



基于群体智能优化的MKL-SVM算法及肺结节识别

李阳 常佳乐 王宇阳

MKL-SVM algorithm for pulmonary nodule recognition based on swarm intelligence optimization

LI Yang, CHANG Jia-yue, WANG Yu-yang

引用本文:

李阳, 常佳乐, 王宇阳. 基于群体智能优化的MKL-SVM算法及肺结节识别[J]. *工程科学学报*, 2021, 43(9): 1157–1165. doi: 10.13374/j.issn2095-9389.2021.01.14.004

LI Yang, CHANG Jia-yue, WANG Yu-yang. MKL-SVM algorithm for pulmonary nodule recognition based on swarm intelligence optimization[J]. *Chinese Journal of Engineering*, 2021, 43(9): 1157–1165. doi: 10.13374/j.issn2095-9389.2021.01.14.004

在线阅读 View online: <https://doi.org/10.13374/j.issn2095-9389.2021.01.14.004>

您可能感兴趣的其他文章

Articles you may be interested in

基于多目标支持向量机的ADHD分类

ADHD classification based on a multi-objective support vector machine

工程科学学报. 2020, 42(4): 441 <https://doi.org/10.13374/j.issn2095-9389.2019.09.12.007>

基于全局优化支持向量机的多类别高炉故障诊断

Multi-class fault diagnosis of BF based on global optimization LS-SVM

工程科学学报. 2017, 39(1): 39 <https://doi.org/10.13374/j.issn2095-9389.2017.01.005>

基于逐层演化的群体智能算法优化

Optimization for swarm intelligence based on layer-by-layer evolution

工程科学学报. 2017, 39(3): 462 <https://doi.org/10.13374/j.issn2095-9389.2017.03.020>

无数学模型的非线性约束单目标系统优化方法改进

Optimization method improvement for nonlinear constrained single objective system without mathematical models

工程科学学报. 2018, 40(11): 1402 <https://doi.org/10.13374/j.issn2095-9389.2018.11.014>

基于改进的支持向量回归机算法的磁记忆量化缺陷反演

Metal magnetic memory quantitative inversion of defects based on optimized support vector machine regression

工程科学学报. 2018, 40(9): 1123 <https://doi.org/10.13374/j.issn2095-9389.2018.09.014>

函数型数据分析与优化极限学习机结合的弹药传输机械臂参数辨识

Parameter identification of a shell transfer arm using FDA and optimized ELM

工程科学学报. 2017, 39(4): 611 <https://doi.org/10.13374/j.issn2095-9389.2017.04.017>

基于群体智能优化的 MKL-SVM 算法及肺结节识别

李 阳, 常佳乐, 王宇阳[✉]

长春工业大学计算机科学与工程学院, 长春 130012

✉通信作者, E-mail: liyanyaya1979@sina.com

摘 要 针对单核学习支持向量机无法兼顾学习能力与泛化能力以及多核函数参数寻优问题, 提出了一种基于群体智能优化的多核学习支持向量机算法。首先, 研究了五种单核函数对支持向量机分类性能的影响, 进一步提出具有全局性质的多项式核和局部性质的拉普拉斯核凸组合形式的多核学习支持向量机算法; 其次, 为增加粒子多样性及快速寻优, 将粒子群优化算法引入了遗传算法中的杂交操作, 并用此改进的群体智能优化算法对多核学习支持向量机进行参数寻优。最后, 分别采用深度特征与手工特征作为识别算法的输入, 研究表明采用深度特征优于手工特征。故本文采用深度特征作为多核学习支持向量机的输入, 以交叉遗传与粒子群混合智能优化算法作为其寻优方式。实验选取合作医院数据集对所提算法进行训练并初步测试, 进一步为了验证所提算法的泛化能力, 选取公开数据集 LUNA16 进行测试。实验结果表明, 本文算法易于跳出局部最优解, 提升了算法的学习能力与泛化能力, 具有较优的分类性能。

关键词 核函数; 支持向量机; 交叉遗传; 粒子群优化; 深度特征

分类号 TP391.4

MKL-SVM algorithm for pulmonary nodule recognition based on swarm intelligence optimization

LI Yang, CHANG Jia-yue, WANG Yu-yang[✉]

School of Computer Science and Engineering, Changchun University of Technology, Changchun 130012, China

✉ Corresponding author, E-mail: liyanyaya1979@sina.com

ABSTRACT To solve the problem that a single kernel learning support vector machine (SVM) cannot consider the learning and generalization abilities and parameter optimization of the multiple kernel function, a multiple kernel learning support vector machine (MKL-SVM) algorithm based on swarm intelligence optimization was proposed. First, the impact of five single kernel functions on the classification indexes of SVM was discussed. These kernel functions include two global kernel functions — the polynomial and sigmoid kernel functions — and three local kernel functions—the radial basis function, exponential kernel function, and Laplacian kernel function. Next, an MKL-SVM algorithm with a convex combination of a polynomial kernel having global properties and a Laplacian kernel having local properties was proposed. Then, to improve particle diversity to avoid falling into local optimal solutions during the iteration, and to reduce the model's training time, the crossover operation in the genetic algorithm was introduced into the particle swarm optimization (PSO) algorithm. This improved swarm intelligence optimization was used to optimize the parameters of the MKL-SVM. Finally, deep learning features based on the classical model VGG16 and handcrafted features according to doctors' suggestions were used as inputs for the recognition algorithm. In this algorithm, transfer learning was used to extract deep learning features and principal component analysis was used to reduce computational complexity through dimensionality reduction. The results show that using deep

收稿日期: 2021-01-14

基金项目: 国家自然科学基金资助项目 (61806024); 吉林省教育厅十三五科研规划项目 (JKH20181041KJ, JJKH20200680KJ); 吉林省科技发展计划项目 (20200401103GX)

learning features is better than handcrafted features. Therefore, this paper adopts the deep learning features as input for the MKL-SVM algorithm and the hybrid swarm intelligent optimization algorithm of crossover genetic and the PSO algorithm as the optimization method. To verify the generalization ability of the proposed algorithm, the public dataset LUNA16 was selected for testing. The experimental results show that the proposed algorithm is easy to jump out of the local optimal solution, improves the learning ability and generalization ability of the algorithm, and has a better classification performance.

