控制与决策
Control and Decision

基于广义罚函数可行性准则的DE算法对不确定数据的处理

王凯光， 高岳， 刘航宇， 周敏

引用本文:
王凯光, 高岳, 刘航宇, 等. 基于广义罚函数可行性准则的DE算法对不确定数据的处理[J]. 控制与决策, 2021, 36(2): 498−504.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2019.0728

基于广义罚函数可行性准则的DE算法对不确定数据的处理

王凯光[1,2], 高岳林[1,2†], 刘航宇[1], 周 敏[1]

(1. 北方民族大学 数学与信息科学学院，银川 750021；
2. 北方民族大学 宁夏智能信息与大数据处理重点实验室，银川 750021)

摘 要: 针对不确定数据集成效率低的问题,构造基于区域分割的广义罚函数可行性准则,分析了分割搜索区域的迭代点特征和可行性准则的性质与优势,据此提出一种基于广义罚函数可行性准则改进的DE算法(DE-GPFFC算法). 机器学习数据集UCI中不确定数据集的数值结果显示:不确定数据集中最优可行点趋向概率0.5分布,其他数据点趋向概率0,1分布,其中趋向于概率0.5分布的数据点位于可行域$\mathrm{int}(D)$,其他数据点位于非可行域$\mathrm{out}(D)$. DE-GPFFC算法使得不确定数据集在可行域边界$\mathrm{Round}(D)$进行跨区域搜索,有效提高了不确定数据分类集成效率.

关键词: 不确定数据；广义罚函数；可行性准则；DE-GPFFC算法；UCI数据

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2019.0728 **开放科学(资源服务)标识码(OSID):**

引用格式: 王凯光,高岳林,刘航宇,等. 基于广义罚函数可行性准则的DE算法对不确定数据的处理[J]. 控制与决策, 2021, 36(2): 498-504.

Application of improved DE algorithm based on generalized penalty function feasibility criteria in uncertain data processing

WANG Kai-guang[1,2], *GAO Yue-lin*[1,2†], *LIU Hang-yu*[1], *ZHOU Min*[1]

(1. School of Mathematics and Information Science，North Minzu University，Yinchuan 750021，China；2. Ningxia Key Laboratory of Intelligent Information and Big Data Processing，North Minzu University，Yinchuan 750021，China)

Abstract: Aiming at the problem of low integration efficiency of uncertain data, this paper constructs a generalized penalty function feasibility criteria (GPFFC) based on region segmentation, analyzes the characteristics of the iterative point feature and the nature and advantages of the screening criterion in the segmented search region, and proposes a improved DE algorithm based on GPFFC (DE-GPFFC). The numerical results of the UCI uncertain data set show that the optimal feasible point tends to have a probability distribution of 0.5 in the uncertain data set, and other data points tend to have a probability of 0,1 distribution, wherein the data points tending to the probability of 0.5 are located in the feasible domain $\mathrm{int}(D)$, other data points are located in the non-feasible domain $\mathrm{out}(D)$. The DE-GPFFC algorithm makes the indeterminate data set cross-region search at the feasible domain boundary $\mathrm{Round}(D)$, which effectively improves the efficiency of uncertainty data classification integration.

Keywords: uncertain data；generalized penalty function；feasibility criteria；DE-GPFFC algorithm；UCI datas

0 引 言

在不确定数据的分类算法中,常见的经典算法有协同聚类算法[1]、贝叶斯广义分析算法(naive Bayes, NB)[2]、带有整数编码的DE算法[3]. 这类基于概率统计原理的数据集成算法所研究的不确定数据样本量是有限的[4]. 对于数据集成处理的具体方式,主要观点集中于如何提高不确定数据的高效率集成,例如Freund[5]所提到的Boosting数据分析方法和Breiman[6]提出的Bagging数据集成分析算法. 对于多种不确定数据处理,Sun等[7]提出了一种将多种不确定数据先分类后集成的模块化数据分析集成策略,通过集成分类器对数据进行处理;Chawla[8]

收稿日期: 2019-05-27；**修回日期:** 2019-10-15.

基金项目: 北方民族大学重大科研专项项目(ZDZX201901)；国家自然科学基金项目(11961001, 61561001)；北方民族大学研究生创新项目(YCX19120)；宁夏高等教育一流学科建设项目(NXYLXK2017B09).

†通讯作者. E-mail: wkg13759842420@foxmail.com.

