面向乳腺肿瘤的诊前问答系统决策模型构建研究*

王世文[1] 李一凡[1] 郑 群[1] 曹旭晨[2]

([1] 天津师范大学管理学院 天津 300387 [2] 天津医科大学肿瘤医院乳腺一科 天津 300181)

〔摘要〕 **目的/意义** 运用决策树分类模型模拟专家问诊思路，预测潜在或已有乳腺肿瘤患者的疾病风险。**方法/过程** 采用 C4.5 经典分类算法和悲观剪枝法，对调研收集的病例数据进行患者预问诊的结果预测。**结果/结论** 生成一棵以“术后化疗 or 放疗在院是否结束”为根节点、拥有 76 个叶子节点的 C4.5 决策树，预测准确率达 95%，并根据分类标签划分为 3 个风险等级。

〔关键词〕 乳腺肿瘤；C4.5 算法；决策树；模型构建

〔中图分类号〕R-058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673-6036.2023.08.010

Study on the Construction of a Decision-making Model of the Pre-clinical Q&A System for Breast Tumor

WANG Shiwen[1], *LI Yifan*[1], *ZHENG Qun*[1], *CAO Xuchen*[2]

[1] *School of Management, Tianjin Normal University, Tianjin 300387, China*; [2] *Department of Breast I, Tianjin Medical University Cancer Hospital, Tianjin 300181, China*

〔**Abstract**〕 **Purpose/Significance** To predict the disease risk of potential or existing breast tumor patients by using a decision tree classification model to simulate the expert consultation idea. **Method/Process** A C4.5 classical classification algorithm and a pessimistic pruning method are used to predict the outcome of patient pre-consultation for case data collected from the study. **Result/Conclusion** A C4.5 decision tree with 76 leaf nodes and “whether postoperative radiotherapy or chemotherapy ends in the hospital” as the root node is generated with 95% prediction accuracy and classified into 3 risk levels according to the classification labels.

〔**Keywords**〕 breast tumor; C4.5 algorithm; decision tree; model construction

1 引言

2020 年世界卫生组织国际癌症研究机构发布数据显示，乳腺癌已成为全球新诊断人数最多的癌症。作为全球第一大癌，其医师资源紧缺且分布不均衡，优质医生资源多集中于大城市三甲医院。受限于医疗资源以及交通条件，不少乳腺肿瘤患者对个人乳腺疾病发展程度缺乏判断，导致治疗不及时，延误救治时机。在就诊前通过问答系统对乳腺肿瘤患者进行疾病初步风险程度评估是了解个人病情、缓解医疗压力的重要方式。因此，本文拟利用乳腺肿瘤科专家门诊医患对话数据，根据名医面对不同患者时所询问病症因素的逻辑顺序，构建面向乳腺肿瘤的诊前问答系统 C4.5 决策树模型。所谓

〔修回日期〕 2023-02-07

〔作者简介〕 王世文，教授，硕士生导师，发表论文 20 余篇；通信作者：李一凡，硕士研究生。

〔基金项目〕 天津市应用基础计划重点项目（项目编号：S18ZC63056）。

“诊前”即患者此前未到过医院就诊乳腺肿瘤相关疾病。该决策树模型可模拟专家问诊思路进行预问诊，进而根据病情信息评估患者风险程度，提供初步的乳腺肿瘤风险评估，帮助患者了解个人病情，对疾病的及时发现和治疗具有重要意义，对医生在患者就诊前提前收集病情信息具有一定辅助作用。

