

基于决策树的乳腺癌病历文本的挖掘与决策

龚乐君¹, 张立鹏¹, 李宇茜¹, 吴向辉¹, 高志宏², 潘传迪², 杨庚¹

(1.江苏省大数据安全与智能处理重点实验室,南京邮电大学计算机学院、软件学院、网络空间安全学院,江苏 南京 210023)

(2.浙江省智慧医疗工程技术研究中心,浙江 温州 325035)

[摘要] 乳腺癌是女性最常见的恶性肿瘤之一,严重威胁着世界范围内女性的健康,临床病历文本携带着经验丰富医生对疾病的诊断信息,对其挖掘,可获得乳腺癌相关的病况,从而可以辅助决策. 本文提交了一种方法从文本处理的角度,使用数据挖掘算法-决策树处理病历文本,挖掘乳腺癌疾病相关信息,对乳腺癌进行 TNM 及临床癌症分期决策,并对决策结果进行验证,同时结合 Neo4j 图数据库建立乳腺癌 TNM-临床分期知识图谱,通过实例展示,该方法可得到乳腺癌的 TNM 与临床癌症分期病况. 表明提交的方法有望用来辅助医生进行决策.

[关键词] 乳腺癌,自然语言处理,决策树,文本挖掘,Neo4j

[中图分类号] TP391 **[文献标志码]** A **[文章编号]** 1001-4616(2019)03-0042-10

Mining and Decision-Making of Breast Cancer Medical Record Text Based on Decision Tree

Gong Lejun¹, Zhang Lipeng¹, Li Yuxi¹, Wu Xianghui¹, Gao Zhihong², Pan Chuandi², Yang Geng¹

(1.Jiangsu Key Lab of Big Data Security & Intelligent Processing, School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

(2.Zhejiang Engineering Research Center of Intelligent Medicine, Wenzhou 325035, China)

Abstract: Breast cancer is one of the most common malignant tumors in women, which seriously threatens the health of women worldwide. Clinical medical records carry the diagnostic information from experienced doctors. Mining these records could receive breast cancer-related conditions. This paper presents a method using data mining algorithm-Decision Tree to process medical records, to obtain breast cancer disease-related information via text processing. We conduct TNM and clinical cancer staging decisions for breast cancer and validate decision results. At the same time, we also combine the Neo4j-map database to establish breast cancer TNM-clinical staging knowledge map. The example shows that this method could obtain TNM and clinical cancer grading conditions for breast cancer. It indicates that the presented method is expected to be used to assist doctors in making decisions.

Key words: breast cancer, natural language processing, decision tree, text mining, Neo4j

乳腺癌是女性最常见的恶性肿瘤之一,严重威胁着世界范围内女性的健康. 随着社会的进步和科技的发展,医疗技术手段有了巨大进步,医学工作者发现了更多的乳腺癌种类,乳腺癌相关的病历数据急剧增加.

大量的乳腺癌病历文本和医者手记,承载着丰富的乳腺癌诊断经验. 对这些乳腺癌病历文本和医者手记进行挖掘,可以快速获得乳腺癌的病况,同时提升诊断速度,是帮助病人最快战胜病魔的第一步^[1-2]. 因此,如何设计一个完善的文本处理系统,帮助医疗工作者进行有关乳腺癌的诊断和病重程度的分析,已是一项迫在眉睫的工程.

早在 2002 年,国际计算语言学大会(ACL)及知识发现和数据挖掘挑战杯(KDD)大会就开始关注医学文本挖掘的发展. 随后一批著名的国际大公司、科研院所纷纷投入此领域. IBM 公司在 2004 年一份关于“制药业 2010: 硅技术的事实”中把医学文本挖掘作为制药行业信息化的关键技术来描述. 英国曼彻斯

收稿日期: 2019-07-16.

基金项目: 国家自然科学基金项目(61502243、61502247、61572263)、浙江省智慧医疗工程技术研究中心项目(2016E10011)、中国博士后基金(2018M632349)、江苏省高校自然科学基金(16KJB520003).

通讯联系人: 龚乐君, 博士, 副教授, 硕士生导师, 研究方向: 数据与文本挖掘, 生物医学信息处理. E-mail: glj98226@163.com

特大学的 NacTeM 是世界上第一个以公众资金资助的文本挖掘中心,该中心相继在生物医学文本挖掘领域出了一大批研究成果^[3-7]. 国际期刊《Artificial Intelligence in Medicine》2014 年还发布了一个 special issue 专门征集生物医学文本挖掘方面的文章^[8]. 近十余年来,国内科研院所及高校相继有团队投入这一研究领域,取得了不俗的成绩. 如郑州大学李慧林使用深度学习技术处理妇科电子病历文本,达到较好的诊断效果^[9]. 来自荷兰的 Hazewinkel Mirjam C 及其团队通过对精神病电子病历使用文本挖掘等方式辅助医疗决策,取得显著成果^[10]. 医学文本挖掘主要利用信息化技术对生物医学文献进行加工整理,这些文献大多以文本的形式存在,将这些非结构的文本数据转化为结构化的数据,抽取有趣的生物医学知识. 病历文本数据的挖掘国际研究日益增多,国内的研究偏少. 针对于文本中提取乳腺癌 TNM 及临床分期的预测则少之又少. 本文的工作则是从自然语言角度出发,使用挖掘算法从临床病历文本挖掘乳腺癌这一疾病的相关信息,对乳腺癌进行 TNM 及临床分期决策,最终辅助医疗工作者进行决策.

