DOI:10.13878/j.cnki.jnuist.2017.04.008

黄志刚[1]　刘虹[1]　刘娟[1]　张岐山[1]

基于 C5.0 算法的胃癌生存预测模型研究

摘要

我国的胃癌发病率高,每年新增胃癌患者占全世界每年新增数量的 42%,胃癌成为我国恶性肿瘤防控的重点.本文针对胃癌数据的特征,给出数据预处理和集成方法;采用 C5.0 分类算法,构建了胃癌生存预测模型,并首次采用美国癌症研究所的 SEER 数据库进行预测实验.实验结果表明:C5.0 预测的精确度、特异性均高于 BP-神经网络算法;胃癌患者的出生地点与最终的存活状态之间存在较强的相关性.该研究是数据挖掘技术在医学领域的一个实际应用,对胃癌的临床诊断具有一定的参考价值,可为医生制定合理的治疗和预防方案提供一定参考.

关键词

数据挖掘;C5.0 分类算法;胃癌;生存预测;SEER 数据库

中图分类号 TP391

文献标志码 A

收稿日期 2017-06-28

资助项目 国家自然科学基金(71473039)

作者简介

黄志刚,男,博士,教授,博士生导师,研究方向为数量经济、金融工程及智能信息处理等.kyc1963@ fzu.edu.cn

1　福州大学 经济与管理学院,福州 350116

0　引言

随着信息和医疗管理网络的发展,电子病历系统成为了医院病历现代化管理的前提条件.如何使用结构化电子病历描述复杂的疾病,为临床诊断和治疗提供科学的决策并辅助临床研究,已成为人们关注的焦点[1].目前,随着医院信息系统的普及,怎样从这些数量众多的电子病历中发现隐藏的有利用价值的信息,已成为电子病历研究中的热点问题.

随着大型医院信息管理系统的发展,数据挖掘方法在电子病历的应用研究主要集中在疾病的诊断、预测、分类或分级和相关因素分析等方面.1)Kusiak 等[2]用两种算法对实体性肺结核进行了诊断分析;Fiaschéé 等[3]运用统计方法和计算智能技术来识别基因诊断指标,对急性移植物抗宿主病(GVHD)准确诊断进行了研究;赵一鸣[4]采用分类回归树方法对结、直肠癌的病例进行了分析;孙清等[5]运用支持向量机-微量元素法建立了胃癌模式识别和诊断的辅助手段;还有采用数据挖掘技术进行胃癌的临床辨治[6]、用药规律[7]和化疗药物不良反应关联的研究[8].2) Wiggins 等[9]用粗糙集理论对心脏病患者术后是否会发生房颤进行了预测;非参数决策树算法被 Das 等[10]用来预测肺癌患者在肺部化疗以后患肺炎的概率;李辉等[11]从肿瘤基因表达谱分析入手建立了肿瘤预测模型.3)Tung 等[12]依据基因数据建立了模糊系统确定小儿淋巴细胞性白血病的具体类型;Mitra 等[13]则是利用智能计算研究了宫颈癌恶性程度的自动分级;李建更等[14]并行分析了胃癌微阵列数据集,采用遗传算法与支持向量机相结合提取特征基因.4)Prather 等[15]通过对医院妇产科中心的数据进行挖掘发现了导致早产的 3 个因素;而 Dutau 等[16]应用决策树研究了诱使儿童慢性或周期性感冒的原因;Zhang 等[17]采用支持向量机评估了胃癌淋巴腺转移在胃癌长期生存中的地位;王文文等[18]将聚类和支持向量机用于胃癌患者住院费用的预测研究.

胃癌是常见的恶性肿瘤之一,它的发病率仅次于肺癌,居世界第二位.我国是胃癌的高发区,每年新增患者达 40 万人,占到全世界发病人数的 42%[19],患病率和死亡率均是世界平均水平的两倍多,约 2~3 分钟就有 1 个中国人死于胃癌,因此胃癌成为我国恶性肿瘤防控的重点[20].常用的胃癌诊断方法对肿瘤标志物检测的灵敏度和特异性均较低.大多数的患者在发现病情时都已经是中到晚期,

因此,如何早期发现病症并合理规范治疗是提高胃癌患者存活率的关键,这对提高胃癌早期诊断和防治有着重要意义.

