·论著·

RF 和 C4.5 决策树在食管癌图像分类中的研究

茹仙古丽·艾尔西丁[1],木拉提·哈米提[2],严传波[2],姚 娟[3]

(1.新疆医科大学基础医学院,新疆 乌鲁木齐 830011;
2.新疆医科大学医学工程技术学院,新疆 乌鲁木齐 830011;
3.新疆医科大学第一附属医院放射科,新疆 乌鲁木齐 830054)

摘 要:目的 探讨 RF 和 C4.5 决策树对 X 线食管造影图像分型中的应用,以及验证分类器对特征的分类能力。**方法** 选取 2018 年 1 月~6 月在新疆医科大学第一附属医院、第二附属医院和第三附属(肿瘤)医院的放射科选取溃疡性、缩窄型和蕈伞型食管癌 X 线图像各 560 张,提取灰度共生矩阵,灰度直方图和混合特征;采用 RF 和 C4.5 决策树通过调整参数进行分类研究。**结果** RF 和 C4.5 决策树对溃疡型和缩窄型食管癌进行分类,灰度共生矩阵的分类准确率分别为 73.30%,67.76%;灰度直方图分类准确率分别为 84.55%,76.16%。而混合特征算法的分类准确率分别为 95.08%,86.87%;对溃疡型和蕈伞型食管癌进行分类,灰度共生矩阵的分类准确率分别为 75.08%,66.96%;灰度直方图分类准确率分别为 83.83%,77.23%。而混合特征算法的分类准确率分别为 80.98%,73.66%。**结论** 灰度直方图特征的分类准确率比灰度共生矩阵特征的平均高 10%,混合特征更适合于溃疡型,缩窄型食管癌的分类。而灰度直方图特征更适合于溃疡型,蕈伞型食管癌的分类;RF 的分类能力比 C4.5 决策树高。此算法可为 X 线食管造影图像的分类提供参考。

关键词:食管癌;随机森林;C4.5 决策树;特征提取

中图分类号:R735.1;TP391.4　　**文献标识码:**A　　DOI:10.3969/j.issn.1006-1959.2018.22.015

文章编号:1006-1959(2018)22-0051-05

Research on RF and C4.5 Decision Tree in Image Classification of Esophageal Cancer

Roxangvl·Arxidin[1],Murat·Hamit[2],YAN Chuan-bo[2],YAO Juan[3]

(1.Basic Medical College,Xinjiang Medical University,Urumqi 830011,Xinjiang,China;
2.College of Medical Engineering Technology,Xinjiang Medical University,Urumqi 830011,Xinjiang,China;
3.Department of Radiology,the First Affiliated Hospital,Xinjiang Medical University,Urumqi 830054,Xinjiang,China)

Abstract:Objective To explore the application of RF and C4.5 decision tree to the classification of X-ray esophageal images and to verify the classifier's ability to classify texture features.**Methods** From January to June 2018, the radiologists of the first affiliated Hospital, the second affiliated Hospital and the third affiliated (tumor) Hospital of Xinjiang Medical University selected 560 X-ray images of ulcerative, constrictive and mushroom esophageal cancer to extract the gray level symbiosis matrix. Grayscale histogram and mixed feature; RF and C4.5 decision tree are used to study the classification by adjusting the parameters.**Results** RF and C4.5 decision tree were used to classify ulcerative and constricted esophageal cancer. The classification accuracy of gray co-occurrence matrix was 73.30% and 67.76%.The classification accuracy of gray histogram was 84.55% and 76.16%,respectively.The classification accuracy of comprehensive feature algorithm was 95.08% and 86.87%, the classification accuracy of ulcerative and mushroom esophageal cancer was 75.08% and 66.96%, respectively, and the classification accuracy of gray histogram was 83.83% and 77.23%, respectively. The classification accuracy of comprehensive feature algorithm was 80.98% and 73.66%,respectively.**Conclusion** The classification accuracy of grayscale histogram is 10% higher than that of gray level co-occurrence matrix. The comprehensive feature is more suitable for classification of ulcerative and constrictive esophageal cancer. The gray histogram features are more suitable for the classification of ulcerative and mushroom esophageal cancer, and the classification ability of RF is higher than that of C4.5 decision tree. This algorithm can provide reference for the classification of X-ray esophageal images.

