,使之跳出该网格.

考虑进化到某代的微粒群,记网格 $C(i,j)$包含的微粒个数为$|C(i,j)|$,微粒群进化一代后,$C(i,j)$中被更新掉的微粒个数为 $n_{C(i,j)}$,那么,该网格中微粒的更新率为$\frac{n_{C(i,j)}}{|C(i,j)|}$,如果$\frac{n_{C(i,j)}}{|C(i,j)|}\leqslant\eta$,其中,$\eta$ 为设定的阈值,那么,随机选择该网格中$\lceil\eta\cdot|C(i,j)|\rceil$个微粒,进行位置扰动.其中,$\lceil\cdot\rceil$表示向上取整函数.

然后,考虑微粒扰动的策略.不妨设进行扰动的微粒为 $\boldsymbol{X}=(x_1,x_2,\cdots,x_N)$,对于第 i 个分量 x_i,产生$[0,1]$之间的随机数,如果该随机数小于或者等于 0.5,那么,对 x_i 以如下方式扰动:

$$x_i=\begin{cases}x_i+0.5, x_i<0.5\\ x_i-0.5, x_i\geqslant 0.5\end{cases} \tag{9}$$

3.6 算法步骤

本文提出的用于异质数据特征选择问题的多目标微粒群优化算法的步骤如下:

Step1 初始化微粒群,设置每个微粒的个体引导者为其本身,储备集为空,迭代次数 $t=0$;

Step2 利用第 3.2 节的方法,计算微粒的适应值;

Step3 利用第 3.3 节的方法,更新储备集;

Step4 判断微粒群进化是否满足终止条件,若是,则终止微粒群进化,输出优化解;

Step5 对每个微粒,执行如下操作:

Step5.1 采用传统的方法,更新微粒的个体引导者;利用第 3.4 节的方法,更新微粒的全局引导者;

Step5.2 由式(1)和(2),更新微粒的速度和位置;

Step6 采用第 3.5 节的方法,对微粒进行扰动,转 Step2.

4 实验

为了验证本文提出的异质数据特征选择方法的性能,选择如下 4 种常用的多目标进化优化算法,进行对比实验,分别是:NSGA-II[14],TV-MOPSO[15],CMOPSO[16],以及 BB-MOPSO[17].鉴于这 4 种算法均用于连续变量优化问题,因此,用于本文的特征选择问题之前,采用第 3.1 节的方法进行问题转化,并采用式(5)计算进化个体(微粒)的适应值.依据这些文献的建议,设置相应算法的参数取值,如表 1 所列.

表 1　参数设置

	种群规模	储备集容量	最大迭代次数	其它参数
NSGA-II	100	100	200	交叉概率 $p_c=0.9$, 变异概率 $p_m=1/N$
TV-MOPSO	50	100	500	$w=0.3(T-t)/T+0.4$, $b=5$, $c_1=-2t/T+2.5$, $c_2=2t/T+0.5$
CMOPSO	100	100	100	$p_m=0.5$, $w=0.4$, $n_1=n_2=30$
BB-MOPSO	100	100	100	
IMOPSO	100	100	100	$w_{\max}=0.995$, $w_{\min}=0.5$, $c_1=c_2=2$, $n_1=n_2=100$, $\eta=0.1$

实验采用的 PC 机配置均为双核 2.69GHz CPU 和 2G RAM,算法采用 Matlab 编程.

采用如下 2 个测度,比较不同方法的性能:

(1)C 测度[18]　考虑 2 种不同的方法得到的 Pareto 解集 A 与 B,其包含元素的个数分别为$|A|$与$|B|$,解集 A 与 B 的 C 测度,记为 $C(A,B)$,表示被 A 中元素占优的 B 中元素的个数,与 B 中元素总个数的比,即:

$$C(A,B)=\frac{|\{b\,|\,b\in B,\exists\, a\in A,\ni:\ a\prec b\}|}{|B|} \tag{10}$$

