

基于克隆选择的差分进化算法及其在 SVM 中的应用

盛明明 黄海燕 赵 玉

(华东理工大学信息科学与工程学院 上海 200237)

摘 要 支持向量机参数是影响其性能的重要因素,但对支持向量机核参数的选取仍没有形成一套成熟的理论,从而严重影响了其广泛的应用。将克隆选择算法引入差分进化算法,对基本克隆选择算法和差分进化算法中的策略进行改进。将两种改进的算法进行融合,提出了一种基于克隆选择的差分进化算法,并将其应用于 SVM 核参数的优化中。测试结果表明,该算法不仅可以有效避免差分进化算法易早熟收敛的问题,而且寻优能力得到显著提高;在 UCI 数据库 wine 数据中的应用表明,利用克隆选择差分进化算法优化 SVM 核参数加快了参数搜索的速度,提高了 SVM 预测精度和泛化能力,具有较高的分类准确率和较好的推广性能。

关键词 克隆选择,差分进化,支持向量机,核参数

中图分类号 TP273 文献标识码 A

Differential Evolution Algorithm Based on Clonal Selection and its Application in SVM

SHENG Ming-ming HUANG Hai-yan ZHAO Yu

(School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China)

Abstract The parameters of support vector machine (SVM) are important factors affecting its performance. However, the absence of a mature theory about the kernel parameter selection of SVM heavily affects its wide application. This paper introduced clonal selection algorithm into differential evolution algorithm, and improved the strategies of basic clonal selection algorithm and differential evolution algorithm. Through combining the two algorithms mentioned above, a differential evolution algorithm based on clonal selection was proposed and applied to optimize the parameters of SVM kernel. The test results show that the algorithm can not only effectively avoid the premature-convergence problem of differential evolution algorithm, but also significantly improve the optimization ability. UCI wine database application data show that the algorithm can accelerate the parameter search speed, and improve the prediction accuracy and generalization ability of SVM. The high accuracy of classification and better generalization performance prove that using clonal selection differential evolution algorithm is a good way to optimize SVM kernel parameter.

Keywords Clone selection, Differential evolution, SVM, Kernel parameter

1 前言

支持向量机 (Support Vector Machine, SVM) 是 Vapnik 等人于 1995 年在统计学习理论上提出的智能优化算法^[1]。与传统的学习方法相比,它具有较好的推广能力和非线性处理能力,并且能够处理高维数据,解决一般学习机不能解决的“维数灾难”和“过学习”问题。作为当前机器学习领域中的一个新的研究热点,其分析和研究仍有许多需要进一步发展和完善的地方。

SVM 的核心部分主要是核参数,它直接影响到算法的运算效率和模型预测的性能。但对于 SVM 核参数的选取仍没有形成一套成熟的理论,传统的参数选取大多依据经验采取试凑的方法,不仅耗时,而且得不到满意的结果,随着支持向量机的发展,传统的方法明显不再适合对支持向量机的参数进行优化^[2-4]。

差分进化 (Differential Evolution, DE) 算法是一种启发式随机搜索优化算法,其思想原理相对简单,算法控制参数较

少,具有较强的全局搜索能力和鲁棒性,而且能提高优化速度,与其他一些算法相比,更容易收敛得到全局最优解^[5-7]。但 DE 算法易陷入局部最优,存在过早收敛的现象。

克隆选择算法 (CLONALG) 是人工免疫系统中一种经典的免疫算法模型,由 De Castro 于 2002 年根据生物免疫系统理论中的克隆选择学说提出^[8]。由于该算法具有并行性、自适应性、学习、识别和记忆等优点,很快被应用到函数优化^[9]、特征选择^[10]、入侵检测^[11]、图像分割^[12]、机器学习^[13] 等领域。

针对 DE 算法的不足和克隆选择的原理以及文献^[14] 提出的算法思想,本文对基本 CLONALG 算法加以改进,再与 DE 算法进行融合,提出基于克隆选择的差分进化 (CDE) 算法,并将该算法应用于 SVM 核参数的优化中。

