DOI:10.3969/j.issn.1005-202X.2019.07.015

医学信号处理与医学仪器

基于SFS-SVM的乳腺癌预测模型的构建

赖胜圣[1],刘虔铖[1],余丽玲[1],刘文平[1],杨蕊梦[2],金浩宇[1]
1.广东食品药品职业学院医疗器械学院,广东 广州 510520;2.广州市第一人民医院/华南理工大学附属第二医院放射科,广东 广州 510180

【摘要】**目的:**构建基于序列前向选择算法(SFS)与支持向量机算法(SVM)分类器融合的乳腺癌预测模型,提高计算机辅助诊断技术对乳腺癌细针穿刺细胞病理的准确率。**方法:**对456组乳腺肿瘤病理数据作为训练集,利用SFS-SVM算法对30个特征进行筛选,得到最优的特征组合,再用112组乳腺肿瘤病理数据作为测试集验证,构建乳腺癌预测模型。该模型的预测精度通过5折交叉验证进行评价。评价指标包括:受试者工作特性曲线(ROC)下面积(AUC)、准确率(ACC)、敏感度和特异度。**结果:**构建了基于SFS-SVM的乳腺癌预测模型,该模型(AUC为98.39%,ACC为97.35%)相对于单独SVM算法(AUC为97.00%,ACC为92.42%)有一定的提高。**结论:**基于SFS特征选择的SVM分类器乳腺癌预测模型能较好地对乳腺癌进行辅助诊断。

【关键词】乳腺癌;预测模型;序列前向选择算法;支持向量机算法
【中图分类号】R318 【文献标志码】A 【文章编号】1005-202X(2019)07-0826-04

Construction of breast cancer prediction model based on SFS-SVM

LAI Shengsheng[1], LIU Qiancheng[1], YU Liling[1], LIU Wenping[1], YANG Ruimeng[2], JIN Haoyu[1]
1. School of Medical Devices, Guangdong Food and Drug Vocational College, Guangzhou 510520, China; 2. Department of Radiology, the Second Affiliated Hospital of South China University of Technology, Guangzhou First People's Hospital, Guangzhou 510180, China

Abstract: Objective To improve the accuracy of computer-aided diagnosis for fine needle aspiration pathology in breast cancer by employing the breast cancer prediction model based on sequential forward feature selection (SFS) algorithm and support vector machine (SVM) classifier. **Methods** The pathological data of 456 breast tumors were used as training set. A total of 30 features were screened by SFS-SVM algorithm to obtain the optimal feature combination, and then the pathological data of 112 breast tumors were used as test set to construct breast cancer prediction model. The prediction accuracy of the constructed model was evaluated with 5-fold cross-validation method. The evaluation indicators included area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, and specificity. **Results** Compared with SVM-based model which had an AUC of 97.00% and an accuracy of 92.42%, the breast cancer prediction model based on SFS-SVM had better performances, achieving an AUC of 98.39% and an accuracy of 97.35%. **Conclusion** The breast cancer prediction model based on SFS-SVM exhibits a good predictive efficacy on the auxiliary diagnosis of breast cancers.

Keywords: breast cancer; prediction model; sequential forward feature selection algorithm; support vector machine algorithm

前 言

根据世界卫生组织及文献报道,导致妇女死亡的5种最常见癌症(按发生频次排列)为乳腺癌、肺癌、胃癌、结肠直肠癌和宫颈癌[1-2]。近年来,在中国尤其是在发达的沿海地区,乳腺癌发病率不断上升,已经严重危及妇女的健康与生命[3]。乳腺肿瘤病灶常规、有效的检查方法之一是针吸细胞学检查,此方法要求医生在显微镜下观察,对细胞的形态、结构等进行分类、测量、判断,容易因为人为因素造成误诊、漏诊等。因此,计算机辅助诊断(Computer-Aided Diagnosis, CAD)应运而生。经过多年的研究与发展,CAD能够协助检测及分析可疑的乳腺癌病灶,并且取得了良好的效果。

