DOI:10.3969/j.issn.1005-202X.2021.03.022

医学人工智能

基于密度分布特征及机器学习诊断COVID-19相关性肺炎

韩冬[1],于勇[2],贺太平[2],段海峰[1],贾永军[1],张喜荣[2],郭佑民[3],于楠[1]
1.陕西中医药大学附属医院医学影像科,陕西 咸阳 712000;2.陕西中医药大学医学技术学院,陕西 咸阳 712000;3.西安交通大学第一附属医院医学影像科,陕西 西安 710061

【摘要】目的:基于密度分布特征及机器学习诊断新型冠状病毒(COVID-19)相关性肺炎。**方法:**回顾性收集经荧光逆转录聚合酶链反应检测确诊COVID-19的患者42例(COVID-19组),社区获得性肺炎43例(对照组)。共获得211份胸部CT图像,以6:4比例分层抽样为训练集(126份)及验证集(85份)。采用一种CAD软件中的肺炎模块获得肺炎不同密度区间所占全肺体积的百分比(P/L%)。密度分布特征降维后采用支持向量机(SVM)建模,并评价4种核函数的SVM模型的诊断效能。**结果:**两组患者的年龄、性别及出现胸膜腔积液的构成比差异均无统计学意义(P>0.05)。肺炎密度分布特征降维后获得32个特征。基于该32个特征建立的4种核函数SVM模型中,多项式SVM模型在验证集的效能最高,受试者特征曲线(ROC)的曲线下面积为0.897(95%可信区间0.828~0.966),P<0.001。准确性为0.906(95%可信区间0.823~0.959),敏感性为0.906,特异性为0.906。**结论:**基于密度分布特征及机器学习诊断COVID-19相关性肺炎有较高的效能,有助于快速筛选COVID-19患者。

【关键词】新型冠状病毒;肺炎;密度分布特征;机器学习
【中图分类号】R318;R563.1 **【文献标志码】**A **【文章编号】**1005-202X(2021)03-0387-05

Diagnosis of COVID-19 associated pneumonia based on density distribution features and machine learning

HAN Dong[1], YU Yong[2], HE Taiping[2], DUAN Haifeng[1], JIA Yongjun[1], ZHANG Xirong[2], GUO Youmin[3], YU Nan[1]
1. Department of Medical Imaging, Affiliated Hospital of Shaanxi University of Chinese Medicine, Xianyang 712000, China; 2. School of Medical Technology, Shaanxi University of Chinese Medicine, Xianyang 712000, China; 3. Department of Medical Imaging, the First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710061, China

Abstract: Objective To diagnose corona virus disease 2019 (COVID-19) associated pneumonia based on density distribution features and machine learning. **Methods** The clinical information of 42 patients with COVID-19 confirmed by RT-PCR (COVID-19 group) and 43 patients with community-acquired pneumonia (control group) were retrospectively collected. A total of 211 chest CT images were obtained, and according to stratified sampling based on a proportion of 6 to 4, the chest images were divided into training set (126) and validation set (85). The percentages of different density intervals of pneumonia in the total lung volume (P/L%) were obtained using a pneumonia module in CAD software. Support vector machine (SVM) was used for modeling after the dimensionality reduction of density distribution features, and the diagnostic efficiency of SVM models with 4 different kernel functions was evaluated. **Results** There was no significant difference in age, gender and constituent ratio of pleural effusion between two groups (P>0.05). A total of 32 features were obtained after the dimensionality reduction of pneumonia density distribution features. Among SVM models with 4 different kernel functions based on these 32 features, polynomial SVM model has the highest efficiency in validation set, and the area under receiver operating characteristic curve was 0.897 (95% confidence interval 0.828-0.966) (P<0.001). The accuracy, sensitivity and specificity of polynomial SVM model were 0.906 (95% confidence interval: 0.823-0.959), 0.906 and 0.906. **Conclusion** The diagnosis of COVID-19 associated pneumonia based on the density distribution features and machine learning has a high efficiency, which is helpful for the rapid screening of COVID-19 patients.