本文针对胃癌确诊时期晚、死亡率高的特点,从SEER提供的数据出发,采用C5.0算法构建胃癌存活时间的预测模型,以期帮助医生做出更为合理的决策和治疗方案.

1 C5.0算法

C5.0是决策树中的经典算法,根据统计学上的置信区间来进行估计,其核心问题是误差估计和修剪标准的设置,算法的基本思路是:1)对于决策树中的每个叶节点,输出变量的多数类别将作为最后的预测类别;2)假设第 n 个叶节点含有 X 个观测,其中有 Y 个错误预测,那么错误率,即误差为 $w_n = X/Y$;3)对第 n 个叶节点的真实误差 e_n 在近似正态分布假设的基础上进行区间估计,置信度设定为 $1-\alpha$,则有

$$P\left(\frac{w_n - e_n}{\sqrt{\frac{w_n(1-w_n)}{X}}} < \left| z_{\alpha/2} \right|\right) = 1-\alpha,$$

其中,$z_{\alpha/2}$ 是临界值,那么第 n 个叶节点 e_n 的置信上限,即C5.0算法的默认置信度为 $1-0.25=75\%$,当 $\alpha = 0.25$ 时,$z_{\alpha/2} = 1.15$.

2 数据处理

2.1 数据的准备

由于我国医院还未完全实现医疗信息化,加之出于隐私数据的保护,国内医疗数据信息公开化程度不高,医疗数据库尚未完善.SEER项目是美国国家癌症研究所监控美国9个注册地的数据,并将这些数据免费提供给以分析研究为目的的机构和实验室.本文的实验数据是从SEER网站(http://www.seer.cancer.gov)上申请的1973—2009年登记的确诊为胃癌的病例.由于本文是研究胃癌患者的生存性和存活时间,因而选取SEER数据集中的DIGOTHR.txt,它包含308 155名随访患者的记录,含有所有消化类癌症的记录,因而需先对数据进行初步的筛选,其规则如表1所示.初步筛选使得所有记录都只患胃癌一种癌症.由于较多属性适用范围是1988年以后,而1998年以前和2003年以后的数据缺失值较多,为了预测模型的准确性,删除1998年以前和2003年以后的记录,选取1998—2002年的数据用于实验.

表1 数据筛选规则

Table 1 Data filtering rules

规则1:	Sequence Number-Central必须是00,这样说明只患有一种癌症. Sequence Number-Central是描述病人一生中可报告的恶性、原位、良性、不明确原发性肿瘤的数量和顺序的编码.如果一个人之前被诊断只有一个恶性肿瘤,随后发现存在第二个恶性肿瘤,则Sequence Number-Central由00变为01.
规则2:	Primary Site是C160—C166,C168—C169,确保患的是胃癌. Site代表原发瘤产生的位点,具体编码参考国际疾病分类肿瘤学第三版(The International Classification of Disease For Oncology,Third Edition,ICD-O-3).胃癌位点的编码范围为C160—C166,C168—C169,具体位点如表2所示.

表2 胃癌位点编码

Table 2 Gastric cancer node cooling

C160	贲门,未特指
C161	胃底
C162	胃体
C163	胃窦
C164	幽门
C165	胃小弯,未特指
C166	胃大弯,未特指
C168	胃部的重叠性病灶
C169	胃,未特指

2.2 数据的集成

由于DIGOTHR.txt所含数据量太大,需要先借助文本分割器将DIGOTHR.txt分成5个较小的文本文件,经过初步筛选,然后再将筛选的结果集成,得到的数据如图1所示.