1 研究方法

本文的工作旨在从临床文本中提取乳腺癌疾病信息,从中挖掘乳腺癌 TNM 与临床癌症分期信息,最终辅助医疗决策. 这个过程涉及到乳腺癌 TNM 与临床分期知识,自然语言处理、决策树算法等.

1.1 乳腺癌 TNM 与临床分期

乳腺癌常有以下症状:乳腺肿块、乳头溢液、皮肤改变、乳头异常、乳晕异常和腋窝淋巴结肿大等症. 按照 2012 年世界卫生组织发布的《WHO 乳腺肿瘤分类》文件,以及组织学的分类标准,乳腺癌可细分为上百种类型. TNM 分期系统是目前国际上最为通用的肿瘤分期系统. 其中,T(Topography),代表原发肿瘤的范围,N(Lymph Node)代表区域淋巴结转移的存在与否及范围,M(Metastasis)代表远处转移的存在与否. 从风险评估的角度上看,使用 TNM 分期标准下的乳腺癌分类更加贴合目的标准^[11-15],并具有最大使用价值和交流价值.

根据文献[16-17],在乳腺癌中,TNM 分期的划分具体内容为如表 1 所示.

表 1 乳腺癌 TNM 分期
Table 1 Breast cancer TNM staging

乳腺癌 TNM 分期			
T	TX		原发肿瘤无法确定(例如已切除)
	T0		原发肿瘤未查出
	Tis	Tis(DCIS)	导管原位癌
		Tis(LCIS)	小叶原位癌
		Tis(Paget)	不伴肿瘤的乳头派杰氏病
	T1	T1mic	微小浸润性癌,最大直径≤0.1cm
		T1a	肿瘤最大直径>0.1cm,≤0.5cm
		T1b	肿瘤最大直径>0.5cm,≤1.0cm
		T1c	肿瘤最大直径>1.0cm,≤2.0cm
	T2		肿瘤最大直径>2.0cm,≤5.0cm
	T3		肿瘤最大直径>5.0cm
	T4	T4a	侵犯胸壁
		T4b	患侧乳房皮肤水肿(包括橘皮样变),溃破或卫星状结节
		T4c	T4a 和 T4b 并存
		T4d	炎性乳腺癌
N	Nx		无法分析
	N0		区域淋巴结无转移
	N1		同侧淋巴结转移,可活动
	N2	N2a	同侧转移性淋巴结相互融合,或与其他组织固定
		N2b	临床无证据显示腋淋巴结转移的情况下,存在临床明显的内乳淋巴结转移
		N3a	同侧锁骨下淋巴结转移及腋淋巴结转移
	N3	N3b	同侧内乳淋巴结转移及腋淋巴结转移
M		N3c	同侧锁骨上淋巴结转移(注:与五版重大变动)
	MX		无法评估
	M0		无远处转移
	M1		有远处转移

癌症临床分期是基于 TNM 分期结果作出的对癌症时间周期的分析,也是对癌症患者所患疾病的病危程度的判断结果. 癌症分期分为 I 期、II 期、III 期和 IV 期 4 大类. 乳腺癌分析中 TNM 分期与临床分级的具体对应细节如表 2 所示.

表 2 乳腺癌 TNM 分期与临床分期的对应情况

Table 2 Correspondence between TNM staging and clinical staging of breast cancer

对应关系描述			
临床分期	肿瘤情况	淋巴结情况	转移情况
0	Tis	N0	M0
I A	T1 *	N0	M0
I B	T0	N1mi	M0
	T1 *	N1mi	M0
II A	T0	N1	M0
	T1 *	N1	M0
	T2	N0	M0
II B	T2	N1	M0
	T3	N0	M0
III A	T0	N2	M0
	T1 *	N2	M0
	T2	N2	M0
	T3	N1	M0
	T3	N2	M0
III B	T4	N0	M0
	T4	N1	M0
	T4	N2	M0
III C	Any T	N3	M0

本文综合以上应用背景结合文本挖掘^[18]及决策树技术^[19],对病历文本进行挖掘,对决策结果进行验证,建立乳腺癌 TNM 分期知识图谱,为医疗工作者在疾病诊断方面的工作增添利器.