Key words: Esophageal cancer;Random forest;C4.5 decision tree;Feature extraction

癌症是严重危害人类健康的慢性疾病,也是威胁生命的主要杀手。其中食管癌是对癌症患者生存质量(quality of life,QOL)影响最大的疾病之一[1]。新疆哈萨克族是食管癌的高发民族,其食管癌死亡率达 155.9/106,高于我国平均水平 15.23/106,是本地区重点防治的恶性肿瘤[2]。随着科学技术的发展和医学影像应用的推广,越来越多的医学图像需要医生解读[3,4]。由于食管癌的早期临床特征不明显,医

基金项目:国家自然科学基金(编号:81460281,81560294,81760330)

作者简介:茹仙古丽·艾尔西丁(1992.10-),女,新疆乌鲁木齐人,硕士研究生,研究方向:医学图像处理及信号分析

通讯作者:木拉提·哈米提(1957.6-),男,新疆乌鲁木齐人,本科,教授,生物医学工程研究所所长,研究方向:医学图像处理及信号分析

生也有可能会因为经验不足或疲劳而产生解读错误，使疾病漏诊[5]。医学影像疾病误诊率可达到 10%~30%[6]。计算机辅助诊断（computer-aided diagnosis，CAD）技术的出现为医生提供有效的诊断决策支持。分析整幅图像时不仅存在大量冗余信息，而且容易降低准确率。所以，将整幅图像缩小到若干小的病灶区域（ROI），然后对图像进行预处理，最后提取特征，这样可以提高计算机处理速度和分类准确率。本研究利用灰度共生矩阵和灰度直方图提取食管图像的特征，构造 RF 和 C4.5 决策树分类器实现图像的分类，以及验证分类器对特征的分类能力。

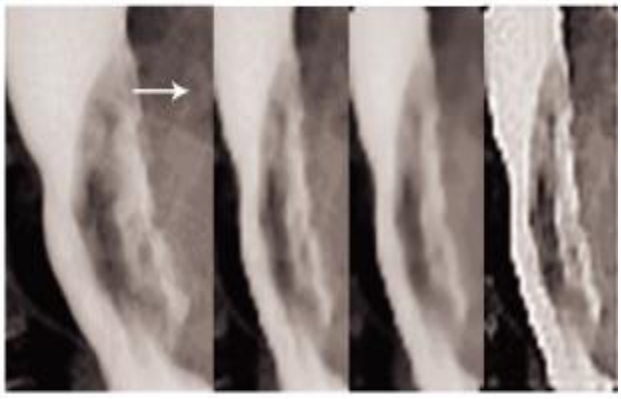
1 资料与方法

1.1 研究对象 选取 2018 年 1 月~6 月在新疆医科大学第一附属医院、第二附属医院和第三附属（肿瘤）医院放射科选取溃疡性、缩窄型和蕈伞型食管癌 X 线图像各 560 张，在临床医师的指导下，人工干预的方式分割出病灶信息区域，并进行归类，病灶区域的提取结果见图 1。

A 溃疡型食管癌 B 缩窄型食管癌 C 蕈伞型食管癌

图 1 感兴趣区域的选择

1.2 图像处理 在医学 X 射线图像放射成像的过程中，由于人体组织结构的复杂性及成像系统的 X 射线散射、电器噪声等不利因素的影响，会导致图像质量下降。最主要表现为对比度差、细节模糊，影响了医生的诊断及分析，为了得到更清晰的图像对图像进行预处理。本研究在临床医师的指导下，首先对图像进行病灶区域分割，并进行归一化。然后用中值滤波去噪[7-10]，中值滤波是一种非线性滤波，适用于滤除脉冲噪声或颗粒噪声，并能保护图像边缘。但是去噪处理之后，出现图像的边缘和轮廓模糊的情况。为了减少这类不利效果的影响，就需要利用高通滤波器对图像进行锐化增强[11,12]，目的是为了使图像的边缘、轮廓线以及图像的细节变得清晰。预处理后的结果见图 2~图 4。



