ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 30,No. 5,Sep. 2015,pp. 1003-1010
DOI:10.16337/j.1004-9037.2015.05.010

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

基于弱监督 ECOC 算法的肺结节辅助检测

苏志远[1,3]　刘慧[1,3]　尹义龙[1,2]

(1.山东财经大学计算机科学与技术学院，济南,250014;2.山东大学计算机科学与技术学院，济南,250101;
3.山东省数字媒体技术重点实验室，济南,250014)

摘　要：肺结节的准确分类与识别是计算机辅助诊断系统在肺癌诊断领域应用的关键,同时也面临巨大的挑战。该技术不仅在特征表示、样本标记等方面存在发展的瓶颈,而且目前缺少准确、有效的分类识别算法。本文提出了一种结合弱监督纠错输出编码(Error-correcting output codes，ECOC)算法和肺结节形状特征表达的肺结节多分类算法。为了提高分类识别的准确率,本文对肺结节的形状特征进行了详细的分析,并提出了一系列准确的形状特征描述向量。在分类识别阶段,算法训练学习了利用专家对肺结节标记信息标记的少量样本,并生成二类分类器,获得编码矩阵。最后,通过计算测试样本编码和编码矩阵每一行的汉明距离,确定样本所属类别。实验结果表明,本文方法能够获得更加准确的分类结果。

关键词：肺结节；分类识别；弱监督；纠错输出编码；肺部图像数据库联盟
中图分类号：TP391.41　　**文献标志码**：A

Pulmonary Nodule Aided Detection Based on Weakly-Supervised ECOC Algorithm

Su Zhiyuan[1,3], Liu Hui[1,3], Yin Yilong[1,2]

(1. School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan, 250014, China;
2. School of Computer Science and Technology, Shandong University, Jinan, 2500101, China; 3. Digital Media Technology Key Lab of Shandong Province, Jinan, 250014, China)

Abstract: Accurate classification and recognition of pulmonary nodules are key process of lung cancer computer-aided diagnosis (CAD) system. Meanwhile, there are still some scientific and technical challenges, including the difficulty of the feature representation and samples labeled, and the lack of accurate and effective recognition and classification algorithms. A multi-classification algorithm is presented combining weakly-supervised ECOC algorithm with pulmonary nodules features expression of shape. In order to improve the classification accuracy, we select a series of accurate shape feature description vectors by deliberating the shape features of pulmonary nodules. During the training phase, the coded matrix is constructed by a series of binary classifiers, which are generated by a small amount of labeled pulmonary nodules from experts. Finally, the Humming distance between the code of testing sample and each row of the coded matrix are calculated to determine the category of the testing sample. Experimental results show that the proposed method can obtain more accurate classification results.

Key words: pulmonary nodule; classification and recognition; weakly-supervised learning; error-correction output codes; lung image database consortium

---

基金项目:国家自然科学基金(61272245)资助项目;山东省科技发展计划(2014GGX101037)资助项目;济南市高校自主创新计划(201401216)资助项目。
收稿日期:2015-06-08;**修订日期**:2015-06-30

引　言

近些年，由于环境污染和吸烟等因素导致肺癌的发病率和死亡率成逐年上升的趋势。如果无法找到有效的方法提早发现并诊断肺癌，预计到 2025 年中国的肺癌患者将增加到 100 万人，从而成为世界肺癌第一大国。国家卫生和计划生育委员会的统计数据显示，目前癌症发病率每年以 26.9%的速度增长[1]，尤其是在雾霾严重地区，如北京、天津、济南等地区，肺癌发病率明显高于全国平均水平，PM2.5 成为肺癌发病的罪魁祸首。肺癌的早期发现和诊疗成为降低肺癌死亡率主要手段[2]。因此，研究高效准确的肺癌辅助检测系统(Computer-aided diagnosis, CAD)具有重要的理论和现实意义。

