不平衡数据集的 CT 结肠镜息肉检测方法

熊 馨, 徐礼胜, 王春武, 康 雁

(东北大学 中荷生物医学与信息工程学院, 110819 沈阳)

摘 要: 目前 CT 结肠镜的息肉检测分类器面临着数据集不平衡问题,数据集中的正样本(息肉)的数量远远小于负样本.针对这个问题,息肉检测分类器采用 SMOTEBoost,结合 SMOTE(Synthetic Minority Over-Sampling Technique)和 Boosting:在数据层面,采用过采样技术 SMOTE 合成少数类样本,减轻数据集中两类样本的不平衡程度;在算法层面,采用 Boosting 方法提高弱分类器的性能,两者结合起来,既改善对少数类样本的预测能力,又保证了对整个数据集的分类精度.为了满足息肉检测对算法实时性的需求,采用 MRMR(Minimum Redundancy Maximum Relevance)方法挑选最大相关、最小冗余的简单特征组成级联第 1 层强分类器,拒绝大多数负样本,极大地提高了分类器的处理速度.实验结果表明:设计的分类器检测直径大于 5 mm 息肉的敏感度达到 90%,每个数据体 6 个假阳.

关键词: 不平衡数据集; CT 结肠镜; 结肠息肉检测; 重采样; Boosting; Cascade AdaBoost

中图分类号: R814 **文献标志码**: A **文章编号**: 0367-6234(2013)11-0112-06

Polyp detection in CT colonography based on imbalanced data sets

XIONG Xin, XU Lisheng, WANG Chunwu, KANG Yan

(School of Sino-Dutch Biomedical & Information Engineering, Northeast University, Shenyang 110819, China)

Abstract: Polyp detection in CT Colongraphy suffers from imbalanced data sets where negative samples (non-polyp) are dominant. In data level, SMOTE (Synthetic Minority Over-Sampling Technique) was applied to alleviate imbalanced degree by synthetic minority samples. In algorithm level, Boosting approach was employed in order to improve classification performance. Having combined Boosting with SMOTE (SMOTEBoost), the proposed classifier not only improved the prediction of the minority samples, but also guaranteed the accuracy over the entire data set. To satisfy real-time requirements for polyp detection, MRMR (Minimum Redundancy Maximum Relevance) was provided to select low-cost simple features for training the first stage of cascade, resulting in refusing the great majority negative samples and speeding procession. The experimental results showed that the classifier could achieve an overall per-polyp sensitivity of 90% (corresponding to the polyp whose diameter is equal to or greater than 5 mm), with false positives of 6 per volume on average.

Key words: imbalanced data sets; CT colonography; colonic polyps' detection; re-sampling; Boosting; Cascade AdaBoost

在欧美,结肠癌是第 2 大癌症相关致死原因[1].而在我国,结肠癌的发病率与死亡率虽然低于胃癌、食管癌、肺癌等常见恶性肿瘤,但是随着人民生活水平的提高、饮食结构的改变,其发病率呈逐年上升趋势.结肠癌通常是由息肉发展而来,及时的检测和清除息肉已被证明能够有效的减少结肠癌死亡率[2].最近几年 CT 结肠镜(CTC)获得广泛关注,是近年来体检及肿瘤早期发现的主要手段[3].但是多层螺旋 CT 扫描的数据量很大,图像层数达 800~2 000(具体的层数取决于病人的身高和 CT 的分辨率).如此大量的数据视觉检查是一个耗时的过程,而且检查结果并不总是可重现的.而采用计算机辅助检测(CAD)技术,自动检测、标记息肉的位置,使检测结果具有可重复性,

收稿日期: 2012-11-12.
基金项目: 国家自然科学基金资助项目 (61071213).
作者简介: 熊 馨(1984—),女,博士研究生;
徐礼胜(1975—),男,教授,博士生导师;
康 雁(1964—),男,教授,博士生导师.
通信作者: 熊馨, xiongxin840826@163.com.

减轻了读片医生的负担.目前 CAD-CTC 系统快速地成熟起来,在实验室环境中,检测具有临床重要性的息肉(直径≥5 mm)[4-5]的敏感度达到85~100%,每个患者的假阳小于 10 例[6-9],这样的性能甚至超过了医生.