目前已有不少学者针对乳腺癌领域的决策模型开展研究。段明月[1]选择决策树（decision tree，DT）的回归树算法构建预测模型对女性乳腺癌 5 年内生存状况进行预测，为临床医生预测乳腺癌患者预后和调整个体化随访策略提供参考。刘绿[2]对比 logistic 回归模型、神经网络模型和决策树模型在乳腺癌彩超影像诊断中的灵敏度、特异度及准确度。余秋燕等[3]研究指出决策树在小样本数据上有优势，相比神经网络、支持向量机、贝叶斯、随机森林算法，决策树模型分类效果最优。决策树作为问答系统的一种决策支持模型，具有清晰的树形结构和较好的分类、预测能力[4]。C4.5 算法作为最常用、最经典的分类算法，其稳定性较好、准确率较高，被广泛应用于预测疾病发生风险、危重疾病的生存时间等医疗领域[5-7]。尽管国内外有关乳腺癌的人工智能研究大都具有较好的准确率，然而大多数研究都基于刻板的临床病历资料或者知识库，并不需要接触患者，面向对象仅是疾病，通过庞大病历库理解和推理，系统给出的方案可能是最正确的。而医生在临床实践中面对的是患者，除疾病外还需考虑医保、婚育等生活因素。另外，大多数决策模型的特征名词比较专业，普通人理解和认识存在障碍，应用层面受限。本研究叶子节点语言更贴近生活，因而对诊前决策辅助更有应用价值。

2 数据获取与处理

2.1 数据获取

2021 年 7 月 12 日—8 月 5 日共 6 次前往天津市肿瘤医院乳腺一科，以乳腺门诊患者为研究对象，以录音方式记录医患对话，并对乳腺影像报告数据系统（breast imaging - reporting and data system，BI-RADS）分析等患者信息进行必要的补充记录，获得门诊对话原始音频数据。目前研究中所有数据均来源于同一医院、同一科室、同一医生的出诊、问诊、触诊数据。数据采集方式获得医院、医生许可，所有研究数据不涉及患者唯一可识别的个人具体信息（如姓名、身份证号、病历号等），采集的患者数据包括性别、症状（外在症状、触诊结果）、检查结果等与病情决策有关的属性，不存在伦理及隐私问题。

2.2 数据处理

2.2.1 确定数据处理原则与清洗标准 为了便于利用与分析，需要对原始音频数据进行文本转写。其间试用多种转文本工具，但效果不佳，存在语义不连贯、语义转写错误等问题，最终决定进行人工转写。在采集的数据源中，剔除数据不完整的问诊，并通过实地调研、医生访谈方式进一步使模糊的医学或药物名词精确化；在问诊录音采集过程中，对来院患者所携带体检报告、病历资料、检查报告等与乳腺肿瘤诊断相关的检查结果、等级、指标等进行补充记录。同时通过访谈和实地调研获得问诊、触诊未涉及的属性，以完善数据属性值。由此，补充患者疾病情况，弥补患者病情程度的随机性的不足。经过对门诊录音数据的整理，最终获得原始问诊文本数据。为降低门诊医患对话口语化随意性的影响，对录音转文本数据进行清洗，获得源病例 208 例，其中女性患者占比 98.56%。

2.2.2 确定属性、属性值、类别及其定义 本研究所选取的数据包含的信息量大，且存在大量非相关属性信息，笔者依据相关医学文献、医生访谈以及数据采集过程中医生问诊、视诊、触诊考虑到的属性因素，最终确定 17 个属性、52 个属性值、6 个分类标签，见表 1。为便于后续绘制决策树，将属性名用英文简称进行标识，将属性取值用数字表示；每个分类标签用“数字 + 英文简称”进行标识。其中属性值“未提及”的含义为该属性在医患对话过程中未谈到且前往实地调研的人员未收集到。分类标签即代表医生在该次问诊结束时得出的诊断结果或处理结果，6 个分类标签的确定均是对采集对话文本问诊结果归类分析所得。其中，分类

标签“手术”的含义为在医患对话过程中，医生对患者的诊疗建议为手术；“没事”的含义为乳腺肿块不需要治疗，乳腺较健康；“进一步做检查”建议患者做乳腺相关检查以帮助后续进一步给出问诊结果；“进一步治疗”表示患者正在放疗或者化疗；“其他情况”表示患者所患疾病不属于乳腺肿瘤科室业务范围。