本文研究的肺癌 CAD 系统的工作流程如图 1 所示，主要分为 3 步：肺结节分割、特征提取和结节病变程度的判定。整个流程中结节的病变程度判定最为重要，研究者已经提出很多有效方法，例如神经网络[3]、支持向量机(Support vector machine, SVM)[4]等。其中，支持向量机在处理小样本学习及非线性等问题中具有一定的优势，已经在模式识别等领域得到了广泛的应用。与传统的神经网络相比，SVM 针对有限样本进行的分类学习，在理论应用上比神经网络具有更强的泛化能力。但是，一般的 SVM 只能对二类问题进行分类，不能对多分类进行有效的操作。为此，许多学者提出了多种解决方法，主要包括一对一(One-vs-One)[5]、一对多(One-vs-Rest)[6]和二叉树等方法。其中 One-vs-One 具有较强的分类能力，但需要多个 SVM 分类器支持。例如，对一个 n 类问题进行分类处理，需要 $n(n-1)/2$ 个 SVM 分类器，并且在获得最后分类结果时，需要利用投票策略进行最终的判定。但是，在判断一类样本类别时，当出现多个类的得票数目相等，就无法进行分类，产生拒分区域。One-vs-Rest 方法对于 n 类问题只需要 n 个 SVM 分类器，但分类器在分界区域存在重叠现象，所以分类效果不理想。二叉树方法具有很高的分类效率，对于 n 类问题只需要 $(n-1)$ 个 SVM 分类器，不会出现不可分现象，但是容易造成误差累积的情况，并且，二叉树的拓扑结构和每个节点所包含子类的设置均会对分类结果产生影响。但是，纠错输出编码(Error-correcting output code, ECOC)在机器学习领域的应用解决了这个难题。同时，文献[7]提出利用纠错码将二分类器扩展到多分类的分类问题中，有效地解决了上述分类问题中的拒分问题。由于纠错输出编码在性能上的优秀表现，许多研究学者对它进行了深入的研究和扩展，文献[8]利

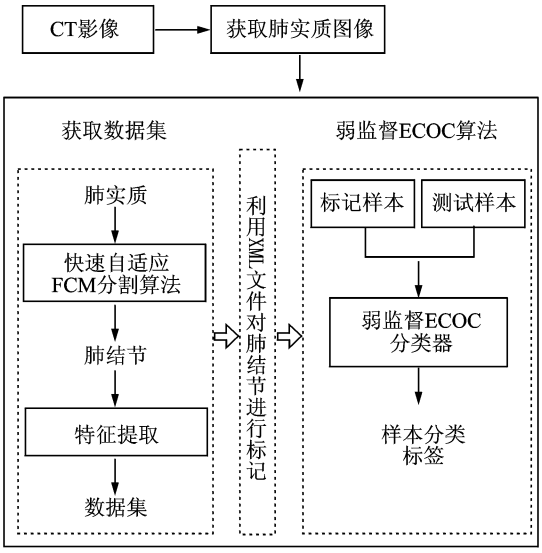


图 1　本文提出的 CAD 系统流程图

Fig. 1　Proposed flow chart of CAD

用反向传播的思想将 ECOC 应用在不同类型的分类问题上。文献[9]则通过优化最大似然目标函数找到了最合适的样本空间编码矩阵。文献[10]提出一种改进的基于 ECOC 算法,可以消除分类过程中的类别重叠无法分类的情况,该方法将纠错输出编码分类算法推向了一个新的高度。

综合上述分析,本文设计了一套行之有效的肺癌 CAD 系统。文献[11]完成了肺结节的分割提取,提出了快速自适应 C 均值模糊聚类算法(Fuzzy c-means, FCM),获得了准确满意的分割结果[11]。在特征提取阶段,本文通过分析肺部图像数据库联盟(Lung image database consortium, LIDC)数据库提供的注释文件制定了一组准确反映肺结节形状信息的特征向量[12]。同时,通过选取部分肺结节建立了具有标签信息的肺结节特征的数据集。通过综合分析,本系统将肺结节综合分成 3 类:恶性结节、阳性结节和假阳性结节。然而,肺结节类型多样、结构复杂,如孤立型、粘连型、毛玻璃型和空洞型等,在训练学习阶段,利用现有数据获得大量完备、准确的标记样本集合十分困难。因此,本文引入弱监督思想[13]的 ECOC 算法,以解决有效标记样本数量不足的缺陷,即采用学习部分标记样本的策略,从而有效改善由于标记样本数量不足导致分类精度下降的情况[14]。