【收稿日期】2019-02-17
【基金项目】广东省高等学校珠江学者岗位计划自主项目(2016);广东省自然科学基金(2018A030313282);广州市科技计划项目(201607010038);广东省医学科学技术研究基金(A2018338, A2019465);广东食品药品职业学院校级科研项目(2015YZ0012, 2015YZ0020)
【作者简介】赖胜圣,硕士,讲师,研究方向:医学图像处理及生理电信号处理,E-mail: laiss@gdyzy.edu.cn
【通信作者】金浩宇,博士,教授,研究方向:医学仪器研发,E-mail: jinhy@gdyzy.edu.cn

刘兴华等[4]提出用Sigmoid核函数的支持向量机算法(SVM)对乳腺癌的辅助诊断准确率达到96.24%,但此方法忽视了多项式核函数时优异的特异度指标,不能令人满意。Mu等[5]用基于Supervised Compact Hyperspheres的分类器对乳腺肿瘤进行良恶性分类,获得较高的准确率。吴辰文等[6]提出一种基于随机森林模型下Gini指标特征的SVM算法分析各个特征对分类结果的重要性,并对乳腺肿瘤分类判别进行验证,其准确率为97.7%,但对训练样本较少的对象会导致算法识别性能降低。近年来出现了用J48决策树算法[7]、二步SVM算法[8]、粒子群算法[9]等方法研究,提高了乳腺肿瘤良恶性分类判别的准确率。另外,有研究者运用扩展卡尔曼滤波器与粒子群算法结合[10]及权重粒子群最小二乘支持向量机[11-12]等各类算法对乳腺癌进行鉴别。这些方法各具特色,优点明确,取得了不错的效果。但在样本量小、非线性、特征数目多的乳腺癌细胞图片的处理上仍存在困难,还有提升的空间。

由于乳腺肿瘤病灶组织发生病变,然而它的细胞显微图像与正常的组织显微图像有所不同,因此需要采用分类能力比较强的算法来进行乳腺肿瘤诊断。本研究提出一种基于序列前向选择算法(Sequential Forward feature Selection, SFS)与SVM分类器融合的方法,用于构建乳腺癌预测模型。

1 材料与方法

1.1 材料

威斯康辛大学威斯康辛诊断乳腺癌数据库(Wisconsin Diagnostic Breast Cancer, WDBC)[13-14]共包括569例乳腺肿瘤,其中良性肿瘤357例,恶性肿瘤212例。每个病例的1组数据包括采用组织中各细胞核的10个特征量的平均值、标准差和最坏值(各特征的3个最大数据的平均值),共30个数据。10个特征量分别是细胞核图像的细胞核半径、质地、周长、面积、光滑性、紧密度、凹陷度、凹陷点数、对称度及断裂度。数据文件中每组数据共分为32个值,第一个字段为病例编号,第二个字段为确诊结构,B(Benign)为良性肿瘤,M(Malignant)为恶性肿瘤,第3~12个字段是该病例肿瘤组织的各细胞核显微图像的10个量化特征的平均值;第13~22个字段是相应的标准差;第23~32个字段是相应的最坏值。这些特征与肿瘤性质有密切关系。为此,需要建立一个确定的模型描述数据库中各个量化特征与肿瘤特征关系,进而可以根据细胞核显微图像的量化特征诊断是否为乳腺癌。