表 1 乳腺科问诊资料属性定义及取值

属性名	属性标识	属性取值
患者性别	P. g	1：女，2：男
乳腺肿瘤手术史	surg	1：是，2：否
两侧乳房对称	sym	1：是，8：否，9：未提及
年龄	P. a	1：不大于 40 岁，2：大于 40 岁
乳头溢液情况	OF	1：血性 and 单孔，2：非血性 or 非单孔，3：无溢液
乳腺超声检查结果	BU	1：未做 B 超，2：3 级及以下，3：4a，4：4b，5：4c，6：5 级及以上
腋下淋巴结坚硬肿大	ALN	1：是，8：否，9：未提及
肿块表面伴有坚硬小结节	HN. SOM	1：是，8：否，9：未提及
肿块是否边界清晰、形状规则	MBS	1：是，8：否，9：未提及
肿块大小	MS	1：大，2：小，9：未提及
肿块活动度	MA	1：活动度好，2：活动度差，9：未提及
肿块质地	MT	1：质地较软，2：质地较硬，9：未提及
怀孕情况	Preg	1：未孕，2：怀孕中，3：已育，9：未提及
乳头凹陷、乳房皮肤橘皮样改变、皮肤溃烂	SN	1：是，8：否，9：未提及
乳房疼痛	BP	1：是，8：否，9：未提及
患者类型	P. t	1：初诊，2：复诊
术后化疗 or 放疗在院是否结束	T	1：是，8：否，5：不涉及
分类标签		1 Oper：手术，2 Rr：定期复查，3 Noth：没事，4 FI：建议进一步检查，5 FT：建议进一步治疗，6 Others：其他情况

进一步按照定义的属性规则对录音文本进行标注，帮助计算机识别语义并训练数据。标注过程中，将过于口语化的表达同义替换为对应属性。文本中下划线斜体词语语义上对应表格中间列的属性值，即对文本进行同义词标注并对文本赋值，见表 2。标注过程为了防止结果存在主观性，在确立属性规则后将文本数据转交第三方人员依照属性规则进行标注，再由研究人员分工进行属性值检查、修改和互检，确保标注规范、所得数据集客观。

表 2 录音对话中的部分同义表达

属性	属性值	自然语言表述
乳腺肿瘤手术史	是（术后）	“刚*烤*完哈，怎么还*烤*糊了呢。”——→术后放疗
		“现在在*吃*什么*药*?”——→术后吃化疗药
		“在我们这*做的手术*?”——→术后
	否（术前）	“结婚了吗?”
		“小孩有吗?”
乳头溢液情况	非血性 or 非单孔	“乳头*流水*吗，我看一下。”
		“乳头老*出血*”
	血性 and 单孔	“乳头溢液要是*单管的、血性的*就得做了”

2.3 决策树构建方法

C4.5 算法作为数据挖掘技术中最常用的分类预测方法，其主要思想是根据信息熵的增益从样本属性中提取最有利于区分实例类别的属性，逐步由根节点向叶子节点构造决策树，可以从生成的决策

树中提取规则[8]。C 4.5 决策树确定每个节点属性的计算方法如下。

设 D 为样本集合，其中第 k 类样本所占的比例为（$k=1, 2, \ldots, y$），y 为样本分类的个数，则 D 的信息熵为：

$$\mathrm{Ent}(D) = -\sum_{k=1}^{y} p_k \log_2 p_k \tag{1}$$

对应数据集 D，选择特征属性 A 作为决策树的判断节点，设属性 A 有 V 个可能的取值 $\{a^1, a^2, a^3, \cdots, a^V\}$，则属性 A 对样本集 D 的条件熵为：

$$\mathrm{Ent}A(D) = \sum_{v=1}^{V} \frac{|D^V|}{D} \mathrm{Ent}(D^V) \tag{2}$$

利用属性 A 划分样本集 D，则信息增益 Gain（A）为：

$$\mathrm{Gain}(A) = \mathrm{Ent}(D) - \mathrm{Ent}A(D) \tag{3}$$

计算属性 A 的“固有值”IV（a），由属性 A 产生的分支节点数目越多，该固定值越大[9]。

$$\mathrm{IV}(a) = -\sum_{v=1}^{V} \frac{|D^V|}{|D|} \log_2 \frac{|D^V|}{|D|} \tag{4}$$

