数据挖掘在喉癌临床研究中的应用*

徐雯[1] 张睿[2] 鄂琪敏[1] 刘加林[1,3] 梁传余[3]

[摘要] **目的**:通过喉癌临床数据的挖掘与分析,为喉癌的早期预测诊断提供决策判断依据。**方法**:在后关系型 Caché 数据库上建立喉癌数据仓库,结合 IBM Cognos 进行联机在线分析,实现喉癌数据多维呈现,并构建喉癌科研数据集于 Weka 环境中实现特征选择及数据挖掘。**结果**:构建喉癌数据仓库;实现喉癌数据多维呈现;通过数据挖掘提取喉癌的集成式和二段式特征子集,发现与喉癌高度相关的潜在特征。**结论**:运用数据挖掘技术对发现喉癌潜在临床知识、辅助临床诊断具有重要的价值。

[关键词] 喉肿瘤;数据挖掘;数据分析

doi:10.13201/j.issn.1001-1781.2015.14.011

[中图分类号] R739.65 [文献标志码] A

The clinical application of data mining in laryngeal cancer

XU Wen[1] *ZHANG Rui*[2] *E Qimin*[1] *LIU Jialin*[1,3] *LIANG Chuanyu*[3]

([1]Department of Medical Informatics, West China School of Medicine, Sichuan University, Chengdu, 610041, China;[2]Information Center, West China Hospital, Sichuan University;[3]Department of Otolaryngology-Head and Neck Surgery, West China Hospital, Sichuan University)

Corresponding author: LIU Jialin, E-mail: dljl8@163.com

Abstract **Objective**: To provide a basis for the prediction and early diagnosis of laryngeal cancer through data mining and analysis. **Method**: We constructed a laryngeal cancer data warehouse on Caché and combined IBM Cognos for on-line analytical processing. After building research data set, software Weka was used for feature selection and data mining. **Result**: Data warehouse of laryngeal cancer was structured and Laryngeal cancer data were multidimensional presented. It is possible to find the high relevant and potential characteristics associated with laryngeal cancer by integration and 2-phase feature subset of data mining to extract the larynx. **Conclusion**: Application of data mining technology could help clinicians to find potential clinical knowledge and make early diagnosis of laryngeal cancer.

Key words laryngeal cancer; data mining; data analysis

喉癌是耳鼻咽喉常见的恶性肿瘤,其发病率占全身恶性肿瘤 5.7%～7.6%[1]。据研究显示,早期诊断治疗可明显改善喉癌患者的预后[2]。目前喉癌的主要诊断方法为喉镜和病理学检查,两种检查手段都使患者造成不同程度的创伤。文献报道癌胚抗原、肿瘤坏死因子等 20 余种血清肿瘤标志物可用于喉癌的诊断及预后监测,但由于缺乏特异性和敏感性,未能在临床上广泛应用。近年来,通过数据挖掘技术获取辅助恶性肿瘤早期预测诊断的临床信息已成为一种趋势[3-4]。而国内对临床数据的挖掘研究较少,未能充分利用日益膨胀的临床数据。目前,国内喉癌领域的临床数据仍处于待挖掘状态,由于喉癌临床高维数据蕴藏着丰富的信息,可能存在有利于预测喉癌的某类隐藏指标,本研究旨在通过数据挖掘技术探索此类指标,提高喉癌诊断的准确率。

本研究在已建立的喉癌临床数据库[5]基础上,结合华西医院的电子病历信息,构建喉癌数据仓库。借助联机分析处理实现数据多维呈现,并依附数据仓库构建喉癌多维科研数据集,结合 Filter 算法对数据集进行特征选择,同时对特征选择结果进行临床分析。

1 材料与方法

1.1 技术平台

喉癌数据仓库基于后关系型数据库 Caché,联机分析处理采用 IBM Cognos。特征选择实验环境为四川大学华西医院数据仓库测试服务器,其硬件环境:CPU 为至强 X5650 处理器,内存 12G。软件环境:Windows 2003 企业版 64 位操作系统,Weka(Waikato Environment for Knowledge Analysis)为 3.6.3 版本。

* 基金项目:国家自然科学基金面上项目(No:71273182);四川省科技支撑计划项目资助(No:2011FZ0035)

