文章编号：2095-2163(2019)02-0091-05 中图分类号：TP311.13 文献标志码：A

基于信息融合的医疗影像辅助决策研究

杜 优

(东华大学 计算机科学与技术学院，上海 201620)

摘 要：乳腺癌是女性的高发癌症，威胁着全球女性的身体健康，因此乳腺癌良恶性研究与决策对于女性乳腺癌的诊断有着重要作用。本文研究提出了一种新的信息融合框架，用于对乳腺癌良恶性进行分类和预测，该框架首先对乳腺超声影像感兴趣区域提取纹理特征，之后对得到的纹理特征数据集使用 4 个基本分类器：朴素贝叶斯、决策树、SVM、KNN 进行分类，对基本分类器的分类结果使用投票法进行决策，最后对分类器信息进行融合，并将 4 个基本分类器的分类结果与基于信息融合的分类器模型结果进行比较。通过实验确定，与单独的分类器相比，基于决策的分类器方法实现了较高的准确度和较低的分类错误率。

关键词：信息融合；灰度共生矩阵；朴素贝叶斯；决策树；SVM；KNN

Research on medical image assistant decision based on information fusion

DU You

(School of Computer Science and Technology, Donghua University, Shanghai 201620, China)

【Abstract】 Breast cancer is a high-risk cancer in women, which threatens the health of women around the world. Therefore, the research and decision-making of breast cancer is important for the diagnosis of breast cancer in women. This paper proposes a new information fusion framework for classification and prediction of benign and malignant breast cancer. The framework first extracts texture features from the region of interest of the breast ultrasound image, and then uses four basics for the obtained texture feature dataset. Classifiers: Naive Bayes, Decision Tree, SVM, KNN are classified, the classification results of the basic classifier are determined using the voting method, finally the classifier information is fused. The classification results between the results of the four basic classifiers and the results of the information fusion classifier model are compared. It is experimentally determined that the decision-based classifier method achieves higher accuracy and lower classification error rate than the separate classifier.

【Key words】 information fusion; gray level co-occurrence matrix; NB; decision tree; SVM; KNN

0 引 言

乳腺癌是女性的高发癌症，世界卫生组织 GLOBOCAN 发布的最新数据表明 2018 年大约有 210 万新诊断的乳腺癌病例，占女性癌症发病率的 25%[1]。虽然中国女性的乳腺癌发病在全球处于较低的水平，但国内的乳腺癌发病趋势在逐年增高[2]。

作为一种乳腺癌的常用检测手段，乳腺超声具有价格便宜、检测率高等优点，因此日常乳腺癌的检测与筛查大多采用乳腺超声的手段。

在国内影像学医师较为稀缺，而又面临较大市场需求的情况下，基于人工智能的辅助决策诊断可以大大减少影像学医师的工作量，同时也可为社会的各类就医群体提供更好医疗保障和服务。

图像特征提取和分类模型构建是医疗影像识别领域的 2 个主要步骤。特征提取主要是提取影像的相关特征，例如纹理特征、颜色特征、形状特征等。分类模型的构建主要是指利用特征信息来构建并学习一套分类准则，在此分类准则下可以对图像进行分类和预测。在医学影像分类中，采用的分类方法主要有支持向量机、决策树等。但是不同的分类器的分类能力不同，例如，对于一张图像某些分类器的识别效果较好，但是某些分类器的识别效果较差。所以，可以通过适当地融合分类器的分类结果，来提高分类模型的准确率。

本文采用了公开乳腺超声数据，对超声数据进行了纹理特征的提取，使用了 4 种基本分类器，对分类器的分类结果进行基于决策的信息融合，最后获得预测与诊断结果。

1 本文方法

本文提出的基于信息融合的医疗影像辅助决策方法的总体流程，见图 1。首先提取出乳腺超声图像感兴趣区域，生成乳腺超声纹理特征数据集，将获得的纹理特征的数据集输入到 4 个不同的分类器