	A	B	C	D	E	F	G
1	Patient_ID_number	Marital_Status_at_DX	Race/Ethnicity	Age_at_diagnosis	Birth_Place	Primary_Site	Histologic_Type_ICD-O-3
2	8990549	5	14	79	7	5	8
3	10630693	4	14	83	7	1	5
4	10636507	2	14	81	5	4	5
5	10637310	5	14	79	7	5	8
6	10639853	2	14	67	7	1	5
7	10642639	5	3	81	244	3	5
8	10647108	2	14	82	11	6	8
9	10647384	2	14	49	455	7	5
10	10648276	2	14	73	7	1	5
11	10649889	5	14	89	431	7	5
12	10650073	2	14	63	7	4	5
13	10651078	2	14	71	5	2	5
14	10651477	2	14	54	451	4	8
15	10651852	2	14	81	14	4	8

图1 初始数据集

Fig. 1 Initial data set

3 C5.0建模及实验

3.1 C5.0建模过程

对预处理后的胃癌数据,采用随机抽样选取1 264个样本作为训练样本集,另外542个样本作为测试样本集进行建模.首先使用C5.0算法根据训练

样本集建立决策树,为避免过拟合问题,C5.0 算法采用自底向上的逐层修剪,修剪后的决策树如图 2 所示.

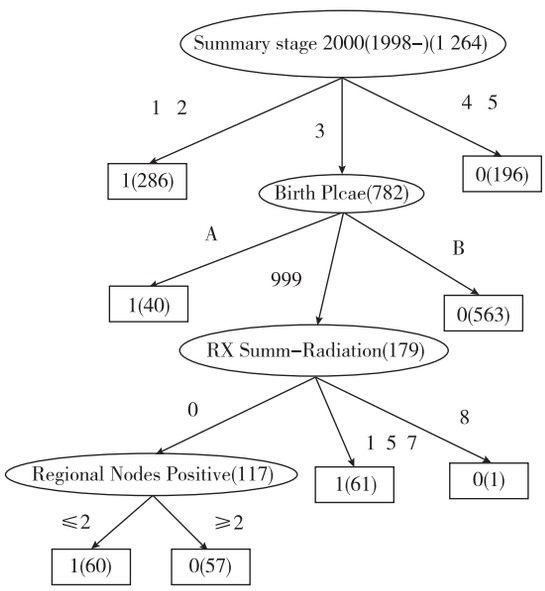

图 2　决策树

Fig. 2　Decision tree

图 2 中,树的最大深度为 4,根节点为癌症总的病变程度,每个节点中括号里的数字代表该节点包含的总样本量.A 与 B 分别指代不同出生地点的编码.

根据修剪后的决策树,通过设置最小样本量和最小置信度,生成规则集.本次实验最小样本量设为 15,最小置信度设为 0.75,产生的规则集由两部分组成,分别代表输出变量的两个类别,具体描述如图 3 所示.根据修剪的决策树共得到 4 条规则,类别 0 含有 3 条,类别 1 有 1 条.由于设置了最小样本量和最小置信度,有些规则就会被舍掉,那么就会存在一些样本不能被覆盖,它们被归为 Default,输出类别为 0.规则后面的括号给出了每条规则覆盖的样本量以及这条规则的正确预测率.例如,类别 0 中的第一条规则共覆盖 563 个样本,它的正确预测率是 88.1%.根据模型预测结果得到 C5.0 算法的混淆矩阵如表 3 所示.

模型的输出结果中也包括输入变量对建模的重要性测度,重要性测度的指标是各个统计检验的 1 - P 的值(P 为概率),它是一个相对值,第 i 个输入变量的重要性定义为

$$Evaluation(i)=\frac{1-P_i}{\sum_i(1-P_i)},\qquad(1)$$

其中,1-P 的值越高,输入变量和输出变量的相关性越大,对输出变量而言也就越重要,所有进入模型的输入变量的重要性之和为 1.

