数据挖掘在健康医疗领域中的应用研究综述

Review of Data Mining Techniques' Application in Medical and Healthcare Field

王若佳[1,2]　魏思仪[1]　赵怡然[1]　王继民[1]
(1.北京大学信息管理系,北京,100871;　2.北京大学海洋研究院,北京,100871)

[摘要]　健康医疗领域的数据挖掘与知识服务已成为健康医疗大数据产业发展的核心需求之一,数据挖掘作为知识提取的关键技术近年来受到较多关注。文章首先对数据挖掘常用于健康医疗领域的模型与算法进行了梳理与说明;然后分别综述了该技术在辅助完成医疗任务、合理管理医疗资源、改进健康信息服务三大方面的应用现状,并归纳了每方面涉及到的细分应用领域、算法及代表性论文;此外,数据挖掘技术在健康医疗领域中的应用局限和问题也不容忽视,文章按照数据采集、数据预处理、算法选择和结果评估的顺序对现有研究中提到的不足进行总结;最后,提出了数据来源多样化,电子病历挖掘语义化,与云计算、人工智能等领域共同发展的三个未来研究方向。

[关键词]　数据挖掘　健康医疗　医疗任务　医疗资源　健康信息服务

[中图分类号]G202;R-05　[文献标识码]A　[文章编号]1003-2797(2018)05-0114-10　DOI:10.13366/j.dik.2018.05.114

[Abstract]　Data mining and knowledge service in medical and healthcare field has become one of the core needs of the development of related industry. Data mining as the key technology of knowledge extraction has gained much attention in recent years. This paper first summarized common models and algorithms of data mining in the field of medicine and healthcare, and then reviewed the application status of this technology in three aspects, including assistance of completing medical tasks, scientific management of medical resources, and improvement of health information service. Meanwhile, this paper concluded the subdivided application areas, algorithms and representative publications. In addition, it also summarized the limitations and problems according to the data mining process. Finally, three future research directions have been put forward.

[Keywords]　Data mining; Medical and healthcare; Medical task; Medical resources; Health information service

随着医院信息系统和健康网站的发展,医疗活动、医学研究和健康信息行为中的数据被存储下来,形成了海量的健康医疗大数据。这类数据的数据量大,存储形式多样,难以用传统数据处理方法进行处理。数据挖掘由于能够分析海量异构数据,越来越多地被应用于健康医疗领域。本文对该领域已发表的国内外相关研究成果进行梳理,总结现阶段数据挖掘在健康医疗领域的主要应用,同时关注其相关应用局限与问题,以期为健康医疗大数据挖掘的进一步研究提供参考。

综述论文的来源数据库及检索方式如表1所示,得到68篇中文文献及3597篇英文文献,由于英文文

[作者简介]　王若佳,博士生,研究方向:Web数据挖掘,健康信息,海洋大数据;魏思仪,本科生,研究方向:健康数据挖掘;赵怡然,硕士生,研究方向:Web数据挖掘,科学评价等;王继民(通讯作者),教授,博士生导师,研究方向:搜索引擎与Web数据挖掘,科学评价等,Email:wjm@pku.edu.cn。

献过多，因此首先按相关性排序，选取最相关的前 500 条；然后依据论文题目和摘要对 568 篇文献进行初步筛选，剔除不是医疗健康领域的文献、未使用数据进行实证研究的文献以及未使用数据挖掘算法的文献，最后得到 216 篇。通过全文阅读，按具体应用领域及目的进行分类，相似论文中选取发表时间较新且代表性较高的文献进行综述。

表 1　综述论文的来源数据库及检索方式

数据库	Web of Science Core Collection	CNKI
检索方式	TOPICS＝ data mining AND（medicine OR medical OR healthcare）	主题＝数据挖掘 AND（医疗 OR 健康）
发文年代	2010—2018	2010—2018
来源类别	SCI-EXPANDED，SSCI，CPCI-S，CPCI-SSH	SCI 来源期刊、EI 来源期刊、核心期刊、CSSCI
检索结果	3597 篇	68 篇
检索时间	2017-10-28	

1　模型与主要算法

数据挖掘是从大量数据中挖掘有趣模式和知识的过程[1]，数据挖掘模型则是对这些模式的精炼与总结，其具体实现需要各种算法的支持。主要的数据挖掘模型与算法如图 1 所示。

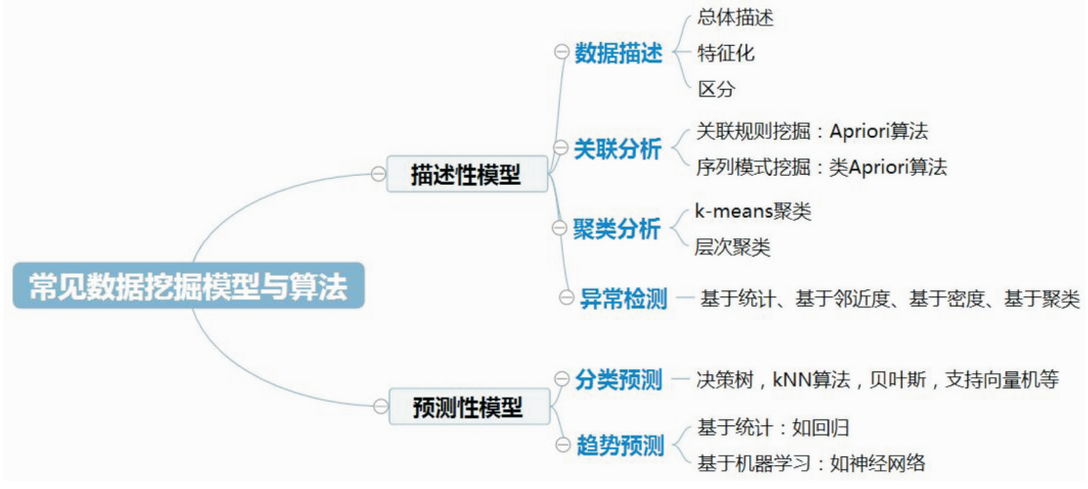

图 1　常见数据挖掘模型与算法汇总

描述性模型用于回答数据集是什么、有什么性质，预测性模型则是对数据现有性质进行归纳，从而预测未来趋势。具体来看，常用于健康医疗领域的模型与算法包括：

（1）数据描述，是对所有数据集的基本描述、特征汇总与对比。例如 De-Arteaga 等人[2]在分析医疗图像检索行为时，统计了日志数据中的检索式数量、仅出现一次的检索式所占比例、最常出现的 10 个词语等，这些特征在一定程度上可反映用户需求。