1.2 自然语言处理

自然语言处理技术集认知科学、计算机科学、语言学、心理学、数学等多学科领域知识于一身,属于人工智能的研究范畴,涉及人脑语言认知机理、语言知识的表达方式,与现实世界之间的关系等^[20]. 文本挖掘处理的对象是由非结构化的文本数据组成的文档集合,这一文档集合是采用人们最基本、最直接、最方便的自然语言来描述的. 因而自然语言处理成为实现文本挖掘的主要技术手段^[21]. 自然语言是人类发展过程中自然产生、约定俗成的用于人类社会交流的语言. 长期以来,人们一直在追求用自然语言与计算机进行通信. 人们不希望花大量时间和精力去学习不很自然的各种机器语言,更愿意用自己的语言来使用计算机. 自然语言处理的研究不仅具有重要的理论价值,而且具有巨大的实用价值. 自然语言处理也称之为计算语言学或自然语言理解^[22],研究的就是人与计算机之间用自然语言进行有效通信的各种理论和方法. 涉及了词法分析、词性标注、语法分析、语义分析等一系列处理流程. 本工作中使用 NLPIR 项目的 API 来辅助构建病历文本分析系统,处理临床文本记录. NLPIR 大数据语义智能分析平台由北京理工大学大数据搜索与挖掘实验室(Big Data Search and Mining Lab.BDSM@BIT)张华平博士主导建立,是信息抽取的主流工具之一^[23-24]. 主要功能包括中文分词;英文分词;词性标注;命名实体识别;新词识别;关键词提取;支持用户专业词典与微博分析. 在网络文本筛选、警务文本等的处理中,NLPIR 都发挥了重要的作用^[25]. 而在病历文本的分析中,NLPIR 相比其他文本挖掘工作,研究则较少. 因而,本文基于 NLPIR 开发病历文本分析系统,完成病历文本的分词、关键词分析、词频分析和新词发现等工作,采用决策树技术,针对文本分析,挖掘乳腺癌疾病相关信息,对乳腺癌进行 TNM 及临床癌症分期决策,这将辅助医生决策.

1.3 决策树

决策树(Decision Tree)是在已知各种情况发生概率的基础上,通过构成决策树来求取净现值的期望值大于等于零的概率,评价项目风险,判断其可行性的决策分析方法,是直观运用概率分析的一种图解

法. 在机器学习中,决策树是一个预测模型,它代表的是对象属性与对象值之间的一种映射关系. Entropy 表示系统的凌乱程度,使用算法 ID3, C4.5 和 C5.0 生成树算法使用熵. 这一度量是基于信息学理论中熵的概念.

决策树一般自上而下生成. 切割的方法多样,都是为了尝试对目标进行最佳的切割. 构成要素有:(1) 决策节点;(2) 方案枝;(3) 状态节点;(4) 概率枝.

决策树的一般决策程序如下:(1) 绘制树状图,根据已知条件排列出各个方案和每一方案的各种自然状态.(2) 将各状态概率及损益值标于概率枝上.(3) 计算各个方案期望值并将其标于该方案对应的状态结点上.(4) 进行剪枝,比较各个方案的期望值,并标于方案枝上,剪掉劣等方案,所剩的最后方案为最佳方案.

单变量的决策树一次仅使用一个特征值,而多变量的决策树同时使用多个特征值^[26-28].

1.4 ID3 算法(决策树算法)

ID3 算法是一种贪心算法,用来构造决策树,它以香农熵理论作为基础.

香农熵定义为信息的期望值. 求得熵,需要先知道信息的定义. 如果待分类的事务可能划分在多个分类之中,则信息定义为

$$H(X) = - \sum_{i=1}^n p(x_i) \log_2 p(x_i). \quad (1)$$

参考信道中数学模型,设 $U = [u_1, u_2, u_3, \dots, u_r]$ 为信源转发消息, $V = [v_1, v_2, v_3, \dots, v_q]$ 为信宿接收消息,则可使用条件概率 $P(V|U)$ 表示信源发送为 U 、信宿接收为 V 的概率,则有以下公式:

$$\sum_j (v_j | u_i) = 1, i = 1, 2, \dots, r; j = 1, 2, \dots, q. \quad (2)$$

信宿收到 v_j 后,关于 U 的不确定性:

$$H(U|v_j) = - \sum_{i=1}^r P(u_i | v_j) \log_2 P(u_i | v_j); j = 1, 2, \dots, q. \quad (3)$$

由此可得条件熵为:

$$H(U|V) = - \sum_{j=1}^q \sum_{i=1}^r P(u_i | v_j) \log_2 P(u_i | v_j). \quad (4)$$

信息增益为:

$$I(U, V) = H(U) - H(U|V), \quad (5)$$

式中, $H(U)$ 为:

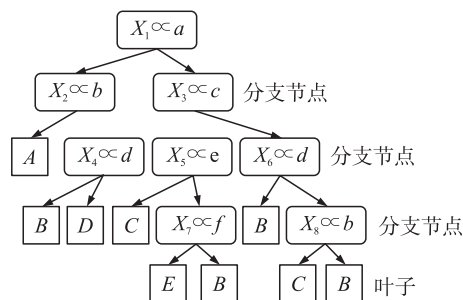
$$H(U) = - \sum_{i=1}^r P(u_i) \log_2 P(u_i). \quad (6)$$

ID3 算法的核心是根据“最大信息熵增益”原则选择划分当前数据集的最好特征,选择熵减少程度最大的特征来划分数据,它比较适用于选择取值较多的属性,数据分得越细确定性越高,信息增益越大. 而病历文本对应的 TNM 分期本身分类详细,导致决策树的属性选择取值较多,和 ID3 算法的思想相吻合.

