

基于改进烟花算法的 SVM 特征选择和参数优化

沈永良, 宋 杰, 万志超

(安徽大学 计算机科学与技术学院, 安徽 合肥 230601)

摘 要: 本文提出了一种改进的基于烟花算法的 SVM 特征选择和参数优化算法. 该算法针对特征选择问题的 0-1 特性, 使用二进制编码的烟花算法, 采用基于 RBF 核函数的 SVM, 在选取尽可能少的特征数目的同时提高了分类准确率. 通过 UCI 数据仿真, 对比结果表明: 该方法避免了过早成熟而陷入局部最优的问题, 可有效地找出合适的特征子集及 SVM 参数, 并取得较好的分类效果.

关键词: 二进制编码; 烟花算法; 特征选择; 参数优化

中图分类号: TP393

文献标识码: A

文章编号: 1000-7180(2018)01-0021-05

Improved Fireworks Algorithm for Support Vector Machine Feature Selection and Parameters Optimization

SHEN Yong-liang, SONG Jie, WAN Zhi-chao

(School of Computer Science and Technology, Anhui University, Heifei Anhui 230601, China)

Abstract: In this paper, we propose a Fireworks Algorithm based method to improve the performance of feature selection and parameters optimization in training SVM. For the 0-1 characteristic of feature selection, the binary coding Fireworks Algorithm and RBF kernel function based SVM are used to improve the accuracy of classification with less features. Compared to previous works, the proposed method can avoid being mature and falling into a local value, and it can effectively find the appropriate feature subset and parameters to get better performance of classification in UCI dataset.

Key words: binary encoding; fireworks algorithm; feature selection; parameters optimization

1 引言

支持向量机(support vector machine, SVM)是通过寻求结构风险最小化来实现实际风险最小化的机器学习算法,其目标是追求在有限信息的条件下得到最优结果. 以往困扰机器学习方法的很多问题,均可通过支持向量机在一定程度上得到解决,其目前前在分类、预测方面获得了广泛的应用^[1-3]. 但是在实际应用中,支持向量机模型参数(如惩罚因子、核函数参数)和特征选择对分类性能有着很大的影响. 所以如何通过优化支持向量机参数和选择特征子集来获得最优分类性能,是目前支持向量机分类问题研究领域的热点. SVM 参数和特征选择问题,

其本质就是优化问题,在相关研究工作中,网格搜索^[4]方法易于实现,但是对于大样本问题其搜索能力差且不能进行特征选择. 文献[5]提出了基于遗传算法的支持向量机的特征选择和参数优化的方法(GA-SVM),GA 虽然克服了容易陷入局部最优的缺点,但是搜索速度慢. 烟花算法是一种新的群体智能优化算法^[6-7],其广泛应用于图像处理、无线传感网络等领域^[8-9]. 本文在传统烟花算法的基础上,构建了离散化的二进制烟花算法,用于特征选择和 SVM 的参数优化,并取得了较好的优化效果.

2 支持向量机原理

支持向量机(SVM)的基本思想是在数据集线

性可分的情况下找出最优分类面. 所谓最优分类就是对于两种线性样本, 在原始空间中, 既可以直接搜索两种样本, 又能使两类样本正确分离, 同时保证最大分类间隔.

在样本空间中, 划分超平面的线性方程为:

$\omega^T x + b = 0$, 其中 $\omega = (\omega_1; \omega_2; \dots; \omega_d)$ 为法向量, 决定了超平面的方向; b 是位移项参数, 决定了超平面与远点之间的距离. 假设有线性可分样本 (x_i, y_i) , $i = 1, 2, \dots, n$, $x_i \in R^d$, $y_i \in \{-1, 1\}$, 此时分类间隔等于 $2 / \|\omega\|$, 使得分类间隔最大, 等价于使 $\|\omega\|^2$ 最小. 满足式(1)且使得 $\|\omega\|^2/2$ 最小的分类面就是最优分类面.

$$y_i(\omega \cdot x_i + b) - 1 \geq 0, \quad i = 1, 2, \dots, n \quad (1)$$

然而在现实任务中, 原始样本空间内往往并不存在一个能正确划分两类样本的超面. 所以需要选择一个核函数, 将非线性的训练数据映射到高维空间, 接着在这个空间中构造最优分类超平面, 以解决在原始空间线性不可分的问题. 核函数的作用是对特征进行从低维到高维的转换, 其中使用最广泛的是高斯核函数, 它在实际使用中具有相当的灵活性, 公式为:

$$\kappa(x_i + x_j) = \exp(-g * \|x_i + x_j\|^2) \quad (2)$$

3 基于二进制编码的烟花算法

传统烟花算法由爆炸算子、变异算子和选择策略构成, 为适应支持向量机特征选择的 0-1 特性, 根据连续烟花算法的优化思想, 本文构造了离散化的二进制烟花算法.

