DOI:10.19651/j.cnki.emt.1802155

基于 BGSA 算法的 SVM 分类模型设计研究

赵东升[1,2] 李艳军[1]
(1.北京京航计算通讯研究所 北京 100074;2.北京市涉密信息载体安全管理工程技术研究中心 北京 100074)

摘 要:为了设计性能更优的支持向量机(SVM)分类模型,对影响其分类性能的参数和样本特征子集进行优化选择,对支持向量机理论和万有引力搜索算法(GSA)进行了研究,提出了一种基于二进制万有引力搜索算法(BGSA)的支持向量机分类模型构建方法,能够对影响支持向量机分类性能的相关参数及有效样本特征子集同时进行优化选择,获得最优组合解,并通过实验对其有效性进行了对比分析和验证。实验结果表明,所提出的 BGSA-SVM 分类模型能够有效提高支持向量机的分类性能,可进一步推广到工程实际中应用。

关键词:万有引力搜索算法;群体智能;支持向量机;参数优化
中图分类号:TN98 **文献标识码**:A **国家标准学科分类代码**:520.1040

Research on the SVM classification model design based on BGSA

Zhao Dongsheng[1,2] Li Yanjun[1]
(1.Beijing Jinghang Research Institute of Computing and Communication, Beijing 100074, China; 2. The Classified Information Carrier Safety Management Engineering Technology Research Center of Beijing, Beijing 100074, China)

Abstract: In order to design a support vector machine (SVM) classification model with better performance, the parameters and sample feature subsets that affect its classification performance are optimized, and the support vector machine theory and gravitational search algorithm (GSA) were studied. The optimal combination solution can be obtained by simultaneously optimizing the relevant parameters and effective sample feature subsets which affect the classification performance of SVM. Its effectiveness is compared and verified by experiments. The experimental results show that the proposed BGSA-SVM classification model can effectively improve the classification performance of support vector machines, which can be further extended to engineering applications.

Keywords: gravitational search algorithm (GSA); group intelligence; support vector machine (SVM); parameter optimization

0 引 言

支持向量机(support vector machine, SVM)是 20 世纪 90 年代发展起来的一种具有较好学习能力和泛化能力的机器学习算法。该算法的构建是基于统计学习理论的结构风险最小化原则,它在解决小样本、高维度和非线性的问题中具有独特的优势。

一些学者的研究表明,SVM 的分类性能与构建 SVM 分类模型时所选的核函数参数值、惩罚系数值和样本特征子集都具有较大的关系。胡冬梅等[1]用粒子群优化(particle swarm optimization, PSO)算法对支持向量机的参数进行优化选择,并通过实验验证了 PSO-SVM 方法作为液晶可变延迟器(LCVR)相位延迟特性标定手段的有效性。梅恒荣等[2]提出了一种基于改进粒子群算法优化 SVM 参数的分类器,并通过实验对其模拟电路故障诊断性能进行了分析验证。王福忠等[3]提出了一种改进粒子群优化支持向量机的故障诊断方法,并将其应用到变压器分接开关的故障诊断中,取得了较好的仿真试验效果。谈利芳等[4]提出利用遗传算法(genetic algorithm, GA)进行特征降维,结合支持向量机进行语音多类情感识别,并通过实验证明了该方法的有效性。黄挺等[5]利用网格搜索法对支持向量机进行参数寻优,取得了较好的彩色图像分割效果。此外,其他学者的相关研究也表明,群智能优化算法在 SVM 模型的构建中,能够展现出较强的寻优能力,获得性能较好的 SVM 分类模型。

万有引力搜索算法(gravitational search algorithm, GSA)是由 Rashedi 等[6]于 2009 年提出的一种群体智能优化算法。该算法的本质是模拟自然界中常见的万有引力现象,将万有引力现象演化成迭代寻优的过程。GSA 算法对种群中每个粒子都赋予质量,根据万有引力定律得到种群