根据式(1),8 个输入变量对 Outcome 的重要性如图 4 所示,其中纵坐标表示的是模型的 8 个输入属性,横坐标为各属性根据式(1)计算得到的 1-P 的值.

```
Rules for 0-contains 3 rule(s)
  Rule 1 for   0  (563,0.881)
    if  Summary stage 2000(1998-)in ["3"]
        and  Birth Place in ["B"]
    then  0
  Rule 2 for 0 (57, 0.754)
    if  Summary stage 2000(1998-)in ["3"]
        and  Birth Place in ["999"]
        and  RX Summ—Radiation in ["0"]
        and  Regional Nodes Positive>2
    then  0
  Rule 3 for 0  (196, 0.954)
    if  Summary stage 2000(1998-) in ["4" "5"]
    then  0
Rules for 1-contains 1 rule(s)
```

图 3　生成规则集描述

Fig. 3　Generating rule set

表 3　C5.0 混淆矩阵

Table 3　C5.0 confusion matrix

C5.0	存活	未存活
预测存活	437	86
预测未存活	144	1 139

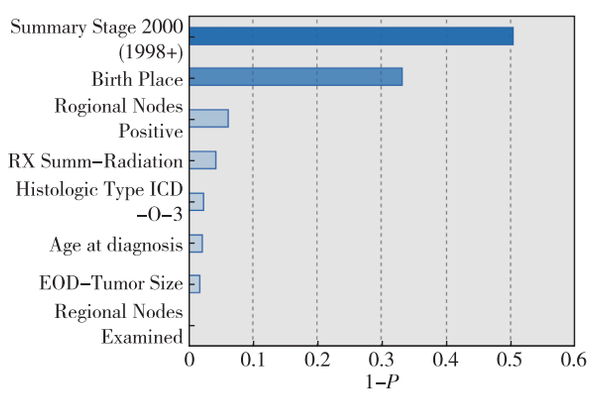

图 4　条件属性的重要性

Fig. 4　Variable importance of conditional attributes

从图 4 中可以发现:1)胃癌患者 5 年后的存活状态主要与癌症总的恶化程度(Summary Stage 2000)、出生地点(Birth Place)、良性区域淋巴结的个数(Regional Nodes Positive)以及第一个疗程中放

射治疗的方法(RX Summ-Radiation) 相关;2)后3个变量和现实是相符的,胃癌确诊时期的早晚一直以来都是影响胃癌存活率的关键因素,而出生地点与最终的存活状态有较大的相关性,相对于以往研究,该结论比较出人意料.

通过分析数据库中5年后没有存活的记录的出生地分布情况发现:我国胃癌的高发地区分布情况与该结果相符,胃癌的高发地区在饮食上都很具有"地方特色",有一个共同点,长时间的食用盐渍食品,比方说腌制的肉类和蔬菜、咸鱼以及海产品等.

3.2 对比实验与分析

比较C5.0算法与BP-神经网络预测结果,结合模型的混淆矩阵计算出精确性、敏感性和特异性,结果如表4所示.结果表明:在胃癌5年生存预测中,C5.0算法的精确度、特异性均高于BP-神经网络,敏感性略低于BP-神经网络;另外,两种算法的敏感性数值均比特异性差,该情况可能是由于样本数据集的非平衡性造成的.因为在筛选后得到的1 806个样本中,没有存活的占了67.83%,而存活的只占了32.17%,导致正类的数量是负类的两倍多;当采用数据建模时,由于预测模型总是追求整体的预测错误率最小,这样整体的高预测正确率往往会掩盖负类的高预测错误率,即模型偏向于正类.这也是实验中特异性总比敏感性高的原因.

表4 精确性、敏感性、特异性比较

Table 4 Comparison of accuracy, sensitivity and specificity %

算法	精确性	敏感性	特异性
C5.0	87.26	75.22	92.98
BP-神经网络	85.71	76.42	90.12

4 结束语

本文针对胃癌患者的生存时间,将C5.0分类算法用于患者生存时间预测模型的构建,给出了数据预处理方法,筛选出与存活时间相关的属性,进而构造了基于C5.0算法的胃癌生存时间预测模型.首次选用SEER数据库中胃癌数据进行胃癌预测实验,将C5.0分类算法和BP-神经网络算法分别进行预测,实验结果表明:1)C5.0算法的精确度、特异性均高于BP-神经网络,而敏感性与BP-神经网络基本持平;2)胃癌属高死亡率的癌症,在预测时可能会造成数据的非平衡;3)胃癌患者的出生地点与最终的存活状态之间存在较强的相关性,即胃癌的发病具有地域特点,因此合理的饮食和良好的生活习惯在一定程度上能提高胃癌的存活率.