3.1 二进制烟花

本文算法中初始烟花采用二进制编码, 共分为两个部分. 在 SVM 方面, 采用 RBF 作为核函数, 需要优化两个参数, 即惩罚因子 c 和核参数 g , 在实验中参数位串长度均设为 20. 在特征选取方面, 特征子集选择位串长度由特征维数决定, 二进制编码中的 1 和 0 分别代表特征被选中和未被选中.

3.2 爆炸算子

针对二进制烟花算法特性, 引入二进制转换算子 $\Phi(x_i, R, L)$. 假设烟花二进制编码为: $x_i = \{\mu_{1i}, \mu_{2i}, \dots, \mu_{Ui}, \nu_{1i}, \nu_{2i}, \dots, \nu_{Vi}, \rho_{1i}, \rho_{2i}, \dots, \rho_{Pi}\}$. 其中, μ 为参数 c 位串, ν 为参数 g 位串, ρ 为特征位串. U, V 分别为参数 c 和 g 位串长度, P 为样本数据特征数目. R 表示二进制字符转换集, L 表示二进制转换步长, 所以二进制转换算子表示为以 x_i 元素为基础字符串, 从转换集 R 中随机选出 L 个二进制字符进行 0-

1 转置. 例如 $x_i = \{1, 1, 0, 1\}$, 转换集 $R = \{2, 3, 4\}$, 转化步长 $L = 2$, 则通过二进制转换算子 $\Phi(x_i, R, L)$ 等概率生成的二进制字符串为 $\{1, 0, 1, 1\}, \{1, 0, 0, 0\}$ 或 $\{1, 1, 1, 0\}$.

3.2.1 爆炸强度

爆照强度是烟花算法中爆炸算子的核心, 模拟现实生活中烟花爆炸的方式. 任何一个烟花爆炸时, 都会产生一簇火花. 通过爆炸强度让适应度数值好的烟花产生的火花个数较多,

这样可以避免寻优时火花总是在最优值附近摆动, 而无法精准地找到最优值. 而对于适应度差的火花, 减少其产生的火花数目, 适度地对其余空间进行探索, 避免“早熟”.

在烟花算法中, 火花个数产生的公式为:

$$S_i = m \frac{Y_{\max} - f(x_i) + \epsilon}{\sum_{i=1}^n (Y_{\max} - f(x_i)) + \epsilon} \quad (3)$$

式中, S_i 为第 i 个烟花产生的火花数目; m 是常数, 用来限制产生的火花总数; $Y_{\max}$ 是当前种群中适应度最差的个体的适应度值; $f(x_i)$ 为个体 x_i 的适应度值; ϵ 为一个极小的常数, 以避免出现分母为零的情况.

为了限制烟花爆炸产生火花的数量过多或过少, 算法中为每个烟花设定如下产生火花数量的限制公式:

$$\hat{S} = \begin{cases} \text{round}(am), & S < am \\ \text{round}(bm), & S > bm, a < b < 1 \\ \text{round}(S_i), & \text{其他} \end{cases} \quad (4)$$

3.2.2 爆炸幅度和位移操作

爆炸幅度的基本思想是让适应度值好的烟花爆炸幅度减小, 使结果有效地收敛到各个极值, 直至找到最优值, 而适应度较差的点, 往往离最优值较远, 让其产生大幅度的变异, 才能有效地到达最优值附近. 第 i ($i = 1, 2, \dots, n$) 个烟花爆炸的半径用式(5)表示:

$$A_i = d \frac{f(x_i) - Y_{\min} + \epsilon}{\sum_{i=1}^n (f(x_i) - Y_{\min}) + \epsilon} \quad (5)$$

式中, A_i 为第 i 个烟花的爆炸幅度范围, d 为烟花二进制编码的维数, 表示最大的爆炸幅度; 参数 $Y_{\min}$ 表示为当前种群中适应度最好的个体的适应度值; $f(x_i)$ 与 ϵ 的意义与式(3)相同.