作者简介：杜 优(1994-)，女，硕士研究生，主要研究方向：人工智能、机器学习、数据分析。

收稿日期：2018-11-26

(朴素贝叶斯、决策树、SVM、KNN)中,将 3 个分类器分类的结果进行决策层的信息融合,得到最终的识别结果。

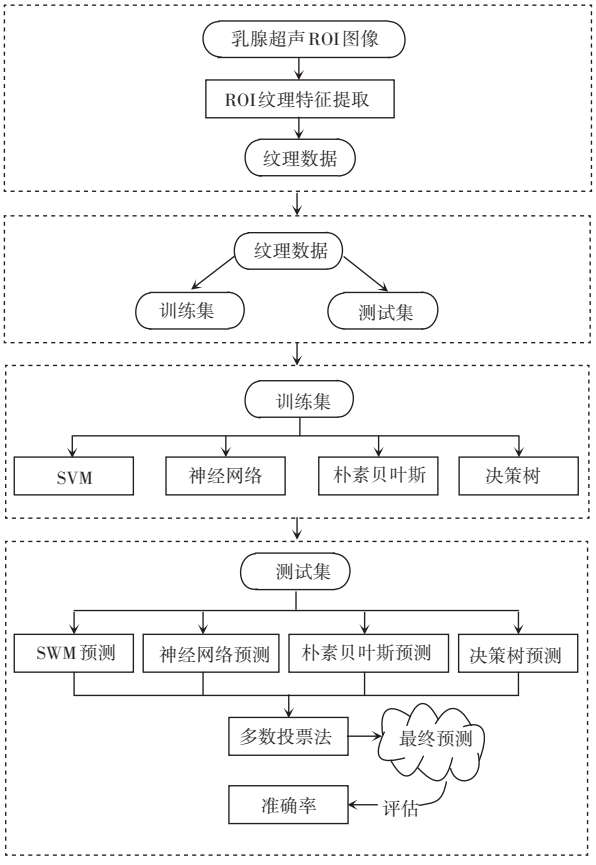

图 1 基于决策信息融合医疗影像流程图

Fig. 1 Flow chart of fusion medical image based on decision information

1.1 纹理特征提取

乳腺超声图像具有特殊性和复杂性,因为需要采集乳腺超声感兴趣区域的特征来进行数据挖掘和数据分析。

灰度共生矩阵的定义如下:假设一张矩形的图片有 N_x 列 N_y 行,出现在每个像素处的灰度级被量化为 N_g 级。设 $L_x = \{1,2,\cdots,N_x\}$ 为列, $L_y = \{1,2,\cdots,N_y\}$ 为行,且 $G_x = \{0,1,\cdots,N_g - 1\}$ 是 N_g 个量化灰度级的集合,集合 $L_x \times L_y$ 是按行列指定排序的图像像素集。图像 I 可以表示为将 G 中的一些灰度级分配给 $L_x \times L_y$ 中的每个像素或坐标对的函数,即 $I: L_x \times L_y \rightarrow G$。纹理上下文信息由相对频率 P_{ij} 的矩阵指定,在图像上距离为 d 的 2 个像素,分别记为 i 和 j。灰度共生矩阵是角度关系和相邻像素之间距离的函数。在本文的研究中使用了 6 个特征,以下等式定义了这些特征。设 $p(i,j)$ 是标准化灰度共生矩阵的第 (i,j) 个条目。矩阵的行和列的平均值和标准差可分别表示为:

$$\mu_x = \sum_i \sum_j i \cdot p(i,j); \tag{1}$$

$$\mu_y = \sum_i \sum_j j \cdot p(i,j); \tag{2}$$

$$\sigma_x = \sum_i \sum_j (i - \mu_x)^2 \cdot p(i,j); \tag{3}$$

$$\sigma_y = \sum_i \sum_j (j - \mu_y)^2 \cdot p(i,j). \tag{4}$$

在此基础上,研究继而给出各特征定义的数学表述具体如下。

(1)对比度(contrast)

$$f_1 = \sum_{n=0}^{N_g-1} n^2 \left\{ \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) \mid |i=j| = n \right\}; \tag{5}$$

(2)非相似性(dissimilarity)

$$f_2 = \sum_i \sum_j |i-j| \cdot p(i,j); \tag{6}$$

(3)同质性(homogeneity)

$$f_3 = \sum_i \sum_j \frac{1}{1 + (i-j)^2} p(i,j); \tag{7}$$

(4)ASM 能量(ASM)

$$f_4 = \sum_i \sum_j p(i,j)^2; \tag{8}$$

(5)能量(energy)

$$f_5 = \sqrt{\sum_i \sum_j p(i,j)^2}; \tag{9}$$

(6)自相关(correlation)

$$f_6 = \frac{\sum_i \sum_j (ij)\, p(i,j) - \mu_x \mu_y}{\delta_x \delta_y}. \tag{10}$$