1.2 SVM原理

SVM是一种典型的非概率两类分类器,在解决小样本、非线性及高维模式识别中有许多特有的优势。其算法思想是将所研究的对象向真实模型的一种逼近,将原始特征用核函数进行变换映射到高维空间,进而分解其特征矩阵[4]。由SVM的定义可知,其由距离超平面最近的点(称为支持向量)决定,对于线性可分两类数据是一条最优分割直线,而对于高维数据点则是一个最优分割超平面。例如给定一数据集 (x_i,y_i), $x_i\in R^n, y_i\in\{-1,+1\}$, $i=1,\cdots,n$,定义分割超平面为 $w^Tx+b=0$,其中 w 和 b 是SVM的参数。在数据点中找到距离分割平面最近的点(支持向量),寻找出最优的 w 和 b 来最大化支持向量到分割超平面的距离,使得支持向量距离该超平面的间隔最大,则有目标函数:

$$\arg\min_{w,b,}\frac{1}{2}\|w\|^2+c\sum_{i=1}^{n}\xi_i,\quad \xi_i\geqslant 0,\ i=1,2,\cdots,n \tag{1}$$

$$\text{s.t.},\ y_i\cdot\left(w^Tx+b\right)\geqslant 1-\xi_i,\ i=1,2,\cdots,n \tag{2}$$

其中, ξ_i 是松弛变量, c 为松弛因子,控制着对噪声的惩罚程度。当数据集线性不可分时,通过核函数将数据映射到高维空间可以使得数据线性可分,计算时候只需要计算核函数 $K\left(x_i,x_j\right)=\left\langle\Phi\left(x_i\right),\Phi\left(x_j\right)\right\rangle$ 。再引入拉格朗日乘子法将目标函数转变为式(2)二次规划的对偶问题:

$$\max_{\alpha}\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_jy_iy_jK\left(x_i,x_j\right) \tag{3}$$

$$\text{s.t.},\ \sum_{i=1}^{n}\alpha_iy_i=0,\quad \alpha_i\geqslant 0,\ i=1,2,\cdots,n \tag{4}$$

其中, α_i 是拉格朗日乘子。这样,通过利用序列最小优化算法求解出 α ,即可求得 w 和 b ,最终可得最优的超平面。

1.3 乳腺癌特征选择及模型流程

WDBC中共有569例乳腺肿瘤数据,其中212例为恶性肿瘤和357例为良性肿瘤。我们将数据随机分为模型训练数据456例(占80%),模型验证数据113例(占20%)。采用SFS方法对乳腺肿瘤病理切片图像提取的特征集合进行降维或特征选择。SFS是一种自下而上的搜索方法,目的是为了去除不相关及多余的特征量,降低特征个数,寻找出最优特征子集,进而能够提高模型的精确度。假设给定一特征集合 $F=(f_1,f_2,\cdots,f_n)$,模型目标函数为 $J(\cdot)$,每次通过特征选择从特征集 F 中选择出一个子集 S,其中该子集 S 对于任何的子集 T 都有 $J(S)>J(T)$。对于SFS算法,特征子集 X 从空集开始,通过5折交叉验证从所有特征中寻找出使得目标函数 J 达到最优的第一个

特征，此后每次只从未选择的特征集中选择一个特征x加入特征子集X使得J最优。重复上述过程，当最佳改进使特征集性能变坏或达到最大允许个数时，也直到J达到最优结果时，停止选择[15]。同时，该算法运算量相对较小，但没考虑特征之间的相关性。

采用SVM作为分类器，SVM工具包采用林智仁教授公开的LIBSVM库[16]，在本研究中，采用的核函数为径向基核函数，并且在MATLAB（2016b）环境下实现模型的构建和评估。