提出了一种将采样过程中所采用的SMOTE算法与Boosting算法相结合的数据集成策略, Flach等[9]结合1BC和2BC两种贝叶斯结构系统提出了一种有效的递归算法,改进了计算机系统与人工生成数据之间的差异,但是对于高维数据集成和处理,递归算法优势稍弱. 为此, Fernández等[10]在递归算法基础上进一步设计了数据吸收层算法框架,将数据维数控制在合理范围,为低维层级收集不确定数据的有效或关键成分提供了一个好的思路. 然而,这些方法具有以下局限性:一是算法操作技术存在漏洞,易造成核心或关键数据的泄密;二是对于不确定数据的高效集成分类效果不明显. 为了提高不确定数据的集成效率,本文引进算法可行性准则来规避这种概率上的数据缺失.

为了解决不确定数据的高效能分类问题,本文在具有全局收敛的差分进化算法[11-12]的基础上,提出一种基于广义罚函数可行性准则的差分进化集成算法(DE-GPFFC算法),为解决不确定数据的高效集成分类提供一个思路.

1　主要思想和预备知识

本文依据的思想原理主要有以下几个:一是差分进化算法的全局收敛特性好,连续空间的搜索区域广泛[11-12];二是内罚函数[13]和外罚函数[14]具有区域分割搜索的全局特性,可松弛搜索区域,扩大种群个体的搜索面积,增加被搜索个体的种群多样性;三是分割搜索空间具有单点拓扑、分支拓扑、离散拓扑的良好空间性质[12].

1.1　差分进化算法

差分进化算法(differential evolution, DE)是由Storn等[11]提出的、为解决切比雪夫不等式的一种采用浮点矢量编码的在连续空间进行搜索的全局优化算法,是通过差分方式进行迭代搜索的全局性进化算法,具有收敛性好、鲁棒性强等优点,包括4种基本操作算子[11]:初始化种群、变异操作、交叉操作以及选择操作.

1.2　内部罚函数规范化

下面对内部罚函数进行广义规范化:对数函数或半指数函数在非线性优化理论中具有较好的收敛性质[13],本文借助对数函数,将内部罚函数项约归为

$$B(x)=-\sum_{i=1}^{j}\ln\Big(-\frac{g_i}{\{\min+\max\}_{\mathrm{int}(D)}g_i}\Big)+\sum_{i=1}^{j}\mathrm{rand}(0,1)x_{ij}. \tag{1}$$

其中: $\{\min+\max\}_{\mathrm{int}(D)}g_i$表示在可行域内部$\mathrm{int}(D)$最大值和最小值的平均值,目的是为了平衡搜索速度; $\mathrm{rand}(0,1)$是一个随机搜索系数,目的是平衡同一空间中的集中数据点与分散数据点之间的间距,减少搜索时间.

1.3　外部罚函数规范化

下面对外部罚函数广义规范化:为了将等式约束和不等式约束平衡在外部罚函数惩罚项上,本文对P_k进行规范化,其具体表现形式为

$$P(x)=P_i(t)\sum_{i=1}^{j}(\max\{\{\min+\max\}_{\mathrm{int}(D)}g_i,0\})^2+P_i(t)\sum_{i=j+1}^{m}(\{\min+\max\}_{\mathrm{Round}(D)}h_i)^2. \tag{2}$$

其中: $\{\min+\max\}_{\mathrm{Round}D}h_i$表示在可行域边界$\mathrm{Round}(D)$最大值和最小值的平均值; $P_i(t)$表示在搜索区域外增加每一个点的搜索概率,保证搜索区域内外的所有点都能够被有效搜索,提高搜索率.

2　基于广义罚函数可行性准则的差分进化集成算法

2.1　广义混合罚函数

单纯使用内部或外部罚函数都存在自身优势,但同时也存在缺陷: 1) 有效解位置在各自罚因子取值不同时往往出现摆动效应,即如果最优点x^*恰好在约束边界上,则采用这两种算法不会使得x^*被寻找,而是无限接近. 2) 外部罚函数的罚因子过大,会使迭代点x^k在最优点x^*附近摆动而被误认为是最优点,从而造成错解;对于内部罚函数,罚因子过小,会使罚函数项的矩阵条件数变大,以至于难以收敛到最优点,即病态效应. 为平衡二者的寻优缺陷,在内外罚函数基础上,引进广义概率系数,构造如下广义混合罚函数:

$$\Psi(x,r_k)=f(x)+r_kB(x)+\mu_kP(x)+P_i(t)\sum_{i=1}^{j}\mathrm{rand}(0,1)x_{ij}. \tag{3}$$