Keywords: novel corona virus; pneumonia; density distribution features; machine learning

【收稿日期】2020-10-16
【基金项目】陕西中医药大学学科创新团队建设项目(2019-QN09;2019-YS04)
【作者简介】韩冬,主治医师,研究方向:机器学习在医学影像的临床应用,E-mail: hundnn@qq.com
【通信作者】于楠,副教授,研究方向:胸部影像学,E-mail: yunan0512@sina.com

前 言

新型冠状病毒病(Corona Virus Disease 2019, COVID-19)被WHO列为突发公共卫生事件,该病普遍易感,对大众正常生产生活造成极大影响[1-2]。实时荧光逆转录聚合酶链反应(RT-PCR)检测是诊断COVID-19的金标准,但敏感性及特异性均较低,检测时间长。尤其当前核酸检测试剂依然紧缺,进一步影响了核酸筛查工作顺利开展。胸部CT检查快速便捷,敏感性高,对筛查COVID-19有重要价值[3-4]。COVID-19典型肺部表现为多发胸膜下的磨玻璃影[5]。但作为病毒性肺炎的一种,与其它肺炎CT表现存在重叠[6]。即使是有经验的影像科医师对COVID-19的诊断特异性仍较低[7]。近年来机器学习在医学影像学领域的数据分析长足发展,对疾病定量评价的效能及效率均有较高优势,其中支持向量机(Support Vector Machine, SVM)在解决非线性分类任务中有良好表现。本研究探讨基于密度分布特征及SVM诊断COVID-19相关性肺炎。

1 材料与方法

1.1 研究对象

回顾性收集整理3家医院发热门诊自2020年1月~3月就诊患者的胸部CT图像,并符合以下纳入和排除标准。纳入标准:①CT图像层厚1~2 mm,有可视的肺炎病变;②接受呼吸道或血液标本实时荧光RT-PCR新型冠状病毒核酸检测;③接受血常规、其他病毒检测、支原体等病原学检测。排除标准:①CT图像存在呼吸运动伪影;②CT图像存在较严重的间质病变;③病原学诊断不明确者。最后共纳入85例患者,男性40例,女性45例,年龄14~89岁,平均年龄(45.92±18.63)岁。根据RT-PCR新冠肺炎核酸检测结果将患者分为COVID-19组和对照组。COVID-19组包括42例患者的132份CT图像,对照组包括43例患者的79份CT图像。将上述211份CT图像按6:4比例分层抽样为训练集(126份)及验证集(85份)。

1.2 肺炎密度分布特征

采用一种计算机辅助分析平台——“数字肺”(Digital Lung DEXIN, China)中肺炎模块对所有患者胸部CT图像进行分析。该模块通过训练2 000例社区获得性肺炎患者的胸部CT图像获得全卷积神经网络模型用于自动分割肺炎区域,并计算肺炎体积及其-1 000~1 000 HU密度分布特征,即肺炎-1 000~1 000 HU(间隔10 HU)的体积所占全肺体积的百分比($V_{pneumonia}/V_{lung}$, P/L%)。根据专业知识以100 HU为阈值,删除100 HU以上的P/L%,使用以下公式使-1 000~100 HU的P/L%之和为1:

新密度区间P/L%=原密度区间P/L%×100/$\sum_{-1000HU}^{100HU}$(P/L%),最终获得-1 000~100 HU不同密度区间(间隔10 HU)的P/L%,共110个密度分布特征。

1.3 统计学分析

数据分析采用R语言(v.3.6.3),以P<0.05为差异具有统计学意义。服从正态分布且方差齐的连续资料比较采用独立样本t检验,否则采用Mann-Whitney U检验。计数资料比较采用χ^2检验或Fisher确切概率法。采用“caret”包中递归特征消除(RFE)对密度分布特征降维,其最优参数组合采用10折交叉验证确定。降维后特征采用“e1071”包进行SVM建模,其核函数分别使用线性、多项式、径向基及Sigmoid函数。采用“pROC”包对SVM模型进行受试者特征(ROC)曲线分析,分别计算训练集及验证集的AUC、准确性、敏感性及特异性。