（2）关联分析，用于探讨研究对象之间的相关性，一般通过挖掘数据中频繁出现的项集实现，如 Ilayaraja 等人[3]用关联分析的方法探讨了患者症状与心脏病危险等级之间的关系。序列模式挖掘是关联分析的高级拓展，该模型在考虑项集出现频次的同时，还需保证项与项之间的顺序不变。陶惠和蒋凡[4]使用该模型研究了患者在不同医院间的转诊模式，由于转诊过程隐含着一定的医院顺序，因此简单的关联分析方法并不适合。

(3)聚类分析,按照某种相似性原则对整体进行分组,相同组中的对象具有较大相似性,而不同组别之间又有一定区分度。如唐晓琳等人[5]对常用健康医疗网站进行了系统聚类,发现参与医疗市场的网站主要包括健康资讯网、医学工具类网站、以及在线药店三种。

(4)异常检测,用于发现与数据中大部分对象表现行为不一致的异常点。例如,Bouarfa 和 Dankelman[6]对 26 个腹腔镜胆囊切除术的工作流程进行挖掘,提出一致的工作流,从而对手术实践方法做出改进;邵笑笑[7]对医保费用数据中的异常和违规行为进行甄别,以完善医疗保险公司的反欺诈机制。

(5)分类预测,是一个两阶段过程。首先建立目标属性与其它属性之间的关系,然后根据这种映射关系判断目标属性未知的对象的类别。王宇燕等人[8]评价了集成分类、决策树和随机森林对结直肠癌患者存活性的预测性能与精度,Sharma 和 Om[9]对比了几种决策树模型对口腔癌患者存活率的预测效果,这对节约医疗资源、降低医疗成本、提高患者满意度等方面都具有实际意义。

(6)趋势预测,与分类预测的区别在于目标属性是数值型数据还是分类型数据。例如,Xu 等人[10]提出的一种基于 Web 数据挖掘的流感检测框架,他们采用了不同的神经网络模型模拟流感样疾病数据和查询数据之间的关系,从而通过搜索引擎预测流感疫情。

2　领域与应用方向

2.1　辅助完成医疗任务

医疗任务的主要内容是对个体的患病状况进行诊治,按照疾病的诊治顺序可分为预防、诊断、治疗和预后四个阶段。数据挖掘技术可有效发掘医疗数据的潜在价值,从而推动医疗任务的完成。

(1)预防

有效控制影响疾病发生的危险因素以及疾病的早期筛查均可以有效预防疾病的发生。常见疾病诱因包括人口统计学特征、家族史、患者体征和生活方式等。Meng[11]等人通过问卷调查收集了糖尿病患者的数据,结合卡方分析识别了影响糖尿病发病率的关键因素,并对比了 logistic 回归、BP 神经网络和 C5.0 决策树三种模型的预测效果。吴生根等人[12]基于福建省手足口病数据,运用卡方自动交互检测(CHAID)作为生长法,分析影响手足口病重症病例发生的危险因素。疾病关联研究中,任仙龙等人[13]使用 apriori 算法对 4585 名社区居民的慢性病调查数据进行分析,发现高血压、糖尿病和高血脂三者之间存在强关联。

(2)诊断

利用病历文本信息和图像数据可辅助疾病诊断。张晔等人[14]采用 logistic 回归对 323 例急性胰腺炎病例进行特征向量选取,并对比选取前后预测模型的准确性,发现特征选取能有效提高模型的性能;此外,文章将支持向量机与 logistic 回归、决策树和人工神经网络进行对比,发现支持向量机在测试集上的准确率更高。Yang 等人[15]使用决策树方法对临床数据进行特征选择,并基于关联规则挖掘方法对临床数据与病理报告进行了分析,从而支持肺癌病理分期诊断。Zubi 等人[16]则采用神经网络、关联规则挖掘的方法对 X 光胸片中的肺癌进行检测和分类。

(3)治疗

疾病治疗方面,国外注重对治疗程序的研究。Villamil 等人[17]采用数据包络分析和过程挖掘评估胃癌治疗的质量,对不同医疗机构按其治疗过程进行有序聚类,以确定不同医疗机构患者的一般治疗模式。Auconi 等人[18]则对 X 光投影测量得到的 22 个变量进行模糊聚类得到牙齿生长特征、面部形态、矫正力方向等因素对正畸效果的影响。

国内则更多集中在中医药物治疗方面。张奇等人[19]应用关联规则分析了李涛教授治疗多发性硬化(MS)所用中药对 MS 患者外周血 T 细胞亚群的影响。丁心香等人[20]对治疗颈性眩晕方剂中的 154 味中药进行了频次统计,通过关联规则挖掘方剂中药物组合规则,采用无监督的熵层次聚类,得到 7 个新方组合。曹锦梅等人[21]使用关联规则分析的方法挖掘年龄、性别、症状之间的关系,并探讨了糖尿病症状与药物使用之间的规则。