2 基于决策树的乳腺癌 TNM-临床癌症分期挖掘

2.1 系统构建

本系统数据来源于病历夹 APP 的经典病历库,作为病历文本集,经过数据清洗后,进行病历文本的分词、关键词分析、新词获取等操作,抽取出乳腺癌病例的疾病有关信息,包括病变部位、病变样本、属性名称和相应的属性值. 在提取文本的基础上,将其展现到窗口的表格中,并在后台结合 TNM 分期标准将原发肿瘤的范围、区域淋巴结的受累程度(转移与否与范围)和远处转移到与否信息提炼出来,结合依据决策树



注: X_i 为特征值; a, b, c, d, e 和 f 为阈值; A, B, C, D 和 E 是分配到每个观测中的类标.

图1 决策树分类器

Fig. 1 Decision tree classifier

技术进行乳腺癌的种类判定和病重程度判定,即进行乳腺癌诊断中的决策步骤. 系统流程框架如图 2 所示.

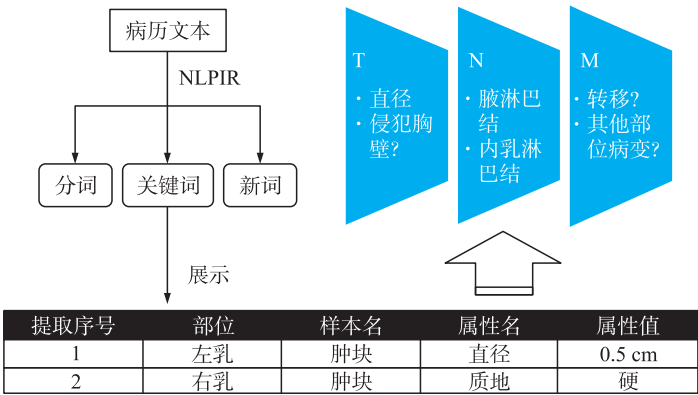


图 2 系统处理流程图

Fig. 2 System processing flow chart

2.2 系统功能实现

本文所开发的病历文本分析系统,将病历文本通过病历文本分词、关键词分析、词频分析等功能,对包含知识的信息进行针对性提取,并按照 TNM 分期依据来进行处理信息,病历文本复杂并有其特殊性,因而要有针对性的处理,详细处理将在相应的功能模块描述. 提取病历文本知识的核心步骤如图 3 所示.

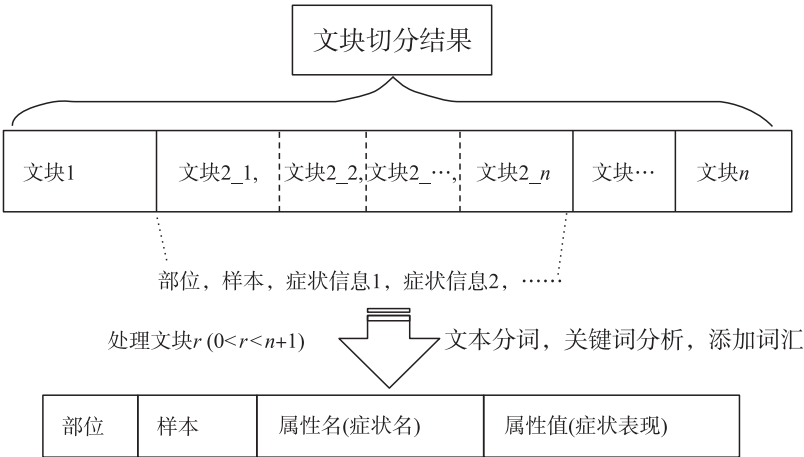


图 3 提取信息的工作流程

Fig. 3 Pipeline for extracting information

- 功能说明:
- 2.2.1 数据清洗
- 为方便下面步骤中文本处理并提高文本处理框的精确度,需要首先完成数据清洗工作^[29-30]. 数据清洗主要包含以下方面:
- (1)同一物体的不同词汇名词统一化;
 - (2)数学符号(如乘号)的统一化;
 - (3)陈述句格式和介词使用的统一化等.
- 数据清洗工作是非常消耗人力和时间的工作,是文本处理提高准确率的基础和根本. 本文所收集的乳腺癌病历文本,来源不一,格式各异. 如出现“肿块”、“肿物”等同义词,对于乳腺癌中肿块大小的描述有“直径”、“长 * 宽 * 高”、“长×宽×高”等对于同一症状的不同角度、格式等描述. 因此,需对所有对抽取知识有影响的词语进行一一清洗.
- 2.2.2 切分文块
- 经过数据清洗后的病历文本,部位词汇、样本词汇、属性名和属性值的相对位置都已符合统一的格

式. 本文使用汉语句号对病历文本进行粗粒度切割,划分出来的大文本块具有统一的部位信息和样本内容特点,但含有多种症状名与具体的症状值所组建的“键-值”对. 接着,使用汉语逗号进行细粒度切割,各个小块分别叙述病历文本的某部位某症状的具体表现情况.

2.2.3 提取信息

通过数据清洗和切分文本块两大步骤,为本环节中文本分词和关键词提取的效率和准确性提供保障. 借助医学知识添加新词,以避免 NLPPIR 项目提供的 API 自身存在的误判问题,更好地发挥工具带给开发人员的辅助作用.