本文是数据挖掘技术在医学领域的一个实际应用,对胃癌的临床诊断具有一定的参考价值,可以为医生制定合理的治疗和预防方案提供一定参考.另一方面,本文的实验数据来源于SEER数据库,虽然在建模前对数据进行了预处理,但数据的主观或客观性的偏差仍对最后的实验结果产生了一定的影响.因此,需要对医学数据的预处理技术和方法开展进一步的研究.针对中国的病历数据进行疾病预测研究也是今后的研究方向.

参考文献

References

[1] 王欣萍,李燕.数据挖掘技术于医学电子病历系统的应用[J].现代预防医学,2008,35(13):2450-2451
WANG Xinping, LI Yan. Application of data mining technology in electronic medical records system[J]. Modern Preventive Medicine, 2008, 35(13): 2450-2451

[2] Kusiak A, Kernstine K H, Kern J A, et al. Data mining: Medical and engineering case studies[C]// Proceedings of the Industrial Engineering Research 2000 Conference, Cleveland, Ohio, 2000: 1-7

[3] Fiasché M, Cuzzola M, Fedele R, et al. Computational intelligence methods for discovering diagnostic gene targets about aGVHD[J]. Frontiers in Artifical Intelligence & Applications, 2009, 204: 271-280

[4] 赵一鸣.分类与回归树:一种适用于临床研究的统计分析方法[J].北京大学学报(医学版),2001,33(6):562-565
ZHAO Yiming. Classification and regression trees: A statistical method suitable for clinical researches[J]. Journal of Peking University (Health Sciences), 2001, 33(6): 562-565

[5] 孙清,鞠建峰,曲庆美,等.支持向量机在胃癌诊断预测中的应用[J].食品与药品,2010,12(11):401-404
SUN Qing, JU Jianfeng, QU Qingmei, et al. Application of support vector machine in prediction of gastric cancer [J]. Food and Drug, 2010, 12(11): 401-404

[6] 马梦妍.基于数据挖掘的舒鹏教授治疗胃癌临床病案的回顾性研究[D].南京:南京中医药大学基础医学院,2016
MA Mengyan. A retrospective study based on data mining, Professor Shupeng clinical case of treatment of gastric cancer[D]. Nanjing: College of Basic Medicine, Nanjing University of Chinese Medicine, 2016

[7] 王泽明,柴可群,陈嘉斌.基于数据挖掘的柴可群治疗胃癌用药规律研究[J].江西中医药大学学报,2017,29(1):38-41
WANG Zeming, CHAI Kequn, CHEN Jiabin. Analysis on the medication rules of CHAI Kequn for the treatment of gastric cancer based on data mining[J]. Journal of Jiangxi University of Traditional Chinese Medicine, 2017,