(2)SP 测度[18]　用来反映某解集在目标空间的分

布性,对于解集 $\boldsymbol{S}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n\}$,其 SP 测度,记为 SP($\boldsymbol{S}$),定义如下:

$$\mathrm{SP}(\boldsymbol{S})=\sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(d^*(\boldsymbol{S})-d(\boldsymbol{X}_j))^2} \quad (11)$$

式中,$d(\boldsymbol{X}_j)=\min\limits_{k\in\{1,2,\cdots,n\},k\neq j}\sum\limits_{i=1}^{2}\left|f_i(\boldsymbol{X}_j)-f_i(\boldsymbol{X}_k)\right|$,$j=1,2,\cdots,n$,为 $\boldsymbol{X}_j$ 在目标空间的拥挤距离,n 为 $\boldsymbol{S}$ 包含的非被占优解的个数.$d^*(\boldsymbol{S})=\dfrac{1}{n}\sum\limits_{j=1}^{n}d(\boldsymbol{X}_j)$为 $\boldsymbol{S}$ 的平均拥挤距离.$f_i(\boldsymbol{X}_j)$为 $\boldsymbol{X}_j$ 在第 i 个目标上的值.

将上述算法分别应用于数据集 Glass,Wine,Vechicle,Ionosphere,Satellite,以及 Sonar 等的分类问题,对于每一个数据集,每种方法均独立运行 20 次,记录每次得到的 Pareto 解集后,计算相应解集的 C 测度和 SP 测度,并统计它们的平均值,分别如表 2 和 3 所列.表 4 列出了 IMOPSO 与上述 4 种比较算法所得 C 和 SP 测度的非参数检验结果,表中,"+"和"-"分别表示在显著性水平为 0.05 时,IMOPSO 的相应测度明显优于和劣于比较算法;"0"表示在设定的显著性水平下,IMOPSO 与比较算法的相应测度无显著差异.

表 2 不同算法得到的 Pareto 解集的 C 测度

	Glass	Wine	Vechicle	Ionosphere	Satellite	Sonar
(IMOPSO,NSGA-II)	1	1	1	1	1	1
(NSGA-II,IMOPSO)	0	0	0	0	0	0
(IMOPSO, TV-MOPSO)	0.0875	0.1500	0.3120	0.3500	0.2350	0.3275
(TV-MOPSO, IMOPSO)	0	0	0	0.0062	0.0092	0.0050
(IMOPSO, CMOPSO)	0.0750	0.1200	0.2500	0.2383	0.1925	0.2800
(CMOPSO, IMOPSO)	0	0	0	0.0143	0.0105	0.0081
(IMOPSO, BB-MOPSO)	0.0500	0.1000	0.0800	0.2200	0.1463	0.2550
(BB-MOPSO, IMOPSO)	0	0.0001	0	0.0232	0.0236	0.0239

由表 2 可以看出,对于所有 6 个测试数据集,IMOPSO 得到的 Pareto 解集能够完全占优 NSGA-II;此外,IMOPSO得到的 Pareto 解集的 C 测度也大于其他 3 种算法,以数据集 Sonar 为例,对于 TV-MOPSO,CMOPSO,以及 BB-MOPSO 而言,IMOPSO 得到的 Pareto 解集的 C 测度分别是 0.3275、0.2800、以及 0.2550;而对于 IMOPSO 而言,这 3 种算法得到的 Pareto 解集的最大 C 测度仅为 0.0239.这说明,与对比方法相比,IMOPSO 得到的 Pareto 解集具有最好的逼近性.

表 3 不同算法得到的 Pareto 解集的 SP 测度

	Glass	Wine	Vechicle	Ionosphere	Satellite	Sonar
NSGA-II	0.3056	0.4214	0.5284	0.2874	0.2149	0.2169
TV-MOPSO	0.1312	0.1071	0.1432	0.1896	0.1732	0.1475
CMOPSO	0.0732	0.0692	0.0789	0.1310	0.1282	0.1035
BB-MOPSO	0.0425	0.0308	0.0481	0.0862	0.0874	0.0568
IMOPSO	0.0317	0.0276	0.0389	0.0824	0.0776	0.0476

由表 3 可以看出,对于这 6 个测试数据集,与对比方法相比,IMOPSO 得到的 Pareto 解集的 SP 测度均最小.这意味着,IMOPSO 得到的 Pareto 解集具有很好的分布性.