基于SFS-SVM的乳腺癌预测模型流程如图1所示。

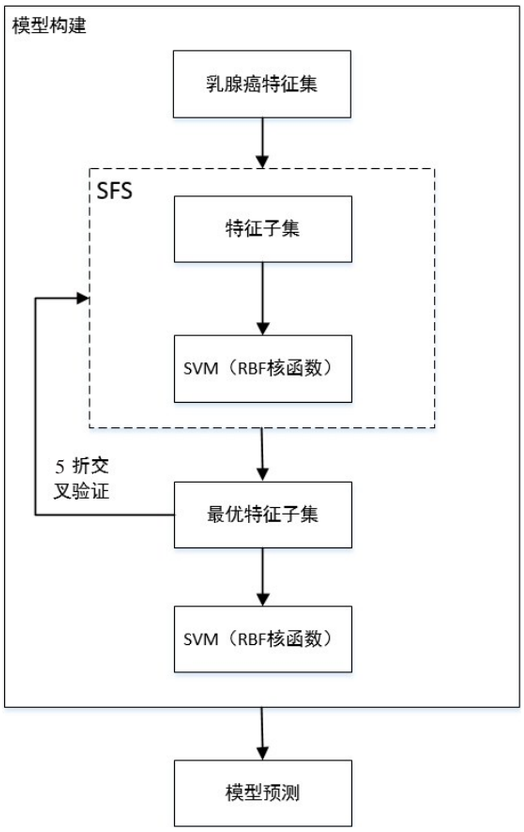

图1 SFS–SVM乳腺癌预测模型流程图

Fig.1 Flow chart of breast cancer prediction model based on sequential forward feature selection (SFS) and support vector machine (SVM)

1.4 评价指标

乳腺癌预测模型的评价指标包括准确率（ACC）、灵敏度（SEN）、特异性（SPE）以及ROC曲线下面积（AUC）。AUC指的是ROC曲线下的面积，其面积越大，则提示分类器的分类效果越好，AUC的值域为[0.5, 1.0]。ACC、SEN、SPE分别定义为：

$$ACC = (TP + TN)/(TP + FP + FN + TN) \quad (5)$$

$$SEN = TP/(TP + FN) \quad (6)$$

$$SPE = TN/(TN + FP) \quad (7)$$

其中，TP为真阳性（True Positive, TP），表示被预测模型正确判别的正样本个数；TN为真阴性（True Negative, TN），表示被预测模型正确判别的负样本个数；FP为假阳性（False Positive, FP），表示被预测模型错误判别的负样本个数；FN为假阴性（False Negative, FN），表示被预测模型错误判别的正样本个数。

2 结 果

2.1 特征选择结果

通过SFS序列前向特征选择算法，对30个具有显著性差异的量化特征进行特征选择，得到最优特征组合，如表1所示。3个最优的量化特征分别是细胞核半径最坏值、质地最坏值及凹陷度最坏值。

表1 SFS特征选择结果（$\bar{x}\pm s$）

Tab.1 Result of SFS for feature selection (*Mean*±*SD*)

肿瘤性质	细胞核半径最坏值	质地最坏值	凹陷度最坏值
恶性	21.13±4.27	29.32±5.42	0.45±0.18
良性	13.38±1.98	23.52±5.49	0.16±0.14

2.2 SFS-SVM及SVM预测准确性比较

由SFS-SVM及SVM构建的乳腺癌预测模型结果分别如表2和表3所示。对于SFS-SVM模型训练组的评估指标结果为：AUC 99.16%、ACC 96.49%、SEN 96.47%、SPE 96.50%；而SVM模型训练组评价指标结果为：AUC 97.09%、ACC 92.48%、SEN 96.91%、SPE 85.02%。对于SFS-SVM模型测试组评估指标结果为：AUC 98.39%、ACC 97.35%、SEN 97.62%、SPE 97.18%；恶性肿瘤里有1例被诊断为良性肿瘤，良性肿瘤中有2例被诊断为恶性肿瘤。对于SVM模型测试组，评价指标结果为：AUC 97.00%、ACC 92.42%、SEN 96.91%、SPE 84.46%。从表2和表3可以看出，基于SFS-SVM的预测模型，从测试组结果可知ACC、SPE、AUC方面比SVM分类方法都有所提升，其中ACC提高4.93%，SPE提高12.72%，AUC提高1.39%。根据表2和表3，画出SFS-SVM模型及SVM模型的ROC曲线如图2所示，可以看出SFS-SVM模型对应的ROC曲线最佳。