KEY WORDS kernel function; support vector machine; crossover genetic; particle swarm optimization; deep learning features

肺癌被认为是一种侵入性疾病,对人类生命健康具有重大威胁,是癌症死亡的主要原因^[1-2]. 计算机断层扫描成像(Computed tomography, CT)技术是检测早期肺癌的重要手段. 在CT图像上,早期肺癌表现为直径小于30 mm的圆形或类圆形致密影,即肺结节. 因此,肺结节的早期识别可以有效地提高肺癌患者的生存率,避免错过最佳治疗机会^[3].

机器学习算法被广泛应用于医学图像处理中,主要分为传统的机器学习算法及深度学习算法. 传统机器学习算法中,包括支持向量机(Support vector machine, SVM)算法、随机森林(Random forest, RF)算法等,其中SVM应用最为广泛^[4-6]. 核函数是SVM的必要理论工具,能将原始数据映射到高维特征空间,实现非线性变换,构造适合给定问题的核函数可以提升分类器的性能. Shankar等^[7]为解决甲状腺疾病的健康诊断问题,采用线性核和高斯核组合形式的多核学习支持向量机(Multiple kernel learning support vector machine, MKL-SVM)算法对甲状腺数据进行分类,多核函数较单一使用线性核函数或高斯核函数的结果更优,准确度达到97.49%. Peng等^[8]提出一种线性核、多项式核和高斯核混合的MKL-SVM算法识别抑郁症,与单核SVM、朴素贝叶斯、决策树等方法相比错误率更低,可降低到16.54%. 核函数的改进在一定程度上可以提高SVM的分类性能,但SVM模型参数选择的影响也不容忽视. 群体智能优化思想因其效率高、适应性强等特点被广泛应用于经典机器学习算法模型的参数寻优. 常用的群体智能优化算法包括粒子群优化算法(Particle swarm optimization, PSO)、遗传算法(Genetic algorithm, GA)、人工鱼群算法(Artificial fish swarm algorithm, AFSA)等^[9-11].

近年来,深度学习被广泛应用于医学领域^[12]. 传统的机器学习方法有着深厚的理论基础,但对于大规模数据集仍存在挑战. 深度学习算法具有更高效的处理能力,但缺乏可解释性,且对于计算设备的要求较高. 因此,对传统机器学习算法的研究仍然存有意义,将深度学习算法与传统机器学习算法相结合也备受关注^[13-15].

本文分别采用根据医生建议设计的手工特征和深度网络提取的深度特征作为所提识别算法的输入,以确定最终的特征输入方式. 为了提高肺结节识别算法的准确度,首先研究了五种单核函数对SVM分类性能的影响,进一步为兼顾学习能力与泛化能力的提升且避免算法陷入局部最优的情况,提出了一种基于群体智能优化的MKL-SVM算法,多核函数加权了具有全局性质的多项式核和局部性质的拉普拉斯核,从而达到模型快速寻优、准确识别的目的.

1 MKL-SVM 算法

1.1 SVM 算法

SVM算法是一种以统计理论为基础的传统机器学习方法,可以最小化结构误差和最大化几何边缘,常用于分类任务和回归分析. SVM以结构风险最小化为准则,在经验风险最小化的同时,兼顾了期望风险最小化^[16]. 在二分类情况中,SVM模型的原问题可以通过拉格朗日乘子法转化为对偶问题:

$$\begin{aligned} \min_{\alpha} & \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{j=1}^l \alpha_j; \\ \text{s.t.} & \sum_{i=1}^l y_i \alpha_i = 0; \\ & 0 \leq \alpha_i \leq C, i = 1, 2, \dots, l. \end{aligned} \quad (1)$$

构造决策函数为

$$\begin{aligned} f(x) &= \text{sgn}(g(x)); \\ g(x) &= \sum_{i=1}^l \alpha_i y_i K(x_i, x_j) + b \end{aligned} \quad (2)$$

偏置 b 的求解如式(3)所示:

$$b = y_j - \sum_{i=1}^l y_i \alpha_i K(x_i, x_j) \quad (3)$$

式中, x_i 和 x_j 为第 i 和 j 个样本输入的特征变量, y_i 和 y_j 为第 i 和 j 个样本对应的标签, α_i 和 α_j 为拉格朗日乘子, $K(x_i, x_j)$ 为核函数, C 为正则化系数,描述了对错分样本的惩罚程度. 很多情况下,训练数据集是线性不可分的,通过核函数映射至高维特征空间,转换为高维空间中的线性问题,在这个

高维空间中寻找最优分类面。

1.2 核函数

核函数直接决定了特征空间的结构,核函数的选择是 SVM 的关键。核函数有诸多形式,但必须满足 Mercer 定理。核函数分为线性核和非线性核,线性核一般应用于线性可分的情况,其特征空间到输入空间的维度是一致的,如式(4)所示:

$$K(x, x') = (x, x') \quad (4)$$

实际问题大多是非线性的,故非线性核更为常用,包括多项式核函数、感知机核函数和径向基核函数等,本文讨论的几种单核函数的具体表示如下:

(1)多项式核 (Polynomial kernel).

$$K_{\text{poly}}(x, x') = ((x, x') + 1)^d \quad (5)$$

式中, d 为多项式阶数,取大于 1 的正整数,随着参数 d 的增大,泛化能力增强,但是当特征空间维数很高时, d 值会很大,使得计算量激增,甚至对某些情况无法得到正确结果。故实际应用中 d 一般取值 2 ~ 3。