这些方案先进行结肠分割,再提取疑似区域,然后由分类器判别这些疑似区域是否为息肉.通常一例数据中最多只有几个息肉,而提取的疑似区域则多达几百个,即正样本和负样本分布严重不平衡.这些息肉检测方案中基本采用标准分类学习算法,而大多数标准分类学习算法假设数据集分布平衡,对于稀疏事件关注比较少,错分少数类比多数类频繁[10].针对结肠 CAD 中正样本和负样本不平衡的问题也有一些研究:文献[9]采用 Cascade AdaBoost 结构,在前面几层拒绝大部分负样本[9];Jinbo Bi 等[11]采用线性规划 Cascade 结构,利用列生成 boosting 来解决线性规划问题,设计的分类器检测肺结节和结肠息肉都达到不错的结果;Dijia Wu 等[12]针对同一病变结构产生多个阳性疑似区域的问题,将 Cascade 和多对象学习结合起来,提出 Min-Max Cascade 框架;Yalin Zheng 等[13]利用带权重的 SVM 检测息肉,给分类错误的少数类样本(即息肉)赋值上更大的权重.

1　方　法

1.1　过采样方法

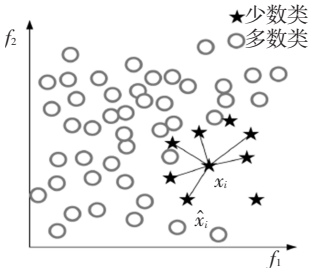
对于少数类样本进行过采样(over-sampling),增加少数类样本数量,称为过采样或上采样(up-sampling).简单的过采样随机地复制少数类样本,直到其所占比例与多数类基本平衡.这种方法表面看起来很公平,但是暗藏问题,因为随机过采样是简单复制原始数据集,会使得其中某一类样本数量过多[14],导致过拟合.而合成采样则不存在这样的问题,SMOTE 是合成采样一个很有效的方法,在各个领域中都有很成功的应用[15].该方法为每个少数类样本确定其 K 最邻近样本,然后在样本与其近邻样本的连线上随机取点,从而生成新的少数类样本.

图 1 说明了 SMOTE 过程.图 1(a)是一个典型的不平衡数据分布,其中:x_i 为选中的一个少数类样本;$\hat{x}_i$ 为 K 的 K(这里 $K=6$) 邻域样本之一. 图 1(b)表明沿着 x_i 和 $\hat{x}_i$ 直线上生成一个合成样本.这样的合成方法避免了某一类样本过于集中,扩大了原始样本集,能够极大的改善分类学习效果.

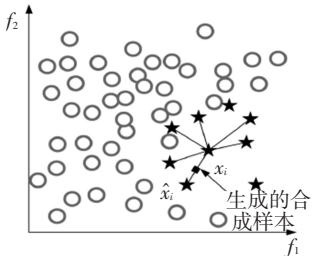
在计算少数类样本最邻近邻域时,连续属性和离散属性需要分开讨论,对连续属性采用欧几里得距离,而对于离散属性采用值距离(Value Difference Metric,VDM)[16].因而合成少数类样本的过程如下:

1)对于连续属性:①计算该样本(少数类样本)与其少数类 k 近邻中的一个样本间的相异度;②用 0 和 1 之间的一个随机数乘以上述相异度;③将得到的相异度加到该样本上,从而创建一个新的少数类样本.

2)对于离散属性:①在该样本与其 K 近邻间采用多数表决方法选取属性值;②将此属性值赋值给新的合成少数类样本.



(a)不平衡数据分布



(b)生成的合成样本

图 1　SMOTE 过程示意图

1.2　Boosting 方法

Boosting 方法是集成学习方法的一种,通过 Boosting 迭代创建集成模型,能够提升弱分类器的性能[17].但是,在标准的 Boosting 算法中,对错分的正样本和负样本给予相同权重,而样本池中负样本远远多于正样本,这样随后训练的弱分类器就偏向负样本,对于正样本的关注度不够,虽然最后的集成分类器能够减少方差和偏差,但是对正样本的判别分类就不是很有效.