2 克隆选择差分进化算法

本文将 CLONALG 算法引入 DE 算法,并对算法策略做了一定的改进,改进的主要特征有:1) 改进 CLONALG 算法

盛明明(1990—),男,硕士生,主要研究方向为故障诊断、智能计算, E-mail: laviewsam@qq.com; 黄海燕(1972—),女,博士,副教授,主要研究方向为控制与优化、复杂工业过程建模; 赵 玉(1989—),男,硕士生,主要研究方向为故障诊断、人工智能。

的克隆策略,采用新的增殖方式;2)对 CLONALG 算法采用新的变异策略;3)改进 DE 算法变异策略。

2.1 差分进化算法

DE 算法主要有 3 个操作:变异、交叉和选择。

1)变异。在差分进化算法众多的变异策略中,常用的是“DE/rand/1”和“DE/best/1”两种方法。本文为提高 DE 算法的收敛性,同时考虑两种变异方式的特点,在新算法中将交叉使用这两种方法。

$$v_i^{k+1} = x_{r_1}^k + F(x_{r_2}^k - x_{r_3}^k) \quad (1)$$

$$v_i^{k+1} = x_{best}^k + F(x_{r_1}^k - x_{r_2}^k) \quad (2)$$

式中, r_1, r_2, r_3 为种群个体中随机产生的 3 个互不相同的向量; F 是缩放因子,当 F 取值较小时算法容易陷入局部收敛,取值过大则降低收敛速度,本文对于 F 的选取采用参数自适应的形式,将 F 控制在截断区间 $[F_{min}, F_{max}]$,并根据进化代数按式(3)进行动态更新,本文取 $F_{max} = 0.8, F_{min} = 0.5$ 。

$$F = F_{max} - (F_{max} - F_{min}) \times \sqrt{\frac{g}{g_{max}}} \quad (3)$$

2)交叉。交叉操作通过以下方法产生实验个体:

$$u_i^{k+1} = \begin{cases} v_i^{k+1}, & rand(0,1) \leq CR \\ x_i^k, & otherwise \end{cases} \quad (4)$$

$CR, rand(0,1)$ 分别为 $[0,1]$ 之间的交叉概率和均匀分布的随机数,本文 CR 取值 0.8。

3)选择。对当前任一个体 x_i^k ,选取实验种群中对应的 u_i^{k+1} ,通过选择算子来产生子代个体 x_i^{k+1} 。

$$x_i^{k+1} = \begin{cases} u_i^{k+1}, & f(u_i^{k+1}) > f(x_i^k) \\ x_i^k, & f(u_i^{k+1}) \leq f(x_i^k) \end{cases} \quad (5)$$

2.2 克隆选择

CLONALG 算法主要包含克隆、变异和选择等操作,具有收敛速度快、多样性好等特点。针对 DE 算法容易形成早熟收敛的缺陷,本文将克隆选择算法引入 DE 过程,并且对 CLONALG 算法的克隆和变异使用新的策略,用以加强算法的局部搜索能力。

克隆操作步骤如下:

①抗原提呈。生成种群规模为 N 的候选抗体集合 X 。

②适应度计算。计算 X 中个体适应度,从中选择 n 个最佳的抗体组成集合 $P_n, n \leq N$ 。

③克隆。对集合 P_n 中的个体进行克隆复制,构成临时集合 C 。为了避免算法过早收敛并且保证算法的收敛速度,本文采用如下克隆策略:保证克隆数目与适应度成正相关的同时,利用进化代数放大后期及过早收敛时个体间的微小差异,进一步增加优秀个体数目,提高优秀个体的变异机会,加强连续寻优能力。

$$N_c(t) = (1/t + \alpha^{(f_t - f_{best})^{1/k}}) \times n \quad (6)$$

其中, $N_c(t)$ 表示抗体 t 要克隆的数量; f_t 是抗体 t 的适应度值; f_{best} 是当前抗体中适应度最优值; α 为小于 1 的常数。本文 α 取值 0.1。