1 肺结节特征提取

为了促进对肺癌 CAD 系统的设计与研发,美国国家癌症研究所(National cancer institute, NCI)建立了肺部计算机断层扫描影像数据库 LIDC,其中包含 1 012 个病例,共含有 1 356 个可供研究人员学习的肺结节。数据库为每一病例建立一个独立的文件,其中包含整肺的 CT 断层扫描影像(DICOM 格式)100～500 张不等,并且为每一病例出示一个 XML 格式的注释文件来标注结节。每一病例中都有 4 名放射学方面专家对其出现的结节情况进行诊断,以结节的边界坐标划分和视觉特征,如毛刺征、分叶征、钙化和结节恶性程度等特征对结节进行描述并存储到注释文件中。如图 2 所示,其特征数值越大,表明恶性程度越高。

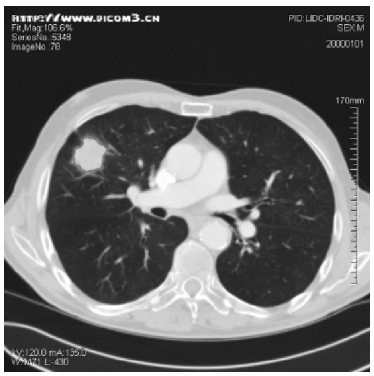


图 2 LIDC 数据库中病例的 CT 图像及肺结节区域

Fig. 2 CT image and pulmonary nodules in LIDC

肺结节通过准确的特征描述子进行表述是获得准确分类识别结果的基石。由于肺结节直径一般在 3～30 mm 之间,非专业人士无法给出精确的诊断结果,导致无法对结节类别进行有效的标记。LIDC 数据库提供了 4 位肺部影像专家对同一病例的诊断结论,并给出了精细度、球形度、钙化程度、恶性程度、边缘、分叶征和毛刺特征等 9 个病变特征描述。通过对 1 000 个病例的统计分析,发现分叶征和毛刺征是最能表征肺结节的恶性程度的病理特征。肺结节病变恶性程度与分叶征和毛刺征表现等级服从同一分布,从图 2 可以得到这个结论。而分叶征和毛刺征病变程度加深在形态学上的主要体现就是形状特征和灰度的变化。但是,由于不同机器拍摄 CT 影像时扫描剂量不固定,同一病例在不同剂量不同机器下获得 CT 影像灰度不同。综上原因,本文选择以形状作为主要特征描述。因为形状信息不会受到机器和扫描剂量等因素的影响。因此本文通过实验总结,选取了一组以形状特征为主的特征向量组对肺结节特征信息进行了提取,包括灰度方差、灰度直方图熵、似圆度、径向均值和方差、边界粗糙度、紧凑度、形状不变矩($H0$, $H2$, $H3$, $H4$)和傅里叶描述子(选取前 20 项)等 31 项描述子。

在获得有效的特征向量之后,由于不同特征的物理意义不同,导致在取值范围上也大不相同。并且每种特征需要在特征表达权值上相同,因此需要对获得的特征向量做归一化处理,使整个特征取值映射到同一取值范围。本文应用高斯归一化算法对获得的特征向量进行归一化处理。高斯归一化可以表示为

$$x_{ik}^{n}=\frac{x'_{ik}-\overline{x'_{ik}}}{s_k} \tag{1}$$

式中：$\overline{x'_{ik}}=\frac{1}{n}\sum_{i=1}^{n}x'_{ik}$，$s_k^2=\frac{1}{n}\sum_{i=1}^{n}(x'_{ik}-\overline{x'_k})^2$。

利用分割提取阶段获得的准确肺结节区域进行特征提取计算，获得肺结节特征数据，如图 3 所示。然后，利用 LIDC 数据库提供的 XML 注释文件对每一肺结节实例进行病变等级标注。图 3 所示肺结节的部分特征数据如表 1 所示。到目前为止已经建立了一个基于 LIDC 数据的 31 维形状特征肺结节标记数据库。