SMOTE 方法通过对少数类样本采用重采样方法,合成新的样本,避免了其中少数类中某一类样本过于集中的问题,扩大了少数类样本集,能够极大地改善分类器的学习效果.

将 Boosting 方法和 SMOTE 相结合,既可以利用 SMOTE 增加少数类样本,来改善分类器对少数类的预测能力,又可以利用 Boosting 来保证分类器对整个数据集的分类精度.在 Boosting 每轮

迭代中,引入 SMOTE,分类器就能从更多的少数类样本中进行学习,进而扩大少数类的决策区域.

在第 t 次迭代,引入 SMOTE,增加 m 个少数类样本,同时更新权重分布 D_t,这样就增加了少数类样本的权重.合成样本在第 t 次迭代结束后就丢弃,并不加入到原始训练集中去,第 $t+1$ 次迭代时,再重新生成新的合成样本,每次迭代结束时在原始训练集上进行误差估计,这样每个弱分类器面对的正样本都不同,训练的弱分类器更具有多样性,进而提高集成分类器的性能.对增加新样本进行权重更新,就是重新归一化权重.假设第 t 次迭代添加合成样本之后的训练数据集为 n_t,则合成样本集 S_t 中的每一个样本的权重为 $1/n_t$(相当于给新合成样本赋予平均的权值,使算法对每一个新合成样本的关注程度达到平均水平).为规范化权值,训练集原有样本权重调整为原来权重的 $(n_t - m)/n_t$,相当于将原来样本的权重适当成比例减少,将减少的权重赋给新产生的合成样本.增加样本之后,权重则更新为

$$D_t^{\text{new}}(i) = \begin{cases} \dfrac{1}{n_t}, & x_i \in S_t; \\ D_t^{\text{old}}(i) \times \dfrac{n_t - m}{n_t}, & x_i \notin S_t. \end{cases} \tag{1}$$

SMOTE 和 Boosting 相结合的算法 SMOTE Boost 如下:

输入:样本集合 $\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$, $x_i \in X$, $y_i \in Y = \{1,\cdots,C\}$ 为类别标鉴,其中,C_m 为少数类;迭代次数为 T.

初始化:$B = \{(i,y): i = 1,\cdots,m, y \neq y_i\}$,

$$D_t(i) = \frac{1}{m}.$$

步骤:for $t = 1,\cdots,T$

1) 调用SMOTE生成N个数据,根据式(1)调整权重分布 D;

2) 利用权重分布 D_t 训练弱学习算法;

3) 计算弱分类器 h_t: $X \to Y$;

4) 计算弱分类器的误差为

$$\varepsilon_t = \sum_{(i,y)\in B} D_t(i,y)(1 - h_t(x_i,y_i) + h_t(x_i,y)).$$

5) 设置 $\beta_t = \varepsilon_t/(1-\varepsilon_t)$,

$$\omega_t = (1/2) \times (1 - h_t(x_i,y) + h_t(x_i,y_i)).$$

6) 更新样本权重分布 D_t 为

$$D_{t+1}(i,y) = (D_t(i,y)/Z_i) \cdot \beta_t^{\omega_t}.$$

其中,Z_t 为归一化因子;

输出:$H(x) = \arg\max_{y\in Y}(\log \frac{1}{\beta_t}) \cdot h_t(x,hy)$.

1.3 Multiple-layer Boosting

P.Viola 等[18]提出 AdaBoost 的 Cascade 框架来解决人脸检测问题.Cascade 是一个退化决策树,每一个结点都是一个 AdaBoost 分类器.正的结果通过第 1 个分类器,触发第 2 个分类器,通过第 2 分类器的正结果,触发第 3 个分类器,以此类推.负的结果出现在任何一个分类器,都会被立刻拒绝掉.这样大量的负样本在开始的几层都被拒绝了,只有少数困难的样本进入到后面的层中,极大的提高了计算速度.因而 Cascade AdaBoost 在实时目标检测中有着广泛的应用.但是该模型是为识别人脸而设计的,假设所有的特征的计算代价是相同的、很低的,并没有考虑到应用拓展之后,不同类型特征的计算代价的巨大差异对分类器性能的影响.