[1]四川大学华西医院 华西临床医学院医学信息学教研室(成都,610041)

[2]四川大学华西医院信息中心

[3]四川大学华西医院耳鼻咽喉头颈外科

通信作者:刘加林,E-mail:dljl8@163.com

1.2　数据仓库的构建

1.2.1　数据来源　本研究数据来源于已构建的喉癌临床数据库及四川大学华西医院电子病历系统，抽取 2011-01－2012-08 期间喉癌与声带息肉（无喉癌病史）出院患者的诊疗数据，整合患者症状体征、实验室检查、耳鼻咽喉专科查体及相应病理活检等信息。

1.2.2　结构设计　数据仓库结构设计主要包括数据逻辑模型、事实表及维度表。逻辑模型采用星型模型；数据仓库共包含患者信息表、诊断表、实验室检查表等 16 个事实表；维度表共建立 30 余个，包括疾病字典表、诊断确诊字典表、手术字典表等。喉癌数据仓库以“患者信息表”为核心，其余事实表均通过“关联字段”与其衔接；而事实表与维度表间通过具有的相同“关联字段”，如“病案号”进行连接。

1.2.3　ETL 过程及数据预处理　由于低质量的数据会对挖掘结果及解读造成重大偏差[6]，我们进行数据的 ETL（Extract，Transform，Load）过程。根据原始数据实际情况，本研究主要进行异构数据整合、异常与错误数据处理、计量与等级资料的统一、缺失值处理等数据预处理工作。最后 ETL 程序会将预处理后的数据保存到数据仓库相关事实表中。

1.3　基于喉癌数据仓库的联机分析处理及数据多维呈现

为方便临床医生了解喉癌患病人群的基线特征，我们将构建联机分析处理（On-Line Analytical Processing，OLAP）系统，以临床医生易理解的方式宏观反映喉癌数据的真实情况，帮助医生达到深入理解数据的目的。

1.4　喉癌科研数据集的构建及特征选择

根据临床分析需求，将数据仓库中多张事实表整合为一张二维平面表，形成科研数据集。考虑到年龄、性别因素对患者生化指标等各方面均有较大影响，最终实验构建喉癌高年龄段男性患者数据集（LaCaOldM）以探索喉癌的高相关特征因素。为减小计算复杂度、提高预测精度，采用集成式及二段式特征选择方法对喉癌科研数据集进行特征选择实验。①LaCaOldM 数据集：喉癌患者（排除复诊和术后患者）作为病例组，选择对全身生化水平影响较小（近似健康人群）的声带息肉患者作为对照组，两组均为年龄≥50 岁的男性患者，并选取患者入院后的首次就诊检查结果作为数据项。喉癌病例组纳入 327 例，声带息肉对照组纳入 214 例；②实验采用 Weka 软件；③使用 5 种单因素 Filter 方法（X2 Statistic，Information Gain，Gain Ratio，Relief F，Symmetrical Uncertainty）对 LaCaOldM 数据集进行特征选择和统计学单因素分析，形成精选后的集成式特征子集，对被选择的喉癌特征属性进行临床分析及评价；最后在集成式特征子集基础上，应用 2 种多因素 Filter 方法（CFS，Correlation-based Feature Selection 和 CON，Consistency-based Feature Selection）去除冗余特征，形成相应的二段式特征子集。

2　结果

2.1　喉癌数据仓库的构建

喉癌数据仓库可实时、持续地获得最新数据，并且灵活方便地连接新接口以整合丰富的临床信息，不仅利于多维数据的联机分析处理，更可根据不同科研需求便捷地获取相应的研究数据集，为临床数据分析及数据挖掘提供数据基础。

2.2　基于喉癌数据仓库的联机分析处理及数据多维呈现

利用 Cognos 提供的 Analysis 及 Report Studio 功能，对喉癌数据的发病年龄、性别等区间分布进行初探。利用 Analysis Studio 选择 2011-01－2012-08 期间喉癌出院患者进行 OLAP 分析，男 478 例，女 16 例，男女患病比例为 30∶1。15～44 岁 38 例，男 35 例，女 3 例；45～59 岁 219 例，男 209 例，女 10 例；60 岁以上 237 例，男 234 例，女 3 例；从年龄分布上看，喉癌随年龄增大其发病例数有升高趋势。

