

基于代价敏感图像层的肺癌细胞病理诊断系统

张 纓

(解放军第八一医院全军肿瘤研究中心, 江苏 南京 210002)



张纓,女,硕士,副主任医师,解放军八一医院病理科副主任。从事病理组织及细胞学诊断工作 20 余年,具备丰富的临床病理诊断经验,主要研究方向为全身各系统肿瘤的病理诊断及基础研究。参加肺癌早期诊断综合治疗研究领域的国家“九五”、“十五”重大科技攻关课题、江苏省应用基础课题及国家自然科学基金课题的研究。2008 年参与研制的“数字化肺癌早期细胞病理电脑诊断系统”通过了江苏省科技厅科技成果鉴定,并申请了计算机软件著作权 1 项及国家发明专利 1 项;2009 年“肺癌新技术研究及临床应用”获得军队医疗成果二等奖;2011 年主持军区“十二五”课题“肝癌特异性 miRNA 的表达及其与细胞周期调控的关系”;参编专著《肝衰竭的基础与临床》;在国家核心学术期刊发表论文 30 余篇,SCI 论文 3 篇。

摘要:将代价敏感方法引入肺癌细胞图像识别,建立基于代价敏感图像层的肺癌细胞病理诊断系统(Cost-sensitive Image-level Lung Cancer Diagnosis System, CILCDS),并研究其临床应用价值。与既往方法比较, CILCDS 在癌与正常细胞的判断以及癌细胞的分类识别诊断过程中取得了最低的错误分类代价 0.37% 以及更高的组织病理结果符合率 97.61%。基于代价敏感图像层技术能够有效减少肺癌细胞病理诊断过程中的假阳性结果,有可能为肺癌细胞病理诊断提供一个更为有效的检测手段。

关键词: 代价敏感;肺肿瘤;诊断;病理学;图像处理,计算机辅助

中图分类号:R734.2 文献标识码:A 文章编号:1004-0242(2014)09-0726-05

doi: 10.11735/j.issn.1004-0242.2014.09.A005

A Cost-Sensitive Image-Level Lung Cancer Diagnosis System

ZHANG Ying

(Cancer Center of PLA, Nanjing Ba-Yi Hospital of PLA, Nanjing 210002, China)

Abstract: Cost-sensitive methods introduced into cytopathological diagnosis, A Cost-sensitive Image-level Lung Cancer Diagnosis System (CILCDS) is proposed, and its clinical value is evaluated. CILCDS obtained lower misdiagnosis costs 0.37% and comparable precision 97.61% in normal/cancerous identification compared with other techniques. It suggests that CILCDS may reduce the false positive results in cytopathological detection of lung cancer and may provide a more effective means for cytopathological diagnosis of lung cancer.

Key words: cost-sensitive; lung neoplasms; diagnosis; pathology; image processing, computer-assisted

肺癌是我国最常见的癌症,肺癌总生存率在过去的 20 年中无明显提高,而乳腺癌、结肠癌和前列

腺癌等生存率则大大提高,其中一个重要原因是早期诊断。十余年来,我们在“肺癌早期诊断立体定位仪”基础上^[1],利用其能对肺部 1cm 左右小病灶快速、三维立体定位穿刺取检标本的优势,与南京大学

收稿日期:2014-04-25;修回日期:2014-05-05

通讯作者:张纓, E-mail: 1738876035@qq.com

软件新技术国家重点实验室研究并创建了“肺癌细胞病理诊断系统”^[2-19],并在临床工作中通过大量样本的实践验证使之得到逐步完善。

1 基于图像处理和人工神经网络的“肺癌早期细胞病理诊断系统”

