基于 mRMR 和 SVM 的弹性图像特征选择与分类

丁建睿,黄剑华,刘家锋,张英涛

(哈尔滨工业大学 计算机科学与技术学院,150001 哈尔滨)

摘 要: 为客观的评价弹性图像,利用图像处理与模式识别技术进行分析. 首先通过彩色变换获取弹性信息,然后提取弹性图像用户感兴趣区域的一阶统计特征和纹理特征,采用"最小冗余最大相关"(mRMR)算法选择优化的特征,最后使用带有核函数的SVM分类器对弹性图像进行分类. 实验结果表明:该方法具有较高的准确率(92%). 采用计算机辅助诊断技术对弹性图像进行定量分析可有助于提高诊断准确率.

关键词: 弹性图像;纹理;特征选择;最小冗余最大相关;支持向量机

中图分类号: TP391 **文献标志码**: A **文章编号**: 0367-6234(2012)05-0081-05

Elastogram features selection and classification based on mRMR and SVM

DING Jian-rui, HUANG Jian-hua, LIU Jia-feng, ZHANG Ying-tao

(School of Computer Science and Engineering, Harbin Institute of Technology, 150001 Harbin, China)

Abstract: For evaluating elastogram objectively, image processing and pattern recogniton techniques are proposed. First the real elasticity information encoded in color was extracted by transform the image from RGB color space to HSV space. Then the statistical features and texture features were extracted from region of interest on the elastogram. The important and reliable features were selected by using Minimum-Redundancy-Maximum-Relevance (mRMR) algorithm. Finally the selected features were input to the SVM classifier to classify the thyroid nodules into benign and malignant. The experiment results confirmed the method had higher accuracy (92%). It is helpful to improve the clinical accuracy by using CAD techniques.

Key words: elastogram; texture; feature selection; Minimum-Redundancy-Maximum-Relevance; support vector machine

特征选择是图像识别系统中的重要组成部分,根据特征选择准则是否依赖于学习算法,特征选择方法可以分为:Filter模型、Wrapper模型和混合模型[1]. 最小冗余最大相关(mRMR)[2]是基于互信息(Mutual Information)的特征选择方法,它根据最大统计依赖性准则来选择特征. 支持向量机(SVM)[3]在很大程度上解决了过学习、非线性及维数灾难等模式识别中存在的问题,是目前针对小样本估计和预测的最佳分类方法[4]. mRMR与SVM结合的特征选择与分类方法已成功应用到地表分类[5]、遥感图像分类[6]和X光图像分类中[7].

弹性成像是测量生物组织的弹性信息并将其可视化的一项新技术,其概念最早由Ophir[8]于1991年提出,经过算法的不断改进,目前已成功应用于临床. T. Shiina等[9]提出的彩色弹性成像技术将弹性图像上的像素根据其弹性幅值编码到256级伪彩色,颜色从红到蓝,代表组织从软到硬,彩色弹性图像半透明的叠加到超声图像上.

本文针对目前临床上评价弹性图像存在的问题,利用CAD技术对弹性图像进行分析,提出了一种新的客观、定量评价弹性图像的方法. 首先从彩色图像中解码得到弹性信息,然后提取病变区域的弹性特征,包括一阶统计特征和纹理特征,为选取与分类最相关且相互间冗余度低的特征子集,采用最小冗余最大相关(mRMR)特征选择算

收稿日期: 2011-03-20.

基金项目: 国家自然科学基金资助项目(60873142);哈尔滨市优秀学科带头人资助项目(2009RFXXS211).

作者简介: 丁建睿(1973—),男,博士研究生.

通信作者: 丁建睿, jrding@hit.edu.cn

法,获得优化特征子集,最后采用带有核函数的SVM分类器对样本进行训练和测试. 实验结果表明该方法具有高准确性和鲁棒性.

1　弹性图像的特征提取

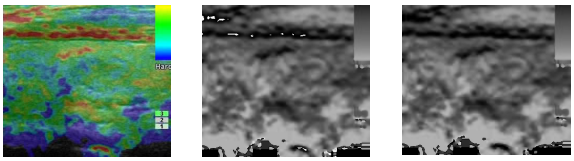
1.1　弹性信息的获取

生物组织的弹性信息是衡量病变中良、恶性的一个重要标准. 然而目前的医学影像手段,包括X射线成像、超声成像、磁共振成像等都无法反映病变的这一生物力学特性. 弹性成像技术的出现解决了这一问题,为医生临床诊断提供了有效的参考依据. 目前利用CAD技术对弹性图像进行分析尚处于起步阶段,部分学者将直方图特征,如:均值、方差、以及病变区域与正常组织的弹性差和比值作为弹性图像的特征[10-11],利用这些特征对弹性图像进行分析和分类,但这些特征都比较简单,且未能反映病变区域弹性信息的空间分布. 本文选择提取弹性图像上病变区域的一阶统计特征来反映弹性信息的总体分布,提取纹理特征来反映病变区域的弹性信息的空间分布. 另外根据图1,将*Hue*归一化到[0,1],选取*Hue*处于蓝色区域的像素为硬度大的组织区域,其与病变区域面积的比值定义为硬组织占病变区域的面积比.