计算属性 A 的信息增益比：

$$\mathrm{Gain\_ratio}(D, a) = \frac{\mathrm{Gain}(D, a)}{\mathrm{IV}(a)} \tag{5}$$

选择具有最高信息增益比的属性作为该集合的测试属性，创建一个节点，并以该属性标记，对属性的每个值创建分枝，并据此划分样本。

为了提高决策树预测的准确率和泛化能力，解决决策树的“过拟合”问题，在训练决策树模型时需要对其进行剪枝处理。C 4.5 决策树算法所采用的剪枝策略是后剪枝法中的悲观剪枝法，即从树的根节点开始搜索，对决策树上的所有内部节点进行计算并分析。重点计算每个内部节点被剪或者被子树代替之后的期望错误率。在剪枝过程中，树中的每棵子树最多需要访问一次，在最坏的情况下，其计算时间复杂度只和非剪枝树的非叶子节点数目成线性关系。因此，相较于其他剪枝方法，悲观剪枝法在实际应用中具有精度高、速度快的优点[10]。

3 决策树模型构建

3.1 决策树模型初建与问题分析

初步研究阶段，针对前述整理的对话文本做进一步病例属性标注，初步标注 15 个属性及 6 个类标签。该决策树构建采用训练集数据 177 条，并从中随机抽取 20 条作为测试集，预问诊抽样的测试集结果准确率为 75%。此研究阶段构建的结果是一棵以“是否有乳腺肿瘤手术史”为根节点且拥有 76 个叶子节点、深度为 12 层的 C4.5 决策树，见图 1。此阶段对医患对话文本仅标注 15 个属性，分别为前文表 1 中前 15 个属性。实地调研发现月经史并不会影响乳腺肿瘤诊断决策，因此在问诊病例文本标注时并未将“月经史”纳入属性。

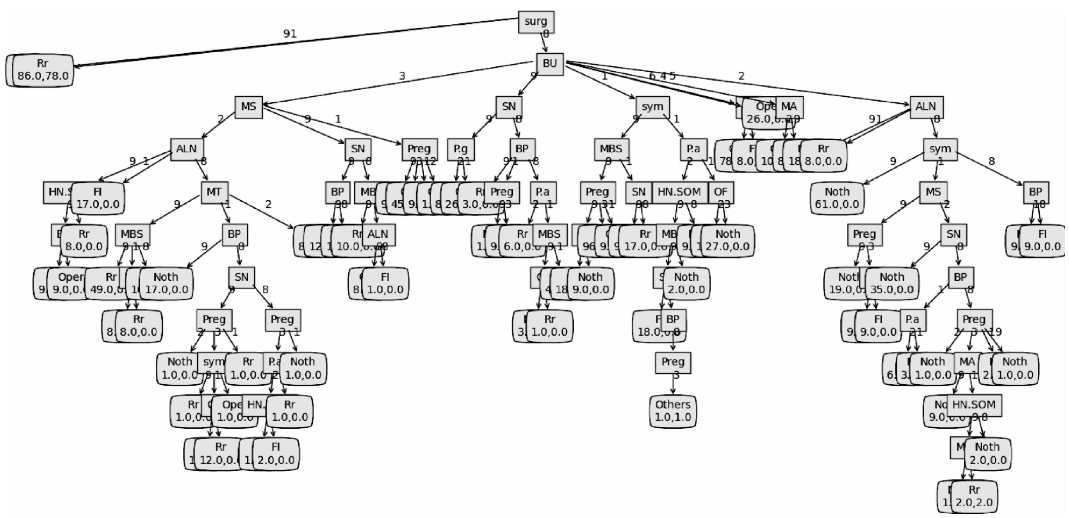

图 1 初步研究阶段决策模型可视化

基于此模型计算的预问诊分类预测值与实际结果不一致的筛选结果（局部），见表 3、表 4，可以看出初步阶段的分类模型预测“手术”“定期复查”“没事”“建议进一步检查”的准确率较高。术后复查患者的叶子节点归属是“定期复查”还是“建议进一步治疗”明显影响决策树准确率。进一步研究讨论得出，所设立的 15 个属性更多适用于术前患者问诊内容，由于术前与术后患者乳腺部位体征差异显著，而判断术后患者的分类结果是“进一步治疗”还是“定期复查”的分岔点属性并未在上述 15 个属性中体现出来。由此可推断，标注属性须调整更新以优化决策树，提高模型分类准确率。