初期,我们将计算机图像处理、模式识别、专家系统和人工神经网络等技术与肺癌细胞病理诊断、组织学诊断有机结合,初步实现了肺癌细胞判定及分类诊断。首先,采用基于图像灰度梯度的直方图和双阈值快速分割法选定有效待检细胞,利用形态学色度学特征对有效待检细胞进行识别;再通过计算机内构的专家系统对细胞核形态、位置及染色质分布等知识依据产生式规则进行推理诊断,初步判別出癌细胞、非癌细胞和核异型细胞;最后,采用人工神经网络模拟人脑结构和信息传导方式建立数据模型,通过特征提取待检细胞的14个特征向量(包括6个形态特征、6个色度特征、整幅图像的红色分量平均值和细胞区域的灰度平均值)模拟人脑思维对样本进行学习 and 训练,实现对肺癌细胞常见的三种类型(鳞癌、腺癌、小细胞癌)的分类诊断,并通过其强大的学习功能与再学习经验累积功能逐步达到错误识别率的最小化。但是,与肿瘤细胞学诊断受到取材局限、细胞病理专家诊断经验差异而导致某些错误结果一样,初期的肺癌细胞病理诊断系统与细胞病理诊断结果的总符合率仅为91.80%,还存在着部分难以判断的核异型细胞和假阳性、假阴性结果。究其原因,可能与多方面因素影响有关,如:炎症所致变形巨噬细胞的误判、涂片背景杂质及染色差异、细胞重叠成堆难以区分等,对于这部分细胞涂片我们建议由有经验病理医生复核以避免误差。

2 改良的“智能化肺癌细胞病理诊断系统”

为了解决初期研究工作中所遇到的肺癌细胞判別诊断上的问题,我们将强化学习、分类器集成和降维方法等计算机技术运用于肺癌细胞病理诊断设计制作了“智能化肺癌细胞病理诊断系统”。基于强化学习的细胞分割方法可有效区分细胞核和细胞浆,

解决了肺癌细胞图像中细胞核与细胞浆之间对比度较低、细胞边缘与背景之间边界模糊及背景杂质噪声产生的影响,能适应不同细胞学涂片的复杂环境及染色背景。采用基于B样条及改进deBoor-cox方法对重叠细胞进行分离和重构,通过确定重叠细胞的分离点、模拟重叠细胞缺失边缘来还原细胞图像,有效解决了穿刺过程中采集细胞数量较少及重叠细胞不易判断的问题,为更准确对细胞进行分析与识别提供了更多的样本。在这个阶段为更好地描述细胞特点提高肺癌细胞分类精度,系统提取40个特征向量,包括:几何形态特征(如:细胞周长、宽度、高度、面积、似圆度、矩形体、伸长度、傅立叶描述子、中心距)、区域几何特征(如:区域内细胞总数目、总面积、细胞面积与似圆度均方差)、色度特征(如:R、G、B和H、I、S)、光密度特征(如:综合光密度、平均灰度、光密度方差、特征灰度)、纹理特征(如:能量、熵、最大概率、对比度、逆差)。我们以细胞特征为属性,搜索产生了四种属性集合用于贝叶斯、支持向量机、K近邻和决策树四种分类器,最终集成产生肺癌分类结果,增加了分类器的可靠性。若四种分类器的结果不一致,则暂定义为未定型细胞。我们将已有术后组织病理明确诊断的细胞涂片中的癌细胞图像收集起来建立了一个直观的各类型肺癌细胞图库,采用基于LDA等降维方法对未定型细胞采用图库比对的方法进一步辅助判別肺癌细胞类别。通过这个阶段的改进,“智能化肺癌细胞病理诊断系统”的肺癌识别诊断率达到了94.80%,核异型细胞识别率为74.20%,假阳性率为1.85%,假阴性率为3.35%。

3 基于代价敏感图像层的“肺癌细胞病理诊断系统”

前期工作中,我们所采用的图像层方法称为多分类多示例自适应增强法,这是一种代价不敏感算法,其潜在条件是假设不同图像之间的错误分类代价是相同的,即假定识别诊断过程中出现的假阳性和假阴性结果所导致的后果相同。这一假设在患者后续实际产生的医疗行为中并不成立,假阳性结果导致的错误划分将会对患者产生更为严重的后果,如:心理压力、过度治疗带来的身心创伤和费用的增加等。由于假阳性和假阴性结果所产生的医疗代价