2 结 果

2.1 一般资料比较

两组患者的年龄和性别差异均无统计学意义(P>0.05),两组患者出现胸膜腔积液的构成比均较低,差异无统计学意义(P=1.000),见表1。

表1 两组患者一般资料比较

Tab.1 Comparison of general data between two groups

参数	COVID-19组	对照组	统计量	P值
年龄/岁	45.62±16.68	46.21±20.56	0.145[a]	0.885
性别(女/男)	24/18	21/22	0.588[b]	0.443
胸膜腔积液(无/有)	41/1	41/2	NA	1.000

a为独立样本t检验;b为卡方检验;NA为Fisher确切概率法,无统计量

2.2 密度分布特征比较

两组患者的胸部CT图像经“数字肺”肺炎模块自动分割肺炎区域,经处理后获得-1 000~100 HU不同密度间隔(间隔10 HU)的P/L%,两组病变的平均直方图差别如图1所示。

2.3 密度分布特征降维

密度分布特征采用RFE降维后10折交叉验证结果表明以下32个特征的模型准确性最高为0.891±0.084,其Kappa值为0.775,如图2。该32个特征按重要性从高到低排序为HU(-870~-861)、HU(-850~-841)、HU(-860~-851)、HU(90~99)、HU(-840~-831)、HU(-880~-871)、HU(-900~-891)、HU(80~89)、HU(-830~-821)、HU(-890~-881)、HU(-910~-901)、HU(-930~-921)、HU(-940~-931)、HU(-920~-911)、HU(-990~-981)、HU(70~79)、HU(-960~-951)、HU(-970~-961)、HU(-980~-971)、HU

