ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 35,No. 3,May 2020,pp. 536－548
DOI:10. 16337/j. 1004-9037. 2020. 03. 016

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

信息增益混合邻域粗糙集的肺部肿瘤高维特征选择算法

陆惠玲[1]，周　涛[1,2,4]，张飞飞[3]，霍兵强[2]

(1. 宁夏医科大学理学院,银川,750004; 2. 北方民族大学计算机科学与工程学院,银川,750021; 3. 中国电信股份有限公司宁夏分公司,银川,750002; 4. 宁夏智能信息与大数据处理重点实验室,银川,750021)

摘　要：针对冗余属性和不相关属性过多对肺部肿瘤诊断的影响以及Pawlak粗糙集只适合处理离散变量而导致原始信息大量丢失的问题,提出混合信息增益和邻域粗糙集的肺部肿瘤高维特征选择算法(Information gain-neighborhood rough set-support vector machine,IG-NRS-SVM)。该算法首先提取3 000例肺部肿瘤CT图像的104维特征构造决策信息表,借助信息增益结果选出高相关的特征子集,再通过邻域粗糙集剔除高冗余的属性,通过两次属性约简得到最优的特征子集,最后采用网格寻优算法优化的支持向量机构建分类识别模型进行肺部肿瘤良恶性的鉴别。从约简和分类识别两个角度验证方法的可行性与有效性,并与不约简算法、Pawlak粗糙集、信息增益和邻域粗糙集约简算法进行对比。结果表明混合算法精确度优于其他对比算法,精确度达到96.17%,并且有效降低了时间复杂度,对肺部肿瘤计算机辅助诊断具有一定的参考价值。

关键词：信息增益;邻域粗糙集;支持向量机;肺部肿瘤;特征选择

中图分类号：TP391.4　　**文献标志码**:A

High-Dimensional Feature Selection Algorithm for Lung Tumors Based on Information Gain and Neighborhood Rough Set

LU Huiling[1], ZHOU Tao[1,2,4], ZHANG Feifei[3], HUO Bingqiang[2]

(1. School of Science, Ningxia Medical University, Yinchuan, 750004, China; 2. School of Computer Science and Engineering, North Minzu University, Yinchuan, 750021, China; 3. China Telecom Corporation Limited Ningxia Branch, Yinchuan, 750002, China; 4. Ningxia Key Laboratory of Intelligent Information and Big Data Processing, Yinchuan, 750021, China)

Abstract: Aiming at the influence of excessive redundant and unrelated attributes on the diagnosis of lung tumors and the fact that Pawlak rough set is only suitable for dealing with discrete variables and causing a large loss of original information, a high-dimensionality of lung tumors with mixed information gain and neighborhood rough set is proposed. The algorithm first extracts the 104-dimensional feature structure decision information table of 3 000 CT images of lung tumors. With the information gain result, the high correlation feature subset is selected, and the high redundancy attribute is eliminated by the neighborhood rough set. The optimal feature subset is obtained through two attribute reductions. Finally, the support vector machine optimized by the grid optimization algorithm is used to construct the classification recognition model to identify the benign and malignant lung tumors. The feasibility and effectiveness of the

基金项目:国家自然科学基金(61561040)资助项目;宁夏312人才计划资助项目;北方民族大学引进人才科研启动(2020KYQD08)资助项目。
收稿日期:2019-10-30;**修订日期**:2019-12-04

method are verified from the two aspects of reduction and classification, and compared with the non-reduction algorithm, Pawlak rough set, information gain and neighborhood rough set reduction algorithm. The results show that the accuracy of the hybrid algorithm is better than other comparison algorithms, the accuracy is 96.17%, and the time complexity is effectively reduced. It has certain reference value for computer-aided diagnosis of lung tumors.

Key words: information gain; neighborhood rough set; support vector machine; lung neoplasms; feature selection

引　言

肺癌的发病率及致死率在癌症中均居首位,肺癌的筛查和早期诊断可以有效预防和控制肺部肿瘤的进一步恶化,电子计算机断层扫描(Computed tomography,CT)可以清晰地反映组织器官的解剖结构,对于早期筛查具有重要意义[1]。虽然影像学检查为肺癌的诊断带来直观结构信息,但大量的影像数据也给临床医生带来沉重的工作负担,漏诊率和误诊率居高不下。以医学影像为基础的计算机辅助诊断系统(Computer aided diagnosis,CAD)借助计算机快速准确的智能化分析能力可以大幅度减轻主观因素带来的失误,提高工作效率。特征级融合是CAD分析过程中非常重要的一部分,可以有效提高辅助诊断的精度,同时从新的角度分析影像数据,与像素级和决策级处理形成互补,共同促进智能化诊断的发展。但是由于特征之间的相关性和冗余性使得“维数灾难”成为高维特征级融合过程中不可避免的问题,有效地特征选择算法可以降低维度和时间复杂度[2]。

信息增益(Information gain,IG)和粗糙集(Rough set,RS)是特征选择是常用的两种算法。IG是衡量包含或者不包含某个特征时为分类器提供信息量的指标,依次求出每个特征对分类器提供的信息量,然后从大到小进行排序,按照一定的规则取前K个特征,从而达到利用信息增益进行特征选择的目的。IG进行特征选择计算复杂度较低,只需单次运算,因此运行效率较高,可以有效剔除冗余、不相关以及噪声特征。但IG作为一种过滤式算法进行特征选择仍然存在问题,它只能衡量某一特征在整个系统中的重要程度,而无法区分该特征属于何种类别,而且并未考虑特征之间的关系,因此,对于不同类别使用相同特征分量的系统而言,采用IG进行特征选择能够取得较好的效果,而对于不同类别具有不同特征分量的系统而言,IG则无法进行特征选择。RS是处理不确定性数据的有效工具,因其无需先验知识的特性,广泛应用于特征选择[3]、模式识别[4]、数据挖掘与知识发现[5]等领域。RS研究的两个重要概念分别是概念近似和属性约简,其中属性约简是在不影响当前识别任务可辨性的前提下降低属性的维度,但是RS最初是在一定基础上构建的等价关系,在许多实际应用中都受到了限制。因此为了避免数据对单一方法的依赖以及更好地剔除数据集中的冗余和不相关属性,很多学者将IG的全局特征选择能力与RS优越的属性约简能力相结合进行高维特征选择,已经成功应用于情感分析[6]、房地产价目分析[7]、肿瘤诊断分类[8]和渔情预测[9]等。但是Pawlak RS只能处理名义型变量,实际应用中的数据往往是连续的数值变量,离散化后的数据集虽然可以适应RS算法等价类的构建,但是也可能会丢失重要信息,并且不同的离散化策略也会影响约简效果[10]。领域粗糙集(Neighborhood rough set,NRS)通过引入领域关系改进了Pawlak RS,可以直接对连续的数值型数据进行处理。IG和RS虽然都可以单独进行特征选择,但是存在一定的局限性,因此将两者的优势相结合进行特征选择具有一定的可行性,借助IG结果选出高相关的特征子集,再通过NRS剔除高冗余的属性,其中,NRS可以克服RS只适合处理离散变量而导致原始信息大量丢失的问题。通过两次属性约简得到最优的特征子集,能更好地剔除数据集中的冗余和不相关特征,提高算法的性能,降低时间复杂度,也可以避免数据对单一方法的依赖。