表 4 不同算法得到的 Pareto 解集的 C 和 SP 测度的非参数检验值

	Glass		Wine		Vechicle		Ionosphere		Satellite		Sonar	
	C	SP	C	SP	C	SP	C	SP	C	SP	C	SP
NSGA-II	+	+	+	+	+	+	+	+	+	+	+	+
TV-MOPSO	+	+	+	+	+	+	+	+	+	+	+	+
CMOPSO	+	+	+	+	+	+	+	0	+	+	+	+
BB-MOPSO	+	+	+	0	+	+	+	+	+	0	+	0

表 4 可知,对于所有 6 个测试数据集,在显著性水平为 0.05 的情况下:(1)IMOPSO 得到的 Pareto 解集的 C 测度显著优于其他算法,原因在于,IMOPSO 引入了微粒扰动策略,能够使得微粒群跳出问题的局部极值,从而提高了所得 Pareto 解集的逼近性;(2)除了 BB-MOPSO 与 CMOPSO 应用于个别测试数据集之外,IMOPSO 得到的 Pareto 解集的 SP 测度均显著优于其他算法,原因在于,IMOPSO 通过高斯采样选择微粒全局引导者,拓宽了微粒全局引导者的选择范围,从而提高了所得 Pareto 解集的分布性.此外,BB-MOPSO 利用微粒个体引导者和全局引导者的关系,采用高斯采样直接更新微粒的位置,这与 IMOPSO 具有一定的相似性,因而对于某些测试数据集,在设定的显著性水平下,得到的 Pareto 解集的 SP 测度与 IMOPSO 没有显著差别.

图 1 展示了 5 种算法得到的 Pareto 前沿.容易看出:(1)与 NSGA-II 相比,TV-MOPSO,CMOPSO,BB-MOPSO,以及 IMOPSO 的逼近性较好;(2)TV-MOPSO 和 CMOPSO 的分布性不及 IMOPSO;(3)尽管对于数据集 Ionosphere,Satellite,以及 Sonar,BB-MOPSO 得到的 Pareto 解集的性能与 IMOPSO 不相上下,但是,对于另外 3 个数据集,BB-MOPSO 得到的 Pareto 解集的性能却仍然劣于 IMOPSO.