在结肠息肉检测中,为了很好地描述提取的 3D 疑似区域,计算的特征包括:简单的灰度特征和体特征,以及复杂的微分几何特征、纹理特征和形状特征.简单的特征能够提高分类器的速度,但是难以满足精度的要求.因而复杂特征在提高分类器性能方面就必不可少了,但是花费的时间多.如果在 Cascade 前面的层中,大部分样本都被拒绝了,而这些样本却都要计算复杂的特征,显然会使得系统的速度变慢.因此,在设计 Cascade 时,本文将特征的复杂度考虑进去.由简单特征构成 Cascade 的第 1 层 C_1,拒绝大部分负样本,提高分类速度,后面的层 C_2 则由所有的特征训练得到[11,19],保证分类的精度,如图 2 所示.

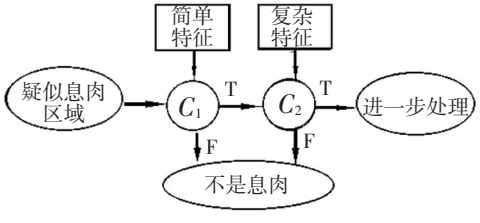


图 2 Multiple-layer Boosting 检测结肠息肉

Cascade 的第 1 层是由简单特征构成,为了提高分类器的效率,没有必要使用所有的简单特征,选取与目标类最大相关、最小冗余的特征即可,采用 MRMR (Minimum Redundancy Maximum Relevance)方法进行挑选.

1.4 MRMR 选择特征

MRMR 是利用信息理论解决特征选择问题[20].在信息理论中,互信息用来衡量两个随机变量的统计依赖性.特征选择可以描述为在目标类上搜寻拥有联合最大依赖性的特征子集.但是 Max-Dependebcy 准则很难实现,估计多变量概率

密度计算很复杂，尤其是对于连续特征.替代选择就是，选择在目标类上有最大相关（maximally relevant），同时有最小冗余（minimally redundant）的特征子集.

为了在特征集$(x_1,\cdots,x_n)$中寻找对于目标类c最好的特征，搜寻最大的互信息$I(x_i;c)$，它表示特征变量x_i和目标类c之间有最大依赖性.可以使用搜寻m个最好的相关特征为

$$\max D(S,c),D=\frac{1}{|S|}\sum_{x_i\in S}I(x_i;c). \quad (2)$$

其中，S为全部特征集.式(2)给出了和目标类最有相关性特征子集，但是并没有描述特征之间的冗余关系.

描述两个特征之间的冗余度为

$$\min R(S,c),R=\frac{1}{|S|}\sum_{x_i,x_j\in S}I(x_i;x_j). \quad (3)$$

联合式(2)、(3)，具有最大相关最小冗余的特征子集为

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

式(4)的实现，首先选择和目标类最相关的一个特征x_k，放入S_1，然后在剩余的特征集$X-S_1$中，寻找下一个与目标类c最大相关，同时和S_1中的特征最小冗余的特征，它和S_1共同构成S_2，这个过程迭代进行，直到找到m个特征，其中$S_1\subset S_2\subset\cdots\subset S_{m-1}\subset S_m$.

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

2　分类器评价标准

评价标准在评价分类器性能和指导分类器模型方面起着重要的作用.传统地，精度是最常用的方法.但是对于数据不平衡问题，精度就不再是合适的评价方法，因为精度依赖多数类，少数类很少影响到它.在医疗领域，这个评价准则应用就很有限，因为某些重要类的代表样本不足，会引起漏诊的.比如，某种癌症在数据集中只占5%（典型的临床数据），而分类器判别无癌症100%，有癌症为0%，则达到95%的精度.从医学的角度来看，这是不可接受的，因为所有的癌症患者都会被漏诊的.

表 1　两类混淆矩阵

类别	预测正类	预测反类
预测正类	TP	FN
预测反类	FP	TN

在包含两类的数据集中，样本数量少，但是识别重要性高的为正样本；反之为负样本.经过分类，训练样本分成4类，为混淆矩阵（confusion matrix）.