(4)预后

预后是指预测疾病的可能病程与结局,既包括判

断疾病的特定后果,也包括预测未来发生某种结局的可能性。

病人自身的身体素质及患病情况是影响预后的常见因素,侯婷[22]基于Ⅰ期子宫内膜样腺癌患者的病理学诊断和治疗记录对年轻患者保留或切除卵巢对其生存预后的影响进行分析,发现保留卵巢的患者接受放疗和淋巴结切除手术的可能性更低。药物不良反应可辅助判断患者预后情况,Kim 等人[23]基于韩国药物不良事件报告挖掘了氟西汀药物的不良反应信号;由于药物监测报告的滞后性,Yang 等人[24]评估了基于社交网络和在线健康社区的药物不良反应识别效果,通过关联规则挖掘方法可以较好地识别某药物及其常见不良反应的规则,从而达到预警作用。患者的预后情况还可通过 DNA 或 RNA 等生物信息进行预测,Xu 等人[25]以 407 例卵巢癌患者的基因资料为原始数据集,综合采用随机森林算法、Cox 比例风险回归等模型分析了与卵巢浆液性癌患者生存期相关的基因;Bai 等人[26]基于同样方法识别了喉癌的潜在生物标志物。

综上,在辅助完成医疗任务方面,数据挖掘技术的应用贯穿于疾病预防、诊断、治疗、预后的全过程。从数据特征来看,大部分相关研究使用的是数值数据或结构化的文本,这类数据结构化程度高,数据量通常较大。从数据来源看,由于国内外健康医疗领域数据开放程度的不同,国内研究一般基于医院的电子病历,其中涉及到中医的研究往往基于著名中医的药方;国外的研究除了使用医院信息系统数据,更多的是基于地方政府公开的数据集、全国某类疾病数据库、药学数据库、在线健康社区等。算法应用方面,logistic 回归常被用于数据预处理阶段,选取特征变量,从而避免选择偏差; apriori 算法被广泛应用于中医的药方规则提取;决策树、神经网络、随机森林算法则常用于病情的预测。算法评估方面,常用的指标包括精确度、灵敏度、特异度等,常采用十折交叉验证的方法进行测试。

表 2　数据挖掘算法用于辅助完成医疗任务的总结

应用领域	具体应用领域	算法	文章作者
预防	疾病诱因识别	logistic 回归、BP 神经网络、决策树 C5.0	Men 等
		决策树 CHAID	吴生根等
	疾病关联分析	apriori 算法	任仙龙等
诊断	辅助临床诊断	支持向量机、logistic 回归	张晔等
		决策树、关联规则挖掘	Yan 等
		神经网络、apriori 算法	Zubi 等
治疗	辅助药物治疗	数据描述、apriori 算法	张奇等
		关联规则挖掘、复杂系统熵聚类、无监督的熵层次聚类	丁心香等
		改进 apriori 算法	曹锦梅等
	临床路径挖掘	数据包络分析	Villamil 等
		数据描述、模糊聚类、网络分析	Auconi 等
预后	药物不良反应	关联规则挖掘	Yang 等
	预后情况预测	数据描述,logistic 回归	侯婷
		随机森林算法、Cox 比例风险回归	Xu,Bai 等

2.2　合理管理医疗资源

医疗资源是指提供医疗服务的生产要素的总称,包括人员、医疗费用、医疗机构、医疗床位和设备等。由于医疗需求增加但医疗资源有限,资源的合理配置成为卫生管理的重点,数据挖掘技术可帮助卫生管理人员深入洞察数据,从而支持决策。

(1)门急诊管理

如何合理安排患者候诊并提高患者满意度是门诊管理的重点。胡敏等人[27]通过分析患者的平均等候时间、问诊时间、患者滞留数量分布,直观地反映了

门诊的患者候诊状况及医生服务情况。陈勇[28]利用神经网络算法探讨了影响门诊患者满意度的相关因素,结果表明只有1名医生出诊会降低门诊患者满意度。彭金燕等人[29]基于患者基本特征、所患疾病、看病次数、付费方式等变量,使用k-means聚类方法将门诊患者聚为四类,并针对每一类提出不同管理策略。

急诊主要针对病情严重或情况紧急的患者,确定病人就诊及处置的优先次序是急诊管理的首要问题。Lin等人[30]首先根据患者基本信息、挂号时间、就诊途径、疾病类型将22990名患者分为6类,然后基于粗糙集理论提取不同类型患者与急救等级之间的规则,从而实现病人分流。Talbert等人[31]对比了决策树、人工神经网络和支持向量机在创伤病人分流中的效果,结果显示支持向量机算法在准确度和特异性方面的表现更好。

(2)住院管理

住院是医院业务中最为繁杂的部分,数据挖掘可用于分析影响住院人次或住院时间的主要因素。Khajehali N和Alizadeh S[32]使用贝叶斯提升集成法探讨了不同抗生素药物分别对不同年龄段肺炎患者住院时间的影响。黄东瑾等人[33]使用logistic回归,探讨了社会学因素、疾病因素和临床因素对老年糖尿病患者住院日分布的影响。

对住院天数的预测还可为医院合理调配人力物力提供科学依据。Xie等人[34]基于保险理赔数据预测了参保人群在未来一年中的住院天数。张晔等人[35]抽取辽宁省某医院去识别化的急性胰腺炎电子病历,建立了基于支持向量回归的胰腺炎患者住院天数预测模型。Rezaei等人[36]发现冠心病患者在合并肺、呼吸障碍及高血压时,住院时间明显变长,支持向量机算法相较于决策树和人工神经网络来说预测准确率最高。

