文章编号 1004-924X(2017)增-0215-06

改进的支持向量机算法在肺结节识别中的应用

李 阳[1,2],赵庆东[1],田 颖[3*]

(1. 长春工业大学 计算机科学与工程学院,吉林 长春 130012;
2. 东北师范大学 数学与统计学院,吉林 长春 130024;
3. 长春市妇产医院,吉林 长春 130042)

摘要:为了解决在肺结节识别过程中数据集正负样本分布不均衡以及参数寻优时间过长的问题,提出了一种 PSO-CSVM 算法。首先从肺 CT 中提取肺结节 ROI 图像,然后对其提取 13 维特征,最后利用提出的基于 PSO 的代价敏感型 SVM 算法进行识别。测试结果显示识别准确率达到 91.11%,敏感度达到 85.71%,特异度达到 93.55%,参数寻优时间为 54.37 s。将提出的算法与遗传寻优算法及网格寻优搜索算法相比较来验证算法的有效性,实验结果表明,PSO-CSVM 算法运行时间较短,准确度和敏感度最优,而且具有运行时间短,识别准确率和检出率高的特点,能够满足医学影像学对肺结节的识别要求。

关 键 词:肺结节识别;代价敏感型支持向量机;粒子群优化算法;RBF 核

中图分类号:TP391.4 **文献标识码**:A **doi**:10.3788/OPE.20172513.0215

Application of improved support vector machine in identification of pulmonary nodule

LI Yang[1,2], ZHAO Qing-dong[1], TIAN Ying[3*]

(1. *College of Computer Science and Engineering, Changchun University of Technology, Changchun* 130012, *China*;
2. *College of Mathematics and Statistics, Northeast Normal University, Changchun* 130024, *China*;
3. *Changchun Obstetrics-Gynecology Hospital, Changchun* 130042, *China*)
* *Corresponding author, E-mail: tianying*781104

Abstract: In order to solve the problems of the unbalanced distribution of positive and negative samplesin data set in pulmonary nodule identification and overtime parameter optimization, a PSO-CSVM algorithm was proposed. ROI image of pulmonary nodule was extracted from the lung CT and then 13-dimensional characteristics was extracted from it. Finally, the proposed PSO-based cost-sensitive type SVM algorithm was used for identification. In the testing the accuracy rate of the identification reached 91.11% and the sensitivity reached 85.71%, specificity reached 93.55%, and the time of parameter optimization was 54.37 s. In order to further verify the effectiveness of the algorithm, the

收稿日期:2017-06-01;**修订日期**:2017-06-22.

基金项目:吉林省科技厅资助项目(No. 20160418080FG);吉林省教育厅资助项目(No. JJKH20170575KJ;No. 2014142)

proposed algorithm was compared with the genetic optimizing algorithm and grid optimizing searching algorithm. The experimental result shows that the run-time of PSO-CSVM is shorter and the accuracy and sensitivity is optimal. It features short run-time, high accuracy rate of identification and detection rate and can meet the requirements of medical imaging for the identification of pulmonary nodule.

Key words: pulmonary nodule recognition; cost-sensitive support vector machine; particle swarm optimization; RBF kernel

