文章编号 1006-8147(2018)06-0480-04

论 著

放射组学在肺癌诊断中的应用

方胜儒,李逸凡,张宇威,蔡 娜,郭 丽
(天津医科大学医学影像学院,天津 300203)

摘要 **目的**:通过放射组学对肺癌病例进行定量特征提取,优化选择,然后通过机器学习方法实现肺癌病例讨论和分析。**方法**:通过公开数据库LIDC中提取224例和医院收集250例肺结节病例,提取共841个放射组学特征;对特征进行正态分析和方差齐性分析,双独立样本t检验进行降维;其余采用秩和分析降维,之后采取Pearson相关系数降维,最后通过机器学习方法进行分类。**结果**:来自LIDC数据库和来自医院的数据在基于随机森林的分类器中的结果分别为AUC=0.657 1、ACC=76.26%,AUC=0.866 7、ACC=76%;在基于支持向量机的分类器中的结果分别为AUC=0.642 9,ACC=76.37%,AUC=0.773 3、ACC=72%。**结论**:在肺癌良恶诊断鉴别中,使用放射组学特征方法可以鉴别良恶性。基于纹理特征的计算机辅助诊断系统可以提高对此类结节的诊断效能。

关键词 计算机辅助诊断技术;肺结节;放射组学;纹理特征
中图分类号 R816.41 **文献标志码** A

The application of radiomics in the diagnosis of lung cancer

FANG Sheng-ru, LI Yi-fan,ZHANG Yu-wei,CAI Na, GUO Li
(School of Medical Imaging, Tianjin Medical University, Tianjin 300203, China)

Abstract **Objective:** To quantitatively extract and optimizeradiomicsfeatures for lung cancer cases and to analyze and discuss lung cancer cases by machine learning method. **Methods:**We obtained images of 224 patients from LIDC databaseand 250 patients from hospital, and 841 radiomicsfeatures were extracted. The features were used to perform dimensionality reduction by the double independent sample t-test, when normality of distribution and Homogeneity variance were calculated. Futhermore, the dimensionality reduction was performed by the rank sum test. And then, the Pearson correlation coefficient was used for further dimensionality reduction.Finally, machine learning method was used for classification. **Results:** In the classifier based on the random forest, the LIDC database showed that ACC=76.26%, AUC=0.657 1 and the data from the Hospital showed that ACC=76%, AUC=0.866 7. In the classifier based on the support vector machine, the LIDC database showed that ACC=76.37%, AUC=0.642 9, and the data from the hospital showed that ACC=72%, AUC=0.773 3. **Conclusion:** In pulmonary nodules, radiomics can be used to identify benign and malignant nodules. Texture-based computer-aided diagnosis systems may improve the diagnostic efficacy on pulmonary nodules.

Key words computer-aided diagnosis; pulmonary nodules; radiomics; texture feature

肺癌是当今世界最常见的恶性肿瘤之一,也是对人类威胁最大的肿瘤性疾病。如果在肿瘤的早期生长阶段即对其进行检查与治疗,就能获得更高的治愈率。因此为提高肺癌病人的生存率,早期筛查成为近年来的热点。在精准医疗的大背景下,放射组学(Radiomics)应运而生。利用多学科的知识对医学影像数据进行分析,应用大量的人工智能提出的数据特征化算法将感兴趣区域的影像数据转化为具有高维度的可发掘的特征空间数据。通过机器学习等高级数据挖掘算法进行大数据处理,对大量的影像数据进行数字化定量定性分析,得到分类模型来综合评价肿瘤的各种分期分型,以达到早期诊断、指导治疗和预测预后的目的[1-3]。

目前,国内外已有很多研究学者对肺部肿瘤进行分析,通过特征提取的方法来进行专家系统的肺部肿瘤的影像诊断,并已有专家能够提取出肺部肿瘤的相关特征信息,为肿瘤诊断提供数据[4-6]。在特征提取阶段,通过数学描述提供肺部病变区域的特征,包括大小、形状、纹理、强度、边缘和其他方面的特征。Gillies在放射组学研究上取得了重要的成果[7-10]。Gillies研究组提取了更多的特性信息,包括肺结节以及病变周围肺组织形成的微环境的大小、灰度值、形状、边缘和纹理(灰度共生、游程长度、小波、Law's特征等)。同样提取了特征的2D和3D数据,并提供了可重复性的测试数据,得出了这些特征对预测恶性肿瘤,疾病进程指标和基因具有相关性的