其中

$$\mu_k=\frac{1}{r_k},$$

$$P(x)=\sum_{i=1}^{j}(\max\{\{\min+\max\}_{\mathrm{int}(D)}g_i,0\})^2+\sum_{i=j+1}^{m}(\{\min+\max\}_{\mathrm{Round}(D)}h_i)^2-$$

$$\sum_{i\notin I}\ln\Big(-\frac{g_i}{\{\min+\max\}_{\mathrm{int}(D)}g_i}\Big),$$

$$I=\{x|g_i(x)\leqslant 0,i=1,2,\ldots,j\},$$

$\{r_k\}$为单调递减的罚因子序列. $\mathrm{Pr}_t(k)$为迭代点的选择概率,其中$t=1,2$,它表示:当罚因子在搜索区域迭代到最优点邻域时,越靠近最优点,迭代点近似最优点程度越大,它对不同类罚函数具有依概率的调节作用,能够调节邻域范围.

2.2 广义罚函数可行性准则

为有效平衡算法搜索过程在混合罚函数解中的摆动问题,设k次迭代后的迭代点为x_k,迭代解x_k精度为充分小的实值$\varepsilon\in(10^{(-4)},10^{(-5)})$. 根据Deb[15]提出的有效解的简要筛法规则,在内部罚函数与外部罚函数及其规范基础上,建立广义罚函数可行性准则如下:

1) 当约束条件下两个有效解x_{k_1}、x_{k_2}位于可行域内部$\mathrm{int}(D)$时,取内部罚函数值较小的有效解为最优近似解,从而保证算法在搜索空间内评估迭代点的递增收敛. 此时,$\mathrm{Pr}_1(k)=x_k\Big/\sum_{i\in I}x_k$,$x_k$满足$f(x_k)=\min(f(x_{k_1}),f(x_{k_2}))$.

2) 当约束条件下两个有效解x_{k_1}、x_{k_2}位于可行域边界$\mathrm{Round}(D)$且为确定解时,取二者均值为一个最优近似解. 此时

$$\mathrm{Pr}_1=P_i(t)\mathrm{rand}(-1,1)\frac{\min\{x_{k_1},x_{k_2}\}}{\sum\limits_{i\in I\bigcup i\notin D}x_k},$$

$$\mathrm{Pr}_2=P_i(t+1)\mathrm{rand}(0,1)\frac{\max\{x_{k_1},x_{k_2}\}}{\sum\limits_{i\in I\bigcup i\notin D}x_k}.$$

3) 当约束条件下两个有效解x_{k_1}、x_{k_2}位于可行域边界$\mathrm{Round}(D)$且为非确定解时,设x_{k_1}为内部迭代解,所有内部迭代解构成可行域内的正的单调递减序列$\{x_{k_1}\}$;x_{k_2}为外部迭代解,所有外部迭代解构成可行域外的正的单调递增序列$\{x_{k_2}\}$. 令$x_{k_1}=\inf\{x_{k_1}\},x_{k_2}=\sup\{x_{k_2}\}$,迭代点的增加可以增强搜索的区域面积,取二者均值$\dfrac{\inf\{x_{k_1}\}+\sup\{x_{k_2}\}}{2}$为一个最优近似解,从而保证算法在搜索空间内评估迭代点的递减收敛. 此时

$$\mathrm{Pr}_1=P_i(t+1)\mathrm{rand}(0,1)\frac{\max\{\inf\{x_{k_1}\},\sup\{x_{k_2}\}\}}{\sum\limits_{i\in\mathrm{int}(I)\bigcup i\notin I}x_k},$$

$$\mathrm{Pr}_2=P_i(t)\mathrm{rand}(-1,1)\frac{\max\{\inf\{x_{k_1}\},\sup\{x_{k_2}\}\}}{\sum\limits_{i\in\mathrm{int}(I)\bigcup i\notin I}x_k}.$$

4) 当约束条件下两个有效解x_{k_1}、x_{k_2}位于可行域外部$\mathrm{out}(D)$时,取外部罚函数值较大的有效解为最优近似解. 此时$\mathrm{Pr}_2=x_k\Big/\sum_{i\in I}x_k$,$x_k$满足$f(x_k)=\max(f(x_{k_1}),f(x_{k_2}))$.

5) 当约束条件的解为无效解时,无论位于搜索区域外部还是内部,都不能作为目标函数的最优近似解. 由于某些无效解与最优近似解之间的搜索关联性,为确保算法在搜索空间持续搜索,需要对无效解进行扩域. 为了保证DE-GPFFC算法在实空间的搜索全局性,对这两个无效解进行依概率操作:以某一个无效解x_1'为圆心,以预设精度$\varepsilon\in(-1,1)$为半径对原搜索区域进行扩域,从而使得无效迭代点在扩域后的松弛可行域中继续有效迭代.