不同,所以在肺癌细胞的判别诊断中权重设定也应该有所区别,我们更需要的是尽可能地避免假阳性结果的误判。代价敏感分类策略可以最小化误分类代价,有效提高识别率^[20]。我们将代价敏感算法引入肺癌细胞病理的识别诊断,形成了基于代价敏感图像层的肺癌细胞病理诊断系统(cost-sensitive image-level lung cancer diagnosis system,CILCDS)。

3.1 方法

收集我院 2008~2011 年期间保存及制片良好、有明确组织病理诊断结果对照的肺穿刺病理细胞学涂片,均采用“肺癌早期诊断立体定位仪”对肺部病灶定位穿刺取检标本,涂片采用湿固定法固定,HE 染色,常规脱水封片。

基于潜在狄利克雷分配模型(latent dirichlet allocation,LDA)方法的基本思想是在最小均方意义下选择最能分开各类数据的特征,使样本在特征空间中类间尽可能彼此远离,类内尽可能紧凑,使样本具有最佳的可分离性,其仅在大样本情况下适用。支持向量机(support vector machine,SVM)虽然在均衡数据的分类中效果良好,但在不均衡数据分类中较为不足。在许多实际应用领域,如:本研究中的假阳性结果和假阴性结果,二者不可能相同,即存在着不均衡的数据集,这种情况下少数类样本的识别率更为重要。代价敏感算法是解决不均衡数据分类的有效手段,代价敏感支持向量机(cost-sensitive support vector machine,CSSVM)能够提高少数类样本的识别率和识别精度。在特征提取及模式识别过程中,CILCDS 同时满足了大样本和少数类小样本的分类分析。CILCDS 中,我们采用 LDA+CSSVM 和 LDA+SVM 方法作为分类器的筛选方法。

为了衡量各种方法在正常/癌细胞分类方法上的结果,我们引入了四种不同的衡量标准:①误分类代价(Cost);②假阳性率(Err_{fp}),即正常样本错分为癌的概率;③假阴性率(Err_{fn}),即癌样本错分为正常的概率;④整体的精度(Accuracy)。在 CILCDS 中,我们引入了 Cost 来衡量整体的错分代价,以控制假阳性和假阴性结果的比例,尽可能减少过度治疗或漏诊对患者所产生的不良后果。实验结果采用 10 折交叉验证法作为验证方法。

3.2 实验结果

通过实验我们发现,在整体错分代价上代价不

敏感方法和代价敏感方法相比有着更高的错分代价(Table 1)。通过与基于轴平行矩形(axis-parallel rectangles,APR)、互信息支持向量机(mutual information support vector machine,mi-SVM)、K 近邻(citation K-nearest neighbors,citation-kNN)、集成多样性密度(ensemble of diversity density,EM-DD)、多分类多示例自适应增强法(multi-class multi-instance adaptive boosting,MCMI-AdaBoost)进行对比,CILCDS 所采用的 LDA+CSSVM 和 LDA+SVM 方法取得了最低的癌细胞错分概率 0.37%和最高的准确度 97.61%,这与前期工作相比有了很大提高。

Table 1 Results of different methods

Method	Cost	Err _{fp} (%)	Err _{fn} (%)	Accuracy (%)
APR	192	2.85	11.81	85.24
Citation-kNN	135	2.21	5.54	92.25
EM-DD	244	4.06	8.86	87.08
MI-SVM	276	4.80	5.90	89.30
mi-SVM	116	1.85	5.90	92.25
MCMI-AdaBoost	134	2.21	5.17	92.62
LDA+SVM	106	1.85	2.21	97.61
LDA+CSSVM	41	0.37	5.75	93.75

尽管假阴性率略高,但我们的目标是在折中假阳性率和假阴性率时最小化整体错分代价,以尽可能减少误诊和漏诊。实验中,我们发现 MCMI-AdaBoost、mi-SVM 和 KNN 整体的准确率也都超过了 90%,说明这三种方法的检测结果与术后组织病理结果符合率也较高,而 mi-SVM 的假阳性率甚至低于 2%,也是比较好的选择。