基金项目 国家自然科学基金青年基金资助项目(81000639),天津医科大学校基金资助项目(2015KYZQ19)
作者简介 方胜儒(1987-),男,硕士在读,研究方向:生物医学工程;通信作者:郭丽,E-mail:gl6290@126.com。

结论[11]。由于纹理特征是细微的特征，在研究过程中是否考虑到采集设备硬件以及图像重建方法，对放射组学纹理特征分析结果有变异的影响。

为临床肺癌诊断技术提供具有参考价值的数据采集工作流程。Kalpathy-Cramer 给出了相关特征介绍，并分析了定量图像特征对肿瘤分割的敏感性以及通过不同特征提取方法计算特征之间的相关性。通过对不同病人之间的类似特征研究其相关性，并对所有病人不相关特征之间相关性的研究，得出了每个独立的特征有很多具有较高的相关性和相同性的结论。目前已经找出了特征内部和特征之间的相应关系，同时也发现了很多特征有一定的相关性，出现冗余特征的情况。

1　资料和方法

1.1　*资料*　本研究所使用的数据来自两个部分，一个是美国国家癌症研究所（National Cancer Institute，NCI）发起的大型公开数据集—肺部图像影像数据库（The LungImage Database Consortium，LIDC），包含从 7 个学术中心和 8 家医学影像公司采集到的 1 018 例患者的肺部 CT 扫描成像结果（扫描层厚 1.25~3 mm，512×512 像素）[12-13]。笔者从中挑选 224 例含有分析结果的数据；另一个是由从医院 CT 检查发现的肺部扫描的数据 250 例。所有的肺部病例的提取分割分析分析都是使用基于 matlab2017b 进行的，具体肺结节分割，肺结节感兴趣区域提取，所有的特征信息提取，均使用 matlab 程序函数编写。

1.2　*方法*

1.2.1　ROI 区域的提取　本实验用到 LIDC 数据集，这个数据集中每个 CT 扫描都有 4 位放射科医生读片评注，医生同时也标注了肺结节轮廓的坐标点，该部分的肺结节分割使用数据库自带的分割数据进行分析。医院的数据是经过 3 名专业的放射科主治医师的筛选分析的分割结果，在提取 ROI 区域进行纹理分析的时候采用分割内部实质的区域进行分析，对医生勾画的区域进行了缩小操作。所有的 ROI 区域的提取使用 matlab2017b 实现。在分割的同时也提取肺结节的形状、大小、边缘毛刺程等形态学信息。

笔者采用逐结节逐层分割的策略，将分割后的感兴趣区整合为一个体积感兴趣区（Volume Of Interest, VOI），VOI 由每层 CT 图像上的感兴趣区根据层面次序依次堆叠而成。如图 1 所示，图中有四种颜色的分割结果，笔者以 4 种颜色全部包含的区域提取 ROI，并做处理。提取体积感兴趣区并做 mask 模板为后面提取特征做前期准备。图 2 给出了提取的一部分肺结节 VOI，mask 模板的图像和对应的信息。

1.2.2　特征信息的提取　特征提取是放射组学分析的基础。针对肺结节的特点，笔者设计了 5 组共 62个放射组学特征构成每个样本的特征空间，特征提取算法的代码全部基于 Matlab R2017b 实现。首先，笔者提取 VOI 的灰度直方图的一阶统计特征，共 14个。此组特征分别对灰度分布、全 VOI 灰度特点、灰度分布的一致性等特点进行描述。其次，形态学的特征作为前期医生诊断的重要依据。本组特征描述结节的 3D 最大长径、基于像素的体积值、基于表面像素的表面积值、圆度、紧密度等指标、维模型分形维数、相关维数、几何学测量特征等特征。最后，笔者提取纹理特征，共 841 个。这一组特征使用灰度共生矩阵（Gray Co-occurrence Matrix, GLCM）算法、灰度游程长矩阵（Gray Level Run-Length Matrix, GLRLM）的方法。其中 GLCM 特征 3D 特征 247 个，