注：从左向右依次为：原始 ROI 图像、归一化后的图像、去噪后的图像、锐化增强后的图像

图 2 溃疡型食管癌

1.3 特征提取

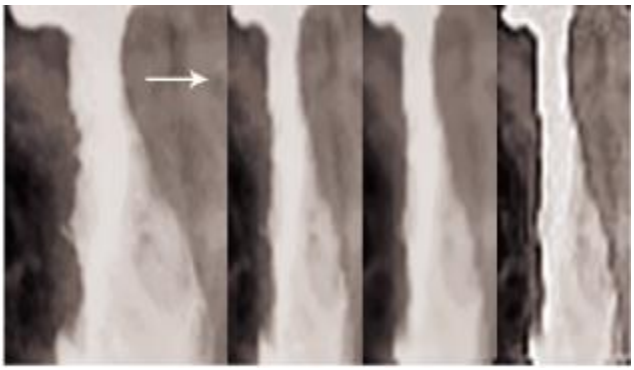
1.3.1 灰度共生矩阵特征 灰度共生矩阵[13,14]将图像中像素及其邻域像素的空间关系和灰度关系结合起来，充分体现了一定空间关系下图像的灰度变化情况，从而达到分析图像纹理特征的目的。首先，对图像进行二层小波分解提取图像的低频信息；然后使用灰度共生矩阵法提取低频信息的特征，取像素距离 d=1，θ={0°，45°，90°，135°}4 个方向，对每个方向上的共生矩阵提取以下 Q_1~Q_4 的特征值。因此，每种纹理形成了能反映自身特征的一组包含 16 个元素的特征向量。

①纹理能量：

$$Q_1=\sum_{g_1}\sum_{g_1}[p(g_1,g_2)]^2 \quad (1)$$

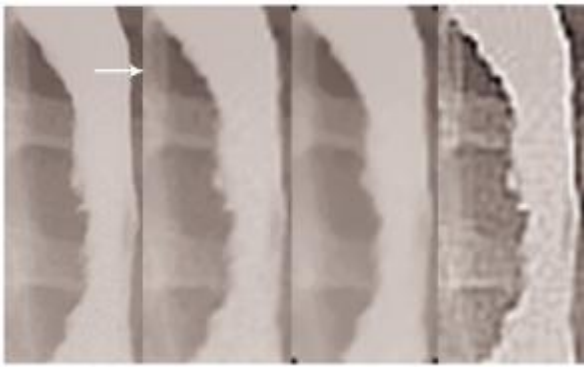
②纹理惯性：

$$Q_2=\sum_{g_1}\sum_{g_2}K^2 p(g_1,g_2)^2 \cdot K=|g_1-g_2| \quad (2)$$



注：从左向右依次为原始 ROI 图像、归一化后的图像、去噪后的图像、锐化增强后的图像

图 3　缩窄型食管癌



注：从左向右依次为原始 ROI 图像、归一化后的图像、去噪后的图像、锐化增强后的图像

图 4　蕈伞型食管癌

③纹理相关性：

$$Q_1=\frac{\sum_{g_1}\sum_{g_2} g_1 g_2 p(g_1-g_2)-\mu_x\mu_y}{\delta_x\delta_y} \quad (3)$$

④纹理熵：

$$Q_4=-\sum_{g_1}\sum_{g_2} p(g_1,g_2)\lg p(g_1,g_2) \quad (4)$$

其中：

$$\mu_x=\sum_{g_1} g_1 \sum_{g_2} p(g_1,g_2),\ \mu_y=\sum_{g_1} g_2 \sum_{g_2} p(g_1,g_2)$$

$$\delta_x^2=\sum_{g_1}(g_1-\mu_x)^2\sum_{g_2} p(g_1,g_2),\ \delta_y^2=\sum_{g_2}(g_2-\mu_y)^2\sum_{g_1} p(g_1,g_2)$$