在确定爆炸幅度后, 需要对爆炸产生的火花进行位移操作, 传统烟花算法使用随机位移的方法:

$$\Delta x_k^i = x_k^i + A_i \cdot \text{rand}(-1, 1) \quad (6)$$

即在 n 个烟花的基础上,首先选取 z 个维度的坐标更新,然后对烟花 $i(i = 1, 2, \dots, n)$ 爆炸产生的第 $j(j = 1, 2, \dots, S_i)$ 个火花的第 $k(k = 1, 2, \dots, z)$ 维进行位移操作。

本文中烟花是由惩罚参数 c 、核参数 g 和特征子集组合而成的二进制结构,结合上文提到的二进制转换算子,烟花 x_i 爆炸后在 A_i 半径内产生第 j 个火花的方法为:

$$x_j = \Phi(x_i, R_0, L_i) \quad (7)$$

式中, $R_0 = \{1, 2, \dots, d\}$ 对应烟花二进制字符编码位置, L_i 为爆炸算子步长。爆炸算子步长是在爆炸半径 A_i 内随机生成的,如下方法:

$$L_i = \varepsilon + \text{rand}(0, A_i) \quad (8)$$

3.3 变异算子

为了进一步提高种群的多样性,在二进制烟花算法中引入变异算子以扩展寻优空间,并避免算法落入局部最优值。生成变异火花采取的策略是适应度越好,即越是接近最优烟花,则其越需要通过变异来增强种群多样性。生成变异火花的方式同 3.2.2 节的方法类似。以烟花 x_i 为父代的变异半径和变异步长分别为:

$$A_i^M = m \cdot \frac{f(x_i) - Y_{\min} + \varepsilon}{\sum_{i=1}^n (f(x_i) - Y_{\min}) + \varepsilon} \quad (9)$$

$$L_i^M = \varepsilon + \text{rand}(0, A_i^M) \quad (10)$$

变异算子的转换集 $R_M = \{R_1, R_2, \dots, R_j, \dots, R_M\}$ 的选取策略是取烟花 x_i 与当前最优烟花相同的部分,如下式所示:

$$R_j = j, x_{i,j} = x_{\text{best},j} \quad (11)$$

式中, x_{best} 表示当前适应度最好的烟花, j 表示烟花 x_i 和烟花 x_{best} 在二进制编码中相同的位置坐标。

由上面三式可得 x_i 为父代生成的变异火花为:

$$x_i^M = \Phi(x_i, R_M, L_i^M) \quad (12)$$

3.4 选择策略

运用爆炸算子和变异算子后,在整个空间就存在了烟花、位移火花、变异火花三种类型的个体,接下来需要选择出一部分作为下一代的烟花。选择的方法是每次留下最优的个体,再从剩余的个体中选择出 $n-1$ 个烟花,用到的是基于曼哈顿距离的选择策略:

$$P_i = \frac{\sum_{j \in K} d(x_i, x_j)}{\sum_{i \in K} \sum_{j \in K} d(x_i, x_j)} \quad (13)$$

式中, $d(x_i, x_j)$ 表示烟花 i 和烟花 j 的曼哈顿距离, K 表示所有个体(烟花、位移火花、变异火花)的集合。

3.5 适应度函数

由于算法的目标是在 SVM 的分类准确率和对应样本选取特征数目之间寻求最佳折中。即通过迭代选出一个烟花个体在选取较少特征数目的同时使得分类器产生的分类准确率越高,其适应度就越高。适应度函数定义如下式:

$$F = W_1 * (\sum_{i=1}^N f_i)^{-1} + W_2 * acc \quad (14)$$

式中, W_1 为特征数目的权重, W_2 为分类准确度的权重, f_i 取 1 或 0,分别代表对应特征被选中或没被选中, acc 为分类准确度。

3.6 算法流程

基于二进制烟花算法的 SVM 参数优化和特征选择算法的流程,具体步骤如下:

输入:初始种群数 n ,火花限制参数 m ,公式(4)中限制参数 a 和 b ,数据集 D ,惩罚函数 $c_{\min}$ 和 $c_{\max}$,核参数 $g_{\min}$ 和 $g_{\max}$,最大迭代次数 T 。

输出:优化的特征子集, SVM 参数(c, g)及对应的分类准确率。

步骤 1:在特定空间随机产生 n 个烟花,根据二进制编码方案将每个个体转化为对应的 SVM 的参数 c, g 和数据的特征子集,利用支持向量机进行学习和训练,根据式(14)计算出个体对应的适应度值 F 。

步骤 2:根据式(3)计算出每个烟花生成的火花个数 S_i ,并参照式(4)来避免烟花爆炸产生过多或者过少的火花。

步骤 3:根据式(8)计算出爆炸算子步长,参照式(5)确定每个烟花的爆炸幅度并通过式(7)生成以 x_i 为父代的火花。

步骤 4:根据本文提出的变异操作方法(见 3.3 节),对烟花进行变异操作。

步骤 5:保留最优适应度的个体为一个烟花,其余 $n-1$ 个依照式(13)确定。

步骤 6:检查结束条件。若满足,则寻优过程结束,返回当前最优的参数 c 、参数 g 、特征子集及对应的分类准确度;否则 $T = T+1$,转至步骤 2。设定结束条件为寻优达到最大迭代次数 T 。

4 实验与分析

为了考察本文算法的可行性和有效性,从 UCI

机器学习知识库中选取多个数据集,采用平台是 Intel Corei5 2.7 GHz CPU,8G RAM,OSX 10.11 操作系统,开发环境为 Matlab2016b,支持向量机软件为 LIBSVM^[10].