收稿日期:2018-10-08

粒子之间的相互作用力,而以这种粒子之间的相互作用力为原动力,通过种群迭代寻优最终收敛到最优位置,具有较强的全局搜索能力。近年来也有不少学者展开了与 GSA 算法相关的研究[7-15],进一步证实该算法具有较好的寻优性能。然而在实际中,往往会遇到许多离散问题,二进制万有引力搜索算法(binary gravitational search algorithm, BGSA)[11]正是一种用于解决这类问题的 GSA 算法。

本文将在研究 SVM 及 BGSA 算法的基础上,针对影响 SVM 分类模型性能的主要因素,提出一种基于 BGSA 算法的 SVM 分类模型构建方法,并通过实验进行对比分析,对该方法的有效性进行验证。

1 SVM

SVM 是一种建立在统计学理论基础上的具有较好学习能力和较强泛化能力的机器学习算法。标准支持向量机分类器是针对两分类问题的,对于多分类问题需构建多分类器来实现。

在两分类问题中,SVM 需要通过求解两类数据样本之间的最优分类超平面来实现分类,该问题即为求解以下优化问题:

$$\begin{cases}\min\varphi(w)=\dfrac{1}{2}\parallel \boldsymbol{w}\parallel^2+C\sum\limits_{i=1}^{n}\xi_i\\ \text{s.t. } y_i(w\cdot x_i+b)-1+\xi_i\geqslant 0,\ \xi_i\geqslant 0,\\ i=1,2,\cdots,n\end{cases}\tag{1}$$

式中:$\boldsymbol{w}$ 为超平面的法向量;C 为错分样本的惩罚系数;ξ_i 近似表示被误分类的样本数;x_i 为待测样本;b 为偏差;$y_i\in\{+1,-1\}$。

为了在非线性情况下得到最优决策面,引入核函数将待处理的问题转化为某个高维空间中的线性问题,从而可以在变换之后的高维空间中求得最优分类面。

利用核函数在特征空间中求解最优分类超平面的优化问题可表示为:

$$\begin{cases}\max W(\alpha)=\sum\limits_{i=1}^{n}\alpha_i-\dfrac{1}{2}\sum\limits_{i,j=1}^{n}y_iy_j\alpha_i\alpha_jK(x_i\cdot x_j)\\ \text{s.t. }\sum\limits_{i=1}^{n}\alpha_iy_i=0,0\leqslant\alpha_i\leqslant C,i=1,\cdots,n\end{cases}\tag{2}$$

式中:$\alpha_i\geqslant 0$,$K(x_i\cdot x_j)$ 表示核函数。

相应的决策函数为:

$$f(x)=\operatorname{sgn}\left[\sum_{i=1}^{n}\alpha_i^*y_iK(x_i\cdot x)+b^*\right]\tag{3}$$

此外,已有学者的研究表明,核函数参数值、惩罚系数值和样本特征子集的选择均是影响 SVM 分类性能的主要因素,因此,本文将对其进行统一编码和优化选择,从而获得性能更好的 SVM 分类模型。

2 GSA 算法

GSA 算法是一种基于万有引力定律进行寻优的智能化方法,它将优化问题的解视为空间中运行的粒子,而粒子之间通过万有引力相互作用,使得较小的粒子朝着质量较大的粒子不断移动,从而通过粒子的这种不断运动来求解优化问题。

根据牛顿万有引力定律,粒子之间的作用力可以表示为[6]:

$$F=G\frac{M_1M_2}{R^2}\tag{4}$$

式中:F 表示引力大小;G 表示万有引力常数;M_1 和 M_2 分别表示两个互相作用的粒子的惯性质量;R 是这两个粒子之间的欧氏距离。

根据牛顿第二定律,粒子加速度的计算公式为:

$$\alpha=F/M\tag{5}$$

万有引力常数 $G(t)$ 的计算公式为:

$$G(t)=G_0\times e^{-\alpha t/T}\tag{6}$$

质量 $M_i(t)$ 的计算公式为[6]:

$$\begin{cases}m_i(t)=\dfrac{fit_i(t)-worst(t)}{best(t)-worst(t)}\\ M_i(t)=m_i(t)/\sum\limits_{j=1}^{N}m_j(t)\end{cases}\tag{7}$$