(2)感知机核 (Sigmoid kernel).

$$K_{\text{sigmoid}}(x, x') = \tanh(a(x, x') + r) \quad (6)$$

式中, a 为斜率, r 为截距常数。Sigmoid 核来源于神经网络,常用作深度学习中的激活函数,使用 Sigmoid 核的 SVM 相当于一个两层的感知机网络。

(3)径向基核函数 (Radial basis function, RBF).

$$K_{\text{rbf}}(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2g^2}\right) \quad (7)$$

式中, g 为 RBF 核的宽度,决定了函数作用范围,

超过此范围,数据的作用就会减弱。RBF 核是一种经典的径向基核,也称为高斯核 (Gaussian kernel),取值仅依赖于特定点距离。

(4)指数核 (Exponential kernel).

$$K_{\text{exponent}}(x, x') = \exp\left(-\frac{\|x - x'\|}{2g^2}\right) \quad (8)$$

指数核也是一种径向基核,将向量之间的 L_2 距离调整为 L_1 距离,降低了对参数的依赖性,但是适用范围较小。

(5)拉普拉斯核 (Laplacian kernel).

$$K_{\text{Laplacian}}(x, x') = \exp\left(-\frac{\|x - x'\|}{g}\right) \quad (9)$$

拉普拉斯核也是一种径向基核,与指数核相类似。

图 1 给出不同核函数全局性与局部性分析的仿真曲线。局部性核函数仅对样本中心点附近的点有较大影响,影响会随距离的增大而减弱,具有较好的拟合效果与较强的学习能力;全局性核函数对与样本中心点距离较远的点也产生影响,有较好的泛化能力。如图 1 所示,图中横轴 X 表示 $x - x'$,纵轴 Y 表示 $K(x, x')$,当 $x \rightarrow x'$ 时, $K(x, x') \rightarrow 1$,即 x 与 x' 越接近,核函数的作用越明显。通过仿真结果分析,多项式核是一种经典的全局性核函数,允许对多项式级的特征连接进行建模。Sigmoid 核也是一种全局性核函数。RBF 核、指数核与拉普拉斯核均为径向基核函数,是典型的局部性核函数。与全局性核函数相比,局部性核函数学习能力较强,而泛化能力较弱。

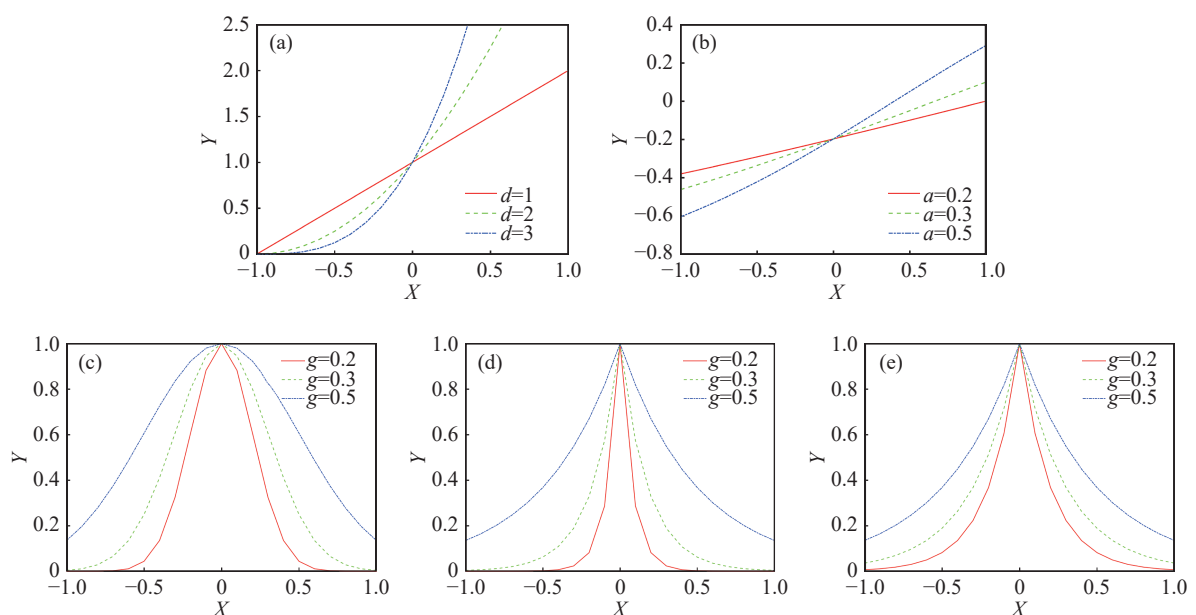


图 1 不同核函数的全局性与局部性分析。(a)多项式核;(b)感知机核;(c)高斯核;(d)指数核;(e)拉普拉斯核

Fig.1 Global and local analyses of various kernel functions: (a) polynomial kernel; (b) sigmoid kernel; (c) Gaussian kernel; (d) exponential kernel; (e) Laplacian kernel

1.3 MKL-SVM 算法

MKL-SVM 通过选择合适的核函数或不同单核函数的加权组合形式, 能够充分利用单个核函数的特性. 经证明核函数的加权凸组合形式满足 Mercer 定理, 仍为核函数, 可用于 SVM 模型^[4], 如式(10)所示:

$$K_{\text{multiple}}(x, x') = \sum_{\alpha=1}^N m_{\alpha} K_{\alpha}(x, x'); \quad (10)$$

$$\sum_{\alpha=1}^N m_{\alpha} = 1, 0 < m_{\alpha} < 1, \alpha = 1, \dots, N$$

式中, m_{α} 是每种基本内核在多核函数中所占的权重, 多核函数共采用 N 种基本内核函数, 其权重之和取为 1.