(3)医疗费用管理

医疗费用反映了医疗服务资源的消耗,同时也是广大患者最为关心的问题。郭慧敏等人[37]利用R软件的arules包探讨了科室、住院天数、性别、有无手术与治疗费用之间的联系。张凯[38]对比了四分位数处理法和k-means聚类在处理血液病医疗费用数据时的区别、优势与不足。除了聚类与关联分析外,决策树在医疗费用管理中应用较广。韩晓梅等人[39]利用决策树对卵巢癌患者进行病例组合分析,给出了相对应的住院费用标准;薛允莲[40]将logistic回归和决策树相结合,探讨了每种病例组合的住院费用;Wang Jing等人[41]则对比了神经网络和决策树模型在胃癌患者住院费用方面的预测效果,结果显示人工神经网络的预测能力和自适应能力都优于决策树。

(4)医疗保险管理

医保欺诈方面,Kirlidog和Asuk[42]总结了土耳其常见的骗保形式,并初步探索了异常检测、聚类和分类模型在保险诈骗中的应用可行性;Kose I等人[43]建立了一个完整的医保欺诈识别框架,从数据处理、模型选择、结果评估和可视化多个方面检测医保数据中的异常。另外,陶惠[44]从大病患者的特征出发,根据实验得到的分类模型判断大病保险的多种影响因素的重要性,对大病保险政策的实施和工作开展有一定指导作用。

(5)其它

数据挖掘在医疗资源管理中还有很多其他应用。药品管理方面,Ramos等人[45]针对沙丁胺醇药品库存不足的问题,通过收集居民年龄构成、温度、空气质量、湿度等数据,预测了该药品在未来一年中的需求量;器官移植方面,Koyuncugil和Ozgulbas[46]开发了一个基于数据挖掘算法的供体选择系统。数据挖掘还可发现患者的转诊模式,如郭浩[47]对转诊社交网络与住院时间和费用的关系进行了研究;陶惠和蒋凡[48]使用改进的Apriori算法挖掘某地常见的转诊序列。

综上,目前将数据挖掘技术用于医疗管理的研究较多,且具体应用领域广泛,包括门诊管理、急诊分流、住院管理、医疗费用管理等。从数据来源看,国内研究一般基于未开放数据,大部分来源于研究人员所在医院的HIS系统;国外研究的数据来源多样,除了医院诊疗数据,还包括当地政府的半开放数据集(需要研究人员申请),以及当地保险公司数据。从挖掘算法看,关联规则挖掘任务的常用算法为apriori;聚类算法多为k-means,但也存在使用层次聚类、SOM自组织映射聚类的研究;分类算法多种多样,常见的如决

策树、logistics 回归、人工神经网络、支持向量机，其它如贝叶斯、粗糙集以及各种集成算法（bagging、boosting、stacking）也得到了应用，具体应用情况如表 3 所示。在应用的过程中，聚类算法往往用于数据预处理阶段，起到将连续型数据处理为分类型数据的作用，从而作为分类模型的输入或输出变量。另外，由于常用分类算法较多，对算法的选择不仅应从算法本身考虑，还应同时兼顾数据特征。部分学者在选择算法时说明了理由，如 Lin[49] 在预测患者创伤等级时选择了粗糙集算法，因为该算法能生成具体规则，可读性强；部分学者在研究中对多个分类算法进行比较，选择了效果最优的模型，总体来看，支持向量机的预测效果较好一些，可能与其泛化能力较强有关。

表 3　数据挖掘算法用于合理管理医疗资源的总结

应用领域	具体应用领域	算法	文章作者
门急诊管理	门诊管理	数据描述	胡敏等
		apriori 算法、人工神经网络	陈勇
		层次聚类、k-means 聚类	彭金燕等
	急诊分流	SOM 自组织映射聚类、k-means 聚类、粗糙集	Lin W T 等
		决策树、人工神经网络、支持向量机	Talbert D A 等
住院管理	住院时间的影响因素	多数投票算法、贝叶斯提升集成法、支持向量机、stacking 算法	Khajehali N 和 Alizadeh S
		logistic 回归	黄东瑾等
	住院天数预测	时间序列，bagged 决策树	Xie Yang 等
		支持向量回归	张晔等
		决策树 C5.0、人工神经网络、支持向量机	Hachesu P R 等
医疗费用管理	医疗费用管理	apriori 算法	郭慧敏等
		四分位数处理法，k-means 聚类、决策树 C4.5	张凯
		单因素方差分析、多元线性逐步回归分析、决策树 CHAID	韩晓梅等
		logistic 回归、决策树 CHAID	薛允莲
		BP 神经网络、决策树	Wang Jing 等
医疗保险管理	医保欺诈	异常检测、聚类和分类模型	Kirlidog M 和 Asuk C
	医保政策支持	k-means 聚类、决策树 C4.5	陶惠
其他	药品管理	简单线性回归、支持向量回归	Ramos M I 等
	患者转诊	线性回归、决策树、apriori 算法	郭浩
		apriori 算法	陶惠和蒋凡

2.3　改进健康信息服务

健康信息服务是指利用现代信息技术，通过对健康信息资源的获取来帮助人们更好地调节控制自身健康问题[50]。数据挖掘方法可用于分析以在线资源为主的健康信息，从而帮助改进健康信息服务。

（1）理解健康信息需求

理解健康信息需求有助于提供更具针对性的健康信息服务。金碧漪等人[51]以雅虎问答中糖尿病相关的 8762 条提问记录为分析对象，通过内容分析和多维尺度分析进行聚类，发现消费者对糖尿病的日常疾病管理、疾病确诊和治疗关注度较高，而对疾病预防关注度较低。Falotico 等人[52]用文本分析结合对应分析的方法分析了肉瘤患者的半结构化访谈样本，发现患病时间长的患者更依赖网络搜索获取健康信息。Ku 等人[53]以艾滋病为例，利用信息增益的特征选择方法和支持向量机、朴素贝叶斯两种分类算法，将论坛用户对艾滋病的关注点归为 4 类，并对比了知识分享型论坛和社会支持型论坛用户的关注差异。