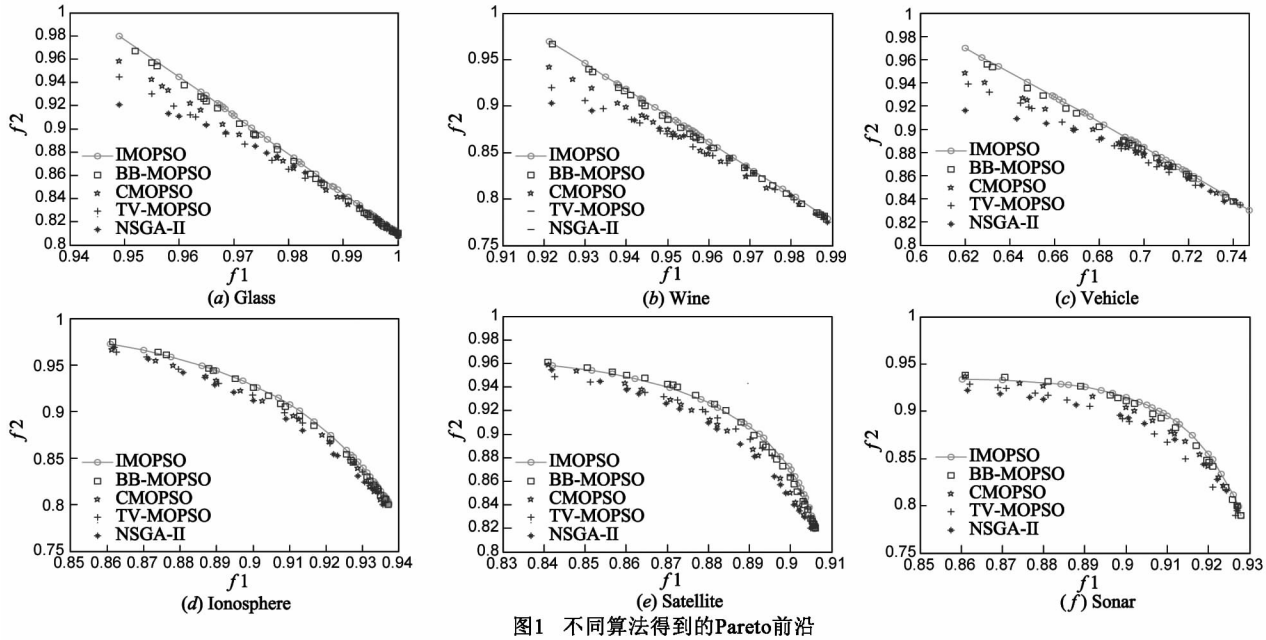


图1　不同算法得到的Pareto前沿

5　总结

异质数据在实际问题中是普遍存在的,数据的不同特征具有不同的可靠性,使得异质数据的特征选择问题变得十分困难,至今没有有效的解决方法.本文将所提方法应用于多个基准数据集分类问题,并与已有的特征选择方法比较.实验结果表明,本文提出的方法在逼近性和分布性等指标上,具有显著的优越性.本文的研究成果为异质数据的特征选择问题,提供了一条行之有效的途径.

需要说明的是,在描述数据特征的可靠性时,本文采用[0,1]之间的精确数值表示.然而,在很多实际问题中,受外部环境、测量设备精度和人类主观等因素的影响,在描述不同特征所对应数据的可信程度(或质量)时,现有技术和方法往往只能帮助决策者定义它的大体数值.相应地,用来描述所得特征子集可靠性的性能指标值变为模糊数.由于存在信息丢失的可能性,此时,传统面向精确数的优化方法不再完全适用.如何处理上述模糊不确定性,是我们下一步需要研究的问题.

参考文献

[1] Roberto H W, George D C, Renato F C, et al. A global-ranking local feature selection method for text categorization[J]. Expert System with Applications, 2012, 39(17): 12851 – 12857.

[2] 朱颢东,钟勇.基于并行二进制免疫量子粒子群优化的特征选择方法[J].控制与决策,2010,25(1):53 – 63.
Zhu Ying-dong, Zhong Yong. Feature selection method based on PBIQPSO[J]. Control and Decision, 2010, 25(1): 53 – 63. (in Chinese)

[3] Peng S H, Xu Q H, Ling X B, et al. Molecular classification of cancer types from micro array data using the combination of genetic algorithms and support vector machines[J]. FEBS Letters, 2003, 555(2): 358 – 362.

[4] Chuang L Y, Yang C H, Li J C. Chaotic maps based on binary particle swarm optimization for feature selection[J]. Applied Soft Computing, 2011, 11(1): 239 – 248.

[5] 宋炜,刘强.基于模拟退火算法的过程挖掘研究[J].电子学报,2009,37(S1):135 – 139.
Song Wei, Liu Qiang. Business process mining based on simulated annealing[J]. Acta Electronica Sinica, 2009, 37(S1): 135 – 139. (in Chinese)

[6] 张昊,陶然,李志勇,蔡镇河.基于 KNN 算法及禁忌搜索算法的特征选择方法在入侵检测中的应用研究[J].电子学报,2009,37(7):1628 – 1632.
Zhang Hao, Tao Ran, Li Zhi-yong, et al. A research and application of feature selection based on KNN and tabu search algorithms in the intrusion detection[J]. Acta Electronica Sinica, 2009, 37(7): 1628 – 1632. (in Chinese)

[7] 李国辉,冯明月,易先清.基于分群粒子群优化的传感器调度方法[J].系统工程与电子技术,2010,32(3):598 – 602.
Li Guo-hui, Feng Ming-yue, Yi Xian-qing. Sensor scheduling method based on grouping particle swarm optimization[J]. Journal of Systems Engineering and Electronics, 2010, 32(3): 598 – 602. (in Chinese)

[8] Zhang W, Liu Y T. Multi-objective reactive power and voltage control based on fuzzy optimization strategy and fuzzy adaptive particle swarm[J]. Journey of Electrical Power and Energy Systems, 2008, 30(9): 525 – 532.