表 3 研究初步阶段决策模型预测差异（局部）

P. g	surg	sym	P. a	OF	BU	ALN	HN. SOM	MBS	MS	MA	MT	Preg	SN	BP	分类标签	预测值
1	1	9	2	3	1	9	9	9	9	9	9	9	9	9	FT	Rr
1	1	9	2	3	9	1	9	9	9	9	9	3	9	9	FT	Rr
1	1	9	2	3	1	8	9	9	9	9	9	3	8	1	FI	Rr

表 4 初步研究阶段 C4.5 决策树分类模型准确率

数据评估 & 分类标签	训练集数据量（条）	测试集准确率（%）
手术	34	100
定期复查	47	100
没事	53	100
建议进一步检查	32	80
建议进一步治疗	10	20
其他情况	1	0
数据总体情况	177	75

3.2 决策树模型优化与验证

在原有属性基础上增加一项“术后化疗 or 放疗在院是否结束”来判断术后患者是“建议进一步治疗”还是“定期复查”，并用字母“T”表示。将“乳腺肿瘤术后出院患者 or（术后出院继续服用周期性其他药物 and 嘱咐在家服用周期药物期间需定期复查）”的患者属性值定义为 1；将“术后未出院 and 需进行后续化疗或放疗其他在院治疗”的患者属性值定义为 8；肿块尚未切除的术前患者不涉及该项属性，用属性值 5 进行标注。再次读取训练集的数据构建 C4.5 决策树，并同样随机抽取其中 20 条作为测试集。准确率提高 10%。调整属性后的分类模型可视化是一棵以“B 超”为根节点、深度 14 层、叶子节点数 86 的C4.5决策树。

对属性调整后的决策树模型，准确率虽有所上升，但考虑到来院患者类型复杂并不能单纯依靠“是否有乳腺肿瘤手术史”进行二分类。参考医院实地调研情况，将来院就诊患者类型细分为初诊和复诊，并将此项可能影响决策模型分类结果的因素补充为新属性，使用“P. t”表示。将“从未去过任何一家医院就诊乳腺肿瘤相关疾病 or 体检发现乳腺异常”的患者类型定义为初诊，属性值为 1；复诊的患者类型则定义为“之前来过本院 or 已在其他医院就诊过 or 术后来医院复查”的患者，属性值为 2。考虑到之前抽取的测试集 6 类标签数据并不均匀，有可能造成准确率的偶然性，因此按照各类标签数量比例再次抽取共 20 条数据作此测试集，计算分类预测准确率。优化后的决策树模型，见图 2。

剪枝后得到的是一棵以“术后化疗 or 放疗在院是否结束”为根节点、深度为 14 的决策树，其叶子节点共 76 个。与乳腺肿瘤相关的类标签，见表 5。可以看出，针对此次抽取的测试集，预测问诊结果为“手术”“定期复查”“没事”“建议进一步检查”“建议进一步治疗“的分类标签更加契合实际结果，较好地模拟了专家问诊思路，总体准确率达到 95%。