29(1):38-41
[8] 郭佳栋,张雪梅,刘影,等.基于数据挖掘技术对胃癌化疗药物不良反应关联性研究[J].药物流行病学杂志,2017,26(1):46-49
GUO Jiadong, ZHANG Xuemei, LIU Ying, et al. Correlation analysis of gastric cancer chemotherapy drugs adverse drug reaction based on data mining technology [J].Chinese Journal of Pharmacoepidemiology,2017,26(1):46-49
[9] Wiggins M C,Firpi H A,Blanco R R,et al.Prediction of atrial fibrillation following cardiac surgery using rough set derived rules[J]//Conf Proc IEEE Eng Med Biol Soc,2006,1(1):4006-4009
[10] Das S K,Zhou S M,Zhang J N,et al.Predicting lung radiotherapy-induced pneumonitis using a model combining parametric Lyman probit with nonparametric decision trees[J].International Journal of Radiation Oncology Biology Physics,2007,68(4):1212-1221
[11] 李辉,王金莲.基于基因表达谱的肿瘤预测模型研究[J].电子学报,2008,36(5):989-992
LI Hui, WANG Jinlian. Study of tumor molecular prediction model based on gene expression profiles[J]. Acta Electronica Sinica,2008,36(5):989-992
[12] Tung W L,Quek C.GenSo-FDSS:A neural-fuzzy decision support system for pediatric ALL cancer subtype identification using gene expression data [J]. Artificial Intelligence in Medicine,2005,33(1):61-88
[13] Mitra P,Mitra S,Pal S K.Evolutionary modular MLP with rough sets and ID3 algorithm for staging of cervical cancer[J].Neural Computing and Applications,2001,10(1):67-76
[14] 李建更,贺益恒,郭庆雷.基于多数据集的胃癌亚型标志基因选择[J].北京工业大学学报,2013,39(10):1590-1595
LI Jiangeng, HE Yiheng, GUO Qinglei. Marker gene selection of gastric cancer subtype based on multi microarray data sets[J].Journal of Beijing University of Technology,2013,39(10):1590-1595
[15] Prather J C, Lobach D F, Goodwin L K, et al. Medical data mining:Knowledge discovery in a clinical data warehouse[C]//Proc AMIA Annu Fall Symp,1997:101-105
[16] Dutau G, Micheau P, Juchet A, et al. Chronic cough in children:Etiology and decision trees[J].Archives de Pediatrie:Organe Officiel de la Societe Francaise de Pediatrie,2001,8(sup 3):610-622
[17] Zhang X P,Wang Z L,Tang L,et a1.Support vector machine model for diagnosis of lymph node metastasis in gastric cancer with multidetector computed tomography:A preliminary study[J].BMC Cancer,2011,11(1):1-6
[18] 王文文,周涛,陆惠玲,等.基于聚类和支持向量机的胃癌患者住院费用建模[J].中国初级卫生保健,2016,30(2):1-4
WANG Wenwen, ZHOU Tao, LU Huiling, et al. A new model for hospitalization expenses of gastric cancer based on clustering and support vector machine [J]. Chinese Primary Health Care,2016,30(2):1-4
[19] 王永川,魏丽娟,刘俊田,等.发达与发展中国家癌症发病率与死亡率的比较与分析[J].中国肿瘤临床,2012,39(10):679-682
WANG Yongchuan,WEI Lijuan,LIU Juntian,et al.Comparison and analysis of the incidence and mortality rate of cancer in developed and developing countries [J]. Chinese Journal of Clinical Oncology, 2012, 39(10):679-682
[20] 王晓瑜.胃癌研究相关文献热点变化分析[J].临床军医杂志,2015,43(9):955-959
WANG Xiaoyu.A biblimentric analysis on gastric cancer research literature [J]. Clinical Journal of Medical Officers,2015,43(9):955-959

Gastric cancer prediction model based on C5.0 classification algorithm

HUANG Zhigang[1] LIU Hong[1] LIU Juan[1] ZHANG Qishan[1]
1 School of Economics and Management,Fuzhou University,Fuzhou 350116

Abstract The incidence of gastric cancer is very high in China,and the number of new patients diagnosed with gastric cancer accounts for 42% of that of the whole world every year,so gastric cancer has become the focus of the prevention and control of malignant tumors in China.In this paper,the C5.0 classification algorithm is used to predict the survival rate of gastric cancer,and experiments are carried out using the SEER database of the American National Cancer Institute.The data preprocessing and data integration methods are given according to the unbalanced characteristics of gastric cancer record data.The prediction experimental results show that,the accuracy and specificity of C5.0 algorithm are high compared with BP-neural network method;and there is an obvious correlation between birth place and survival state of gastric cancer patients.This study is a practical application of data mining technology in the field of medicine,which has certain reference value for the clinical diagnosis of gastric cancer;it can provide reference for doctors to formulate reasonable treatment and prevention program.

Key words data mining;C5.0 classification algorithm;gastric cancer;survival prediction;SEER