2.3 可行性准则的分析与证明

为了从数学角度更加直观地看到可行性准则的4条性质与算法优势,Wang等[12]利用Heisenberg测不准量子原理在数值实验基础上已经证明了种群个体收敛速度与收敛精度之间的几何关系,这个关系符合测不准原理. 本文利用这个关系,以数学角度从必要性和存在性两个方面对可行性准则的区域分割进行分析.

引理1 设f_ε是定义在完备赋范线性空间中的连续可微函数且$f_\varepsilon(v_1,\ldots,v_n)\in L_2(\boldsymbol{R}^n),M\in\mathrm{Sp}(2n+P_t,\boldsymbol{R})$. M^n为完备赋范线性空间,$\{S_i^n|i=1,2,\ldots,n\}$为广义n维完备赋范线性子空间且$M^n=S_1^n\bigcup S_2^n\bigcup\ldots\bigcup S_n^n$形成了一个关于$M^n$的开覆盖,则$\forall\lambda_i^t\in\{\lambda_i^t|i=1,2,\ldots,n\},\exists S_i^n,\lim\limits_{t\to\infty}\lambda_i^t\sim\lambda_i$或$\lim\limits_{t\to\infty}\lambda_i^t=(\lambda_\varepsilon)_i,\lambda_i\in S_i^n$,当$\det(B)\neq 0$时,有下式成立:

$$\Delta_v^2\cdot\Delta_{x_\beta^\varepsilon}^2\geqslant\Big(\frac{\sqrt{\lim\limits_{t\to\infty}\lambda_1^t}+\ldots+\sqrt{\lim\limits_{t\to\infty}\lambda_n^t}}{2}\Big)^2=\Big(\frac{\sqrt{(\lambda_\varepsilon)_1}+\ldots+\sqrt{(\lambda_\varepsilon)_n}}{2}\Big)^2.\tag{4}$$

引理1的详细证明见文献[12]. 它从理论上说明了个体迭代过程中收敛精度与速度之间的几何关系呈现出类似于量子行为的Heisenberg测不准原理,具有共轭性质,这对于分析区域分割的最优点的跨区域分布和依概率跳跃具有重要意义,通过概率的跳跃作用强制内、外罚函数在不同的分割区域进行搜索.

设区域存在两个局部最优可行点x_1和x_2,其分布满足

$$\Delta_v^2\cdot\Delta_{x_\beta^\varepsilon}^2\geqslant\Big(\frac{\sqrt{x_1}+\sqrt{x_2}}{2}\Big)^2,$$

且

$$\sqrt{\lim_{t\to\infty} x_1^t} = \sqrt{x_1},\ \sqrt{\lim_{t\to\infty} x_2^t} = \sqrt{x_2},$$

其中t是算法迭代次数. 根据引理1对筛法准则的性质给出以下说明:

1) 内罚函数与外罚函数都具有不可跨区域检测性质,不利于边界点的搜索和可行性检测,而增加依概率条件后就能够依据迭代点x_k与全局最优点x^*的近似程度对近似最优解x_k进行平衡处理,于是,准则2)或准则3)在寻优条件上起到了平衡作用;

2) 当迭代点x_k完全落在可行域内部$\mathrm{int}(D)$或外部$\mathrm{out}(D)$时,准则1)或准则4)能对寻优区域中的迭代点进行依概率搜索,因此,可以避免单一寻优区域造成近似全局最优点x^*的丢失,又能保证迭代点在全区域的搜索,避免由于解的取值的单一性而造成精度不高的近似解出现的情况;

3) 由于准则1)或准则4)采用依概率选择策略,这实质上是一种精英保留策略的强化策略,可以在更强约束的优化问题中应用该准则;

4) 由于内部罚函数倾向于选择距离可行域边界较远的迭代点,外部罚函数倾向于选择距离可行域边界较近的迭代点,不利于最优点的全局收敛,该准则可有效避免迭代点的分散,在差分进化算法的诱导作用下可较快找到全局最优点;

5) 可行性准则在增加搜索概率的基础上,实现跨区域搜索,有利于大范围的迭代点检验.