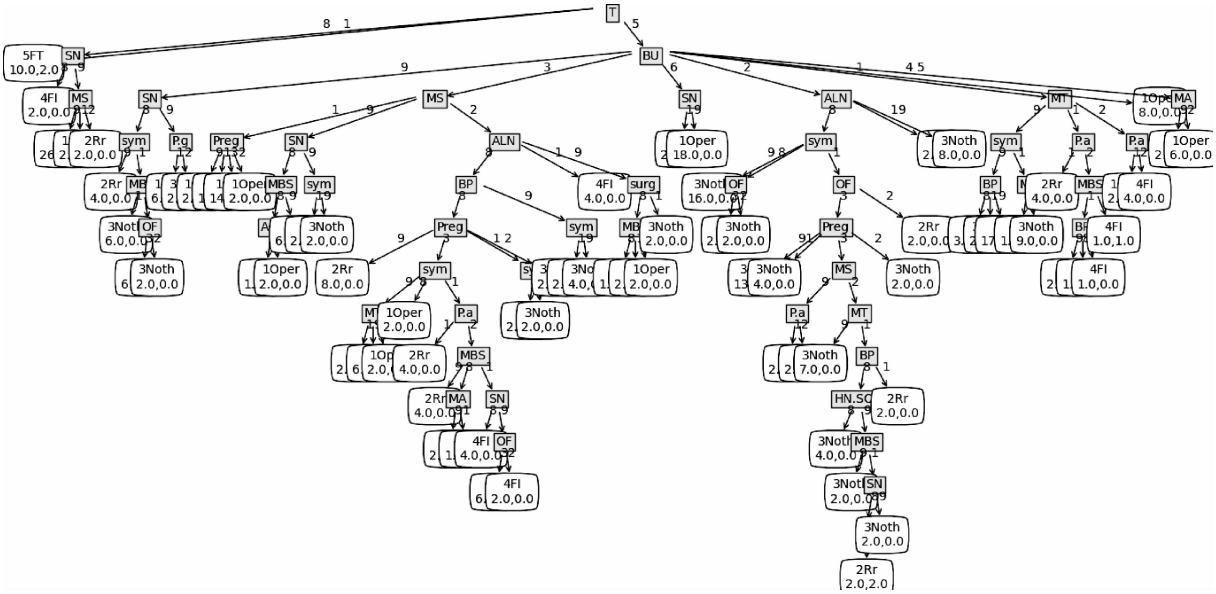

图 2 优化后决策模型可视化

表 5 C4.5 决策树调整后分类模型准确率

数据评估 & 分类标签	训练集数据量（条）	测试集准确率（%）
手术	34	100
定期复查	52	100
没事	51	100
建议进一步检查	33	100
建议进一步治疗	6	100
其他情况	1	0
数据总体情况	177	95

4 结语

本研究所构建的面向乳腺肿瘤的诊前问答系统决策模型分类规则易于理解、分类准确率较高，具有一定疾病风险预测价值。此诊前决策模型有 6 个分类标签：“手术”“定期复查”“没事”“建议进一步检查”“建议进一步治疗”“其他情况”，其中前 3 者在医生当次问诊中不仅给出问诊处理结果还对患者情况了解完备，整个问诊形成闭环，即已作出判断。而问诊结果为“建议进一步检查”的流程则包含于前述闭环中，需要在获得检查结果后至少再次问诊，医生才可对患者的乳腺情况作出最终判断。“建议进一步治疗”的患者，即乳腺肿瘤术后人群，原患病部位发病风险在手术后已大大降低，需继续使用药物或者进行放疗来防止复发。“其他情况”患者（如发生乳腺炎等疾病）不属于乳腺肿瘤风险范围。

由此，笔者将此决策树模型的前 4 个分类标签分别定义为不同的乳腺肿瘤病发风险程度：“建议进一步检查”定义为“风险不明”；“没事”对应 1 级风险；“定期复查”对应 2 级风险；“手术”对应 3 级风险。风险等级越高乳腺肿瘤的严重程度越高。在就诊前通过问答系统对可能患有的乳腺肿瘤疾病预先进行风险评估，使患者及时了解个人病情、抓住治疗时机。在减缓大城市医生问诊压力的同时，也能帮助关心乳腺健康的人群进行预防与筛查，实现早发现早治疗。

在构建决策树的过程中，门诊患者样本过少、录音环境嘈杂等因素导致转写的可用文本较少，故将所有数据用作训练集，并从中抽取 20 条数据用作测试集。测试集数据均出自训练集，分类预测结果难免会受部分影响；数据来自“同一个医院、同一个科室的同一名医生”既是目前研究的特点，也是后续扩大研究范围进一步提升预测模型的方向。有利于对决策过程中的变量进行有效控制，使研究本身能够有一定深度。因此，本研究成果提供了一

（下转第 65 页）

教育和健康意识培育，学生个人要提高健康信息获取主动性，各类网络媒介要根据自身媒介特性，扬长避短，深耕优势内容模块，为大学生提供契合需求的健康信息。

参考文献

1 朱庆华，杨梦晴，赵宇翔，等. 健康信息行为研究：溯源、范畴与展望［J］. 中国图书馆学报，2022，48（2）：94－107.