因此,本文综合IG和NRS的优势,提出一种混合高维特征选择算法。该算法首先计算所有特征的信息增益,挑选出信息增益值大于平均信息增益值的特征项,初步排除原始特征集合中相关性较小或者噪声特征,降低后续特征选择的时间复杂度。其次,将IG约简后的特征子集作为NRS的输入进行二次约简,得到最终的特征子集。最后利用支持向量机(Suppotr vector machines,SVM)作为分类器验证约简前后算法的性能。

1 基本原理

1.1 信息增益

信息增益通过分析增加或者删除某一特征前后系统信息量(熵)的变化多少来衡量特征的重要性,关于期望信息、熵和信息增益等概念的定义如下:

定义1[11] 假定S是数据样本,定义了m个不同类$C_i(i=1,2,\cdots,m)$,设S_i为C_i的样本数。a_j为属性A的值($j=1,2,\cdots,v$)。样本分类所需的期望信息为

$$I(S_1,S_2,\cdots,S_m)=-\sum_{i=1}^{m}\frac{S_i}{S}\log_2\frac{S_i}{S} \tag{1}$$

定义2 设用属性A将S划分为v个子集$\{S_1',S_2',\cdots,S_v'\}$,根据$A$划分成子集的期望信息,即熵为

$$E(A)=\sum_{j=1}^{v}\frac{S_j'}{S}\left[-\sum_{i=1}^{m}\frac{S_{ij}'}{S_j'}\log\frac{S_{ij}'}{S_j'}\right] \tag{2}$$

式中:S_{ij}'为子集S_j'中类C_i的样本数。

选择属性A的信息增益为

$$\mathrm{Gain}(A)=I(S_1,S_2,\cdots,S_m)-E(A) \tag{3}$$

1.2 邻域粗糙集

粗糙集(Rough set,RS)是一种分析具有不确定、不完整和不一致等不完备特性的软计算方法,揭示信息中存在的潜在规律,已经成功应用于近似推理、分析决策和模式识别等领域。为了不断完善RS的理论基础、拓展应用范围,专家学者们相继提出了很多改进算法,例如粒度RS、NRS、加权RS、覆盖RS、灰色RS、决策RS、模糊RS、优势RS等[12]。其中NRS是在Pawlak RS的基础上引入邻域关系而提出的处理连续型数据的模型,避免了Pawlak RS离散化过程中对原始信息的丢失[13],相关定义如下:

定义3 给定决策信息系统$S=(U,A,V,f)$,其中$U=(x_1,x_2,\cdots,x_n)$称为论域,表示全体样本构成的集合;$A=C\cup D,C=\{c_1,c_2,\cdots,c_n\}$表示条件属性构成的集合,$D=\{d_1,d_2,\cdots,d_n\}$表示决策属性构成的集合;$V=\bigcup_{a\in A}V_a$表示属性值域$V_a$的并集;$f:U\times A\to V$表示映射关系的信息函数。定义$x_i\in U$的邻域为

$$\delta_B(x_i)=\{x_j\,|\,x_j\in U,\Delta_B(x_i,x_j)\leqslant\delta\} \tag{4}$$

式中$\Delta_B(x_i,x_j)=\left(\sum_{i=1}^{N}\left|f(x_1,a_i)-f(x_2,a_i)\right|^p\right)^{1/p}$为$P$范数距离函数。

定义4[14] 邻域决策系统的上近似和下近似决策属性D将论域U划分为N个等价类$X_1,X_2,\cdots,X_N$,$B\subset A$生成U上的邻域关系,则决策属性D关于子集B的上、下近似及边界定义为

上近似

$$\overline{N_B}D=\bigcup_{i=1}^{N}\overline{N_B}X_i \tag{5}$$

下近似亦称为正域,记作$\mathrm{Pos}_B(D)$

$$\mathrm{Pos}_B(D)=\underline{N_B}D=\bigcup_{i=1}^{N}\underline{N_B}X_i \tag{6}$$

式中:$\overline{N_B}X=\left\{x_i\middle|\delta_B(x_i)\cap X\neq\varphi,x_i\in U\right\}$,$\underline{N_B}X=\left\{x_i\middle|\delta_B(x_i)\subseteq X,x_i\in U\right\}$。

边界域

$$BN(D)=\overline{N_B}D-\underline{N_B}D \tag{7}$$

负域:$\mathrm{Neg}_B(D)=U-\overline{N_B}D$。

定义5[15] 重要度定义特征c相对于A的重要度为

$$\mathrm{Sig}_B(c,B,D)=\gamma_{B\cup c}(D)-\gamma_B(D) \tag{8}$$

式中,$\gamma_B(\mathrm{D})=\dfrac{\left|\mathrm{Pos}_B(D)\right|}{|U|}$称为决策属性$D$相对于子集$B$的依赖度,属性重要度反映了条件属性对决策属性的贡献程度,取值0~1,值越接近于1表明属性越重要。

定义6 给定$\mathrm{NDT}^{\delta}=(U,N_C^{\delta}\cup D,f,V)(\delta>0)$,$B\subseteq C$,若$B$满足以下两个条件,则称$B$为$C$的一个约简:

(1)$\gamma_{N_B^{\delta}}=\gamma_{N_C^{\delta}}$;