2 基于群体智能优化的 MKL-SVM 识别算法

2.1 群体智能优化算法

群体智能优化是一种启发式算法, 主要模拟自然界中各种生物的生活行为, 如昆虫、鱼群、鸟群等, 它们通过一种合作的方式觅食, 群体中的各个生物通过累积经验不断更新搜索方向^[17]. 群体智能优化算法灵活、高效、适应性强且易于实现, 广泛应用于各种问题.

GA 算法由 Holland 于 1975 年提出, 是一种借鉴生物界自然选择与进化机制而发展的全局优化算法. GA 算法随机产生初始解, 通过一定的选择、交叉、变异操作, 逐步迭代产生解. GA 算法的适应性广泛, 不依赖于问题的具体领域, 隐含并行搜索特性, 减少了陷入局部最优解的情况. PSO 算法最早由 Kennedy 和 Eberhart 于 1995 年提出, 从鸟群觅食行为特征中得到启发, 并应用于求解优化问题中. 算法中每个粒子代表问题的一个潜在解, 且对应于一个适应度值, 粒子的速度决定其移动方向和距离, 速度随粒子的变换经验进行调整, 从而达到寻优的目的. 与网格搜索算法相比, PSO 算法无需遍历所有组合参数, 能节约模型的训练时间; PSO 算法与 GA 算法相比, 无需进行选择、交叉、变异的操作, 通过粒子在解空间追踪最优的粒子进行搜索. 虽然 PSO 算法能够更快地得到最优解, 但其缺点是随着迭代次数的增加, 种群多样性减少, 易引起粒子早熟现象, 从而易陷入局部最优^[18]. 为提高 PSO 算法的整体性能, 可以从参数设置、收敛性及与其他算法结合等方面进行改进^[19]. Choudhary 等^[20] 为声表面波工艺参数优化研

究了 PSO 和 GA 的混合算法, 利用 GA 算法中的遗传变异算子来避免粒子出现早熟收敛现象, 实验结果表明混合算法在最优解、均值、计算时间等指标上均表现优异. Koessler 等^[21] 为提高算法的鲁棒性, 提出了模式搜索和 PSO 混合的优化算法, 先运行 PSO 寻求最佳位置后, 再利用模式搜索方法以进一步最小化适应度函数, 实验结果表明混合算法在提高精度和鲁棒性方面取得了成功.

本文借鉴 GA 算法中的杂交操作, 以增加 PSO 算法中粒子的多样性, 即在 PSO 算法每次迭代过程中根据杂交概率选取一定数量的粒子放入杂交池, 用两两杂交产生相同数目的子代粒子替换亲代粒子. 子代粒子的位置和速度计算公式如式(11)所示:

$$C_x = \beta \cdot P_1(x) + (1 - \beta) \cdot P_2(x)$$

$$C_v = \frac{P_1(v) + P_2(v)}{|P_1(v) + P_2(v)|} \cdot |P_1(v)| \quad (11)$$

式中, C_x 代表子代粒子的位置, C_v 代表子代粒子的速度, P_1 和 P_2 代表亲代粒子, x 代表粒子位置, v 代表粒子速度, β 为 0 到 1 之间的随机数, 代表杂交概率. 提出的基于交叉遗传与粒子群混合的群体智能优化 (GAPSO) 算法的过程如图 2 所示.

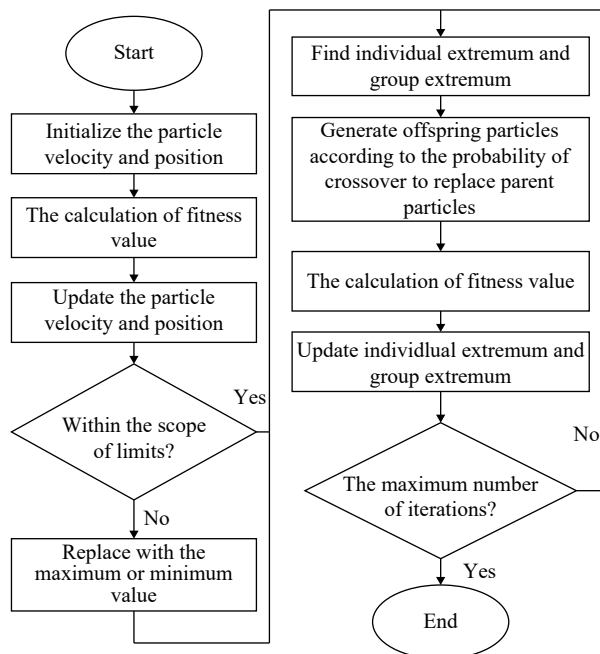


图 2 GAPSO 的算法流程图

Fig.2 Flowchart of the GAPSO algorithm

2.2 群体智能优化的 MKL-SVM 算法

本文进一步提出了一种改进的 MKL-SVM 算法, 基本内核采用具有较强泛化能力的多项式核函数和具有较强学习能力的拉普拉斯核函数进行

线性凸组合,可以更好地兼顾样本的全局特征与局部特征,提高学习能力与泛化能力. 多核函数的具体构成如式(12)所示:

$$K(x, x') = \lambda K_{\text{poly}}(x, x') + (1 - \lambda) K_{\text{Laplacian}}(x, x') \quad (12)$$

式中, λ 能够自由调节不同的内核在多核函数中的权重,范围取(0,1).

基于上述群体智能优化思想,本文最终使用提出的 GAPS0算法作为 MKL-SVM 的惩罚参数 C 、权重 λ 、核函数参数 d 和 g 的寻优方法. 将 5 折交叉验证下的 ACC 作为 GAPS0 优化算法的适应度函数值,最终达到准确分类的目的,ACC 定义如式(13)所示:

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (13)$$

式中, TP 为识别出的真阳性结节, TN 为识别出的真阴性,即非结节, FP 为假阳性结节, FN 为假阴性结节.