式中:$fit_i(t)$ 表示粒子的适应度值;$best(t)$ 表示时刻 t 的最佳解;$worst(t)$ 表示时刻 t 的最差解。

在 t 时刻,物体 j 在第 d 维上受到粒子 i 的引力计算公式为[6]:

$$F_{ij}^{\ d}(t)=G(t)\frac{M_{pi}(t)\times M_{aj}(t)}{R_{ij}(t)+\varepsilon}(x_j^d(t)-x_i^d(t))\tag{8}$$

式中:ε 表示一个无穷小的常量;$M_{pi}(t)$ 和 $M_{aj}(t)$ 分别表示粒子 i 和粒子 j 的惯性质量。

在 t 时刻,第 d 维上的粒子受到的作用力为[6]:

$$F_i^{\ d}(t)=\sum_{j\in Kbest,j\neq i}rand_jF_{ij}^{\ d}(t)\tag{9}$$

在算法迭代过程中,粒子的速度和位置的计算公式为:

$$\begin{cases}v_i^{\ d}(t+1)=rand_i\times v_i^{\ d}(t)+a_i^{\ d}(t)\\ x_i^{\ d}(t+1)=x_i^{\ d}(t)+v_i^{\ d}(t+1)\end{cases}\tag{10}$$

二进制万有引力算法(BGSA)是为解决实际中的离散问题而构建的,在算法迭代过程中,粒子速度转换计算公式为:

$$S(v_{ij}^k(t))=|\tanh(v_{ij}^k(t))|\tag{11}$$

粒子位置的更新规则为:

$$\begin{cases}if\ rand<S(v_{ij}^k(t+1))\\ x_{ij}^k(t+1)=complement(x_{ij}^k(t))\\ else\ x_{ij}^k(t+1)=x_{ij}^k(t)\end{cases}\tag{12}$$

BGSA 算法的主要流程如下:

1) 初始化粒子的位置与加速度,进行粒子种群大小、迭代次数等参数的初始化;

2) 计算每个粒子的适应度值,更新重力常数;

3) 计算每个粒子的质量及加速度;

4）计算每个粒子的速度，并进行转换；

5）按照粒子位置的更新规则更新粒子的位置；

6）判断是否满足算法迭代终止条件，如果不满足，继续迭代，否则，终止迭代，输出最终结果。

3　BGSA-SVM 分类模型

在利用 BGSA 进行寻优之前，首先要对 SVM 参数和样本特征集合进行统一编码，因此，该编码主要由两部分构成：1)SVM 参数的编码；2)样本特征集合的编码，如果与某个样本特征相对应的编码值为“1”，则表示该样本特征被选取，否则，表示该样本特征将被舍弃。

编码经过 BGSA 算法优化选择之后，需将第 1 部分提取并进行转换之后才能送入 SVM 分类模型。编码相互转换公式为：

$$P = \frac{Val_{\max} - Val_{\min}}{2^{len} - 1} \times d + Val_{\min} \tag{13}$$

式中：$Val_{\max}$ 和 $Val_{\min}$ 分别表示参数的最大值和最小值；len 是编码长度；d 是编码的十进制值。

综合考虑 SVM 分类精度和特征子集两部分因素，本文设计的 BGSA 适应度计算公式为：

$$fitness = w_1 \times Acc + w_2 \times \left(1 - \frac{Fse}{Fal}\right) \tag{14}$$

式中：Acc 表示分类正确率；w_1 和 $w_2 = 1 - w_1$ 分别表示分类正确率和特征子集的权重系数；Fal 和 Fse 分别表示特征集合的大小和选取的特征子集的大小。在优化选择过程中，如果能够选取较小的样本特征子集同时获得较高的分类正确率，很好的权衡这两部分，就能获得较高的适应度值。w_1 和 w_2 的值可根据实际应用情况进行设置。