图 3 肺结节分割结果

Fig. 3 Segmentation results of pulmonary nodules

表 1 图 3 所示肺结节的部分特征数据

Table 1 Partial feature data of nodules from Fig. 3

特征	结节								
	1#	2#	3#	4#	5#	6#	7#	8#	9#
灰度方差	0.137 8	0.127 9	0.103 5	0.161 4	0.189 7	0.182 8	0.113 7	0.194 1	0.107 8
灰度直方图熵	4.558 9	4.340 5	4.162 8	4.358 0	3.674 8	3.980 7	4.688 1	3.435 9	4.197 0
似圆度	0.487 8	0.540 2	0.775 3	0.655 1	0.745 8	1.011 5	0.031 8	1.013 9	0.504 9
径向均值	0.693 4	0.595 8	0.803 5	0.625 0	0.746 0	0.642 8	0.574 5	0.826 5	0.573 2
径向方差	0.154 4	0.231 4	0.101 9	0.168 0	0.170 0	0.243 0	0.186 2	0.121 5	0.244 5
边界粗糙度	0.114 1	0.162 3	0.093 4	0.151 3	0.307 3	0.236 2	0.159 8	0.390 5	0.234 1
形状不变矩 $H0$	3 087.000 0	571.000 0	1 764.000 0	285.000 0	376.000 0	218.000 0	141.000 0	102.000 0	380.000 0
形状不变矩 $H1$	0.180 4	0.287 5	0.168 1	0.205 8	0.186 4	0.161 5	1.357 4	0.166 2	0.350 3
形状不变矩 $H2$	0.001 7	0.055 2	9.281 2E−05	0.008 3	0.007 2	0.006 062 3	0.084 2	0.001 9	0.094 9
形状不变矩 $H3$	0.001 1	0.004 758 6	0.000 622 66	0.003 8	6.586 4E−05	0.000 022 58	0.267 3	0.000 151 06	0.001 0
紧凑度	0.333 2	0.164 4	0.597 5	0.361 4	0.424 8	0.306 8	0.254 1	0.624 7	0.164 4

2 弱监督分类识别算法

2.1 纠错输出编码

纠错输出编码算法可以将多分类问题转化为一系列二类分类器。给定一个 K 类的分类问题，依照 One-vs-Rest 策略为每一类都训练一个二类分类器，分类器个数定义为 L。因此每个类别可以获得一个编码长度为 L 的码字。将所有的码字排列在一起形成一个 $K\times L$ 的编码矩阵 $\boldsymbol{M}$。如图 4 给出的是一个 5 类问题的 ECOC 分类器。其中方阵为编码矩阵 $\boldsymbol{M}$，该矩阵利用 7 个二类分类器 $\{f_1, f_2, \cdots, f_7\}$ 对 5 类问题 $\{C_1, C_2, \cdots, C_5\}$ 进行训练学习，分别生成其对应的码字。在每一个二类分类器 f_i 中，如果测试样本被完全认定为所属类别则编码为 $+1$，否则编码为 -1。在图 4 的例子中，假设带有标签的 n 个训练样本为 $\{(\boldsymbol{x}_1, l(\boldsymbol{x}_1)), (\boldsymbol{x}_2, l(\boldsymbol{x}_2)), \cdots, (\boldsymbol{x}_n, l(\boldsymbol{x}_n))\}$，经过训练得到分类矩阵。其中，编码为 $+1$ 表示正例，编码为 -1 表示反例。

ECOC	f_1	f_2	f_3	f_4	f_5	f_6	f_7
C_1	1	−1	−1	−1	1	1	−1
C_2	−1	1	−1	−1	−1	1	−1
C_3	−1	−1	1	−1	−1	−1	1
C_4	−1	−1	−1	1	−1	−1	−1
C_5	−1	−1	−1	−1	1	−1	1

二类分类器　　类

图 4　一个 5 类问题的 ECOC 分类器

Fig. 4　ECOC design for 5-class problem

在编码过程中,利用 L 个二类分离器可以获得测试数据集中每一样本的编码码字。通过计算测试样本编码和编码矩阵的每一行的汉明距离,根据距离最小的原则确定样本所属类别,即当码字之间距离最小,则其具有相同的标签,且该分类器结合了通信论中关于信道传输的纠错解码技术。当分类器中某些 SVM 分类器产生错,也可以通过纠错机制获得准确的分类结果。当两码字之间的距离为 d 时,可以利用 ECOC 算法的纠错机制对 $(d-1)/2$ 位误码进行纠正。