我们定义总体错分代价为,研究结果(Figure 1)表明,与其他方法相较,CILCDS 中 Cost_{cn} 和 Cost_{nc} 的比值对 Cost_{total} 影响最小,且 Cost_{total} 值最低。我们进一步将 MCMI-AdaBoost、mckNN 和 CILCDS 进行比较,实验结果(Figure 2)表明 CILCDS 在三种方法中 Cost_{cn}、Cost_{nc}、Cost_{cc} 和 Cost_{total} 值均为最低,取得了最低的整体错分代价。

4 结 语

不同类型的分类错误往往对应着不同的代价,肺癌细胞识别诊断过程中判定的假阳性和假阴性结果亦对应着不同的代价,而通常假阳性结果较之假

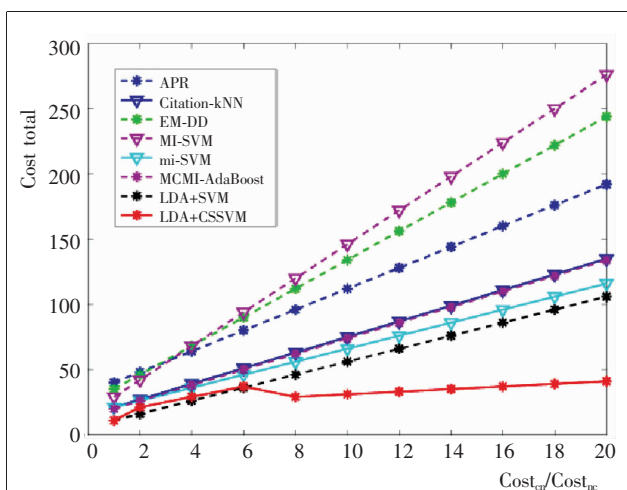


Figure 1 Influence of different Costcn/Costnc ratios on the experiment results

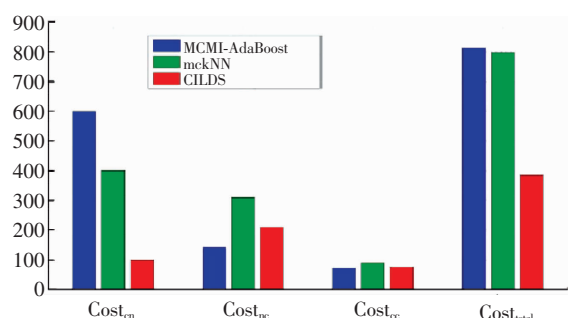


Figure 2 Costcn, Costnc, Costcc and Costtotal of MCMI-AdaBoost, Citation-KNN and CILCDS

阴性结果对患者产生的医疗代价更高，也更易导致医疗纠纷。代价敏感分类就是为不同类型的错误分配不同的代价，注重在分类时尽量避免产生高代价的错误，以达到分类时产生的错误总代价最低的目标，这种思想更加符合现实中的分类问题。将基于代价敏感图像层技术应用于肺癌细胞的识别诊断能够有效减少肺癌细胞病理诊断过程中的假阳性结果，提高识别精度。实验结果的综合分析提示，CILCDS 由于引入了 Cost 来衡量整体的错分代价，较好地控制了假阳性和假阴性结果的比例，在癌与正常细胞判断以及癌细胞分类识别诊断过程中能够取得更低的错误分类代价以及较好的组织病理结果符合率，与多种实验方法比较能显示出一定的优势，更适用于肺癌细胞病理诊断。

十余年来，我们在“肺癌细胞病理诊断系统”的研制过程中不断摸索改进，并在临床工作中通过大

量样本的实践验证使之更加优化完善。但是，任何肿瘤的早期诊断都不是靠完全单一的方法来完成的，肺癌早期诊断更是如此，必定是多种诊断方法的互补^[21,22]。随着我们研究工作的不断深入和实验样本的扩充，肺癌细胞病理诊断系统的稳定性和可靠性将进一步得以提升，有可能成为临床肺癌诊断较好的辅助手段之一。