本文提出的 BGSA-SVM 分类模型构建步骤可总结如下：

1）数据预处理：读取样本数据，并进行样本数据预处理；

2）编码：对 SVM 模型的待优化选择的参数和样本数据特征集合统一编码；

3）利用 BGSA 算法进行种群编码优化选择；

4）编码转换：通过编码转换得到 SVM 模型参数；

5）根据编码构建新的训练样本数据集合和测试样本数据集合；

6）进行训练得到新的 SVM 分类模型；

7）利用训练获得的 SVM 分类模型对测试数据集合进行分类；

8）进行适应度值评估；

9）迭代终止条件判断：如果满足，则终止迭代，否则，返回步骤 3)继续迭代；

10）输出最终结果。

4　实验结果及分析

利用二进制粒子群优化算法（binary particle swarm optimization，BPSO）构建 BPSO-SVM 分类模型，与本文提出的 BGSA-SVM 分类模型进行对比分析实验。在实验中，所选取的标准样本数据如表 1 所示，有两分类问题数据（Australian、Heart、Ionosphere），也有多分类问题数据（Iris、Wine、Pendigits）。

表 1　样本数据

序号	数据名称	数据类别数	数据样本总个数	特征数量
1	Australian	2	690	14
2	Heart	2	270	13
3	Ionosphere	2	351	34
4	Iris	3	150	4
5	Wine	3	178	13
6	Pendigits	10	7494	16

实验环境在 Windows 7 操作系统上构建，利用 MATLAB 平台进行算法编码与实现。在实验中，特征子集部分的权重系数 w_2 取值为 0.02，根据本文设计的适应度值计算方法，该部分的最大取值必然会小于 0.02，所以平均适应度值实验结果保留 4 位小数以便进行更精准的对比分析。共进行 10 次独立实验，两分类数据 Heart 和多分类数据 Pendigits 的平均迭代寻优曲线分别如图 1 和 2 所示，平均实验结果以均值±标准差的形式，如表 2 所示。

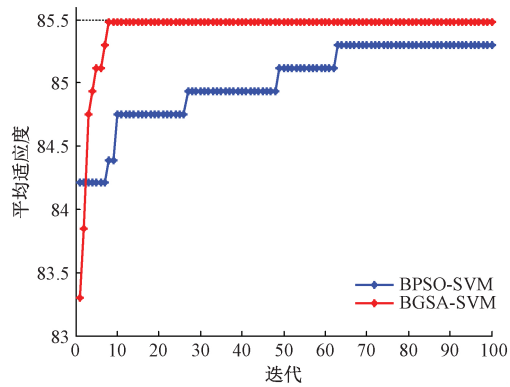

图 1　两分类数据 Heart 的平均迭代寻优曲线

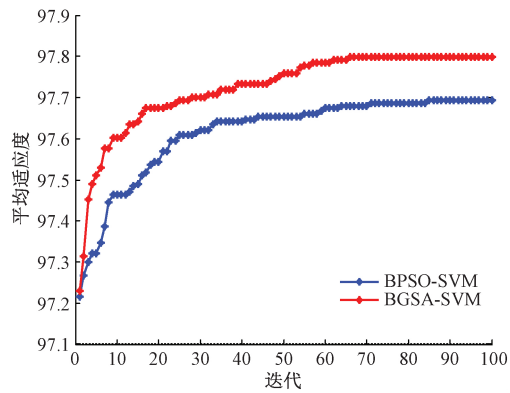

图 2　多分类数据 Pendigits 的平均迭代寻优曲线

表 2 实验结果

数据名称	特征数量	平均优选样本特征数量		平均适应度值	
		BPSO-SVM	BGSA-SVM	BPSO-SVM	BGSA-SVM
Australian	14	6.32±0.82	6.32±0.51	88.485 9±0.598 9	89.053 9±0.366 5
Heart	13	5.70±0.82	5.50±0.87	85.297 9±0.000 5	85.479 9±0.573 5
Ionosphere	34	7.60±1.35	2.70±0.82	98.001 1±0.005 5	98.007 4±0.002 1
Iris	4	1.00±0.00	1.00±0.00	98.015 0±0.000 0	98.015 0±0.000 0
Wine	13	3.00±0.00	2.80±0.42	98.005 1±0.000 0	98.005 8±0.001 4
Pendigits	16	13.91±1.06	13.90±0.32	97.709 1±0.075 7	97.790 5±0.020 6