(2)对于任一属性$a\in B$,$\gamma_{N_B^{\delta}}>\gamma_{N_{B-\{a\}}^{\delta}}$。

2 IG混合NRS的高维特征选择算法

2.1 算法流程

混合高维特征选择算法首先对肺部肿瘤图像进行预处理,提取特征构建原始特征库,归一化原始特征得到实验数据,计算所有特征的信息增益值,挑选信息增益值大于平均信息增益值的特征,得到第1次约简结果,即得到高相关的特征子集,再通过NRS剔除高冗余的属性,得到最终的特征子集,最后利用SVM作为分类器进行分类识别。本文算法流程如图1所示。

图1 算法流程图

Fig.1 Flow chart of proposed algorithm

2.2 算法步骤

(1)数据获取:收集宁夏某三甲医院核医学科肺部肿瘤CT图像3 000例,每例数据包括临床诊断结果、影像资料和检查所见等,根据医嘱中的临床结论对CT图像进行标签的设定(分为良性和恶性),为了避免数据量过少导致模型训练不充分,本文对肺部肿瘤只进行良恶性的分类,并不考虑肿瘤的分级[16]。最终得到1 500例恶性肺部肿瘤CT图像和1 500例良性肺部肿瘤CT图像;

(2)图像预处理:根据医嘱中的临床结论从相应的DICOM文件中获取肺部肿瘤CT图像,并且对应编号,去伪彩转化为灰度图像。从全局灰度图像中根据临床标记截取对应的病灶区域作为ROI图像,将ROI图像归一化成50像素×50像素大小的实验数据。数据预处理过程见图2所示。

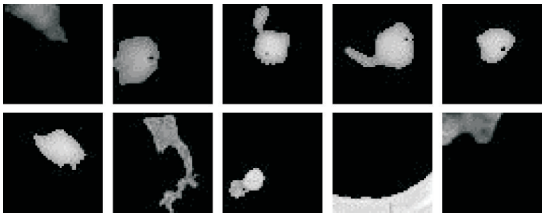


图2 肺部肿瘤图像预处理过程

Fig.2 Preprocessing process of lung tumor image

(3)图像分割:为了对肺部肿瘤图像的形状、纹理和灰度等特征进行全面、准确的测量,采用最大类间方差法(OTSU算法)[17]对预处理后的ROI区域进行分割,其基本原理是将直方图在某一阈值处分割成两组,一组对应背景,一组对应目标。如图3所示,给出本文分割前后的5组实例;图4给出两例患者的肺部肿瘤CT图像(图4(a)是恶性肺部肿瘤CT图像ROI区域,图4(b)是良性肺部肿瘤CT图像ROI区域)。

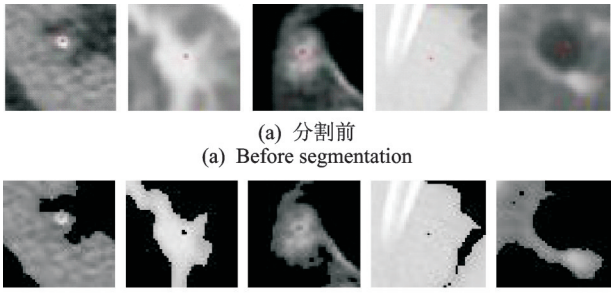


图3 OTSU算法分割前后肺部肿瘤ROI区域实例

Fig.3 Examples of ROI region of lung tumors before and after OTSU segmentation

(4)特征提取:提取分割后ROI图像目标区域的特征,包括形状特征、纹理特征和灰度特征共104维特征,具体特征见文献[18]中表1所示。构建的决策信息表大小为3 000×105,最后一列为决策属性,代表肿瘤的良恶性("-1"代表良性肿瘤,"1"代表恶性肿瘤)。表1给出图4所示的两例患者的肺部肿瘤ROI区域的104维特征。

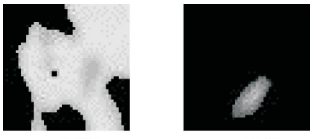


图4 肺部肿瘤CT图像ROI区域示例

Fig.4 Examples of ROI regions in CT images of lung tumors

表1 两例患者的肺部肿瘤ROI区域的104维特征值

Table 1 104-dimensional eigenvalues of ROI region of lung tumors in two patients

肿瘤类型	纹理特征	角点特征	Hu矩特征	小波特征			统计特征	几何特征	灰度共生矩阵特征			
				能量	范数	标准差			0°	45°	90°	135°
恶性肺部肿瘤CT图像ROI区域特征	12.616 8	22	0.218 4	1.000 0	675.529 5	60.220 7	83.789 6	99.000 0	0.175 4	0.154 6	0.163 6	0.151 5
	50.011 1		0.011 6	0.001 4	24.837 9	3.544 6	3 225.494 7	1 715.000 0	2.268 9	2.411 7	2.347 0	2.457 6
	0.453 7		0.000 9	0.006 4	53.948 0	7.706 9	56.793 4	2.198 9	1.045 3	1.511 5	1.286 9	2.049 1
			0.000 1	0.003 9	42.260 7	6.037 2	−0.660 3	0.729 2	0.083 8	0.082 2	0.082 5	0.080 1
			0.000 0	0.003 3	39.082 2	5.577 4	1.661 8	0.979 6	0.171 7	0.147 4	0.158 0	0.144 4
			0.000 0	0.001 6	26.623 5	3.803 0	1 423 674	1.000 0	1.045 3	1.511 5	1.286 9	2.049 1
			0.000 0	0.001 6	26.777 4	3.824 9	5.401 8		11.824 1	11.830 5	11.774 3	11.824 2
				0.002 7	34.840 0	4.976 1			1.985 4	2.055 8	2.019 6	2.068 2
									0.791 8	0.928 0	0.874 6	0.993 5
									141.271	139.468 7	139.605 1	138.594 4
									1.035 6	1.373 1	1.211 6	1.793 2
									−0.529 9	−0.441 2	−0.479 7	−0.411 4