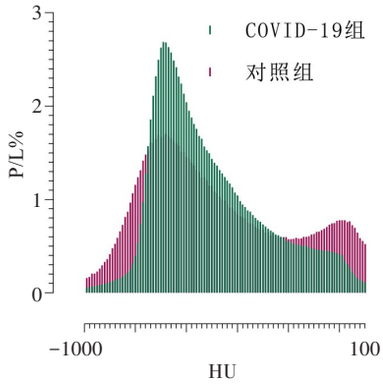


图1 两组肺炎P/L%的平均直方图比较

Fig.1 Comparison of mean histogram of P/L% between two groups of pneumonia

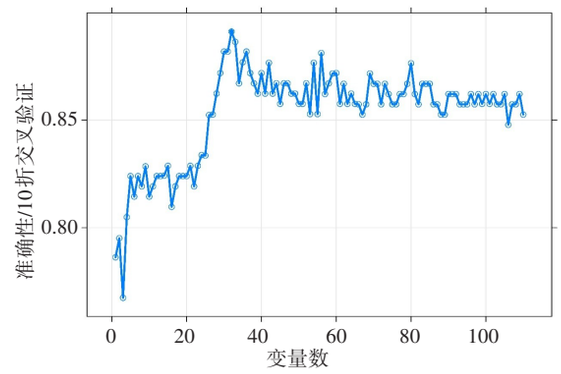


图2 递归特征消除(RFE),不同特征数模型与其准确性的关系

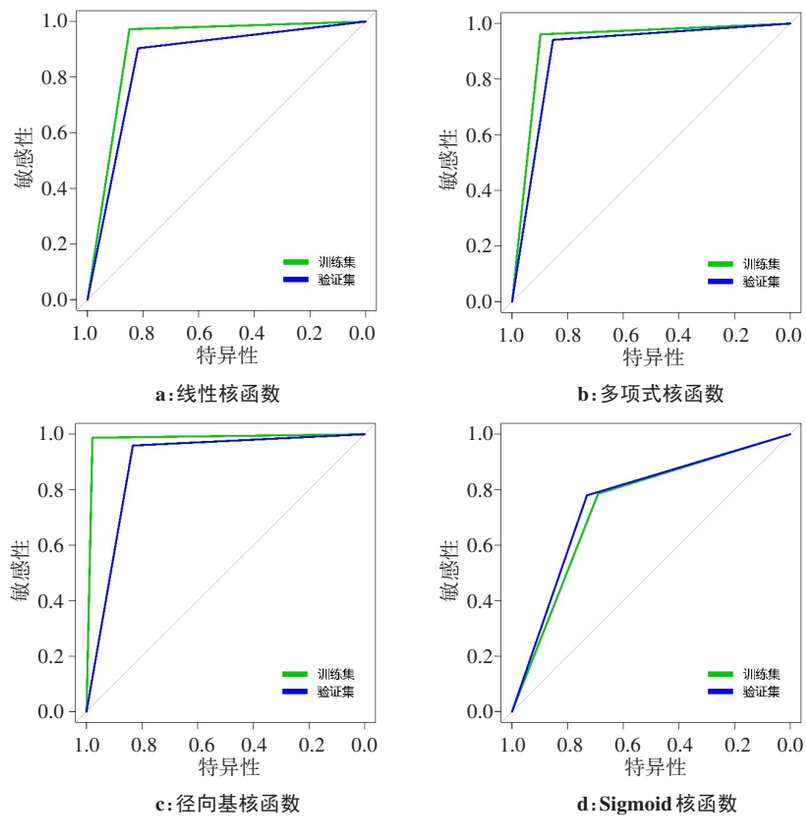
Fig.2 Recursive feature elimination, the relationship between models with different number of features and their accuracies

当特征为32个时准确性最高(即蓝色实心点)

(-950~-941)、HU(60~69)、HU(-920~-811)、HU(-1 000~-991)、HU(-810~-801)、HU(50~59)、HU(40~49)、HU(30~39)、HU(-800~-791)、HU(-630~-621)、HU(-650~-641)、HU(-640~-631)及HU(-620~-611)。值得注意的是上述32个密度分布特征包括了3个连续的密度区间,即HU(-1 000~-791)、HU(-650~-611)及HU(30~100)。

2.4 SVM建模及验证

以上述降维后的32个密度分布特征作为自变量进行SVM建模,分别使用线性、多项式、径向基及Sigmoid共4种核函数建模,4种SVM模型的ROC曲线如图3,其AUC(95%可信区间)、准确性(95%可信区间)、敏感性及特异性如图4。其中多项式SVM模型在验证集的AUC及准确性最高,分别为0.897(95%可信区间0.828~0.966, P<0.001),准确性为0.906(95%可信区间0.823~0.959),敏感性为0.906,特异性为0.906。