2.4 基于广义罚函数可行性准则的差分进化集成算法的实现

在DE-GPFFC算法中,设空间维数为$D = n$,种群个体可以用$\{(x_i^t, \delta_i)|i = 1,2,\ldots,\mathrm{NP}\}$来表示,$x_i^t$为$n$维决策变量,$\delta_i$为$n$维个体迭代步长变量,初始种群在$n$维空间约束中均匀分布. DE-GPFFC算法的搜索步长计算公式如下所示:

$$\begin{aligned}\delta_i^{t+1} &= \delta_i^t + \exp(\tau\cdot N(0,1) + \tau'\cdot N_i(0,1)),\\ x_i^{t+1} &= x_i^t + \delta_i^{t+1}\cdot N_i(0,1).\end{aligned} \tag{5}$$

其中:t为迭代次数;τ和τ'为种群个体自适应学习率,为了保证计算的准确性,本文按照Schwefel[16]提出的方法进行计算,即$\tau = (\sqrt{2\sqrt{n}})^{-1}$和$\tau' = (\sqrt{2n})^{-1}$;$N(0,1)$和$N_i(0,1)$均是以0为均值、以1为方差的实数均匀高斯分布. DE-GPFFC算法步骤如下:

step 1: 初始化变量. 根据种群规模和个体数量,设$t = 0$,产生包含ζ个父代个体和ξ个子代个体的初始种群,每一个体与$\{(x_i^t, \delta_i)|i = 1,2,\ldots,\mathrm{NP}\}$相对应;

step 2: 计算初始变量的适应度函数值. 计算ζ个父代个体的适应度函数值并记录最大最小值,同时计算初始条件下位于可行域内部和外部各迭代点的适应度函数值并记录最大最小值.

step 3: 根据变异操作并结合式(5)产生所对应的ζ个变异个体.

step 4: 计算产生所对应的ζ个变异个体的适应度函数值,根据可行性准则,依概率计算不同区域的迭代点的适应度函数值.

step 5: 根据变异、选择操作对产生所对应的ζ个变异个体进行选择操作和交叉操作,产生ξ个子代个体并计算其适应度函数值,记录符合精度的迭代点x_k.

step 6: 将ζ个变异个体和经选择操作与交叉操作产生的ξ个子代个体组成新的$(\zeta+\xi)$个种群个体,按照可行性准则选择第t代ζ个个体作为第$t+1$代父代个体,记录符合精度的迭代点x_{k+1}.

step 7: 产生分布在可行域内外两部分一系列迭代点列$\{x_{k_1}\}$、$\{x_{k_2}\}$.

step 8: 判断迭代点列$\{x_{k_1}\}$、$\{x_{k_2}\}$是否为跨区域迭代点.

step 9: 若其中一列迭代点中的某个迭代点$x_{k_i}$$(i \in \{1,2\})$在可行域边界$\mathrm{Round}(D)$从外区域跨向内区域,则以该迭代点$x_{k_i}(i \in \{1,2\})$为圆心,以不超过误差$\varepsilon \in [-1,1]$的距离为半径,产生一列单调递减的迭代点列$\{x_{k_i}\}$,执行step 12;否则执行step 10.

step 10: 若其中一列迭代点中的某个迭代点$x_{k_i}$$(i \in \{1,2\})$在可行域边界$\mathrm{Round}(D)$从内区域跨向外区域,则以该迭代点$x_{k_i}(i \in \{1,2\})$为圆心,以不超过误差$\varepsilon \in [-1,1]$的距离为半径,产生一列单调递增的迭代点列$\{x_{k_i}\}$,执行step 12.

step 11: 按照可行性准则中的依概率公式计算$\mathrm{Pr}_1, \mathrm{Pr}_2$.

step 12: 将step 9的运算结果代入式(1),计算适应度函数值以及迭代点,判断是否满足精度要求.

step 13: 如果迭代点满足可行性准则以及精度要求,则终止算法流程;否则令$t = t+1$,返回step 3.

3 实证分析

3.1 验证数据集

在本次实验中,根据DE-GPFFC算法将UCI机器学习数据集成分为Single-Link数据集成(SLCE)和Complete-Link数据集成(CLCE),将两种集成方法所得结果与KNN-SK(KNN数据填充)[17]、SKNN-SK(SKNN数据填充)[18]所得结果进行集成比较. 本文在

计算机上作随机删除处理(按百分之五比例随机删除),实验结果均为同一集成方法在同一数据集处理200次所得结果.

3.2 验证指标

为了验证DE-GPFFC算法在不确定数据集成分类的效果,本文采用CA (classification accuracy)[19]、ARI (adjusted rand index)[19] 和 NMI (normalized mutual information)三类验证指标,其具体描述如下.