续表

肿瘤类型	纹理特征	角点特征	Hu矩特征	小波特征			统计特征	几何特征	灰度共生矩阵特征			
				能量	范数	标准差			0°	45°	90°	135°
									0.948 7	0.941 5	0.812 8	0.913 2
									1.419 1	1.573 9	1.467 0	1.548 1
	10.966 4	14	0.594 4	0.996 3	332.563 7	35.621 4	17.698 8	56.000 0	0.620 7	0.610 9	0.615 5	0.609 7
	25.325 0		0.197 3	0.080 5	94.546 0	13.498 0	1 122.488 3	266.000 0	0.878 7	0.930 1	0.909 9	0.935 6
	0.376 7		0.029 9	0.006 6	27.023 8	3.860 5	33.503 6	1.065 9	0.338 4	0.631 8	0.525 7	0.674 3
良性			0.006	0.027 1	54.890 9	7.841 5	1.419 5	0.636 4	0.246 3	0.237 3	0.241 1	0.235 8
肺部			0.000 0	0.005 4	24.402 0	3.486 0	3.060 7	0.863 6	0.620 2	0.610 2	0.614 7	0.609 1
肿瘤			−0.000 2	0.003 0	18.187 1	2.598 0	577 300	1.000 0	0.338 4	0.631 8	0.525 7	0.674 3
CT			−0.000 1	0.003 8	20.665 6	2.952 0	2.068 0		4.046 1	4.035 4	4.031 0	4.035 4
图像				0.004 6	22.622 2	3.231 7			0.802 8	0.851 0	0.825 6	0.861 2
ROI									0.327 0	0.397 9	0.392 1	0.390 1
区域									25.714	24.991 9	25.210 9	24.884 2
特征									0.363 9	0.640 0	0.546 3	0.676 7
									−0.693 5	−0.611 3	−0.639 0	−0.603 1
									0.921 3	0.911 2	0.895 2	0.953 5
									2.251 0	2.395 9	2.364 5	2.420 5

(5)特征归一化:为了得到精确的数据处理结果,对原始数据进行归一化消除数据数量级和量纲的差异,本文采用常用的最大最小值法,使得归一化后的数据均落在[0,1]之间,其公式为

$$f(x_i)=\frac{x_i-x_{\min}}{x_{\max}-x_{\min}} \qquad i=1,2,\cdots,n \tag{9}$$

式中$x_{\max}$和$x_{\min}$分别为样本数组的最大值和最小值。

(6)基于信息增益的特征选择。

输入:连续型决策信息表$S=(U,A,V,f)$,其中$U=(x_1,x_2,\cdots,x_n)$称为论域,表示全体样本构成的集合;$A=C\cup D$,C表示条件属性构成的集合(即步骤(5)中经过归一化处理的104维特征集合),D表示决策属性构成的集合(即肺部肿瘤的良恶性,用数字“1”代表恶性肺部肿瘤,“−1”代表良性肺部肿瘤);V表示属性值域的并集;f表示映射关系的信息函数;

输出:约简后的属性集red_1;

步骤:①初始化集合$\mathrm{red}_1=\varphi$,计算每个条件属性的信息增益$\mathrm{Gain}(C_i)$,计算每个条件属性信息增益的平均值average;

②选择信息增益最大的属性c_i,$\mathrm{red}_1=\mathrm{red}_1\cup\{c_i\}$,并在条件属性集$C$中去掉该属性;

③如果信息增益最大的属性c_i的信息增益值小于平均值average,则停止得到red_1,否则调至步骤②。

(7)基于邻域粗糙集的特征选择:NRS属性约简是在不影响决策系统本身决策能力的前提下删除冗余属性,其约简算法使用的是前向贪心算法[19],其主要算法模型见图5,其主要步骤如下:

输入:信息增益约简后的属性集合$\mathrm{red}_1=(U,A',V,f)$,其中$A'=C'\cup D$,$C'$表示步骤(6)中信息增益值大于等于平均信息增益值的条件属性的集合,确定邻域半径δ的集合,设置重要度下限为0.001;

输出:两次约简集合red;

步骤：①初始化两次约简集合red＝φ，样本smp＝U；

②对$\forall a_i \in (C'-\text{red})$利用式(6)计算正域$\text{Pos}_{\text{red}+a_i}^{\text{smp}}(D)$；

③对于$a \in B$，选择a_k使得正域$\text{Pos}_{a_k}(D)$最大；

④利用式(8)计算属性重要度$\text{Sig}(a_k, \text{red}, D)$；

⑤若$\text{Sig}(a_k, \text{red}, D)$大于设定的重要度下限值，则输出约简结果red，结束程序，否则，记录k值，令：red＝red＋a_k，$I = I - \text{Pos}_{a_k}$，然后返回步骤②继续计算，直至输出约简结果red。

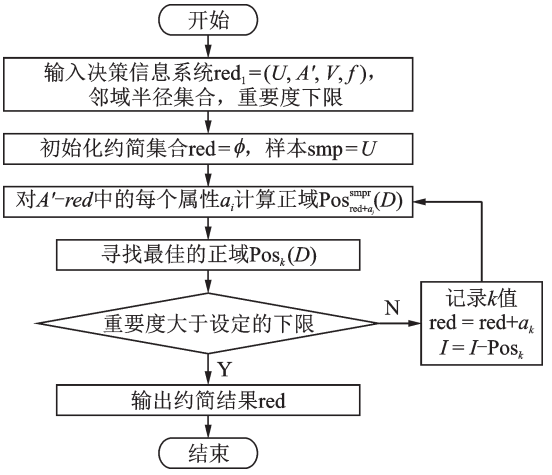


图5 前向贪心算法流程图

Fig.5 Flow chart of forward greedy algorithm

(8)基于SVM的分类识别。

SVM是有监督的数据挖掘算法，其基本思想是使用结构风险最小化原理在属性空间构建最优决策超平面对样本进行分割，使得分类器达到全局最优，可以较好地解决小样本、过学习、高纬度和局部极值等实际应用难题，具有很强的泛化能力与分类识别能力[20]，可以有效解决以医学影像为基础的CAD诊断过程中“非线性、高维度”的难题。因此，本研究采用SVM对两次约简的结果进行分类识别，其中核函数选择径向基核函数(Radial basis function，RBF)，采用网格寻优算法(Grid search，GS)优化SVM的c和g参数，构建分类识别模型进行肺部肿瘤良恶性的鉴别。