本文采用灰度共生矩阵来提取乳腺超声感兴趣区域的纹理特征,共采集了对比度(contrast)、非相似性(dissimilarity)、同质性(homogeneity)、ASM 能量(ASM)、能量(energy)、自相关(correlation)6 个特征,每个特征扩展为 4 个维度,共 24 个特征,最终生成纹理数据集。例如对于 id 为 us1 的乳腺超声图像提取纹理特征之后得到的数据集示例详见表 1。

1.2 分类器分类

机器学习实质是研究如何根据过去的观察结果自动学习做出准确的预测。分类算法是机器学习当中的常用算法,对于数据的预测与分类有着重要的意义。本文选取的分类算法主要有朴素贝叶斯、决策树、SVM、CNN。基本分类器分类方法可探讨论述如下。

1.2.1 朴素贝叶斯分类器

朴素贝叶斯分类器是一种概率分类器,同时也

是基于贝特斯定理的分类技术，并假设预测变量之间具有独立性。朴素贝叶斯分类器假定类中特定特征的存在与其它特征的存在无关。

表 1　乳腺影像感兴趣区域特征提取结果示例

Tab. 1　Example of feature extraction results of regions of interest in breast imaging

id	us1
contrast - 0 - 0	488.330 8
dissimilarity - 0 - 0	16.595 46
homogeneity - 0 - 0	0.059 729
ASM - 0 - 0	0.000 23
energy - 0 - 0	0.015 157
correlation - 0 - 0	0.721 348
contrast - 0 - 1	625.332 7
dissimilarity - 0 - 1	18.718 6
homogeneity - 0 - 1	0.056 009
ASM - 0 - 1	0.000 217
energy - 0 - 1	0.014 746
correlation - 0 - 1	0.649 198
contrast - 0 - 2	713.317
dissimilarity - 0 - 2	20.643 29
homogeneity - 0 - 2	0.060 398
ASM - 0 - 2	0.000 307
energy - 0 - 2	0.017 532
correlation - 0 - 2	0.617 2
contrast - 0 - 3	747.272
dissimilarity - 0 - 3	20.846 01
homogeneity - 0 - 3	0.050 368
ASM - 0 - 3	0.000 199
energy - 0 - 3	0.014 114
correlation - 0 - 3	0.580 856

贝叶斯定理提供了一种从 $P(c)$，$P(x)$ 和 $P(x \mid c)$ 来计算后验概率 $P(c \mid x)$ 的方法，可将其解析为如下数学公式：

$$P(c \mid x) = \frac{P(x \mid c)\,P(c)}{P(x)}, \tag{11}$$

1.2.2　KNN 分类器

KNN 是用于分类和回归的十大数据挖掘算法之一，是懒惰学习器的代表。KNN 根据分配给测试样本的 KNN 标签进行预测。工作原理可阐述如下。

在一组训练数据集 D 上选择 K 的初始值，因为没有标准的方法来设置 K 的值，使得 K 的初始值将根据实验结果进行随机选择。根据样本数据的所需结果确定 K 的值。此后再使用欧几里得距离公式测量采样点 X 与其 K 个邻居之间的距离。这些样本之间的距离可定义为：

$$d(X,Y) = sqrt\left(\sum_{i=1}^{n}(x_i - y_i)\right), \tag{12}$$

最后根据得到的距离来进行排序，选择距离最小的 K 个点，由此完成分类、回归或其它任务。

1.2.3　决策树分类器

决策树以树结构的形式构建分类或回归模型。通过将数据集划分为越来越小的子集，同时逐步开发相关的决策树。最终结果是具有决策节点和叶节点的树。决策节点具有 2 个或更多分支，叶节点表示分类或决定。树中最顶层的决策节点，对应于成为根节点的最佳预测器。决策树可以处理分类和数值数据。