从图 1 和 2 的平均迭代寻优曲线可知,与 BPSO-SVM 分类模型相比,本文提出的 BGSA-SVM 分类模型在迭代寻优过程中,具有较快的收敛速度,较强的跳出局部最优能力和较好的全局寻优能力。

从表 2 的实验结果可知,本文提出的 BGSA-SVM 分类模型能够优选出相对较少的样本特征数量同时获得较大的适应度值,可以更好地权衡样本特征子集与分类正确率,无论针对两分类问题还是多分类问题均能获得较好的结果,具有更好的性能。

5 结 论

本文在研究影响 SVM 分类性能因素的基础上,引入 BGSA 对 SVM 模型参数及样本数据特征子集同时进行优化选择,提出了 BGSA-SVM 分类模型构建方法,并通过对比分析实验对其有效性进行了验证。实验结果表明,本文提出的 BGSA-SVM 分类模型能够有效提高支持向量机的分类性能,可进一步推广到工程实际中应用。

参考文献

[1] 胡冬梅,宋路,牛国成.基于支持向量机的波片相位延迟测量新方法[J].仪器仪表学报,2016,37(7):1517-1523.

[2] 梅恒荣,殷礼胜,刘冬梅,等.改进粒子群算法优化的 SVM 模拟电路故障诊断[J].电子测量与仪器学报,2017,31(8):1239-1246.

[3] 王福忠,石秀立.改进 PSO-SVM 算法的变压器分接开关故障诊断[J].电子测量技术,2016,39(11):190-194.

[4] 谈利芳,刘蓉,黄刚,等.基于遗传优化的多级 SVM 语音情感识别[J].电子测量技术,2017,40(10):122-126.

[5] 黄挺,王元庆,张自豪.基于 GS-SVM 的彩色图像分割算法[J].电子测量技术,2017,40(7):105-108,112.

[6] RASHEDI E, NEZAMABADI S, SARYAZDI S. GSA: A gravitational search algorithm [J]. Information Sciences, 2009, 179(13): 2232-2248.

[7] RASHEDI E, NEZAMABADI-POUR H, SARYAZDI S. Filter modeling using gravitational search algorithm [J]. Engineering Applications of Artificial Intelligence, 2011, 24(1):117-122.

[8] DUMAN S, GÜVENÇ U, SÖNMEZ Y, et al. Optimal power flow using gravitational search algorithm[J]. Energy Conversion & Management, 2012, 59(59):86-95.

[9] BHATTACHARYA A, ROY P K. Solution of multi-objective optimal power flow using gravitational search algorithm [J]. IET Generation Transmission & Distribution, 2012, 6(8): 751-763.

[10] SHUKLA A, SINGH S N. Multi-objective unit commitment with renewable energy using GSA algorithm[J]. Inae Letters, 2016, 1(1):21-27.

[11] RASHEDI E, NEZAMABADI-POUR H, SARYAZDI S. BGSA: binary gravitational search algorithm [J]. Natural Computing, 2010, 9(3): 727-745.

[12] LIU C, NIU P, LI G, et al. A hybrid heat rate forecasting model using optimized LSSVM based on improved GSA[J]. Neural Processing Letters, 2017, 45(1):299-318.

[13] DAS P K, BEHERA H S, PANIGRAHI B K. A hybridization of an improved particle swarm optimization and gravitational search algorithm for multi-robot path planning[J]. Swarm & Evolutionary Computation, 2016, 28:14-28.

[14] ELAZIM S M A, ALI E S. Optimal SSSC design for damping power systems oscillations via gravitational search algorithm[J]. International Journal of Electrical Power & Energy Systems, 2016, 82:161-168.

[15] MAHMOUDI S M, AGHAIE M, BAHONAR M, et al. A novel optimization method, gravitational search algorithm (GSA), for PWR core optimization [J]. Annals of Nuclear Energy, 2016, 95:23-34.

作者简介

赵东升(通信作者),博士、工程师,主要研究方向为模式识别、智能算法、半实物仿真技术、自动化测试技术等。

E-mail:mailzds@126.com

李艳军,硕士、工程师,主要研究方向为智能算法、自动化测试技术等。

E-mail:1569288823@qq.com