2.2　结合弱监督思想的 ECOC 算法

尽管利用 ECOC 作为多分类框架对多分类问题进行分解并得到了很好的应用实践,但是在肺结节的分类应用过程中,仍然存在着大量无法区分模棱两可的数据,描述特征之间存在重合的现象,使训练过程不能得到完全准确的编码矩阵。本文结合了弱监督算法 PL-ECOC 算法[10]并同时为系统提供了准确标记的肺结节训练实例。该方法通过 ECOC 算法对准确标记的肺结节样本进行学习。该标记样本通过 LIDC 数据库中提供的注释文件标记肺结节获得标记特征数据集。利用标记特征样本集的信息为获得二进制编码矩阵提供监督信息。根据该明确的监督信息可以有效降低样本中模棱两可的干扰信息。因此,该系统可以获得准确的类别标记结果。在实验结果分析中也同样证明了该系统的高效性。

弱监督 ECOC 算法作为一个完善的多分类机制,仍然延续的是 One-vs-Rest 策略。给定训练样本集:$S=\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_n, y_n)\}$;其中 $\boldsymbol{x}_i \in \mathbf{R}^m$, y_i 为数据样本 $\boldsymbol{x}_i$ 的标签。ECOC 分类器可以将样本实例通过 $X \xrightarrow{\text{ECOC}} \boldsymbol{M}\{+1, -1\}$ 映射到编码矩阵,即在编码阶段利用该映射生成一个 $C \times L$ 的编码矩阵。在编码矩阵 $\boldsymbol{M}$ 中,每一行代表一个 L 位的码字。另一方面,每一列定义为一个二类分类器 f_i。通过上述定义,每一个二类分类器,可以根据训练样本的标记 y 判断是否完全为正例或反例来生成训练样本集到编码矩阵的映射。

在解码阶段,给定测试数据集 $\boldsymbol{x}_i^{\mathrm{T}}$,通过计算可以生成一个 L 位的编码矩阵 $f(\boldsymbol{x}_i^{\mathrm{T}})$,该矩阵包含 L 个二类分类器:$f(\boldsymbol{x}_i^{\mathrm{T}})=\{f_1(\boldsymbol{x}_i^{\mathrm{T}}), f_2(\boldsymbol{x}_i^{\mathrm{T}}), \cdots, f_n(\boldsymbol{x}_i^{\mathrm{T}})\}$,对比测试样本的码字和编码矩阵之间的汉明距离,如果测试样本的码字和编码矩阵中 $\boldsymbol{M}(i, :)$ 距离最小,即测试样本 $\boldsymbol{x}_i^{\mathrm{T}}$ 的类别标签与 C_i 其相同。判断式为:$d(\boldsymbol{x})=\arg\min_{1\le i\le C}\{\mathrm{dist}(f(\boldsymbol{x}_i^{\mathrm{T}}), \boldsymbol{M}(i, :))\}$,其中 $\mathrm{dist}(\boldsymbol{x}, \boldsymbol{y})$ 为汉明距离,即码字中不同码字的位数,可表示为 $\mathrm{dist}(\boldsymbol{x}, \boldsymbol{y})=\sum_{k=1}^{n}(\boldsymbol{x}_k \oplus y_k)$,其中 $\oplus$ 为模 2 运算。

利用上述算法,本文在训练阶段向分类识别系统提供标记的肺结节特征数据 $(\boldsymbol{x}_i, y_i)$,在这里需要说明

的是本文构造了一个新的数据结构，该数据结构包含特征数据和分类标签。通过训练阶段后可以获得 L 位的二类分类器，进而获得编码矩阵。在训练二类分类器时，按照样本是否完全属于该类的原则计算生成二类分类器，如果假设成立，则赋值为+1，否则赋值为−1。下面给出基于弱监督 ECOC 分类器的肺结节分类识算法。

算法 1　基于弱监督 ECOC 分类器的分类识别算法

输入：标记的肺结节特征训练集 $\{\boldsymbol{x}_i, y_i\}$，编码长度 L，二分类监督学习器 B，二分类训练集阈值 thr $=\mathrm{ceil}(0.1\times num_{\mathrm{train}})$，肺结节特征测试样本 $\boldsymbol{x}^{\mathrm{T}}$。