根据混淆矩阵，常用的分类器评价指标为

$$\begin{cases}\text{Precision}=\text{TP}/(\text{TP}+\text{FP}),\\ \text{Recall}=\text{TP}/(\text{TP}+\text{FN}),\\ \text{F-value}=\dfrac{(1+\beta^2)\cdot\text{Recall}\cdot\text{Precision}}{\beta^2\cdot\text{Recall}+\text{Precision}}.\end{cases}$$

其中，β为Precision相对Recall的重要程度，通常情况下取$\beta=1$. F-value结合了Recall和Precision，被用来评估分类器对少数类的性能，它是Recall和Precision的调和均值，其取值接近两者中的较小值，因此较大的F-value表示Recall和Precision的值都比较大.

ROC分析在临床CAD中经常应用[21]，ROC曲线描述了两个指标的关系：真阳比例（True Positive Fraction，TPF）和假阳比例（False Positive Fraction，FPF）为

$$\begin{cases}\text{TPF}=\dfrac{\text{TP}}{\text{TP}+\text{FN}},\\ \text{FPF}=\dfrac{\text{FP}}{\text{TN}+\text{FP}}.\end{cases}$$

改变分类器的决策变量上的决策阈值，可以改变类的预测.一个阈值对应一对（TPF，FPF）或者（sensitivity，1−sensitivity），通过这些点对，x轴为TPF，y轴为FPF，就可以绘制ROC曲线.

3　实验与讨论

目前还没有关于CT结肠镜息肉的公开数据库，本文的实验数据来自东软医疗系统有限公司，采集了200个患者（包括俯卧位和仰卧位）的CT结肠镜数据.每层数据像素的尺寸为0.73 mm×0.73 mm，层厚为1.33 mm.所有数据都经过对比剂标记残便，由两名有经验的医生浏览全部的数据，标记出72个息肉，尺寸在[2.2，18.8] mm之间，其中有临床重要性的息肉直径≥5 mm有40个.将正样本和负样本随机分成两部分：70%用作训练；30%用作独立测试.在训练数据中，取50%数据用作训练Boosting分类器，另外50%数据用作调整每个结点的最优参数.在训练Boosting分类器的时候，由于负样本远远大于正样本，为了提高训练速度，对负样本随机向下采样，使得正样本和负样本的比率达到1∶30.

3.1　重采样对分类器性能的影响

本实验用来验证少数类样本重采样对分类器性能的影响，比较了标准Boosting，先SMOTE再Boosting和SMOTEBoost的性能，如表2所示.从表

2 中可以看出,对于标准 Boosting 来说,正样本的数量远远小于负样本,假阳(FP)的数目比真阳(TP)的数目还多,因而 Precision 只有 0.315,F-value 只有 0.227.利用 SMOTE,增加正样本的数目,扩大训练样本集,然后再用 Boosting 训练分类器,则能够极大的提高 Precision、Recall 和 F-value.

SMOTE%表示使用 SMOTE 时,增加正样本的数目占原始正样本数目的百分比.对于先 SMOTE 再 Boosting,SMOTE%取值越大,合成的正样本数目越多,样本集正负样本的不平衡程度越小,Precision、Recall 和 F-value 的值越大.当 SMOTE%=600%,合成的正样本是原始正样本的 6 倍,此时 F-value 达到了 0.815.而 SMOTEBoost 则在每次 Boosting 迭代时,添加新的正样本,其 Precision、Recall 和 F-value 的值比标准 Boosting 的大.同时也可以看出参数 SMOTE%对其的影响,并不是 SMOTE%的取值越大,SMOTEBoost 的性能越好.即使 SMOTE%取最大值 600%,正样本和负样本的比率也没达到 1∶1,但是 SMOTEBoost 早在 SMOTE% = 200% 时已经过拟合了.SMOTEBoost 在 SMOTE%取 100%时,性能是最好的,Precision、Recall 和 F-value 的值比先 SMOTE 再 Boosting 最好的值还好,因而在后续的实验中,SMOTEBoost 都是取 SMOTE%为 100%.