1) CA: 是衡量DE-GPFFC算法在正确集成分类所对应集成类别的样本比例的统计数学量,计算公式为

$$\mathrm{CA}=\frac{1}{n}\sum_{i=1}^{k}\alpha_i. \tag{6}$$

其中: α_i 是DE-GPFFC算法在正确集成分类所对应集成类别的样本数量, k 是集成类, n 是样本总数.

2) ARI: 是衡量在考虑了相同数据类和相异数据类后,DE-GPFFC算法在正确集成分类所对应集成类别的样本比例的统计数学量,计算公式为

$$\mathrm{ARI}=\frac{\sum\limits_{i=1}^{I}\sum\limits_{j=1}^{J}C_{n_{ij}}^2-\eta}{\frac{1}{2}(\rho+\phi)-\eta}, \tag{7}$$

$$\rho=\sum_{i=1}^{I}C_{n_i}^2,\ \phi=\sum_{j=1}^{J}C_{n_j}^2,\ \eta=\frac{2\rho\phi}{n(n-1)}. \tag{8}$$

其中: n_{ij} 是数据集成结果中第 i 个数据集包含原数据集类为 j 的样本数量; $C_{n_{ij}}^2$ 是数据集中同时包含第 i 个数据集和第 j 个数据集的排序数量,且 $C_{n_{ij}}^2=C_{n_i}*C_{n_j}$, C_{n_i}、C_{n_j} 分别为第 i 个数据集和第 j 个数据集的排序数量; n_i 是数据集成结果中第 i 个数据集的样本数量; n_j 是原数据集类为 j 的样本数量; n 是样本总数; I、J 是数据集成结果中得到的不同类数据集个数和原数据集类个数.

3) NMI: 是衡量DE-GPFFC算法在正确集成分类所对应集成类别的样本数量的相似度,计算公式为

$$\mathrm{NMI}=\frac{\sum\limits_{i=1}^{I}\sum\limits_{j=1}^{J}\frac{n_{ij}^2 n}{n_i n_j}}{\sqrt{\sum\limits_{i=1}^{I}\left(\frac{n_i^2}{n}\right)\cdot\left(\sum\limits_{j=1}^{J}\frac{n_j^2}{n}\right)}}, \tag{9}$$

其中各变量含义同式(7)和(8).

以上3类验证指标极限值设置为1,若不确定数据集成后数据结构与原数据结构接近,则表示数据集值较大,也即DE-GPFFC算法对不确定数据集成效果较好.

3.3 实验分析

KNN-SK与SKNN-SK选取的临近个数为 $K=10$, DE-GPFFC算法与其他集成聚类算法在本文验证指标下的实验结果及不确定数据方差如表1、图1所示.

由图1可知,全部数据关于DE-GPFFC算法的CA值、ARI值、NMI值均呈现: 平均最优值与平均次最优值稳定在0 ~ 0.2周围且误差不超过±0.2; 方差最优值与方差次最优值基本稳定在0.01周围,符合数据稳定性标准且误差不超过±0.1,符合稳定性要求. 从表1中数据集分析可得出如下结论: DE-GPFFC算法在均值与方差上的数值结果呈现两极分化,不确定数据集由于分为两类概率趋势而被高效集成,从而显示了DE-GPFFC算法对不确定数据集成的优越性.

表1 DE-GPFFC算法关于CA、ARI、NMI的值的比较

data sets	CA				ARI				NMI			
	SLCE	CLCE	KNN	SKNN	SLCE	CLCE	KNN	SKNN	SLCE	CLCE	KNN	SKNN
Der.	0.030	**0.522**	0.912	0.003	0.774	0.014	**0.600**	0.038	0.402	0.864	**0.410**	0.230
C. A.	0.121	0.880	**0.470**	0.203	0.673	0.001	0.590	**0.519**	0.020	0.353	**0.621**	0.008
Aut.	0.080	0.755	**0.500**	0.633	0.930	0.112	**0.544**	0.177	0.701	0.006	0.049	**0.572**
Spo.	0.714	0.006	**0.688**	0.330	0.959	0.010	0.007	**0.223**	0.900	0.247	0.233	**0.355**
(CMC)	0.033	0.793	0.921	**0.530**	0.558	0.662	0.101	**0.503**	0.200	0.241	0.793	**0.590**
Soy.	0.055	**0.499**	0.061	0.041	0.711	0.109	**0.665**	0.093	**0.203**	0.008	0.922	0.115
Glass	**0.450**	0.379	0.115	0.573	0.880	**0.517**	0.002	0.433	0.019	0.703	0.114	**0.559**
平均值	CA				ARI				NMI			
	0.211	0.547	**0.523**	0.330	0.783	0.203	**0.358**	0.283	0.349	0.346	**0.448**	0.347
方差	0.070	0.080	0.120	**0.060**	**0.020**	0.070	0.090	0.030	0.110	0.100	0.110	**0.050**