本文将彩色弹性图像从RGB彩色空间变换到HSV彩色空间,其中*Hue*分量反映色彩信息,可以用来表示弹性信息. 从RGB中获取*Hue*分量的计算方法可表示为

$$\begin{cases} \text{if} \quad R \geqslant G \geqslant B, \quad H = 60° \times \dfrac{G-B}{R-B}; \\ \text{if} \quad G > R \geqslant B, \quad H = 60° \times \left(2 - \dfrac{R-B}{G-B}\right); \\ \text{if} \quad G \geqslant B > R, \quad H = 60° \times \left(2 + \dfrac{B-R}{G-R}\right); \\ \text{if} \quad B > G > R, \quad H = 60° \times \left(4 - \dfrac{G-R}{B-R}\right); \\ \text{if} \quad B > R \geqslant G, \quad H = 60° \times \left(4 + \dfrac{R-G}{B-G}\right); \\ \text{if} \quad R \geqslant B > G, \quad H = 60° \times \left(6 - \dfrac{B-G}{R-G}\right). \end{cases} \tag{1}$$

*Hue*从0~360°分别对应着颜色从红,黄,绿,蓝到红,由于对弹性信息的彩色编码为从红到蓝,其中:蓝色代表组织的弹性小,红色代表组织的弹性大,而从式(1)中可以看出300°~360°和0~60°的红色部分有重叠,为了准确的获取弹性信息,需要进行处理,对于$R \geqslant B > G$的像素,将其对应的*Hue*赋值为0,效果如图1所示.



(a)弹性图像　(b)未经过处理的Hue　(c)经过处理的Hue

图1　反映病变区域的弹性图像

在图1(b)中由于*Hue*分量的重叠问题,造成一些在图1(a)中为红色的像素具有较高的*Hue*,没有正确的反映弹性信息,而在图1(c)中由于进行了处理,图1(a)中所有的红色像素在图1(c)中均为低*Hue*.

1.2　特征提取

一阶统计特征反映了病变区域全局的弹性信息,本文采用均值(Mean)、众数(Mode)、方差(Variance)、偏斜率(Skewness)、峰度(Kurtosis)、熵(Entropy)、能量(Energy)、光滑度(Smoothness)作为一阶统计特征(特征编号为F1-F8).

硬组织区域面积比定义为病变区域内*Hue* > 0.5的像素之和与病变区域面积之比,相比对弹性图像的评分法,该特征定量的给出了病变区域内部软硬组织的比例(特征编号为F9).

图像的纹理特征提供了像素灰度的空间分布信息,对于弹性图像来说,病变区域的纹理特征反映了该区域弹性信息的空间分布,即病变区域组织的软硬分布以及生长、浸润状况. 对图像纹理特征的描述分为统计描述方法和结构化描述方法,由于统计描述方法计算简单而被广泛使用,共生矩阵是一种常用的图像纹理统计描述方法.

共生矩阵[12]定义为距离为d,方向为θ的灰度级i和j的联合概率密度,它不仅反映了灰度的分布特性,也反映了具有相同灰度级的位置分布特性,是有关图像灰度变化的二阶统计特征,其元素$C_{d,\theta}(i,j,d,\theta)$定义为

$$C_{d,\theta}(i,j,d,\theta) = \| \{((x_1,y_1),(x_2,y_2)) \mid x_2 - x_1 = d\cos\theta, y_2 - y_1 = d\sin\theta, I(x_1,y_1) = i, I(x_2,y_2) = j\} \|.$$

式中:(x_1,y_1)、(x_2,y_2)分别为弹性图像中病变区域的像素;$I(\cdot)$为像素的*Hue*;$\| \cdot \|$为满足条件的像素对的个数. 本文提取4个方向上($\theta = 0°, 45°, 90°, 135°$),4个距离($d = 1,2,3,4$)的16个共生矩阵,为减少计算复杂性并保留图像细节,64个*Hue*用来计算共生矩阵. 从共生矩阵中提取对比度(contrast),相关(correlation),能量(energy)和一致性(homogeneity)4个特征,为减少特征空间的维数,对同一距离的特征进行平均,一共从共生矩阵中提取16个特征(特征编号为F10-F25).