（2）理解健康决策行为

健康决策行为与患者实际接受的治疗息息相关。以就诊行为为例，发掘其影响因素有助于调整健康信息教育内容，制定适宜的护理干预策略。Oh 等

人[54,55]基于癌症患者和心脏病患者的访谈和问卷数据,采用决策树算法预测患者的就医时间和治疗策略选择,并用 logistic 回归分析患者选择的影响因素。研究发现心脏病患者的就医时间受症状严重程度、相似疾病的过往就医经验以及就医障碍等因素影响;治疗策略选择受医疗治疗方案有效性、患者对其它理疗方案的信心等因素影响。

(3)改进健康信息检索

健康信息检索是健康信息的重要获取途径之一。检索任务识别方面,孙丽[56]基于问卷调查和检索过程录屏数据,分别用人工神经网络、决策树和支持向量机的方法构建了分类器,依据检索过程的特征指标预测了检索任务类型,有助于完善关键词推荐和优化结果列表排序。检索结果数量方面,De-Arteaga 等人[57]基于专业医学图像搜索引擎 ARRS GoldMiner 的检索日志,利用支持向量机、logistic 回归、随机森林等算法根据检索式特征预测医学图像检索结果的数量,有助于改进推荐检索式,返回适量的检索结果。

(4)优化健康信息组织

健康类网站与论坛是健康信息的重要来源,合理的内容组织有助于用户快速定位所需信息。Chen 等人[58]研究了糖尿病相关文章的自动分类方法,对 11216 篇文章进行了特征提取和人工标注,基于深度信念网络进行分类,实验证明其准确率高于支持向量机算法。罗文馨等人[59]基于 30 个常见疾病主题,从医学新闻网站上采集对应文档,运用 Word2Vec 技术对各疾病的相关文档构造词向量,计算向量距离,从而判断疾病关联,有助于提高信息服务平台的内容组织和导航质量。

综上,与前两大应用领域相比,数据挖掘在改进健康信息服务领域的应用相对较少。在改进健康信息服务领域,研究数据主要来自健康网站、健康论坛、问卷与访谈结果、检索日志或检索过程的录屏。国内外研究在数据源上不存在显著差异。数据类型以文本型为主,数据结构化程度较低。在数据预处理阶段往往需要进行词或文档的向量化表示,常用方法有人工定义、词袋模型等。词袋模型得到的特征词往往较多,结合信息增益进行特征选择能提高模型整体效果[60]。预测型任务中,常用的算法如支持向量机、logistic 回归、决策树、随机森林、深度信念网络等。其中,支持向量机适用于高维数据,当输入数据为高维文档向量或词向量时往往有较好的表现。

表 4　数据挖掘算法用于改进健康信息服务的总结

应用领域	具体应用领域	算法	文章作者
理解健康信息需求	健康信息需求聚类	多维尺度分析	金碧漪等
		对应分析	Falotico R 等
	异常信息需求识别	信息增益、支持向量机、朴素贝叶斯	Ku Yungchang 等
理解健康决策行为	治疗决策影响因素识别	决策树、logistic 回归	OH H S 等
改进健康信息检索	识别检索任务	人工神经网络、决策树、支持向量机	孙丽
	预测检索结果数量	支持向量机、logistic 回归、决策树	De-Arteaga M 等
优化健康信息组织	文章主题分类	深度信念网络、支持向量机	Chen Xinhuan 等
	疾病关联探测	Word2Vec	罗文馨等

3　问题与应用局限

数据挖掘技术在健康医疗领域的应用仍存在较多问题与局限,具体如下:

(1)数据孤岛与高质量数据缺乏

数据采集是数据挖掘的第一步。目前国内医院信息系统数据往往存在不完整、不规范等问题,因此在分析之前需要初步评估数据的准确性。数据孤岛是制约健康医疗数据采集的另一大问题。一方面,医疗机构之间无法实现数据互通;另一方面,多数人不愿意公开自己在移动医疗或健康监测平台上的数据,因为当他们认为自己从移动医疗中的获益程度比不上隐私泄露的风险时,其信息共享的意愿会急剧下降[61,62]。

(2)维度灾难与样本分布不平衡

医疗数据往往维度较高,而数据分布较为稀疏,

因此预处理阶段需要去除相关性较低的属性以提高挖掘速度或准确度。此外,在分类预测任务中,健康医疗数据往往面临着样本类别分布不平衡的问题。直接利用不平衡样本进行数据挖掘,会导致样本数量大的类别预测准确率较高、样本数量小的类别预测准确率较低。针对这一问题,常用的方法有欠采样、bagging、boosting 等。此外,Wang 等人[63]还采用了 Synthetic Minority Over-sampling 技术和 Cost-Sensitive Classifier 技术。

(3)算法选择困难

挖掘算法是数据挖掘的核心,当应用于健康医疗领域时,研究人员应综合考虑数据集和挖掘任务的特征选择适宜的算法。例如,支持向量机适用于高维数据集,在低维数据集上的表现则相对较弱;决策树和 logistic 回归的结果可解读性较强,虽然有时预测准确度不如神经网络等复杂模型,但常被应用于注重结果解读的健康医疗数据挖掘。

(4)交叉验证与评估指标的选择

结果评估是数据挖掘过程中的一个重要步骤。对于非监督模型,一种评估方法是直观分析,如从每一类中随机抽出样本进行人工评估;另一种是对其内部信息进行分析,如衡量簇内样本点之间的距离、衡量样本到其它簇的距离是否足够远等[64]。对于监督模型,可通过减少大样本类别的样本数得到平衡样本,还可以使用 k 折交叉验证的方法,充分利用每个样本。此外,从评估指标的选择来看,应充分利用混淆矩阵求解灵敏度(sensitivity)和特效性(specificity)。

4 结论与展望

随着数据挖掘技术的发展,其在健康医疗领域的相关研究呈上升趋势。通过回顾国内外相关文献,本文综述了数据挖掘在健康医疗领域的应用与研究进展,同时对主要模型与算法、应用的局限与问题进行了总结。