2 贾明霞，徐跃权，石尧. 国内外用户网络健康信息行为研究动态［J］. 医学信息学杂志，2022，43（9）：42－46.

3 AGYEMANG－DUAH W，ARTHUR－HOLMES F，PEPRAH C，et al. Dynamics of health information－seeking behaviour among older adults with very low incomes in Ghana：a qualitative study［J］. BMC public health，2020，20（1）：928.

4 HASSAN S，MASOUD O. Online health information seeking and health literacy among non－medical college students：gender differences［J］. Journal of public health，2021，29（6），1267－1273.

5 罗晓兰. 患者网络健康信息沟通意愿及行为调查［J］. 医学与哲学，2022，43（16）：34－37.

6 周培宇，梁昌勇，马一鸣. COVID－19 背景下基于 IMB 模型的中老年人在线健康信息搜寻行为影响机制研究［J］. 中国管理科学，2022，30（3）：76－84.

7 杨霞，王晓梅. 河南省大学生健康信息行为分析［J］. 医学与社会，2020，33（12）：48－51，89.

8 沈默. 大学生网络健康信息搜寻行为及其影响因素研究［D］. 杭州：浙江大学，2018.

9 PIAN W，SONG S，ZHANG Y. Consumer health information needs：a systematic review of measures［J］. Information processing and management，2020，57（2）：102077.

10 RUESCH J，BATESON G，PINSKER E C，et al. Communication：the social matrix of psychiatry［M］. London：Routledge，2017.

11 PETTY R E，CACIOPPO J T. The elaboration likelihood model of persuasion［J］. Advances in experimental social psychology，1986，19（1）：123－205.

（上接第 59 页）

个基础模型（原型）。在后续研究中，将在此基础上引入强化学习，扩大采集数据的医生数量，同时引入必要的检查结果数据，从而丰富模型，拓展模型应用的普遍性；使模型在就诊前为患者和医生提供辅助性建议，从而应用到相应智能问答系统中，为诊断结果提出初步建议，进而实现辅助决策。

参考文献

1 段明月. 决策树模型在预测乳腺癌 5 年生存状况研究中的应用［D］. 长春：吉林大学，2020.

2 刘绿. Logistic 回归模型、神经网络模型和决策树模型在乳腺癌的彩超影像诊断中的比较研究［D］. 衡阳：南华大学，2013.

3 余秋燕，赵莹，孙继佳，等. 典型机器学习算法在脂肪肝分类预测研究中的实现与比较［J］. 数理医药学杂志，2019，32（1）：1－3.

4 迟辉，高颖. 基于决策树法探析高颖教授辨治失眠主方主症规律［J］. 世界中医药，2021，16（16）：2484－2492.

5 林文怡，宛小燕，刘元元. 常见新近决策树算法及其在卫生领域中的应用［J］. 现代预防医学，2019，46（23）：4233－4237，4242.

6 江泽飞，许凤锐. 肿瘤医生眼中的人工智能［J］. 精准医学杂志，2018，33（1）：9－11，14.

7 TUPASELA A，DI NUCCI E. Concordance as evidence in the Watson for oncology decision－support system［J］. AI & society，2020，35（4）：811－818.

8 王守选，叶柏龙，李伟健，等. 决策树、朴素贝叶斯和朴素贝叶斯树的比较［J］. 计算机系统应用，2012，21（12）：221－224.

9 郭星晨，王青青，王亚. C4.5 决策树算法在医疗数据分类中的应用研究［J］. 安庆师范大学学报（自然科学版），2021，27（2）：49－53.

10 李萍. 悲观剪枝算法在学生成绩决策树中的应用［J］. 电脑开发与应用，2014，27（5）：4－6.