3 仿真实验

3.1 实验环境

(1)硬件环境：处理器是Inter(R) Core(TM)i3-2130 CPU @3.40 GHz，安装内存4.00 GB，500 GB内存。

(2)软件环境：Matlab R2012a，LibSVM，Windows 7操作系统。

3.2 分类器评价指标

早期诊断准确性的性能评价包括敏感性和特异性2大指标，但是这2个指标很难全面描述分类器的整体性能。因此，本文对约简阶段评价指标为约简长度，分类识别阶段的评价指标包括：精确度(Accuracy)、灵敏性(Sensitivity)、特异性(Specificity)、F值(F-score值)、马修斯相关性系数(Matthews correlation coefficient，MCC)、平衡F分数(Balanced F Score，F1Score)、约登指数(Youden index，YI)和时间(Time)，具体介绍如下：

精确度(Accuracy)是最常见的评价指标，精确度越高，分类器性能越好，计算公式为

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

灵敏性(Sensitivity)和特异性(Specificity)，分别用来衡量分类器对正例和负例的识别能力，值越大，识别性能越高，计算公式为

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{11}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{12}$$

F值是查全率与查准率加权调和平均，用来权衡精确度和召回率，计算公式为

$$F=\frac{2\times TP}{2\times TP+FP+FN} \tag{13}$$

MCC是描述实际分类与预测分类之间的相关系数,全面考虑了真阳性、真阴性、假阳性和假阴性,是一种更加均衡的指标,它的取值范围是[−1,1],值越接近于1表示受试对象的预测越准确,计算公式为

$$\mathrm{MCC}=\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{14}$$

F1Score是统计学中衡量二分类模型精确度的一种更为全面的评价指标,是准确率和召回率的一种加权平均,它的取值范围为[0,1],值越接近1代表模型的精确度越高,计算公式为

$$\mathrm{F1Score}=\frac{TP}{2TP+FN+FP}+\frac{TN}{2TN+FP+FN} \tag{15}$$

YI又称正确指数,用灵敏度与特异度之和减1的值表示,它的取值范围为[0,1],值越接近1,模型预测的真实性越好,计算公式为

$$\mathrm{YI}=\mathrm{Sensitivity}+\mathrm{Specificity}-1 \tag{16}$$

式中:TP代表实际为恶性并且被正确预测为恶性的样本数,FP代表实际为良性但是被错误预测为恶性的样本数;TN代表实际为良性并且被正确地划分为良性的样本数;FN代表实际为恶性但是被错误预测为良性的样本数。

3.3 实验结果及分析

经过IG的和NRS两阶段约简后的特征子集,从理论层面可以有效降低原始决策信息表的维度,时间复杂度和空间复杂度。通过IG可以初步筛选降低数据噪声,剔除相关性较小的属性,经过NRS二次约简可以有效剔除高冗余的属性。为了进一步验证文本提出的两阶段约简高维特征选择算法的可行性和有效性,以3 000例(1 500例良性,1 500例恶性)肺部肿瘤CT图像为研究对象,获取ROI区域后分别提取形状、纹理和灰度特征共104维构造原始特征集合,采用IG和NRS进行两阶段约简,约简结果利用SVM进行分类识别。

实验1 不同约简算法约简过程的比较

利用不同的算法对原始决策信息表进行约简,并采用约简长度比较不同算法的优劣,具体结果见表2和图6所示。

表2 不同约简算法约简过程的比较

Table 2 Comparison of reduction processes of different reduction algorithms

约简算法	约简后剩余属性	约简长度
Pawlak RS	{c1,c2,c3,c4,c6,c11,c12,c16,c19,c20,c21,c22,c24,c25,c27,c36,c37,c38,c39,c43,c45,c46,c47,c50,c51,c56,c57,c59,c60,c64,c72,c74,c76,c77,c82,c85,c86,c90,c95,c98,c99}	41
IG	{c1,c2,c5,c36,c37,c38,c39,c40,c41,c42,c44,c45,c46,c49,c50,c51,c52,c53,c54,c55,c56,c57,c58,c59,c61,c62,c63,c64,c65,c66,c67,c68,c69,c70,c71,c72,c74,c75,c76,c77,c78,c79,c80,c81,c82,c83,c84,c85,c87,c88,c89,c90,c91,c92,c93,c94,c95,c96,c97,c98,c100}	61
NRS	{c1,c2,c3,c4,c12,c28,c37,c43,c46,c47,c55,c56,c60,c61,c63,c70,c73,c74,c86,c87,c89,c99,c100}	23
IG-NRS	{c1,c2,5,c37,c38,c39,c42,c46,c49,c50,c51,c52,c53,c55,c56,c57,c59,c61,c62,c63,c64,c65,c66,c69,c70,c72,c74,c75,c79,c83,c85,c87,c88,c89,c92,c95,c96,c98,c100}	39

从表3和图6可见，采用不同算法对原始决策信息表进行约简后相比不约简时，信息表的维度都有较大程度的降低，本文算法的约简长度仅高于NRS约简算法，相比原始信息表维度降低65维。

实验2 不同约简算法分类结果的比较

对实验一不同算法的约简结果分别利用SVM五折交叉(即每次从1 500例良(恶)性肿瘤图像中选取300例作为测试集，其余1 200例作为训练集)进行分类识别，从精确度、敏感度、特异性、*F*值、MCC、F1Score、Youden和总时间8个指标评价算法的优劣。具体结果见表3。

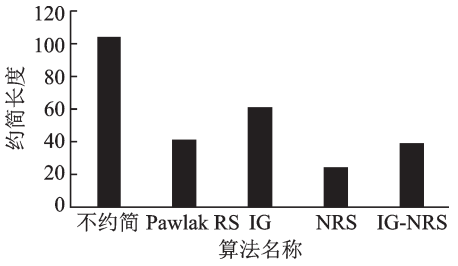


图6 不同约简算法约简结果(约简长度)的对比

Fig.6 Comparison of reduction results (reduction length) of different reduction algorithms