2.3　喉癌数据集特征选择结果及临床分析

运用 5 种单因素和 2 种多因素 Filter 方法分析喉癌特征属性，分别形成 LaCaOldM 数据集的集成式和二段式特征子集（表 1）。二段式特征子集是在集成式特征子集的基础上，去除可能存在的冗余特征而形成，特征选择结果与喉癌具有高相关性。运用 5 种单因素 Filter 方法选择出来的集成式特征子集，其特征属性可以代表喉癌的某些特点，如患者年龄、痰中带血、白蛋白检查等信息都与已知临床知识相匹配。但仍有部分特征属性与临床预期不相符，如血清氯离子、纤维蛋白原（fibrinogen，Fbg）、血清 β 羟基丁酸（β-OHBA）等，这些指标在喉癌组与声带息肉组间差异具有统计学意义（$P<0.01$）。见图 1。

3　讨论

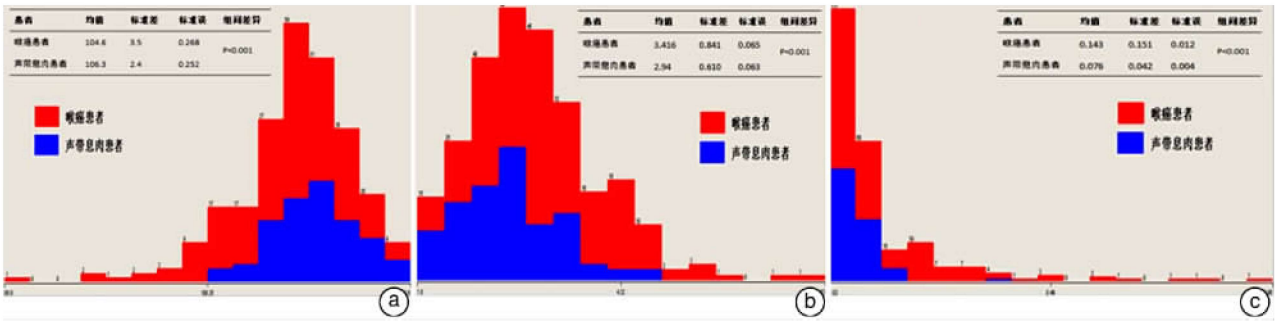
本研究在已建立的一体化喉癌临床数据库基础上，结合电子病历系统中患者的诊疗数据，构建喉癌数据仓库。通过联机分析处理实现数据多维呈现，并构建喉癌数据集。从多角度对喉癌的高相关特征进行分析。

研究中的喉癌高相关性特征主要从基本信息、症状体征信息和检验信息 3 个方面进行描述。基本信息中，年龄因素与喉癌相关性较高，喉癌患者 OLAP 呈现中显示，喉癌患者的年龄集中在 45 岁以上，男性明显多于女性；症状体征信息方面，如呼吸困难，痰中带血，喉部新生物，声带息肉等信息与

表 1 喉癌 LaCaOldM 数据集提取的特征子集结果

特征子集	特征选择方法	特征子集数	特征子集项目(各项目权重无差异)	生成的数据集
集成式特征子集	基于单因素分析后的经验型特征子集	38	基本信息:年龄;检验信息:Fbg 血清 β-OHBA 测定,红细胞计数,血红蛋白,红细胞压积,平均血小板体积,血小板分布宽度,白细胞计数,中性分叶核粒细胞绝对值,淋巴细胞绝对值,单核细胞绝对值,总蛋白,白蛋白,白球比例,尿酸,氯,阴离子间隙,乙肝 e 抗体半定量;症状体征信息:吞咽疼痛,呼吸困难,喉梗阻,喉部新生物,声带息肉,气紧,活动后气紧,痰中带血,体重下降,说话多,声带新生物,吞咽困难,咳嗽,水肿,精神差,说话费力,发声吃力,睡眠差,咳痰。	LaCaOld-MExp
二段式特征子集	基于 CFS 方法的二段式特征子集	11	基本信息:年龄;检验信息:血红蛋白,血小板分布宽度;症状体征信息:吞咽疼痛,呼吸困难,喉部新生物,声带息肉,体重下降,吞咽困难,精神差,发声吃力。	LaCaAl-lExpCsf
二段式特征子集	基于 CON 方法的二段式特征子集	24	基本信息:年龄;检验信息:Fbg,血清 β-OHBA 测定,红细胞计数,红细胞压积,平均血小板体积,血小板分布宽度,中性分叶核粒细胞绝对值,淋巴细胞绝对值,单核细胞绝对值,白蛋白,尿酸;症状体征信息:呼吸困难,喉部新生物,声带息肉,气紧,活动后气紧,痰中带血,吞咽困难,咳嗽,说话费力,发声吃力,咳痰。	LaCaAl-lExpCon