构建决策树的核心算法，称为 ID3。其中，采用自上而下的贪婪搜索方式，通过可能的分支空间进行无回溯。

信息增益基于在属性上拆分数据集后熵的减小来构建决策树，对其设计步骤可简洁分述如下。

(1)计算目标的熵，如式(13)所示：

$$E(T) = \sum_{i=1}^{c} - P_i \log_2 P_i, \tag{13}$$

(2)将数据集拆分为不同的属性，计算每个分支的熵，而后按比例添加，来获得拆分的总熵，在分割之前从熵中减去所得的熵，最终信息增益或熵减小。研究可得数学运算公式如下：

$$G(T, X) = E(T) - E(T, X). \tag{14}$$

(3)选择具有最大信息增益的属性作为决策节点，将数据集除以其分支，并且在每个分支上重复相同的过程。

(4)熵为 0 的分支为叶子节点。

(5)熵大于 0 的分支需要进一步决策并分裂。

(6)ID3 算法在非叶子分支上运行，直到所有数据都被分类。

1.2.4　SVM 分类器

支持向量机通过在高维或者无限维空间中构造超平面来解决分类、回归或其它任务。SVM 通过一个与任何类的最近训练数据点具有最大距离的超平面来实现分类、回归或其它任务。通常，边缘越大，分类器的泛化误差越低。虽然通常情况下，原始数据集可以在有限空间中进行描述，但是可能会面临要区分的集合不是线性可分的问题。所以，有学者提出将原始的有限维空间映射到更高维空间，使得分类或预测更加容易。高维空间中的超平面则定义为在该空间中具有向量的内积是恒定的点集。

1.3 决策层信息融合

信息融合是指将来自于不同的信息源、多格式信息等进行合并,从而产生更加完整、准确的信息或决策。在数据挖掘领域信息融合得到了广泛的应用,主要应用在基于特征层的信息融合、基于决策层的信息融合和基于数据源的信息融合等方面。本文所使用的信息融合方法即为基于决策层的信息融合。

由于单一的分类器可能会因为分类器的分类原理而有不同的结果,为了提高分类的效果,本文在使用 4 种不同的分类器划定分类后,将不同分类器的分类结果进行信息融合,如此就能得到一个更准确、更可靠的分类模型。

对于乳腺超声影像来说,基于决策层的信息融合指的是多分类器结果融合。本文使用投票法来对多分类器结果进行信息融合。由于每个分类器对每一种分类结果均为概率输出,则本文使用的投票法为对每一种分类器的概率输出进行累加之后比较每一种分类的概率大小,从而计算求出分类结果,比如,对于一个分类为 0 和 1 二分类,分类为 0 的概率,投票法公式为:

$$p_0 = p_{0(ID-3)} + p_{0(SVM)} + p_{0(KNN)} + p_{0(NB)}, \quad (15)$$

对于分类为 1 的概率,投票法公式为:

$$p_1 = p_{1(ID-3)} + p_{1(SVM)} + p_{1(KNN)} + p_{1(NB)}. \quad (16)$$

之后对 p_0 和 p_1 进行归一化,即可得到最终的分类结果。

2 实验结果及分析

本文首先对乳腺超声感兴趣区域图像进行纹理特征提取,得到新的纹理数据集,而后对纹理数据集进行随机化并以 7:3 来进行划分,分为训练集和测试集。

朴素贝叶斯分类器、KNN 分类器、决策树分类器、SVM 分类器以及使用投票法进行基于决策的分类模型的 ROC 曲线如图 2 所示。

在仿真实验中,观察分析后可得到如下实验结果:高斯贝叶斯分类器在训练集上的准确率为 72%,在测试集上的准确率为 78.67%, *AUC* 值为 85.76%;KNN 分类器在训练集上的准确率为 95.43%,在测试集上的准确率为 89.33%, *AUC* 值为 96.77%;决策树分类器在训练集上的准确率为 100%,在测试集上的准确率为 90.67%, *AUC* 值为 90.14%;此外,SVM 分类器在训练集上的准确率为 84.57%,在测试集上的准确率为 85.33%, *AUC* 值为 94.83%;使用投票法对 4 种分类器做基于决策的数据融合分类模型,在训练集上的准确率为 99.43%,在测试集上的准确率为 93.33%, *AUC* 值为 97.65%。