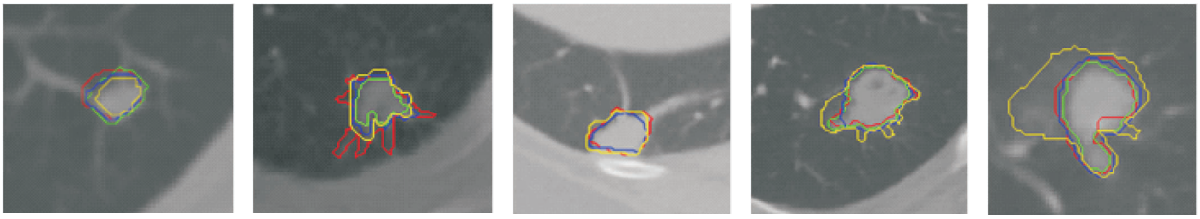

图 1　LIDC 数据库的肺结节分割图像

Fig 1　Segmentation images of lung nodule in LIDC data

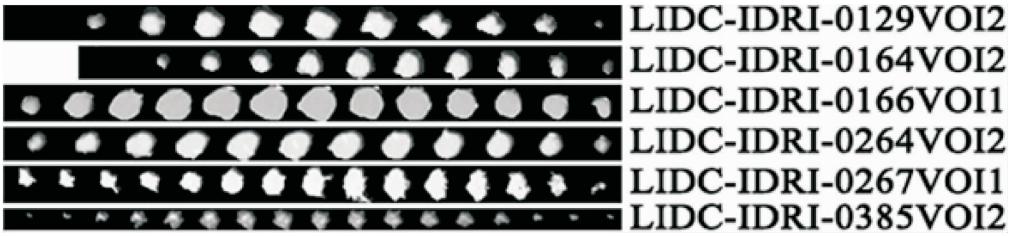

图 2　获取 VOI 区域的 mask 模板图像

Fig 2　The mask template images for the VOI region

和 GLRLM 特征 2D 特征 55 个。这两个特征都是基于二阶统计的特征描述子。肺结节在 CT 表现上有着肉眼可见的纹理，因此 GLCM、GLRLM 特征在描述结节内部的纹理特点时有着独特的优势。Laws 特征 482 个，Laws 纹理特征是一种典型的基于模板卷积的纹理描述特征，测量单个像素和邻域灰度分布统计分析方法。LoG 特征 27 个。使用高斯滤波将孤立的噪声点和较小的结构组织滤除，然后利用无方向性的拉普拉斯算子实现。多尺度 3D 小波分解（Multilevel 3D Wavelet Decomposition at Level）特征 16 个。借助正交小波对图像进行小波分解，得到不同分辨率的一系列图像。均为三维特征算法，笔者采用对 13 个方向取均值的方法来处理。

1.2.3 特征降维与分类 对于所有特征进行正态分析和方差齐性分析，通过分析的特征采用双独立样本 t 检验进行降维；其余采用秩和分析进行降维，之后采取 Pearson 相关系数进一步降维。然后，对不同维度的特征通过支持向量机和随机森林的分类器在不同的 Pearson 相关性系数阈值的条件下进行了讨论，并根据最佳的 Pearson 相关系数建立了预测模型。最后，通过 10 折检验选择最佳模型，并在检验集中对其临床分类效果进行了预测。

2 结果

对于所有提取到的放射组学特征，我们需要对其进行统计学差异分析。首先，需要分别对于良性数据与恶性数据单独进行正态性分析。我们对所有的放射组学特征进行了 Lilliefors 正态检验，良性结节与恶性结节的检验结果 P 值（P <0.05）的特征，共发现了 57 个特征通过正态分布检验，再对其进行齐性方差剩余 49 个特征，在良性结节和恶性结节中 30 个特征表现出了统计学差异。其余不符合正态分布的放射组学特征需要进行秩和检验。对于这 57 个符合正态分布的放射组学特征的 Hrtley 方差齐性分析 P 值（P<0.05）。通过秩和检验我们得到了 427 个特征。综上，笔者通过统计分析一共获得了 457 特征，各个特征的残留率如表 1 所示。

在对数据进行秩和检验与双独立样本 t 检验之后，我们对于筛选出来的数据进行皮尔逊相关系数检验以对放射组学特征进行进一步的降维。我们根据由不同的阈值分类得出的放射组学特征进行了 50 次分类器训练，并对其准确度进行了分析。由此笔者选取了 0.14 为皮尔逊相关系数的阈值以筛选在良恶性肺结节中相关性极弱的放射组学特征。