3 实验数据及结果分析

实验数据采用吉林省某大型三甲专科医院的 20 组病例共约 700 幅图像,一部分用于模型训练,一部分用于模型测试. 识别的前期准备,包括图像预处理、肺结节感兴趣区域 (Region of interests, ROI) 获取和 ROI 特征提取等,具体步骤与文献 [4] 相同,共提取出 270 个肺结节候选 ROI(80 个结节, 190 个非结节),每个 ROI 分别提取 13 维手工特征和降至 49 维的深度特征作为所提识别算法的输入数据. 这些数据经过随机打散后分为 170 个训练样本与 100 个测试样本,并进行归一化处理. 进一步,为了验证模型的泛化能力,采用公开数据集 LUNA16 作为测试集 [22]. 实验平台采用 MATLAB, 利用 libsvm 工具箱进行仿真实验.

在实验参数选取中,设置种群粒子数为 20,最大迭代次数为 200,粒子维度为 3,分别代表权值 λ 、正则化系数 C 和多核函数中拉普拉斯核的参数 g ,其中 λ 取值范围为 [0, 1],速度为 [-0.6, 0.6];参数 C 取值范围为 [2^{-9} , 2^9],速度为 [$-2^9 \times 0.6$, $2^9 \times 0.6$];参数 g 取值范围为 [2^{-7} , 2^7],速度为 [$-2^7 \times 0.6$, $2^7 \times 0.6$];杂交概率设置为 0.6,多项式核函数的阶数选取 2 及 3. 实验采用 ACC、SEN、SPE 三个评价指标,分别代表总体识别正确率、结节的检出率和非结节识别正确率,进一步采用 PRE、 F_1 -score、MCC 综合指标,表达式分别如下所示:

$$\text{SEN(Recall)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (14)$$

$$\text{SPE} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (15)$$

$$\text{PRE} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (16)$$

$$F_1\text{-score} = \frac{2 \cdot \text{SEN} \cdot \text{PRE}}{\text{SEN} + \text{PRE}} \quad (17)$$

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \quad (18)$$

式中, PRE 表示模型对负样本的区分能力, F_1 -score 表示 PRE 和 SEN 的调和平均评估指标, MCC 表示预测分类与实际分类结果之间的相关系数.

3.1 手工特征实验

手工特征为根据医生建议设计的特征,与文献 [4] 相同,将 7 个形态学特征、2 个灰度特征以及 4 个纹理特征共计 13 维手工特征作为提出识别算法的输入. 其中选取的 7 个形态学特征为:面积、直径、周长、矩形度、扁度、圆形度和细长度. 灰度特征采用灰度均值和灰度方差. 纹理特征采用灰度共生矩阵的能量、对比度、熵和逆差矩. 为保证实验的有效性,每组实验重复进行 10 次,实验结果取统计均值. 测试阶段的受试者操作特征曲线 (Receiver operating curve, ROC) 与查准率-查全率曲线 (Precision-recall curve, PR) 如图 3 所示. 表 1 列出了在合作医院测试数据集上,前述五种单核函数 SVM 算法在所提 GAPS0 寻优方式下的结果和本文 MKL-SVM 算法不同寻优方式下的识别结果. 其中, ACC_mean 为平均准确度, ACC_max 代表最大 ACC, ACC_max 相应的 SEN 和 SPE 分别用 MASEN 和 MASPE 表示, AUC 代表 ROC 曲线下的面积, AP 代表 PR 曲线下的面积. ROC 曲线代表真阳率与假阳率的比例,左上顶点越接近于 (0,1) 点,代表分类器性能越佳. PR 曲线代表 PRE 与 SEN 的关系,更易受到样本分布的影响,右上顶点越接近于 (1,1) 点,针对不均衡样本的分类器性能越佳. 由表 1 可知,仅采用 GAPS0 算法对单核 SVM 寻优时,全局性质的多项式核函数的 ACC 和 SEN 均高于 Sigmoid 核,性能较优. 局部性质的拉普拉斯核函数的 ACC_mean 为 90.60%,高于其他四种核函数,其 ACC_max 为 91.00%, MASEN 为 92.59%, MASPE 为 90.41%,性能均较优. 故本文算法采用多项式核函数和拉普拉斯核函数凸组合的多核函数. 如图 3 所示,本文算法的 ROC 曲线左上顶点更趋近于 (0,1) 点, PR 曲线右上顶点更趋近于 (1,1) 点, AUC 与 AP 值均更接近于 1,具

有较优分类性能. 综合比较, 本文算法的 ACC_mean、F₁-score、MCC 及 AP 均为其中最大值, 比单

独使用多项式核函数和拉普拉斯核函数的实验结果更优.