2　特征选择和分类

2.1　特征选择

本文从弹性图像的病变区域总共提取了 25 个特征,特征之间的相关性和冗余性会降低分类的准确率,同时医学图像通常属于小样本学习,特征过多将提高分类器的复杂度,造成过拟合,降低分类器的泛化能力,因此需要对特征集合进行选择和优化.

本文采用"最小冗余最大相关"(mRMR)方法进行特征选择. 特征选择的目的是从特征空间中寻找与目标类别有最大相关性且相互之间具有最少冗余性的 m 个特征[13],最大相关和最小冗余的定义为

$$\max \quad D(S,c), \quad D = \frac{1}{|S|}\sum_{x_i \in S} I(x_i;c);$$

$$\min \quad R(S), \quad R = \frac{1}{|S|^2}\sum_{x_i,x_j \in S} I(x_i,x_j).$$

式中:S 为特征集合;c 为目标类别;$I(x_i;c)$ 为特征 i 和目标类别 c 之间的互信息;$I(x_i,x_j)$ 为特征 i 与特征 j 之间的互信息.

给定两个随机变量 x 和 y,它们之间的互信息根据其概率密度函数 $p(x)$,$p(y)$ 和 $p(x,y)$ 分别定义为

$$I(x;y) = \iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}\mathrm{d}x\mathrm{d}y. \quad (2)$$

对于多元变量 S_m 和目标类别 c,互信息定义为

$$I(S_m;c) = \iint p(S_m,c)\log\frac{p(S_m,c)}{p(S_m)p(c)}\mathrm{d}S_m\mathrm{d}c. \quad (3)$$

将式(2),(3)进行组合,可以得到"最小冗余最大相关"(mRMR)的特征选择标准为

$$\max \quad \Phi(D,R), \quad \Phi = D - R. \quad (4)$$

式(4)表示应该选择与类别最大相关而与候选特征最小冗余的特征. 假定已确定一个有 m 个特征的数据集 S_m,下一步需要从数据集 $\{S - S_m\}$ 中选择使得式(4)最大化的第 $m+1$ 个特征为

$$\max_{x_i \in S-S_m}\left[I(x_i;c) - \frac{1}{m}\sum_{x_j \in S_m} I(x_i;x_j)\right].$$

2.2　SVM 分类器

本文采用带有核函数的 SVM(KSVM)分类器对弹性图像进行分类,训练样本被 KSVM 分类器映射到高维空间以获得优化的分类平面,KSVM 具有泛化能力强和可以通过将样本映射到高维空间以解决非线性分类问题的优点.

两类问题可以通过利用 KSVM 最小化来进行求解

$$\Phi(w,\xi) = \frac{1}{2}(ww) + C\left(\sum_{l=1}^{L}\xi_l\right).$$

其中:$\forall \xi_l \geqslant 0$, $wx_l + b \geqslant 1 - \xi_l$ if $y_l = -1$;且

$$wx_l + b \leqslant -1 + \xi_l \text{ if } y_l = -1.$$

式中:w 为需要求解的分隔平面;ξ 为软边缘;x_l 为训练样本;y_l 为 x_l 的已知类别;L 为训练样本的个数;C 为常数.

上述问题可以利用拉格朗日方法变换为寻找参数向量 α^0 以最大化为

$$w(\alpha) = \sum_{l=1}^{L}\alpha_l - \frac{1}{2}\sum_{i,j}^{L}\alpha_i\alpha_j y_i y_j K(x_i,x_j),$$

满足

$$\sum_{l=1}^{L}\alpha_l y_l = 0,\ 0 \leqslant \alpha_l \leqslant C.$$

式中 $K(x_i,x_j)$ 为核函数,对于每个训练样本 x_i,有与之对应的参数 α_i^0,如果 $\alpha_i^0 \neq 0$,该训练样本称为支持向量(support vector). 训练结束后,对于测试样本 x 的类别为

$$F(x,\alpha^0) = \mathrm{sign}\left(\sum_{l=1}^{N_{SV}} y_l\alpha_l^0 K(x,x_S) + b\right).$$

式中:x_S 为支持向量;N_{SV} 为支持向量的个数;$K(x,x_S)$ 为核函数.