4 结 语

本研究所构建的基于SFS-SVM算法的乳腺癌预测模型，采用序列前向特征选择算法进行特征选择，去除与样本分类无关的特征量，实现高维特征的降维，并且通过5折交叉检验的验证方法尽量保证模型的鲁棒性。经过WDBC数据集的验证，相对于传统的SVM分类器[17]

表2 SFS–SVM预测结果(%)

Tab.2 Prediction results of SFS–SVM (%)

模型	*n*	AUC	ACC	SEN	SPE
训练集	456	99.16	96.49	96.47	96.50
测试集	113	98.39	97.35	97.62	97.18

AUC:ROC曲线下面积;ACC:准确率;SEN:灵敏度;SPE:特异性

表3 SVM预测结果(%)

Tab.3 Prediction results of SVM (%)

模型	*n*	AUC	ACC	SEN	SPE
训练集	456	97.09	92.48	96.91	85.02
测试集	113	97.00	92.42	96.91	84.46

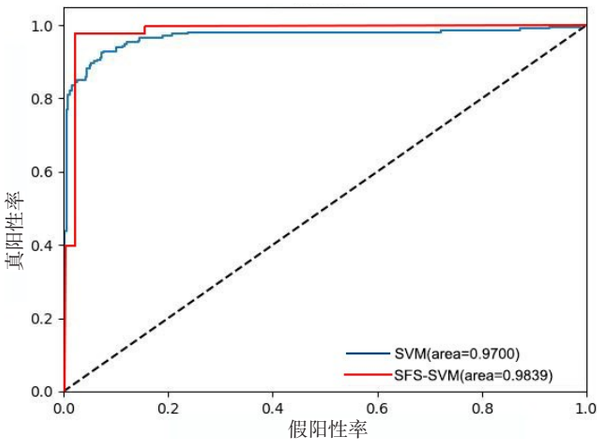

图2 SFS–SVM及SVM预测模型的ROC曲线

Fig.2 ROC curve of prediction model based on SFS–SVM *vs* SVM

以及另外一些改进型的SVM方法,如RS-SVM[18]、PSVM、NSVM、LP-SVM、LS-SVM、SS-SVM[19],本文所提出的SFS-SVM算法的3项评估指标ACC、SEN及SPE都取得了较高值,说明此方法可以成为准确、可信的乳腺癌辅助诊断工具,具有良好的前景。本研究用SFS-SVM算法预测模型在预测乳腺癌的同时,能计算出各个细胞核特征对乳腺癌预测贡献的大小,去除了大量的冗余信息。根据结果可知,细胞核半径最坏值、质地最坏值及凹陷度最坏值是基于SFS-SVM算法乳腺癌预测模型的重要指标,这3个特征具有决定性意义。但文献[20]采用多表面方法树算法筛选出3个最优特征子集为面积最坏值、平滑最坏值和质地平均值。这说明依据不同算法构建的乳腺癌预测模型,所选择出的最优特征亦有所区别,特征选择的结果直接影响着分类器性能。

目前,我们把SFS-SVM模型仅应用于针吸穿刺细胞检查的临床数据,根据最新的报道[15],SFS-SVM算法模型可以有效应用于宫颈癌放疗中直肠毒性预测,这给我们很好的启示,SFS-SVM算法模型可以应用于其它疾病的影像图像的辅助诊断。

【参考文献】

[1] WINGO P A, TONG T, BOLDEN S. Cancer statistics[J]. CA Cancer J Clin, 1995, 45(1): 8-30.

[2] JEMAL A, SIEGEL R, WARD E, et al. Cancer statistics 2009[J]. CA Cancer J Clin, 2009, 59(4): 225-249.