结合关键信息在病历文本中的具体位置和词性,如:本文所取症状词汇的词性为名词等,症状表现形式词汇词性为形容词或动词短语,位置在名词前后相邻位置,病历文本分析系统可快速锁定病历文本中包含的知识,并经过既定格式进行整理和输出.

2.2.4 提取 TNM 分期标准字

根据 T、N、M 不同衡量标准,特别关注肿块直径、受累淋巴结的位置和数量以及从文中搜寻是否发生远处转移的证据. 结合前半阶段的症状信息提取工作,通过参照影响 3 个指标的各症状信息确定 T、N、M 3 个标准下的衡量结果,进而产生乳腺癌 TNM 分期结果.

2.2.5 基于决策树的 TNM 分期预测

在决策树算法依赖的分类依据中,设置 3 个衡量标准和结论的数据格式,为“T,N,M,Diagnose”,前 3 列的内容判断可在文本处理中进行,即将决策树算法融入到文本提取中.

如:

@ feature

T,N,M,Diagnose

@ data

$d \leq 0.1$, 区域淋巴结无转移,无远处转移,

T1micN0M0

$0.1 < d \leq 0.5$, 区域淋巴结无转移,无远处转

移,T1aN0M0

$0.5 < d \leq 1.0$, 区域淋巴结无转移,无远处转

0,区域淋巴结无转移,无远处转移,T2N0M0

.....

在此基础上,本文使用 ID3 算法的乳腺癌 TNM 分期决策树伪代码如图 4 所示。

2.2.6 基于决策树的癌症分期预测

使用决策树技术,在乳腺癌 TNM 分期预测完毕后,开始癌症分期的决策步骤. 预先建立如下决策树使用的 TNM 分期和病重程度(分期)决策之间的对应关系. 如:

@ feature

TNM,Time

@ data

T1micN0M0, I 期

T2N0M0, II A 期

T3N0M0, II B 期

T4bN0M0, III B 期

.....

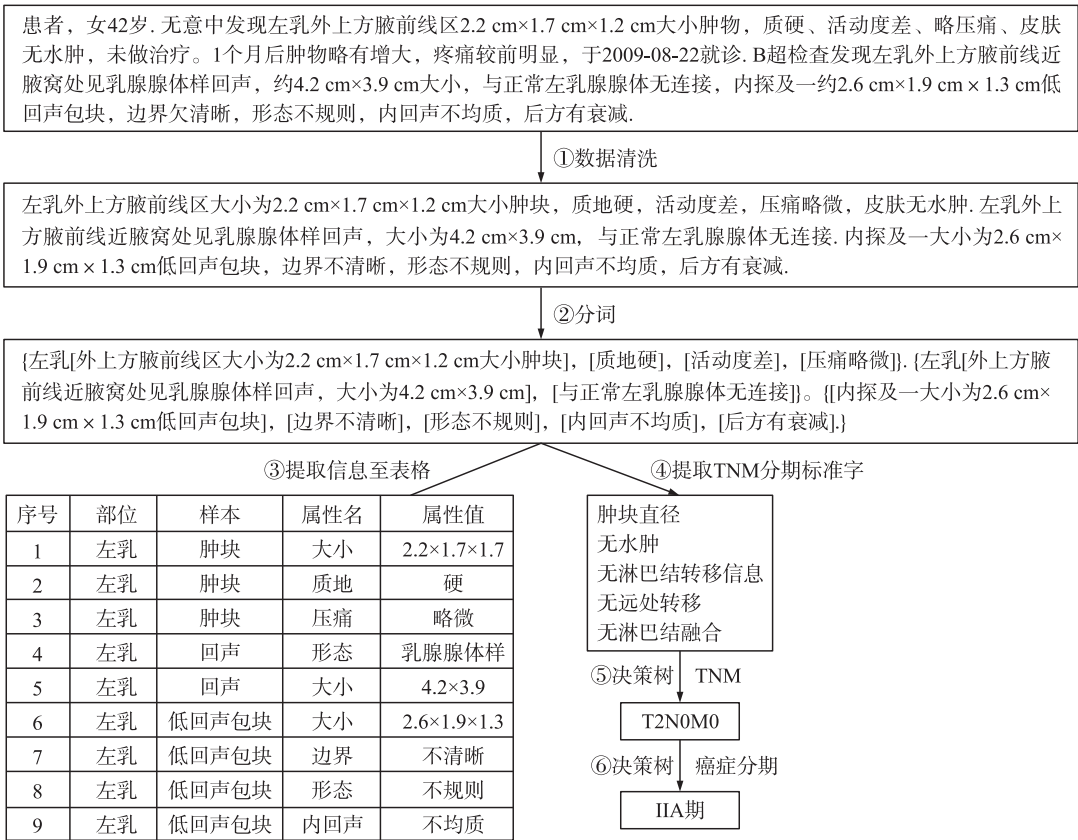
以上介绍的方法其详细实现细节可用一个实例表示,如下所示:

```

TNM分期 决策树ID3算法
@featureList TNM特征列表
@ featureValueTableList TNM特征值列表
Input:
testData= “侵犯胸壁, 区域淋巴结无转移, 无远处转移”
1.for t=1 to featureValueTableList.size()
2. 计算信息熵(使用公式(1))
3. 更新新信息熵和原信息熵
4. 计算信息增益(使用公式(5))
5. 更新最大增益下标
6. 增加乳腺癌TNM分期决策树节点
7.end
8.for cn=1 to featureList.size()
9. 递归寻找含input特征值的节点
10.end
11.返回T4aN0M0
Output: T4aN0M0
    
```