笔者将以上的放射组学特征又区分为二维放射组学特征和三维放射组学特征。其中二维放射组学特征包括了一维放射组学特征、基本形状大小特征、二维灰度游程矩阵（GLRL-2D）、Laws 图像纹理特征（Law-Textures）、LoG 二阶边缘信息特征；三维放射组学特征包括了三维灰度共生矩阵、三维灰度区域大小矩阵（GLSZM-3D）、多尺度三维小波特征；而这些特征合称混合放射组学特征。我们通过二维放射组学特征、三维放射组学特征、混合放射组学特征对于基于随机森林的肺结节良恶性分类器进行了分析。如图 3 所示，在 3 种特征分析中混合特征的识别精度要比其他两个高。

同样还分析了不同数据库利用混合特征进行分类的结果讨论。绘制了对于 LIDC 数据的基于支持向量机的肺结节良恶性分类器和基于随机森林的肺结节良恶性分类器的处理结果，以及肿瘤医院数据的基于支持向量机的肺结节良恶性分类器和基于随机森林的肺结节良恶性分类器的处理结果的 ROC 曲线图（图 4）。

图 4 中，来自 LIDC 数据库的数据的基于随机森林的肺结节良恶性分类器的，其中 AUC（Area Under Curve）被定义为 ROC（Receiver Operating Characteristic）曲线下的面积，ACC（Accuracy）为准确率。AUC=0.657 1、ACC=76.26%，基于支持向量机的肺结节良恶性分类器的 AUC=0.642 9、ACC=76.37%；来自肿瘤医院的数据的基于随机森林的肺结节良恶性分类器的 AUC=0.866 7、ACC=76%，基于支持向量机的肺结节良恶性分类器的 AUC=0.773 3、ACC=72%。由此可以发现，来自 LIDC 数据库的数据其分类准确度较高但是其 AUC 较低，而来自肿瘤医院的数据则正好与之相反。根据反复试验的数

表 1 特征残留率

Tab 1 Feature residual rates

	stat1	ss	glcm	glrl	glszm	laws	LoG	wav
正态分布检验	0	0.1666	0.1012	0.0909	0	0.01	0.4074	0.625
方差齐性检验	0	0.0833	0.0971	0.0909	0	0.01	0.2592	0.5
双独立 t 检验	0	0	0.0728	0	0	0.008	0.2222	0.125
秩和检验	0.7857	0.5833	0.4979	0.2727	0.3636	0.51	0.4814	0.125
总体残留	0.7857	0.5833	0.5708	0.2727	0.3636	0.518	0.7037	0.25
Pearson 系数	0.7142	0.5833	0.5222	0.2727	0.1818	0.026	0.7037	0.25