表3 不同算法分类结果的比较

Table 3 Comparison of classification results by different algorithms

算法名称	五折交叉次数	精确度/%	敏感度/%	特异性/%	*F*值	MCC	F1Score	Youden	总时间/s
SVM	1	93.00	91.67	94.33	0.929 1	0.860 3	0.930 0	0.860 0	258.520 9
	2	96.17	94.33	98.00	0.961 0	0.924 0	0.961 7	0.923 3	262.402 9
	3	95.17	95.67	94.67	0.951 9	0.903 4	0.951 7	0.903 3	263.297 6
	4	96.50	95.67	97.33	0.964 7	0.930 1	0.965 0	0.930 0	259.020 5
	5	98.00	96.67	99.33	0.979 7	0.960 3	0.980 0	0.960 0	287.365 5
	平均值	95.77	94.80	96.73	0.957 3	0.915 6	0.957 7	0.915 3	266.121 5
Pawlak RS-SVM	1	91.33	90.67	92.00	0.912 8	0.826 7	0.913 3	0.826 7	103.635 6
	2	95.67	94.00	97.33	0.955 9	0.913 8	0.956 7	0.913 3	108.421 2
	3	95.33	95.67	95.00	0.953 5	0.906 7	0.953 3	0.906 7	92.429 4
	4	96.33	95.33	97.33	0.963 0	0.926 9	0.963 3	0.926 7	96.925 1
	5	98.00	96.67	99.33	0.979 7	0.960 3	0.980 0	0.960 0	100.656 5
	平均值	95.33	94.47	96.20	0.953 0	0.906 9	0.953 3	0.906 7	100.413 6
IG-SVM	1	93.83	91.33	96.33	0.936 8	0.877 8	0.938 3	0.876 7	83.656 8
	2	95.83	95.00	96.67	0.958 0	0.916 8	0.958 3	0.916 7	83.758 7
	3	96.33	96.00	96.67	0.963 2	0.926 7	0.963 3	0.926 7	87.776 5
	4	96.67	96.67	96.67	0.966 7	0.933 3	0.966 7	0.933 3	89.170 3
	5	97.33	96.67	98.00	0.973 2	0.946 8	0.973 3	0.946 7	84.395 9
	平均值	96.00	95.13	96.87	0.959 6	0.920 3	0.960 0	0.920 0	85.751 6
NRS-SVM	1	93.50	92.67	94.33	0.934 5	0.870 1	0.935 0	0.870 0	80.852 7
	2	95.17	93.33	97.00	0.950 8	0.903 9	0.951 7	0.903 3	67.172 7
	3	95.67	95.33	96.00	0.956 5	0.913 4	0.956 7	0.913 3	65.589 7
	4	97.00	95.67	98.33	0.969 6	0.940 3	0.970 0	0.940 0	69.191 3
	5	98.17	97.33	99.00	0.981 5	0.963 5	0.981 7	0.963 3	69.048 6
	平均值	95.90	94.87	96.93	0.958 6	0.918 2	0.959 0	0.918 0	70.371 0
IG-NRS-SVM	1	94.17	91.67	96.67	0.940 2	0.884 4	0.941 6	0.883 3	58.558 5
	2	95.17	93.33	97.00	0.950 8	0.903 9	0.951 7	0.903 3	57.996 7
	3	96.33	95.33	97.33	0.963 0	0.926 9	0.963 3	0.926 7	60.605 6
	4	97.17	96.67	97.67	0.971 5	0.943 4	0.971 7	0.943 3	62.214 1
	5	98.00	96.67	99.33	0.979 7	0.960 3	0.980 0	0.960 0	72.177 0
	平均值	96.17	94.73	97.60	0.961 0	0.923 8	0.961 7	0.923 3	62.310 4

由表4可见,同种算法不同交叉次数各评价指标存在差异,为了全面衡量算法的性能,采用五折交叉的平均值作为该算法的最终分类结果。除过Pawlak RS-SVM,本文算法敏感度相比其他算法有较小程度的降低,其他指标都优于其他算法,精确度、特异性、F值、MCC、F1Score和Youden分别提高0.17%～0.83%、0.67%～1.4%、0.001 5～0.008 1、0.003 5～0.016 9、0.001 7～0.008 3和0.003～0.016 7,时间降低8.06～203.81 s。由于精确度和时间是最常用的评价指标,为了更加清晰地表示不同算法在精确度和时间2个指标上的差异,将这2个指标的平均值绘制柱状图,分别见图7,8所示。

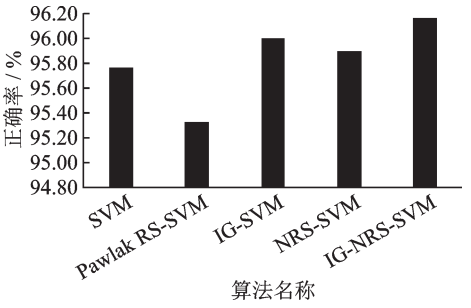


图7 不同算法分类精确度的比较

Fig.7 Comparison of classification accuracy of different algorithms

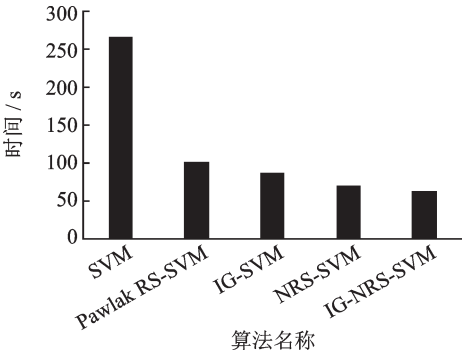


图8 不同算法分类时间的比较

Fig.8 Comparison of classification time by different algorithms

由图7和图8可见,本文算法的精确度最高,Pawlak RS-SVM模型的精确度最低。因为Pawlak RS的理论基础是等价关系的划分,因此只能分析离散型的数据,无法分析连续型数据。Pawlak RS在分析连续型数据之前必须进行离散化处理,不仅增加了算法的时间复杂度,也会缩小数据之间的差异性,很大程度会影响最终的约简结果。离散化方法不同,离散化结果也不尽相同,给Pawlak RS约简带来一定的影响[21],加之离散化后的特征集合无法全面刻画肺部肿瘤ROI区域。本文算法的时间复杂度相比不约简时降低4.27倍,也低于其他对比算法,由此可见本文算法可以提高肺部肿瘤高维特征选择算法的精确度,有效降低算法的时间复杂度,具有一定的推广价值。

本文分类阶段采用五折交叉,每折交叉测试集为600例,包括300例良性,300例恶性,按照表4的实验结果可知,不同算法五折交叉结果精确度范围为91.33%～98.17%,则分类错误的样本数范围为11～52个,由于篇幅限制,只列举每个算法第5折交叉实验结果中分错的样本示例(对3 000例实验样本进行编号,其中恶性肿瘤1 500例,编号为1～1500,良性肿瘤1 500例,编号为1501～3000。则第5折交叉的测试集样本编号为:恶性肿瘤样本编号为1201～1500,良性肿瘤样本编号为2701～3000),结果见表4,其中"样本编号"列中的数字代表不同算法中被错误分类的样本编号;"分类错误的肺部肿瘤ROI区域"是指不同算法中被错误分类的样本,即分割后的肺部肿瘤ROI区域。