应用研究方面,数据挖掘不仅能够辅助完成预防、诊断、治疗、预后等医疗任务,并通过辅助门急诊、住院、医疗费用、医疗保险等的管理为医疗资源的合理配置提供参考,还能帮助理解健康信息需求和行为,优化健康信息获取,进而改善健康信息服务。算法方面,目前健康医疗领域的大部分研究仍然采用复杂度相对较低的传统算法。一方面可能因为复杂算法本身存在一定局限,如神经网络等复杂算法虽然往往能得到更好的效果,但也存在过程不易理解、时间和空间复杂度较高等问题;另一方面因为健康医疗领域与数据挖掘领域的研究人员存在知识结构差异,健康医疗领域研究者由于数据挖掘知识相对有限,难以应用前沿的挖掘算法;数据挖掘领域研究者则缺乏相应的健康医疗领域知识,在挖掘结果的解读与价值评估方面存在困难。

未来,可从以下三个方面进一步研究。

(1)数据来源多样化。健康医疗数据的种类较多,医疗过程数据大部分来源于医院信息系统;医学科研数据来自于专门设计的医学研究或疾病监测,数据质量高,具有一定的针对性;自我量化数据主要是用户的体征信息,一般通过可穿戴设备等终端进行采集,具有方便实时的特点;用户生成数据多为文本数据,如健康社区中与医生的互动、社交网络中与病友的交流等,对这些数据的分析更有助于改善“以人为中心”的医疗服务。现有研究大多只针对其中一种数据源进行探讨,也有相关研究结合了公共卫生数据与搜索引擎数据对流感进行预测[65],结果表明搜索数据可反映出传统数据无法预测的流感最新变异趋势,说明不同数据之间包含的信息具有一定程度的互补性,如何挖掘并利用这种数据源之间的信息互补,从而提高准确度与实时性是未来的研究方向之一。

(2)电子病历挖掘语义化。电子病历是患者就医过程的记录,包含大量潜在知识。目前基于电子病历的语义挖掘研究较少,一方面在于电子病历中大部分信息以非结构化的文本形式保存,无法被计算机理解和处理;另一方面在于如果仅是对电子病历进行浅显的数据挖掘,对医疗活动并没有太大帮助。未来,可构建医学术语词表,建立医学实体之间的语义关系,从概念层面进行数据挖掘,为医务人员提供临床决策的辅助和支持。

(3)宏观来看,数据挖掘的基础是数据,未来应夯实健康医疗数据挖掘的应用基础,综合开发利用以居民电子健康档案、电子病历、电子处方等为核心的基

础数据库，加强健康医疗海量数据存储清洗、分析挖掘、安全隐私保护等关键技术攻关。此外，云计算、大数据等技术的发展也为数据挖掘技术的实施提供了便利条件，而人工智能的崛起为数据挖掘的应用指明了方向。

参考文献

1 Fayyad U M, Piatetsky-Shapiro G, Smyth P. From Data Mining to Knowledge Discovery in Databases[J]. AI Magazine, 1996, 17(3):37-54.

2 De-Arteaga M, Eggel I, Jr C E K, et al. Analyzing Medical Image Search Behavior: Semantics and Prediction of Query Results[J]. Journal of Digital Imaging, 2015, 28(5):537-546.

3 Ilayaraja M, Meyyappan T. Efficient Data Mining Method to Predict the Risk of Heart Diseases Through Frequent Itemsets[J]. Procedia Computer Science, 2015, 70:586-592.

4 陶惠，蒋凡. 改进的序列模式挖掘在医院转诊中的应用[J]. 计算机系统应用，2015，24(10):253-258.

5 唐晓琳，余世英，吴江. 基于URL共现分析的医疗健康类网站竞争态势研究[J]. 情报杂志，2016，35(4):98-104.

6 Bouarfa L, Dankelman J. Workflow Mining and Outlier Detection from Clinical Activity Logs[J]. Journal of Biomedical Informatics, 2012, 45(6):1185-1190.

7 邵笑笑. 基于医保费用的分析与异常检测研究[D]. 成都:电子科技大学，2016.

8 王宇燕，王杜娟，王延章，等. 改进随机森林的集成分类方法预测结直肠癌存活性[J]. 管理科学，2017，30(1):95-106.

9 Sharma N, Om H. Data Mining Models for Predicting Oral Cancer Survivability[J]. Network Modeling Analysis in Health Informatics & Bioinformatics, 2013, 2(4):285-295.

10 Xu W, Han Z, Ma J. A Neural Network Based Approach to Detect Influenza Epidemics Using Search Engine Query Data[C]// IEEE, Machine Learning and Cybernetics (ICMLC), Qingdao, 2010:1408-1412.

11 Meng X H, Huang Y X, Rao D P, et al. Comparison of Three Data Mining Models for Predicting Diabetes or Prediabetes by Risk Factors[J]. Kaohsiung Journal of Medical Sciences, 2013, 29(2):93-99.

12 吴生根，陈武，欧剑鸣，等. 福建省2008—2013年手足口病重症病例危险因素的分类决策树分析[J]. 中国卫生统计，2015，32(6):1040-1041.

13 任仙龙，胡冬梅，王文娟，等. 关联规则在社区居民慢性病患病率分析中的应用[J]. 中国卫生统计，2013，30(6):818-820.

14 张晔，张晗，尹玢璨，等. 基于电子病历利用支持向量机构建疾病预测模型——以重度急性胰腺炎早期预警为例[J]. 现代图书情报技术，2016，32(2):83-89.

15 Yang H F, Chen Y P. Data Mining in Lung Cancer Pathologic Staging Diagnosis: Correlation Between Clinical and Pathology Information[J]. Expert Systems with Applications, 2015, 42(15-16):6168-6176.