参考文献：

- [1] Ye YK, Shao C, Ge XZ, et al. Study and clinical application of a stereoscopic apparatus for early diagnosis of lung cancer [J]. Chinese Journal of Surgery, 1992, 30 (5):303-305. [叶玉坤, 邵冲, 葛效忠, 等. 肺癌早期诊断双相立体定位仪的研制及临床应用 [J]. 中华外科杂志, 1992, 30 (5):303-305.]
- [2] Mao YQ, Li N, Lu XQ, et al. The researching and implementation of morphological recognition in the system of diagnosing the early stage lung cancers [J]. Computer Engineering, 1999, 25(8):21-25. [冒宇清, 李宁, 陆新泉, 等. 肺癌早期诊断系统中形态学识别的研究与实现[J]. 计算机工程, 1999, 25(8):21-25.]
- [3] Lu JF, Yang JY, Tang ZM, et al. Design of a separating algorithm for overlapping cell images[J]. Journal of Computer Research and Development, 2000, 37(2):228-232. [陆建峰, 杨静宇, 唐振民, 等. 重叠细胞图像分离算法的设计[J]. 计算机研究与发展, 2000, 37(2):228-232.]
- [4] Yang YB, Li N, Lu XQ, et al. The researching and implementation of chromatic recognition for the early-stage lung cancer diagnosis [J]. Computer Engineering and Applications, 2000, 36(3):44-46. [杨玉彬, 李宁, 陆新泉, 等. 肺癌早期诊断系统中色度识别的研究与实现[J]. 计算机工程与应用, 2000, 36(3):44-46.]
- [5] Lu XQ, Li N, Chen SF, et al. The researching and implementation of morphology and chromatics in the system of diagnosing the early stage lung cancer [J]. Pattern Recognition and Artificial Intelligence, 2000, 13(1):116-120. [陆新泉, 李宁, 陈世福, 等. 形态学和色度学在肺癌早期诊断系统中的研究与实现 [J]. 模式识别与人工智能, 2000, 13(1):116-120.]
- [6] Lu XQ, Li N, Chen SF, et al. The design and implementation of diagnosis system for early stage lung cancer[J]. Journal of Chinese Computer Systems, 2000, 21(1):70-74. [陆新泉, 李宁, 陈世福, 等. 肺癌早期诊断系统的设计与实现[J]. 小型微型计算机系统, 2000, 21(1):70-74.]
- [7] Hua L, Yang YB, Li N, et al. Knowledge-based system for