表 2 SMOTE 对 Boosting 的影响

方法	取值	Precision	Recall	F-value
标准 boosting	-	0.135	0.717	0.227
先 SMOTE 再 boosting	100%	0.186	0.869	0.306
	200%	0.416	0.894	0.568
	300%	0.542	0.894	0.675
	400%	0.632	0.903	0.743
	500%	0.691	0.923	0.790
	600%	0.730	0.922	0.815
SMOTEBoost	100%	0.796	0.903	0.846
	200%	0.809	0.856	0.832
	300%	0.669	0.844	0.747
	400%	0.600	0.844	0.701
	500%	0.539	0.842	0.657
	600%	0.539	0.840	0.656

3.2 Multiple-layer Boosting 分类器实验结果

采用 MRMR 挑选 K 个简单特征构成第 1 层强分类器的级联分类器(分类器Ⅰ),与所有层的强分类器由全部的特征构成的级联分类器(分类器Ⅱ)的时间比较,如表 3 所示.表 3 中的时间,是判别一套 CT 数据所需的平均时间.平均一套数据有 448 个疑似区域,分类器Ⅰ在第 1 层就拒绝 422 个负样本,剩余的 26 个疑似区域由后面的层进行判别.而对这 422 个疑似区域计算所有特征,也导致判别整套数据需要 3 057.1 s.分类器Ⅱ的第 1 层是由 MRMR 选出的 K 个简单特征训练得到,拒绝了 389 个疑似区域,剩余的 59 个需要计算复杂特征再进一步进行判别,最终花费 2 562.1 s,和分类器Ⅰ相比,所需时间减少了 495 s,提高了处理的速度.

为了绘制分类器Ⅰ和分类器Ⅱ的 ROC 曲线,需要调整每一层强分类器的阈值.当每层强分类器的阈值都大于 0 时,检测率(TPR)和假阳率(FPR)均为 0,所以阈值的调整范围为[-10,0].为了获得逐渐增加的 TPR 和 FPR,从最后一层开始倒序删除每层强分类器,同时调整当前最后一层强分类器的阈值,绘制的 ROC 曲线如图 3 所示.从图 3 中可以看出,分类器Ⅱ在 ROC 曲线下的面积 AUC 小于分类器Ⅰ的,这可以看作是分类器Ⅱ为了提高速度,在性能方面做出的牺牲,但是分类器Ⅱ对于检测直径≥5 mm 的息肉的敏感度(sensitivity)能达到 90%,平均每例数据 6 个假阳.

表 3 分类器Ⅰ和分类器Ⅱ的时间比较

分类器	时间/s	疑似区域总数	第 1 层拒绝数目
分类器Ⅰ	3 057.1	448	422
分类器Ⅱ	2 562.1	448	389

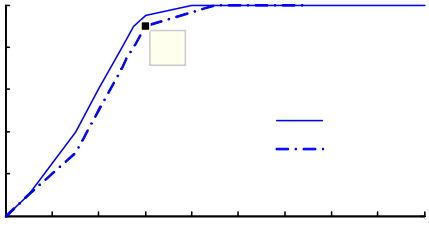


图 3 分类器Ⅰ和分类器Ⅱ的 ROC 曲线

4 结 论

1)本文采用 SMOTE 和 Boosting 结合的方法来解决结肠息肉数据集不平衡的问题,利用 SMOTE 方法合成少数类样本,增加少数类样本的数量,改善了分类器对少数类的预测能力;同时利用 Boosting 保证分类器对整个数据集的分类精度.

2)为了满足结肠 CAD 应用的实时需求,本文采用 MRMR 方法来挑选最大相关、最小冗余的

简单特征构成级联的第 1 层强分类器，拒绝掉绝大多数负样本，提高了级联分类器的处理速度，平均处理一套数据的时间减少了 495 s.

3）本文设计级联分类器检测直径≥5 mm 息肉的敏感度达到 90%，每个数据体有 6 个假阳.实验中假阳的类型主要是厚的褶皱和标记的残便，在保证敏感度的前提下，应进一步降低假阳.