[9] Moslehi G, Mahnam M. A Pareto approach to multi-objective

flexible job-shop scheduling problem using particle swarm optimization and local search[J]. International Journal of Production Economics, 2011, 129(1): 14 – 22.

[10] Mohammad J A, Davar G. Automatic detection of erythemato-squamous diseases using PSO-SVM based on association rules [J]. Engineering Applications of Artificial Intelligence, 2013, 26(1): 603 – 608.

[11] Alper U, Alper M. A discrete particle swarm optimization method for feature selection in binary classified problems[J]. European Journal of Operational Research, 2010, 206(3): 528 – 539.

[12] Jin C, Jin S W, Qin L N. Attribute selection method based on a hybrid BPNN and PSO algorithms[J]. Applied Soft Computing, 2012, 12(8): 2147 – 2155.

[13] 李中凯,谭建荣,冯毅雄等.基于拥挤距离排序的多目标粒子群优化算法及应用[J].计算机集成制造系统,2008, 14(7): 1329 – 1336.
Li Zhong-kai, Tan Jian-rong, Feng Yi-xiong, et al. Multi-objective particle swarm optimization based on crowding distance sorting and its application[J]. Computer Integrated Manufacturing Systems, 2008, 14(7): 1329 – 1336. (in Chinese)

[14] Deb K, Pratap A, Agarwal S, et al. A fast and elitist multiobjective genetic algorithm: NSGA-II[J]. IEEE Transactions on Evolutionary Computation, 2002, 6(2): 182 – 197.

[15] Tripathi P K, Bandyopadhyay S, Pal S K. Multi-objective particle swarm optimization with time variant inertia and acceleration coefficients[J]. Information Sciences, 2007, 177(22): 5033 – 5049.

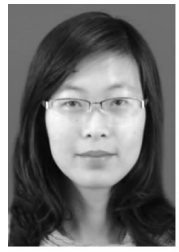
[16] Coello C C A, Pulido G T, Lechuga M S. Handing multiple objectives with particle swarm optimization[J]. IEEE Transactions on Evolutionary Computation, 2004, 8(3): 256 – 279.

[17] Zhang Y, Gong D W, Ding Z H. A bare-bones multi-objective particle swarm optimization algorithm for environmental/economic dispatch[J]. Information Sciences, 2012, 192(1): 213 – 227.

[18] Gong D W, Zhang Y, Zhang J H. Multi-objective Particle swarm optimization based on minimal particle angle[J]. Lecture Notes in Computer Science, 2005, 3644(1): 571 – 580.

作者简介

巩敦卫 男,1970 年 3 月出生,江苏铜山人,中国矿业大学教授、博导、中国电子学会高级会员、中国煤炭学会高级会员. 1992 年、1995 年和 1999 年分别在中国矿业大学、北京航空航天大学、中国矿业大学获理学学士、工学硕士和工学博士学位. 主要研究方向:基于搜索的软件工程、智能优化与控制.
E-mail: dwgong@vip.163.com



胡 滢 女,1988 年 4 月出生,安徽黄山人. 2011 年在安徽理工大学获工学学士学位,现为中国矿业大学硕士研究生,主要研究方向:基于微粒群优化的特征选择.
E-mail: hy200712008@126.com

张 勇 男,1979 年 9 月出生,山东莱芜人,中国矿业大学副教授. 2003 年在聊城大学获工学学士学位,2006 年和 2009 年在中国矿业大学分别获工学硕士和工学博士学位. 主要研究方向:微粒群优化及其应用.
E-mail: yongzh401@126.com