④变异。本文为避免算法陷入局部最优,对克隆后抗体采取自身位置搜索的变异策略,即对集合 C 中的每个个体,按给定方向和距离进行自身位置搜索,生成变异抗体群 C^* 。

$$x_i^{*k} = x_i + v_i \quad (7)$$

$$v_i = \beta \times \lambda \times x_i \quad (8)$$

其中, x_i^{*k} 表示变异后产生的新抗体; v_i 表示搜索步长; β 和 λ 分别为搜索方向和搜索比例。为保证多方向全面寻优,本文抗体搜索方向按式(9)确定;同时为了保证算法前期具有较大的变异空间,以及实现算法后期的小范围局部搜索,本文搜索比例由式(10)进行实时更新,其中 θ 为小于 1 的正数,本文 θ 取值 0.3。这样个体多样性在变异环节得到极大丰富的同时也避免了算法的“早熟”现象。

$$\beta(t) = (-1)^t \quad (9)$$

$$\lambda(t) = 1 - \theta^{(1 - \frac{t}{g_{max}})^2} \quad (10)$$

⑤选择。选择变异后适应度最佳的个体进入下一代。

2.3 算法流程

本文提出的 CDE 算法实现步骤如下:

Step1 初始化种群规模 Np 、最大进化代数 g_{max} 。令迭代次数 $g = 1$ 。

Step2 计算个体适应度,求出最优适应值并记录下当前代最优个体 x_{best}^g 。

Step3 如果最优适应度达到理论最优值或迭代次数 g 等于最大迭代次数 g_{max} ,输出结果;否则,执行 Step4。

Step4 按照适应度大小对所有个体排序,将中间的 1/3 个体作为抗体进行克隆选择操作,并对 x_{best}^g 进行更新。

Step5 根据当前进化代数,奇数代,随机选择种群中 3 个不同的个体按式(1)进行变异操作;偶数代,随机选择种群中 2 个不同的个体,按式(2)进行变异操作。生成变异个体 v_i^g 。

Step6 按式(4)进行交叉操作,生成实验个体 u_i^g 。

Step7 按式(5)进行选择操作,生成 $g+1$ 代个体 x_i^{g+1} 。

Step8 $g = g + 1$,返回 Step2。

2.4 算法性能评估

为了测试新算法的优化效果,采用如下 4 种典型的 Benchmark 函数来测试算法的性能。函数 f_1 和 f_2 主要测试算法的寻优精度以及收敛速度,函数 f_3 和 f_4 主要检测算法的全局搜索性能。

1) Sphere 函数

$$f_1(x) = \sum_{i=1}^n x_i^2 \quad (11)$$

全局最优值: $x_i = 0, f(x) = 0$

2) Rosenbrock 函数

$$f_2(x) = \sum_{i=1}^n 100 \times (x_{i+1} - x_i^2)^2 + (1 - x_i)^2 \quad (12)$$

全局最优值: $x_i = 0, f(x) = 0$

3) Rastrigin 函数

$$f_3(x) = \sum_{i=1}^n [x_i^2 - 10 \cos(2\pi x_i)] + 10 \quad (13)$$

全局最优值: $x_i = 0, f(x) = 0$

4) Griewank 函数

$$f_4(x) = \frac{1}{4000} \sum_{i=1}^n x_i^2 - \prod_{i=1}^n \cos\left(\frac{x_i}{\sqrt{i}}\right) + 1 \quad (14)$$

全局最优值: $x_i = 0, f(x) = 0$

利用本文提出的 CDE 算法与标准 DE 算法和基本 CLONALG 算法分别对上述函数进行寻优。实验中,函数维数为 30,种群规模 $Np = 100$,最大迭代次数 $g_{max} = 1000$,每种算法各运行 100 次,以减少偶然性,保证测试质量,结果如表 1 所列。

表1 DE/CLONALG/CDE 函数测试比较

函数		f ₁	f ₂	f ₃	f ₄
平均值	DE	3.46e-17	22.8239	152.23	6.05e-03
	CLONALG	6.31e-02	25.0838	17.01	7.39e-01
	CDE	5.97e-156	2.61e-05	0	0
标准差	DE	1.75e-17	5.67e-01	7.02e+00	1.12e-03
	CLONALG	4.35e-02	2.06e+00	3.15e+00	9.54e-02
	CDE	2.67e-156	4.59e-05	0	0
最快收敛代数	DE	31	317	44	140
	CLONALG	124	76	291	135
	CDE	19	15	153	24
平均收敛代数	DE	92	503	76	288
	CLONALG	203	155	574	267
	CDE	46	72	236	56