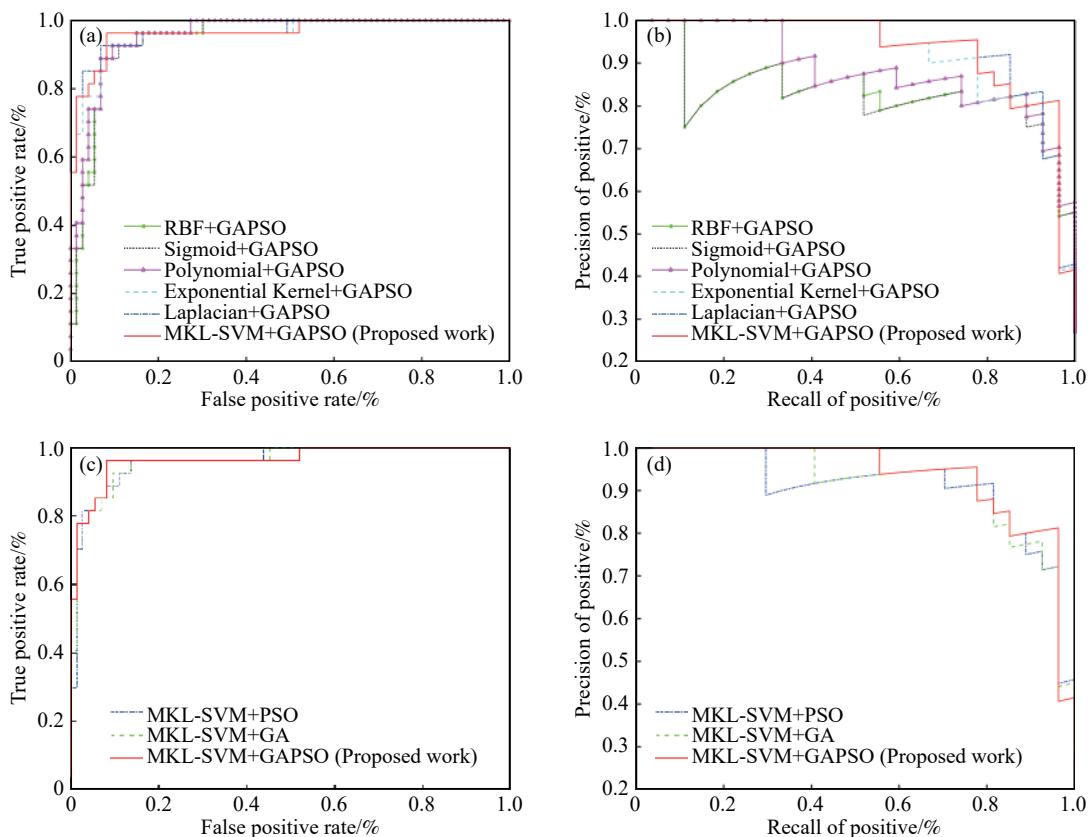


图 3 不同算法的 ROC 曲线图及 PR 曲线图. (a)不同核函数 SVM 算法的 ROC 曲线; (b)不同核函数 SVM 算法的 PR 曲线; (c)不同寻优方式 MKL-SVM 算法的 ROC 曲线; (d)不同寻优方式 MKL-SVM 算法的 PR 曲线

Fig.3 ROC and PR curves of various algorithms: (a) ROC curves of SVM algorithms with various kernel functions; (b) PR curves of SVM algorithms with various kernel functions; (c) ROC curves of the MKL-SVM algorithm with various optimization algorithms; (d) PR curves of the MKL-SVM algorithm with various optimization algorithms

表 1 不同核函数的实验结果

Table 1 Experimental results of various kernel functions

Algorithm	ACC_mean/%	ACC_max/%	MASEN/%	MASPE/%	F ₁ _score/%	MCC/%	AUC	AP
Polynomial kernel + GAPSO	90.00	90.00	85.19	91.78	82.14	75.30	0.9584	0.8506
Sigmoid kernel + GAPSO	89.00	89.00	77.78	93.15	81.26	74.35	0.9482	0.7990
RBF kernel + GAPSO	90.50	91.00	88.89	91.78	82.85	76.39	0.9498	0.8022
Exponential kernel + GAPSO	90.40	91.00	92.59	90.41	83.71	77.60	0.9604	0.8470
Laplacian kernel + GAPSO	90.60	91.00	92.59	90.41	84.18	78.23	0.9655	0.8464
MKL-SVM + PSO	90.80	92.00	88.89	93.15	83.68	77.44	0.9609	0.8726
MKL-SVM + GA	89.50	90.00	92.59	89.04	82.50	75.93	0.9619	0.8830
MKL-SVM + GAPSO	91.10	92.00	88.89	93.15	84.30	78.29	0.9650	0.8984

采用 GA、PSO 和 GAPSO 对 MKL-SVM 算法进行参数寻优, 采用所提 GAPSO 算法的 ACC_mean 为 91.10%, 高于 GA 算法和 PSO 算法; GAPSO 算法和 PSO 算法的 ACC_max 均为 92.00%, 高于 GA 算法 2%, MASEN 低于 GA 算法 3.7%, MASPE 高于 GA 算法

4.11%. 综合比较, GAPSO 算法的总体识别准确度更高, 但敏感度略低, 这可能由于适应度函数选为 ACC 导致的. 图 4 为训练阶段本文算法的适应度曲线.

由图 4 可见, 本文算法学习能力较优, 能够使粒子跳出局部最优, 更易寻求全局最优解, 10 次

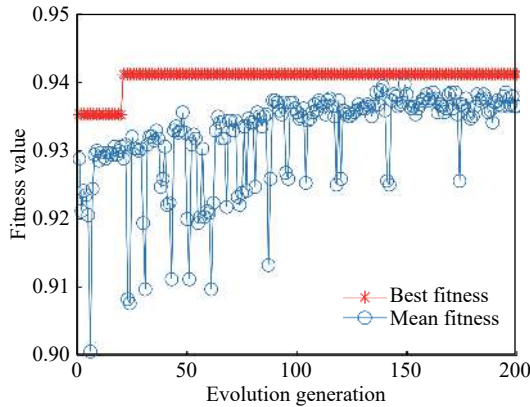


图 4 本文算法的适应度曲线

Fig.4 Fitness curve of the proposed algorithm

训练结果的均值为 94.06%。

3.2 深度特征实验

深度结构模型可以获取更丰富的特征信息, 故本文进一步采用深度网络模型 VGG16^[23] 对合

作医院数据集提取深度特征. VGG16 模型的低层是由卷积层和最大池化层交替组成, 高层是全连接层, 利用 CNN 特征提取的核心是卷积层和池化层, 将提取得到的特征图作为全连接层的输入, 包含肺结节最丰富的语义信息, 可以更全面地描述特征. 因此, 本文采用原始的 VGG16 模型, 模型共分为 6 部分, 前 5 部分为卷积网络, 最后为全连接网络, 具体结构见文献 [23]. 本文将预训练的 VGG16 第一个完全连接层前的权重, 迁移到目标网络中进行特征提取, 从而得到 $7 \times 7 \times 512 = 25088$ 维深度特征. 为避免产生维数灾难并且降低计算量, 采用主成分分析 (Principal component analysis, PCA) 方法将深度特征进行降维. 根据累计方差贡献率及 VGG16 模型的结构, 同时降低计算复杂度, 最终选取了 $7 \times 7 = 49$ 维的深度特征与本文提出的算法结合, 实验结果如表 2 所示.