图 4 乳腺癌 TNM 分期决策树算法

Fig. 4 Breast cancer TNM staging decision tree algorithm



2.3 系统结果展示

本文使用 Java 语言开发了病历文本系统,文本处理系统页面展示如图 5 所示. 在系统指定的文本框输入病历文本后,可点击相应按钮进行分词操作、关键词分析操作和新词获取操作,在已有文本处理结果的基础上,可进行信息的组合展示,并根据 TNM 标准的分类条目对病历文本中提取出的信息进行“翻译”工作,对应到 TNM 分期中的某一结论. 如加上 TNM 分期与病重程度的对应关系,可对此病例进行病重程度的分析和预测.



图 5 文本处理系统主界面

Fig 5 Text processing system main interface

3 基于 Neo4j 图数据库的决策验证

本文利用 Neo4j 图形数据库将 TNM 分期标准中的 T 分类、N 分类和 M 分类为 3 个属性. 创建完毕后, 展示如下:

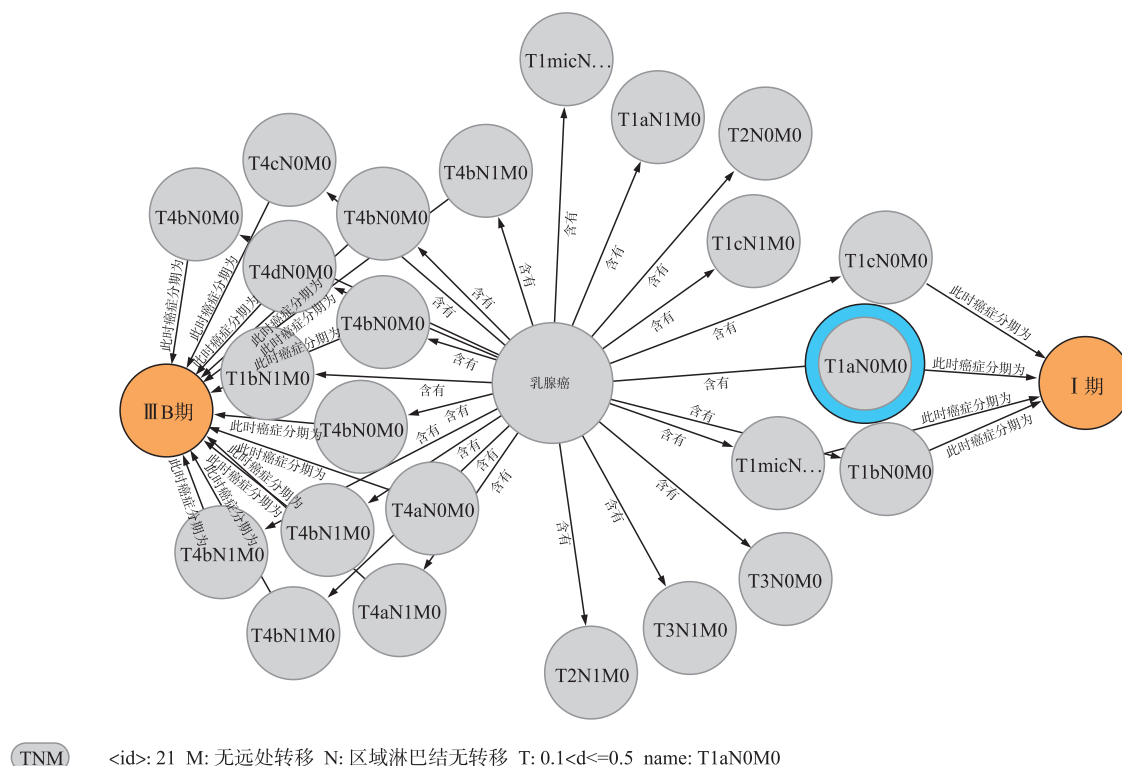


图 6 Neo4j 图数据库创建展示

Fig. 6 Neo4j map database creation show

利用 Java 集成编译器 Eclipse 中建立 Maven 项目, 使用 jdbc 桥接方式, 可与 Neo4j 进行连接. 如只关注 Neo4j 的使用细节, 对应使用的 Cypher 查询语句为:

`MATCH p=(a)-[:含有]->(b)-[:此时癌症分期为]->(c) WHERE b.T="侵犯胸壁" AND b.N="区域淋巴结无转移" AND b.M="无远处转移" RETURN p`

查询结果如图 7 所示.

根据查询结果可知满足“侵犯胸壁”、“区域淋巴结无转移”、“无远处转移”条件时, TNM 分期为 T4aN0M0, 癌症分期为 III B 期.