1.3.2 灰度直方图特征　灰度直方图是个灰度级的离散函数，可以用式(5)表示图像灰度直方图的定义[15,16]。

$$H(i)=\frac{n_i}{N},\ i=0,1,2,\cdots\cdots L-1 \quad (5)$$

注：i 表示灰度级，L 表示灰度级种类数，n_i 表示图像中具有灰度级 i 的像素的个数，N 表示图像总的像素数

式(5)描述的是图像中具有该灰度级的像素的个数占图像总像素的百分比，即图像中具有灰度级 i 的像素出现的频率。本文仅讨论实验中提取的以下几种描述能力较强的参数作为特征统计量。

①均值(mean)：反映一幅图像的平均灰度值。

$$\mu=\sum_{i=0}^{L-1} iH(i) \quad (6)$$

②方差(variance)：反映一幅图像的灰度在数值上的离散分布情况。

$$\sigma^2=\sum_{i=0}^{L-1}(i-\mu)^2 H(i) \quad (7)$$

③歪斜度(skewness)：反映图像直方图分布的不对称程度。歪斜度越大表示直方图分布越不对称，反之越对称。

$$\mu_x=\frac{1}{\sigma^3}\sum_{i=0}^{L-1}(i-\mu)^3 H(i) \quad (8)$$

④峰态(kurtosis)：反映图像的灰度分布在接近均值时的大致状态，用以判断图像的灰度分布是否非常集中于平均灰度附近。峰态越小，表示越集中；反之表示越分散。

$$\mu_k=\frac{1}{\sigma^4}\sum_{i=0}^{L-1}(i-\mu)^4 H(i)-3 \quad (9)$$

⑤能量(energy)：反映灰度分布的均匀程度，灰度分布较均匀时能量较大，反之较小。

$$\mu_{en}=\sum_{i=0}^{L-1}[H(i)]^2 \quad (10)$$

2 图像分类

2.1 RF 分类算法　RF 算法[17,18]是由 Brieman 在 2001 年提出的一个集成学习算法框架。训练时，通过 Bagging 方法随机抽取样本集和特征集训练不同的决策树；分类时，每棵树对类别进行“投票”决定最终分类结果[19]。随机森林特征选择过程是通过迭代生成随机森林，每轮迭代后对特征重要性进行排序，剔除不重要的特征，直至符合结束条件。

2.2 C4.5 决策树分类算法　C4.5 决策树算法[20-22]是判断给定样本与某种属性相关联的决策过程的一种表示方法，从数据中生成分类器的一个特别有效的方法是生成一棵决策树，该方法广泛应用于数据挖掘和机器学习等领域，用来解决与分类相关的问题。决策树表示法是应用最广泛的逻辑方法。目前生成决策树方法的算法主要有三种：CART 算法，ID3 算法，C4.5 算法。其中 C4.5 算法具有分类速度快且精度高的特点，是发展得比较完善的一种决策树算法[23]。

3 结果

本研究构造 RF 分类器和 C4.5 决策树分类器，采用十折交叉验证法。在实验过程中，调整 RF 分类器的数字特征（0≤n≤2）和 C4.5 决策树分类器的参数 c（0.1≤c≤5）对提取的特征进行分类。分类结果见图 5。RF 和 C4.5 决策树分类器应用于两种分类方式，分类准确率，见表 1。