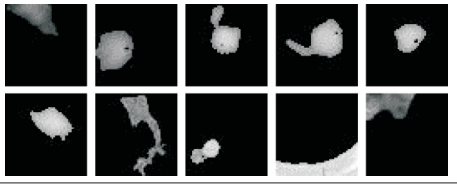
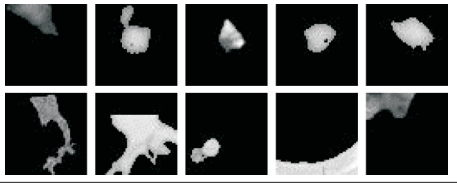
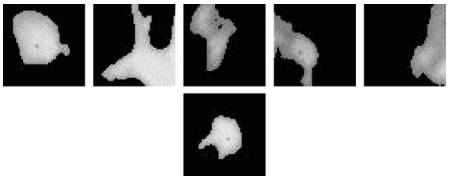
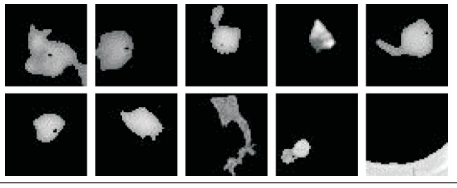
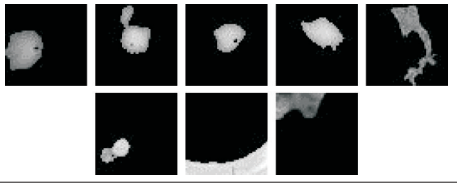
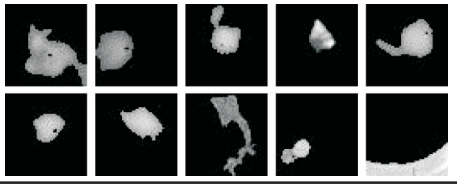
由表3可见,不同算法第5折交叉实验结果被错误分类的样本具有以下特点:

(1)恶性肿瘤被错分的样本数量大于良性肿瘤,可能的原因是恶性肿瘤较为复杂,具有不同的分级,不同等级肿瘤的形状、纹理和灰度特征也不尽相同。本文为了增大样本量,并未考虑肿瘤的分级,因此,恶性肿瘤中的不同等级的样本差异性较大,被错分的概率也较大;

(2)被错分的恶性肿瘤病灶部位所占像素普遍小于被错分的良性肿瘤,可能的原因是恶性肺部肿瘤相对良性肿瘤增长较快,表现在CT图像中就是在ROI区域中所占像素相对较大,如图4所示,恶性肿瘤像素占整个ROI区域的比例超过50%,而良性所占比例较小。图5中被错分的恶性肿瘤样本ROI区域中病灶部位所占像素几乎都小于50%,而良性肿瘤病灶像素所占比例相对分类正确的较大,因此,被错分的良性肿瘤病灶所占比例普遍大于恶性肿瘤;

表4 不同算法第五折交叉分类错误的肺部肿瘤ROI区域

Table 4 ROI Regions of pulmonary tumors with fifth-fold cross classification error in different algorithms

算法名称	样本种类	样本编号	分类错误的肺部肿瘤ROI区域
SVM	实际为良性被错分为恶性的样本	2787、2983	
	实际为恶性被错分为良性的样本	1220、1221、1241、1260、1263、1296、1349、1378、1412、1489	
Pawlak RS-SVM	实际为良性被错分为恶性的样本	2869、2983	
	实际为恶性被错分为良性的样本	1220、1241、1259、1263、1296、1349、1369、1378、1412、1489	
IG-SVM	实际为良性被错分为恶性的样本	2815、2817、2869、2955、2966、2983	
	实际为恶性被错分为良性的样本	1201、1221、1241、1259、1260、1263、1296、1349、1378、1412	
NRS-SVM	实际为良性被错分为恶性的样本	2815、2869、2983	
	实际为恶性被错分为良性的样本	1221、1241、1263、1296、1349、1378、1412、1489	
IG-NRS-SVM	实际为良性被错分为恶性的样本	2815、2817	
	实际为恶性被错分为良性的样本	1201、1221、1241、1259、1260、1263、1296、1349、1378、1412	

(3)不同算法中被错误分类的样本存在交叉,恶性肿瘤样本交叉较多,本文数据处理过程是根据医嘱中临床结论来标记肿瘤良恶性的,存在一定的误差,可能的原因一方面是恶性肿瘤的分级,另一方面是临床医生的判定存在一定的漏诊和误诊,因此,不同算法中多次出现相同的被分类错误的样本。

4 结束语

为了提高肺部肿瘤计算机辅助诊断算法的性能,分析了IG和NRS的优缺点,提出一种混合IG和NRS的肺部肿瘤高维特征选择算法,从理论层面分析两阶段约简算法的可行性。为了验证算法的有效性,提取3 000例肺部肿瘤CT图像的104维特征构造决策信息表,借助IG和NRS两次属性约简得到最优的特征子集,最后采用SVM进行分类识别。通过与不约简算法、Pawlak RS、IG和NRS约简算法进行比较可知,该算法可以提高算法的精确度,有效降低时间复杂度,并且综合对比不同方法构建的肺部肿瘤高维特征选择算法的性能,确保本文方法的优越性,从模型方法的逐步选择上保证结果的科学性,对肺部肿瘤计算机辅助诊断具有一定的参考价值。

本文提出的算法虽然在一定程度上提高了算法的精确度,降低了时间复杂度,但是仍然存在一些不足之处。(1)由于医学图像的病灶部位在整个图像中所占像素比例很小,很有可能就是几个到十几个像素,因此特征级处理之前必须获取ROI并提取特征才能更好地刻画病灶区域。本文的ROI区域通过手动获取,可能会存在人为因素导致的误差,因此后期的研究可以考虑结合医嘱,让计算机自动识别ROI区域;(2)算法的性能有待进一步提高,后期可以改善实验环境,配备高性能的服务器或者利用云平台降低由于客观因素导致的时间复杂度。

参考文献:

[1] 张忠凤,张春. 多层螺旋CT对肺癌的诊断价值[J]. 实用临床医学,2016,17(1): 53-54.
ZHANG Zhongfeng, ZHANG Chun. Multi-slice spiral CT in diagnosis of lung cancer[J]. Practical Clinical Medicine, 2016, 17(1): 53-54.