a:线性核函数　b:多项式核函数

c:径向基核函数　d:Sigmoid核函数

图3 4种SVM模型在训练集及验证集的ROC曲线

Fig.3 ROC curve of 4 SVM models in training set and validation set

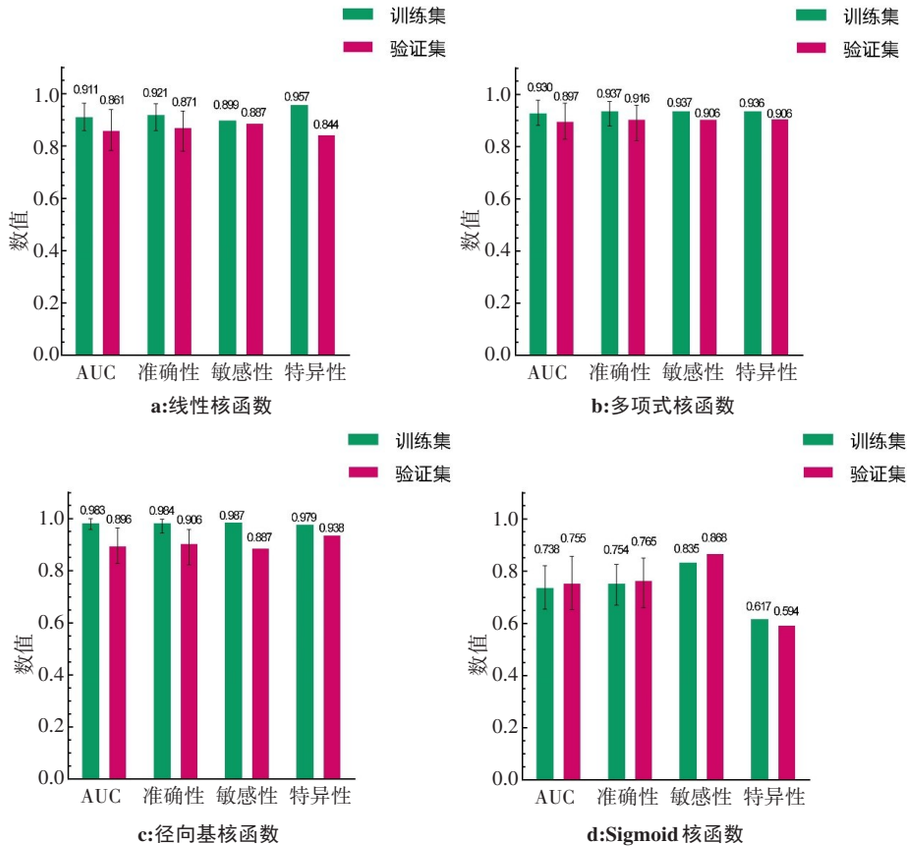


图4 4种SVM模型在训练集及验证集的AUC(95%可信区间)、准确性(95%可信区间)、敏感性及特异性

Fig.4 AUC (95% confidence interval), accuracy (95% confidence interval), sensitivity and specificity of 4 SVM models in training set and validation set

3 讨 论

本研究结果表明肺炎密度分布特征在COVID-19相关性肺炎及对照组中存在差别,尤其是降维后的32个密度分布特征中。采用上述特征建立的SVM鉴别诊断模型表现出不同的诊断效能,其中多项式SVM模型无论在训练集还是验证集的AUC、准确性、敏感性及特异性方面均有良好的鉴别诊断效能。

几乎所有肺部病灶在CT图像上非常容易被发现,故胸部CT对COVID-19相关性肺炎的检测有较高的敏感性,且高于RT-PCR[7-8]。虽然COVID-19相关性肺炎在CT图像有一些典型表现,如早期出现单侧或双肺胸膜下斑片状磨玻璃影;进展期肺炎病灶数目增多,范围增大,进一步累及多个肺叶,实变与磨玻璃影共存;重症患者表现为双肺弥漫性病变,表现为多发的磨玻璃影及"铺路石征"等[9],但即使有以上CT表现,COVID-19相关性肺炎的胸部CT表现仍与其他各种肺感染性病变重叠,诊断特异性不高[10-12]。

肺炎作为一种弥漫性病变,其CT图像分析采用传统半定量方法存在一定局限性。定量CT软件系统可进行肺部病变范围的自动分割及定量,并能对比病变动态变化。本研究通过定量CT软件系统肺炎模块自动提取获得肺炎的密度分布特征经处理后共110个。各参数间存在不同程度的相关性。将所有特征用于建模,一方面模型极易过拟合,另一方面模型复杂度及运算成本增加。本研究采用RFE的主要原理是采用上述110个参数不断构建模型,筛选对鉴别两组病变最重要特征,然后对剩余特征再次迭代,直到遍历所有特征。本研究降维后剩余32个特征,包括HU(-1 000~-791)、HU(-650~-611)及HU(30~100)3个连续的密度区间。根据CT值设定,HU(-1 000~-791)及HU(-650~-611)密度区间代表磨玻璃区域,HU(30~100)密度区间则代表实性区域。表明除上述密度区间外,两组病变的密度分布在其它密度区间存在重叠,因此在RFE降维过程中被剔除。

近期有研究提出在COVID-19人群中,采用全卷积神经网络自动计算了HU(-700~-500)、HU(-500~-200)及HU(-200~60)3个定量特征,用于反映磨玻璃区域体积占比、磨玻璃-实性过渡区域占比及实性区域占比,最终结果表明在第0~4天胸部CT图像磨玻璃区域体积占比及实性区域占比的变化对患者向重症进展具有较强的预测能力,效能优于基于临床及实验室检查指标[13]。表明采用密度分布特征定量评