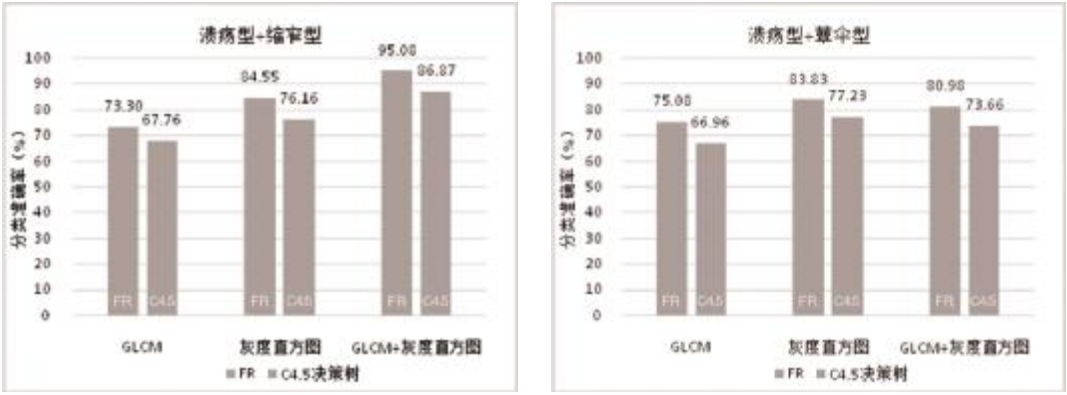


图 5 分类结果

表 1 分类准确率

分类	GLCM				灰度直方图				GLCM+灰度直方图			
	溃疡型	缩窄型	溃疡型	蕈伞型	溃疡型	缩窄型	溃疡型	蕈伞型	溃疡型	缩窄型	溃疡型	蕈伞型
RF	73.40	75.20	77.30	73.20	85.40	83.80	85.40	82.30	96.30	93.90	82.90	79.10
C4.5 决策树	65.50	70.00	67.50	66.80	73.80	78.60	79.10	75.40	88.20	85.50	74.80	72.50

4 讨论

本次研究显示：①用溃疡型和缩窄型两种食管癌进行分类，RF（n=2）和 C4.5 决策树（c=0.35）对灰度共生矩阵特征的分类准确率分别为 73.30% 和 67.76%；RF（n=2）和 C4.5 决策树（c=0.4）对灰度直方图特征的分类准确率分别为 84.55% 和 76.16%；RF（n=0）和 C4.5 决策树（c=0.3）对混合特征的分类准确率分别为 95.08% 和 86.87%。RF 分类器对三种特征的分类能力比 C4.5 决策树分类器的高，对混合特征的分类准确率最好。②用溃疡型和蕈伞型两种食管癌进行分类，RF（n=0）和 C4.5 决策树（c=0.3）对灰度共生矩阵特征的分类准确率分别为 75.08% 和 66.96%；RF（n=2）和 C4.5 决策树（c=0.25）对灰度直方图特征的分类准确率分别为 83.83% 和 77.23%；RF（n=0）和 C4.5 决策树（c=0.2）对混合特征的分类准确率分别为 80.98% 和 73.66%。RF 分类器对灰度直方图特征的分类效果比灰度共生矩阵特征的高 9.51%，比混合特征的高 3.21%；灰度直方图特征更适合于这两种食管癌的准确分类。

本研究结果显示，两种分类器适用于两种分类方式 RF 分类器的分类效果比 C4.5 决策树分类器的高；这可能是由于 C4.5 决策树虽然是一种简单且快速的非参数分类方法，还具有很好的准确率。然而当数据复杂或者存在噪声时，易出现过拟合问题，使得分类准确率下降。随机森林是以决策树为基本分类器的一个集成学习模型，它克服了决策树过拟合问题，对噪声和异常值有较好的容忍性，对高维数据分类问题具有良好的可扩展性和并行性。总的分类效果来讲，溃疡型食管癌的分类效果最高，这表明溃疡性食管癌与缩窄型或蕈伞型食管癌在灰度共生和灰度直方图特征上有很大的差异。这可能是由于溃疡性食管癌的病灶区存在长条状溃疡造成图像的灰度改变所致。

食管癌是常见的消化道恶性肿瘤，新疆哈萨克族人群的食管癌发病率与其他民族相比居首位。本研究选取溃疡型、缩窄型和蕈伞型 X 线食管造影图像为研究对象，使用基于灰度共生矩阵，灰度直方图和混合的特征提取方法，通过构造 RF 和 C4.5 决策树分类器对特征的分类能力进行验证。结果表明，灰度直方图特征的分类效率优于灰度共生矩阵特征；RF 与 C4.5 决策树分类器分类能力进行比较，RF 的分类能力较佳，更适合于本研究所使用的研究对象进行分类。这将新疆哈萨克族 X 线食管造影图像的分类提供一种新的思路和参考。

参考文献：

[1]徐悦洋，卫莉，杨长永.癌症患者自我感受负担现状的研究进展[J].中国实用护理杂志，2018，34(13)：1032-1035.