16 Zubi Z S, Saad R A. Using Some Data Mining Techniques for Early Diagnosis of Lung Cancer[C]// World Scientific and Engineering Academy and Society (WSEAS), International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases. Cambridge UK, 2011:32-37.

17 Villamil M D P, Barrera D, Velasco N, et al. Strategies for The Quality Assessment of The Health Care Service Providers in The Treatment of Gastric Cancer in Colombia[J]. Bmc Health Services Research, 2017, 17(1):654-669.

18 Auconi P, Scazzocchio M, Cozza P, et al. Prediction of Class III Treatment Outcomes Through Orthodontic Data Mining[J]. European Journal of Orthodontics, 2015, 37(3):257-267.

19 张奇，李涛，许勇钢，等. 基于关联规则挖掘治疗多发性硬化所用中药对患者T细胞亚群的影响[J]. 中国中西医结合杂志，2016，36(4):425-429.

20 丁心香，王爱国，郑昆仑，等. 基于无监督数据挖掘中药内服治疗颈性眩晕的组方用药规律分析[J]. 中国中药杂志，2016，41(5):955-959.

21 曹锦梅，凌灿，赵小龙，等. 基于关联规则分析治疗2型糖尿病临床用药规律[J]. 西南师范大学学报(自然科学版)，2013，38(10):82-87.

22 侯婷. 数据挖掘在蛋白质翻译后修饰及疾病诊断和预后中的应用[D]. 上海:华东理工大学，2017.

23 Kim S, Park K, Kim M S, et al. Data-Mining for Detecting Signals of Adverse Drug Reactions of Fluoxetine Using the Korea Adverse Event Reporting System (KAERS) Database[J]. Psychiatry Research, 2017, 256:237-242.

24 Yang H D, Yang C C. Using Health Consumer Contributed Data to Detect Adverse Drug Reactions by Association Mining with Temporal Analysis[J]. Acm Transactions on Intelligent Systems & Technology, 2015, 6(4):1-30.

25 Xu M, Guo J C, Zhang J, et al. Protein-Coding Genes, Long Non-Coding RNAs Combined with MicroRNAs as a Novel Clinical Multi-Dimension Transcriptome Signature to Predict Prognosis in Ovarian Cancer[J]. Oncotarget, 2017, 8(42):72847-72859.

26 Bai Z G, Shi E H, Wang Q W, et al. A Potential Panel of Two-Long Non-Coding RNA Signature to Predict Recurrence of Patients with Laryngeal Cancer[J]. Oncotarget, 2017, 8(41): 69641-69650.

27 胡敏，王鹏，于京杰. 基于移动互联网和数据挖掘技术的门诊排队流程设计[J]. 医学研究生学报，2015(2):192-194.

28 陈勇. 基于数据挖掘技术的门诊医疗管理研究[D]. 天津:河北工业大学，2015.

29 彭金燕，张大亮，孙飞超. 基于医患互动的患者分类及管理策略研究[J]. 南京医科大学学报(社会科学版)，2012，12(3):190-193.

30,49 Lin W T, Wu Y C, Zheng J S, et al. Analysis by Data Mining in The Emergency Medicine Triage Database at A Taiwanese Regional Hospital[J]. Expert Systems with Applications, 2011, 38(9):11078-11084.

31 Talbert D A, Honeycutt M, Talbert S. A Machine Learning and Data Mining Framework to Enable Evolutionary Improvement in Trauma Triage[C]// MLDM, Machine Learning and Data Mining in Pattern Recognition, New York USA, 2011:348-361.

32 Khajehali N, Alizadeh S. Extract Critical Factors Affecting the Length of Hospital Stay of Pneumonia Patient by Data Mining (Case Study: an Iranian Hospital) [J]. Artificial Intelligence in Medicine, 2017,83:2-13.

33 黄东瑾，谢玲珠，郑仰纯,等. 基于病案首页数据挖掘的老年糖尿病患者住院日影响因素分析[J]. 广东医学，2016，37(13):1952-1956.

34 Xie Y, Schreier G, Hoy M, et al. Analyzing Health Insurance Claims on Different Timescales to Predict Days in Hospital[J]. Journal of Biomedical Informatics, 2016, 60:187-196.

35 张晔，张晗，尹玢璨,等. 基于支持向量方法机构建急性胰腺炎患者住院天数预测模型[J]. 医学信息学杂志，2016，37(2):57-60.

36 Hachesu P R, Ahmadi M, Alizadeh S, et al. Use of Data Mining Techniques to Determine and Predict Length of Stay of Cardiac Patients[J]. Healthcare Informatics Research, 2013, 19(2): 121-129.

37 郭慧敏，杜军，黄路非. 基于 R 的 Apriori 算法在高额住院费用中的应用研究[J]. 中国卫生统计，2017，34(2):315-317.

38 张凯. 数据挖掘技术在医疗费用数据中的应用研究[D]. 北京:北京邮电大学，2015.

39 韩晓梅，刘志云，阿布都沙拉木·依米提,等. 基于 DRGs 的卵巢癌患者住院费用分析[J]. 中国卫生统计，2016，33(2):298-300.

40 薛允莲. logistic 回归结合决策树技术在冠心病患者住院费用组合分析中的应用[J]. 中国卫生统计，2015，32(6):988-989.

41 Wang J, Li M, Hu Y T, et al. Comparison of Hospital Charge Prediction Models for Gastric Cancer Patients: Neural Network vs. Decision Tree Models[J]. Bmc Health Services Research, 2009, 9(1):161-166.

42 Kirlidog M, Asuk C. A Fraud Detection Approach with Data Mining in Health Insurance[J]. Procedia - Social and Behavioral Sciences, 2012, 62:989-994.