4 决策树评估

在机器学习和统计中, 常使用 AUC 算法和 ROC 曲线, 基于混淆矩阵, 判断二元分类系统下分类结果的性能. 混淆矩阵分别用“0”和“1”代表负样本和正样本. FP 代表实际类标签为“0”, 但预测类标签为“1”的样本数量. 假正率 (false positive rate, FPR) 是实际标签为“0”的样本中, 被预测错误的比例. 真正率 (true positive rate, TPR) 是实际标签为“1”的样本中, 被预测正确的比例. 精确率 (Accurate Rate) 是正确预测为正类的样本数占实际正类的比重. 召回率 (The Recall Rate) 是正确预基于项目的多元分类^[31] 特点, 在此运用决策树算法的意义在于信息可能不完

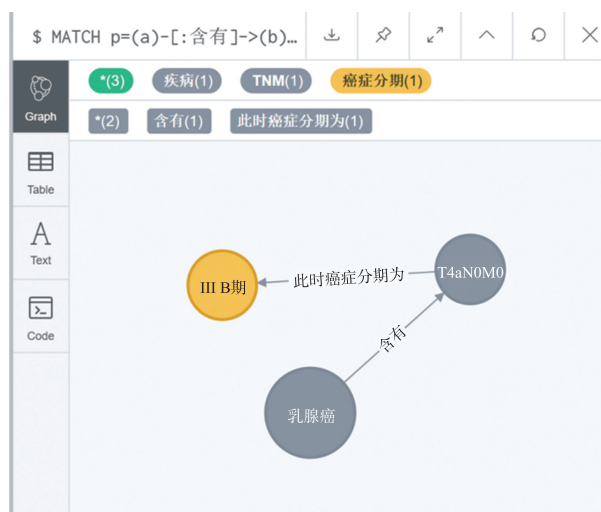


图 7 Neo4j 图数据库创建展示

Fig. 7 Display of Neo4j graph database

整的情况下,加入投票的思想,得到最贴近正确答案的结论. 因此,True Negative 定义为数据获取不饱和但预测正确,False Positive 定义为数据获取饱和但预测不正确,False Negative 定义为数据获取不饱和且预测不正确,True Positive 定义为数据获取饱和且预测正确. 使用交叉验证法^[32]可对分类结果进行合理的评估和验证.

针对以上提交的方法,我们采用了病历夹中经典的 8 例乳腺癌病历文本作为测试数据集,经数据清洗、文本挖掘、信息整合和决策树算法后,结果如表 3 所示.

表 3 系统性能
Table 3 System's performances

序号	系统预测	原文判断	TN	FP	FN	TP	精确率 TP/(TP+FP)	召回率 TP/(TP+FN)	准确率 (TP+TN)/(TP+FP+TN+FN)
1	T2N0M0	T2N0M0	8	0	0	5	1	1	1
2	T2N0M0	T1N0M0	11	1	0	1	0.5	1	0.923
3	T3N3aM0	T3N3M0	0	0	0	13	1	1	1
4	T2N0M0	T2N1M0	4	0	1	8	1	0.889	0.923
5	T2N2aM0	T2N2M0	0	0	0	13	1	1	1
6	T2N0M0	T2N0M0	5	1	1	7	0.875	0.875	0.846
7	T2N0M0	T2N0M0	4	1	0	8	0.889	1	0.923
8	T4cN3aM1	T4N3M1	0	0	0	13	1	1	1
合计			32	3	2	67	0.957	0.971	0.951

针对测试数据集,提交的这一方法获得了 97.1%的召回率及 95.7%的精确率,表明这一方法有望提取乳腺癌分期及临床癌症分期信息.

5 总结

本工作收集关于乳腺癌的病历,随即进行数据清洗工作后,通过使用 NLPiR 汉语分析工具对病例文档进行词性分词、关键词分析、词频分析等工作,结合乳腺癌病历文档的格式,将病例中的肿块信息、癌细胞转移情况等信息展示到窗体中的表格. 同时,根据乳腺癌 TNM 分期中区分不同分期的分类标准,使用决策树技术对病历中的乳腺癌情况进行初步诊断,并结合 TNM 不同分期和癌症病重程度的临床分期对应关系,对病人患病情况做预警分析. 最后,通过图数据库 Neo4j,系统建立了乳腺癌 TNM 分期知识图谱,对决策树下的决策工作进行验证.

本工作旨在帮助医生进行快速的乳腺癌分析、关键信息的抓取. 在乳腺癌病历分析的工作中,使用数据挖掘,结合机器学习中的决策树算法,达到了预期的目的,这项工作是极其有意义的. 该项目还有很大的研究和发展空间,如果在后续工作中,增大乳腺癌病历的具体数量,使用更多机器学习的算法和思想修改文本分析系统的算法和代码,该工作将会展现出更大的学术潜力和使用价值.