（下转第 63 页）

(上接第 54 页)

[2]杨芳,木拉提·哈米提,严传波,等.PCA 和 SVM 在新疆哈萨克族食管癌图像分类中的研究与应用[J].科技通报,2016,32(3):53-57.

[3]Martin Spahn.X-ray detectors in medical imaging[J].Nuclear Instruments and Methods in Physics Research Section A Accelerators Spectrometers Detectors and Associated Equipment,2013,731(5):57-63.

[4]Setyowati E,Suparta GB,Poedjomartono B.Phantom image dimension analysis on computed tomography image [J].AIP Conference Proceedings,2016,1755(1):1-4.

[5]房俊飞,李文武,刘桂芬,等.腹段食管癌漏诊的常见原因及预防措施[J].肿瘤研究与临床,2002,14(4):281.

[6]顾晴,熊长明,柳志红,等.结缔组织病合并肺栓塞的临床特征及误诊原因分析[J].中华医学杂志,2015,95(2):120-122.

[7]黄梦涛,胡永才.改进自适应中值滤波的低照度烟雾图像去噪[J].计算机工程与设计,2018,39(6):1659-1663.

[8]张娟.融合中值滤波与小波软阈值去噪模型的新元矿视频监控图像滤波方法[J].金属矿山,2017,(12):103-107.

[9]龚梦龙.中值滤波结合小波变换在光谱去噪中的应用[J].科技与创新,2018(12):152-154.

[10]陈晓,唐诗华.改进的中值滤波在图像去噪中的应用[J].地理空间信息,2015,13(6):77-78.

[11]杨作宝,侯凌燕,杨大利.人脸识别的光照预处理算法[J].北京信息科技大学学报(自然科学版),2015,30(6):77-82.

[12]武惠杰,郭天兴.高通滤波器性能研究[J].电力电容器与无功补偿,2014,35(2):5-8.

[13]宋卫华,张青.灰度共生矩阵算法研究[J].黄山学院学报,2014,16(3):34-37.

[14]任国贞,江涛.基于灰度共生矩阵的纹理提取方法研究[J].计算机应用与软件,2014,31(11):190-192,325.

[15]牛冲,牛昱光,李寒,等.基于图像灰度直方图特征的草莓病虫害识别[J].江苏农业科学,2017,45(4):169-172.

[16]任民宏,陈波,鲁秋菊.基于灰度直方图和高斯混合模型多特征肤色识别算法[J].陕西理工学院学报,2017,33(5):43-46,64.

[17]尹华,胡玉平.基于随机森林的不平衡特征选择算法[J].中山大学学报(自然科学版),2014,53(5):59-65.

[18]宁霄,赵鹏.随机森林算法在树木年轮图像分割中的应用[J].林业工程学报,2018,3(4):125-130.

[19]吴帅,赵方.基于随机森林的老年人居住偏好预测研究[J].计算机工程与科学,2018,40(5):924-930.

[20]Tzirakis P,Tjortjis C.T3C: improving a decision tree classification algorithm's interval splits on continuous attributes[J].Advances in Data Analysis and Classification,2017,11(2):353-370.

[21]杨茂,翟冠强.基于决策树理论的风电功率实时预测方法[J].电测与仪表,2018,55(11):120-124.

[22]王小乐,张玉锋,袁媛.基于决策树的卫星故障诊断知识挖掘方法[J].电子设计工程,2018,26(3):165-169.

[23]马伟杰.基于 C4.5 决策树算法的网络学习行为研究[J].科教导刊(电子版),2016,116(8):150-151.

收稿日期:2018-8-21;修回日期:2018-9-10

编辑/雷华