输出：测试样本 $\boldsymbol{x}^{\mathrm{T}}$ 的标签 $\boldsymbol{y}^{\mathrm{T}}$。

初始化：$l=0$；

While $l\neq L$ do

任意生成 L 位的列码字 $\boldsymbol{n}=[n_1, n_2, \cdots, n_c]^{\mathrm{T}}$；

参照标记样本计算 n_i，初始化二类分类器训练集 $T=0$；

如果训练样本完全属于 $\boldsymbol{y}^{+}$，则设置编码矩阵 $\boldsymbol{M}(i, +1)$；反之如果完全属于 y^{-}，则设置编码矩阵 $\boldsymbol{M}(i, -1)$；

if $|T|\geqslant$ thr then

$l=l+1$；

$\boldsymbol{M}(:, l)=\boldsymbol{n}$；

Endif

Endwhile

那么，将 $B(T)$ 赋值给二类分类器。

输出：二类分类器 $f(\boldsymbol{x})$；计算测试样本和每一个码字之间的距离 $d(\boldsymbol{x})$；返回测试样本 $\boldsymbol{x}^{\mathrm{T}}$ 的标签 $\boldsymbol{y}^{\mathrm{T}}$。

3　实验结果与分析

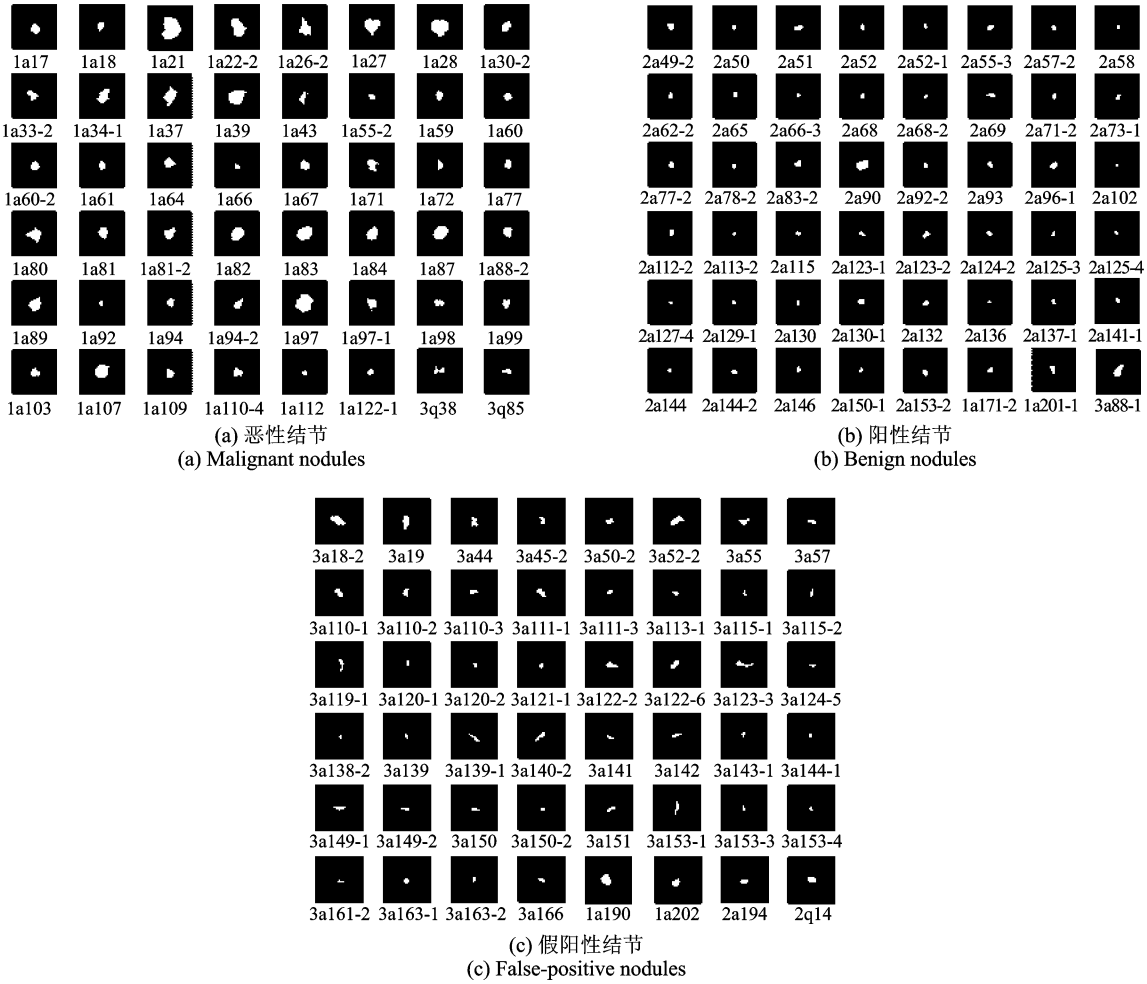
本文使用 LIDC 数据库提供的 CT 影像作为实验数据，并选择了 188 个病例。其中包括 147 例恶性结节，149 例良性结节和 156 例假阳性结节。在整个数据库中，4 位放射科专家对每个结节进行分析诊断，并对恶性等级进行评估标记为 1～5，即随着数值的增大恶性程度将会增大。因此，本文将恶性等级标记为 4，5 的结节定义为恶性肿瘤。同样地，将注释文件中恶性等级为 1，2 和 3 定义为良性结节。其中，假阳性结节病例是大多病例中提取的如血管等区域。本文为系统训练阶段提供 144 例标记信息作为监督信息，包括 47 例恶性结节，42 例良性结节和 55 例假阳性结节。为了证明弱监督 ECOC 分类器在肺结节分类识别应用方面的准确性。图 5 给出了部分结节分类结果。