从表1结果可以看出:

①对 Sphere 函数进行优化时,CDE 算法与 DE 算法均表现出较快的收敛速度,但 CDE 算法收敛精度明显高出其他两种算法。

②对 Rosenbrock 函数的优化,DE 算法收敛时间较长,CLONALG 算法时间虽短但精度最低,CDE 算法收敛速度最快、精度最高,说明 CDE 算法能够很好地避免陷入局部最优。

③从 Rastrigin 函数的优化结果可以看到,DE 算法收敛速度最快,但收敛精度最低,CDE 算法以比较适中的收敛速度达到了理想的收敛精度。

④在 Griewank 函数的优化中,CDE 算法以快而准的特点领先于其他两种算法。

综上,本文提出的 CDE 算法克服了标准 DE 算法和 CLONALG 算法的自身缺陷,并在函数测试中性能突出,具有收敛速度快、精度高的特点,寻优能力比较乐观。

3 基于 CDE 算法的 SVM 参数优化

3.1 支持向量机

支持向量机的基本思想是通过非线性映射将原空间转换到一个新的特征空间,然后在特征空间中求出最优分类超平面,并且在机器的学习能力和模型的复杂性之间寻找最佳折衷,使得超平面与各个样本间的距离最大化,从而获取最好的推广能力。

设样本集为:

$$(x_i, y_i), x_i \in R^n, y_i \in \{-1, 1\}, i=1, 2, \dots, l$$

其中, x_i 为输入向量, y_i 为对应的输出值。

当样本点近似满足线性分类时,便归结为以下优化问题:

$$\begin{cases} \min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l \xi_i \\ \text{s. t. } y_i((w, x) + b) \geq 1 - \xi_i, i=1, \dots, l \end{cases} \quad (15)$$

其中, $\|w\|^2$ 称为结构风险,代表模型的复杂程度,使函数更为平坦,从而提高泛化能力; $C \sum_{i=1}^l \xi_i$ 称为经验风险,代表模型的误差; C 为惩罚系数,协调错分样本比例与算法复杂度间的关系; ξ_i 为松弛变量。引入拉格朗日函数将其转化为对偶问题:

$$\begin{cases} \max \sum_{i=1}^l a_i - \frac{1}{2} \sum_{i,j=1}^l y_i y_j a_i a_j \langle x_i, x_j \rangle \\ \text{s. t. } \sum_{i=1}^l a_i y_i = 0, 0 \leq a_i \leq C, i=1, \dots, l \end{cases} \quad (16)$$

求解式(16)得到优化系数 a_i ,其中不为零的 a_i 对应的样本为支持向量,然后根据 KKT 条件,得到参数 b ,则最优分类函数为:

$$f(x) = \text{sign} \left[\sum_{i=1}^n y_i a_i \langle x_i, x \rangle + b \right] \quad (17)$$

对于非线性问题,引入核函数 $K(x_i, x_j)$ 将样本空间映射到高维特征空间中再进行求解,可得非线性问题的指示函数为:

$$f(x) = \text{sign} \left[\sum_{i=1}^n y_i a_i K(x_i, x_j) + b \right] \quad (18)$$

径向基(RBF)核函数具有较好的学习能力,对于高维、低维或大样本、小样本均适用,且具有较宽的收敛域,与其它核函数相比参数少,易于掌握,许多文献都已证明了其良好的应用性。故本文选取径向基核函数作为支持向量分类器的核函数:

$$K(x_i, x_j) = \exp \left(- \frac{\|x_i - x_j\|^2}{\sigma^2} \right) \quad (19)$$

其中, σ 为核参数,是 RBF 核的宽度,实际上是隐含地改变映射函数,从而改变样本特征子空间分布的复杂程度。

本文进行的 SVM 参数优化即指寻找最优的参数组合 (C, σ) ,使 SVM 具有最好的分类性能。

3.2 SVM 参数优化步骤

本文通过 CDE 算法优化 SVM 的惩罚参数 C 和核参数 σ 来提高 SVM 的分类准确率。基于 CDE 算法的 SVM 参数优化具体步骤为:

Step1 初始化 CDE 算法参数以及 SVM 惩罚参数 C 和核参数 σ 的上下限值,并以此随机产生 (C, σ) 。

Step2 以当前的 (C, σ) 参数组合值作为 SVM 的参数,利用 SVM 对样本数据进行训练和检验,并得到检验结果,即样本的分类结果。

Step3 将 Step2 得到的分类结果与实际分类结果进行对比,计算目标函数值是否达到理论最优值或迭代次数 g 是否等于最大迭代次数 g_{\max} ,是则输出结果;否则执行下一步。

Step4 $g = g + 1$,进入下一代进化。

Step5 运用 CDE 算法对当前种群进行参数寻优。

Step6 计算产生新的 (C, σ) 并转至 Step2。

Step7 得到 SVM 最优参数 (C, σ) ,利用 SVM 对样本数据进行训练和检验,以此进行分类。

其具体流程如图1所示。

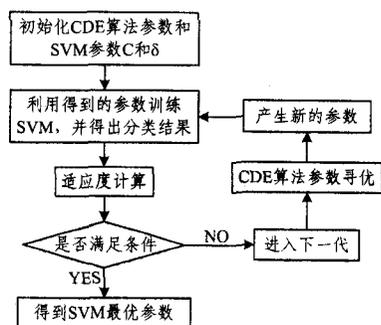


图1 基于 CDE 算法的 SVM 模型参数选择流程

3.3 仿真和实验分析

为了验证 CDE 算法对支持向量机参数优化的可行性和有效性,本文选取 UCI 数据库中的 wine 数据在 Matlab 环境下进行实验验证。wine data set 数据中包含 178 个样本,每个样本含有 13 个特征分量(化学成分),在这 178 个样本中,1~59 属于第一类(类别标签为 1),60~130 属于第二类(类别标签为 2),131~178 属于第三类(类别标签为 3)。现将每个类别分成两组,重新组合数据,一部分作为训练集(train_wine),一部分作为测试集(test_wine);训练集由第一类的 1~30、第二类的 60~95、第三类的 131~153 组成;测试集由各类余下样本组成。

(下转第 48 页)

- [11] Fu S, Desmarais M C. Fast Markov blanket discovery algorithm via local learning within single pass[C]// 21st Conference of the Canadian Society for Computational Studies of Intelligence (Canadian AI). Springer, 2008
- [12] Zeng Y X, Xiang H Y, Mao H. Dynamic ordering-based search algorithm for Markov blanket discovery[C]// 15th Pacific-Asia Conference on Data Mining, 2011. Shenzhen, China: Springer, 2011
- [13] Acid S, De Campos L M, Castellano J G. Learning Bayesian network classifiers; Searching in a space of partially directed acyclic graphs[J]. Machine Learning, 2005, 59(3): 213-235
- [14] Fu Shun-kai, Minn M C D S, Lv Tian-yi. A Survey of Advances

- [15] Koller D, Friedman N. Probabilistic graphical models: Principles and Techniques[M]. MIT Press, 2009
- [16] Bromberg F, Margaritis D, Honavar V. Efficient Markov network structure discovery using independence tests[J]. Journal of Artificial Intelligence Research, 2009, 35(1): 449-484
- [17] Fu S, Desmarais M C. Tradeoff analysis of different Markov blanket local learning approaches[C]// 12th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD). Osaka, Japan; Springer, 2008
- [18] Duda R O, Hart P E. Pattern Classification and Scene Analysis [M]. John Wiley & Sons, 1973-2-9
- [19] Zhang H, Jiang L, Su J. Hidden Naive Bayes[C]// AAAI. 2005

(上接第 21 页)

对 CDE 算法中的参数进行初始化: 最大进化代数 $g_{\max} = 100$, 种群规模 $Np = 40$, 其他参数依据前文设定。SVM 惩罚参数 C 和核参数 σ 均设置在 $[2^{-10}, 2^{15}]$ 。最终的仿真结果如图 2—图 4 所示。