注:CFS,基于相关性的特征选择算法;CON,基于一致性的特征选择算法。



a:血清氯离子在喉癌数据集上的分布;b:纤维蛋白原在喉癌数据集上的分布;c:血清 β 羟基丁酸在喉癌数据集上的分布。

图 1 喉癌组与声带息肉组生化水平之间的差异

喉癌存在高相关性;检验信息方面,如 Fbg、血清 β-OHBA 测定等实验室指标都与喉癌关联程度较高。临床应用中,可将这些指标作为喉癌高危人群的临床参考,对喉癌进行预防、筛查,有利于降低喉癌的患病风险与早期诊断,提高喉癌患者的生存质量。

同时,研究发现部分喉癌特征指标与临床预期不符,经查阅文献发现相关研究甚少,对此我们认为可能产生新的临床见解。①血清氯离子:近年来的研究证明血清氯离子与肿瘤细胞增殖凋亡有关[7-8],但国内外尚少有关血清氯离子与喉癌关系的研究报道。余文发等[9]在研究氯通道阻断剂时发现阻断氯通道可诱导 Hep-2 细胞凋亡,初步证明氯离子水平与 Hcp-2 细胞的增殖相关。同时,研究报道血清氯离子的正常范围可能受到年龄因素的影响,但年龄因素排除后,喉癌组氯离子水平仍低于对照组,表明住院患者在喉癌确诊前血清氯离子水平已经降低,对于出现此结果的原因可进一步探究。②Fbg:由于恶性肿瘤患者的血液常常存在明显的高凝状态[10],其对肿瘤诊断具有一定的临床意义。有研究指出不同临床分期喉癌患者 Fbg 检测数据显示,随着喉癌恶性程度的发展,Fbg 发挥的作用更为明显(Wojtukiewicz 等,2003);其含量也会随着喉癌肿瘤的消退而下降至正常,而在治疗无效或复发者体内维持较高水平[11]。鉴于 Fbg 与喉癌的相关性,Fbg 是否可作为监测喉癌病程发展和评估预后的标志物需要进行临床验证。③β-OHBA:目前尚未出现有关 β-OHBA 与喉癌的报道。但 β-OHBA 在喉癌组与对照组间差异明显,提示其对喉癌及声带息肉具有较好的鉴别价值,临床上应对其进行深入研究及验证。

参考文献

[1] 王晓舟, 王钦. S100A4 和上皮性钙黏素在喉鳞状细胞癌中的表达及意义[J]. 临床耳鼻咽喉头颈外科杂志, 2010, 24(21): 993-995.

2 型糖尿病伴阻塞性睡眠呼吸暂停低通气综合征的临床特点和蛋白羰基水平分析

苏丽清[1] 迟海燕[1] 李吉洲[1] 王海静[1] 孙常青[1]