[2] GAO Wanfu, HU Liang, ZHANG Ping, et al. Feature selection by integrating two groups of feature evaluation criteria[J]. Expert Systems with Applications, 2018, 110: 11-19.

[3] YUE Xiaodong, CHEN Yufei, MIAO Duoqian, et al. Tri-partition neighborhood covering reduction for robust classification[J]. International Journal of Approximate Reasoning, 2017, 83: 371-384 .

[4] WANG Qi, QIAN Yuhua, LIANG Xinyan, et al. Local neighborhood rough set[J]. Knowledge-Based Systems, 2018, 153: 53-64.

[5] LI Huaxiong, ZHANG Libo, ZHOU Xianzhong, et al. Cost-sensitive sequential three-way decision modeling using a deep neural network[J]. International Journal of Approximate Reasoning, 2017, 85: 68-78.

[6] 蒲国林. 基于粗糙集与信息增益的情感特征选择方法[J]. 微电子学与计算机,2016,33(1): 96-99.
PU Guolin. A sentiment feature selection method based on rough set and information gain[J]. Microelectronics & Computer, 2016, 33(1): 96-99.

[7] 徐分,蒋芸,王勇,等. 基于粗糙集和信息增益的属性约简改进方法[J]. 计算机工程与设计, 2009,30(24): 5698-5700.
XU Fen, JIANG Yun, WANG Yong, et al. Improved algorithm for attribute reduction based on rough sets and information gain [J]. Computer Engineering and Design, 2009, 30(24): 5698-5700.

[8] DAI Jianhua, XU Qing. Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification[J]. Applied Soft Computing, 2013, 12(1): 211-221.

[9] 徐立萍,姜志旺. 基于粗糙集及信息增益的数据挖掘预测算法[J]. 中国科技论文,2012,7(7): 552-555,559.
XU Liping, JIANG Zhiwei. A new data mining and prediction algorithm based on rough set and information gain[J]. China Science Paper, 2012, 7(7): 552-555, 559.

[10] CHEN Yumin, ZHANG Zunjun, ZHENG Jianzhong, et al. Gene selection for tumor classification using neighborhood rough

sets and entropy measures[J]. Journal of Biomedical Informatics, 2017, 67: 59-68.

[11] SZIDÓNIA L, LÁSZLÓ L. Gabor feature selection based on information gain[J]. Procedia Engineering, 2017, 181: 892-898.

[12] 张飞飞,周涛,陆惠玲,等. 特征级融合方法及其在医学图像方面的应用[J]. 计算机应用与软件,2019,36(4): 1-9,45.
ZHANG Feifei, ZHOU Tao, LU Huiling, et al. Feature-level fusion and its application in medical image[J]. Computer Applications and Software, 2019, 36(4): 1-9, 45.

[13] 桑秀丽,李哲,吕梁. 基于NRS与改进LS-SVM解析模型的乳腺肿瘤分类诊断[J]. 统计与决策,2017,1: 84-86.
SANG Xiuli, LI Zhe, LÜ Liang. Classification diagnosis of breast tumor based on NRS and improved LS-SVM analytical model [J]. Statistics & Decision, 2017, 1: 84-86.

[14] LIU Jinghua, LIN Yaojin, LI Yuwen, et al. Online multi-label streaming feature selection based on neighborhood rough set[J]. Pattern Recognition, 2018, 84: 273-287.

[15] YANG Xibei, LIANG Shaochen, YU Hualong, et al. Pseudo-label neighborhood rough set: Measures and attribute reductions [J]. International Journal of Approximate Reasoning, 2019, 105: 112-129.

[16] 梁蒙蒙,周涛,夏勇,等. 基于PSO-ConvK卷积神经网络的肺部肿瘤图像识别[J]. 山东大学学报:工学版,2018,48(5): 77-84.
LIANG Mengmeng, ZHOU Tao, XIA Yong, et al. Lung tumor images recognition based on PSO-ConvK convolutional neural network[J]. Journal of Shandong University: Engineering Science, 2018, 48(5): 77-84.

[17] 袁小翠,吴禄慎,陈华伟. 基于Otsu方法的钢轨图像分割[J]. 光学精密工程,2016,24(7): 1772-1781.
YUAN Xiaocui, WU Lushen, CHEN Huawei. Rail image segementation based on Otsu threshold method[J]. Optics and Precision Engineering, 2016, 24(7): 1772-1781.

[18] 张飞飞,周涛,陆惠玲,等. 基于贝叶斯粗糙集的肺部肿瘤CT图像高维特征选择算法[J]. 生物医学工程研究,2018,37(4): 404-409.
ZHANG Feifei, ZHOU Tao, LU Huiling, et al. An algorithm for high dimension feature selection of lung tumor CT image based on Bayesian rough set[J]. Journal of Biomedical Engineering Research, 2018, 37(4): 404-409.

[19] CHEN Yumin, ZHANG Zunjun, ZHENG Jianzhong, et al. Gene selection for tumor classification using neighborhood rough sets and entropy measures[J]. Journal of Biomedical Informatics, 2017, 67: 59-68.

[20] 田德红,何建敏,张保强. 基于NRS-SVM模型的航空弹药消耗预测研究[J]. 南京航空航天大学学报,2018,50(5): 666-671.
TIAN Dehong, HE Jianmin, ZHANG Baoqiang. Research on aviation ammunition consumption prediction based on NRS-SVM model[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2018, 50(5): 666-671.

[21] 晏伟峰. 邻域粗糙集及其基于邻域粗糙集的分类算法[D]. 南昌: 江西师范大学,2011.
YAN Weifeng. Neighborhood rough set and its classification algorithm based on neighborhood rough set[D]. Nanchang: Jiangxi Normal University, 2011.

作者简介:

陆惠玲(1976-),女,副教授,研究方向:医学数字图像处理、智能算法。

周涛(1977-),男,教授,研究方向:医学数字图像处理、智能算法,E-mail: zhoutaonxmu@126.com。

张飞飞(1991-),女,硕士研究生,研究方向:粗糙集、医学图像特征级融合。

霍兵强(1994-),男,硕士研究生,研究方向:残差神经网络、医学图像处理。

(编辑:陈珺)