参考文献

[1] CHOWDHURY T, WHELAN P, GHITA O, *et al.* A fully automatic CAD-CTC system based on curvature analysis for standard and low-dose CT data [J]. IEEE Transactions on Biomedical Engineering, 2008, 55(3): 888-901.

[2] JOHNSON C, HARA A. CT colonography: the next colon screening examination? [J]. Radiology, 2000, 216(2): 331-341.

[3] PICKHARDT J, CHOI R, HWANG I, *et al.* Computed tomographic virtual colonoscopy to screen for colorectal neoplasia in asymptomatic adults [J]. The New England Journal of Medicine, 2003, 349(23):2191-2200.

[4] SUMMERS R, BEAULIEU C, PUSANIK L, *et al.* Automated polyp detector for CT colonography feasibility study [J].Radiology, 2000, 216(1): 284-290.

[5] YOSHIDA H, NAPPI J. Three dimensional computer aided diagnosis scheme for detection of colonic polyps [J]. IEEE Transactions on Medical Imaging, 2001, 20(12):1261-1274.

[6] SUMMERS R, YAO J, PICKHARDT P, *et al.* Computed tomographic virtual colonoscopy computer-aided polyp detection in a screening population [J]. Gastroenterology, 2005, 129(6):1832-1844, 2005.

[7] YOSHIDA H, MASUTANI Y, MACENEY P, *et al.* Computerized detection of colonic polyps at CT colonography on the basis of volumetric features: pilot study [J]. Radiology, 2002, 222(2) 327-336.

[8] PAIK D, BEAULIEU C, RUBIN G, *et al.* Surface normal overlap: a computer-aided detection algorithm with application to colonic polyps and lung nodules in helical CT [J]. IEEE Transactions on Medical Imaging, 2004, 23(6):661-675.

[9] SLABAUGH G, YANG X, YE X, *et al.* A robust and fast system for CTC computer-aided detection of colorectal lesions [J]. Algorithms,2010,5(1):21-43.

[10] SUN Yanmin, KAMEL M, WONG A, *et al.* Cost-sensitive boosting for classification of imbalanced data [J]. Pattern Recognition, 2007, 40(12):3358-3378.

[11] BI Jianbo, PERIASWAMY S, OKADA K, *et al.* Computer aided detection via asymmetric cascade of sparse hyperplane classifiers [C]//Proceedings of the Twelfth Annual SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY: ACM, 2006:837-844.

[12] WU Dijia, BI Jinbo, BOYER K. A min-max framework of cascaded classifier with multiple instance learning for computer aided diagnosis [C]//Proceedings of Computer Vision and Pattern Recognition. Miami, FL: IEEE Xplore, 2009:1359-1366.

[13] ZHENG Yalin, YANG Xiaoyun, BEDDOE G, *et al.* Reduction of false positives in polyp detection using weighted support vector machines [C]//Proceedings of the 29th Annual International Conference of the IEEE EMBS. Lyon, Francer: IEEE Xplore, 2007:4433-4436.

[14] HE Haibo, GARCIA E. Learning from imbalanced data [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9):1263-1284.

[15] CHAWLA N, BOWYER K, HALL L, *et al.* SMOTE: synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research,2002,16(1): 321-357.

[16] CHAWLA N, LAZAREVIC A, HALL L, *et al.* SMOTEBoost: improving prediction of the minority class in boosting [C]//Proceedings of Seventh European Conf. Principles and Practice of Knowledge Discovery in Databases. Berlin, Heidelberg: Springer-Verlag, 2003: 107-119.

[17] FREUND Y. Boosting a weak learning algorithm by majority [J]. Information and Computation, 1995, 121(2):256-285.

[18] VIOLA P, JONES M. Robust real-time face detection [J]. International Journal of Computer Vision, 2004, 57(2): 137-154.

[19] PAISITKRIANGKRAI S. Robust Object Detection with Efficient Features and Effective Classifiers [D]. Randwick: The University of New South Wales, 2011.

[20] PENG H, LONG F, 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.

[21] OBUCHOWSKI A. Receiver operating characteristic curves and their use in radiology [J]. Radiology, 2003, 229(1):3-8.

（编辑　张　红）