在实验中,参数 c 的搜索范围为 $(2^{-5}, 2^{15})$, 参数 g 的搜索范围为 $(2^{-5}, 2^3)$, 算法维度 d 为 c, g 位串长度和加上特征数目,初始烟花数目 $n = 5, m = 50, a = 0.04, b = 0.8$, 最大爆炸幅度即算法维度 d 为 c, g 位串长度和加上特征数目,最大迭代次数 $T = 200$, 分类准确度权重 $W_1 = 0.8$, 特征数目权重 $W_2 = 0.2$. 数据集、类别、样本数和特征数如表 1 所示.

表 1 实验所用数据集				
序号	数据集	类别	样本数	特征数
1	Australia	2	690	14
2	Sonar	2	208	60
3	Wine	3	178	13
4	Ionosphere	2	351	34
5	Breast cancer	2	699	10
6	Vowel	11	990	13
7	German	2	1 000	24

为了验证算法的效果,针对 UCI 数据集,实验中采用 10 次折迭交叉来确认分类精确度. 其中 9 个子集为训练模型,剩余的 1 个子集为独立的测试数据集. 程序运行 10 次,每 1 个子集依次作为测试数据集,求得 10 次运行准确率的平均数,即可得到该数据集在此次实验中的选择特征数和分类准确率,得到的实验数据均用平均数±标准差的形式表示. 除了和网格搜索算法对比外,本文还引入了分别基于 GA 和 PSO 的特征选择和 SVM 参数联合优化算法^[11]和本文算法作对比,记为 GA-SVM 和 PSO-SVM.

由图 1 可以看出,本文提出方法通过迭代搜索到的适应度最大,其特征选择和参数优化能力对比 GA-SVM 和 PSO-SVM 能力更强,其中 GA-SVM 过早进入了早熟状态而陷入局部最优.

如表 2 所示,网格搜索算法在样本相对简单的情况下往往能得到的较高的准确率,当样本趋于复杂化时,由于其无法进行特征选择,分类准确率下降明显. 本文方法与之相比无论在准确率方面,还是在特征数目方面,都具有更好的性能.

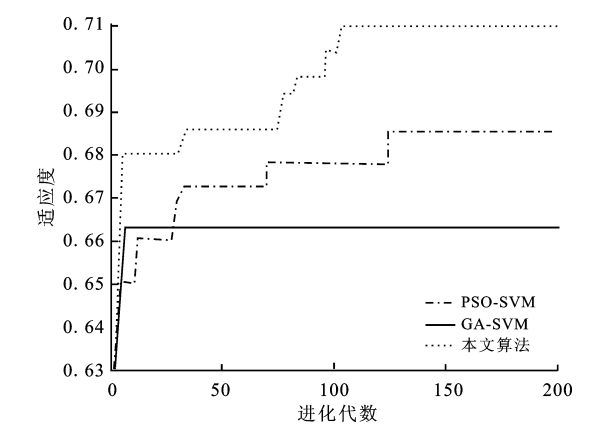


图 1 三种算法在数据集 German 上的最佳适应度

表 2 预测结果对比			
数据集	二进制烟花算法		网格搜索
	选择特征数	分类准确率	分类准确率
Australia	6.3±2.05	90.54±2.35	84.74±4.52
Sonar	24.4±3.35	96.04±2.14	88.50±8.32
Wine	5.3±0.52	97.19±1.12	93.16±2.88
Ionosphere	14.2±3.52	96.02±1.64	90.29±3.56
Breast cancer	4.6±0.94	98.86±1.86	92.43±2.50
Vowel	8.6±2.28	96.54±3.21	91.09±4.77
German	13.1±2.12	85.84±1.98	76.90±1.73

表 3 是本文算法与 GA-SVM 算法和 PSO-SVM 算法在 7 个 UCI 数据集上的对比实验结果, 可以看出,在数据集 Australia、Sonar、Ionosphere、Breast cancer、Vowel、German 上,本文提出的方法获得了更高的分类准确率、最少的特征选择. 以特征数目较多的数据集 Sonar 为例,本文方法选出平均特征数目为 24.4,相较于 GA-SVM 算法特征数目少了 25.4%,分类准确率提高了 3.04 个百分点. 在数据集 Wine 上,本文方法选择特征数虽略高于 GA-SVM 算法,但在分类准确率方面提高了 2.84 个百分点. 与 PSO-SVM 的算法相比,本文方法在数据集 Breast cancer 地分类准确度和数据集 Vowel 的特征选择数目虽略差于 PSO-SVM 算法,但是在其他数据集上,本文方法均获得了更好的效果.