据证明，来自 LIDC 的数据准确度相比于来自与肿瘤医院的数据的准确度高约 3%。由此我们推测不同来源的肺结节数据对于分类器的建立有一定的影响。

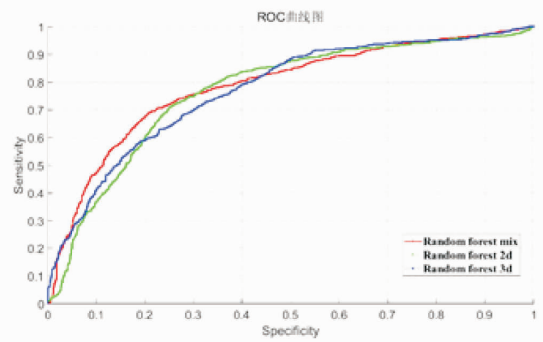

图 3　基于随机森林中三种特征分类的 ROC 曲线

Fig 3　ROC curve based on three feature in random forest

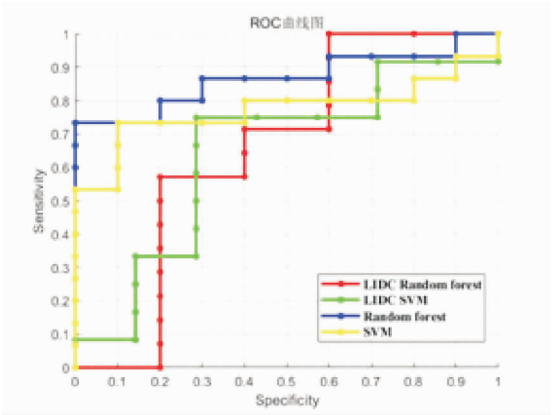

图 4　肿瘤医院数据与 LIDC 数据库 ROC 比较

Fig 4　Comparison of hospital data with ROC of LIDC database

3　讨论

笔者分析二维、三维和混合特征的放射组学信息特征的差异，采用 SVM 和随机森林两种分类方法的差异。从结果中得出混合放射组学特征的分类准确度相对于二维放射组学特征的分类准确度略微有一定优势，且这两者对于肺结节的良恶性区分能力高于三维放射组学特征的分类准确度。混合放射组学特征的数量大于三维放射组学特征，而三维放射组学特征数量也大于二维放射组学特征。总体上，混合数据特征仍优于二维特征和三维特征。所以在今后的处理过程中，我们需要发现更适合病例的特征进行分析，会大大提高分类的准确度。

基于随机森林的肺结节良恶性分类器中相比于基于 SVM 的肺结节良恶性分类器而言，其在两者共同的最优阈值 (Pearson correlation coefficient= 0.14)处具有更高的分类准确度，且在总体表现上也优于后者。在最优阈值之前，两类分类器在相同的放射组学特征数量时分类能力互有高低，但是在最优阈值处以及之后，在相同的放射组学特征数量的情况之下，基于随机森林的肺结节良恶性分类器的分类准确度明显优于基于支持向量机的肺结节良恶性分类器。根据 ROC 曲线所示，基于支持向量机的肺结节良恶性分类器曲线下面积（AUC=0.866 7），而基于随机森林的肺结节良恶性分类器曲线下面积（AUC=0.773 3）。由此结果显示，基于随机森林的肺结节良恶性分类器相对于基于支持向量机的肺结节良恶性分类器具有更好的分类效果。

参考文献：

[1] 郑钧正.从基因组学到放射组学的启示[J].医学研究杂志,2016,45(2):1

[2] 苏会芳,周国锋,谢传森,等.放射组学的兴起和研究进展[J].中华医学杂志, 2015, 95(7):553

[3] Liang C S, Huang Y Q, He L, et al. The development and validation of a CT-based radiomics signature for the preoperative discrimination of stage I-II and stage III-IV colorectal cancer[J].Oncotarget, 2016, 7(21):31401

[4] Larue R T H M, Defraene G, De Ruysscher D, et al. Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures[J]. Br J Radiol, 2017, 90(1070): 20160665

[5] P aul D, Su R, RomainM, et al. Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier[J].Comput Med Imaging Graph,2017,60(9):42

[6] Gutenko I, Dmitriev K, Kaufman A E, et al. AnaFe:visual analytics of image derived temporal features focusing on the spleen[J].IEEE Trans Vis Comput Graph,2017,23(1):171

[7] Liu Y, Kim J, Balagurunathan Y, et al.Radiomic features are associated with EGFR mutation status in lung adenocarcinomas[J]. Clin Lung Cancer,2016,17(5):441

[8] Zhou M, Chaudhury B, Hall L O, et al. Identifying spatial imaging biomarkers of glioblastomamultiforme for survival group prediction[J]. J Magn Reson Imaging, 2016,46(1):115

[9] Gillies R J, Kinahan P E, Hricak H. Radiomics: images are more than pictures, they are data[J].Radiology,2016,278(2):563

[10] Hawkins S, Wang H, Liu Y, et al. Predicting malignant nodules from screening CT scans[J]. J Thorac Oncol, 2016, 11(12):2120

[11] Shen C, Liu Z, Guan M, et al. 2D and 3D CT radiomicsfeatures prognostic performance comparison in non-small cell lung cancer[J]. Transl Oncol, 2017, 10(6):886

[12] 3Rd A S, Mclennan G, Bidaut L, et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans[J]. Med Phys, 2012, 38(2):915

[13] Clark K, Vendt B, Smith K, et al. The cancer imaging archive (TCIA): maintaining and operating a public information repository[J]. J Digital Imaging, 2013, 26(6):1045

（2018-05-22 收稿）