表 2 深度特征结合本文算法的实验结果

Table 2 Results of the proposed algorithm combined with deep learning features

Algorithm	ACC_mean/%	ACC_max/%	MASEN/%	MASPE/%	F ₁ _score/%	MCC/%	AUC	AP
Handcrafted features + MKL-SVM + GAPSO	91.10	92.00	88.89	93.15	84.30	78.29	0.9650	0.8984
Deep learning features + MKL-SVM + GA	88.00	88.00	81.82	91.04	81.82	72.86	0.9038	0.8755
Deep learning features + MKL-SVM + PSO	89.80	91.00	75.76	97.01	82.57	76.81	0.9484	0.9038
Deep learning features + MKL-SVM + GAPSO (Proposed work)	91.50	94.00	81.82	100	85.81	80.69	0.9588	0.9043

由表 2 可知, 采用深度特征作为本文算法的输入向量, ACC_mean 为 91.50%, ACC_max 为 94.00%, 均高于 GA 及 PSO 寻优方式. 与手工特征相比, ACC_max 比手工特征提高 2%, MASPE 可达 100%, 但 MASEN 仅为 81.82%. 深度特征与本文算法结合的 ACC_mean、F₁-score、MCC 及 AP 均为其中最大值, 在 ACC 和 SPE 有明显提升, 但 SEN 检出率较低. 图 5 为本文算法采用深度特征在训练阶段的适应度曲线, 10 次适应度函数的均值为 86.35%, 但测试阶段的 10 次平均准确度为 91.50%, 泛化能力有所提升.

综上所述, 本文提出基于 GAPSO 的 MKL-SVM 算法进行参数寻优, 易于跳出局部最优解, 提升了学习能力与泛化能力, 达到模型快速寻优、准确识别的目的. 进一步, 本文将深度特征作为所提识别算法的输入, 能够提升识别准确度, 作为最终的肺结节识别模式.

3.3 公开数据集验证模型泛化能力实验

为了评估模型的有效性和泛化性, 进一步采用了 LUNA16 数据集^[22] 进行验证实验. 该数据集是公开的肺图像数据库联盟与图像数据库资源

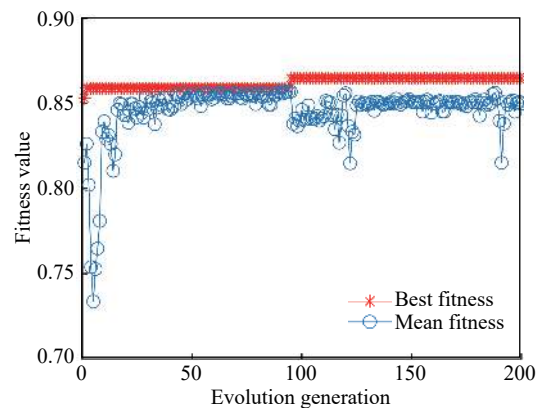


图 5 深度特征结合本文算法的适应度曲线

Fig.5 Fitness curve combining deep learning features with the proposed algorithm

计划 (Lung image database consortium and image database resource initiative, LIDC-IDRI) 的子集, 包括 888 组低剂量肺部 CT 影像, 其中至少 3 名放射科专家标注的总共 1186 个结节. 经本文的图像预处理方法后, 共选取 1140 个肺结节 ROI (650 个结节, 490 个非结节), 并用 VGG16 模型提取 49 维深度特征作为提出算法的输入, 以验证模型推广能

力. 表 3 列出了本文算法与近年来当前主流算法在 LIDC-IDRI 和 LUNA16 的实验结果. 当前, 深度学习方式备受关注, Zhao 等^[24]利用迁移学习方式对 ResNet 等模型进行微调, 以达到肺结节分类的目的. Masood 等^[25]提出基于多维区域的全卷积网

络用于肺结节检测, 实验结果表明提出模型的灵敏度与分类精度均较高. Mastouri 等^[14]将双线性深度学习网络与 SVM 算法相结合以实现肺结节分类, 结果表明将深度网络与传统机器学习方式相结合能够提升分类准确度.

表 3 所提算法与当前主流算法的性能比较

Table 3 Performance comparison of the proposed algorithm with current state-of-the-art methods

References	Year	Datasets	Methods	ACC/%	SEN/%	SPE/%	AUC
Zhao <i>et al.</i> [24]	2019	LIDC-IDRI (743 images)	Transfer learning CNNs	85.00	94.00	—	0.94
Masood <i>et al.</i> [25]	2020	LIDC-IDRI (892 images)	Enhanced multidimensional region-based fully CNN	97.91	98.1	93.2	0.9813
Mastouri <i>et al.</i> [14]	2020	LUNA16 (3186 images)	Bilinear CNN + SVM	91.99	91.85	92.27	0.959
Proposed work	2021	LUNA16 (1140 images)	Deep learning features+ Improved MKL-SVM	95.29	94.85	95.89	0.9803

由表 3 可知, 所提算法的 AUC 能够达到 0.9803, 进一步验证了本文算法的泛化能力. 与文献 [24] 和文献 [14] 相比, 本文算法的评价指标均较高, ACC 与 SEN 有所提升, 具有较佳的分类性能. 本文算法较文献 [25] 性能略低, 但计算复杂度小于文献 [25] 中所提方法, 能够节约时间成本, 后续可以通过增加标准数据库扩充训练样本, 以达到提升分类性能的目的.