[3] 徐光炜, 胡永昇, 阚秀. 中国10万妇女乳腺癌筛查初探[J]. 中国肿瘤, 2010, 19(9): 565-568.
XU G W, HU Y S, KAN X. The preliminary report of breast cancer screening for 100 000 women in China[J]. China Cancer, 2010, 19(9): 565-568.

[4] 刘兴华, 蔡从中, 袁前飞, 等. 基于支持向量机的乳腺癌辅助诊断[J]. 重庆大学学报, 2007, 30(6): 140-144.
LIU X H, CAI C Z, YUAN Q F, et al. Computer-aided diagnosis of breast cancer based on support vector machine [J]. Journal of Chongqing University, 2007, 30(6): 140-144.

[5] MU T T, ASOKE K. East cancer diagnosis from fine-needle aspiration using supervised compact hyperspheres and establishment of confidence of malignancy [C]//16th European Signal Processing Conference. 2008.

[6] 吴辰文, 李长生, 王伟, 等. 一种改进的SVM算法在乳腺癌诊断方面的应用[J]. 计算机工程与科学, 2017, 39(3): 562-566.
WU C W, LI C S, WANG W, et al. Application of an improved support vector machine algorithm in the diagnosis of breast cancer [J]. Computer Engineering & Science, 2017, 39(3): 562-566.

[7] DEVI R D, DEVI M I. Outlier detection algorithm combined with decision tree classifier for early diagnosis of breast cancer[J]. Int J Adv Eng Technol, 2016, 12: 93-98.

[8] ZENG N, WANG Z, ZHANG H, et al. A novel switching delayed PSO algorithm for estimating unknown parameters of lateral flow immunoassay[J]. Cogn Comput, 2016, 8(2): 143-152.

[9] KUMAR U K, NIKHIL M B, SUMANGALI K. Prediction of breast cancer using voting classifier technique[C]//In Proceedings of the IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials, Chennai, India. 2017.

[10] LATCHOUMI T P, PARTHIBAN L. Abnormality detection using weighed particle swarm optimization and smooth support vector machine[J]. Biomed Res, 2017, 28(11): 4749-4751.

[11] OSMAN A H. An enhanced breast cancer diagnosis scheme based on two-step-SVM technique[J]. Int J Adv Compute Sci Appl, 2017, 8(4): 158-165.

[12] YUE W B, WANG Z D, CHEN H W. Machine learning with applications in breast cancer diagnosis and prognosis[J]. Designs, 2018, 2(13): 1-17.

[13] FRANK A, ASUNCION A. UCI machine learning repository[DB/OL]. http://archive.ics.uci.edu /ml.

[14] WOLBERG W H, MAMGASARIAN O L. Multisurface method of pattern separation for medical diagnosis applied to breast cytology[J]. Proc Natl Acad Sci U S A, 1990, 87(12): 9193-9196.

[15] CHEN J, CHEN H, ZHING Z, et al. Investigating rectal toxicity associated dosimetric features with deformable accumulated rectal surface dose maps for cervical cancer radiotherapy[J]. Radiat Oncol, 2018, 13(1): 125.

[16] CHANG C C, LIN C J. LIBSVM: a library for support vector machines [M]. ACM, 2011.

[17] BENNETT K P, BLUE J A. A support vector machine approach to decision trees[C]//In Proceedings of the IEEE International Joint Conference on Neural Networks. 1998: 2396-2401.

[18] CHEN H L, YANG B, LIU J, et al. A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis[J]. Expert Syst Appl, 2011, 38(7): 9014-9022.

[19] AZAR A T, El-SAID S A. Performance analysis of support vector machines classifiers in breast cancer mammography recognition[J]. Neural Comput Appl, 2014, 24(5): 1163-1177.

[20] WOLBERG W H, STREET W N, HEISEY D M, et al. Computerized breast cancer diagnosis and prognosis from fine-needle aspirates[J]. Arch Surg, 1995, 130(5): 511-516.

(编辑:陈丽霞)