为了进一步证明本系统性能，比较了传统的分类识别方法和本文提出的部分标记的弱监督分类方法。通过调整训练样本的比例，得到在不同比例下训练样本对分类结果准确率的影响。表 2 给出了本文系统算法与 One-vs-One，One-vs-Rest 和传统的 FCM 分类方法[15]的性能比较。其中，Mean 为均值，表示平均分类准确率，Std 为标准差，表示该方法的稳定性[16]。

表 2　不同分类方法在肺结节的特征数据集上的性能比较

Table 2　Performance comparison of different classifications on feature of pulmonary nodules dataset

算　法	标记样本所占比例				Mean±Std
	0.1	0.2	0.3	0.4	
FCM	0.734 6	0.734 6	0.734 6	0.734 6	0.734 6±0.04 2
One-vs-Rest	0.786 4	0.795 2	0.807 8	0.809 2	0.803 7±0.03 4
One-vs-One	0.819 3	0.822 5	0.840 6	0.851 2	0.843 2±0.01 9
本文方法	0.830 4	0.837 2	0.839 4	0.841 2	0.834 4±0.01 0



(a) 恶性结节
(a) Malignant nodules

(b) 阳性结节
(b) Benign nodules

(c) 假阳性结节
(c) False-positive nodules

图 5 部分肺结节分类结果

Fig. 5 Classification results section of pulmonary nodules

表 2 的实验结果显示,传统的 FCM 聚类算法在标记样本比例发生变化时,由于不是监督算法其准确率没有呈现出相应的变化。而 One-vs-Rest 和 One-vs-One 的算法则随着监督信息中标记样本所占比例的增加,其算法准确率逐渐提高。但是在监督标记样本比例较低的情况下并没有表现出优秀的识别性能。本文提出的算法随着监督标记样本比例提高的同时其识别准确率也成逐渐提高的趋势,同时在监督标记样本比例较低的情况下同样表现优秀。但在标记样本例增加到一定的程度时,性能又有下降的趋势,其原因可能是随着监督信息的增多,标记样本中的模糊分类样本的误差堆积产生的。综合上述实验结果可知,本文提出的系统方案能够获得较好的分类识别效果。比较 3 类传统的分类算法,可以发现该方法尤其在标记样本量较少的情况下性能仍然表现优异,而其他分类方法则随着训练样本比例的增加,性能得到提高并趋于平衡。因此,该方法能有效解决肺结节标记样本不足的情况。

4 结束语

对于普通的监督算法而言,随着监督样本的增加,其性能可以得到有效改善。但现实情况下往往存在有效标记样本不足的情况,无法为分类算法提供足够的监督信息。本文提出了一种针对肺结节分类