1 引　言

肺CAD系统通常由图像预处理、肺实质分割、肺结节的感兴趣区域(Region of Interesting, ROI)分割、ROI特征提取以及肺结节识别等部分组成,其中肺结节识别是系统中的重要组成部分。文献[1]提出了一种3D模型,采用三维分割技术提取肺结节,在识别阶段采用神经网络方法来区分结节和非结节以提高识别能力;Shen等[2]利用多组卷积神经网络对LIDC-IDRI数据集的1 010个疑似恶性结节样本进行识别,该方法与传统方法的不同之处在于它不依赖于肺结节的精确分割,以直接构建原始结节的模型作为输入。首先用卷积神经网络方法进行特征提取,然后利用机器学习方法来对疑似结节进行识别,其准确率达到87.14%,而且该方法能够表征结节的边缘及直径信息;Froz等[3]采用"人工爬虫"模型,即人工生命算法对LIDC-IDRI数据集的1 402个样本提取方向纹理特征,最后利用支持向量机对结节识别,其敏感度达到91.86%,特异度达到94.78%。为了解决在肺结节有效识别中的多个异质特征子集,高维无关特征以及数据集正负样本分布不平衡等问题,文献[4]提出了一种多核框架进行特征选择,用于从特征子集进行异构特征融合及选择。实验结果表明,所提方法在几何平均值(G-mean)和受试者工作特征曲线(Receiver Operating Characteristic, ROC)下面积(AUC)指标优于其他方法;文献[5]提出了一种将稀疏张量表示方法用于CT图像的预处理阶段,来达到去除肺部CT图像中的噪声,增强了图像的有用信息,且获得更高精度图像的目的;文献[6]利用改进的遗传算法从所提取的22个特征中,选择了7个特征作为最优特征子集,并且用最佳特征子集来训练支持向量机分类器模型,通过减少假阳并保留真实结节来提高分类器的性能,其特异度达到了95.5%;Li等[7]提出一种混合核支持向量机算法对肺结节进行识别,并利用网格搜索算法对参数进行寻优,敏感度和准确度分别为92.59%和92%,但网格搜索法计算量过于庞大,运算时间过长。

鉴于上述分析,针对结节识别过程中数据集正负样本分布不均衡导致分类平面向数据较少一类的样本侧倾斜,以及运算时间较长的问题,提出PSO-CSVM算法,即在训练阶段拟采用粒子群算法(Particle Swarm Optimization, PSO)对代价敏感型支持向量机(Cost-sensitive Support Vector Machine, CSVM)识别器进行参数寻优,然后将寻得的最优参数组代入测试集进行测试,最终实现肺结节的识别。

2 PSO-CSVM算法

2.1 PSO算法

PSO算法是由Kennedy和Eberhart在1995年提出的一种群体智能的优化算法。在算法中,每个优化问题的解类似于搜索空间中的一个粒子,每个粒子对应一个适应度函数值。每个粒子通过跟踪个体极值和群体极值来更新自己的位置。PSO算法首先随机生成一个粒子种群,然后通过粒子在解空间中追随当前最优的粒子进行迭代搜索,直到达到要求,停止搜索。在每次迭代过程中,粒子更新速度和位置的公式为[8]:

$$\mathbf{V}_{\mathrm{id}}^{k+1} = \omega\mathbf{V}_{\mathrm{id}}^{k} + c_1 r_1(\mathbf{P}_{\mathrm{id}}^{k} - \mathbf{X}_{\mathrm{id}}^{k}) + c_2 r_2(\mathbf{P}_{\mathrm{gd}}^{k} - \mathbf{X}_{\mathrm{id}}^{k}), \tag{1}$$

$$\mathbf{X}_{\mathrm{id}}^{k+1} = \mathbf{X}_{\mathrm{id}}^{k} + \mathbf{V}_{\mathrm{id}}^{k+1}. \tag{2}$$

其中:$\mathbf{V}_i=[V_{i1}, V_{i2}, \cdots, V_{iD}]^{\mathrm{T}}$ 表示第 i 个粒子的速度,$\mathbf{X}_i=[x_{i1}, x_{i2}, \cdots, x_{iD}]^{\mathrm{T}}$ 表示第 i 个粒子的位置,$\mathbf{P}_i=[P_{i1}, P_{i2}, \cdots, P_{iD}]^{\mathrm{T}}$ 为其个体极值,$\mathbf{P}_g=[P_{g1}, P_{g2}, \cdots, P_{gD}]^{\mathrm{T}}$ 为种群的全局极值,$d=1, 2, \cdots, D$, $i=1, 2, \cdots, n$, D为搜索空间维数,ω 为惯性权重,体现粒子继承先前的速度能力,惯性权重较大有利于全局搜索,惯性权值较小利于局部搜