4 结论

针对单一核函数无法兼顾学习能力和泛化能力以及模型参数寻优的问题, 提出了基于群体智能优化的 MKL-SVM 识别算法. 分别采用手工特征与深度特征作为输入, 验证提出算法的有效性, 实现了肺结节准确识别的目的. 具体创新如下:

(1) 采用多项式核与拉普拉斯核的加权组合形式作为 MKL-SVM 的核函数, 提出的算法兼顾了学习能力与泛化能力.

(2) 为解决 PSO 算法迭代后期易缺乏粒子多样性的问题, 提出 GAPS0 混合智能算法进行参数寻优, 更易寻求全局最优解.

(3) 采用深度特征与提出算法相结合, 能够提高模型的分类性能.

在未来的研究中, 针对敏感度略低的问题, 将进一步探讨多评价指标兼顾, 以保证识别准确度和敏感度的进一步提升, 避免结节漏检. 同时改进深度卷积神经网络进行特征提取并融合手工特征, 以保留完整的特征信息, 提升识别指标. 此外, 通过增加标准数据库扩充训练样本, 可以达到提升分类性能的目的.

参 考 文 献

- [1] Shen W, Zhou M, Yang F, et al. Multi-crop Convolutional Neural Networks for lung nodule malignancy suspiciousness classification. *Pattern Recognit*, 2017, 61: 663
- [2] Ferlay J, Colombet M, Soerjomataram I, et al. Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018. *Eur J Cancer*, 2018, 103: 356
- [3] Siegel R L, Miller K D, Jemal A. Cancer statistics, 2018. *CA: A Cancer J Clin*, 2018, 68(1): 7
- [4] Li Y, Zhu Z, Hou A, et al. Pulmonary nodule recognition based on multiple kernel learning support vector machine-PSO. *Comput Math Methods Med*, 2018, 2018: 1461470
- [5] Renita D B, Christopher C S. Novel real time content based medical image retrieval scheme with GWO-SVM. *Multimed Tools Appl*, 2020, 79(23-24): 17227
- [6] Jia D Y, Li Z Y, Zhang C W. Detection of cervical cancer cells based on strong feature CNN-SVM network. *Neurocomputing*, 2020, 411: 112
- [7] Shankar K, Lakshmanaprabu S K, Gupta D, et al. Optimal feature-based multi-kernel SVM approach for thyroid disease classification. *J Supercomput*, 2020, 76(2): 1128
- [8] Peng Z C, Hu Q H, Dang J W. Multi-kernel SVM based depression recognition using social media data. *Int J Mach Learn Cybern*, 2019, 10(1): 43
- [9] Valdez F. A review of optimization swarm intelligence-inspired algorithms with type-2 fuzzy logic parameter adaptation. *Soft Comput*, 2020, 24(1): 215
- [10] Zhou T, Lu H L, Hu F Y, et al. A model of high-dimensional feature reduction based on variable precision rough set and genetic algorithm in medical image. *Math Probl Eng*, 2020, 2020: 1
- [11] Gao Y K, Xie L B, Zhang Z D, et al. Twin support vector machine based on improved artificial fish swarm algorithm with application to flame recognition. *Appl Intell*, 2020, 50(8): 2312
- [12] Litjens G, Kooi T, Bejnordi B E, et al. A survey on deep learning in medical image analysis. *Med Image Anal*, 2017, 42: 60

- [13] Zhang B H, Qi S L, Monkam P, et al. Ensemble learners of multiple deep CNNs for pulmonary nodules classification using CT images. *IEEE Access*, 2019, 7: 110358
- [14] Mastouri R, Khlifa N, Neji H, et al. A bilinear convolutional neural network for lung nodules classification on CT images. *Int J Comput Assist Radiol Surg*, 2021, 16(1): 91
- [15] Özyurt F, Sert E, Avci E, et al. Brain tumor detection based on Convolutional Neural Network with neutrosophic expert maximum fuzzy sure entropy. *Measurement*, 2019, 147: 106830
- [16] Hao P Y. Dual possibilistic regression analysis using support vector networks. *Fuzzy Sets Syst*, 2020, 387: 1
- [17] Yan B L, Zhao Z, Zhou Y C, et al. A particle swarm optimization algorithm with random learning mechanism and Levy flight for optimization of atomic clusters. *Comput Phys Commun*, 2017, 219: 79
- [18] Wu Z, Zhang S, Wang T. A cooperative particle swarm optimization with constriction factor based on simulated annealing. *Computing*, 2018, 100(8): 861
- [19] Freitas D, Lopes L G, Morgado-Dias F. Particle swarm optimisation: A historical review up to the current developments. *Entropy*, 2020, 22(3): 362
- [20] Choudhary A, Kumar M, Gupta M K, et al. Mathematical modeling and intelligent optimization of submerged arc welding process parameters using hybrid PSO-GA evolutionary algorithms. *Neural Comput Appl*, 2020, 32(10): 5761
- [21] Koessler E, Almomani A. Hybrid particle swarm optimization and pattern search algorithm. *Optim Eng*, 2020: 1
- [22] Setio A A A, Traverso A, Bel T, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. *Med Image Anal*, 2017, 42: 1
- [23] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition//*Proceedings of 2015 International Conference on Learning Representations*. California, 2015: 1
- [24] Zhao X, Qi S, Zhang B, et al. Deep CNN models for pulmonary nodule classification: Model modification, model integration, and transfer learning. *J Xray Sci Technol*, 2019, 27(4): 615
- [25] Masood A, Sheng B, Yang P, et al. Automated decision support system for lung cancer detection and classification via enhanced RFCN with multilayer fusion RPN. *IEEE Trans Ind Inform*, 2020, 16(12): 7791