识别的系统，在系统中结合 ECOC 分类器和弱监督思想，有效地解决了可用肺结节标记样本不足的问题。其中，在构建肺癌辅助诊断系统的过程中，通过分析和解释 LIDC 数据库中的注释文件，提取并构建了一个标记的肺结节形状特征数据集，进而利用该数据集对肺结节进行诊断识别。实验结果表明，该系统能有效地提高肺结节分类识别的准确率及鲁棒性，特别是在监督样本量较少的情况下性能表现依然优秀，在一定程度上解决了肺结节标记样本不足的问题。本文将进一步丰富和完善对肺结节的特征表示，同时优化改进分类算法，获得更加准确的分类结果。

参考文献：

[1] Siegel R, Ma J M, Zou Z H, et al. Cancer statistics [J]. A Cancer Journal for Clinicians, 2014, 64(1): 9-29.

[2] Ayman E B, Garth M B, Georgy G, et al. Computer-aided diagnosis systems for lung cancer: Challenges and methodologies [J]. International Journal of Biomedical Imaging, 2013: 942353.

[3] Zaidi N A, Squire D M. Local adaptive SVM for object recognition [C] //Digital image computing: Techniques and Application (DICTA), 2010 International Conference on. Sydney, Australia: IEEE, 2010:196-201.

[4] Tanchotsrinon C, Phimoltares S, Maneeroj S. Facial expression recognition using graph-base features and artificial neural networks [C] // Imaging Systems and Techniques (IST), 2011 IEEE International Conference on. Penang, Malaysia: IEEE, 2011:331-334.

[5] Anthony G, Gregg H, Tshilidzi M. Image classification using SVMs: One-against-one vs one-against-all [C] //Proceeding of the 28th Asian Conference on Remote Sensing. Kuala Lumpur, Malaysia: IEEE, 2007:12-16.

[6] Hong J H, Cho S B. Aprobabilistic multi-class strategy of one-vs-rest support vector machines for cancer classification [C]. // Advances in Neural Information Processing (ICONIP 2006). Brazilian: Neurocomputing, 2008: 3275-3281.

[7] Dietterichand T G, Bakiri G. Solving multiclass learning problems via error-correcting output codes [J]. Journal of Artificial Intelligence Research, 1995, 2:263-286.

[8] Alpaydin E, Mayoraz E. Learning error-correcting output codes form data [C] // Proceeding of the 9th Internet Conference on Artificial Neural Networks. Edinburgh, UK: IET, 1999:743-748.

[9] Utschick W, Weichselberger W. Stochastic organization of output codes in multiclass learning problems [J]. Neural Compute, 2001, 13(5):1065-1102.

[10] Zhang M L. Disambiguation-free partial label learning [C] //Proceeding of the 14th SIAM International Conference on Data Mining(SDM14). Philadelphia, PA:[s. n.], 2014:37-45.

[11] Liu H, Zhang C M, Su Z Y, et al. Research on a pulmonary nodule segmentation method combining fast self-adaptive FCM and classification [J]. Computational and Mathematical Methods in Medicine, 2015: 18576.

[12] Samuel G A, Geoffrey M. The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans [J]. Medical Physics, 2011, 38:915-931.

[13] 汪荆琪，徐林莉. 一种基于多视图数据的半监督特征选择和聚类算法 [J]. 数据采集与处理, 2015, 30(1):106-116.
Wang Jingqi, Xu Linli. Semi-supervised feature selection and clustering for multi-view data [J]. Journal of Data Acquisition and Processing, 2015, 30(1):106-116.

[14] Cour T, Sapp B, Taskar B. Learning from partial labels [J]. Journal of Machine Learning Research, 2011, 12:1501-1536.

[15] Bouchachia A, Pedrycz W. Data clustering with partial supervision [J]. Data Mining and Knowledge Discovery, 2006, 12(1):47-78.

[16] Foody G M. Harshness in image classification accuracy assessment [J]. International Journal of Remote Sensing, 2008, 29(11):3137-3158.

作者简介：

苏志远(1988-)，男，硕士研究生，研究方向：医学图像处理、机器学习及应用，E-mail:suzhiyuanlt@163.com。

刘慧(1978-)，女，教授，研究方向：医学图像处理、计算机辅助诊断、信息检索机器学习及其应用。

尹义龙(1972-)，男，教授，博士生导师，研究方向：机器学习及应用。