本文采用 RBF 核函数,其定义为

$$K(x_i,x_j) = \exp(-\gamma\|x_i - x_j\|^2), \quad \gamma > 0.$$

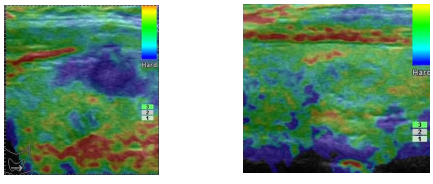
通过网格搜索法,最终选择性能最优的 C 和 γ 值.

3　实验结果与分析

3.1　图像获取

为了验证本文的方法在弹性图像上应用效果,本文对 125 例甲状腺弹性图像进行了分析处理. 本文采用的所有甲状腺弹性图像由哈尔滨医科大学附属第二医院提供,并由哈尔滨医科大学附属第二医院超声科专家对图像中的病变区域进行手工标注,所有病例均经过病理检验,其中:56 例恶性,69 例良性. 超声图像和弹性图像均采用配备线性探头,中心频率为 6 - 13MHz 的日立 Vision900 商用超声设备获得.

图 2(a),(b)分别为一例恶性肿瘤和一例良性肿瘤的弹性图像.



(a)恶性肿瘤　　(b)良性肿瘤

图 2　一例恶性肿瘤和一例良性肿瘤的弹性图像

3.2 分类结果

首先,影像医师在超声图像上勾勒病变区域,该区域被自动映射到对应的弹性图像上;彩色弹性图像从 RGB 空间变换到 HSV 空间,特征提取自病变 *Hue* 分量图像中的病变区;采用 mRMR 算法进行特征选择后,取排名前 5 位的特征作为样本采用 KSVM 进行训练和测试,这些特征分别为 F9:硬区域面积比;F20:共生矩阵(d = 3)的能量(Energy)特征;F21:共生矩阵(d = 3)的同质(Homogeneity)特征;F5:一阶统计特征中的峰度(Kurtosis)特征;F23:共生矩阵(d = 4)的相关(Correlation)特征. 其中硬区域面积比可以认为是评分法的一种量化形式,而共生矩阵的特征及一维统计特征反映了弹性信息的空间分布和总体分布,这些特征反映了病变区域的软硬程度及硬度分布情况.

在实验中采用"留一"(leave-one-out)测试法[14],一幅图像用来测试,其他图像用来训练,该过程不断迭代,直至所有图像都被测试过为止.

方法的性能采用分类准确率来评价,定义正确分类和错误分类的恶性病例个数为真阳性(True Positive, TP)和假阴性(False Negative, FN),正确和错误分类的良性病例个数为真阴性(True Negative, TN)和假阳性(False Positive, FP),分类准确率定义为:(TP + TN)/(TP + TN + FP + FN).

为了证明 5 个特征是否为最佳特征集合,分别对 4 个特征和 6 个特征的情况进行了实验,实验结果如表 1 所示.

表 1 分类准确率

特征集合	真阳性	假阴性	真阴性	假阳性	准确率/%
所有特征	47	9	63	6	88.0
前 6 位	52	4	61	8	90.4
前 5 位	52	4	63	6	92.0
前 4 位	52	4	60	9	89.6

由于特征之间存在冗余和相关,当选取所有特征进行分类时并不能达到最佳效果,当选择由 mRMR 算法所选择的前 5 位特征进行分类,在实验所用的样本集合上达到最佳效果,可见 mRMR 算法有效的去除了特征之间的冗余和相关.

3.3 鲁棒性测试

为进一步测试方法的鲁棒性,从相同病例中选取 125 幅未经训练的图像作为测试样本,利用已经训练好的分类器进行分类,测试结果如表 2 所示.

表 2 相同病例不同图像的测试结果

真阳性	假阴性	真阴性	假阳性	准确率/%
48	8	62	7	88

由于 125 例用作测试的图像未参加训练,性能有所下降,但仍然达到了 88% 的准确率,证明采用优化特征集合所得到的分类器具有良好的泛化能力,验证了方法的稳定性和可靠性.

3.4 与仪器计算的弹性比值进行比较

VISION 900 可以根据影像医师勾勒的正常组织区域和病变区域自动计算两区域的平均弹性比值,称为弹性计算(strain ratio),该值越大说明病变区域与正常组织的弹性差异越大,病变为恶性的可能性就越大. 用该值进行分类的结果和采用本文方法分类的结果比较如表 3 所示.