- early diagnosis of lung cancer [J]. Application Research of Computers, 2000, 17(2):90-92. [花蕾, 杨育斌, 李宁, 等. 基于知识的肺癌早期细胞病理诊断系统[J]. 计算机应用研究, 2000, 17(2):90-92.]
- [8] Lu XQ, Li N, Chen SF. Research on application of morphological, color features and neural network to the recognition of lung cancer cells [J]. Journal of Computer-Aided Design & Computer Graphics, 2001, 13(1):87-92. [陆新泉, 李宁, 陈世福. 形态、颜色特征及神经网络在肺癌细胞识别中的应用研究 [J]. 计算机辅助设计与图形学学报, 2001, 13(1):87-92.]
- [9] Lu XQ, Li N, Chen SF, et al. Approach to the recognition of lung cancer cells on the basis of neural network [J]. Computer Applications and Software, 2001, 18(10):1-6. [陆新泉, 李宁, 陈世福, 等. 基于神经网络的肺癌细胞识别方法的研究[J]. 计算机应用与软件, 2001, 18(10):1-6.]
- [10] Zhou ZH, Jiang Y, Yang YB, et al. Lung cancer cell identification based on artificial neural network ensembles [J]. Artif Intell Med, 2002, 24(1): 25-36.
- [11] Yang YB, Chen SF, Lin H, et al. A chromatic image understanding system for lung cancer cell identification based on fuzzy knowledge [J]. Lecture Notes in Computer Science, 2004, 3029: 392-401.
- [12] Zhang H, Wang CJ, Ye YK, et al. SARGES: design and implementation of an expert system for diagnosis SARS [J]. Computer Engineering and Applications, 2004, 40 (18): 217-220. [张华, 王崇骏, 叶玉坤, 等. SARGES:SARS 医疗辅助诊断专家系统的设计和实现[J]. 计算机工程与应用, 2004, 40(18):217-220.]
- [13] Zhang Y, Ye YK, Wang D, et al. Clinical application of image processing and neural network in cytopathological diagnosis of lung cancer [J]. Chinese Journal of Thoracic and Cardiovascular Surgery, 2005, 21(4):238-240. [张纓, 叶玉坤, 汪栋, 等. 图像处理和人工神经网络在肺癌细胞病理诊断中的应用[J]. 中华胸心血管外科杂志, 2005, 21(4):238-240.]
- [14] Wang D, Ye YK, Zhang Y, et al. Implementation and application of a digital cytopathological lung cancer diagnosing system [J]. Chinese Journal of Surgery, 2007, 45(20): 1417-1419. [汪栋, 叶玉坤, 张纓, 等. 数字化肺癌细胞病理诊断系统的研制和临床应用 [J]. 中华外科杂志, 2007, 45(20):1417-1419.]
- [15] Zhang Y, Wang D, Ye YK, et al. Study on computer-aided diagnosis of cytopathological lung cancer [J]. Chinese Clinical Oncology, 2008, 13(10):926-928. [张纓, 汪栋, 叶玉坤, 等. 计算机辅助肺癌细胞病理诊断的初步研究[J]. 临床肿瘤学杂志, 2008, 13(10):926-928.]
- [16] Shi YH, Gao Y, Wang RL, et al. Transductive cost-sensitive lung cancer image classification [J]. Applied intelligence, 2013, 38(1):16-28.
- [17] Zhang Y, Wang D, Ye YK, et al. A cost-sensitive classification strategy and its application to lung cancer cell recognition [J]. Journal of Medical Postgraduates, 2012, 25 (6):567-570. [张纓, 汪栋, 叶玉坤, 等. 代价敏感分类策略在肺癌细胞识别诊断中的应用 [J]. 医学研究生学报, 2012, 25(6):567-570.]
- [18] Lu JF, Yang JY, Hu XL, et al. Creation of a computer-assisted real-time localization system for early diagnosis of lung cancer [J]. Chinese Clinical Oncology, 2001, 6(4):314-316. [陆建峰, 杨静宇, 胡雪蕾, 等. 计算机辅助肺部小病灶穿刺定位系统的建立[J]. 临床肿瘤学杂志, 2001, 6(4): 314-316.]
- [19] Zhu H, Wang D, Ye YK, et al. Implementation of double-C-arm synchronous real-time X-ray positioning system computer aided for aspiration biopsy of small lung lesion [J]. Chinese Journal of Radiology, 2007, 41(12):1353-1356. [朱宏, 汪栋, 叶玉坤, 等. 肺小病灶穿刺诊断同步实时定位跟踪 X 线立体定位系统的研究[J]. 中华放射学杂志, 2007, 41(12):1353-1356.]
- [20] Jiang XT, Zhou XH. Network intrusion detection based on genetic neural network misclassification cost sensitive [J]. Computer Systems & Applications, 2011, 20(6):49-51. [蒋贤特, 周晓慧. 基于遗传神经网络的误分类代价敏感网络入侵检测[J]. 计算机系统应用, 2011, 20(6):49-51.]
- [21] Ye YK. The molecular events in the development of lung cancer and early diagnosis of lung cancer [J]. Chin J Surg, 2000, 38(7):485-487. [叶玉坤. 肺癌发生发展中的分子事件与肺癌早期诊断[J]. 中华外科杂志, 2000, 38(7):485-487.]
- [22] Wang D, Ye YK, Su CQ, et al. Detection of multiple gene mutations in non-small-cell lung cancer by suspension microarray [J]. National Medical Journal of China, 2009, 89 (48):3393-3396. [汪栋, 叶玉坤, 苏长青, 等. 液相芯片技术检测非小细胞肺癌多基因突变 [J]. 中华医学杂志, 2009, 89(48):3393-3396.]