[摘要] **目的**:观察 2 型糖尿病中阻塞性睡眠呼吸暂停低通气综合征(OSAHS)的患病情况,临床特点及血清蛋白羰基水平。**方法**:选择 2 型糖尿病患者 203 例,进行多导联睡眠呼吸监测,记录 AHI、年龄、身高、体重指数(BMI)、腰围、糖尿病病程,测定空腹血糖、糖化血红蛋白(HbA1c)、血清蛋白羰基(PCO)水平。**结果**:2 型糖尿病患者中 OSAHS 患病率为 79.2%,其中重度 30.4%,中度 45.4%,轻度 24.2%。伴 OSAHS 患者的 BMI、腰围、空腹血糖、HbA1c 及血清 PCO 水平均高于未合并 OSAHS 者,差别有统计学意义($P<0.01$)。回归分析发现 HbA1c 与 OSAHS 患病风险呈独立正相关($P<0.05$,$OR=6.11$),HbA1c、BMI、腰围、病程、血清 PCO 水平均与 AHI 独立正相关,HbA1c 为最主要的独立危险因素($P<0.05$)。**结论**:2 型糖尿病患者中 OSAHS 的患病率较高,合并 OSAHS 的患者血糖控制差,体内蛋白质氧化损伤加重。

[关键词] 2 型糖尿病;阻塞性睡眠呼吸暂停;患病率;糖化血红蛋白;血清蛋白羰基
doi:10.13201/j.issn.1001-1781.2015.14.012
[中图分类号] R563.8 [文献标志码] A

Clinical features, levels of protein carbonyl in serum of obstrucitive sleep apnea syndrome with type 2 diabetes mellitus

SU Liqing CHI Haiyan LI Jizhou WANG Haijing SUN Changqing
([1]Department of Endocrinology, Weihai Municipal Hospital, Weihai, 264200, China)
Corresponding author: CHI Haiyan, Email: zhaoweichihaiyan@163.com

Abstract **Objective**: To explore the prevalence, clinical feature and levels of protein carbonyl(PCO) in serum of type 2 diabetes mellitus combining obstrucitive sleep apnea syndrome(OSAHS). **Method**: Two hundred and three patients with type 2 diabetes were taken multi lead sleep detection and their AHI, age, height, body mass index (BMI), waistline, duration of diabetes, fast blood glucose, HbA1c level and level of PCO in serum were recorded. **Result**: The prevalence of OSAHS was 79.2% in 203 patients, serious apnea 30.4%, moderate apnea 45.4%,

[1]威海市立医院内分泌科(山东威海,264200)
通信作者:迟海燕,E-mail:zhaoweichihaiyan@163.com

[2] 马玥莹,刘良发,黄德亮,等. 晚期喉癌患者术后生存回顾性分析[J]. 临床耳鼻咽喉头颈外科杂志,2013,27(15):844-846.
[3] NAM S, PARK T. Pathway-based evaluation in early onset colorectal cancer suggests focal adhesion and immunosuppression along with Epithelial-Mesenchymal transition[J]. PloS One, 2012, 7: 1-14.
[4] AHMED K, ABDULLAH-AL-EMRAN, JESMIN T, et al. Early detection of lung cancer risk using data mining[J]. Asian Pacifc J Cancer Prev, 2013, 14: 595-598.
[5] 鄂琪敏,刘加林,黎勇,等. 基于 Web 的一体化喉癌临床数据库的构建及应用[J]. 临床耳鼻咽喉头颈外科杂志,2014,28(15):1181-1184.
[6] TING S, SHUM C, KWOK S, et al. Data Mining in Biomedicine: Current Applications and Further Directions for Research[J]. J Software Engineering Appl, 2009, 2: 150-159.
[7] RENAUDO A, L'HOSTE S, GUIZOUARN H, et al. Cancer cell cycle modulated by a functional coupling between sigma-1 receptors and Cl-channels[J]. J Biological Chem, 2007, 282: 2259-2267.
[8] 杨胜萍,张青云. 七种恶性肿瘤血清离子检测水平分析[J]. 中国现代医学杂志,2014,24(1):9-12.
[9] YU W F, ZHAO Y L, WANG K, et al. Inhibition of cell proliferation and arrest of cell cycle progression by blocking chloride channels in human laryngeal cancer cell line Hep-2[J]. Neoplasma, 2009, 56: 224-229.
[10] YAMASHITA H, KITAYAMA J, TAGURI M, et al. Effect of preoperative hyperfibrinogenemia on recurrence of colorectal cancer without a systemic inflammatory response[J]. World J Surg, 2009, 33: 1298-1305.
[11] 李大伟,董频. 喉鳞癌患者凝血功能与肿瘤复发的关系[J]. 现代肿瘤医学,2010,18(9):1721-1722.

(收稿日期:2015-01-30)