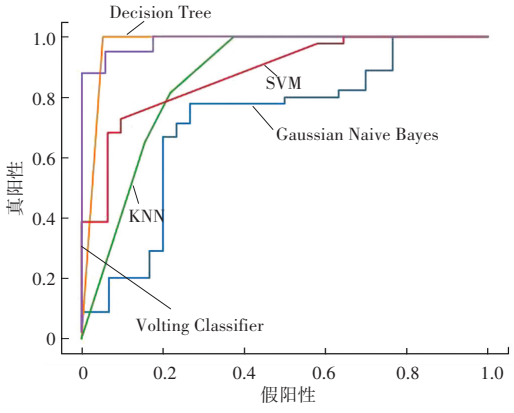

图 2 4 种基本分类器和融合分类模型结果示意对比图

Fig. 2 Schematic comparison of the results of the four basic classifiers and the fusion classification model

4 种基本分类器和投票法分类模型的 *AUC* 值折线图和准确率折线图如图 3、图 4 所示。

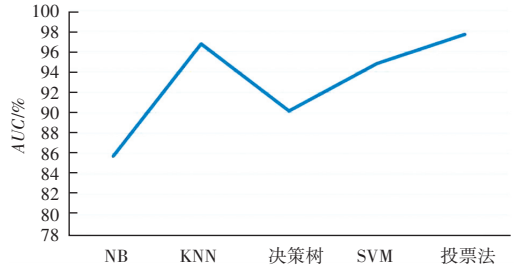

图 3 4 种基本分类器和融合分类模型 *AUC* 值对比图

Fig. 3 Comparison of *AUC* values of four basic classifiers and fusion classification models

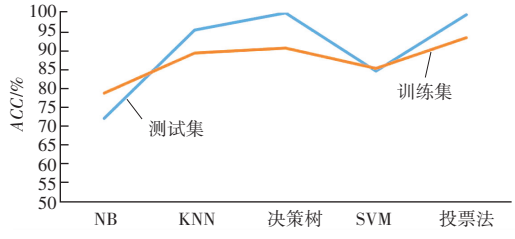

图 4 4 种基本分类器和融合分类模型准确率对比图

Fig. 4 Comparison of the accuracy of the four basic classifiers and the fusion classification model

由图 3 和图 4 可知,相比于单个分类器的分类效果,本文使用的融合了多个分类器的投票法模型有更高的 *AUC* 值和准确率。本文使用的投票法模型具有较高的准确率,对临床上的乳腺超声影像识别有着重要意义。

3 结束语

本文针对乳腺超声影像感兴趣区域进行了纹理

特征的提取，组成了新的纹理数据集，并对纹理数据集使用基本分类器进行分类，接下来又对基本分类器使用投票法进行信息融合，来获得新的分类模型。此后的实验结果表明：基于信息融合的分类模型相比 4 个基本分类器有更好的分类效果和更高的准确率，对于临床上乳腺癌的良恶性分类有较高的实用价值。

参考文献

[1] 王晓明，王玲. 电动机的 DSP 控制—TI 公司 DSP 应用[M]. 北京：北京航空航天大学出版社，2004.

[2] Texas Instruments. TMS320LF/LC240xA DSP Controllers Reference Guide - System and Peripherals [Z]. Texas：Texas Instruments，2006.

[3] 贺晓建，王福明. 基于灰度共生矩阵的纹理分析方法研究[J]. 山西电子技术，2010(4)：89-90，93.

[4] 徐天伟. 基于灰度共生矩阵的医学 PET 图像纹理分析研究[J]. 电脑知识与技术，2017，13(5)：219-220.

[5] 张欣，梁宗保. 多分类器融合算法研究与应用[J]. 湘潭大学自然科学学报，2011，33(2)：99-103.

[6] GUO Zhe，LI Xiang，HUANG Heng，et al. Medical image segmentation based on multi-modal convolutional neural network：Study on image fusion schemes[J]. arXiv preprint arXiv：1711.00049v2. 2017.

[7] LIU Jingxin，ZHANG Tongzhou，ZHENG Caixia，et al. Recognition of breast section cancer cells based on double-layer information fusion[J]. China Medical Devices，2018，33(1)：20-23，33.