[参考文献]

[1] 史双,路潜,杨萍,等. 乳腺癌就诊延误的研究现状[J]. 中华护理杂志,2015(4):88-91.
[2] 陈万青,张思维,郑荣寿,等. 中国 2009 年恶性肿瘤发病和死亡分析[J]. 中国肿瘤,2013(1):5-15.
[3] TSURUOKA Y, MIWA M, HAMAMOTO K, et al. Discovering and visualizing indirect associations between biomedical concepts[J]. Bioinformatics,2011,27(13):111-119.
[4] OKAZAKI N, ANANIADOU S, TSUJII J. Building a high-quality sense inventory for improved abbreviation disambiguation[J]. Bioinformatics,2010,26(9):1246-1253.
[5] WANG X, TSUJII J, ANANIADOU S. Disambiguating the species of biomedical named entities using natural language parsers[J]. Bioinformatics,2010,26(5):661-667.
[6] WANG X, RAK R, RESTIFICAR A, et al. Detecting experimental techniques and selecting relevant documents for protein-protein interactions from biomedical literature[J]. BMC bioinformatics,2011,12(Suppl 8):S11.
[7] HANNA S. Text mining and information analysis of health documents[J]. Artificial intelligence in medicine,2014,61(3):127-130.

- [8] 王浩畅,赵铁军. 生物医学文本挖掘技术的研究与进展[J]. 中文信息学报,2008,22(3):89-98.
- [9] 李慧林. 基于电子病历的疾病预测方法研究及应用[D]. 郑州:郑州大学,2018.
- [10] Hazewinkel Mirjam C, de Winter Remco F P, van Est Roel W, et al. Text analysis of electronic medical records to predict seclusion in psychiatric wards: proof of concept. [J]. Frontiers in psychiatry, 2019, 10: 188-192.
- [11] 李德辉,范焕芳,孙春霞. 乳腺癌中医证型与 TNM 分期相关性的 Meta 分析[J]. 中国老年学杂志, 2017(15): 135-137.
- [12] 傅春燕,陈述政,潘颖. 乳腺癌中医症候分类与 TNM 分期相关性研究[J]. 中国现代医生, 2013(4): 121-123.
- [13] 薛卫成. 介绍乳腺癌 TNM 分期系统(第 7 版)[J]. 诊断病理学杂志, 2010(4): 6-9.
- [14] 刘雨馨,王振光,武凤玉,等. 乳腺癌术前 TNM 分期与术后 ~(18)F-FDGPET/CT 阳性显像相关性分析[J]. 影像研究与医学应用, 2018(11): 86-88.
- [15] 代文杰,张爽. 从 AJCC 第 8 版乳腺癌预后分期解读看外科临床新进展[J]. 临床外科杂志, 2018(1): 21-23.
- [16] 薛卫成,阚秀. 介绍乳腺癌 TNM 分期系统(第 6 版)[J]. 诊断病理学杂志, 2008, 15(3): 161-164.
- [17] 刘艳辉,张芬. 新辅助化疗后的乳腺癌 AJCC TNM 分级与预后关系的评价[J]. 循证医学, 2007(3): 149-151.
- [18] 王若佳,魏思仪,赵怡然,等. 数据挖掘在健康医疗领域中的应用研究综述[J]. 图书情报知识, 2018(11): 116-125.
- [19] 姜欣,徐六通,张雷. C4.5 决策树展示算法的设计[J]. 计算机工程与应用, 2003, 8(4): 93-95.
- [20] 王灿辉,张敏,马少平. 自然语言处理在信息检索中的应用综述[J]. 中文信息学报, 2007(2): 37-47.
- [21] 刘颖. 计算语言学[M]. 北京:清华大学出版社, 2002.
- [22] 宗成庆. 统计自然语言处理[M]. 北京:清华大学出版社, 2008.
- [23] 车万翔,刘挺,李生. 实体关系自动抽取[J]. 中文信息学报, 2005(2): 2-7.
- [24] 李保利,陈玉忠,俞士汶. 信息抽取研究综述[J]. 计算机工程与应用, 2003(10): 4-8.
- [25] 孙琳. 基于 NLP 汉语分词系统和 BFSU PowerConc 1.0 的警务汉语词频与搭配研究——以禁毒案件为例[J]. 现代语文(语言研究版), 2016(12): 142-147.
- [26] CHEN S B, RAO P. Land degradation monitoring using multitemporal Landsat TM/ETM data in a transition zone between grassland and cropland of northeast China[J]. International journal of remote sensing, 2008, 29(7): 2055-2073.
- [27] PAL M, MATHER P M. An assessment of the effectiveness of decision tree methods for land cover classification[J]. Remote sensing of environment, 2003, 86(4): 554-565.
- [28] 苗夺谦,王珏. 基于粗糙集的多变量决策树构造方法[J]. 软件学报, 1997(6): 26-32.
- [29] 马克. 数据清洗在统计调查实践中的应用[J]. 调研世界, 2018(10): 1-2.
- [30] 郝爽,李国良,冯建华,等. 结构化数据清洗技术综述[J]. 清华大学学报(自然科学版), 2018, 58(12): 3-16.
- [31] 饶萍,王建力,王勇. 基于多特征决策树的建设用地信息提取[J]. 农业工程学报, 2014(12): 241-248.
- [32] 刘学艺,李平,郜传厚. 极限学习机的快速留一交叉验证算法[J]. 上海交通大学学报, 2011, 45(8): 49-54.

[责任编辑:顾晓天]