5 结束语

本文在连续烟花算法的启发下,将传统烟花算法的爆炸算子离散化,构建了二进制烟花算法. 该算法对比传统优化算法,避免了优化过程中早熟和陷入局部最优的问题,实验表明该算法能得到更高的适应度值,精确得到了待分类数据的特征

表 3 预测结果对比

数据集	二进制烟花算法		遗传算法		粒子群算法	
	选择特征数	分类准确率	选择特征数	分类准确率	选择特征数	分类准确率
Australia	6.3±2.05	90.54±2.35	6.7±2.16	86.81±3.64	8.9±2.21	90.03±4.12
Sonar	24.4±3.35	96.04±2.14	32.7±4.00	93.00±2.36	35.1±2.96	94.26±5.13
Wine	5.3±0.52	97.19±1.12	4.6±0.72	94.35±1.76	7.46±1.86	97.10±2.03
Ionosphere	14.2±3.52	96.02±1.64	15.4±3.32	92.57±2.02	18±2.88	95.01±2.16
Breast cancer	4.6±0.94	98.86±1.86	4.9±0.99	94.04±1.98	6.00±1.29	99.18±2.01
Vowel	8.6±2.28	96.54±3.21	9.0±1.38	94.67±2.78	7.10±0.64	96.20±3.61
German	13.1±2.12	85.84±1.98	14.3±2.70	80.80±2.10	17.4±3.28	82.34±1.76

子集和 SVM 参数. 对于算法在实际问题中的应用,也是下一步要研究的方向.

参考文献:

[1] 李婷婷,姬东鸿. 基于 SVM 和 CRF 多特征组合的微博情感分析[J]. 计算机应用研究, 2015, 32(4): 978-981.

[2] 李小琳,孙玥,刘洋. 基于 SVM 修正的模糊时间序列模型在沪指预测中的应用[J]. 中国科学技术大学学报, 2016, 46(3): 238-246.

[3] Uricar M, Timofte R, Rothe R, et al. Structured output svm prediction of apparent age, gender and smile from deep features[C]//Chalearn Looking at People and Faces of the World: Face Analysis Workshop and Challenge. [s. l.]. 2016:730-738.

[4] Hsu C W, Chang C C, Lin C J. A Practical guide to support vector classsication[J]. 国立台湾大学资讯工程学系, 2003, 67(5):58-67.

[5] Huang C L, Wang C J. A GA based feature selection and parameters optimization for support vector machines. expert systems with applications[J]. Expert Systems with Applications, 2006, 31(2):231-240.

[6] Tan Y, Zhu Y. Fireworks algorithm for optimization [C]//International Conference on Advances in Swarm

Intelligence. Berlin, Springer-Verlag, 2010:355-364.

[7] 谭营. 烟花算法引论[M]. 北京:科学出版社, 2015.

[8] Misra P R, Si T. Image segmentation using clustering with fireworks algorithm [C] // Intelligent Systems and Control (ISCO), 2017 11th International Conference on, Bilbao, Spain, IEEE, 2017: 97-102.

[9] Arsic A, Tuba M, Jordanski M. Fireworks algorithm applied to wireless sensor networks localization problem[C]// IEEE Congress on Evolutionary Computation. Press. USA, IEEE, 2016:4038-4044.

[10] Chang C C, Lin C J. LIBSVM: A library for support vector machines[J]. Acm Transactions on Intelligent Systems & Technology, 2011, 2(3):1-27.

[11] 张俊才,张静. 使用粒子群算法进行特征选择及对支持向量机参数的优化[J]. 微电子学与计算机, 2012, 29(7):138-141.

作者简介:

沈永良 男, (1992-), 硕士研究生. 研究方向为嵌入式系统智能计算. E-mail: syl296022756@qq.com.

宋杰 男, (1966-), 博士研究生, 副教授. 研究方向为嵌入式系统、智能计算、生物信息.

万志超 男, (1991-), 硕士研究生. 研究方向为嵌入式系统、智能计算.