价COVID-19相关性肺炎具有一定临床实用价值。近期的一项大样本研究采用深度学习技术构建了基于胸部CT图像检测COVID-19的三维卷积神经网络模型，在鉴别COVID-19、社区获得性肺炎以及非肺炎人群中表现出较高效能，表明深度学习技术在肺炎图像分割及分类方面有巨大应用潜力[14]。深度学习技术对图像直接分类是其高阶应用，需要搭建专用的神经网络架构，技术要求较高，同时运算硬件要求也较高。另一方面深度学习技术如卷积神经网格提取特征（卷积及池化等）的不可解释性导致阅读者对问题及任务的理解存在一定难度。

本研究尚存在以下局限性：①本研究的样本量有限，部分患者多次检查的CT图像被纳入分析。随着治疗或病情变化，患者胸部CT图像可能发生不典型变化。②对照组个别患者仅进行了一次RT-PCR检测，可能存在假阴性。但此类患者在后续的严密观察中肺部病变已吸收或没有新发病灶。③本研究中COVID-19组病例来自3家医院，部分患者的疫区史、实验室检查及临床信息缺失。

总之，基于密度分布特征及机器学习诊断COVID-19相关性肺炎有较高的效能，有助于快速筛选COVID-19患者。

致谢：感谢西北大学陈一兵博士对本文数据进行分析和指导。

【参考文献】

[1] GUAN W J, NI Z Y, HU Y, et al. Clinical characteristics of coronavirus disease 2019 in China[J]. N Engl J Med, 2020, 382(18): 1708-1720.
[2] PHELAN A L, KATZ R, GOSTIN L O. The novel coronavirus originating in Wuhan, China: challenges for global health governance[J]. JAMA, 2020, 323(8): 709-710.
[3] ZHAO W, ZHONG Z, XIE X, et al. CT scans of patients with 2019 novel coronavirus (COVID-19) pneumonia[J]. Theranostics, 2020, 10(10): 4606-4613.
[4] YOON S H, LEE K H, KIM J Y, et al. Chest radiographic and CT findings of the 2019 novel coronavirus disease (COVID-19): analysis of nine patients treated in Korea[J]. Korean J Radiol, 2020, 21(4): 494-500.
[5] LEI J, LI J, LI X, et al. CT imaging of the 2019 novel coronavirus (2019-nCoV) pneumonia[J]. Radiology, 2020, 295(1): 18.
[6] FRANQUET T. Imaging of pulmonary viral pneumonia[J]. Radiology, 2011, 260(1): 18-39.
[7] XIE X, ZHONG Z, ZHAO W, et al. Chest CT for typical 2019-nCoV pneumonia: relationship to negative RT-PCR testing[J]. Radiology, 2020, 296(2): E41-E45.
[8] FANG Y, ZHANG H, XIE J, et al. Sensitivity of chest CT for COVID-19: comparison to RT-PCR[J]. Radiology, 2020, 296(2): E115-E117.
[9] 管汉雄, 熊颖, 申楠茜, 等. 新型冠状病毒肺炎(COVID-19)临床影像学特征[J]. 放射学实践, 2020, 35(2): 125-130.
GUAN H X, XIONG Y, SHEN N X, et al. Novel coronavirus pneumonia (COVID-19) clinical imaging features[J]. Radiologic Practice, 2020, 35(2): 125-130.
[10] BERNHEIM A, MEI X, HUANG M, et al. Chest CT findings in coronavirus disease-19 (COVID-10): relationship to duration of infection[J]. Radiology, 2020, 295(3): 200463.
[11] PAN F, YE T, SUN P, et al. Time course of lung changes of chest CT during recovery from 2019 novel coronavirus (COVID-19) pneumonia[J]. Radiology, 2020, 295(3): 715-721.
[12] ZU Z Y, JIANG M D, XU P P, et al. Coronavirus disease 2019 (COVID-19): A perspective from China[J]. Radiology, 2020, 296 (2): E15-E25.
[13] LIU F, ZHANG Q, HUANG C, et al. CT quantification of pneumonia lesions in early days predicts progression to severe illness in a cohort of COVID-19 patients[J]. Theranostics, 2020, 10(12): 5613-5622.
[14] LI L, QIN L, XU Z, et al. Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT[J]. Radiology, 2020, 296(2): 200905.

（编辑：黄开颜）