[8] MINAVATHI，MURALI S，DINESH M S. Information fusion from mammogram and ultrasound images for better classification of breast mass[C]// Proceedings of International Conference on Advances in Computing. India：Springer，2012：943-953.

[9] SUN Li，LI Lihua，XX Weidong，et al. A novel classification scheme for breast masses based on multi-view information fusion[C]// International Conference on Bioinformatics and Biomedical Engineering. Chengdu，China：IEEE，2010：1-4.

[10] 孙利. 基于多分类器和双视角信息融合的乳腺钼靶图像病灶分类算法研究[D]. 杭州：杭州电子科技大学，2010.

[11] 马继丰. 基于决策层信息融合的手写汉字识别研究[D]. 西安：西安科技大学，2007.

[12] 许良凤，徐小兵，胡敏，等. 基于多分类器融合的玉米叶部病害识别[J]. 农业工程学报，2015，31(14)：194-201.

[13] 张欣，梁宗保. 多分类器融合算法研究与应用[J]. 湘潭大学自然科学学报，2011，33(2)：99-103.

[14] 董火明，高隽，汪荣贵. 多分类器融合的人脸识别与身份认证[J]. 系统仿真学报，2004，16(8)：1849-1853.

[15] 苏燕妮，汪源源. 乳腺肿瘤超声图像中感兴趣区域的自动检测[J]. 中国生物医学工程学报，2010，29(2)：178-184.

[16] 朱晓琳，李秀英，朱鹰. 高频彩色超声波检查在早期乳腺癌诊断中的应用[J]. 肿瘤研究与临床，2001，13(6)：419-420.

(上接第 90 页)

通过上文的分析，在藏文音节结构中，辅音有可能是占位辅音、也有可能是非占位辅音，对应占位辅音编码和非占位辅音编码，元音只能是非占位字符。结合上文藏文文字特点和拼写规律，同时结合藏文正字法知识，在对藏文音节结构进行判定时，如音节结构中有占位辅音，则该占位辅音可能为前加字、上加字、基字、后加字或者再后加字；如音节结构中有非占位辅音，则非占位辅音可能为基字或下加字。由此可以归纳总结藏文音节的编码特点：现代藏文音节中最多只能存在一个纵向结构的字丁组合；如果该纵向结构字丁组合中仅存在一个占位辅音，那么连同元音在内最多包含 3 个连续的非占位字符；前加字、后加字和再后加字都是占位辅音[7]。所以，在对藏文音节进行判定过程中，判定的关键是判定出该音节结构中的占位辅音和非占位辅音，确定基字后，再结合藏文正字法知识，进一步对藏文的音节结构进行判定。

5　结束语

本文主要对现代藏文音节结构进行了分析和研究，通过分析藏文的文字特点和拼写规律同时结合藏文正字法知识，对藏文正确的拼写结构做出了归纳和总结，接下来则依据小字符集编码方案对藏文音节不同的部件实现划分和确定，最终达到对藏文音节结构进行判断，从而识别藏文音节结构中不同的部件。但是由于梵文音节和外来的新造词的存在，结果会对藏文音节结构的判定产生一定的影响，这在未来工作中需要展开后续的研究和深化。

参考文献

[1] 周季文. 藏文拼音教材[M]. 北京：民族出版社，1983.

[2] 陈小莹，艾金勇. 基于小字符集藏文拉丁转写系统的设计与实现[J]. 中文信息学报，2016，30(3)：74-78.

[3] 完么扎西，尼玛扎西. 小字符集现代藏文排序技术的研究[J]. 计算机工程与应用，2013，49(8)：146-150.

[4] 黄小兰，黄鹤鸣，钟小莉. 现代藏文音节的划分与确定[J]. 计算机应用与软件，2012，29(9)：62-65.

[5] 关白，才科扎西. 现代藏文音节字自动校对研究[J]. 计算机工程与应用，2012，48(29)：151-156.

[6] 黄鹤鸣，达飞鹏. 基于排序的现代藏文音节判定[J]. 计算机应用，2009，29(7)：2003-2005，2008.

[7] 陈小莹，艾金勇. 基于小字符集编码的藏文音节结构判定[J]. 西北民族大学学报(自然科学版)，2015，36(4)：33-36.