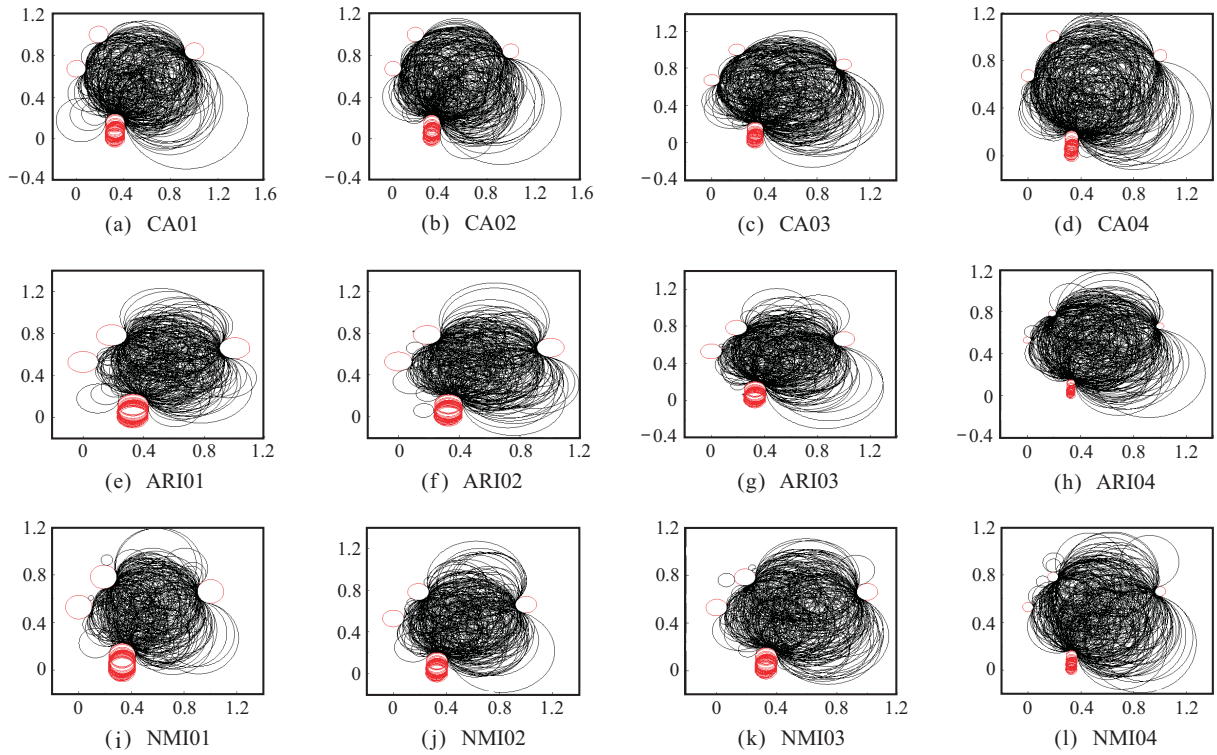

图 1　DE-GPFFC算法关于CA、ARI、NMI的收敛集成结果

Dermatology 数据和 Credit Approval 数据关于 DE-GPFFC 算法的 CA 值、ARI 值、NMI 值呈现如下情况: 最优值与次最优值趋向概率 1 分布, 其他数据点趋向概率 0 分布, 其中趋向于概率 1 分布的数据点位于可行域 $\mathrm{int}(D)$, 趋向于概率 0 分布的数据点位于非可行域 $\mathrm{out}(D)$. Automobile 数据、Sponge 数据和 Soybean 数据关于 DE-GPFFC 算法的 CA 值、ARI 值、NMI 值呈现如下情况: 不确定数据集中最优可行点趋向概率 0.5 分布, 其他数据点趋向概率 0,1 分布, 其中趋向于概率 0.5 分布的数据点位于可行域 $\mathrm{int}(D)$, 其他数据点位于非可行域 $\mathrm{out}(D)$. DE-GPFFC 算法使得不确定数据集在可行域边界 $\mathrm{Round}(D)$ 进行跨区域搜索, 有效提高了不确定数据分类集成效率.