43 Kose I, Gokturk M, Kilic K. An Interactive Machine-Learning-Based Electronic Fraud and Abuse Detection System in Healthcare Insurance[J]. Applied Soft Computing, 2015, 36:283-299.

44 陶惠. 数据挖掘技术在医保中的研究与应用[D]. 合肥:中国科学技术大学，2015.

45 Ramos M I, Cubillas J J, Feito F R. Improvement of the Prediction of Drugs Demand Using Spatial Data Mining Tools[J]. Journal of Medical Systems, 2016, 40(1):6.

46 Koyuncugil A S, Ozgulbas N. Donor Research and Matching System Based on Data Mining in Organ Transplantation[J]. Journal of Medical Systems, 2010, 34(3):251.

47 郭浩. 医疗保险数据分析和应用研究[D]. 合肥:中国科学技术大学，2016.

48 陶惠，蒋凡. 改进的序列模式挖掘在医院转诊中的应用[J]. 计算机系统应用，2015，24(10):253-258.

50 沈丽宁. 国外健康信息服务现状扫描及启示[J]. 医学信息学杂志,2010,31(06):38-40,51.

51 金碧漪,许鑫. 社会化问答社区中糖尿病健康信息的需求分析[J]. 中华医学图书情报杂志,2014,23(12):37-42.

52 Falotico R, Liberati C, Zappa P. Identifying Oncological Patient Information Needs to Improve e‐Health Communication: a preliminary text‐mining analysis[J]. Quality & Reliability Engineering International, 2015, 31(7):1115-1126.

53,60 Ku Y C, Chiu C C, Zhang Y L, et al. Text Mining Self-Disclosing Health Information for Public Health Service[J]. Journal of the Association for Information Science & Technology, 2014, 65(5):928-947.

54 Oh H S, Park H A. Decision Tree Model of the Treatment-Seeking Behaviors among Korean Cancer Patients[J]. Cancer Nursing, 2004, 27(4):259-266.

（下转第 9 页）

工作和情报学自产生到近几十年的发展过程中，基本上都是作为后勤服务保障性系统出现的。大数据环境带来了巨大改变，情报学善于捕捉、处理和利用数据的传统将使它在整个人文社会科学研究中发挥引领作用，甚至在未来一段时间内，可能会对社会各行各业带来重要影响，因此强调重视情报学领域复合型、交叉型人才的培养。但从历史经验来看，新的环境在带来机遇的同时，也必然会存在各种挑战和困难。对于情报学研究而言，获取信息的价值是信息增值的核心过程，而问题引导才是从大数据中提炼价值的核心。情报学者在解决自身学科问题基础上，可以展开跨学科跨领域研究，在更广阔的背景下进一步促进情报学的发展。

参考文献

1,6 马费成.推进大数据、人工智能等信息技术与人文社会科学研究深度融合[N].光明日报,2018-07-29(06).

2 李广建,化柏林.大数据分析与情报分析关系辨析[J].中国图书馆学报,2014,40(5):14-22.

3 胡易容,张克.从"数字化生存"到"符号的栖居"——论数字人文学的符号学界面[J].华南师范大学学报(社会科学版),2016(2):31-36.

4 马费成.在改变中探索和创新[J].情报科学,2018,36(1):3-4.

5 郝龙,李凤翔.社会科学大数据计算——大数据时代计算社会科学的核心议题[J].图书馆学研究,2017(22):20-29,35.

7 程学旗,靳小龙,王元卓,等.大数据系统和分析技术综述[J].软件学报,2014,25(9):1889-1908.

8 彭宇,庞景月,刘大同,等.大数据:内涵、技术体系与展望[J].电子测量与仪器学报,2015,29(04):469-482.

(收稿日期:2018-08-13)

(上接第 123 页)

55 Oh H S, Park H A. Decision-Tree Model of Treatment-Seeking Behaviors after Detecting Symptoms by Korean Stroke Patients [J]. Journal of Korean Academy of Nursing, 2006, 36(4):662-670.

56 孙丽. 任务类型对网络健康信息搜寻行为的影响及其预测模型研究[D]. 长春:吉林大学, 2015.

57 De-Arteaga M, Eggel I, Jr C E K, et al. Analyzing Medical Image Search Behavior: Semantics and Prediction of Query Results [J]. Journal of Digital Imaging, 2015, 28(5):537-546.

58 Chen X H, Zhang Y, Xing C X, et al. Diabetes-Related Topic Detection in Chinese Health Websites Using Deep Learning[C]// Springer, International Conference on Smart Health. Beijing China, 2014:13-24.

59 罗文馨, 陈翀, 邓思艺. 基于 Word2Vec 及大众健康信息源的疾病关联探测[J]. 现代图书情报技术, 2016, 32(9):78-87.

61 Richardson J E, Ancker J S. Public Perspectives of Mobile Phones' Effects on Healthcare Quality and Medical Data Security and Privacy: A 2-Year Nationwide Survey[C]// AMIA. AMIA Annual Symposium proceedings. San Francisco, 2015: 1076-1082.

62 Gaylin D S, Moiduddin A, Mohamoud S, et al. Public Attitudes about Health Information Technology, and its Relationship to Health Care Quality, Costs, and Privacy[J]. Health Services Research, 2011, 46(3):920-938.

63 Wang K J, Makond B, Wang K M. An Improved Survivability Prognosis of Breast Cancer by Using Sampling and Feature Selection Technique to Solve Imbalanced Patient Classification Data [J]. BMC Medical Informatics & Decision Making, 2013, 13(1):1-14.

64 Liu Y C, Li Z M, Xiong H, et al. Understanding of Internal Clustering Validation Measures[C]// IEEE, International Conference on Data Mining. Sydney, 2011:911-916.

65 王若佳, 李培. 基于互联网搜索数据的流感监测模型比较与优化[J]. 图书情报工作, 2016(18):122-132.

(收稿日期:2018-01-02)