索。惯性权重的经验取值为 $\omega_{start}=0.9$, $\omega_{end}=0.4$;k 为当前迭代次数;c_1 和 c_2 为非负的常数,称为加速度因子,两者中当 c_1 较大时,会导致粒子在搜索范围内很难收敛;当 c_2 较大时,会导致粒子过早结束搜索。因此 c_1 和 c_2 的选择对于寻优过程至关重要,在 PSO 算法中 c_1 和 c_2 的取值常取 $c_1=c_2=2$ 左右。r_1 和 r_2 是分布于[0,1]之间的随机数。

2.2　CSVM 算法

支持向量机利用最优化方法解决机器学习的问题,寻求模型的复杂性和学习能力之间的最佳折衷,识别未知样本[9]。传统 SVM 中正类与负类采用相同的惩罚参数 C,C 值决定了最大类间间隔与最小训练错误间的折中程度,当数据集严重不均衡时,如正类的数目远小于负类的数目,正类与负类采用相同的惩罚参数,将使得正类的误差之和小于负类的误差之和,相当于加大了负类的惩罚力度,导致分割平面向正类的一侧移动。为此,可以对正类与负类分别引入不同的惩罚系数 C_+ 和 C_-,从而可以灵活调整假阳性与假阴性的错误分类代价,这种算法被称为代价敏感型支持向量机(CSVM)[10]。设训练样本 $T=\{(x_i,y_i)\}(i=1,2,\cdots,l)$,$x_i$ 为 SVM 的输入特征,且 $x_i\in R^N$,$y_i\in\{-1,+1\}$为类别标签,$i=1,2,\cdots,l$,l 为训练样本个数。$y_i=1$ 时对应于结节情况,而 $y_i=-1$ 对应于非结节情况。基于二分类目标函数 CSVM 实现非线性分划的识别算法,其模型的原问题可表示为:

$$\min_{w,b,\xi}\frac{1}{2}\|w\|^2+C_+\sum_{y_i=+1}\xi_i+C_-\sum_{y_i=-1}\xi_i,\quad(3)$$

$$s.t.\quad y_i((w\cdot\Phi(x_i))+b)\geqslant 1-\xi_i,$$

$$i=1,2,\cdots,l,\quad(4)$$

$$\xi_i\geqslant 0,i=1,2,\cdots,l.\quad(5)$$

其中:C_+ 是对应肺结节样本的惩罚系数,C_- 是对应非结节样本的惩罚系数,ξ_i 为松弛变量,b 为常数偏置。通过拉格朗日乘子可将原问题转化为对偶问题:

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}y_iy_jK(x_i,y_i)\alpha_i\alpha_j-\sum_{j=1}^{l}\alpha_j,\quad(6)$$

$$s.t.\sum_{i=1}^{l}y_i\alpha_i=0$$

$$0\leqslant\alpha_i\leqslant C_+,y=+1$$

$$0\leqslant\alpha_i\leqslant C_-,y=-1.\quad(7)$$

则决策函数为:

$$f(x)=\mathrm{sgn}(g(x)).\quad(8)$$

其中:

$$g(x)=\sum_{i=1}^{l}\alpha_iy_iK(x_i,x_j)+b,\quad(9)$$

式(9)中偏置 b 求解方式如下:

$$b=y_i-\sum_{i=1}^{l}y_i\alpha_iK(x_i,x_j),\quad(10)$$

式(10)中 $K(x_i,x_j)$为核函数。RBF 核是归一化核,它的学习能力较强。其函数表达式为:

$$K_{rbf}(x_i,x_j)=\exp\left\{\frac{-\|x_i-x_j\|^2}{2g^2}\right\}.\quad(11)$$

式(11)中 g 为核的宽度,g 的大小直接影响核函数的学习能力,当 g 过小时,整个系统将丧失泛

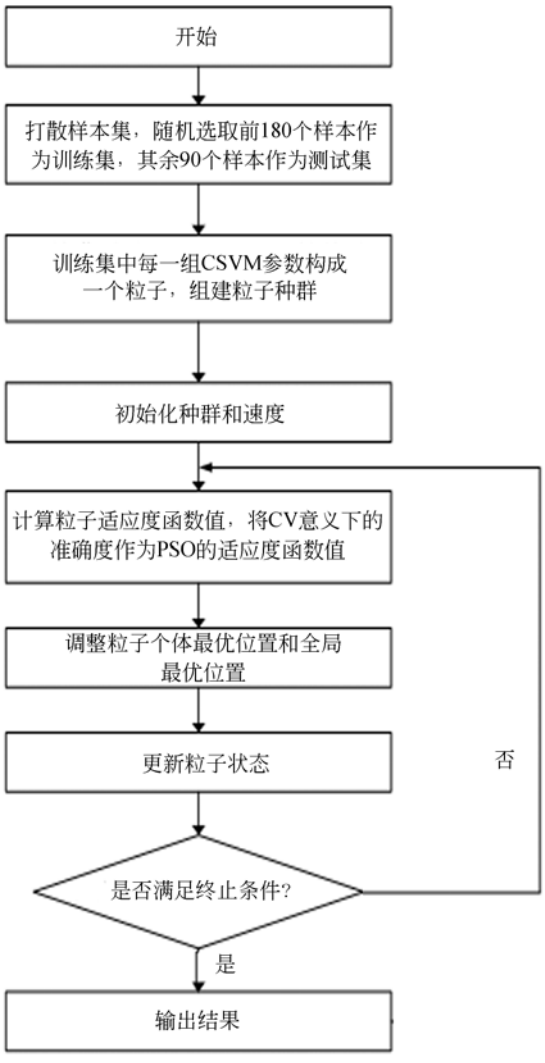


图 1　PSO-CSVM 算法流程图

Fig. 1　Flowchart of PSO-CSVM algorithm

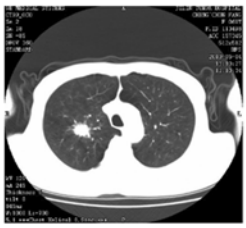
化能力，导致“过学习”的发生；当 g 过大时，其学习推广能力或对新样本点的准确判断能力大大降低，因此 RBF 核的学习能力随参数 g 的变大而削弱。

2.3 PSO-CSVM 算法

为了得到精度更高的代价敏感型 SVM 识别器，需要对 CSVM 建模的各个参数进行优化。本文利用 PSO 优化算法的全局搜索能力，对 CSVM 建模过程中的参数取值进行优化调整，以期更精确、快速地获得最佳的 CSVM 识别器。将 PSO 算法的思想引入 CSVM，提出 PSO-CSVM 算法，将训练集 CV 意义下的准确率作为适应度函数值，在训练集上得到的对应最高确率的参数组作为最优参数组。PSO-CSVM 算法流程图如图 1 所示。

3 实验分析及结果

实验中所采用的数据来自吉林省某三甲医院，共有 20 组病例约 700 幅 CT 图像，这些 CT 图像经过预处理、肺实质分割及左右肺分开等处理，共提取出 270 个孤立型候选结节的 ROI，其中包含 80 个结节，190 个假阳；然后从 ROI 中提取出共 13 维特征，其中包括 7 个形态特征，2 个灰度特征以及 4 个纹理特征[7]；最终将特征向量作为算法的输入来进行肺结节的识别。ROI 的提取步骤如下图 2 所示。



(a)原始肺CT图像
(a)Origin CT image

(b)预处理图像
(b)Preprocessed image

(c)肺实质分割后的右肺
(c)Right lung of lung parenchymal segmentation

(d)提取出的结节
(d)Extracted the nodule

(e)提取出的假阳
(e)Extracted the non-nodule

图 2　提取 ROI 步骤

Fig. 2　ROI extraction steps

实验平台采用 MATLAB，并利用 libsvm 工具箱进行仿真，工具箱中参数 w_i 为惩罚参数 C 的权重，本算法中正类样本的惩罚参数 C_+ 的权重设为 w_i，负类样本的惩罚参数 C_- 的权重为 w_{-1}；将 w_{-1} 设置为 1，参数 w_i 通过寻优得到，当 w_i 达到 2.4 左右识别效果最佳，C_+ 通过 $C_- \times w_i$ 得到。首先将样本随机打散，将所提取的 270 个样本的 13 维特征向量进行归一化处理，归一化的映射函数为：

$$x_{\text{normal}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}. \tag{12}$$