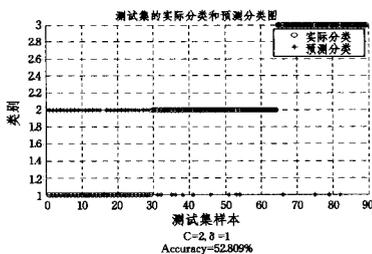


图 2 基于 SVM Wine data set 数据分类图

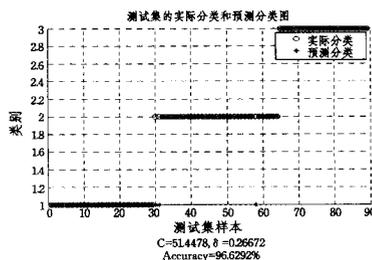


图 3 基于 DE-SVM Wine data set 数据分类图

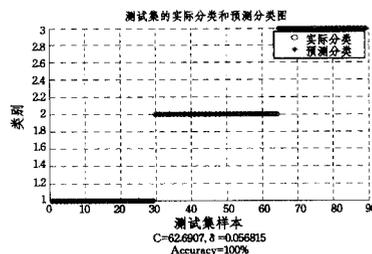


图 4 基于 CDE-SVM Wine data set 数据分类图

通过图 2 可看出,在不使用任何算法对 SVM 进行参数优化(默认参数)的前提下, SVM 对 Wine data set 的分类存在随机性而且效果比较差;图 3 中经过 DE 优化的 SVM 分类准确率达到了 96.63%,但仍有个别样本点未能正确分类;而图 4 中基于 CDE-SVM 模型分类准确率达到了 100%,所有样本点全部得到了准确分类,由此可以看出 CDE 算法在数据分类中也同样体现出了有效性和优越性。

结束语 本文将克隆选择算法引入差分进化算法中,提

出了基于克隆选择的差分进化算法,并将该算法应用于支持向量机的参数优化中。实验结果表明,基于克隆选择的差分进化算法不仅能有效避免早熟收敛现象,而且具有较强的寻优能力,同时经过 CDE 算法优化过的 SVM 分类精度得到了很大的增强,进一步提高了 SVM 的学习和泛化能力。

参 考 文 献

- [1] 赵海洋,徐敏强,王金东.改进二叉树支持向量机及其故障诊断方法研究[J].振动工程学报,2013(5):764-770
- [2] 彭光金,司海涛,俞集辉,等.改进的支持向量机算法及其应用[J].计算机工程与应用,2011,47(18):218-211
- [3] 于明,艾月乔.基于人工蜂群算法的支持向量机参数优化及应用[J].光电子·激光,2012,23(2):374-378
- [4] 庄平,白振林,许云峰.基于蚁群算法的支持向量机参数选择方法研究[J].计算机仿真,2011,28(5):216-219
- [5] Das S, Suganthan P N. Differential evolution: a survey of the state-of-the-art[J]. IEEE Transactions on Evolutionary Computation, 2011, 15(1): 4-31
- [6] Angira R, Babu B V. Optimization of process synthesis and design problems: a modified differential evolution approach [J]. Chemical Engineering Science, 2006, 61(14): 4707-4721
- [7] Babu B V, Angira R. Modified differential evolution (MDE) for optimization of nonlinear chemical processes [J]. Computer and Chemical Engineering, 2006, 30(6): 989-1002
- [8] de Castro L N, Tmanis J. Artificial Immune Systems: A New Computational Intelligence Approach [M]. British: Springer Press, 2002
- [9] Leandro N de C, Fernando J Von Z. Learning and optimization using the clonal selection principle[J]. IEEE Transactions on Evolutionary Computation, 2002(3): 239-251
- [10] 张向荣,焦李成.基于免疫克隆选择算法的特征选择[J].复旦学报(自然科学版),2004,43(5):926-929
- [11] 王俊,田玉玲.用于入侵检测的动态克隆选择算法的研究[J].计算机与数字工程,2010(6):108-110
- [12] 刘倩,仇宾.基于克隆选择算法的花卉图像分割[J].计算机工程与应用,2012,48(14):185-189
- [13] 徐佳,张卫.人工免疫系统中的抗体生成与匹配算法[J].计算机工程,2010,36(9):181-183
- [14] 胡超杰,章斌.一种采用克隆选择的免疫差分进化算法[J].计算机应用研究,2013,30(6):1640-1642