4　结　论

本文基于 DE 算法并结合混合罚函数的广义规范化, 对不确定数据处理算法进行了研究和分析, 构造了基于广义混合罚函数筛法准则的差分进化集成算法, 以改善不确定数据在数据集成方面的分类问题或提高数据集成技术. 通过 UCI 机器学习真实数据集进行了实证分析, 其验证结果符合不确定数据分类预期, 且集成效果较好.

参考文献(References)

[1] Everitt B. Cluster analysis[J]. Quality and Quantity, 1980, 14(1): 75-100.

[2] 叶云, 石聪聪, 余勇, 等. 保护隐私的分布式朴素贝叶斯挖掘[J]. 应用科学学报, 2017, 35(1): 1-10.
(Ye Y, Shi C C, Yu Y, et al. Privacy-preserving distributed naive Bayes data mining[J]. Journal of Applied Sciences, 2017, 35(1): 1-10.)

[3] 王凯光, 高岳林. 十进制整数编码的DE算法模式集定理研究[J]. 应用数学, 2019, 32(2): 443-451.
(Wang K G, Gao Y L. The schema sets theorem of DE algorithm for decimal integer coding[J]. Mathematica Applicata, 2019, 32(2): 443-451.)

[4] Díez-Pastor J F, Rodríguez J J, García-Osorio C, et al. Random balance: Ensembles of variable priors classifiers for imbalanced data[J]. Knowledge-Based Systems, 2015, 85: 96-111.

[5] Freund Y. Boosting a weak learning algorithm by majority[J]. Information and Computation, 1995, 121(2): 256-285.

[6] Breiman L. Bagging predictors, machine learning research: Four current directors[J]. ResearchGate, 1996, 24(2): 123-140.

[7] Sun Z B, Song Q B, Zhu X Y, et al. A novel ensemble method for classifying imbalanced data[J]. Pattern Recognition, 2015, 48(5): 1623-1637.

[8] Chawla N V, Lazarevic A, Hall L O, et al. SMOTEBoost: Improving prediction of the minority class in Boosting[J]. Lecture Notes in Computer Ence, 2003, 2838: 107-119.

[9] Flach P A, Lachiche N. Naive Bayesian classification

of structured data[J]. Machine Learning, 2004, 57(3): 233-269.

[10] Fernández A, López V, Galar M, et al. Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches[J]. Knowledge-Based Systems, 2013, 42(2): 97-110.

[11] Storn R, Price K. Differential evolution — A simple and efficient heuristic for global optimization over continuous spaces[J]. Journal of Global Optimization, 1997, 11(4): 341-359.

[12] Wang K, Gao Y. Topology structure implied in β-Hilbert space, heisenberg uncertainty quantum characteristics and numerical simulation of the DE algorithm[J]. Mathematics, 2019, 7(4): 330-345.

[13] Wright M. The interior-point revolution in optimization: History, recent developments, and lasting consequences[J]. Bulletin of the American Mathematical Society, 2005, 42(1): 39-57.

[14] Bard J. Engineering optimization: Theory and practice, third edition[J]. IIE Transactions, 1997, 29(9): 802-803.

[15] Deb K. An efficient constraint handling method for genetic algorithms[J]. Computer Methods in Applied Mechanics and Engineering, 2000, 186(2/3/4): 311-338.

[16] Schwefel H P P. Evolution and optimum seeking[M]. New Jersey: John Wiley & Sons Inc., 1995: 100-185.

[17] Silva L O, Zárate L E. A brief review of the main approaches for treatment of missing data[J]. Intelligent Data Analysis, 2014, 18(6): 1177-1198.

[18] Batista G E A P A, Monard M C. An analysis of four missing data treatment methods for supervised learning[J]. Applied Artificial Intelligence, 2003, 17(5/6): 519-533.

[19] Liang J Y, Bai L, Dang C Y, et al. The K-means-type algorithms versus imbalanced data distributions[J]. IEEE Transactions on Fuzzy Systems, 2012, 20(4): 728-745.

作者简介

王凯光 (1994−), 男, 助理研究员, 硕士, 从事最优化理论与方法、智能信息处理、智能优化理论与应用的研究, E-mail: wkg13759842420@foxmail.com;

高岳林 (1963−), 男, 教授, 博士生导师, 从事最优化理论与方法、智能计算与应用等研究, E-mail: gaoyuelin@263.net;

刘航宇 (1994−), 男, 硕士生, 从事智能计算、数据处理的研究, E-mail: hangyu1815@163.com;

周敏 (1994−), 女, 硕士生, 从事智能计算理论与应用的研究, E-mail: 18408613876@163.com.

(责任编辑：李君玲)

下 期 要 目