其中：x_{normal} 是归一化后的数据，$x_{\min}$、$x_{\max}$ 分别是原始数据 x 的最小值和最大值。然后选取其中 180 个样本作为训练集，90 个样本作为测试集，训练集与测试集中正、负样本个数如表 1 所示。

表 1　数据集正、负样本分布

Tab. 1　Positive and negative sample distribution of the data set

	训练样本集	测试样本集
样本总数	180	90
正样本数	52	28
负样本数	128	62

参数 C_- 和 g 的范围均设置为 $2^{-9} \sim 2^{9}$ 且在训练阶段采用五折交叉验证；PSO 算法进行寻优时，种群粒子数设置为 20，算法迭代进化次数为 200，c_1 初始为 1.5，c_2 初始为 1.8，当迭代次数到达 200 时，迭代终止，输出结果。

实验指标采用准确度，敏感度与特异度来衡量，准确率用 ACC 表示，敏感度用 SEN 表示，特异度用 SPE 表示。ACC，SEN 与 SPE 的表达式

分别为:

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

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

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

其中,TP 为识别出的真实肺结节;FP 为非结节而被误判为结节;FN 为真实肺结节被误判为非结节;TN 是识别出的非结节。

表 2 列出了几种不同算法所得到的不同结果,可以看到,PSO-CSVM 算法得到的识别准确率为 91.11%,敏感度为 85.71%,特异度为 93.55%,Grid-CSVM 算法得到的识别准确率为 88.88%,敏感度为82.14%,特异度为 91.94%,GA-CSVM 算法得到的识别准确率为 84.44%,敏感度为 76.86%,特异度为 91.94%,PSO-SVM 算法得到的识别准确率为 87.78%,敏感度为 82.14%,特异度为 90.32%,几种算法的 ROC 曲线如图 3 所示,ROC 曲线下对应的面积为:PSO-SVM 算法为 0.079 5,GA-CSVM 算法为 0.153 2,Grid-CSVM 算法为 0.053 6,PSO-CSVM 算法为 0.043 8,实验结果表明,PSO-CSVM 算法与遗传算法及网格搜索法相比较,无论是在 ACC 方面还是 SEN 方面的表现均优于二者,得到的正负类样本的 C_+ 和 C_- 的比值,即 w_i 为 2.4,这与负类和正类样本数量的比值 2.375 近似相等。

表 2　几种算法的比较

Tab.2　Comparison between different algorithms

算法	最优参数组	ACC/%	SEN/%	SPE/%	运行时间/s
PSO-SVM	C_-=7.417, g=5.12	87.78	82.14	90.32	53.64s
GA-CSVM	C_-=39.86, C_+=95.66, g=23.102	84.44	76.86	91.94	59.18
Grid-CSVM	C_-=32, C_+= 76.80, g=0.25	88.89	82.14	91.94	113.25
PSO-CSVM	C_-=11.70, C_+= 28.08, g=0.22	91.11	85.71	93.55	54.37