表 3 与仪器计算的弹性比值的比较

方法	真阳性	假阴性	真阴性	假阳性	准确率/%
前 5 位特征	52	4	63	6	92.0
弹性比值	44	12	65	4	87.2

弹性比值法试图用定量的方法来评价彩色弹性图像,但由于病变区域与正常组织区域需要影像医师手动进行选择,同样具有主观性,容易造成假阳性和假阴性. 实验结果表明,本文提出的方法可以客观、定量的评价弹性图像,分类性能优于弹性比值法.

3.5 与评分法的性能比较

弹性图像评分法是影像医师根据弹性图像病变区域的色彩分布,人为主观给出的分值,用来评价图像的良恶性程度,甲状腺弹性图像中通常采用 4 分评分法,分值越高,其恶性程度越高. 为了与该方法进行比较,超声专家对 125 例甲状腺弹性图像分别进行了评分,评分法得到的结果与本文提出方法的结果比较如表 4 所示.

表 4 与评分法的比较

方法	真阳性	假阴性	真阴性	假阳性	准确率/%
前 5 位特征	52	4	63	6	92.0
评分法	44	12	60	9	83.2

由于评分法对医师要求很高,并且与环境、心理等众多因素有关,因此具有很强的主观性,而本文提出的方法可以对弹性图像进行客观、定量的评价,性能远远高于评分法的结果.

4 结 论

1) 实验结果表明该方法达到了预期的效果,与现存的方法相比,具有更高的准确率和可靠性.

2) 有效特征的选择降低了分类器的复杂度,

减少了计算量,提高了泛化能力,为本方法在实时医学图像处理系统中的应用奠定了基础.

3) 利用该方法对甲状腺弹性图像进行定量的分析,能够帮助医生客观、准确的判断病变的性质,为进一步的诊断提供了有效的参考依据.

参考文献:

[1] LIU Huan, YU Lei. Toward integrating feature selection algorithms for classification and clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(4): 491 – 502.

[2] DING C, PENG H. Minimum redundancy feature selection from microarray gene expression data[J]. Journal of Bioinformatics and Computational Biology, 2005, 3(2): 185 – 205.

[3] VAPNIK V. The nature of statistical learning theory [M]. New York: Springer Verlag, 2000.

[4] 顾志伟,吴秀清,荆浩, 等. 一种基于特征选择的医学图像检索方法 [J]. 中国生物医学工程学报, 2007, 26 (1): 30 – 34.

[5] 李士进,陶剑,林林, 等. 面向宏观地表分类的特征选择算法比较研究 [J]. 计算机工程与应用, 2008, 44(21):130 – 132.

[6] 刘峰,龚健雅. 一种基于多特征的高光谱遥感图像分类方法[J]. 地理与地理信息科学,2009, 25(3): 19 – 22.

[7] YOON Sejong, KIM Saejoon. Mutual information-based SVM-RFE for diagnostic classification of digitized mammograms [J]. Pattern Recognition Letter, 2009, 30(16):1489 – 1495.

[8] OPHIR J, CESPEDES I, PONNEKANTI H, *et al.* Elastography: a quantitative method for imaging the elasticity of biological tissues [J]. Ultrasonic imaging, 1991, 13 (2): 111 – 134.

[9] SHIINA T, YAMAKAWA M, NITTA N, *et al.* Clinical assessment of real-time, freehand elasticity imaging system based on the combined autocorrelation method [C]//Proceedings of the IEEE Ultrasonics Sympos. Washington: IEEE Xplore, 2003: 664 – 667.

[10] Shirley SELVAN M K, SHENBAGADEVI S, SURESH S. Feature extraction for characterization of breast lesions in ultrasound echography and elastography [J]. Journal of Computer Science, 2010, 6(1): 67 – 74.

[11] Huang Chiun-Sheng, MOON Woo-Kyung, SHEN Wei-Chih, *et al.* Analysis of elastographic and B-Mode features at sonoelastography for breast tumor classification [J]. Ultrasound in Medicine and Biology, 2009, 35(11):1794 – 1802.

[12] NEWELL D, NIE K, CHEN J H, *et al.* Selection of diagnostic features on breast MRI to differentiate between malignant and benign lesions using computer-aided diagnosis: differences in lesions presenting as mass and non-mass-like enhancement [J]. European Radiology, 2010, 20(4): 771 – 781.

[13] PENG H, LONG F H, DING C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226 – 1238.

[14] CHANG Ming-wei, LIN Chih-jen. Leave-one-out bounds for support vector regression model selection [J]. Neural Computation, 2005, 17(5):1188 – 1222.

(编辑　张　红)