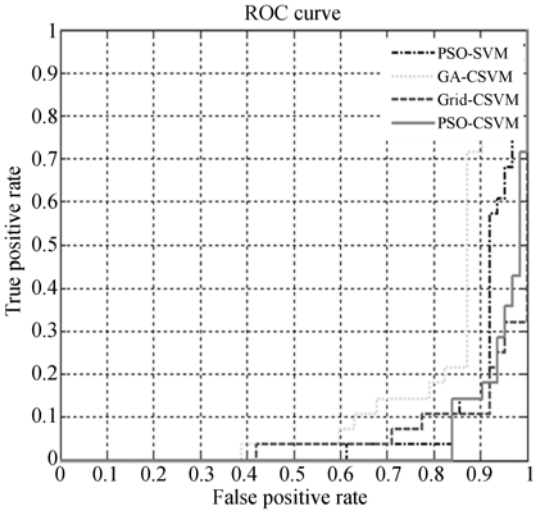


图 3　ROC 曲线

Fig.3　ROC Curve

4　结　论

本文提出了 PSO-CSVM 算法并应用于肺结节的识别,解决了数据集严重不均衡导致在分类过程中,分类平面向某一类数据方向倾斜,影响分类效果的问题,减少了漏检情况的发生;与 GA-CSVM 算法相比,PSO-CSVM 算法无需选择、交叉、变异等过程,算法简单;与 Grid-CSVM 算法相比,PSO-CSVM 算法的计算量小,节省了时间,且更容易得到全局最优解。实验结果表明,利用 PSO-CSVM 算法对肺结节进行识别,其识别准确率达到了 91.11%,敏感度达到 85.71%,特异度达到 93.55%,运行时间为 54.37 s,能够在一定程度上满足医学影像学对肺结节的实时性识别要求,正

负类样本惩罚系数的比值与负正类样本数量的比值近似相等，并且可将该算法推广至其他数据集。

参考文献：

[1] CASCIO D, MAGRO R, FAUCI F, *et al.*. Automatic detection of lung nodules in CT datasets based on stable 3D mass-spring models.[J]. *Computers in Biology & Medicine*, 2012, 42(11):1098-109.

[2] SHEN W, ZHOU M, YANG F, *et al.*. Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification[J]. *Pattern Recognition*, 2016, 61:663-673.

[3] FROZ B R, FILHO A O D C, SILVA A C, *et al.*. Lung nodule classification using artificial crawlers, directional texture and support vector machine[J]. *Expert Systems with Applications*, 2017, 69:176-188.

[4] CAO P, LIU X, YANG J, *et al.*. A multi-kernel based framework for heterogeneous feature selection and over-sampling for computer-aided detection of pulmonary nodules[J]. *Pattern Recognition*, 2017, 64(C):327-346.

[5] 李勇，苗壮，王青竹，等. 纹理引导的稀疏张量表示及在肺 CT 图像中的应用[J]. 光学 精密工程，2015，23(2):550-556.
LI Y, MIAO ZH, WANG Q ZH, *et al.*. Sparse tensor representation of texture guidance and its application in lung CT images [J]. *Opt. Precision Eng.*, 2015, 23(2):550-556. (in Chinese)

[6] SUN S, LI W, KANG Y. Lung nodule detection based on GA and SVM[C]// *International Conference on Biomedical Engineering and Informatics*, IEEE, 2016:96-100.

[7] LI Y, WEN D, WANG K, *et al.*. Mixed Kernel Function SVM for Pulmonary Nodule Recognition [C]. *Image Analysis and Processing - ICIAP 2013*. Springer Berlin Heidelberg, 2013:449-458.

[8] EBERHART R, KENNEDY J. A new optimizer using particle swarm theory [C]// *International Symposium on MICRO Machine and Human Science*, IEEE, 1995:39-43.

[9] 李姜，郭立红. 基于改进支持向量机的目标威胁估计[J]. 光学 精密工程，2014，22(5):1354-1362.
LI J, GUO L H. Target threat estimation based on improved support vector machine [J]. *Opt. Precision Eng.*, 2014, 22(5):1354-1362. (in Chinese)

[10] ZHANG J, CAO M Y, GAI W, *et al.*. Performance comparison of ESVM and CSVM for classifying the Lung Nodules on CT Scans[C]// *Seventh International Conference on Image and Graphics*, IEEE, 2013:409-413.

作者简介：

李　阳(1979－)，女，吉林长春人，博士，副教授，硕士导师，分别于 2003 年、2006 年、2014 年在吉林大学获得学士、硕士、博士学位，现为长春工业大学教师、东北师范大学统计学博士后，主要从事模式识别及图像处理方面的研究。E-mail：liyangyaya1979@sina.com

赵庆东(1989－)，女，吉林长春人，硕士研究生。2013 年于武汉轻工大学获得学士学位，研究方向为图像处理及模式识别。E-mail：1151715878@qq.com

田　颖(1978－)，女，吉林长春人，硕士，副主任医师，2002 年获得延边大学学士学位，2006 年获得吉林大学硕士学位，主要从事癌症及妇科疾病研究。E-mail：tianying781104.student@sina.com